5-1-2014

Development of SRADE tool and analysis of quality scores of the reads of Next-Generation Sequencing data

Chaitanya Krishna Kotha

Development of SRADE tool and analysis of quality scores of the reads of Next-Generation Sequencing data

by

Chaitanya Krishna Kotha

A thesis submitted in partial fulfillment of the of the requirements for the Degree of Master of Science in Bioinformatics

Department of Bioinformatics

College of Science

Rochester Institute of Technology

Rochester, NY

May 1, 2014

<u>**Committee Approval**</u>**:**

Dr. Gary R. Skuse

Professor of Biological Sciences

Thesis Advisor/Chair

Date

Dr. Feng Cui

Assistant Professor

Committee Member

Date

Dr. Gregory Babbitt

Assistant Professor

Committee Member

Date

ACKNOWLEDGEMENTS

I would like to express my gratitude towards my advisor Dr. Gary R. Skuse for his valuable guidance and support in every aspect during the entire course of this study. I am really grateful for his excellent supervision and continuous encouragement in my pursuit of my academic goals.

My deep gratitude to Dr. Feng Cui and Dr. Gregory Babbitt for their insightful suggestions and guidance with their professional knowledge and experience that helped me to complete the study.

I would also like to thank my friends for helping me to pursue my goals, providing me with valuable information and to successfully complete my degree.

Finally, I would like to thank my parents and family members for their unconditional love, humor and support all throughout the years. I dedicate this thesis to them.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Development of SRADE tool and analysis of quality scores of the reads of Next-Generation Sequencing data

Capillary Electrophoresis (CE) based on Sanger sequencing has given the ability to extract and explain the genetic information among any given biological system. Even though it brought about a major breakthrough in the field of biology, it had limitations like speed, throughput, scaling and resolution that paved the way for the invention of new technology named as Next-Generation Sequencing (NGS). With the invention of NGS technology, there has been a lot of insight into the genomes, transcriptomes and epigenomes of many of the species on earth. As time passed by, a lot of information has been generated using the NGS technology and new methods have been developed with each method having its merits and de-merits. Some of the most popular sequencing methods that were developed were Illumina sequencing, 454 pyrosequencing, SOLiD sequencing and Ion Torrent Semiconductor sequencing. All the information generated by these sequencing methods are stored in databases and of all the available databases, one of the most important one is National Center for Biotechnology Information (NCBI) integrated with Sequence Read Archive (SRA).

The sequencing data from the Sequence Read Archive is downloaded through a web interface and converted into the required and useful format using SRA toolkit provided by NCBI. Using the OS Architecture of the SRA toolkit, the data that is stored in '.sra' format is converted into tab delimited text and saved into a text file with '.txt' extension. The data obtained from the files have a lot of redundant information and only a particular data is required for analysis. So, in

order to reduce the redundant information and in order to obtain only the desired data, an algorithm is developed that acts using a User Interface (UI), where the user can select the desired data for analysis. This ensures less computational time, high accuracy and memory efficiency. The User Interface developed is named as SRADE (Sequence Read Archive Data Extractor).

The data obtained from the SRA files have information regarding the sequence reads, quality of the reads, their position and their length that can be used for mapping. The information obtained from different types of sequencing methods may be different and the quality of the reads may be different. Therefore a comparison of the quality of the results developed from multiple runs of the same sequencing method as well as different sequencing methods is done, so as to find the differences, the best method for sequencing the genes and to find a cost effective way to determine the reads with high quality score and low quality score. For the purpose of comparison, a "whole exome sequencing of 1000 Genomes project of Illumina" with data from four runs are being considered along with "1000 Genomes whole exome project of Illumina and AB_SOLiD are being studied.

1. INTRODUCTION

a. Objective

The main objective of this research is to present a bioinformatics framework conceived to analyze the data generated by different Next-Generation sequencing methods in the simplest and the cost efficient way possible.

In order to achieve this goal, Next-Generation Sequencing (NGS) data is obtained from the Sequence Read Archive (SRA) of the National Center for Biotechnology Information (NCBI) database through a web interface. The data downloaded can be viewed with the help of Vdb-view software provided in the SRA toolkit by the National Center for Biotechnology Information (NCBI). Vdb-view software only helps in viewing the data without giving the user, the advantage to copy the data or search for required information.

In order to modify the data, and make it useful for the user, an Operating System architecture provided in the SRA toolkit is used that can convert the data from the file into a useful format. Therefore vdb-dump from the OS architecture of the SRA toolkit is used to convert the file in SRA format into text delimited format. The data obtained from conversion contains lot of redundant and repetitive information that occupies a lot of space in the memory and makes it hard to work with. Therefore a User Interface (UI) is developed to extract only the information required by the user. The information obtained is imported into the R statistical software in order to compare different Next-Generation sequencing methods by generating graphs and maps. The data also contains reads that can be used for mapping with the reference gene using free source toolkits available online like Bowtie and Maq.

a. Overview

Genes play an important role in all the living organisms and in order to get a better understanding of the gene, genomes, transcriptomes and epigenomes several methods have been developed of which the Next-Generation Sequencing (NGS) methods have proven to be the most efficient techniques. Some of the major sequencing methods that generate large amount of data are Illumina sequencing, 454 pyrosequencing, SOLiD sequencing and Ion Torrent semiconductor sequencing.

All the data generated is stored in large databases on the cloud. Of all the databases available, the National Center for Biotechnology Information (NCBI) has a vast amount of data stored in SRA format. The files in SRA format are present in the Sequence Read Archive (SRA) integrated to the National Center for Biotechnology Information (NCBI). Sequence Read Archive (SRA) was developed in association with NCBI, DDBJ and EBI under INSDC (International Nucleotide Sequence Database Collaboration). Chapter 1 gives a brief description of the objectives of the study and the overview of the information available in other chapters.

Chapter 2 provides a description of what is contained in the data, the databases available to download the data and the Sequence Resource Archive (SRA) that is being considered in the present research. The chapter also includes information about the different file formats in which the data can be stored and accessed and also the process of retrieving the data from Sequence Read Archive (SRA). This chapter also provides information about the R statistical environment application used to perform statistical and mathematical operations and to display graphics such as graphs and plots.

The materials and methods including SRA toolkit, SRADE tool and R statistical environment that are considered in the present study are explained in detail in chapter 3. This includes the process of conversion of data in SRA format into tab delimited text format using SRA toolkit, followed by using desktop data extraction application called SRADE (Sequence Read Archive Data Extractor) and finally followed by the sequence of commands used to visualize the useful data obtained from the Next-Generation Sequencing (NGS) files obtained from the National Center for Biotechnology Information (NCBI).

Chapter 4 gives a detailed explanation of the genes considered for the study and the observations of the results obtained by using the materials and methods explained in chapter 3. The samples considered for the study are the data from "whole exome sequencing for the 1000 genomes project" from multiple runs of Illumina and the data from "1000 Genome project whole exome sequencing of GIH population" from Illumina and ABI_SOLiD. All the data is analyzed using details plots and observations.

Chapter 5 provides a detailed discussion of the observations, the reasons that result in the exhibition of the observations and the resulting factors. The final chapter (i.e. chapter 6) presents a list of references that are used for the study.

2. BACKGROUND

a. Next-Generation Sequencing

The invention of Capillary Electrophoresis (CE), which is based on Sanger sequencing has created the opportunity to extract and to analyze the genetic information among any given biological organism and system. Even though it was a major breakthrough in the field of biology, there were several limitations to its application like speed, throughput, scaling and resolution that paved the way for the invention of new technology named as Next-Generation Sequencing (NGS). With the invention of NGS technology, there has been a great insight into the genomes, transcriptomes and epigenomes of all the species on earth. As the time passed by, a lot of information has been generated using the Next-Generation Sequencing (NGS) technology and new methods have been developed with each method having its merits and de-merits. Some of the most popular sequencing methods that were developed were Illumina sequencing, 454 pyrosequencing, SOLiD sequencing and Ion Torrent Semiconductor sequencing. All the information generated by these sequencing methods are stored in databases and of all the available databases, one of the important one is National Center for Biotechnology Information (NCBI), that is integrated with Sequence Read Archive (SRA).

The backend principle of the Next-Generation Sequencing techniques is similar to that of the Capillary Electrophoresis (CE). The principal underlying NGS techniques is:

"Identification of the bases present in the small fragments of DNA by the emission of signals as each of the fragment is re-synthesized from the DNA template strand." (Retrieved from www.illumina.com)

This principle is extended in a parallel way across some millions of reactions, rather than being limited to only a few fragments or even a single DNA. The process involves the fragmentation of genomic DNA sample (gDNA) into smaller segments called ‘reads’, which in turn are mapped to the reference genome using tools. This mapping of the reads to the reference genome is used to determine the sequence of the entire chromosome in a particular genomic DNA sample.

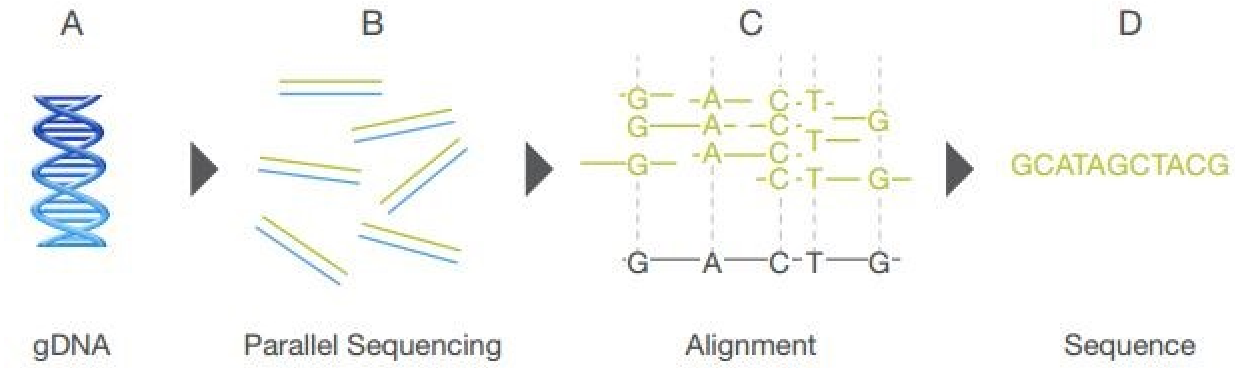


Figure 1: Overview of Whole-Genome Resequencing. (*Illumina, Inc.)*

Figure 1 gives the overview of resequencing of an entire genome where A represents the genomic DNA sample being considered, B represents the strands or fragments of the genomic DNA after fragmentation or parallel sequencing, C represents the alignment or mapping of the reads to the reference genome and finally D represents the regenerated sequence of the entire chromosome.

b. Sequence Read Archive (SRA)

The amount of information obtained from researches all over the globe is rapidly increasing day-to-day. In order to store the data and meet the requirements of the research community to easily access and manipulate the data, Sequence Read Archive (SRA) was developed by National Center for Biotechnology Information (NCBI), European Bioinformatics Institute (EBI) and DNA Data Bank of Japan (DDBJ) under the under the auspices of International Nucleotide Sequence Database Collaboration (INSDC). The major purpose of developing this archive is to store and archive short read data and also to access the metadata stored by the researchers with pin point accuracy.

The information contained in the Sequence Read Archive (SRA) is obtained from some of the major Next Generation Sequencing methods namely:

- Roche 454 pyrosequencing
- Illumina Native
- Illumina SRF
- AB SOLiD Native
- AB SOLiD SRF

The submission of the data can be done in two ways. The first one being the web-based interface which is used by independent research groups for occasional submission of data. The second and the most commonly used one is the automated submission pipeline. The automated submission pipeline is used my most of the centers that generate large amount of data every day with Sequence Read Format (SRF) being the common file format.

The information that is stored in the Sequence Read Archive includes:

- Experimental metadata
- Alignments
- Small scale assemblies

- Reads and quality scores
- Intensity data

Accessing data present in the Sequence Read Archive (SRA) can be done in two ways. The first one is the web – interface used by independent researchers that allows the user to access any data type, reads and the quality scores. The second one is the SRA System Development Kits (SDK) that provides the user with an Application Programming Interfaces (APIs). These APIs allow the user to access and manipulate large data files.

In order to access the files using a web – interface, the user can go into the National Center for Biotechnology Information (NCBI) website and select the Sequence Read Archive (SRA) from the drop down menu that is present to the left of the search column and type the required query. If the user wants to access RNA-seq data files, typing "RNA-seq" in the search column gives the RNA-seq files of all organisms. The 'read' SDK in the SRA development kit allows the user to access the data in SRA format and convert into some of the major short read formats along with providing security by preventing accidental modification of data.

c. The R Project for Statistical Computing

R is a free end software environment and language to perform statistical computing and graphics that are being considered for analysis of the data in this study. It is a part of the GNU operating system project similar to the S language and environment, but with some important differences. This software provides a number of statistical techniques like linear and non-linear modeling, clustering, time-series analysis, classification, classical-statistical tests and others and also graphical techniques and tools to draw and to visualize the data. One of the major strengths of R environment lies in the graphical techniques where a well-designed publication quality plots can be generated along with the mathematical formula and symbols where ever needed.

R software environment is available as a Free software but under the terms and conditions of the Free Software foundation’s GNU’s general public license in the form of source code. One of the major advantages of using R statistical software compared to other software tools available in the market is that R compiles and runs on a vast variety of operating systems like UNIX, Windows and MacOS. It is an integrated suite of graphical display, data manipulation and calculation and includes:

- Facility of effectively store and handle data,
- Suite for calculation on arrays and its matrices,
- Large collection of integrated data analysis tools,
- A well-developed programming language that includes loops user-defined functions, conditions and other facilities.

As said, R is a true programming computer language and therefore it provides the user to add new functions by defining them and can also be linked to other programming languages such as C, C++ and FORTRAN at run-time.

One of the other major advantages of using R environment is that packages can be developed and extended when required. R project provides the user with 8 major packages and all other packages are made available through the CRAN family of the internet sites that cover a vast variety of techniques in modern statistics.

3. MATERIALS AND METHODS

The Next-Generation Sequencing (NGS) data for this study is downloaded using the Sequence Read Archive (SRA) web – interface through the National Center for Biotechnology Information (NCBI) webpage. A particular gene of interest from a species of greater interest is chosen and the data is downloaded. The downloaded file is in SRA format. For the purpose of modifying the data into useful format, SRA toolkit is used. Once the required data is available the following steps are followed to analyze the data. The analysis of the data is discussed under the results and the observation sections.

a. SRA toolkit

In order to access the data that is downloaded in SRA format, a toolkit is provided that helps to access and modify the data. The toolkit can be accessible from the command line prompt or installed form the pre-compiled binaries. The toolkit can also be built from the source code using the SRA development kit (SDK). The most common tools that are contained in the toolkit include:

- fastq-dump → Used to dump SRA files into fastq format
- illumina-dump → Used to dump SRA files into illumina format
- vdb-dump → Used to dump SRA files into vdb format
- sra-dbcc → Used to check database components

The tools provided above are OS architectures and can be accessed using the command line or the terminal. SRA toolkit provides one more tool to view or read the SRA files. The toolkit is named as ‘vdb-view’ and can be installed only on a windows platform. This tool provides the user only the option to view the data contained in the file. All other options like copy or modify are not available within the tool. For such reason, one of the OS architecture needs to be used to convert the data from the files into readable format. Therefore vdb-dump is used in this research to convert the files from SRA format to VDB format. The command line argument for converting the data is as follows:

```
vdb-dump [options] <path/file> [<path/file> ...]
```

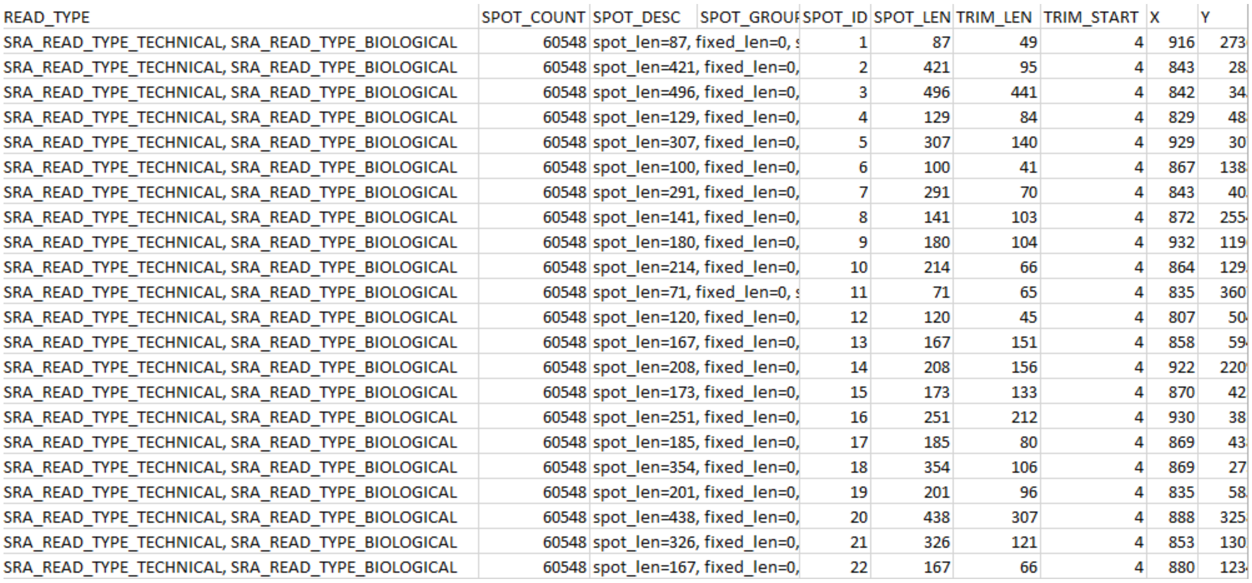
Options:

- *-f | --format (csv, xml, json, piped, tab, fasta, fastq)*
- *-A | --schema_dump*
- *-D | --dna_bases*
- *-X | -- in_hex*

The file can further be converted into the desired format. For the purpose of the present research, the files in SRA format are converted into tab delimited text format. The time taken for converting the data is proportional to the size of the file and to some extent on the destination format. Once the required data is available in readable format, the next step of the research is to analyze the data to obtain proper results.

The major advantage of using these OS architectures is that they convert the data into other formats that can be accessible by most of the other software tools. One major disadvantage of using the OS architectures to convert the data is that they do not generate the column names. In order to determine the column names, the user needs to access the file using vdb-view and write them manually in the converted files.

An example of the file obtained from the conversion of the SRA file using SRA toolkit is given in Figure 2.



READ_TYPE	SPOT_COUNT	SPOT_DESC	SPOT_GROUP	SPOT_ID	SPOT_LEN	TRIM_LEN	TRIM_START	X	Y
SRA_READ_TYPE_TECHNICAL, SRA_READ_TYPE_BIOLOGICAL	60548	spot_len=87, fixed_len=0, s		1	87	49	4	916	273
SRA_READ_TYPE_TECHNICAL, SRA_READ_TYPE_BIOLOGICAL	60548	spot_len=421, fixed_len=0,		2	421	95	4	843	28
SRA_READ_TYPE_TECHNICAL, SRA_READ_TYPE_BIOLOGICAL	60548	spot_len=496, fixed_len=0,		3	496	441	4	842	34
SRA_READ_TYPE_TECHNICAL, SRA_READ_TYPE_BIOLOGICAL	60548	spot_len=129, fixed_len=0,		4	129	84	4	829	48
SRA_READ_TYPE_TECHNICAL, SRA_READ_TYPE_BIOLOGICAL	60548	spot_len=307, fixed_len=0,		5	307	140	4	929	30
SRA_READ_TYPE_TECHNICAL, SRA_READ_TYPE_BIOLOGICAL	60548	spot_len=100, fixed_len=0,		6	100	41	4	867	138
SRA_READ_TYPE_TECHNICAL, SRA_READ_TYPE_BIOLOGICAL	60548	spot_len=291, fixed_len=0,		7	291	70	4	843	40
SRA_READ_TYPE_TECHNICAL, SRA_READ_TYPE_BIOLOGICAL	60548	spot_len=141, fixed_len=0,		8	141	103	4	872	255
SRA_READ_TYPE_TECHNICAL, SRA_READ_TYPE_BIOLOGICAL	60548	spot_len=180, fixed_len=0,		9	180	104	4	932	119
SRA_READ_TYPE_TECHNICAL, SRA_READ_TYPE_BIOLOGICAL	60548	spot_len=214, fixed_len=0,		10	214	66	4	864	129
SRA_READ_TYPE_TECHNICAL, SRA_READ_TYPE_BIOLOGICAL	60548	spot_len=71, fixed_len=0, s		11	71	65	4	835	360
SRA_READ_TYPE_TECHNICAL, SRA_READ_TYPE_BIOLOGICAL	60548	spot_len=120, fixed_len=0,		12	120	45	4	807	50
SRA_READ_TYPE_TECHNICAL, SRA_READ_TYPE_BIOLOGICAL	60548	spot_len=167, fixed_len=0,		13	167	151	4	858	59
SRA_READ_TYPE_TECHNICAL, SRA_READ_TYPE_BIOLOGICAL	60548	spot_len=208, fixed_len=0,		14	208	156	4	922	220
SRA_READ_TYPE_TECHNICAL, SRA_READ_TYPE_BIOLOGICAL	60548	spot_len=173, fixed_len=0,		15	173	133	4	870	42
SRA_READ_TYPE_TECHNICAL, SRA_READ_TYPE_BIOLOGICAL	60548	spot_len=251, fixed_len=0,		16	251	212	4	930	38
SRA_READ_TYPE_TECHNICAL, SRA_READ_TYPE_BIOLOGICAL	60548	spot_len=185, fixed_len=0,		17	185	80	4	869	43
SRA_READ_TYPE_TECHNICAL, SRA_READ_TYPE_BIOLOGICAL	60548	spot_len=354, fixed_len=0,		18	354	106	4	869	27
SRA_READ_TYPE_TECHNICAL, SRA_READ_TYPE_BIOLOGICAL	60548	spot_len=201, fixed_len=0,		19	201	96	4	835	58
SRA_READ_TYPE_TECHNICAL, SRA_READ_TYPE_BIOLOGICAL	60548	spot_len=438, fixed_len=0,		20	438	307	4	888	325
SRA_READ_TYPE_TECHNICAL, SRA_READ_TYPE_BIOLOGICAL	60548	spot_len=326, fixed_len=0,		21	326	121	4	853	130
SRA_READ_TYPE_TECHNICAL, SRA_READ_TYPE_BIOLOGICAL	60548	spot_len=167, fixed_len=0,		22	167	66	4	880	123

Figure 2: Data Sample from Whole Exome Sequencing for the 1000 Genomes Project.

Figure 2 represents an excel worksheet that contains the data obtained from Whole Exome Sequencing for the 1000 Genomes Project using Illumina technique. Here each column represents a single data type and contains an identifier for each member of that data type. Each row represents a cluster within the flow cell used for sequencing. Each flow cell contains several tiles and each tile may contain around a million clusters depending on the type of organism being considered.

The latest Illumina flow cell contains 8 lanes on the flow cell with thousands of tiles. The last columns in the data file namely X and Y (not represented in figure 1) where X represents the X – coordinate of the cluster in the flow cell and Y represents the Y – coordinate of the cluster in the flow cell.

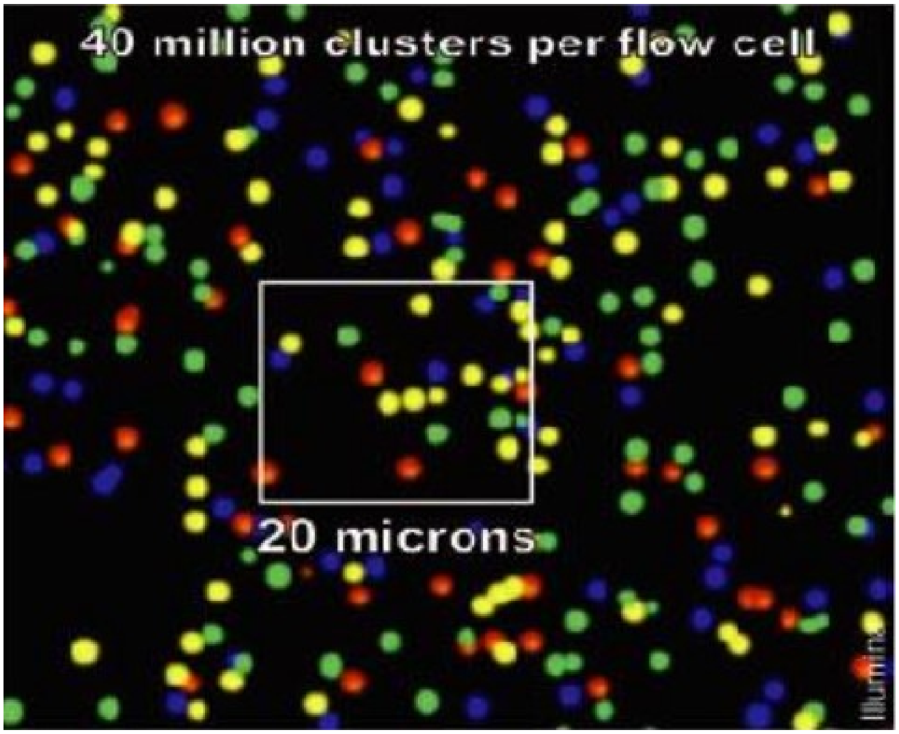


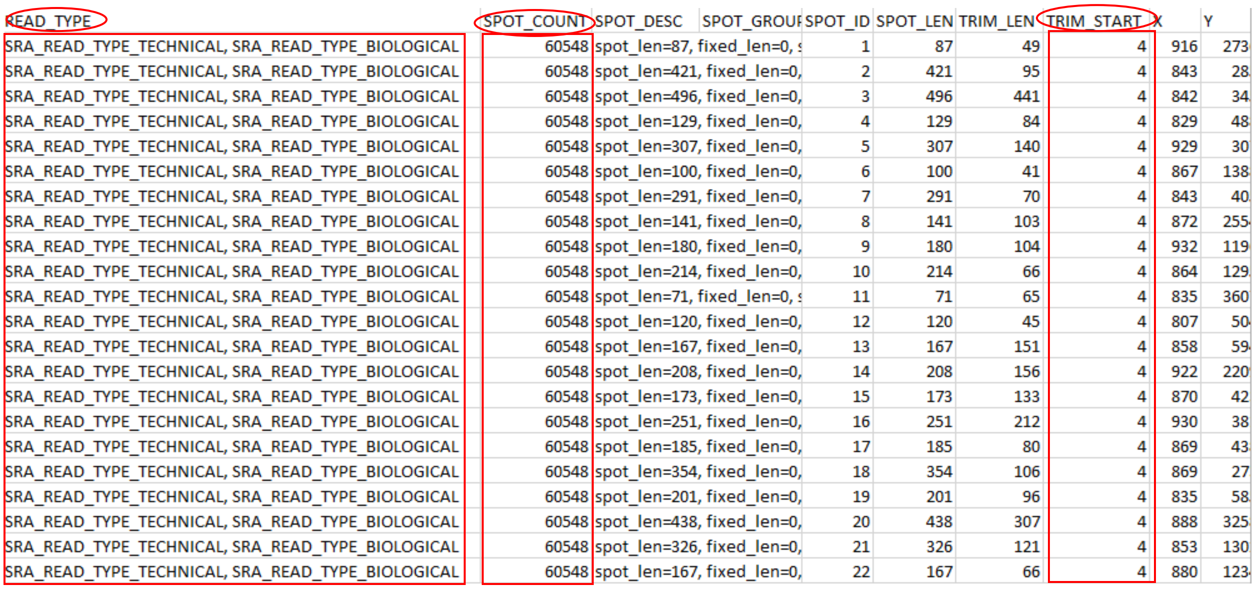
Figure 3: Flow cell of Illumina *(Illumina, Inc.)*

Figure 3 represents a flow cell developed in Illumina where each spot represents a cluster formed within the flow cell and the color represents the type of nucleic acid attached to it. Each nucleic acid is colored with a different fluorescent dye to identify what kind of nucleic acid is being added. The color is observed by laser excitation and the base is identified. In Illumina, red color represents Adenine; yellow color represents Guanine; Blue color represents Cytosine and green color represents Thiamine.

The color of fluorescent dye in reference to the type of base may differ with the type of sequencing technique being used and also some factors depending on the type of experiment being conducted. There can be a maximum of around 40 million clusters per flow cell in Illumina flow cell.

b. SRADE

The data that is obtained by using SRA Toolkit contains lot or redundant information as seen in Figure 4 All the redundant and repeating information is rounded with red border.



READ_TYPE	SPOT_COUNT	SPOT_DESC	SPOT_GROUI	SPOT_ID	SPOT_LEN	TRIM_LEN	TRIM_START	X	Y
SRA_READ_TYPE_TECHNICAL, SRA_READ_TYPE_BIOLOGICAL	60548	spot_len=87, fixed_len=0, s		1	87	49	4	916	273
SRA_READ_TYPE_TECHNICAL, SRA_READ_TYPE_BIOLOGICAL	60548	spot_len=421, fixed_len=0,		2	421	95	4	843	28
SRA_READ_TYPE_TECHNICAL, SRA_READ_TYPE_BIOLOGICAL	60548	spot_len=496, fixed_len=0,		3	496	441	4	842	34
SRA_READ_TYPE_TECHNICAL, SRA_READ_TYPE_BIOLOGICAL	60548	spot_len=129, fixed_len=0,		4	129	84	4	829	48
SRA_READ_TYPE_TECHNICAL, SRA_READ_TYPE_BIOLOGICAL	60548	spot_len=307, fixed_len=0,		5	307	140	4	929	30
SRA_READ_TYPE_TECHNICAL, SRA_READ_TYPE_BIOLOGICAL	60548	spot_len=100, fixed_len=0,		6	100	41	4	867	138
SRA_READ_TYPE_TECHNICAL, SRA_READ_TYPE_BIOLOGICAL	60548	spot_len=291, fixed_len=0,		7	291	70	4	843	40
SRA_READ_TYPE_TECHNICAL, SRA_READ_TYPE_BIOLOGICAL	60548	spot_len=141, fixed_len=0,		8	141	103	4	872	255
SRA_READ_TYPE_TECHNICAL, SRA_READ_TYPE_BIOLOGICAL	60548	spot_len=180, fixed_len=0,		9	180	104	4	932	119
SRA_READ_TYPE_TECHNICAL, SRA_READ_TYPE_BIOLOGICAL	60548	spot_len=214, fixed_len=0,		10	214	66	4	864	129
SRA_READ_TYPE_TECHNICAL, SRA_READ_TYPE_BIOLOGICAL	60548	spot_len=71, fixed_len=0, s		11	71	65	4	835	360
SRA_READ_TYPE_TECHNICAL, SRA_READ_TYPE_BIOLOGICAL	60548	spot_len=120, fixed_len=0,		12	120	45	4	807	50
SRA_READ_TYPE_TECHNICAL, SRA_READ_TYPE_BIOLOGICAL	60548	spot_len=167, fixed_len=0,		13	167	151	4	858	59
SRA_READ_TYPE_TECHNICAL, SRA_READ_TYPE_BIOLOGICAL	60548	spot_len=208, fixed_len=0,		14	208	156	4	922	220
SRA_READ_TYPE_TECHNICAL, SRA_READ_TYPE_BIOLOGICAL	60548	spot_len=173, fixed_len=0,		15	173	133	4	870	42
SRA_READ_TYPE_TECHNICAL, SRA_READ_TYPE_BIOLOGICAL	60548	spot_len=251, fixed_len=0,		16	251	212	4	930	38
SRA_READ_TYPE_TECHNICAL, SRA_READ_TYPE_BIOLOGICAL	60548	spot_len=185, fixed_len=0,		17	185	80	4	869	43
SRA_READ_TYPE_TECHNICAL, SRA_READ_TYPE_BIOLOGICAL	60548	spot_len=354, fixed_len=0,		18	354	106	4	869	27
SRA_READ_TYPE_TECHNICAL, SRA_READ_TYPE_BIOLOGICAL	60548	spot_len=201, fixed_len=0,		19	201	96	4	835	58
SRA_READ_TYPE_TECHNICAL, SRA_READ_TYPE_BIOLOGICAL	60548	spot_len=438, fixed_len=0,		20	438	307	4	888	325
SRA_READ_TYPE_TECHNICAL, SRA_READ_TYPE_BIOLOGICAL	60548	spot_len=326, fixed_len=0,		21	326	121	4	853	130
SRA_READ_TYPE_TECHNICAL, SRA_READ_TYPE_BIOLOGICAL	60548	spot_len=167, fixed_len=0,		22	167	66	4	880	123

Figure 4. Data Sample from Whole Exome Sequencing for the 1000 Genomes Project showing redundant and repeating data.

This redundant data occupies lot of space in the memory and sometimes makes it difficult to view or modify the data because of its large size especially when human genomes are being considered for the study. Data from the human genomes contain millions of reads and therefore they occupy Gigabytes of space in memory which is difficult to open using a normal application like ‘WordPad’ or ‘text document’ or some other Microsoft Word tools. Therefore in order to extract only the required data from the files after conversion into tab delimited text format, I have developed an application named SRADE (Sequence Read Archive Data Extractor) that takes a tab delimited text file as an input.

Once the data is converted into the tab delimited text format using the SRA toolkit, the SRADE tool is used in extension to it so as to extract particular information from the resulting data. It acts as a Graphical User Interface where the user can easily access the files and extract required information without using any commands or tools.

SRADE allows the user to choose specific data required such as the columns and the particular rows of the data. For ease of availability and access of this application, it is made available as a desktop application that can be downloaded from the following web link:

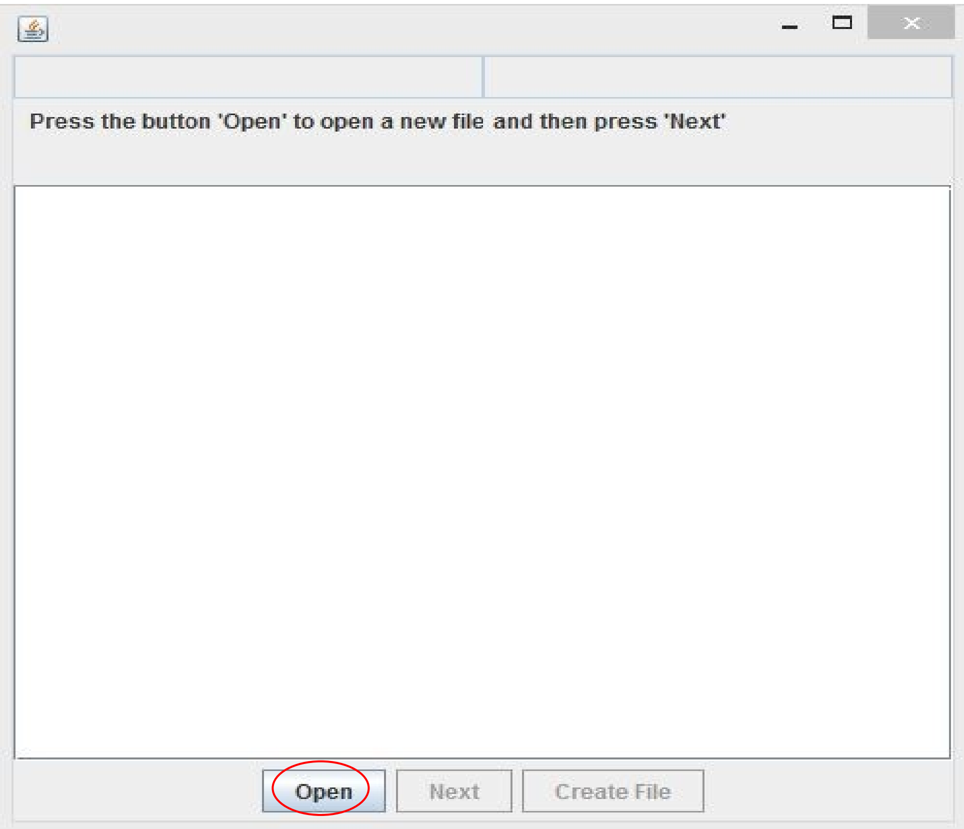
https://sites.google.com/site/ngsdextractor/

The website contain the application along with a sample input file, user manual and documentation of the application. In order to download the application the user has to have an internet connection to download it onto a local computer. The above link has also been posted into the Omic-tools website and can be accessed using the link:

http://omictools.com/educational-resources/text-mining-solutions/data-extractor-s4337.html

The application is developed based on Java 1.7 version. The reason for use of java to code this application is to make the desktop application available on any of the operating system that can run java applications like Windows, MacOS and Linux. The popup menu is displayed in Figure 5. SRADE contains a User Interface (UI) that gives user the flexibility to extract the required information without the use of any internet connection. The application is named as SRADE.jar and can be opened by double-clicking on the icon

.

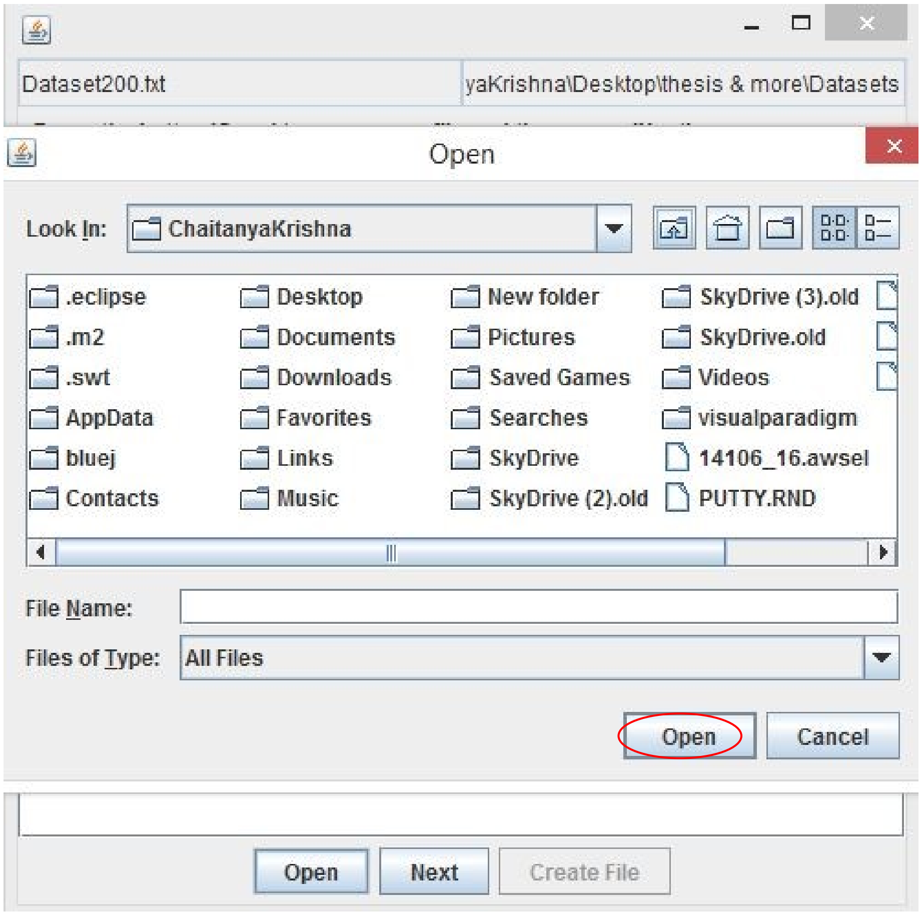


Press the button 'Open' to open a new file and then press 'Next'

Figure 5: Opening screenshot of SRADE

Figure 5 gives the screenshot of SRADE as soon as it is opened. In order to open a new file the “Open” button which appears on the bottom of the frame needs to be clicked which gives a popup menu with the system files where the user can select the desired file from the system memory.

Once the desired file is selected the “Next” button on the bottom of the screen gets activated in order to go to the next step for data extraction. In case if the user made a mistake in selecting the desired file then the button “Open” can be pressed to again show the popup menu.



Dataset200.txt yaKrishna\Desktop\thesis & more\Datasets

Look In: ChaitanyaKrishna

File Name:

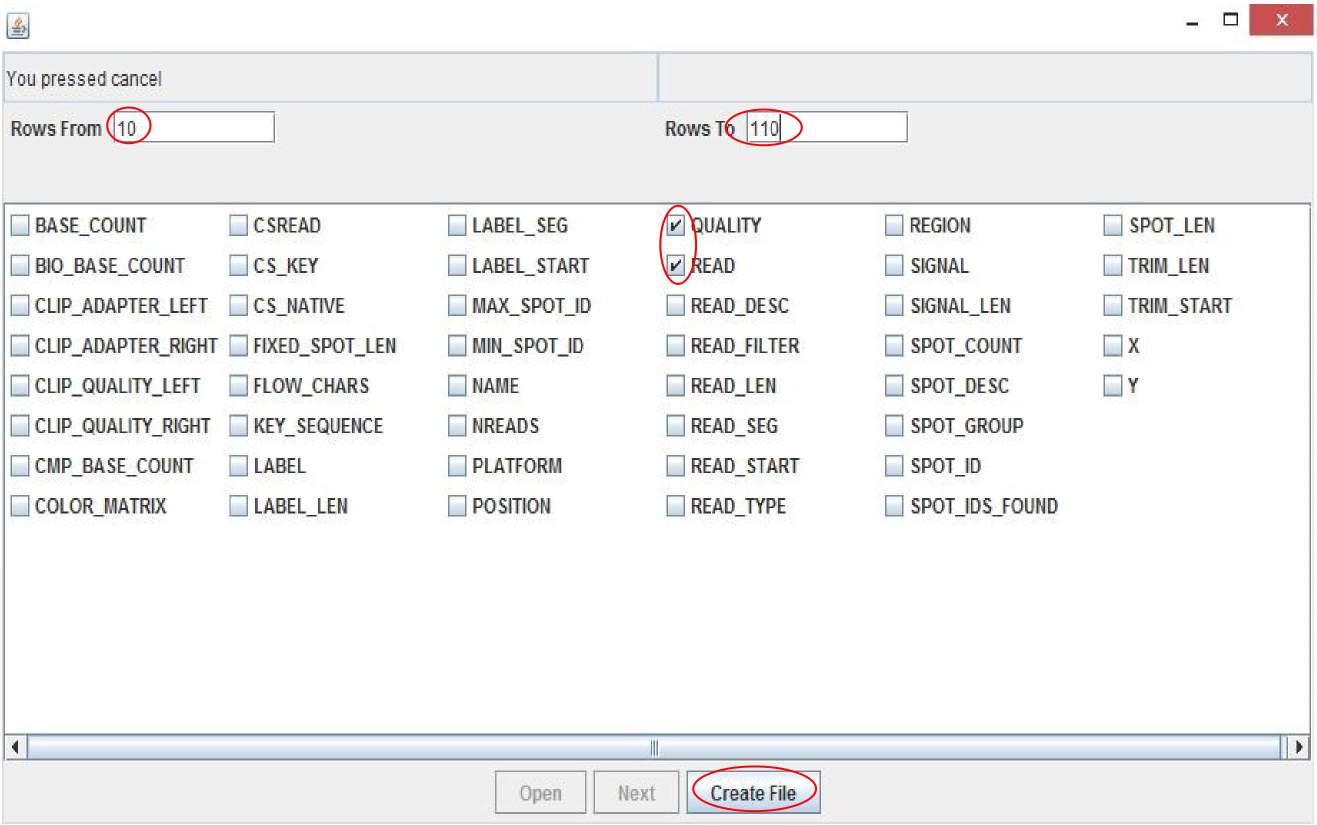
Files of Type: All Files

Figure 6: Screenshot of SRADE with the popup menu for file selection

After pressing "Next" button on bottom of the frame the application lists all the column names with checkboxes so that the user can select the desired column to be written in the output file. On top of the frame there is an option for the user to select the required number of rows in the output file. The rows can be selected in three ways:

- If the user requires all the rows from the input file then both the text boxes can be left blank and the application automatically outputs all the rows from the input files respective to the columns selected.

- If the user requires only a particular number of rows from the starting say 100 rows from starting, then the user can just specify the value 100 in the text field corresponding to the “Rows To” label.
- If the user requires rows from a particular number to a particular number then the user has to specify the starting row number in the text field corresponding to the “Rows From” label and the ending row number in the text field corresponding to the “Rows To” label.



You pressed cancel

Rows From 10

Rows To 110

☐ BASE_COUNT
☐ BIO_BASE_COUNT
☐ CLIP_ADAPTER_LEFT
☐ CLIP_ADAPTER_RIGHT
☐ CLIP_QUALITY_LEFT
☐ CLIP_QUALITY_RIGHT
☐ CMP_BASE_COUNT
☐ COLOR_MATRIX
☐ CSREAD
☐ CS_KEY
☐ CS_NATIVE
☐ FIXED_SPOT_LEN
☐ FLOW_CHARS
☐ KEY_SEQUENCE
☐ LABEL
☐ LABEL_LEN
☐ LABEL_SEG
☐ LABEL_START
☐ MAX_SPOT_ID
☐ MIN_SPOT_ID
☐ NAME
☐ NREADS
☐ PLATFORM
☐ POSITION
☑ QUALITY
☑ READ
☐ READ_DESC
☐ READ_FILTER
☐ READ_LEN
☐ READ_SEG
☐ READ_START
☐ READ_TYPE
☐ REGION
☐ SIGNAL
☐ SIGNAL_LEN
☐ SPOT_COUNT
☐ SPOT_DESC
☐ SPOT_GROUP
☐ SPOT_ID
☐ SPOT_IDS_FOUND
☐ SPOT_LEN
☐ TRIM_LEN
☐ TRIM_START
☐ X
☐ Y

Figure 7: Screenshot of SRADE for selection of required information

The application does not allow non-numeric values being entered in the text filed where the number of rows needs to be specified. In case if the user represents a non-numeric value, then the application gives an error mentioning that non-numeric values are not accepted.

Figure 7 shows the screenshot of SRADE where the desired number of rows can be selected and also the desired columns. If the user does not select any column name from the list, then the application automatically writes out all the columns from the input file. Here, there is no relation between the column names and the row specification. Once the user selects the required rows and columns that are needed in the output file, then on pressing the "Create File" button on the bottom of the frame creates the output file at a destination that is same as the input file and the output file name being "Result_<input file name>.txt". If the output file is successfully written, a popup appears saying that the output file has been successfully saved as; followed by the output file name.

After the output file has been created, the application returns to the starting screen that is shown in Figure 5 where the user can do the same for another input file. If the user does not want to continue with the application at any point of time, the application can be close by clicking on the window close option "x" on the top-right corner of the frame. The application can be minimized by clicking the "–" button on the top-right corner of the frame. The application can also be maximized to fit the entire screen by pressing the square button on the top-right corner of the window frame.

c. Analysis of Data

Once the required data is obtained by using SRADE, the data is analyzed using R statistical environment tool. The main consideration of this study is to determine the use of the information obtained by conversion of the files from SRA format to the tab delimited text format. Some of the useful information contained in the files is:

- Read
- Quality of the read
- Position of the read
- X and Y coordinates of the read on the flowcell

Reads can be used for mapping where the reads obtained from the files can use mapped against a reference gene using mapping tools such as Bowtie, Maq and other of which Bowtie, Bowtie2 and Maq are the free tools available for use. Each of the read has several bases and each base has a quality associated to it. So based on this information the quality of the reads can be used to determine the quality of the Next-Generation Sequencing technique. So for this purpose, analysis of a particular gene from different sequencing platforms is done. To do the analysis a particular gene is considered and the corresponding files from different platforms are taken and the process of conversion into tab delimited text format is done, followed by data extraction with only the quality of the reads being extracted using either the desktop data extraction application.

The resulting output file contains the quality of the reads separated by comma. Now the text file is imported into the R statistical environment in the form of a table using the following command:

mydata = (“input file path and name”, sep = “,”) This command imports the file and stores the data in a tabular format in the variable ‘mydata’. The part of the command “sep” is used to separate each of the quality score corresponding to each of the base in the read into a separate column. In order to display the data the command “*mydata”* can be typed. In order to display the heat map of the data that has been obtained, first the table is converted into a matrix by using the command:

myMatrix = data.matrix (mydata)

Once the data is converted into matrix, the matrix can be displayed using the command “*mydata_matrix”* and heat map can be generated using the command:

mydata_heatmap = heatmap (mydata_matrix, Rowv = NA, Colv = NA, col = heat.colors (256), scale = “column”, margins = c (5, 10))

On typing the command the heat map is displayed and if the user wants to scale by the rows then the scale can be modified. The margins and the color can also be modified according the requirement. Now to obtain the average quality of each of the entire read, the following command is given:

readQuality = rowMeans (mydata)

This gives the mean quality of each of the read and the mean values can be displayed by using the command “*readQuality*”. Now the desired graphs and plots can be drawn using the respective commands. A plot of read quality data can be drawn by using the command:

plot = plot(readQuality)

Similarly, a box-plot can be drawn by using the command:

```
boxplot = boxplot (readQuality)
```

In a similar way other required plots can be developed. Taking into account, the reads with high average quality score, heat maps are generated as to determine the variation in quality scores. This is done by converting the data into matrix in R statistical environment and developing a heat map. The matrix is developed using the command:

```
myMatrix = data.matrix (mydata)
```

The heat map is generated using the command:

```
heatmap <- heatmap (myMatrix, Rowv=NA, Colv=NA, col = heat.colors (256), scale="column",
margins=c (5, 10))
```

By comparing the plots from the samples of genes from different sequencing platforms, a difference can be observed among the plots. The next step of the study is to search for factors that are responsible for the difference in quality of the reads in different sequencing platforms. In order to extract only the required information such as the reads with quality score less than a minimum value or the reads that have high quality scores; an SQL package of R named “sqldf” can be used to write SQL statements to extract the data.

For example if the user wants to get the reads with quality scores less than 20 (quality score < 20) considering the reads are in table named “read_table” and the average quality scores of the reads in other table named “avg_table”, the command to get the desired output is:

```
read_output = sqldf(“SELECT read_table.reads FROM read_table, avg_table WHERE
average < 20”)
```

Here read_table.reads is the column in read_table that contains all the reads and average is the column in avg_table that contains all the average quality score of each of the read corresponding to the read present in read_table. This query is generated by taking into account that the two tables do not have any primary key and foreign key relation. If two tables have columns that are in common, then the appropriate command can be given to get the desired output. It is recommended that the two tables have some common identifier so that the data does not mix up.

Suppose if the read_table has an identifier for each of the read in a column named read_id and the avg_table has a corresponding column named avg_id that represents the quality score of each of the read in read_table, then the command to obtain the reads with quality score less than 20 can be written as:

```
read_output1 = sqldf (“SELECT read_table.reads FROM read_table, avg_table WHERE
read_table.read_id = avg_table.avg_id AND average < 20”)
```

This ensures that there is no error while getting the output. Similar commands can be written to obtain the desired results. The resulting data can be written into files on memory and can be used to purify the reads with low quality.

4. RESULTS & OBSERVATIONS

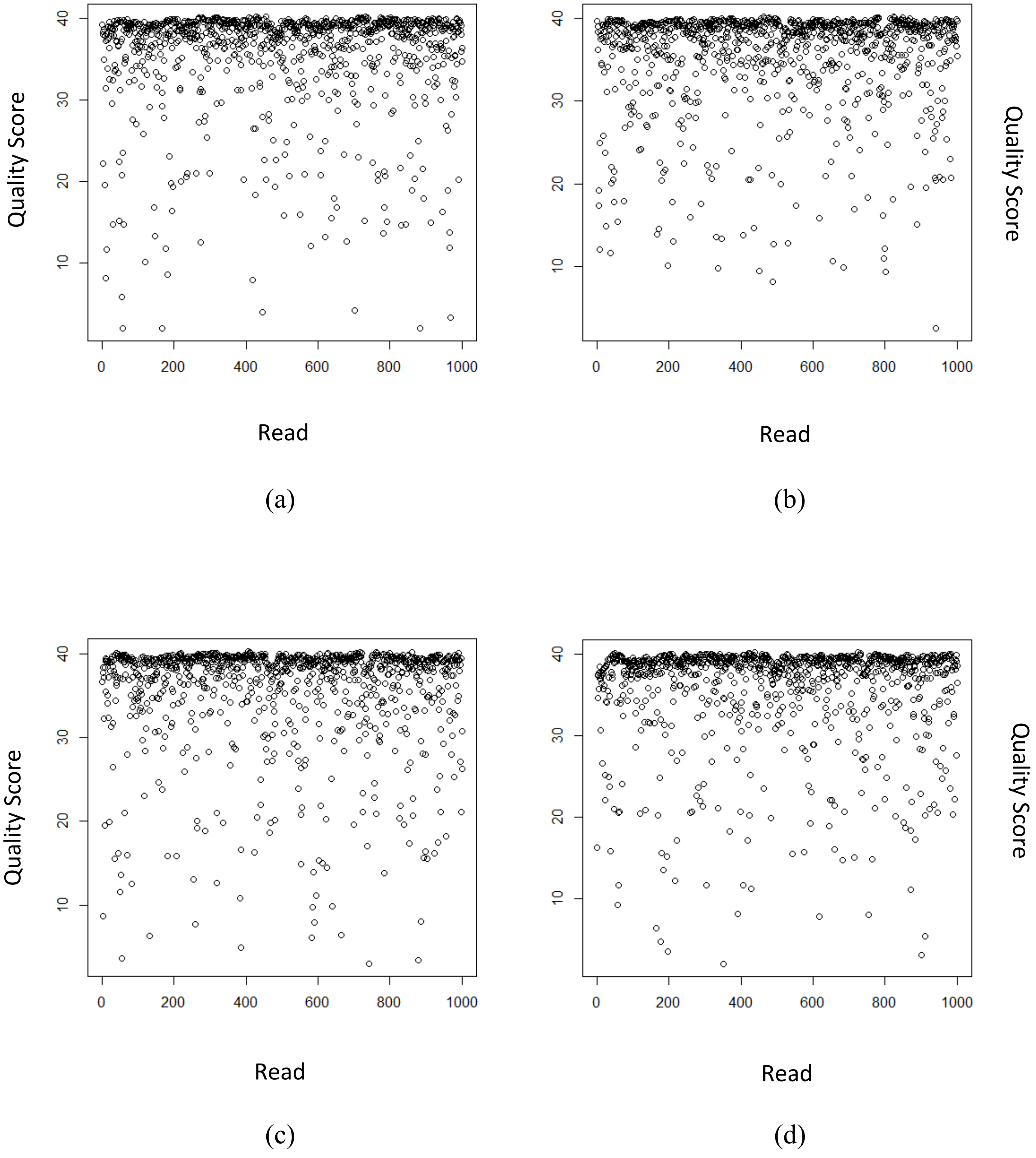
a. Comparison of data from multiple runs of Illumina

For this part of the study, a whole exome sequencing of the 1000 Genomes project is considered with the data from four runs of Illumina sequencing of the same sample is downloaded and processed according to the methods discussed in chapter 3. The experiment design of the sample is “Whole Exome Sequencing for the 1000 Genomes Project via in-solution hybrid selection” with the sample being a generic sample form a normal human being. The instrument model used for sequencing of the sample is Illumina HiSeq 2000. The id’s of the runs being considered are (a) SRR397677, (b) SRR397681, (c) SRR397686 and (d) SRR399240 respectively.

Id	Run	# of Spots	# of Bases
a	SRR397677	2,410,402	366.4M
b	SRR397681	2,442,325	371.2M
c	SRR397686	2,394,221	363.9M
d	SRR399240	2,392,264	363.6M

Table 1: Experimental attribute data of each of the runs of Illumina

Table 1 represents the experiment attribute data of each of the run of illumine for the sample being considered for the study. For the ease of representation in a plot, only 1000 read qualities from the middle of the dataset are being considered and for a box plot first ten thousand reads are being considered.



(a)

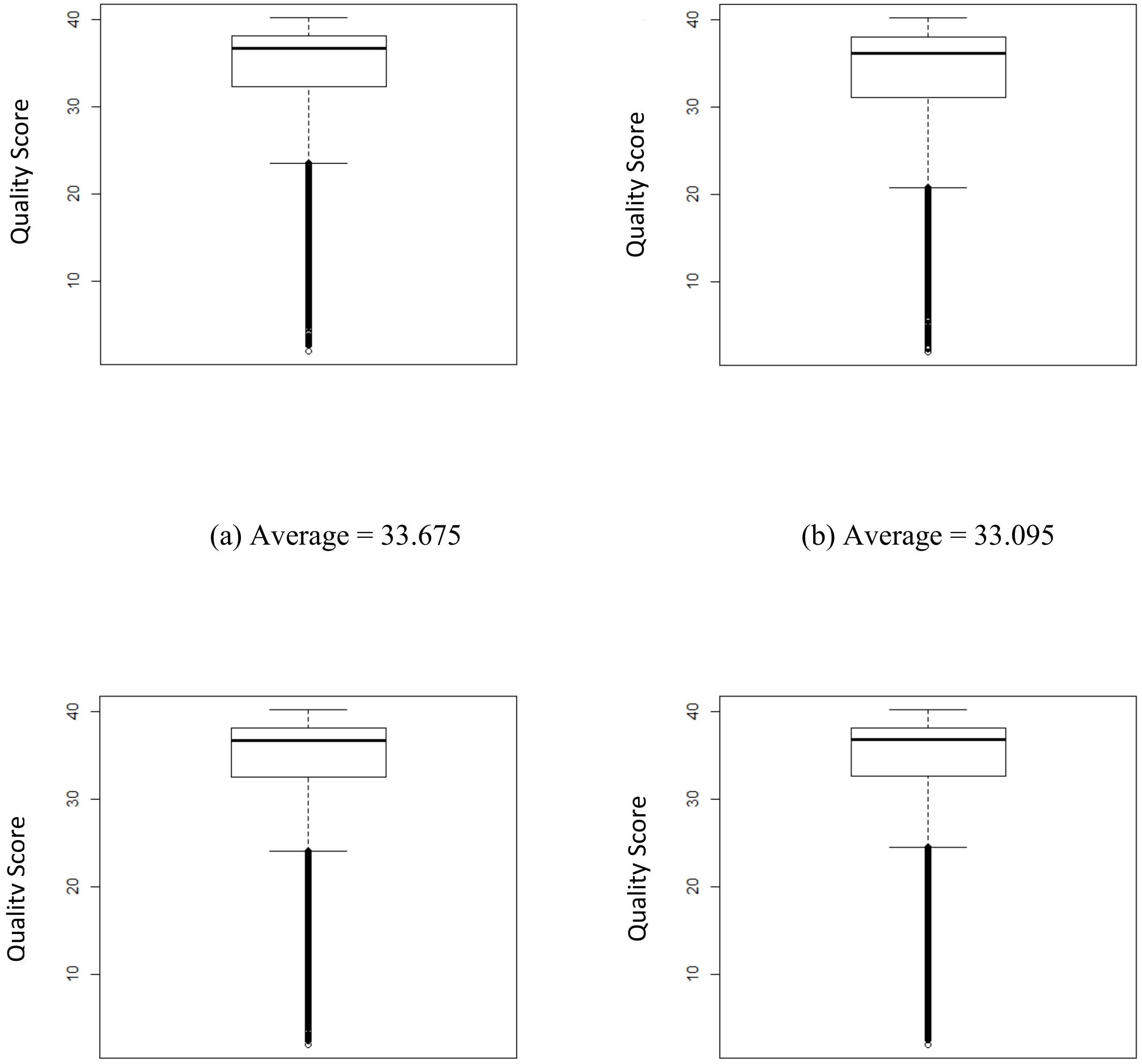
(b)

(c)

(d)

Figure 8: Plot of the quality of reads from different runs of Illumina.

Here the X axis represents the read spot in flow cell and Y axis represents the average quality score of each of the read in the flow cell.



(a) Average = 33.675

(b) Average = 33.095

(d) A

(c) Average = 33.823

(d) Average = 33.734

Figure 9: Box plot of quality of reads from different runs of Illumina

A difference between the plots can be observed where the dots are spread differently in different graphs. This provides some insights into the variation in different runs of sequencing. In the box plot, values on the Y-axis represent the average quality score of the read. There is a very minor difference observed among each of the plot but there exists a difference. The difference is similar when data from 8 Illumina is considered. The prefix number before the name of the sequencing technique represents the number of runs that are done during the sequencing of the sample. Most of the sequencing is done only for a single run but in order to obtain results that remove anomalies, multiple runs are done on the sample.

Observations:

- The first difference observed is that each of the run produces a different result for the total number of spots of being observed in the flow cell.
- The other most important difference is the total number of bases observed during the runs. Each of the runs differs at least by a few hundred bases.
- The reads are not generated in the same order in multiple runs.
- The quality of each of the base in the read is different to the adjacent read and sometime very low quality of one of the base my result in very poor quality of the entire read.
- The quality of the reads in this case do not go greater than 50 which can result in low mapping quality of the reads to the reference genome.
- The length of each of the read is not always the same. Some of the reads have more number of bases compared to the other adjacent reads.
- The plots represent variation in the quality of the reads with dots differentially scattered around the plot without any similarity.
- The box plot shows slight variation in the average quality and the range in which most of the qualities of the reads fall into.
- Few of the runs have large outliers where the quality of score of the read is very low.

b. Comparison of data from different sequencing techniques

For the comparison of the data between two different sequencing techniques, two of the highly used techniques namely Illumia and ABI_SOLiD are considered with the sample being "1000 genomes whole exome sequencing of different populations." Some of the reasons for considering this sample are because of the diversity of the ancestry, similarity in climatic conditions, samples generated from two of the highly used sequencing techniques and others. For consistency in the results and to understand if the results are same for multiple samples, a total of 5 different samples taken from families belonging to different population are considered for this study. The population considered for this study are namely GIH, PEL, ACB, CLM and LWK respectively. A brief description of comparison of each of the sample is done below.

i. Comparison of data from GIH population

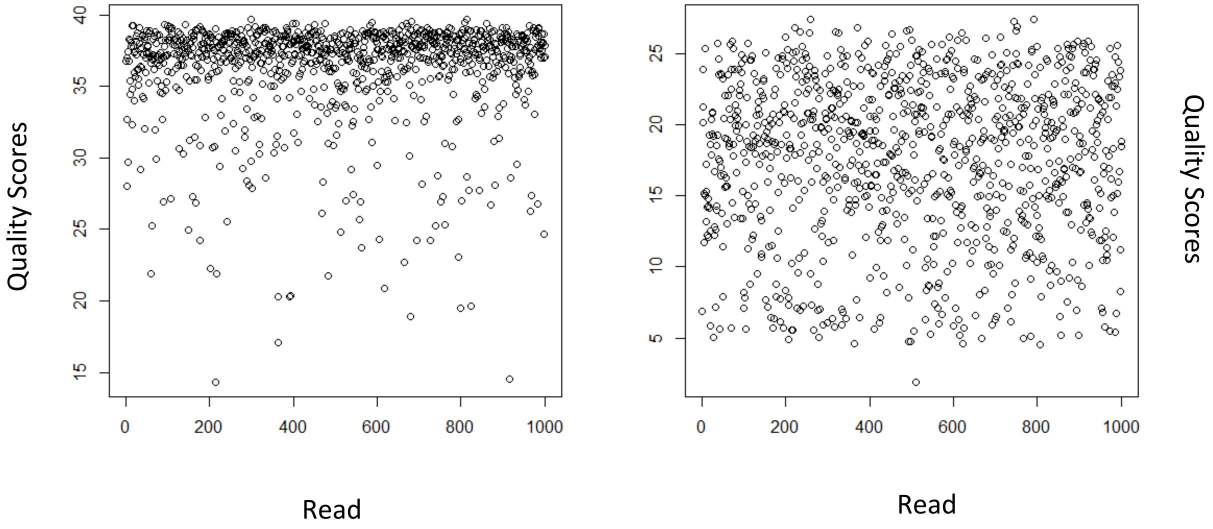
GIH stands for Gujarati Indian from Huston, Texas. The population considered in the study is people form Huston, Texas region with ancestors from Gujarat, India. In order to generate a plot for comparison of the quality of the reads thousand reads from the middle of the dataset are considered and to generate dot plots ten thousand reads from the middle of the dataset are taken into account.

The ID's from the National Center for Biotechnology Information are (a) SRR768537 for the data from one run of Illumina sequencing and (b) SRR398919 for the data from one run of ABI_SOLiD sequencing. Samples for both the sequencing techniques are selected using hybrid selection from the HapMap. The experimental attribute data and the plots of the data are given below:

Id	Run	# of Spots	# of Bases
a	SRR768537	32,920,493	6.6G
b	SRR398919	114,682,480	9.7G

Table 2: Experimental attribute data of Illumina and ABI_SOLiD respectively for GIH population.

Table 2 represents the experimental attribute data of the 1000 genomes whole exome sequencing of GIH population sample of Illumina and ABI_SOLiD respectively.

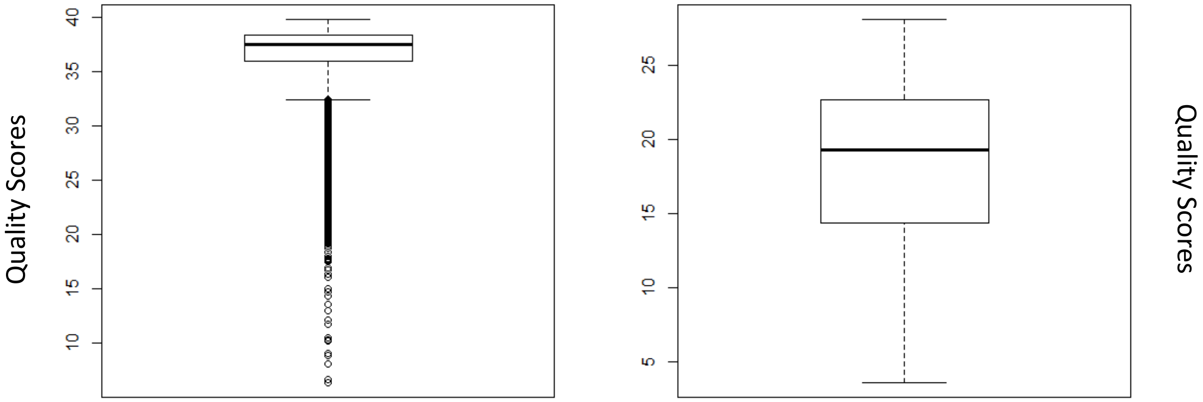


(a)

(b)

Figure 10: Plot of quality of reads from (a) Illumina and (b) ABI_SOLiD for GIH population

Figure 10 represents the plot of quality reads obtained from Illumina and ABI_SOLiD respectively. The X-axis represents the reads and the Y-axis represents the quality scores of each of the respective reads after sequencing. The plots represent a large variation in the distribution of data.



(a)Average = 36.344

(b) Average = 18.227

Figure 11: Box plot of quality of reads of from (a) Illumina and (b) ABI_SOLiD for GIH population

Figure 11 represents the box plots of the quality of reads obtained from Illumina and ABI_SOLiD respectively. The Y-axis represents the average quality score of the reads. The box plots represents that there is a large variation in the quality of the reads obtained from different sequencing methods.

The heat maps are generated for the reads that have the highest average quality. The maps are represented below.

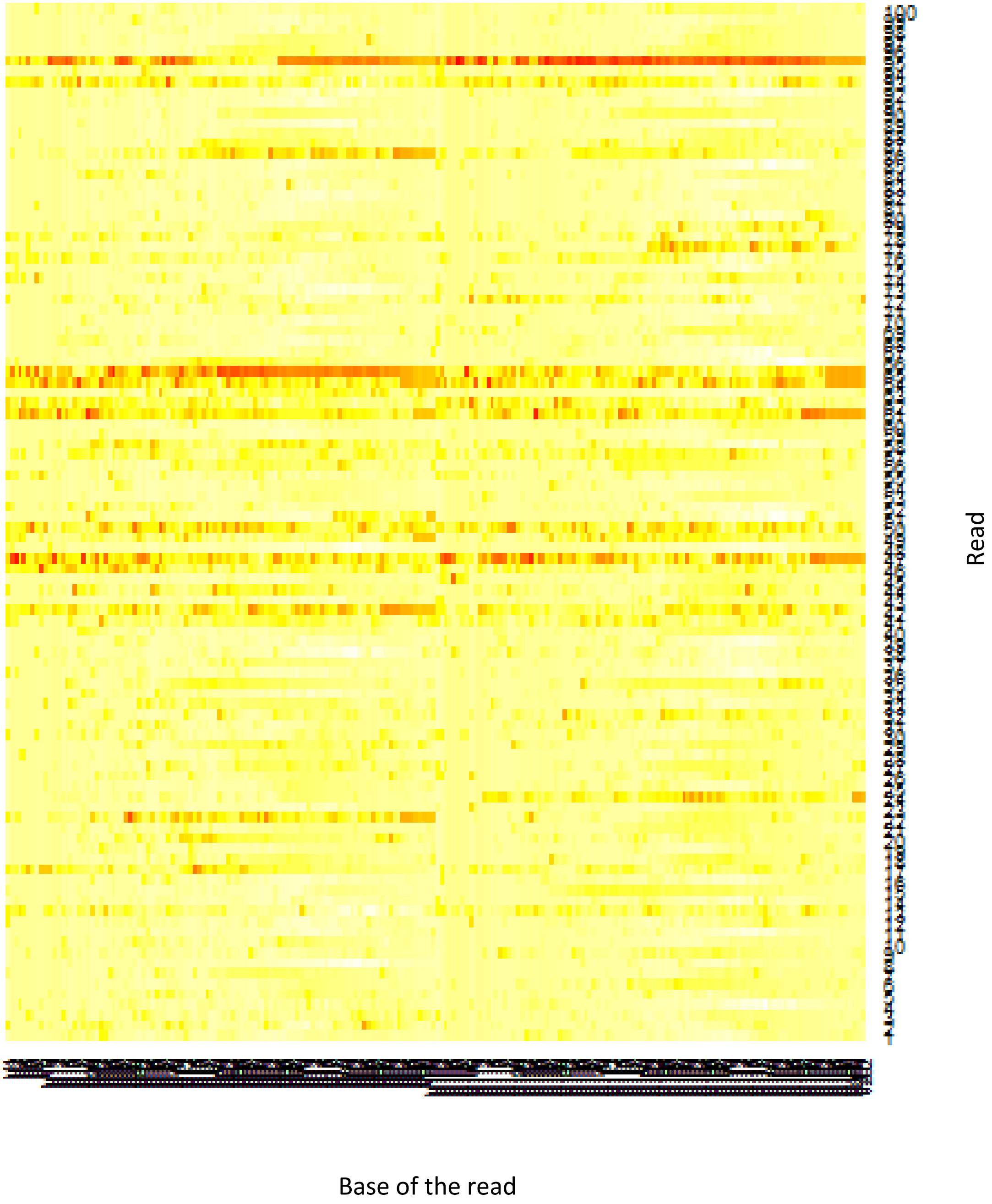


Figure 12: Heat map of reads with high quality score from Illumina sample

In the heat maps X-axis represents each of the individual read and the Y-axis represents the quality of each of the base of a particular read.

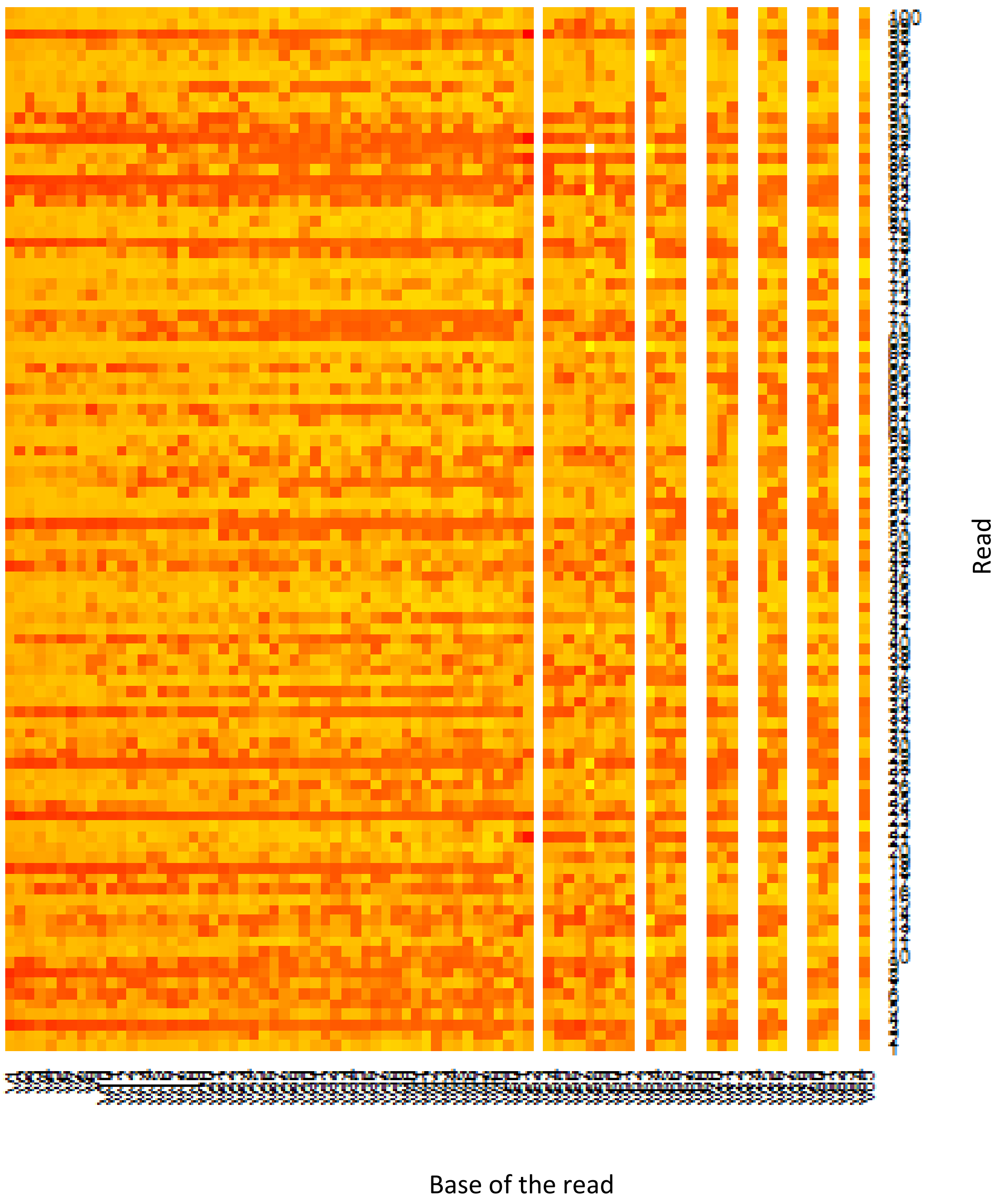


Figure 13: Heat map of reads with high quality score from ABI_SOLiD sample

The heat map from ABI_SOLiD sample has bases that have scores similar to other bases even though the average score of the read is very less compared to the sample generated form Illumina sequencing.

ii. Comparison of data from PEL population

PEL stands for Peruvian in Lima, Peru. The population considered in the study is people form Lima, Peru region with ancestors from Peru. In order to generate a plot for comparison of the quality of the reads thousand reads from the middle of the dataset are considered and to generate dot plots ten thousand reads from the middle of the dataset are taken into account.

The ID's from the National Center for Biotechnology Information are (a) SRR716639 for the data from one run of Illumina sequencing and (b) SRR360369 for the data from one run of ABI_SOLiD sequencing. Samples for both the sequencing techniques are selected using hybrid selection from the HapMap. The experimental attribute data and the plots of the data are given below:

Id	**Run**	**# of Spots**	**# of Bases**
a	SRR716639	22,884,794	4.6G
b	SRR360369	101,542,951	8.6G

Table 3: Experimental attribute data of Illumina and ABI_SOLiD respectively for PEL population

Table 3 represents the experimental attribute data of the 1000 genomes whole exome sequencing of PEL population sample of Illumina and ABI_SOLiD respectively.

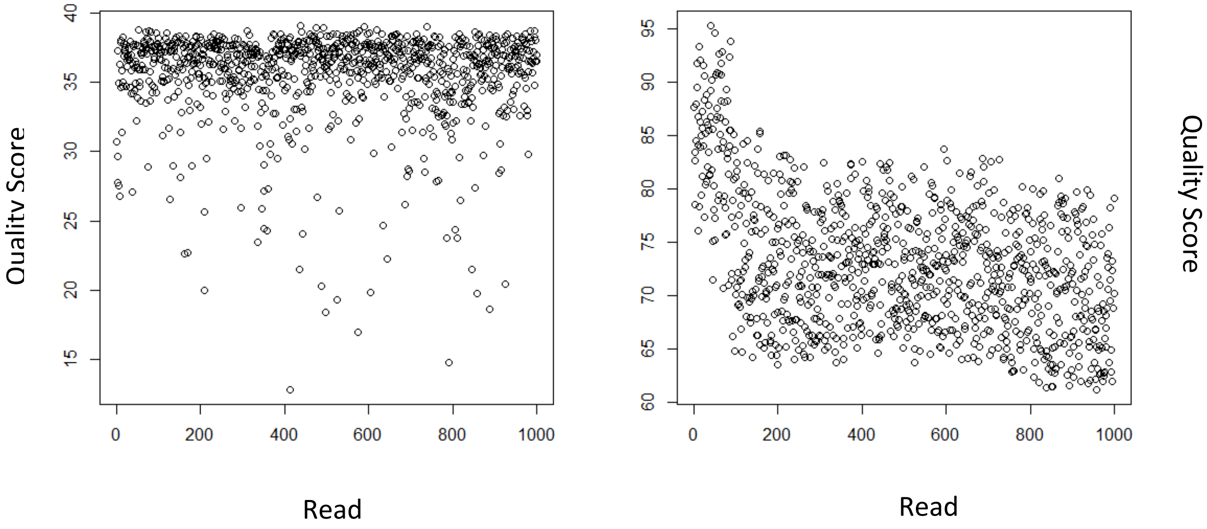
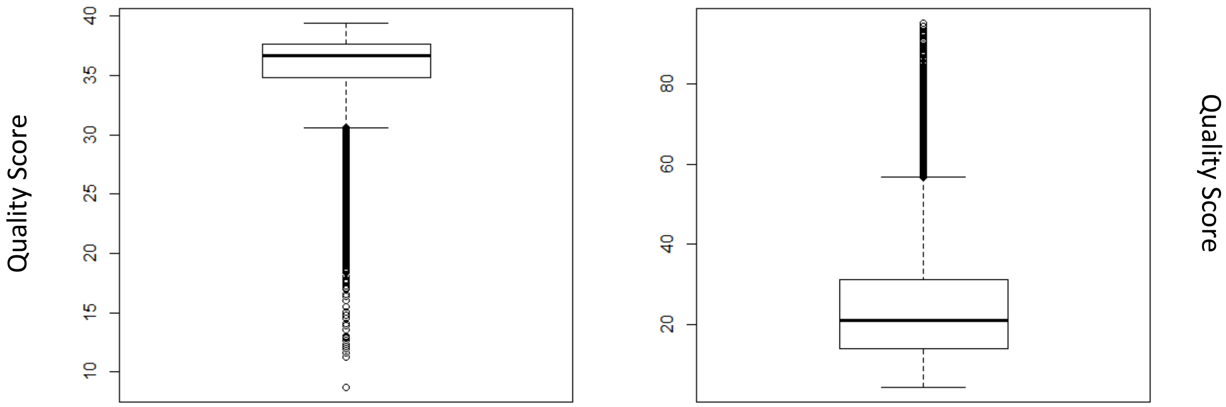


Figure 14: Plot of quality of reads from (a) Illumina and (b) ABI_SOLiD for PEL population



(a)Average = 35.588

(b) Average = 28.646

Figure 15: Box plot of quality of reads of from (a) Illumina and (b) ABI_SOLiD for PEL population

Figure 14 represents the plot of quality reads obtained from Illumina and ABI_SOLiD respectively. The X-axis represents the reads and the Y-axis represents the quality scores of each of the respective reads after sequencing.

Figure 15 represents the box plots of the quality of reads obtained from Illumina and ABI_SOLiD respectively. The Y-axis represents the average quality score of the reads. The box plots represents that there is a large variation in the quality of the reads obtained from different sequencing methods.

iii. Comparison of data from ACB population

ACB stands for African Caribbean in Barbados. The population considered in the study is people form Barbados whose ancestry is from Barbados. In order to generate a plot for comparison of the quality of the reads thousand reads from the middle of the dataset are considered and to generate dot plots ten thousand reads from the middle of the dataset are taken into account.

The ID's from the National Center for Biotechnology Information are (a) SRR710428 for the data from one run of Illumina sequencing and (b) SRR389212 for the data from one run of ABI_SOLiD sequencing. Samples for both the sequencing techniques are selected using hybrid selection from the HapMap. The experimental attribute data and the plots of the data are given below:

Id	Run	# of Spots	# of Bases
a	SRR710428	29,919,752	6G
b	SRR389212	114,960,325	9.8G

Table 4: Experimental attribute data of Illumina and ABI_SOLiD respectively for ACB population

Table 4 represents the experimental attribute data of the 1000 genomes whole exome sequencing of ACB population sample of Illumina and ABI_SOLiD respectively.

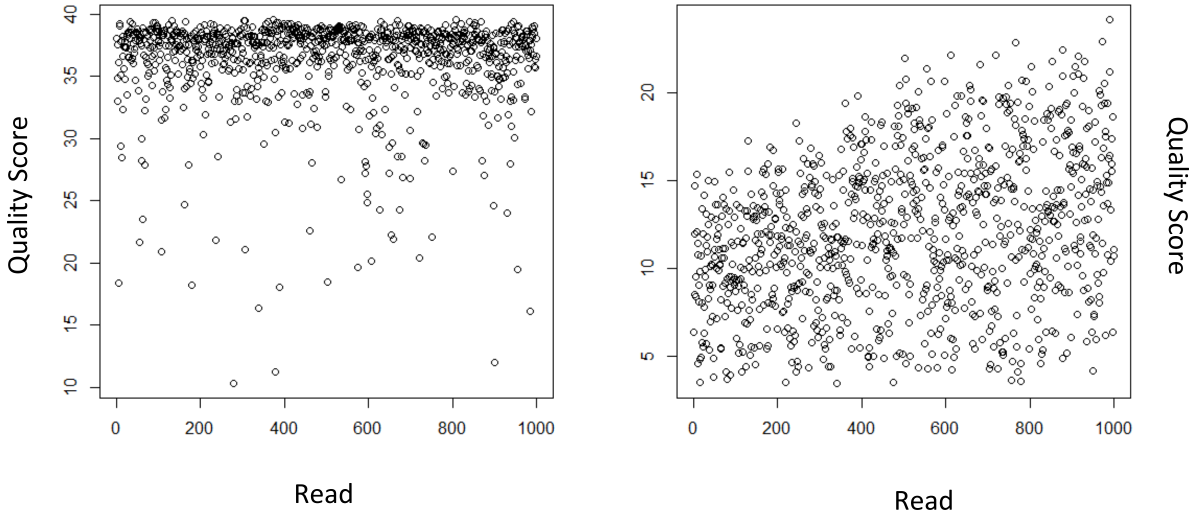
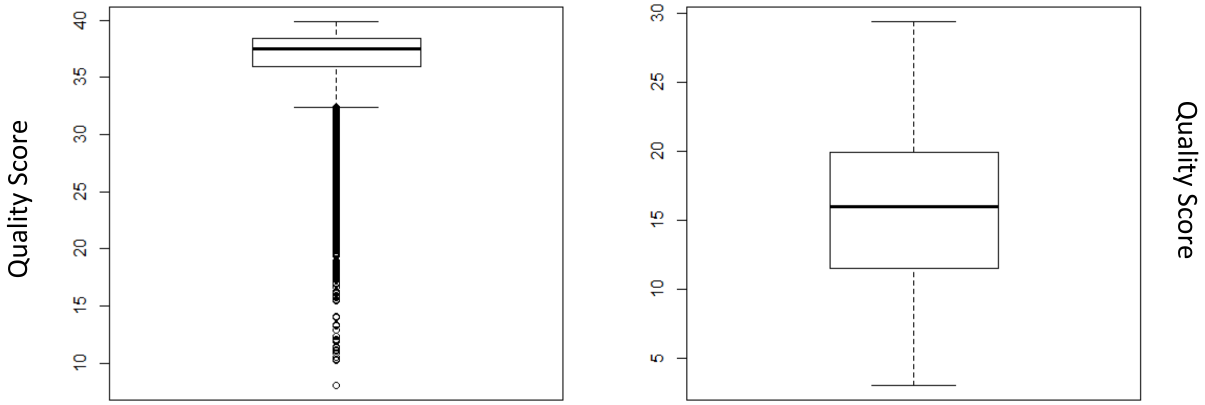


Figure 16: Plot of quality of reads from (a) Illumina and (b) ABI_SOLiD for ACB population

Figure 16 represents the plot of quality reads obtained from Illumina and ABI_SOLiD respectively. The X-axis represents the reads and the Y-axis represents the quality scores of each of the respective reads after sequencing



(a)Average = 36.402

(b) Average = 15.726

Figure 17: Box plot of quality of reads of from (a) Illumina and (b) ABI_SOLiD for ACB population

Figure 17 represents the box plots of the quality of reads obtained from Illumina and ABI_SOLiD respectively. The Y-axis represents the average quality score of the reads. The box plots represents that there is a large variation in the quality of the reads obtained from different sequencing methods.

iv. Comparison of data from CLM population

CLM stands for Colombia from Medellin, Colombia. The population considered in the study is people form Medellin, Colombia whose ancestry is from Colombia. In order to generate a plot for comparison of the quality of the reads thousand reads from the middle of the dataset are considered and to generate dot plots ten thousand reads from the middle of the dataset are taken into account.

The ID's from the National Center for Biotechnology Information are (a) SRR707199 for the data from one run of Illumina sequencing and (b) SRR171729 for the data from one run of ABI_SOLiD sequencing. Samples for both the sequencing techniques are selected using hybrid selection from the HapMap. The experimental attribute data and the plots of the data are given below:

Id	**Run**	**# of Spots**	**# of Bases**
a	SRR707199	34,657,468	7G
b	SRR171729	121,324,022	6.1G

Table 5: Experimental attribute data of Illumina and ABI_SOLiD respectively for CLM population

Table 5 represents the experimental attribute data of the 1000 genomes whole exome sequencing of CLM population sample of Illumina and ABI_SOLiD respectively.

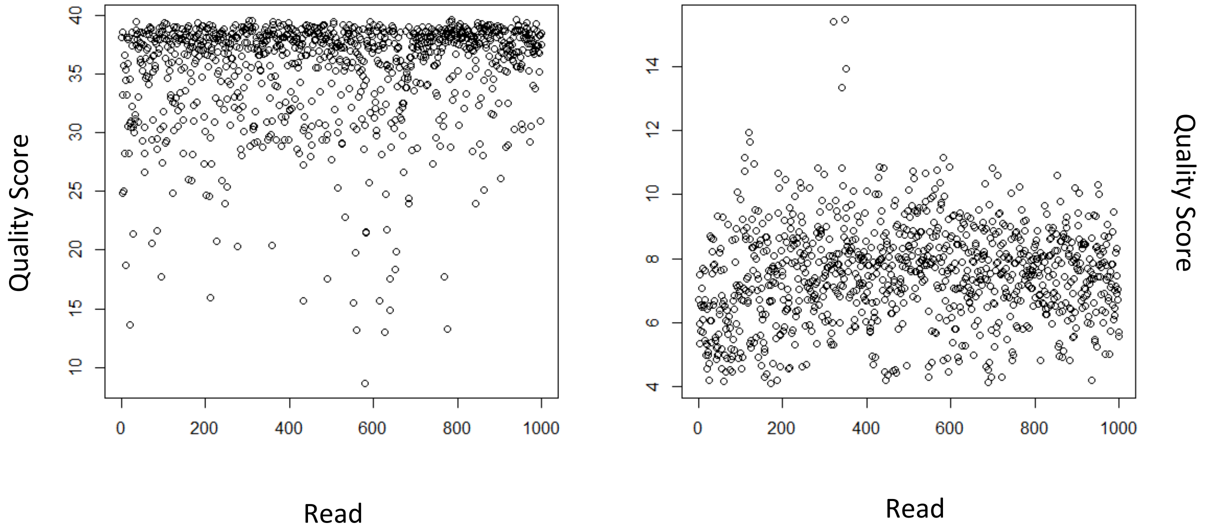
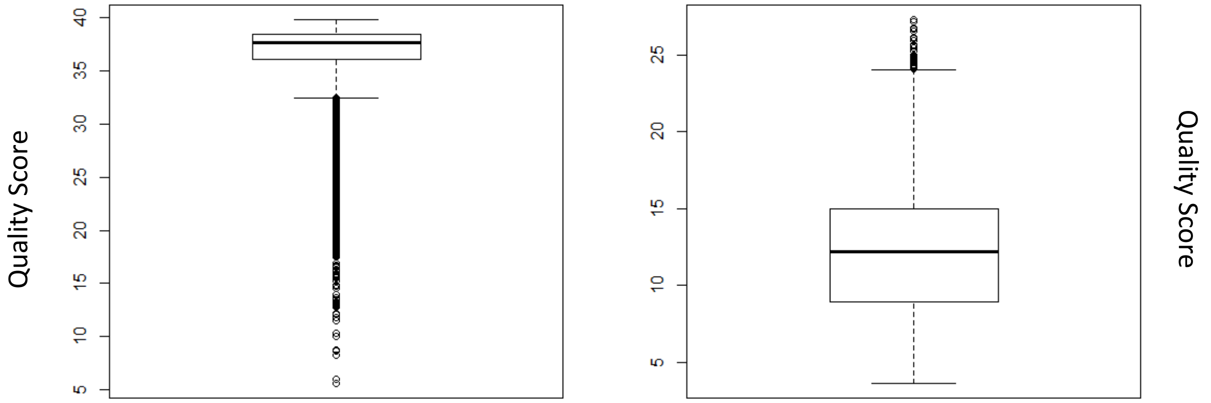


Figure 18: Plot of quality of reads from (a) Illumina and (b) ABI_SOLiD for CLM population



(a)Average = 36.408

(b) Average = 12.291

Figure 19: Box plot of quality of reads of from (a) Illumina and (b) ABI_SOLiD for CLM population

Figure 18 represents the plot of quality reads obtained from Illumina and ABI_SOLiD respectively. The X-axis represents the reads and the Y-axis represents the quality scores of each of the respective reads after sequencing.

Figure 19 represents the box plots of the quality of reads obtained from Illumina and ABI_SOLiD respectively. The Y-axis represents the average quality score of the reads. The box plots represents that there is a large variation in the quality of the reads obtained from different sequencing methods.

v. Comparison of data from LWK population

LWK stands for Luhya in Webuye, Kenya. The population considered in the study is people form Webuye, Kenya whose ancestry is from Luhya. In order to generate a plot for comparison of the quality of the reads thousand reads from the middle of the dataset are considered and to generate dot plots ten thousand reads from the middle of the dataset are taken into account.

The ID's from the National Center for Biotechnology Information are (a) SRR766057 for the data from one run of Illumina sequencing and (b) SRR223504 for the data from one run of ABI_SOLiD sequencing. Samples for both the sequencing techniques are selected using hybrid selection from the HapMap. The experimental attribute data and the plots of the data are given below:

Id	Run	# of Spots	# of Bases
a	SRR766057	24,206,858	4.9G
b	SRR223504	145,164,469	7.3G

Table 6: Experimental attribute data of Illumina and ABI_SOLiD respectively of LWK population

Table 6 represents the experimental attribute data of the 1000 genomes whole exome sequencing of LWK population sample of Illumina and ABI_SOLiD respectively.

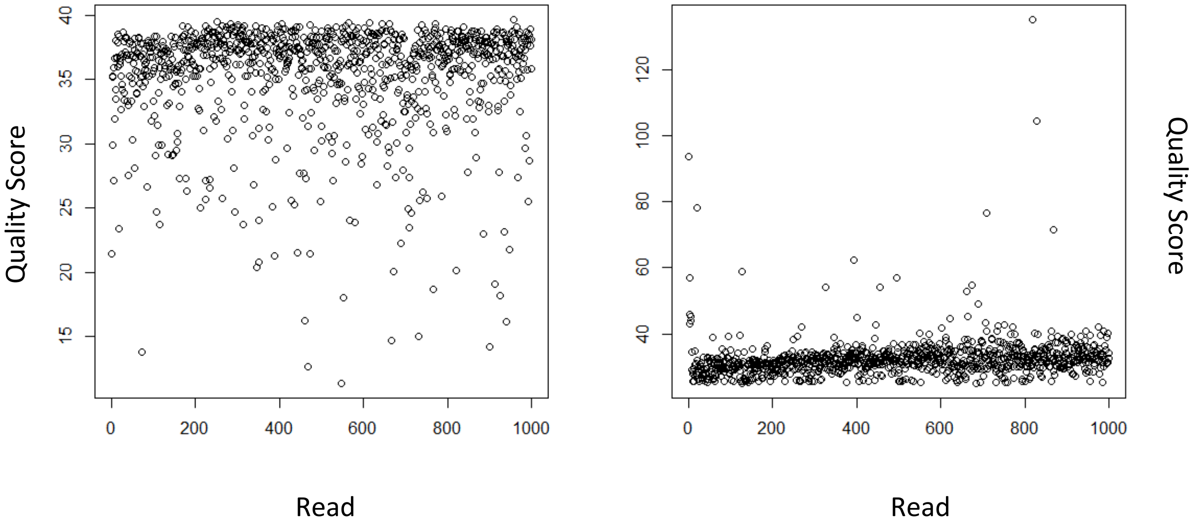
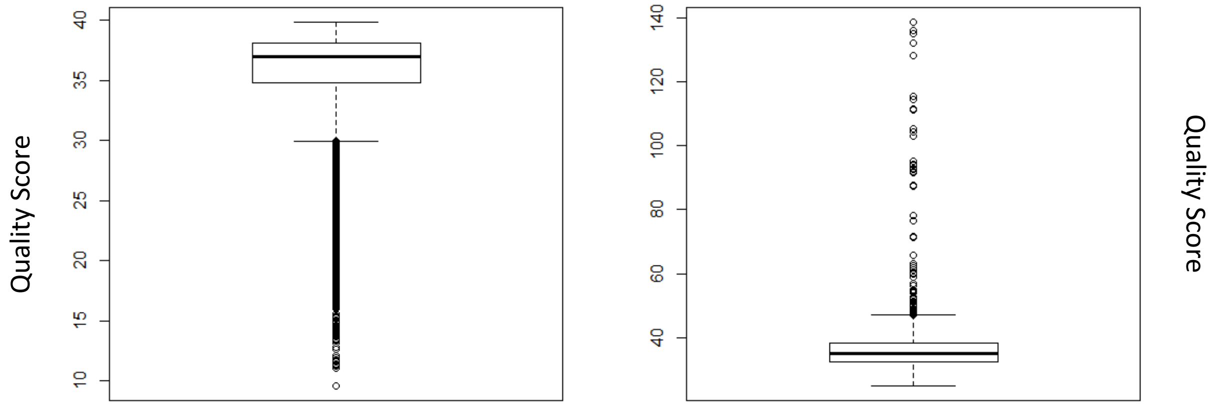


Figure 20: Plot of quality of reads from (a) Illumina and (b) ABI_SOLiD of LWK population

Figure 20 represents the plot of quality reads obtained from Illumina and ABI_SOLiD respectively. The X-axis represents the reads and the Y-axis represents the quality scores of each of the respective reads after sequencing.



(a)Average = 35.551 (b) Average = 35.824

Figure 21: Box plot of quality of reads of from (a) Illumina and (b) ABI_SOLiD of LWK population

Figure 21 represents the box plots of the quality of reads obtained from Illumina and ABI_SOLiD respectively. The Y-axis represents the average quality score of the reads. The box plots represents that there is a large variation in the quality of the reads obtained from different sequencing methods.

Observations:

- The observations are similar to the comparison of multiple runs. The first difference observed is that each of the technique produces a different result for the total number of spots of being observed in the flow cell.

- The other most important difference is the total number of bases observed in both the techniques is very vast.
- The reads are not generated in the same order.
- The quality of each of the base in the read is different to the adjacent read and sometime very low quality of one of the base my result in very poor quality of the entire read.
- The quality of the reads in this case do not go greater than 50 which can result in low mapping quality of the reads to the reference genome.
- The length of each of the read is not always the same. Some of the reads have more number of bases compared to the other adjacent reads.
- The plots represent variation in the quality of the reads with dots differentially scattered around the plot without any similarity.
- The box plot represents a significant variation in the average quality and the range in which most of the qualities of the reads fall into.
- Except for the LWK population samples, all other samples for comparison of two different sequencing platforms, Illumina sequencing seems to produce reads with good quality scores whereas ABI_SOLiD sequencing seems to produce reads with low quality scores.
- For the LWK population samples, the quality of the bases in some cases is observed to reach 255 which is practically impossible. This may be due to some sequencing error or may be ideal condition.

5. DISCUSSION

The results and the observations show a lot of variations and discuss extensively about the quality scores of the reads. The first and foremost thing is to know how the qualities of the bases are calculated. The quality is originally called "Phred quality score", which was developed by the Phred program. This scoring rubric was developed for the sole purpose of automation of the sequencing of the DNA in the Human Genome Project. The quality score is generated for each of the nucleotide base in the reads after sequencing. Since the human genome project, the phred quality scores have become widely accepted and applied to compare the efficacy of sequencing methods.

The quality scores 'Q' are defined as a property that is logarithmically related to the base-calling error probabilities 'P'.

$Q = -10 \log_{10} P$

The quality scores are calculated by determining some parameters related to the peak resolution and the peak shape of each of the nucleotide base and these parameters are used as basis to look up for a matching quality score in the huge lookup tables. The quality scores in the huge lookup tables are generated from the traces of the sequence where the exact and the correct sequence are known.

In the observations found in Chapter 4, some of the reads are found to have very low quality scores. The reason for this is that each of the read is synthesized from the 5' end base by base and the steps are prone to error, which accumulate and make it hard to perform image processing. Therefore low quality of the reads is observed in some reads. One other phenomenon for generating low quality reads is that during dephasing large clusters containing thousands of identical molecules undergo reverse-strand synthesis. Each of the molecules contains one base followed by dye followed by a blocker initially and after imaging the dye and blocker are removed. In an ideal situation only a small fraction of the molecules miss out a cycle and retain an extra unblocked base. But in a real time situation, cumulative effective of the process of dephasing over cycles results in a cluster with many molecules having an extra unblocked base resulting in miscalls. This results in the decrease in the quality score of the entire read.

The same phenomenon of dephasing may also result in different runs having different base number, different order of the reads, difference in lengths of the reads and other factors. The difference in length of the reads is due to the fact that some of the reads may retain extra molecule in each cycle resulting in the detection of the extra bases in the analysis pipeline. Also the sequencing techniques are always not perfect. There might be some mechanical errors during the sequencing process and therefore it is advised to perform multiple runs on the sample to obtain better results.

One of the other observation is that the quality scores of the bases in most of the cases are observed to be lower than 50. A quality score of 50 does not mean that it is 50 % accurate, but it means that there is 1 in 100,000 probability of having an incorrect base call. Therefore the base accuracy is determined to be as 99.999% for a quality score of 50. If the quality score is found to be 10 then the probability of finding an incorrect base call is 1 in 10 and therefore the base accuracy is 90%.

The phenomenon of dephasing also applies to the Illumina and ABI_SOLiD having differences in the quality of the reads for the same sample. In ABI_SOLiD there might be reads with multiple unblocked molecules being accumulated resulting in lowering the overall quality score of the read. The other reason for the difference is that the Illumina system output 6M sequence per each run with a standard density and the ABI_SOLiD outputs around 30M sequences per run with high density. During the sequencing of microRNA, ABI_SOLiD generates a large number of “isomiRs”, which are a result of over fitting to the reference genome.

One of the other reasons for the difference in the reads in Illumina and ABI_SOLiD is due the errors while doing library preparation. Illumina though expensive, has a flexible and automated library preparation procedure resulting in reduction of manual error. On the other hand ABI_SOLiD has a very complex library preparation technique which involves the use of emulsion PCR and therefore may result in some errors. Apart from this base space versus color space can also result in the difference in determining the exact reads. The base space can be easily read and understood when compared to the color space which requires high expertise to understand and may also result in errors.

Finally by observing the reasons for the differences and the quality of the reads generated by two of the highly used sequencing platforms, Illumina proves to be the best technique to obtain better quality results even though it has some limitations over the others. In order to control the quality of the reads while sequencing several tools are developed. The limitations of the tools is that the speed of processing of the data is not proportional to the speed at which the volume of data is increasing and also the tools depend on prior information of the contaminating species that is not available in advance. Therefore it is not possible to correct the error rate while generating the data itself. One of the best, accurate, fast and holistic tools available to control the data quality is QC-Chain that helps in determining the contaminated reads and removing them by fast read processing. The advantage of this tool is that it is based on a parallel computation technique resulting in a very high processing speed of the data. Some of the other tools available are Quake, Coval, Sniper and others.

The difference between these tools and this part of the study is that the tools either take all the reads and map them to the reference species genome and therefore separate the reads based on the type of the species or either consider the read data distribution along all the cycles during sequencing and find the reads with the least quality scores.

A study on the quality of the reads during each and every cycle of sequencing was studied by Xi Yang *et al*. in 2013. The researchers were able to collect the quality of each read during each and every cycle of sequencing and plotted a graph as shown below.

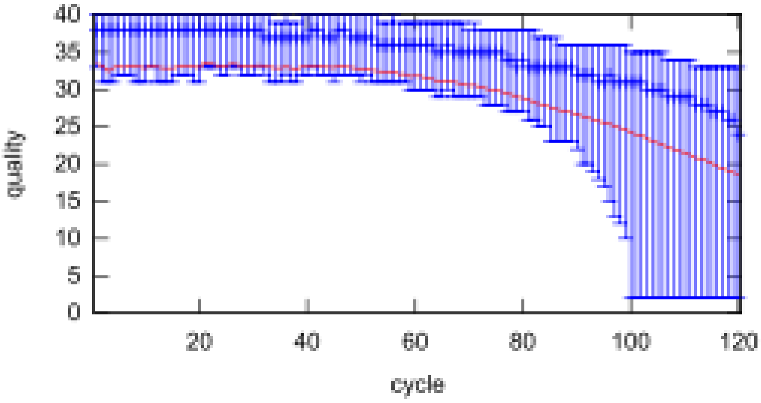


Figure 22: Box chart for read quality along read. (Xi Yang *et al*., 2013)

Figure 22 shows a box chart for the read quality of a single read considered by their study where the quality of the read keeps decreasing as the number of cycles increases strengthening the concept of dephasing of the reads during sequencing. They have also determined the correlation between two adjacent reads and plotted the correlation plot as below.

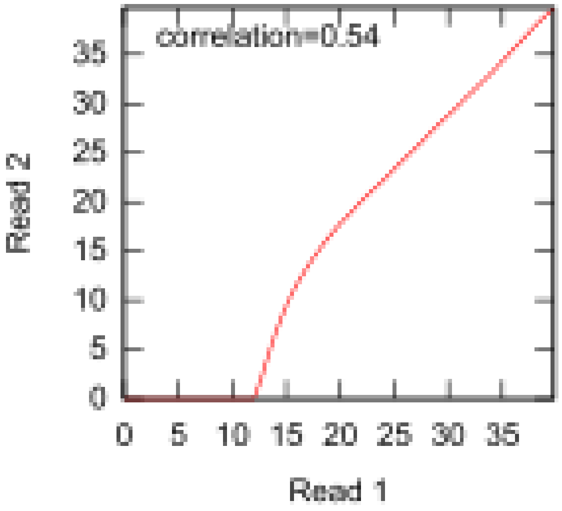


Figure 23: Plot of correlation between two read ends. (Xi Yang *et al*., 2013)

From figure 23 it can be observed that the correlation between two adjacent reads may not be high and can vary differently depending on the base present at each position of the read.

The main advantage of this study on difference in quality scores of the reads is that the technique described is easy and does not cost any money because all the tools used and the application developed are available free of cost for the user. The only time taking process here is the conversion of the data from the files in SRA format to the tab delimited text format. This study is unique in a way that the data obtained from SRA conversion into normal text format has not been analyzed ever before to determine the reads with low quality score and to compare two sequencing platforms for the same sample.

6. References

Andrew. B. S. (Oct 2011). Exome sequencing: a transformative technology. *The Lancet Neurology.* 10: 942-946. Retrieved from www.thelancet.com/neurology

Ayat. H., Dourk. B., Amanda. E. T. and Umit. V. C. (2012). Benchmarking short sequence mapping tools. *BMC Bioinformatics.* 14:184. DOI: 10.1186/1471-2105-14-184

Ben. L. and Steve. L. S. (Mar 2012). Fast gapped-read alignment with Bowtie 2. *Nat Methods.* 9(4): 357-359. DOI: 10.1038/nmeth.1923

Chandra. S. P., Rafal. S. and Andrzej. T. (2011). Sequencing technologies and genome sequencing. *Journal of Applies Genetics.* DOI: 10.1007/s 13353-011-0057-x

David. R. B. (2006). Whole-genome re-sequencing. *Science Direct.* DOI: 10.1016/j.gde.2006.10. 009

David. R. K., Michael. C. S. and Steven. L. S. (2010). Quake: quality-aware detection and correction of sequencing errors. *Genome Biol.* 11(11): R116. DOI: 10.1186/gb-2010 11-11-r116

Ewing. B. and Green. P. (1998). Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.* 8(3): 186-194. Retrieved from http://genome.cshlp.org/content/8/3/186.long

Ewing. B., Hiller. L., Wendl. M. C. and Green. P. (1998). Base-calling of automated sequencer traces using phred. I. Accuracy assessment". *Genome Res.* 8 (3): 175-185. Retrieved from http://genome.cshlp.org/content/8/3/175.long

Fatih. O., Adam. R. P., Dan. R. J., Jeffrey. G. R., Lauryn. E. S., Peter. M. … & Patrice. M. M. (2009). Direct RNA sequencing. *Nature.* DOI: 10.1038/nature08390

Fujimoto. A., Nakagawa. H., Hosono. N., Nakano. K., Abe. T., Boroevich. A. K. … & Tsunoda. T. (2010). Whole-genome sequencing and comprehensive variant analysis of Japanese individuals using massively parallel sequencing. *Nature Genetics.* DOI:10.1038/ng.691

Heng. L., Jue. R. and Richard. D. (2008). Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.* 18(11): 1851-1858. DOI: 10.1101/gr.078212.108

Jay S. and Henlee. J. (Oct 2008). Next-generation DNA sequencing. *Nature Biotechnology.* 26.10: 1135-45. Retrieved from http://www.nature.com/naturebiotechnology

Jorge. S. R. (2009). Next-Generation sequencing. *Breast Cancer Res.* 11(Suppl 3): S12. DOI: 10.1186/bcr2431

Matrin. S., Guy. C. and Hideaki. S. (Jan 2010). Archiving next generation sequencing data. *Nucleic Acids Res.* D870-D871. DOI: 10.1093/nar.gkp1078

Qian. Z., Xiaoquan. S., Anhui. W., Jian. X. and Kang. N. (2013). QC-Chain: Fast and Holistic quality control method for next-generation sequencing data. *PLos Once.* 8(4): e60234. DOI: 10.1371/journal.pone.060234

Rasko. L., Hideaki. S. and Martin. S. (2010). The sequence read archive. *Nucleic Acids Res.* DOI: 10.1093/nar/gkg1019

Richterich. P. (1998). Estimation of errors in “raw” DNA sequences: a validation study. *Genome Res.* 8(3): 251-259. Retrieved from www.genome.cshlp.org

Sebastian. D. and Szymon. G. (2013). Data compression of sequencing data. *Algorithms for Molecular Biology.* 8: 25. Retrieved from http://www.almob.org/content/8/1/25

SRA Handbook, National Center for Biotechnology Information, 2010. Bethesada.

Xiaoqing. Y., Kishore. G., Joseph. W., Martina. V., Zhenghe. W., Sanford. M., … & Shuying. S. (2012). How do alignment programs perform on sequencing data with varying qualities and from repetitive regions? *BioData Min.* 5: 6. DOI: 10.1186/1756-0381-5-6

Xi. Y., Di. L., Fei. L., Jun. W., Jing. Z., Xue. X. … & Baoli. Z. (2013). HTQC: a fast quality control toolkit for Illumina sequencing data. *BMC Bioinformatics.* 14:33. Retrieved from http://www.biomedcentral.com/1471-2105/14/33

Yuichi. K., Martin. S. and Rasko. L. (2011). The sequence read archive: explosive growth of sequencing data. *Nucleic Acids Res.* DOI: 10.1093/nar/gkr854