

T-cell repertoire profiling using Next Generation Sequencing

Dissertation submitted in partial fulfillment for the degree of

Master of Science in Biotechnology

Submitted by

Sonali Tripathy



KIIT School of Biotechnology, Campus-11

KIIT to be deemed University

Bhubaneswar, Odisha, India

Under the Supervision of

Dr. Rajasekhara Reddy,

Chief Technology Officer

Clevergene Biocorp Private Limited

Bengaluru, Karnataka

CERTIFICATE

This is to certify the dissertation entitled “*T-cell repertoire profiling using Next Generation Sequencing*” Submitted by *Sonali Tripathy* in partial fulfilment of the requirement for the degree of Master of Science in Biotechnology, KIIT School of Biotechnology, KIIT to be deemed University, Bhubaneswar bearing Roll No. 1661026 & Registration No. 16529450259 is a bona fide research work carried out by her under my guidance and supervision from *30.04.2018* to *11.05.2018*.

Date:

Place: Bengaluru




(Rajasekhara Reddy. R)

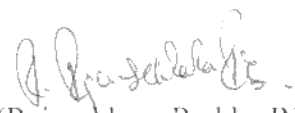
CERTIFICATE

This is to certify that the dissertation entitled "*T-cell repertoire profiling using Next Generation Sequencing*" submitted by *Sonali Tripathy* Roll No. 1661026 Registration No. 16529450259 to the KIIT School of Biotechnology, KIIT to be deemed University, Bhubaneswar-751024, for the degree of Master of Science in Biotechnology is her original work, based on the results of the experiments and investigations carried out independently by her during the period from **30.04.2018** to **11.05.2018** of study under my guidance.

This is also to certify that the above said work has not previously submitted for the award of any degree, diploma, fellowship in any Indian or foreign University.

Date:

Place: Bengaluru


(Rajasekhara Reddy. R)

DECLARATION

I hereby declare that the dissertation entitled “*T-cell repertoire profiling using Next Generation Sequencing*” submitted by me, for the degree of Master of Science to KIIT University is a record of bonafide work carried by me under the supervision of *Dr. Rajasekhara Reddy, Chief Technology Officer, Clevergene Biocorp Private Limited, Bengaluru, Karnataka, India.*

Date: 11.5.18

Place: Bengaluru

Sonali Tripathy

Acknowledgement

This dissertation had been successfully completed by putting lot of time and effort to ensure that this project meets the objectives that were stated.

It is a great pleasure to acknowledge my deepest thanks and gratitude to Dr. Rajasekhara Reddy, Chief Technology Officer, Clevergene Biocorp Private Limited and Tony Hose, Co founder and CEO, Clevergene Biocorp Private Limited, for suggesting the topic of this essay and their kind supervision. It is a great honour to work under their supervision.

I would like to express my deepest thanks and sincere appreciation to Ms. Diksha Soni, Scientific Officer, Clevergene Biocorp Private Limited, for her encouragement, creative and comprehensive advice during the laboratory work

I would like to express my sincere gratitude and appreciation to Ms. Anupam Joshi and Mr. Nahush BN, Bioinformatics Scientists, Clevergene Biocorp Private Limited, for their generous advice and support during the analysis of sequenced data.

Lastly, I would like to thank my University, KIIT School of Biotechnology, for coming up with the idea of dissertation in final semesters for the partial fulfillment of degree for Master of Science in Biotechnology which exposes the students to work in a new environment and learn new techniques and concepts.

Date: 11.5.18

Place: Bengaluru

Sonali Tripathy

Abbreviations

TCR	T-cell receptor
CD4	cluster of differentiation 4
CD8	cluster of differentiation 8
MHC	major histocompatibility complex
CD3	cluster of differentiation 3
APC	Antigen- presenting cell
HSC	Hematopoietic stem cell
CDR3	Complementarity-determining region 3
PCR	polymerase chain reaction
NGS	Next-generation sequencing
pMHC	Peptide-loaded major histocompatibility complex
DNA	Deoxyribonucleic acid
ABI	Applied biosystems

RNA	Ribonucleic acid
cDNA	complementary Deoxyribonucleic acid
IDT	Integrated DNA Technologies
TE buffer	Tris acetic acid ethylene diamine tetra acetic acid buffer
EDTA	Ethylenediaminetetraacetic acid
rpm	Revolutions per min
ETOH	Ethyl Alcohol
UV	Ultra Violet rays
SYBR	Synergy Brands
TAE	Tris-acetate-Ethylenediaminetetraacetic acid
BR	Broad-Range
RIN	RNA Integrity Number
dNTP	Deoxyribonucleotide triphosphate
dATP	Deoxyadenosine triphosphate

DTT	Dithiothreitol
RNase	Ribonuclease
HF buffer	High Fidelity buffer
dsDNA	double-stranded DNA
Gb	Giga Byte
SNPs	single nucleotide polymorphisms

List of Figures

Page no.

Figure 1: Qubit Fluorometer	11
Figure 2: DNA and RNA chips used in Bioanalyzer.....	13
Figure 3: Agilent 2100 Bioanalyzer.....	14
Figure 4: Bioanalyzer chip vortexer with adaptor.....	13
Figure 5: Electropherogram of Bioanalyzer.....	15
Figure 6: Library Multiplexing Overview.....	19
Figure 7: Illumina next generation sequencing platform.....	21
Figure 8: MiSeq flowcell.....	22
Figure 9: MiSeq reagent cartridge.....	22
Figure 10: Gel image of total RNA.....	24
Figure 11: Bioanalyzer electrophoresis profile of RNA sample.....	25
Figure 12: Bioanalyzer electrophoresis profile of T-cell library.....	25
Figure 13: Generation of clusters during sequencing.....	26
Figure 14: Distribution of sequence data quality.....	27
Figure 15: Percentage of N bases at each cycle.....	27
Figure 16: Distribution of GC percentage for read 1 and read 2 sequence data.....	28
Figure 17: Adapter contamination in read 1 and read 2.....	28

Table of Contents	Page no.
Abstract.....	1
1. Introduction.....	2
1.1 Scope and Objectives.....	6
2. Methodology.....	7
2.1 Data Mining.....	7
2.2 Primer synthesis.....	7
2.3 Extraction of RNA.....	9
2.4 QC for RNA sample.....	10
2.4.1 Quality check by Agarose Gel Electrophoresis.....	10
2.4.2 Quantity check by Qubit Fluorometer.....	11
2.4.2.1 Preparing standards and sample.....	12
2.4.3 Quality check by Agilent 2100 Bioanalyzer.....	13
2.4.3.1 Setting up chip priming station.....	15
2.4.3.2 Preparing the gel.....	16
2.4.3.3 Preparing the gel dye mix.....	16
2.4.3.4 Loading the gel dye mix.....	16
2.4.3.5 Loading the marker.....	16
2.4.3.6 Loading the ladder and sample.....	16
2.5 Library Preparation.....	17
2.5.1 First strand cDNA synthesis.....	17
2.5.2 Adapter ligation.....	17
2.5.3 Purification of ligation reaction using AmPure XP beads.....	17
2.5.4 Second strand cDNA synthesis.....	18
2.5.5 PCR for V(D)J enrichment.....	18

2.5.6 Final PCR.....	18
2.6 QC of sequencing libraries.....	18
2.7 Sequencing the libraries on Illumina MiSeq.....	19
3. Data Analysis.....	23
4. Results.....	24
4.1 Sequencing.....	26
4.2 Sequence Data Quality.....	26
5. What did I learn?.....	29
6. How to complement in Corporate?.....	30
References.....	31

Abstract

A highly diverse T-cell receptor (TCR) repertoire is a fundamental property of an effective immune system and is associated with efficient control of viral infections and other pathogens. Analysis of T-cell repertoire help to understand the dramatic changes in cellular immunity that transpire through the course of ageing which results from age dependent involution of thymus as well as by viral infection. The T-cell repertoire is highly diverse and occurs by recombination of V(D)J gene segments. Direct measurement of the TCR diversity is impossible because the diversity is high and frequency distribution of individual TCRs is heavily skewed. Because of this extreme diversity, there has been development of specialized methods which aim to characterize the T-cell repertoire in depth. Next generation sequencing based technologies are widely employed for analysis of human cell repertoire. Here we aim to standardize the laboratory procedures to generate NGS library for T-cell repertoire from blood by identifying a suitable protocol from the published reports followed by generating a sequencing library, quality check of library and data analysis of sequenced library. The methodology for this began with extraction of RNA from blood, quality check for RNA sample, library preparation, quality check for libraries, sequencing of libraries on Illumina MiSeq and data analysis. The primers used during the work were synthesized by reconstituting the same in 1X TE buffer to make concentration of 100 μ M stock. After high-throughput sequencing of the libraries, it was observed that the library generated reads ensuring that it worked properly. The quality of data was checked with MultiQC. After quality check of data, the low quality reads were removed using Trimgalore. The read length was found to be good with few adapter sequences. The reads were then mapped onto reference genome (here human whole genome was taken as reference) using software called bowtie2 which said that the reads came from the regions expected. Other analysis pipelines used were the sole proprietor of the company and were not disclosed.

1 INTRODUCTION

In humans and closely related species cellular immunity is mediated by T cells or T lymphocytes, which participate in the detection and neutralization of pathogenic threats [1]. Helper T cells not only help to activate the B cells to secrete antibodies and macrophages to destroy ingested microbe but also activate the Cytotoxic T cells to kill the infected targeted cells. T cells with functionally stable TCRs express both CD4 and CD8 co-receptors and are termed as double positive (CD4+and CD8+). The double positive T cells are exposed to wide variety of self antigens. T cell antigen receptors are found only on cell membrane. Variable domains of the chain form an antigen binding site. T cell receptor has only one antigen binding site.

Helper T cells have co receptors called CD4 which bind to MHC class II molecule. Cytotoxic T cells recognize target cells bearing antigens associated with class I MHC molecule. They can bind to any cell in the body that has been invaded by pathogens. T cell receptor is associated with a group of molecules called CD3 complex which is necessary for T cell activation. These molecules are also called signal transducers because they help to convert the extracellular binding of antigen and receptor to internal cellular signals. T cell receptors or TCRs selectively bind antigens that are displayed by the major histocompatibility complex (MHC) molecules on the surface of antigen presenting cells (APCs) [2].

The sum of all the TCRs of an individual is termed as the T-cell repertoire or TCR profile [3]. The recognition of antigens by TCRs activates the T cells, causing them to proliferate and initiate immune responses by releasing cytokines [1].

Majority of TCRs are heterodimers composed of two distinct subunit chains (α and β) both of which contain variable domain. In humans, these are encoded by single copy of genes [1]. The TCR- β gene locus, which spans 620 kb on chromosome 7, contains 50 variable (V), 2 diversity (D) and 13 joining (J) gene segments. A tremendous diversity of TCRs are required to recognize wide range of pathogenic threats one might encounter [4].

TCR diversity is generated during the early stages of T cell development. T cell progenitors are derived from hematopoietic stem cells (HSCs) in the thymus, as these cells divide recombination occurs between V and J segments and V, D and J segments in TCR- α and TCR- β genes respectively. Commonly referred to as “V(D)J recombination”, this process yields a population of T cells with sufficient TCR diversity to recognize wide range of peptides. The region of TCR- β that spans the V-D and D- J junctions is referred to as “complementarity determining region 3” (CDR3) is unique to each TCR- β variant and is frequently used for quantifying TCR diversity and high throughput profiling experiments [1].

The complementarity determining regions (CDRs) are part of variable regions in antibodies and T cell receptors generated by B cells and T cells respectively where these molecules bind to their specific antigen [5]. The CDR3 regions of both TCR- α and TCR- β chain straddles the V(D)J junction, the primary site of antigen contact. The CDR3 region is most affected by recombination and CDR3 region of β chain accounts for most of the variation within an individual’s T cell repertoire. Antigenic cross reactivity of T cells results in a discrepancy between the number of different nucleotide or amino acid TCR combinations in the host and number of different antigens recognized by the T-cell repertoire [6].

T-cell repertoire analysis help to understand the dramatic changes in cellular immunity that transpire through the course of ageing which result from age dependent involution of thymus as well as by chronic persistent viral infections. The TCR diversity has been associated with autoimmunity and it is also important in estimating the repertoire of antibody classes, accessing the size of metagenome in microbial communities and measuring the rate of evolution of pathogenic virus [6]. The TCR repertoire can change greatly with the onset and progression of diseases, which is why there has been an interest among the scientists to determine the immune repertoire status under different disease conditions such as cancer, autoimmune, inflammatory and infectious diseases [3].

The main challenge when studying the immune repertoire is its diversity. Estimating the diversity of T cell repertoire is difficult for many reasons. First, the repertoire is very diverse; second, the relationship between diversity of TCR- α and TCR- β sequences and the actual TCR diversity is unclear. In the past, different techniques were employed to study the T cell repertoire. Monoclonal antibodies allowed the analysis of specific V genes by fluorescence microscopy or flow cytometry, polymerase chain reaction (PCR) strategies in parallel with spectratyping techniques were able to provide a low resolution overview of the repertoire [6]. Knowledge on TCR amino acid sequences enables tracking of specific T cell clones in tissues which contributes to virus specific T cell immunity and enables diagnosis of various T cell related disorder. Despite these methods, technical limitations made it difficult to create a comprehensive review of the human T cell repertoire until methods based on next generation sequencing (NGS) were developed.

Massively parallel high throughput sequencing allows millions of T cell receptor genes to be characterized from a single sample of blood or tissue [7]. T-cell repertoire sequencing involves PCR amplification and sequencing of CDR3 region from one of the TCR subunits. The CDR3 region has the most variability and directly contacts the peptide MHC (pMHC). Most of the TCR-seq has focused on the TCR β chain because the locus contains the D gene segment which is absent from the α chain locus. As a result, TCR β chain has a greater potential for diversity than the α chain. The TCR β chain tends to be more interrogated because in peripheral blood which is the most accessible source of T cells, more than 90% of T cells are $\alpha\beta$ T cells [4].

TCR-seq studies have provided new insights into healthy human T-cell repertoire, such as estimation of repertoire size. In context of disease, TCR-seq has been instrumental in characterizing the immune repertoire after hematopoietic stem cell transplantation and the method has been used to develop biomarkers and diagnostics for various infections and neoplastic diseases [4].

Considering Sanger Sequencing as the first generation, new generations of DNA sequencing has been introduced consequently. The development of next generation sequencing technologies (NGS) has contributed to reducing costs and producing massive sequencing data [8].

The many commercially available NGS platforms, Roche 454 Pyrosequencing, which is based on light signal upon incorporation of nucleotide by polymerase. Illumina sequencing technology relies on reversible terminator technology for rapid and large scale sequencing. ABI Solid sequencing uses DNA Ligase for sequencing rather than DNA Polymerase. Ion semiconductor sequencing is based on detection of hydrogen ions released during polymerization of DNA. When it comes to scale and quality of data Illumina sequencing is considered to be the best in NGS technologies.

Advances in next generation sequencing technologies in which millions of sequences can be read simultaneously have been transformative for immune repertoire analysis. The extent to which cellular immunity affects human health, T cell repertoire sequencing has become a critical tool in biomedical discovery (for example, profiling the TCRs of human infiltrating lymphocytes to develop new diagnostics) and clinical management of patients (for example, using the TCR repertoire diversity to manage patients after transplantation) [4]. From a scientific standpoint, a critical avenue for future research is to address the pressing need for methodological advances that will enable routine profiling of T-cell repertoire.

Considering the growing importance of T-cell repertoire sequencing in both research and clinical field, this project is targeted to standardize the wet lab protocol for generating illumina sequencing library of T-cell repertoire from human RNA.

1.1 Scope

To standardize the laboratory procedure to generate next generation sequencing (NGS) library for T-cell repertoire.

Objectives

- ❖ To identify suitable protocol from the published reports by datamining.
- ❖ Generate sequencing library using human RNA.
- ❖ Checking the quality of library by sequencing and analyzing the data.

2 METHODOLOGY

2.1 Data Mining

To review the reported research regarding the T-cell repertoire sequencing using NGS, PubMed was searched using different combination of key words “T-cell”, “repertoire”, “next”, “generation”, “sequencing”, “Illumina”. There are different approaches for T-cell repertoire sequencing but many of them use multiple primers to amplify the T-cell receptor genes. Using multiple primers will complicate the experiment due to variation in primer efficiencies. For the current work, procedure of Heather et. al. (11th January,2016) was selected (with modifications) because it uses single primer to amplify the T cell receptor gene and it can be used to target both alpha and beta receptor regions.

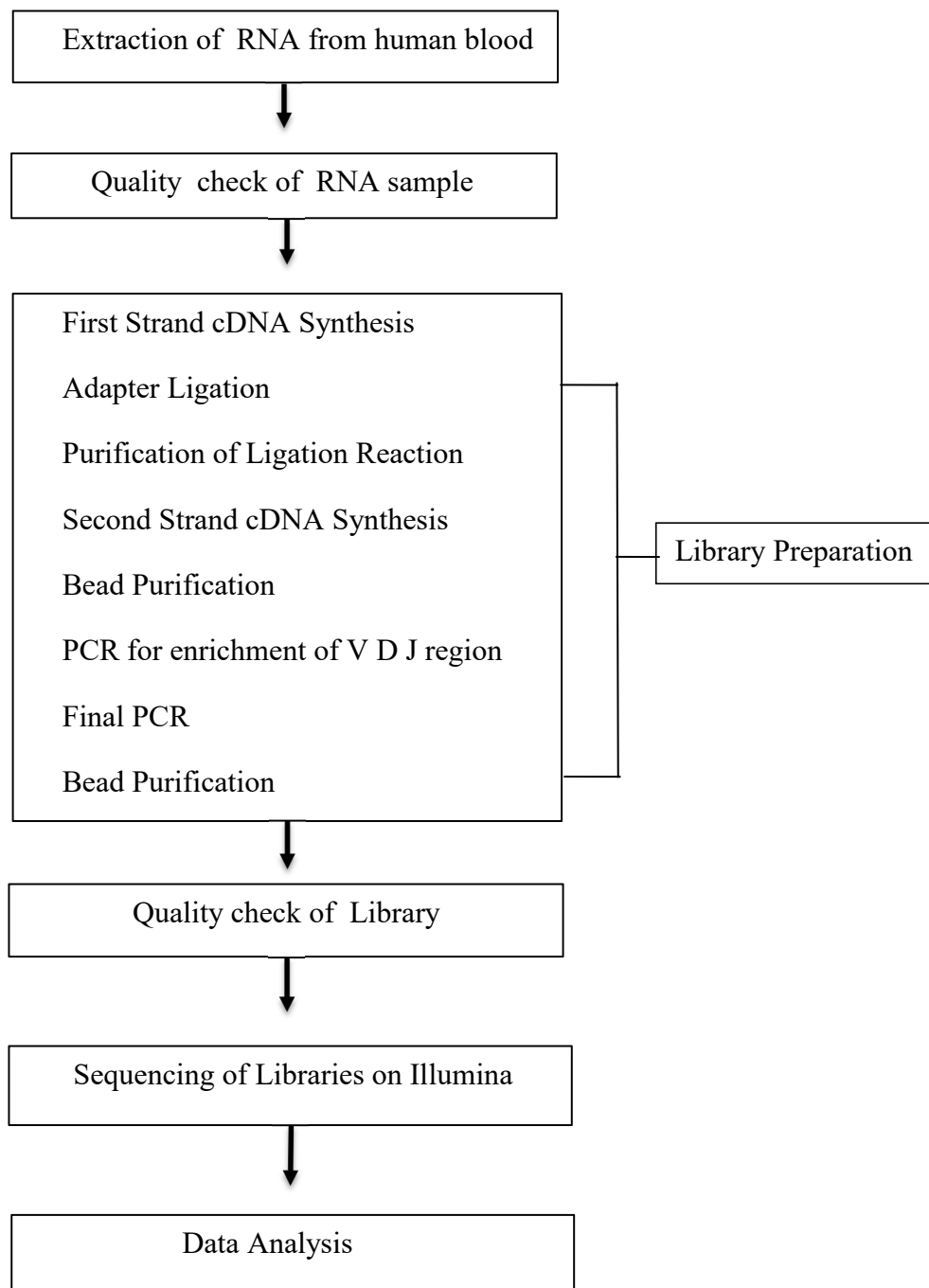
2.2 Primer Synthesis

The primers and adapter used in this project work i.e. α RC2, β RC2, SP1-6N-1X- α RC1, SP1-6N-1X- β RC1.1, SP1-6N-1X- β RC1.2 and SP2-6N respectively were taken from the article. The primers were synthesized by Integrated DNA Technologies (IDT, Singapore). These primers were reconstituted in 1X TE buffer to make concentration of 100 μ M stock. Two other primers that contain partial Illumina adapter sequences were also synthesized for indexed Illumina library generation.

Table 1 Primers and their sequence used in the experiment

Primer name	Sequence
α RC2	GAGTCTCTCAGCTGGTACACG
β RC2	ACACAGCGACCTCGGGTGGGAA
SP2-6N	[Phos]NNNNNNAGATCGGAAGAGCACACGTCTGAACTCCAGT CAC[SpC3]
SP1-6N-Ix- α RC1	ACACTCTTTCCTACACGACGCTCTTCCGATCTNNNNNNxxxx xxACGGCAGGGTCAGGGTTCTGGATAT
SP1-6N-Ix- β RC1.1	ACACTCTTTCCTACACGACGCTCTTCCGATCTNNNNNNxxxx xxGGTGGGAACACCTTGTTTCAGGTCCTC
SP1-6N-Ix- β RC1.2	ACACTCTTTCCTACACGACGCTCTTCCGATCTNNNNNNxxxx xxGGTGGGAACACGTTTTTCAGGTCCTC

Experiment Workflow



2.3 Extraction of RNA

RNA was extracted from human blood using TRIzol reagent.

Principle: TRIzol Reagent (Thermo Fisher Scientific) is a ready-to-use reagent for the isolation of total RNA/DNA/Protein from cells and tissues. The reagent, a monophasic solution of phenol and guanidine isothiocyanate, is an improvement to the single-step RNA isolation method developed by Chomczynski and Sacchi (Gauthier, Madison and Michel, 1997). During sample homogenization or lysis, TRIzol Reagent maintains the integrity of the RNA, while disrupting cells and dissolving cell components. Addition of chloroform followed by centrifugation, separates the solution into an aqueous phase and an organic phase. RNA remains exclusively in the aqueous phase. After transfer of the aqueous phase, the RNA is recovered by precipitation with isopropyl alcohol.

Procedure:

1. 1 ml of self blood was collected in an EDTA vacutainer.
2. The sample was transferred to nuclease free 1.5 ml micro tube and centrifuged at 1000 rpm for 2 minutes.
3. After removing the plasma from the samples, Buffy coat was collected and transferred to a fresh 1.5ml tube.
4. To the buffy coat 1ml of TRIzol was added and vortexed briefly and centrifuged at 12000 g for 10 min at 4°C.
5. The supernatant was discarded and to the pellet 0.2 ml chloroform was added. The tube was vortexed vigorously for about 15-20 sec and incubated at room temperature for about 5 min.
6. The sample was centrifuged at 12000 g for 15 min at 4 °C. The upper aqueous phase containing the total RNA was transferred to a fresh tube.
7. To the tube about 0.5 ml Isopropyl alcohol per 1ml TRIzol reagent was added and incubated at room temperature for about 5 min and centrifuged at 12000 g for 8 min at 4 °C.
8. The supernatant was discarded and the pellet was washed with 1 ml 75% ETOH and centrifuged at 12000 g for 5 min at 4 °C.
9. The pellet was air dried at room temperature and resuspended in 10 µl TE buffer.

2.4 QC of RNA sample

Good quality RNA yields good sequencing library hence the quality check of the RNA sample is essential for NGS library preparation.

2.4.1 Quality check by Agarose Gel Electrophoresis

Principle:

Agarose gel electrophoresis is a routinely used method for separating proteins, DNA or RNA. (Kryndushkin et al., 2003). Nucleic acid molecules are size separated by the aid of an electric field where negatively charged molecules migrate toward anode (positive) pole. The migration flow is determined solely by the molecular weight where small weight molecules migrate faster than larger ones (Sambrook & Russel 2001). In addition to size separation, nucleic acid fractionation using agarose gel electrophoresis can be an initial step for further purification of a band of interest. Extension of the technique includes excising the desired “band” from a stained gel viewed with a UV transilluminator (Sharp et al., 1973). In order to visualize nucleic acid molecules in agarose gels, ethidium bromide or SYBR Safe are commonly used dyes. Illumination of the agarose gels with 300-nm UV light is subsequently used for visualizing the stained nucleic acids.

Procedure:

1. 2% gel was prepared by dissolving 0.6 g of SeaKem LE Agarose(Lonza) 500 g, Catalog #50004, Rockland, ME, USA in 30 ml of 1X TAE buffer (by Thermo Fisher Scientific) in a flask.
2. The flask was heated in a microwave for 2 min to let the agarose dissolve properly.
3. The gel was then allowed to cool followed by addition of 3 μ l SYBRTM Safe DNA Gel Stain (by Thermo Fisher Scientific) Catalog #S33102.
4. The gel was then poured on the gel casting tray with comb inserted in it and was allowed to solidify for 20 min.
5. After solidification of gel, the casting tray was placed in a buffer tank containing 700-800 ml of 1X TAE buffer.
6. 1 μ l of TrackitTM Cyan/Orange Loading Buffer (6X) Catalog #10482028 was added to parafilm and 5 μ l of RNA sample was added and mixed by setting the pipette to 6 μ l.

7. 6 μ l of sample was added to the first well and 5 μ l of 100 bp ladder from Chromous Biotech Pvt. Ltd. Catalog #LAN02 was added to the second well.

8. The gel was then run at 100 V for 20 min.

9. The gel was viewed under LUMIX-BOX™ Blue Light LED epi-illuminator Catalog #APSVE100. It features a blue light wavelength of 470 nm that keeps the user safe from UV radiations.

2.4.2 Quantity check by Qubit 2.0 Fluorometer

Principle:

The Qubit 2.0 Fluorometer is a benchtop fluorometer for the quantitation of DNA, RNA, and protein, using the highly sensitive and accurate fluorescence-based Qubit quantitation assays. Use of the state-of-the-art dyes selective for dsDNA, RNA, and protein minimizes the effects of contaminants in the sample that affect the quantitation. Further, the very latest illumination and detection technologies used in the Qubit 2.0 Fluorometer for attaining the highest sensitivity allows to use as little as 1 μ l of sample and still achieve high levels of accuracy, even with very dilute samples. Each dye is specific for one type of molecule: DNA, RNA and Protein. These dyes have low fluorescence and upon binding to the target (DNA, RNA and Protein), they become highly fluorescent.



Figure 1 Qubit 2.0 Fluorometer

Procedure:

The sample was quantified using Qubit RNA Broad Range(BR) assay kit. The assay kit is designed for initial RNA sample concentrations from 1 ng/ μ L to 1 μ g/ μ L providing an assay range from 20–1,000 ng. The kit provides concentrated assay reagent, dilution buffer, and pre-diluted RNA standards.

2.4.2.1 Preparing standards and sample

1. 0.5 ml Qubit assay tubes were set up for standards and sample. 2 tubes for standards and 1 for sample.
2. The lid of the tube was labeled.
3. Qubit working solution was prepared by diluting the Qubit RNA BR Reagent 1:200 in Qubit RNA BR Buffer.

Buffer: $199 \times 3 \times 1.1 = 656.7 \mu\text{l}$

Reagent: $1 \times 3 \times 1.1 = 3.3 \mu\text{l}$ (The final volume in each tube was 200 μl)

4. In each standard tube, 10 μl of Qubit RNA BR standard 1 and 10 μl of Qubit RNA BR standard 2 was taken followed by 190 μl of working solution in each tube.
5. In the sample tube, 1 μl of sample was taken followed by 199 μl of working solution.
6. The tubes were vortexed and a quick spin was given to collect any of the liquids adhering to the walls of the tube.
7. The tubes were then allowed to incubate at room temperature for 2 min.
8. The readings were then taken on Qubit Fluorometer.
9. On the home screen of Qubit Fluorometer, RNA Broad range as the assay type was selected. The standards screen was displayed and both of the standards were read by inserting the tube and pressing on “read standards”.
10. In the same way, the sample tube was also inserted and the concentration of the sample was read.

2.4.3 Quality check by Agilent 2100 Bioanalyzer

The quality of RNA sample was assessed based on RNA Integrity Number (RIN) and it was generated by running the sample on Agilent Bioanalyzer RNA 6000 nano chip. A RIN value > 7 suggests that the RNA is of good quality.

Principle:

Agilent Bioanalyzer offers “lab on a chip” microfluidics platform for analysis of DNA, RNA and protein. The instrument offers sizing, quantification and quality control, and delivers results in high quality digital data. It is an automated instrument and can analyze up to 12 samples in less than 30 minutes. A number of kits are available for use with Bioanalyzer. The RNA kits offer the ability to check RNA quality including assignment of RNA Integrity Number(RIN).



Figure 2 DNA and RNA chips used in Bioanalyzer



Figure 3 Agilent 2100 Bioanalyzer



Figure 4 Bioanalyzer chip vortexer with adaptor.

Calculation of RNA Integrity Number(RIN)

The RNA Integrity Number(RIN) was developed to remove individual interpretation in RNA quality control. The RIN software algorithm allows for the classification of eukaryotic total RNA based on a numbering system from 1 to 10, with 1 being the most degraded profile and 10 being the most intact. For development of RIN algorithm, adaptive learning tools such as neural networks provided by quantum bioinformatics were employed. They allowed the determination of critical features that can be extracted from an electrophoretic trace. These features are part of an electropherogram that can be analyzed using an appropriate integrator. They can be signal areas, intensities, ratios etc.

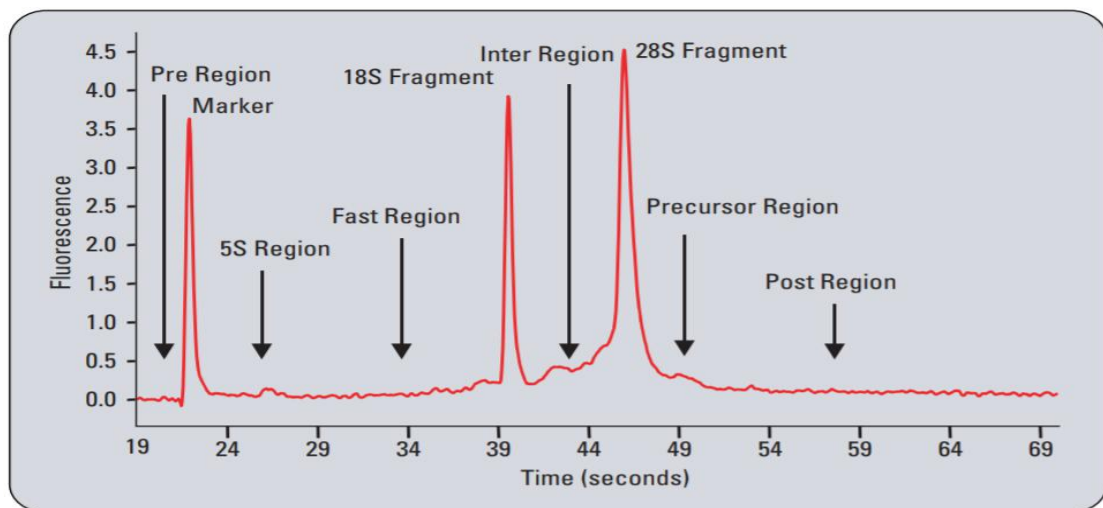


Figure 5 Electropherogram showing the regions that are indicative of RNA quality.

Procedure:

2.4.3.1 Setting up the chip priming station

1. The old syringe was unscrewed from the lid of chip priming station.
2. The old syringe was released from the clip and discarded.
3. The plastic cap of new syringe was removed and was inserted into the clip.
4. It was then slid into the hole of the lock adapter and screwed tightly to the chip priming station.
5. The chip priming station was opened by pulling the latch.
6. The syringe clip was adjusted.

2.4.3.2 Preparing the gel

1. 550 μl of RNA gel matrix was pipetted into a spin filter.
2. Centrifuged at 15000 g at room temperature for 10 min.
3. 65 μl of filtered gel was aliquoted to RNase free tubes. The tubes were stored at 4 °C.

2.4.3.3 Preparing the gel dye mix

1. The RNA dye concentrate was allowed to equilibrate to room temperature for 30 min.
2. The RNA dye concentrate was vortexed for 10 sec, centrifuges briefly and 1 μl of it was added to 65 μl aliquot of filtered gel.
3. The solution was vortexed and centrifuged at 13000 g at room temperature for 10 min.

2.4.3.4 Loading the gel dye mix

1. A new RNA chip was put on the chip priming station.
2. 9 μl of gel dye mix was put into the well marked 'G'.
3. The plunger was positioned at 1 ml and the chip priming station was closed.
4. The plunger was pressed till it was held by the clip.
5. The clip was released after 30 sec and the plunger was pulled back to 1 ml position.
6. The chip priming station was opened and 9 μl of gel dye mix was added to the wells marked 'G'.

2.4.3.5 Loading the marker

1. 5 μl of RNA marker was added in all 12 samples well as well in the ladder well.

2.4.3.6 Loading the ladder and samples

1. 1 μl of ladder was added to the ladder well.
2. 1 μl of sample was added to each 12 samples well. 1 μl of RNA marker was added to each unused sample well.
3. The chip was then placed horizontally in the chip vortexer and vortexed for 1min at 2400 rpm.
4. The chip was the run on Agilent 2100 Bioanalyzer within 5 min.

2.5 Library Preparation

2.5.1 First strand cDNA synthesis

In a PCR tube RNA was mixed with 1 μ l of 10 μ M each RC2 primers (α RC2 and β RC2), 1 μ l of dNTP mix (10 mM) and heated at 65°C for 5 min. The tube was then immediately put on ice for rapid cooling. 1 μ l of Dithiothreitol (DTT) (0.1 μ M), 1 μ l of Murine RNase inhibitor (20-40 U/ μ l), 1 μ l of ProtoScript II Reverse Transcriptase (200 U/ μ l), 8 μ l of 5X FS buffer were added and incubated at 55°C for 20 min, then heat inactivated at 70°C for 15 min, and then hold at 4°C. Reverse transcribed reactions were then purified and eluted in 10 μ l of elution buffer using Qiagen PCR purification column.

2.5.2 Adapter Ligation

SP2-6N adapter was ligated to at the 3' end of cDNA in 60 μ l of 1X T4 RNA ligase buffer (2 μ l), T4 RNA ligase 1(5 μ l), dATP (1 mM) (2 μ l), SP2-6N ligation oligonucleotide (1 μ M) (2 μ l) and 20 μ l of purified cDNA. Ligation reaction was incubated at 16°C for 90 min, heat inactivated at 65°C for 10 min and the hold at 4°C. The products were diluted with 40 μ l of nuclease free water and purified using 1X AmPure XP beads. The final product was eluted in 30 μ l of 0.1X TE buffer.

2.5.3 Purification of Ligation Reaction using AmPure XP beads

30 μ l (1X) resuspended AmPure XP beads was added to the ligated product and was mixed well by vortexing 10 times. It was then incubated at room temperature for about 5 min. A quick spin was given to collect any of liquid adhering to the walls of the tube. The tube was then placed on the magnetic rack to separate the beads from supernatant. The supernatant was discarded after the solution became clear. 200 μ l of freshly prepared 80% Ethanol was added to the tube containing beads. The tube was incubated at room temperature for 30 sec after which the supernatant was removed carefully without disturbing the beads. The beads were air dried for 5 min with the tube on the magnetic stand with lid open. The tube was then removed from the magnetic stand and DNA was eluted from beads by adding 25 μ l of 0.1X TE buffer. It was then mixed well by vortexing/pipetting up and down followed by incubation at room temperature. The tube was put on a magnetic rack for the solution to become

clear. Without disturbing the bead pellet the supernatant was transferred to a clean PCR tube.

2.5.4 Second strand cDNA synthesis

The ligated cDNA underwent second strand synthesis in 50 μ l of dNTPs (10 mM) (1 μ l), TN701 custom amplicon primer (10 μ M) (1 μ l), Phusion (1 U) (0.5 μ l) in 5X HF buffer (10 μ l) and nuclease free water (7.5 μ l). The thermal conditions were 95°C for 3 min, slow ramp to 80°C for 10 sec, another slow ramp to 58°C for 45 sec, final extension at 72 °C for 5 min and hold at 4 °C.

2.5.5 PCR for V D J enrichment

The above reaction was bead purified (1X) and the reaction conditions were the same as 2.5.4 except the primers used were SP1-6N-1x- α RC1, SP-6N-1x- β RC1.1 and SP-6N-1x- β RC1.2. This reaction targeted the V D J regions.

2.5.6 Final PCR

The purified products were pooled together and subjected to PCR to generate sequencing library: dNTP (10 mM) (0.5 μ l), TN701 and TN501 custom amplicon primers (10 μ M each) (1 μ l each) and Kappa (1 U) (25 μ l) and nuclease free water (3 μ l). The PCR program started with initial denaturation at 95 °C for 3 min before 35 cycles of 95°C for 5 min, 95°C for 30 sec, 55°C for 45 sec, 72°C for 3 sec, final extension at 72°C for 7 min and the hold at 4 °C followed by 0.9X bead purification.

2.6 QC of sequencing libraries

Agilent Bioanalyzer DNA 7500 kit was used for the quantification of cDNA libraries. This ensures optimization of Next Generation Sequencing(NGS) libraries prior to sequencing run.

The libraries were denatured and diluted and sequenced on Illumina MiSeq, using 2 x 300 paired end sequencing kits.

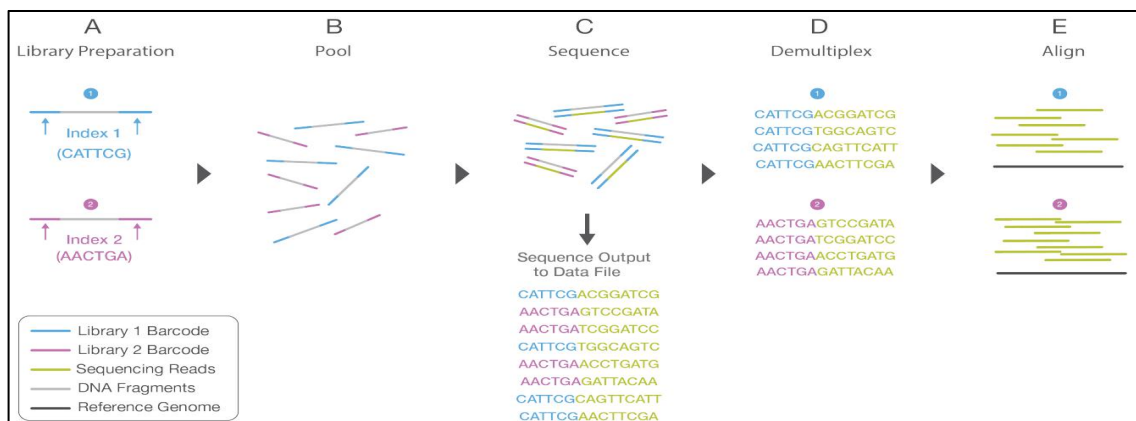
Illumina Next Generation Sequencing(NGS) Technology

Illumina sequencing by synthesis chemistry is the most widely accepted next generation sequencing technology. This sequencing technology uses four

fluorescently labeled nucleotides to sequence millions of clusters on flow cell surface in parallel. During each sequencing cycle, a single labeled dNTP is added to the nucleic acid chain based on the complementarity. The labeled nucleotide serves as a terminator for polymerization, so after each dNTP incorporation, the fluorescent dye is imaged to identify the base and then enzymatically cleaved to allow the incorporation of next nucleotide. Base calls are made directly from signal intensity measured during each cycle which reduces the error rates compared to other technologies.

2.7 Sequencing the libraries on Illumina MiSeq

Illumina offers ready to use next generation sequencing kits. The kit contains ready to use reagent cartridges which is sufficient for 600 cycles. The maximum output for 2 x 300 bp read length is 15 Gb and the total time taken to complete the run is ~3 days. Since we need only a fraction of data for this work, the current library was multiplexed with other libraries.



Multiplexing allows a large number of libraries to be pooled and sequenced simultaneously during a single sequencing run. With multiplexed libraries, unique index sequences are added to each fragment during library preparation so that each read can be identified and sorted before final data analysis.

Figure 6 Library Multiplexing Overview - (A) Unique index sequences are added to different libraries during library preparation. (B) Libraries are pooled together and loaded into the same flow cell lane. (C) Libraries are sequenced together during a single sequence run. (D) A demultiplexing algorithm sorts the reads into different files according to their indexes. (E) Each set of reads is aligned to the appropriate reference sequence.

Workflow of Illumina Sequencing

Diluting and pooling the library



Cluster Generation
Hybridization of sample to flow cell
Amplification of sample
Linearization of fragments
Blocking of fragments
Hybridization of sequencing primers



Sequencing



Data Analysis



1. Flow cell compartment - Houses the flow cell throughout the run.
2. Touch screen monitor - Enables instrument configuration and run setup.
3. Optics module - Enables imaging of flow cell.
4. Reagent compartment - Holds reagents, wash solutions and wash bottles.



Figure 7 Illumina MiSeq Next Generation Sequencing platform.



Figure 8 MiSeq Flow cell



Figure 9 MiSeq reagent cartridge with numbered reservoirs.

3 Data Analysis

The sequence data quality was checked by analyzing the reads with FastQC [9] and MultiQC [10] programs. GC% and adapter contamination was checked. The data was processed to remove the low quality reads and trim adapter sequences. The quality filtered reads were mapped to human genome using STAR v2 RNAseq aligner [11] to check the whether the reads are generated from the V(D)J regions.

The reads will be binned based on V(D)J recombination and the quantity of each combination will be assessed based on number of reads in each bin. The variation between samples will be assessed to identify the differentially expressed V(D)J recombination.

4 Results

The extracted total RNA concentration was 0.5 $\mu\text{g}/\mu\text{l}$ and the RNA quality was good as per the agarose gel image (Figure 10) showing two distinct rRNA bands. Bioanalyzer RNA integrity number was 10 indicating high quality intact RNA (Figure 11) with two peaks of rRNA.

The library concentration was 1.28 $\text{ng}/\mu\text{l}$ and the Bioanalyzer report showed that in addition to T cell library there are adapter dimers (peak at 135 bp) (Figure 12).

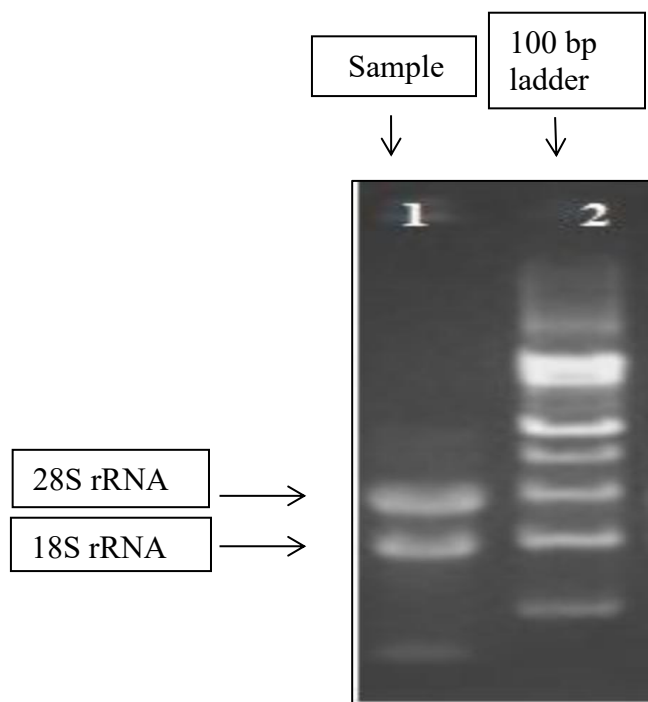


Figure 10 Gel image of total RNA

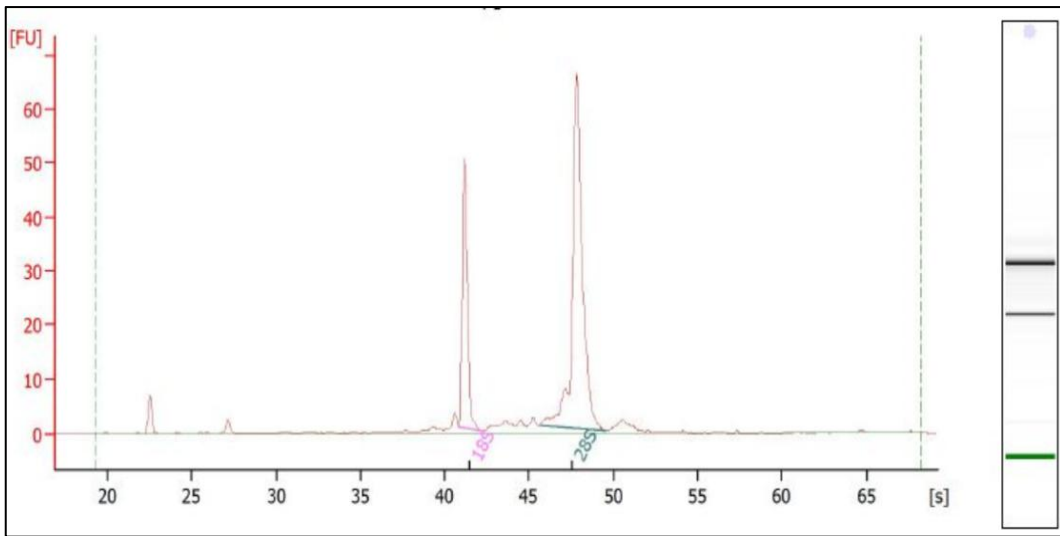


Figure 11 Bioanalyzer electrophoresis profile of RNA sample.

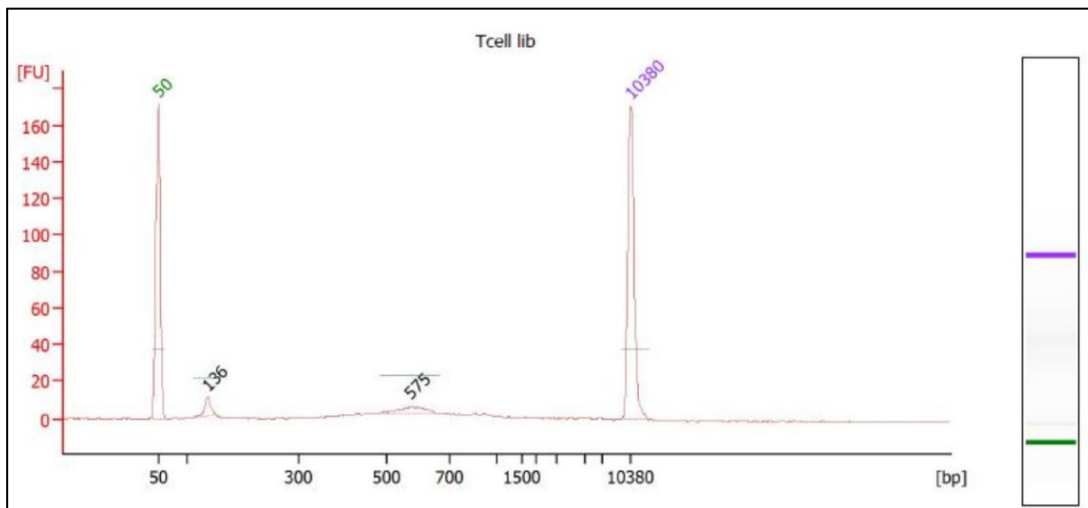


Figure 12 Bioanalyzer electrophoresis profile of T cell library

4.1 Sequencing

In addition to T- cell repertoire library, other libraries were also included in the sequencing run to utilize full capacity of the flow cell. The number of clusters were 983 ± 13 k/mm² indicating good performance of libraries and optimal clustering. The total yield of the run was 13.19 GB. The quality parameters of the run, error rate and phasing/pre-phasing were within the recommended thresholds i.e. <3 and <0.2/0.1 respectively. After demultiplexing, the T-cell repertoire library yielded 357926 reads.

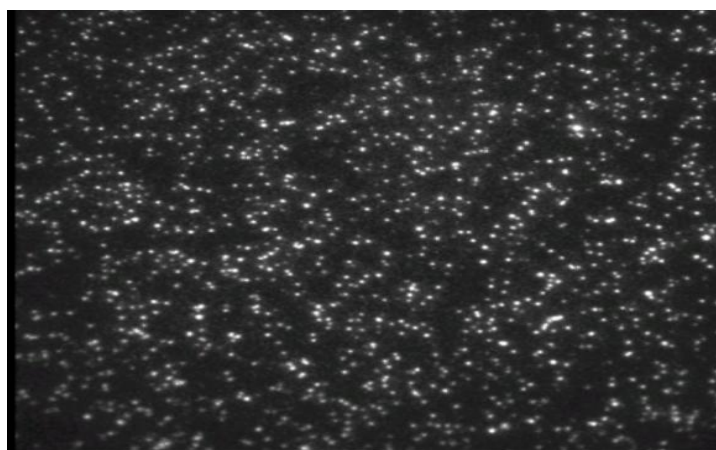


Figure 13 Generation of clusters during sequencing

4.2 Sequence Data Quality

The sequence data was of high quality, as 81% of the run data has quality values (Q-values) greater than 30 (Figure 14). In the T-cell repertoire library there were no ambiguous calls (“N”) detected (Figure 15). The % GC of read 1 and read 2 are almost same indicating good quality base calls (Figure 16). There was considerable adapter sequence contamination in the data (Figure 17). The adapter contamination was anticipated based on the Bioanalyzer results. After quality trimming the sequence data was used for T-cell repertoire analysis pipeline.

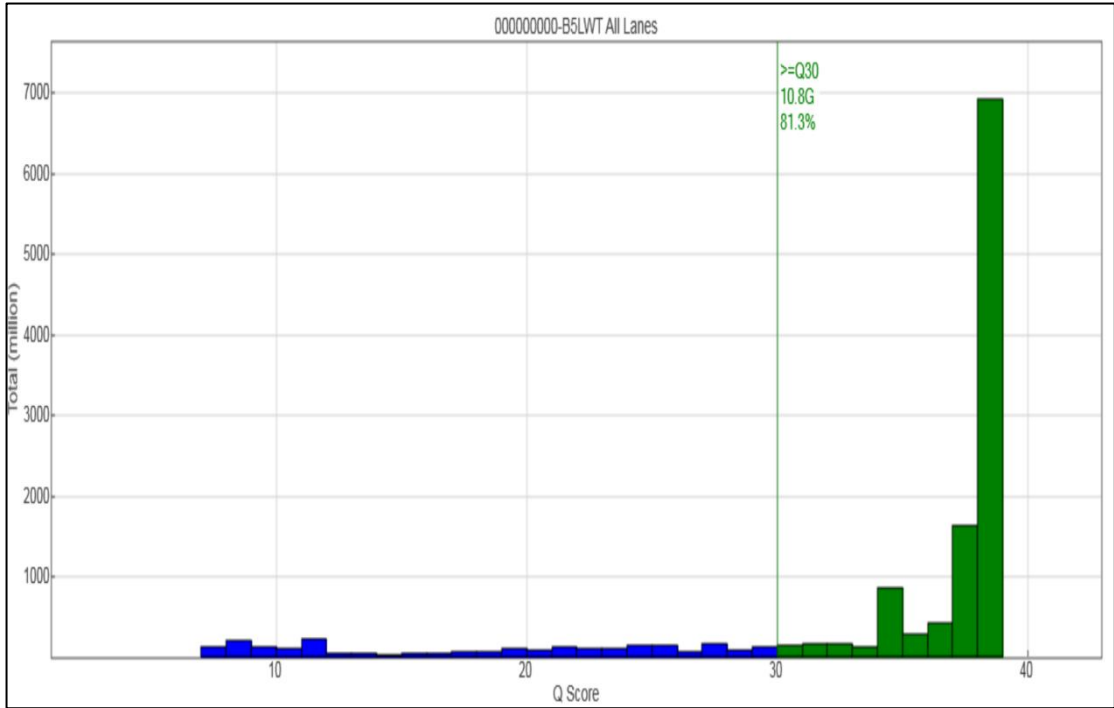


Figure 14 Distribution of sequence data quality scores.

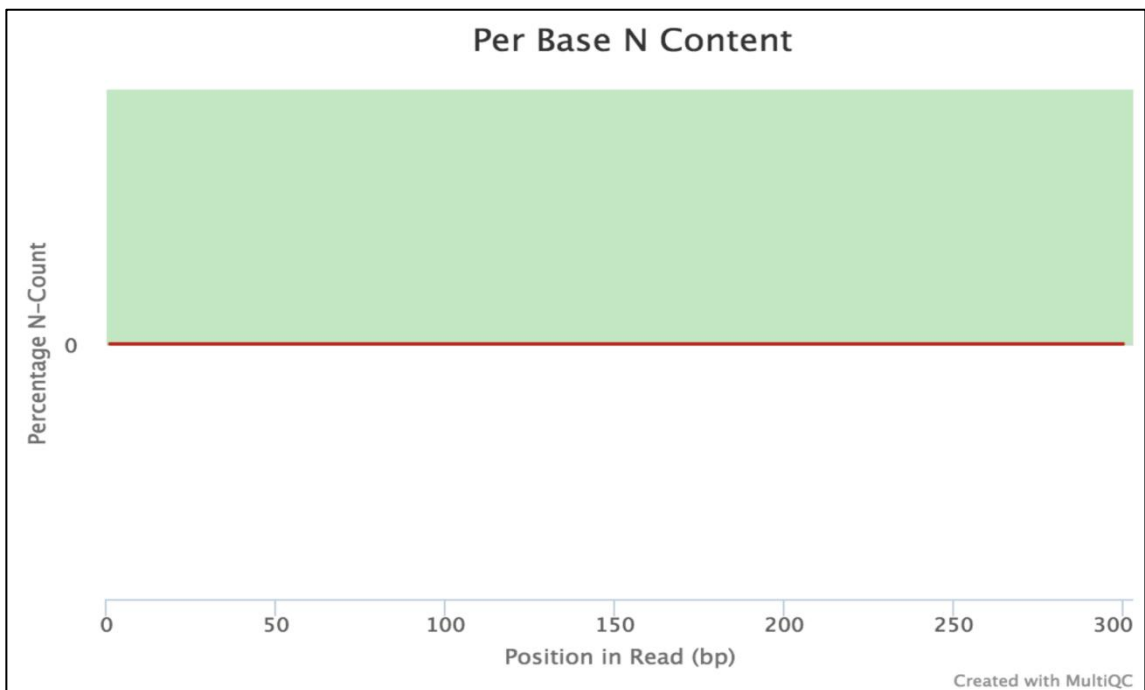


Figure 15 Percentage of N bases at each cycle.

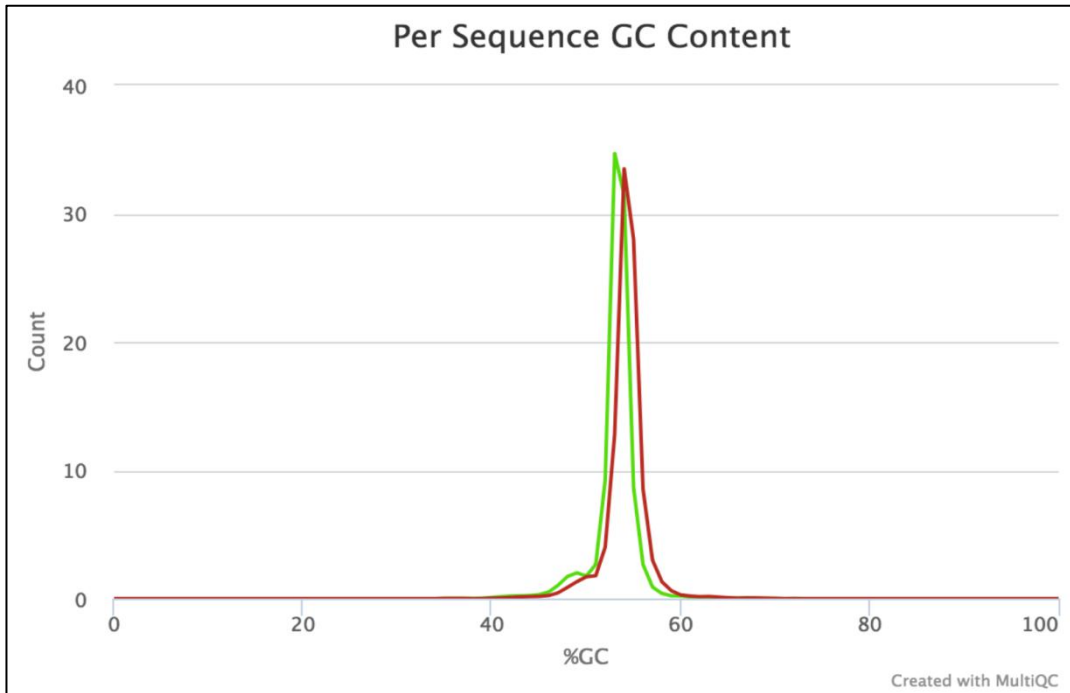


Figure 16 Distribution of GC percentage for read1 and read 2 sequence data.

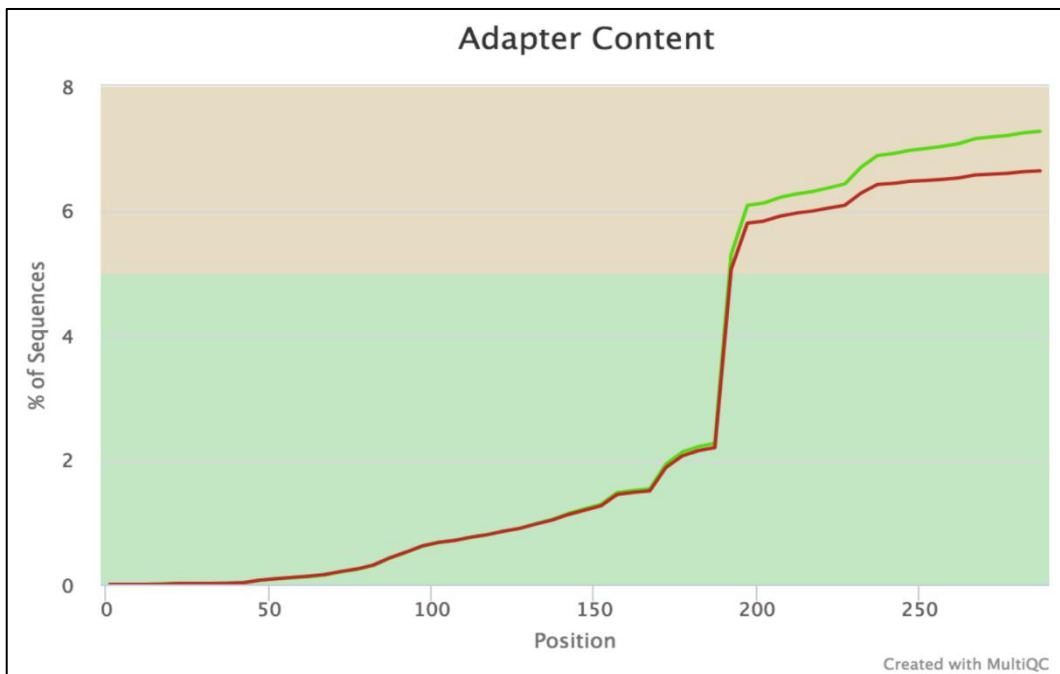


Figure 17 Adapter contamination in read 1 and read 2.

5 What Did I learn?

During the period of dissertation, I was able to know how a Molecular biology lab in a corporate organization functions. How the protocols are developed and standardized and documented. As part of the project was I was able to perform literature search and understand the published protocols.

When it comes to Molecular biology techniques, I worked precisely with instruments like Bioanalyzer, Qubit Fluorometer and MiSeq. These instruments are regularly used in a Molecular biology lab to assess the quality and quantity of DNA and RNA. I was able to understand the steps involved in next generation sequencing technology.

My project work was mainly concentrated on standardizing a custom library preparation method for T-cell repertoire sequencing. This helped me learn more about how sample quality is checked before starting taking it to the experiment. How to plan a day's activity in the lab so I can stop my experiment at a safe point and continue it on the following day.

I have grasped the importance of data quality check and filtering. As this impacts the downstream results it was essential for me to know about the quality parameters and their thresholds. These helped me understand how can we declare whether a data set is good or bad.

Knowing about the cutting edge technology like NGS has helped me understand the potential and possibilities the field of Molecular biology and genomics offer. I was able to understand the importance of T-cell repertoire sequencing and its application in clinical sector.

In summary, I was able to work with a group of genomic scientists who integrated wet lab and bioinformatics techniques to develop methods that can be offered as products. I learned how to be part of such a diverse organization.

6 How to complement in Corporate?

The interest in writing the Master's dissertation at an organization is constantly increasing. There are several advantages of writing a thesis in co-operation with a company. It helps us to create valuable contacts with the corporate world while finishing our studies at the same time. We get to solve a real problem and participate in the planning and development of real arrangements, services or products in an organization. We get supervision and guidance from several directions which makes our work easier. We get a chance to promote our knowledge outside the academic world. The high practical orientation and results oriented cooperation within a team are a major incentive for students to spend the six month Master thesis phase at a company.

References

1. A SMARTer Approach to T-Cell Receptor Profiling [Internet]. [cited 2018 Apr 30]. Available from: http://www.clontech.com/IN/Products/cDNA_Synthesis_and_Library_Construction/NGS_Learning_Resources/Technical_Notes/Human_TCR_Profiling?sitex=10060:22372:US
2. T cell [Internet]. Wikipedia. 2018 [cited 2018 Apr 30]. Available from: https://en.wikipedia.org/w/index.php?title=T_cell&oldid=837433822
3. Rosati E, Dowds CM, Liaskou E, Henriksen EKK, Karlsen TH, Franke A. Overview of methodologies for T-cell receptor repertoire analysis. BMC Biotechnol [Internet]. 2017 [cited 2018 Apr 30];17. Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5504616/>
4. Woodsworth DJ, Castellarin M, Holt RA. Sequence analysis of T-cell repertoires in health and disease. Genome Med. 2013;5:98.
5. Complementarity-determining region [Internet]. Wikipedia. 2017 [cited 2018 Apr 30]. Available from: https://en.wikipedia.org/w/index.php?title=Complementarity-determining_region&oldid=801793124
6. Laydon DJ, Bangham CRM, Asquith B. Estimating T-cell repertoire diversity: limitations of classical estimators and a new approach. Phil Trans R Soc B. 2015;370:20140291.
7. Heather JM, Ismail M, Oakes T, Chain B. High-throughput sequencing of the T-cell receptor repertoire: pitfalls and opportunities. Brief Bioinform. 2017;
8. Ari S, Arkan M. Next-Generation Sequencing: Advantages, Disadvantages, and Future. 2016. p. 109–35.
9. Andrews S. FastQC A Quality Control tool for High Throughput Sequence Data [Internet]. <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>. [cited 2016 Jun 29]. Available from: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>

10. Ewels P, Magnusson M, Lundin S, Källér M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*. 2016 Oct 1;32(19):3047–8.

11. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinforma Oxf Engl*. 2013 Jan 1;29(1):15–21.