

# **Human Mitochondrial Genome Analysis using Next Generation Sequencing Technology**

Dissertation submitted in partial fulfillment for the degree of  
Master of Science in Biotechnology

Submitted By

**SWETA GHOSH**



KIIT School of Biotechnology, Campus- 11

KIIT to be Deemed University

Bhubaneswar, Odisha, India

Under the Supervision of

**Dr. Rajasekhara Reddy**

**Chief Technology Officer (CTO)**

**Cleverage Biocorp Pvt. Ltd**

**Bengaluru, Karnataka.**

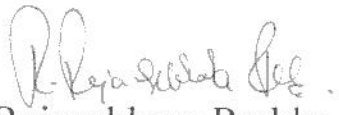
## CERTIFICATE

This is to certify the dissertation entitled “**Human Mitochondrial Genome Analysis using Next Generation Sequencing Technology**” Submitted by **SWETA GHOSH** in partial fulfillment 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. 1661031& Registration No. 16646851465 is a bona fide research work carried out by her under my guidance and supervision from 18.12.2017 to 15.05.2018.

Date: 15.5.2018

Place: BENGALURU



  
(Rajasekhara Reddy .R)

CERTIFICATE

This is to certify that the dissertation entitled “**Human Mitochondrial Genome Analysis using Next Generation Sequencing Technology**” submitted by SWETA GHOSH *Roll No. 1661031 Registration No. 16646851465* 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 18.12.2017 to 15.05.2018 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: 15.5.2018

Place: BENGALURU



  
(Rajasekhara Reddy .R)

## DECLARATION

I hereby declare that the dissertation entitled “**Human Mitochondrial Genome Analysis using Next Generation Sequencing Technology**” 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 Pvt. Ltd, Bengaluru, Karnataka.***

*Date:*

*Place: Bengaluru*

*Name & Signature*

## **Acknowledgments**

I would like to express my gratitude to Dr. Rajasekhara Reddy, Chief Technology Officer (CTO), the project supervisor at Clevergene Biocorp Pvt. Ltd, Bangalore, Karnataka. for allowing me to complete my M.Sc. Dissertation work. I thank him for his encouragement and valuable guidance in every step.

I am grateful to Mr. Tony Jose, Chief Executive Officer (CEO), Ms. Diksha Soni, and Ms. Amrita Bhattacharjee, Scientific Officers for their support, cooperation and being patient to my queries.

I would like to thank the whole team and my fellow mates of Clevergene Biocorp Pvt. Ltd for their constant support and cooperation.

A special note of thanks to my professors at KIIT School of Biotechnology, KIIT University, Bhubaneswar for helping me to gain a deeper insight in the field of genomics .

*Date:*

*Place:*

*Name & Signature*

## Abbreviations

ATP	Adenosine Triphosphate
BR	Broad Range
CE	Capillary Electrophoresis
CPEO	Chronic Progressive External Ophthalmoplegia
D-loop	Displacement Loop
H-strand	Heavy Strand
IDT	Integrated DNA Technologies
L-strand	Light Strand
MCS	MiSeq Control Software
MELAS	Mitochondrial Encephalomyopathy, Lactic Acidosis, and Stroke-like episodes
MERRF	Myoclonic Epilepsy with Ragged-Red Fibers
mtDNA	Mitochondrial Deoxyribo Nucleic Acid

nDNA	Nuclear Deoxyribo Nucleic Acid
NGS	Next Generation Sequencing
PCR	Polymerase Chain Reaction
RC	Respiratory Chain
rCRS	Revised Cambridge Reference Sequence
SAV	Sequence Analysis Viewer
TAE	Tris-Acetate-EDTA
TBE	Tris-Borate-EDTA

## List of Figures

<b>Figure-1:</b> Schematic diagram of mitochondrial DNA.....	3
<b>Figure-2:</b> Work flow of experiment.....	11
<b>Figure-3:</b> Qubit 2 flurometer.....	16
<b>Figure-4:</b> Flow of PCR cycle.....	18
<b>Figure-5:</b> Workflow of DNA library prep.kit for Illumina.....	21
<b>Figure-6:</b> Gel Photograph of CEPH DNA.....	27
<b>Figure-7:</b> Gel picture showing PCR bands .....	28
<b>Figure-8:</b> Gel photograph after fragmentation .....	29
<b>Figure-9:</b> Gel photograph of final library .....	30
<b>Figure-10:</b> Bioanalyzer electropherogram of final library .....	31
<b>Figure-11:</b> Photograph showing cluster densities.....	32
<b>Figure-12:</b> Q score distribution Plot.....	33
<b>Figure-13:</b> Q30 graph of Illumina reads.....	34
<b>Figure-14:</b> Quality score distribution over all sequences.....	35
<b>Figure-15:</b> N content across all bases.....	36
<b>Figure-16:</b> Distribution of sequence lengths over all sequences.....	37
<b>Figure-17:</b> Allignment plot.....	38



## Table of Contents

Abstract.....	1
Introduction.....	2
Scope and Objectives.....	10
Methodology.....	11
Literature mining.....	12
DNA quality check.....	12
Agarose gel electrophoresis and quantitation.....	14
PCR amplification and amplicon check.....	17
Clean up and quantitation.....	20
Fragmentation and fragment check.....	20
DNA library preparation.....	21
Library quality check.....	22
Sequencing.....	25
Data Analysis.....	26
Results.....	27
What did you learn.....	39
References.....	40

## **Abstract**

The availability of NGS technologies has provided mitochondrial genome sequences with high coverage, thereby enabling decoding of a number of human mitochondrial diseases. The disease associated with mitochondria can be identified quickly by their inheritance patterns and sequencing the mtDNA. In NGS technology, large depth coverage, could be effectively used to map heteroplasmic variations. It is capable of reliably detecting and quantifying heteroplasmies down to the 1%.

The ultimate goal of our research effort is to generate mitochondrial genome sequence information from human DNA samples. In order to accomplish this goal, we have employed PCR amplification reactions to generate sufficient template of mtDNA for NGS.

The experiment can be broadly divided into wet lab and dry lab. Till library preparation it was conducted in a wet lab and the data quality check and analysis was conducted in a bioinformatics facility. The NGS application requires versatility when it comes to assay development. The assay needs to be tested and validated multiple times before offering it as a product.

The current work has showed effective results; however, it needs to be further optimized to get rid of the nonspecific amplification in MTL-1 primer. In conclusion, this work had helped me understand the overall next generation sequencing workflow and its applications.

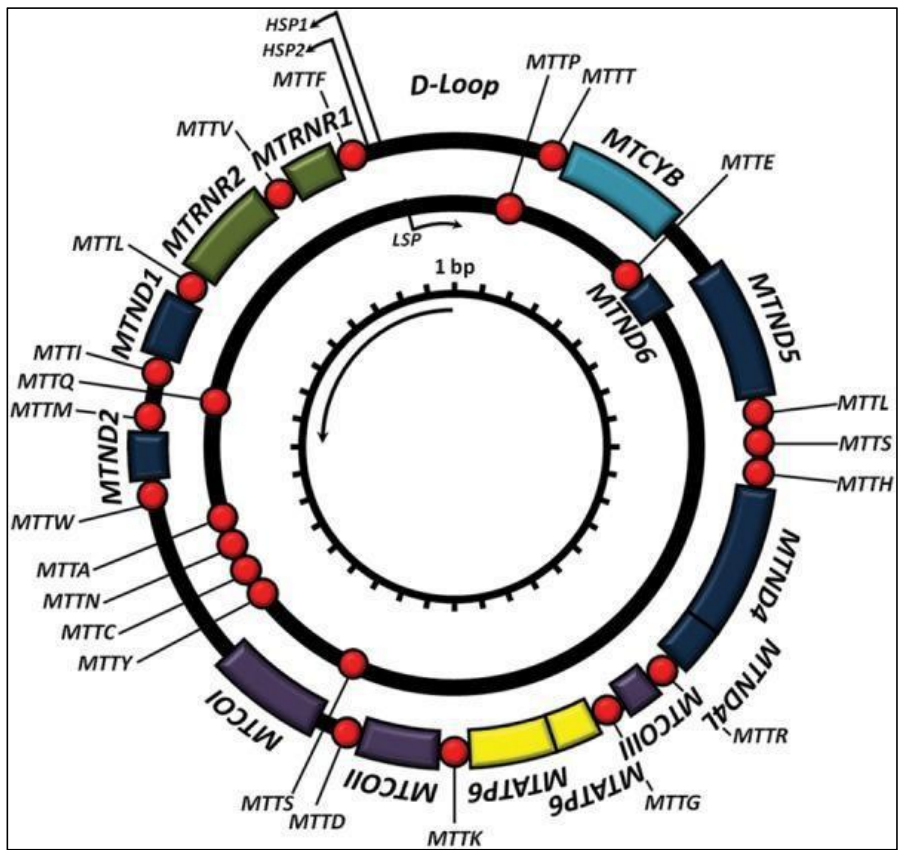
## 1. Introduction

The mitochondrion is a highly specialized organelle, within cells that converts energy from food into a form that cells can use. They are principally charged with the production of cellular energy through oxidative phosphorylation. This process uses oxygen and sugars to create adenosine triphosphate (ATP). A set of enzyme complexes, designated as complexes I-V, carry out oxidative phosphorylation within mitochondria. In addition to energy production, mitochondria play a role in several other cellular activities, for example, mitochondria help regulate apoptosis, calcium signaling, regulation of cellular metabolism and steroid synthesis [1]. They are also necessary for the production of substances such as cholesterol and heme.

Although most DNA is packaged in chromosomes within the nucleus, mitochondria also have a small amount of their own mitochondrial DNA (mtDNA) [2]. However, the simplistic elegance of biochemical adenosine triphosphate (ATP) production belies a, complex, synergistic relationship between two genomes: the mtDNA and the nuclear genome (nDNA) [3].

In eukaryotes, mtDNA is 16.6 kb, circular double-stranded DNA molecule. (Fig.1) [2]. The two strands of mtDNA are distinguished by their nucleotide composition; Guanine rich heavy (H-strand) and cytosine-rich light strand (L-strand). mtDNA length varies between species (15 000–17 000 bp) [4]. Each cell contains more than one mitochondrion and within each mitochondrion multiple copies of mtDNA exist ranging up to 20 [5].

mtDNA contains 37 genes, 28 on the H-strand and 9 on the L-strand. Thirteen of the genes encode polypeptide components of the mitochondrial respiratory chain (RC). Among 13 genes, 3 subunits are of cytochrome c oxidase (also called complex IV), 7 subunits of NADH-CoQ reductase (also called complex I), 2 subunits of F<sub>0</sub> ATPase (part of the ATP synthase complex, also called complex V), and the cytochrome b subunit of CoQH<sub>2</sub>-cytochrome c reductase. Twenty-four genes encode mature RNA products i.e, 22 tRNA molecules, 16s rRNA (large ribosomal subunit) and 12s rRNA (small ribosomal subunit) [6].



**Figure 1.** Mitochondrial DNA. schematic diagram of the 16.6-kb, circular, double-stranded mtDNA molecule, where the outer circle represents the heavy strand and the inner circle the light strand. Shown are the genes encoding the mitochondrial RC: *MTND1–6*, *MTCOI–II*, *MTATP6* and *8* and *MTCYB*; the two ribosomal RNAs (green boxes) and each of the 22 tRNAs (red spheres) [2].

mtDNA is extremely efficient when it comes to packaging genes, with ~93% representing coding region [7]. mtDNA genes lack intronic regions and some genes, notably *MTATP6* and *MTATP8*, have overlapping regions. Most genes are contiguous, separated by one or two non-coding base pairs [8]. mtDNA contains only one significant non-coding region i.e. the displacement loop (D-loop). The D-loop is the site of mtDNA replication initiation (origin of heavy strand synthesis, OH) and is also the site of both H-strand transcription promoters (HSP1 and HSP2) [9].

The concepts, which distinguish mitochondrial genetics (which is population genetics) from mendelian genetics, are highly relevant to the etiology and pathogenesis of mitochondrial respiratory chain disorders [10]. Human mitochondrial DNA is inherited strictly from mothers. Thus, barring mutation, a mother passes along her mtDNA type to her children, and therefore siblings and maternal relatives have an identical mtDNA sequence. Since even distantly related maternal relatives should possess the same mtDNA type, this extends the number of useful reference samples that may be used to confirm the identity of a person [11].

Mitochondrial DNA acquires mutations at 6 to 7 times the rate of nuclear DNA because it lacks protective histones, it is in close proximity to the electron transport chain, exposed to high concentrations of free radicals which can damage the nucleotides, it lack DNA repair mechanisms, which results in mutant tRNA, rRNA and protein transcripts [12]. Mutations in mtDNA may lead to mitochondrial dysfunction, which manifests in a broad spectrum of diseases affecting various tissues like brain, heart, liver, and skeletal muscles. The clinical symptoms of the disease depend on various factors like the cell type affected, heteroplasmy levels of the mutated DNA [13]. Mitochondrial DNA mutations have been found in various cancers including breast, colon, and liver cancers.

The small genome size has enabled to understand the diversity of mitochondrial variations. This has been complimented by a number of informatics resources such as MITOMAP which have systematically curated information from various resources on mitochondrial variations [14].

**Table 1.** The list of disorders that are reported to be caused due to different gene alterations in the mitochondrial genome [14].

<b>Disorder</b>	<b>Gene</b>	<b>Gene Name</b>
Leigh syndrome	MTATP6	mitochondrial ATP synthase membrane subunit 6
	MTND3	mitochondrial NADH dehydrogenase 3
	MTND4	mitochondrial NADH dehydrogenase 4
	MTND5	mitochondrial NADH dehydrogenase 5
	MTCO3	mitochondrial cytochrome C oxidase 3
	MTTW	mitochondrial tRNA tryptophan
Dystonia	MTND1	mitochondrial NADH dehydrogenase 1
	MTND6	mitochondrial NADH dehydrogenase 6
MELAS (Mitochondrial encephalomyopathy, lactic acidosis, and stroke-like episodes)	MTND1	mitochondrial NADH dehydrogenase 1
	MTND4	mitochondrial NADH dehydrogenase 4
	MTND5	mitochondrial NADH dehydrogenase 5
	MTND6	mitochondrial NADH dehydrogenase 6
	MTCYB	mitochondrial cytochrome b
	MTCO3	mitochondrial cytochrome C oxidase

	MTTL1	mitochondrial tRNA leucine 1
	MTTQ	mitochondrial tRNA glutamine
Multisystem disorder	MTCO1	mitochondrial cytochrome oxidase 1
	MTCO2	mitochondrial cytochrome oxidase 1
MERRF (Myoclonic epilepsy with ragged-red fibers)	MTTK	mitochondrial tRNA lysine
	MTTG	mitochondrial tRNA glycine
	MTTH	mitochondrial tRNA histidine
Exercise Intolerance	MTCYB	mitochondrial cytochrome b
	MTCO1	mitochondrial cytochrome oxidase 1
	MTCO2	mitochondrial cytochrome oxidase 2
	MTCO3	mitochondrial cytochrome oxidase 3
	MTTL1	mitochondrial tRNA leucine 1
	MTTA	mitochondrial tRNA alanine
	MTTS1	mitochondrial tRNA serine 1
	MTTE	mitochondrial tRNA glutamic acid

CPEO (Chronic Progressive external ophthalmoplegia)	MTTA	mitochondrial tRNA alanine
	MTTK	mitochondrial tRNA lysine
Diabetes Mellitus	MTND4	mitochondrial NADH dehydrogenase 4
	MTTL1	mitochondrial tRNA leucine 1
	MTTI	mitochondrial tRNA isoleucine
	MTTK	mitochondrial tRNA lysine
	MTTS2	mitochondrial tRNA serine 2
	MTRNR1DM	mitochondrial melatonin receptor 1D
Alzheimer & parkinson disease	MTND1	mitochondrial NADH dehydrogenase 1
	MTND2	mitochondrial NADH dehydrogenase 2
	MTTQ	mitochondrial tRNA glutamine
	MTRNR2AD PD	mitochondrial nucleophosmin 2A
Idiopathic Sideroblastic Anemia	MTCO1	mitochondrial cytochrome oxidase 1
Maternally Inherited Hypertrophic Cardiomyopathy	MTTI	mitochondrial tRNA isoleucine
	MTTW	mitochondrial tRNA tryptophan



		MTTK	mitochondrial tRNA lysine
		MTTG	mitochondrial tRNA glycine
		MTTH	mitochondrial tRNA histidine
Fatal Cardiomyopathy	Infantile	MTTI	mitochondrial tRNA isoleucine

Advancements in DNA sequencing were accompanied by new methods for fast and efficient isolation, amplification, assembly and annotation of mtDNAs. For example, enhanced long-range polymerase chain reaction (PCR) techniques allowed for the amplification of entire mitochondrial DNA, the results of which could then be sequenced stepwise using a ‘primer walking’ approach [9]. As sanger sequencing technique improved, it became easier and more affordable, especially for small laboratory groups, to sequence entire mtDNAs. It is valid for heteroplasmy quantification for heteroplasmies  $\geq 10\%$  but any variant lower than this is difficult to detect with Sanger sequencing [11].

The game changer in mitochondrial genomics, was the introduction of massively parallel sequencing platforms, which are cheaper, faster and can generate more data than sanger based methods [7]. The availability of NGS technologies has provided mitochondrial genome sequences with high coverage, thereby enabling decoding of a number of human mitochondrial diseases. In NGS technology, large depth coverage, could be effectively used to map heteroplasmic variations [15]. It is capable of reliably detecting and quantifying heteroplasmies down to the 1% [12].

With the advent NGS accuracy of diagnosing mitochondrial disorders has improved many folds. A simplified amplification and library preparation procedure can reduce the cost of such kind of test. Considering this, the current work is aimed to design and standardize library preparation work flow for human mitochondrial sequencing.

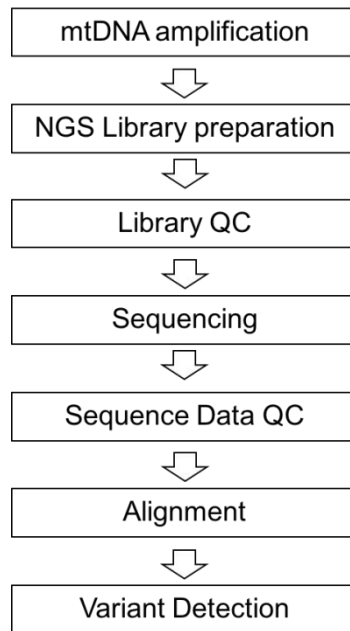
## **Scope**

- To standardize the laboratory protocol for sequencing human mitochondrial DNA using Illumina Next Generation Sequencing platform.

## **Objectives**

- Standardizing long PCR reaction to generate two overlapping amplicons of human mtDNA.
- Illumina sequencing library preparation using the mtDNA amplicons.
- Sequencing the mtDNA library.

## 2. Methodology



**Figure 2.** The experiment work flow.

The experiment was divided into four parts steps for the ease of execution. These steps are

1. Library Preparation: The sequencing library was prepared by random fragmentation of the mtDNA amplicons followed by 5' and 3' sequencing adapter ligation.
2. Library quality and quantity check to ensure that the library is able to generate good data.
3. Sequencing: Illumina next generation sequencing technology was used to generate high quality paired end reads [14-15].
4. Data Analysis: Data quality check and filtering were performed. QC passed sequence reads were aligned to the reference mitochondrial genome and SNP, insertion deletion (indel) identification were performed.

## Literature mining

As first step published scientific literature was mined to check for procedure followed to analyze mtDNA using NGS technology. , NCBI pubmed and google scholar were searched using combination of keywords “human”, “mtDNA”, “next generation sequencing”, “illumina”, “primer”, “amplification”. The primers for the experiment (Table 2) were designed by Clevergene Biocorp pvt. Ltd.

**Table 2.** Primers used to amplify human mtDNA

Primer ID	Sequence
MTL-F1	5’ - AAA GCA CAT ACC AAG GCC AC -3’
MTL-R1	5’ - TTG GCT CTC CTT GCA AAG TT -3’
MTL-F2	5’ - TAT CCG CCA TCC CAT ACA TT -3’
MTL-R2	5’ - AAT GTT GAG CCG TAG ATG CC -3’

### 2.1 DNA Source

CEPH 1347-02 DNA (Catalog no. 403062, Thermo Fisher Scientific) was used for this study. CEPH individual 1347-02 DNA was used, as it was widely genotyped, and data was available for comparison. This data can be used to check the quality of this experiment results.

### 2.2 DNA Quality Check

The DNA quality with reference to its fragmentation, was checked by 1% agarose gel electrophoresis. If the DNA is fragmented it would form a smear rather than an intact high molecular weight band.

### **2.2.1. Agarose Gel Electrophoresis**

#### **Principle:**

Agarose gel electrophoresis is the easiest and most popular way of separating and analyzing DNA, RNA and proteins [12]. Here DNA molecules are separated on the basis of charge and size by applying an electric field to the electrophoretic apparatus. The migration rate of the linear DNA fragments through agarose gel is proportional to the voltage applied to the system. As voltage increases, the speed of DNA also increases. But voltage should be limited because it heats and finally causes the gel to melt. Shorter molecules migrate more easily and move faster than longer molecules through the pores of the gel and this process is called sieving. Agarose makes an inert matrix and most agarose gels are made between 0.7% and 2% of agarose. A 1% gel will show good separation for large DNA fragments (5-10kb) and a 2% gel will show good resolution for small fragments with size range of 0.2-1kb. Low percentage gels are very weak but high percentage gels are usually brittle and do not set evenly.

In order to visualize nucleic acid molecules in agarose gels, ethidium bromide or SYBR Green are commonly used. They are intercalating agents which bind to DNA by inserting themselves between the stacked bases in double-stranded DNA and allows the convenient detection of DNA fragments in gel when exposed to specific wave length of light. After the running of DNA through a gel, any band containing more than ~20 ng DNA becomes distinctly visible under.

The most commonly used buffers for DNA electrophoresis are Tris-acetate-EDTA (TAE) and Tris-borate-EDTA( TBE). The migration rate of DNA fragments in both of these buffers is somewhat different due to the differences in ionic strength. These buffers provide the ions for supporting conductivity.

## **Procedure**

### **Running Buffer Preparation**

1litre of 1X TAE Buffer was prepared from 50X TAE Buffer (Catalog no. B49, Thermo Scientific) by initially taking 600ml of distilled water in a measuring jar and adding 20 ml of 50X TAE Buffer and then the volume was made up to 1000 ml. From the 1X TAE buffer 700 ml was poured into the electrophoresis tank (midi submarine gel system, Genaxy).

### **Agarose Gel Preparation**

For preparing 1% Agarose gel, 0.3 gm of SeaKem LE Agarose (Catalog no. 50004, Lonza) was mixed with 30ml of 1X TAE Buffer along with 3  $\mu$ l of 10,000X SYBR Safe DNA Gel stain (Catalog no. S33102, Thermo Fisher Scientific). The mixture was heated in microwave for 5 minutes until agarose was melted and no granules were visible. The molten agarose gel was kept at room temperature for a while. Gel casting tray was fixed and combs were fitted. Molten agarose was poured into the tray and let it at room temperature for 30 minutes. After solidification the gel was kept in the electrophoresis tank and combs were removed with care.

### **Preparation of 6X Loading Dye**

For preparing 5ml of 6X loading dye, 1.5gms of 30% (v/v) Glucose, 25mg of 0.25% (w/v) Bromophenol blue and 25mg of 0.25% (w/v) Xylene cyanol FF were added to nuclease free water. Stock was stored at 4<sup>0</sup>C.

### **Sample Loading**

DNA sample and ladder (Lambda DNA-Hind III Digest, NEB) were loaded in the well after mixing them with 6X loading dye and run at 150V for 20 mins. Gel was then visualized on blue light (Catalog no. APSVE100, Lumix-box Blue Light LED epi-illuminator).

## 2.2.2 DNA Quantitation

### Principle:

Qubit dsDNA BR (Broad-Range) Assay Kit (Catalog no. 32853, Thermo Fischer Scientific, USA) was used to quantitate the DNA sample. The kit provides concentrated assay reagent, dilution buffer, and pre-diluted DNA standards. The concentrations were read using the Qubit 2.0 fluorometer (Invitrogen, Life Technologies). The Qubit dsDNA BR Assay Kit measurements give better indication of sample quantity than that produced by spectrophotometer. The Qubit fluorometer uses fluorescent dyes to determine the concentration of nucleic acids and proteins in a sample.

In Qubit 2.0 fluorometer each dye is specific for a type of molecule, DNA, RNA or protein. These dyes have extremely low fluorescence until they bind to their targets (DNA, RNA or protein). Upon binding, they become intensely fluorescent. The difference in fluorescence between bound and unbound dye is several orders of magnitude. Qubit DNA dye used for the high sensitivity assay has extremely low fluorescence until it binds to DNA. Upon binding to DNA, probably by intercalation between the bases, it assumes a more rigid shape and becomes intensely fluorescent [9-10]. Once added to a solution of DNA, the Qubit DNA dye binds to the DNA within seconds and reaches equilibrium in less than two minutes.

At a specific amount of the dye, the amount of fluorescence signal from this mixture is directly proportional to the concentration of DNA in the solution. The Qubit fluorometer can pick up this fluorescence signal and convert it into a DNA concentration measurement using DNA standards of known concentration. The Qubit fluorometer uses DNA standards to derive the relationship between DNA concentration and fluorescence. It then uses this relationship to calculate the concentration of a sample, based on its fluorescence when mixed with the Qubit dye.



**Figure 3.** Qubit 2.0 fluorometer.

### **Procedure:**

#### Preparing Samples and Standards

1. Three 0.5ml tubes for 2 standards and a samples were taken.
2. Master mix was prepared in a 1.5 ml tube by adding 3  $\mu$ l Qubit dsDNA BR Reagent and 579  $\mu$ l (190 $\times$ 2+199)  $\mu$ l Qubit dsDNA BR buffer, diluting it to 1:200.
3. 190  $\mu$ l of Qubit working solution and 10  $\mu$ l of each Qubit standards (S2 and S2) was added to each of the tubes used for standards.
4. 199  $\mu$ l of working solution and 1  $\mu$ l of sample was added to the assay tube.
5. Mixed by vortexing for 2-3 seconds. Incubated for 2 mins at room temperature.
6. Assay tube was inserted into the instrument followed by 2 standards in the right order for callibration of the Qubit fluorometer.
7. Reading Standards and Samples
8. On the home screen of the Qubit® 2.0 fuorometer, pressed DNA, then selected dsDNA Broad Range as the assay type.



9. On the Standards screen, pressed yes to read the standards.
10. Tube containing Standard (S1) and Standard (S2) were placed into the sample chamber one after other. When the reading was completed tube was removed. Once the calibration is completed the concentration of samples was measured.
11. The concentration value was recorded in ng/ $\mu$ l.

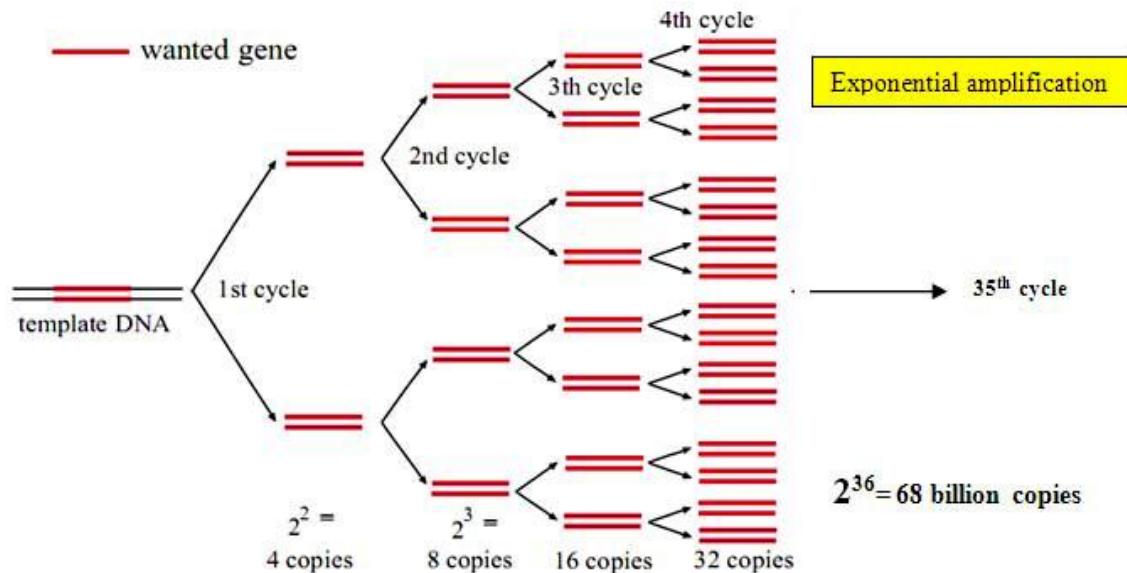
## **Primer Synthesis and Reconstitution**

The primers used in this work i.e, MTL-F1, MTL-F2, MTL-R1 and MTL-R2 were synthesized by Integrated DNA Technologies (IDT, Singapore). The primers were reconstituted to 100  $\mu$ M, diluting it with 1X TE Buffer (Ambion™ Thermo Fisher Scientific). Further reconstituted primers were diluted to 10  $\mu$ M (working concentration) with Nuclease free water. Stock primers were stored at -20<sup>0</sup>C for future use.

## **2.3 PCR Amplification**

### **Principle:**

Polymerase Chain Reaction (PCR) is used to amplify a target DNA fragment using short oligos that anneal to the region of interest due to complementarity. For amplification, DNA dependent DNA polymerase, dNTPs and buffer containing cofactors will be provided along with the template DNA. The reaction takes place in a thermal cycler which modulates the temperature and number of cycles. PCR typically include three temperature steps *viz.* denaturation, annealing and extension. During denaturation double stranded DNA will be denatures to single stranded molecule using high temperature i.e. >90 <sup>0</sup>C. Then the temperature will be reduced to ~ 55<sup>0</sup>C to anneal primers to the target region. Annealing temperature is depends of primer T<sub>m</sub> (meting temperature). Once the annealing is completed temperature will be changed based on the optimal polymerase working temperature. This denaturation, annealing and extension will be repeated 30-35 times (cycles) to obtain millions of copies of the amplified target or amplicon.



**Figure 4.** Representing the Polymerase Chain Reaction cycles

## Procedure

KAPA HiFi HotStart ReadyMix (#KK2600, Kapa Biosystems, South Africa) was used. HiFi HotStart DNA Polymerase is a ready mix format, containing all reaction components except primers (MTL-F1/MTL-F2 or MTL-R1/MTL-R2) and template DNA. The ready mix contained KAPA HiFi HotStart DNA Polymerase (0.5 U per 25  $\mu$ l reaction) in a proprietary reaction buffer containing dNTPs (0.3 mM of each dNTP at 1X), MgCl<sub>2</sub> (2.5 mM at 1X) and stabilizers. The PCR reaction was prepared as mentioned in table 3. Mixed well by pipetting. The tubes were placed in thermal cycler (Veriti 96 well Thermal Cycler, Applied Biosystems) and following programs were run.

**Table 3.** Reagents used in PCR reaction.

Reagents	Volume ( $\mu$ l)
DNA template (10 ng/ $\mu$ l)	5
Kappa Master mix	10
Diluted Forward Primer (10 $\mu$ M)	1
Diluted Reverse Primer (10 $\mu$ M)	1
Nuclease free water	3
Total Reaction volume	20

**Thermal Cycling conditions:**

Initial Denaturation : 5 minutes @ 95 °C

Denaturation : 15 secs @ 95 °C

Annealing : 10 secs @ 68 °C (slow ramping 0.2°C/sec from 68 °C to 60 °C )

15 secs @ 60 °C

Extension: 11 minutes @ 72 °C

Final Extension : 15 minutes @ 72 °C

Hold : 4 °C

### **2.3.1 Amplicon Check by Agarose Gel Electrophoresis**

PCR amplicons were visualized by loading 5  $\mu$ l of amplicon product in a 1% Agarose gel (0.3gms of agarose in 30ml of 1X TAE buffer). Run in 1X TAE buffer at 150V for 20 mins.

### **2.4 PCR Clean up**

To remove the leftover primers and primer dimers, a 0.9X clean-up of the amplicons were performed using Ampure XP Beads.

#### **Procedure**

1. 13.5  $\mu$ l of AMPure XP beads (Agencourt Bioscience Corporation, Beckman Coulter) were added to 15  $\mu$ l product.
2. Incubated at RT for 5 minutes. Placed the tube on magnetic stand till the solution became clear.
3. Discarded the supernatant, beads were washed with 200  $\mu$ l Ethanol (80%).
4. Beads were air dried to remove alcohol.
5. Eluted the amplicons in 20  $\mu$ l of 0.1X TE buffer.

### **2.5 Quantitation of PCR amplicons**

Qubit dsDNA HS (High-Sensitivity) Assay Kit (Catalog no. 32854, Thermo Fischer Scientific, USA) was used to quantitate the PCR product accurately using Qubit 2.0 Fluorometer.

### **2.6 Fragmentation**

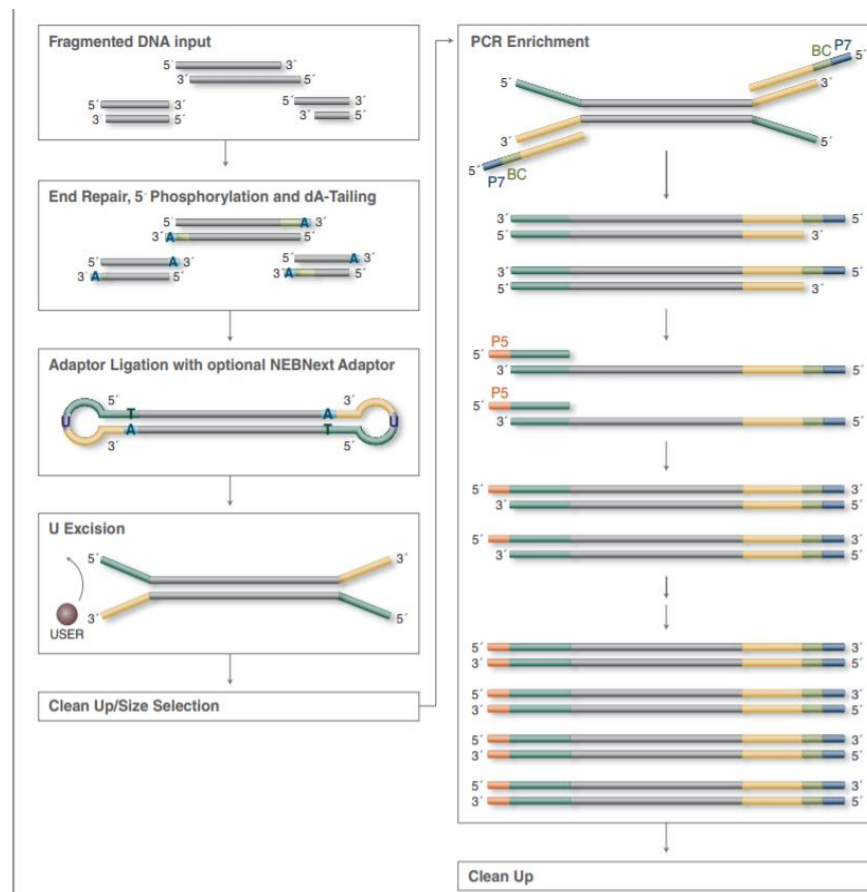
The Biruptor 200 sonicator was used to generate 350bp DNA fragments from the pooled amplicons. The burst was set to low and cycles were set for 30 seconds burst with 90 seconds pause. Ice cold water was used in the sonicator bath. Water was changed after every 3 cycles to maintain low temperature.

### 2.6.1 Fragment Check

Fragmented amplicons were visualized on a 2% Agarose run in 1X TAE buffer (40mM Tris-acetate, 1mM EDTA) at 150V for 20 mins.

### 2.7 DNA Library preparation

Sequencing library was generated using NEBNext Ultra II DNA Library Prep Kit for Illumina (New England Biolabs). The kit can generate high quality libraries from a broad range of input amounts (500 pg to 1 µg).



**Figure 5.** Workflow demonstrating NEBNext Ultra II DNA Library Prep Kit for Illumina

## **Procedure**

### **End Repair**

To a sterile nuclease free tube 25  $\mu$ l of Fragmented amplicon product along with 1.5  $\mu$ l NEBNext Ultra II End Prep Enzyme Mix and 3.5  $\mu$ l NEBNext Ultra II End Prep Reaction Buffer were added. The reaction was incubated at 20°C for 30 mins and heat inactivated at 65°C for 30 mins then held at 4°C followed by heated lid set to  $\geq$  75°C.

### **Adaptor Ligation**

15  $\mu$ l of NEBNext Ultra II Ligation Master Mix, 0.5  $\mu$ l NEBNext Ligation Enhancer and 1.25  $\mu$ l NEBNext Adaptor for Illumina were added to 30  $\mu$ l of End repair reaction mixture. The reaction was incubated at 20°C for 15 mins in a thermal cycler with the heated off. 1.5  $\mu$ l of USER Enzyme was added to the ligation mixture, mixed well and incubated at 37°C for 15 mins with the heated lid set to  $\geq$  50°C.

### **Size Selection/Cleanup of Adapter Ligated DNA**

43.5  $\mu$ l of Ampure XP beads were added to 48.25  $\mu$ l of Adaptor ligated DNA. The reaction was incubated at room temperature for 5 mins and then placed on magnetic stand. Clear supernatant was transferred to a fresh tube and added 82.5  $\mu$ l of Ampure XP beads. The reaction was incubated for 5 min at RT. Placed it on magnetic stand. The supernatant was discarded. Beads were washed twice with 200  $\mu$ l Ethanol (80%). The beads were then air dried. DNA was then eluted in 7.5  $\mu$ l of 0.1X TE buffer.

### **PCR Enrichment of Adapter Ligated DNA**

To 7.5  $\mu$ l of Adaptor Ligated DNA, 25  $\mu$ l NEBNext Ultra II Q5 Master Mix, 5  $\mu$ l Index Primer/i7 Primer, 5  $\mu$ l Universal PCR Primer/i5 Primer were added. The tube was placed in thermal cycler and cycling conditions were set to, initial denaturation at 98°C for 30 secs, 12 cycles of denaturation at 98°C for 10 secs, extension at 65°C for 75 secs, final extension at 65°C for 5 mins then hold at 4°C.

## **Clean up**

38.25  $\mu$ l of AMPure XP beads were added to 42.5  $\mu$ l PCR product. The reaction was incubated at RT for 5 min. The tube were placed on magnetic stand for 2mins. Supernatant were discarded. . Beads were washed twice with 200  $\mu$ l Ethanol (80 and air dried. The amplicons were eluted in 20  $\mu$ l of 0.1X TE buffer.

## **2.7.1 Library Quality Check**

### **Qubit fluorometric Quantitation**

Qubit dsDNA HS (Broad-Range) Assay Kit was used to quantitate the prepared library.

### **Library size estimation**

The library size if essential for sequencing because the number of molecule that are to be loaded into the flow cell needs to be estimated. In other words, molarity of the library needs to be calculated. To assess the library size, we can perform the gel electrophoresis but the size resolution is less in standard agarose gel when compared to polyacrylamide gels and capillary electrophoresis. To get best possible resolution of library size Bioanalyzer was used.

The Agilent 2100 Bioanalyzer is a micro-capillary based electrophoretic cell platform that allows rapid and sensitive investigation of nucleic acid samples. The system provides precise information on automated sizing, quantitation and fragment size distributions in a digital format.

### **Procedure**

1. DNA dye concentrate was vortexed and 25  $\mu$ l of the dye was added to DNA gel matrix vial.
2. The gel mix was vortexed and transferred to spin filter and centrifuged at 1500 g for 10 mins. Solution was protected from light and stored at 4 °C.
3. In the Bioanalyser DNA 7500 chip 9  $\mu$ l of gel-dye mix was added to the well marked **g**.
4. The chip was placed on the base of chip priming station and syringe was adjusted as per the manufacture's recommendation.
5. Plunger was positioned at 1ml and then closed the chip priming station. After 35 seconds clip was released. The plunger was slowly pulled back to 1 ml position after 5 seconds.

6. The chip priming station was opened and added 9  $\mu\text{l}$  of gel-dye mix in the other two wells marked as **g**.
7. 5  $\mu\text{l}$  of marker was added in the sample and ladder wells.
8. 1  $\mu\text{l}$  of DNA ladder and sample were added to their respective wells. 1  $\mu\text{l}$  of nuclease free water was added to the remaining wells.
9. Chip vortexed for 1min.
10. Lid of the Agilent 2100 Bioanalyzer and chip was placed carefully into the receptacle and closed the lid.
11. The 2100 expert software screen displayed the chip icon at the top left of the Instrument context that implied that a chip has been inserted.
12. In the Instrument context, selected the appropriate assay from the Assay menu and entered sample information.
13. Clicked the start button.
14. After the run was finished, chip was removed from the receptacle of the bioanalyzer.
15. One of the wells of the electrode cleaner was filled with 350  $\mu\text{l}$  deionized analysis-grade water and placed the electrode cleaner in the Agilent 2100 Bioanalyzer.
16. Closed the lid and left it for about 10 seconds.
17. Opened the lid and removed the electrode cleaner.



## **2.8 Sequencing**

To sequence mtDNA with 100X depth we would need 1.6 Mb of data. Illumina MiSeq instrument is used to generate sequence data.

The MiSeq reagent cartridge is a single-use consumable consisting of foil sealed reservoirs prefilled with clustering and sequencing reagents sufficient for sequencing 1 flow cell. Each reservoir on the cartridge is numbered.

### **2.8.1 Library Dilution and Denaturation**

#### **Procedure**

1. Prepared a fresh dilution of 1N NaOH by mixing 900  $\mu$ l nuclease free water and 10 M NaOH (100  $\mu$ l) in a 1.5 ml micro-centrifuge tube. The tubes were inverted several times to mix.
2. From 1N NaOH it was further diluted to 0.2N NaOH, by mixing 200  $\mu$ l of 1N NaOH and 800  $\mu$ l of Nuclease free water.
3. The library was diluted to 4 nM based on the Bioanalyzer and Qubit results.
4. Mixed 5  $\mu$ l of 4nM library and 5  $\mu$ l of 0.2N NaOH in a micro-centrifuge tube.
5. Vortexed briefly and then centrifuged at 280 $\times$ g for 1 min.
6. Incubated at room temperature for 5 mins.
7. Added 990  $\mu$ l chilled HT1 to the tube containing denatured library. The result is 1ml of a 20 pM denatured library.
8. The library was pooled with other sequencing libraries as per the manufacturer's instructions.
9. 3% PhiX control library was spiked in the final library pool.
10. 600ul of the final diluted library pool was loaded onto the miseq reagent cartridge in position 17, which is labeled Load Samples.
11. The reagent cartridge and flow cell were loaded onto the MiSeq instrument as per the manufacturer's recommendations.

12. The sample sheet containing run parameters and sample index barcodes was loaded in MiSeq Control Software (MCS) and run was performed.

13. The run was monitored with the use of Illumina Sequence Analysis Viewer (SAV) software.

14. Once the run is completed the mandatory MiSeq wash was performed after replacing the reagent Cartridge with wash cartridge containing distilled water.

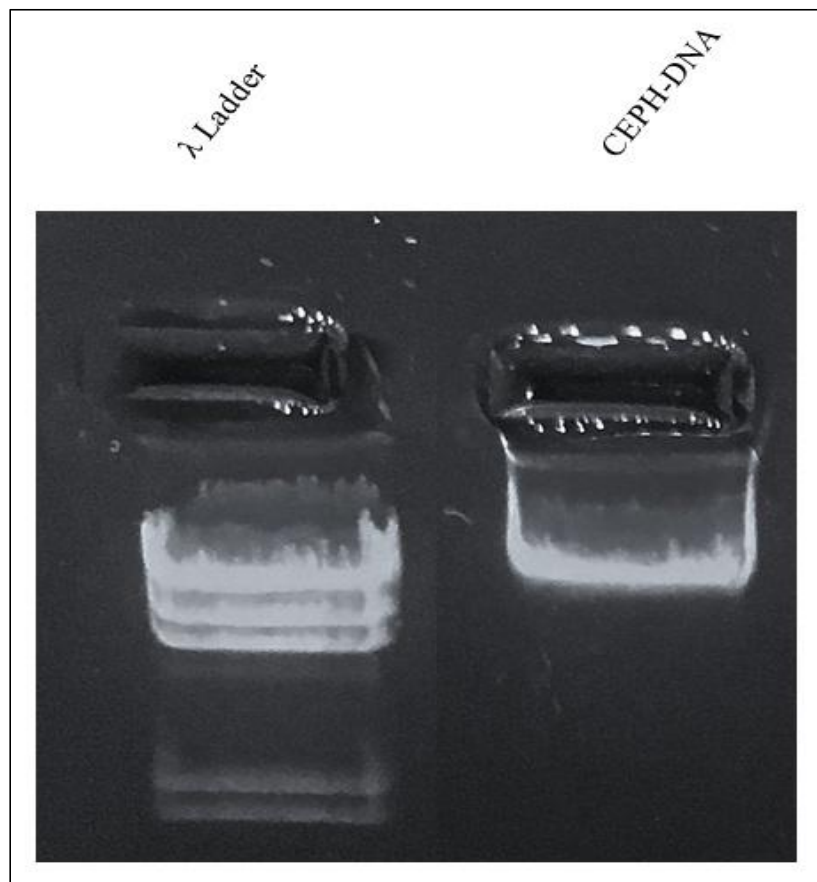
## **2.9 Data Analysis**

The sequence data quality was checked using FastQC and MultiQC software. The sequence reads were mapped to revised Cambridge Reference Sequence (rCRS) using Bowtie 2.

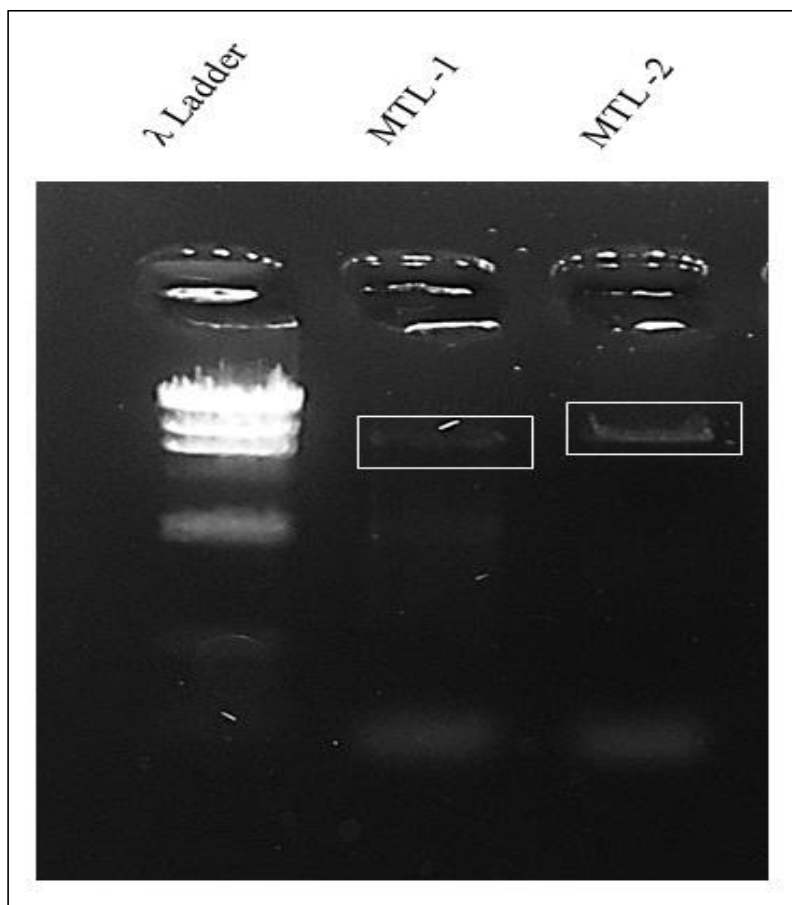
## Results

### DNA Quality Check

The CEPH DNA showed intact high molecular weight band (fig 6) on 0.8% agarose gel, indicating no degradation and suitable for mtDNA amplification. As per the Qubit 2 reading the concentration of the given CEPH-DNA was found to be 10 ng/ $\mu$ l. The expected amplicon size for MTL-1 and MTL-2 were 9.1 kb and 11.2 kb respectively. There were DNA bands in the gel at expected size range (Figure 7). However, MTL-1 showed nonspecific amplification. The goal of the study being sequencing the mtDNA, the non specific amplification do not interfere with the goal hence was not removed from the amplicons.

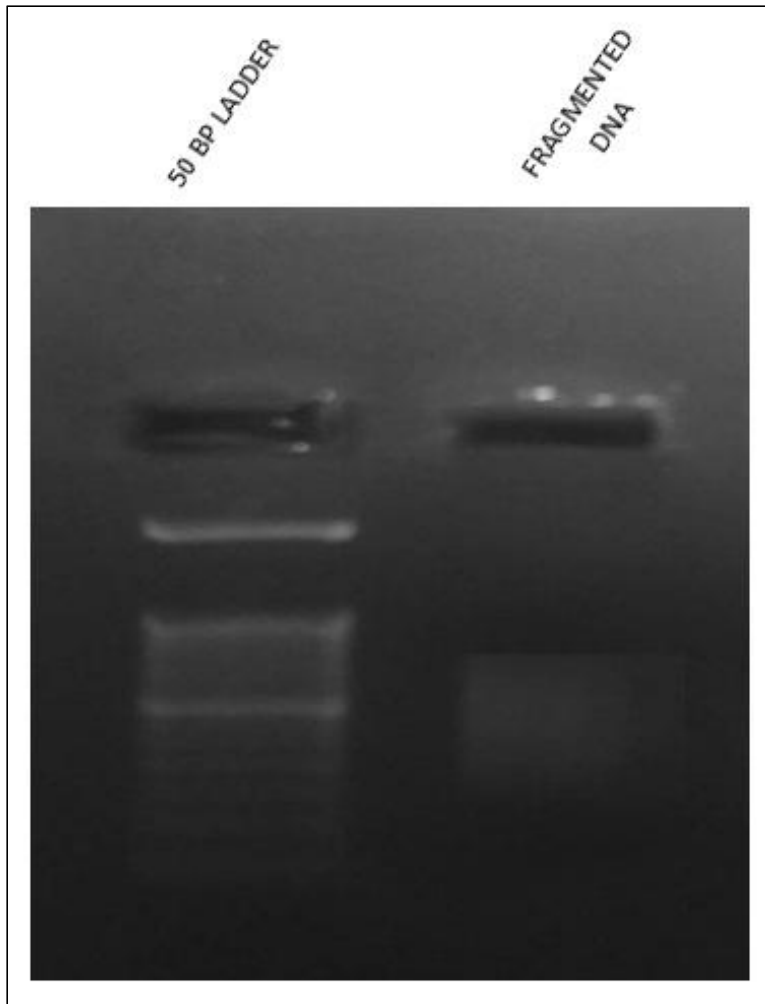


**Figure 6.** Gel Photograph showing band in 0.8% Agarose gel.



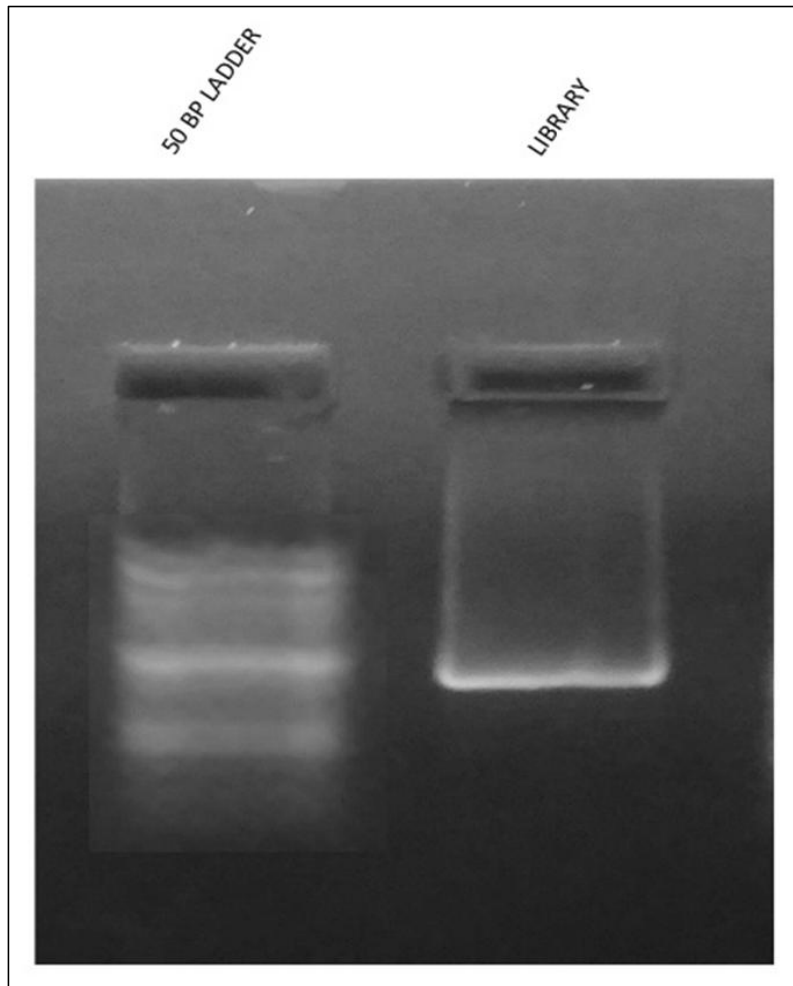
**Figure 7.** Gel picture showing PCR bands in lane 2 and lane 3 in 0.8% Agarose gel

The fragmented DNA was in the range of 300-500 bp (Figure 8) which was suitable for Illumina sequencing library preparation.



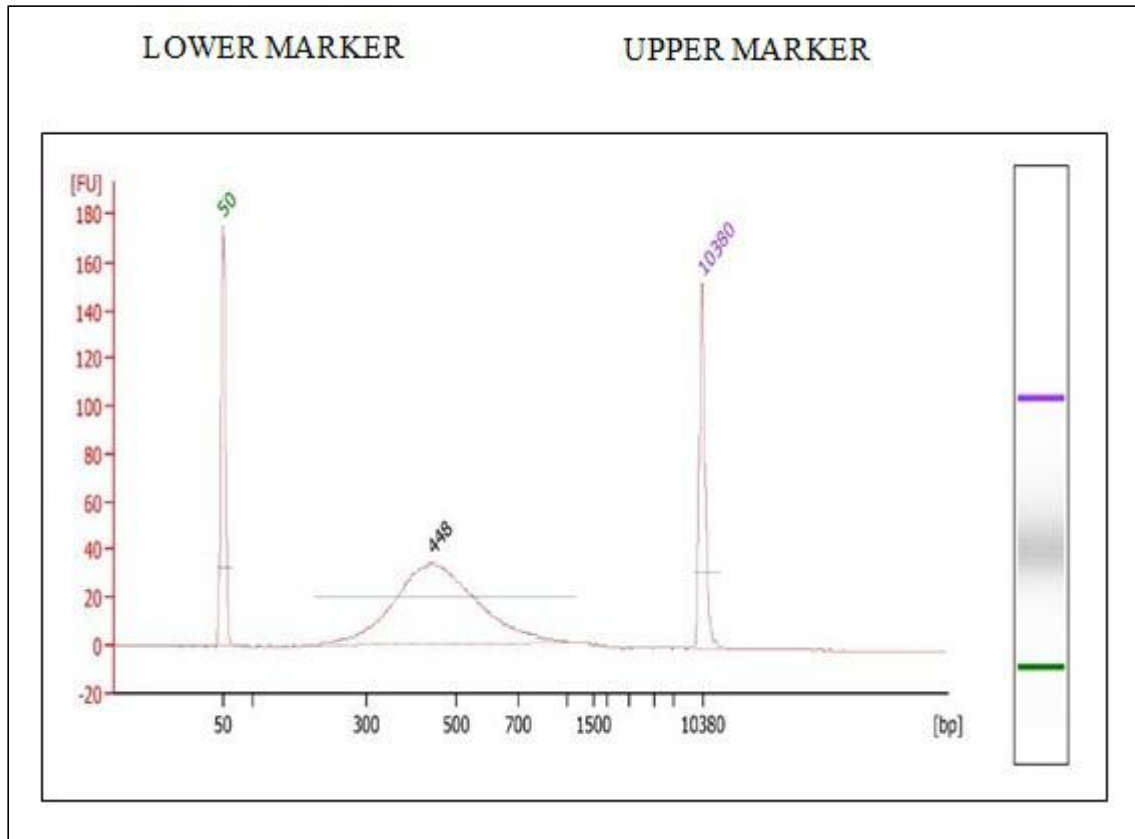
**Figure 8.** Gel photograph showing size range of DNA after fragmentation in 2% Agarose gel.

The sequencing library showed an intact band at 500 bp (Figure 9) indicating the library prep procedure was successful.



**Figure 9.** Gel Photograph showing final library size and integrity in 2% Agarose gel.

The sequence library showed no adapter contamination and good library size distribution in the Bioanalyzer electrophoresis (Figure 10). The library passed quality requirement with respect to size and adapter content. As per Qubit 2 readings the library concentration was 17 ng/ $\mu$ l, which is more than required concentration hence the library passed both quality and quantity threshold.



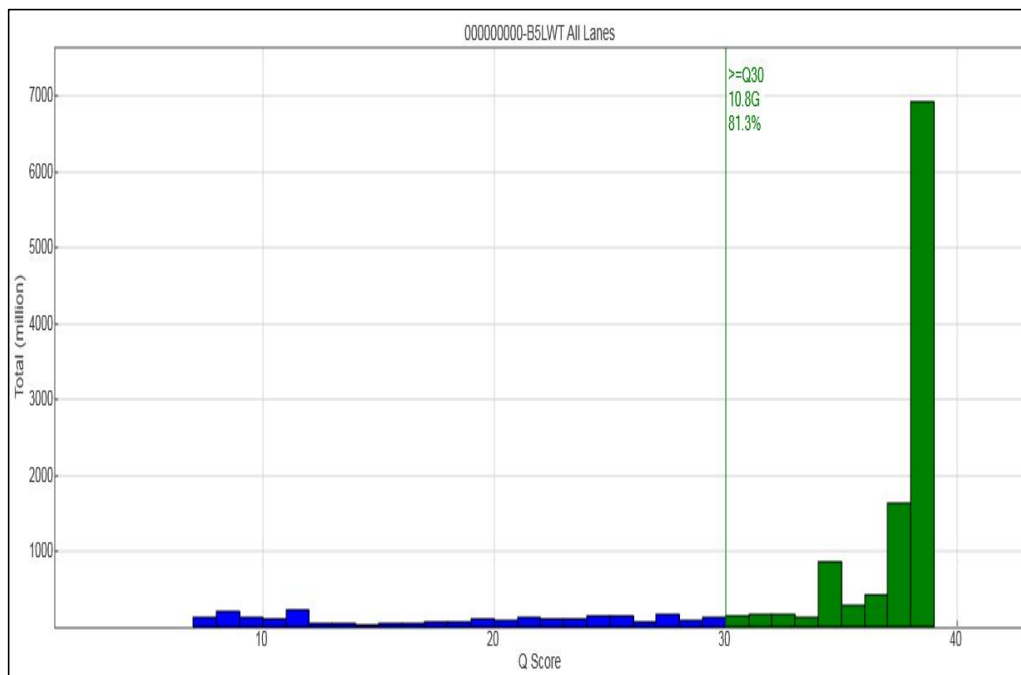
**Figure 10.** Bioanalyzer Electropherogram of final library using Agilent DNA 7500 Kit showing average size of library to be 448 bp

## Sequencing

In addition to mtDNA library other libraries were also included in the sequencing run to utilize the flow cell capacity to the fullest. The number of clusters were  $983 \pm 13$  k/mm<sup>2</sup> (Figure 11) indicating good performance of the libraries and optimal clustering. The total yield of the run is 13.19 giga bases and more than 80% of the data has quality score > 30 (Figure 12). The quality parameters of the run, error rate and phasing/pre-phasing were within the recommended thresholds. After demultiplexing, the mtDNA library yielded 179335 paired end reads which is equivalent to 6494 X depth of mtDNA.



**Figure 11.** Photograph showing cluster densities



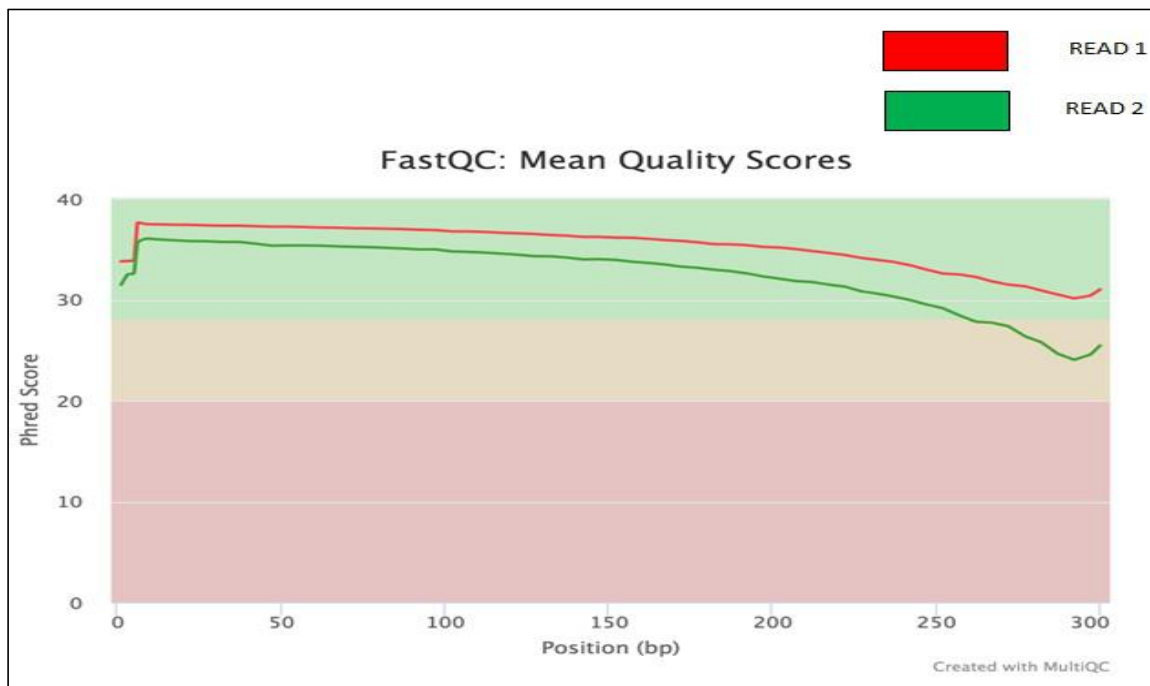
**Figure 12.** Q Score distribution Plot



FastQC and MultiQC analysis revealed that the data is of good quality, 81.3% of run data has Qvalue greater than Q30. Majority of the reads having quality score > 30 and read length > 290 (Figures 12 and ).

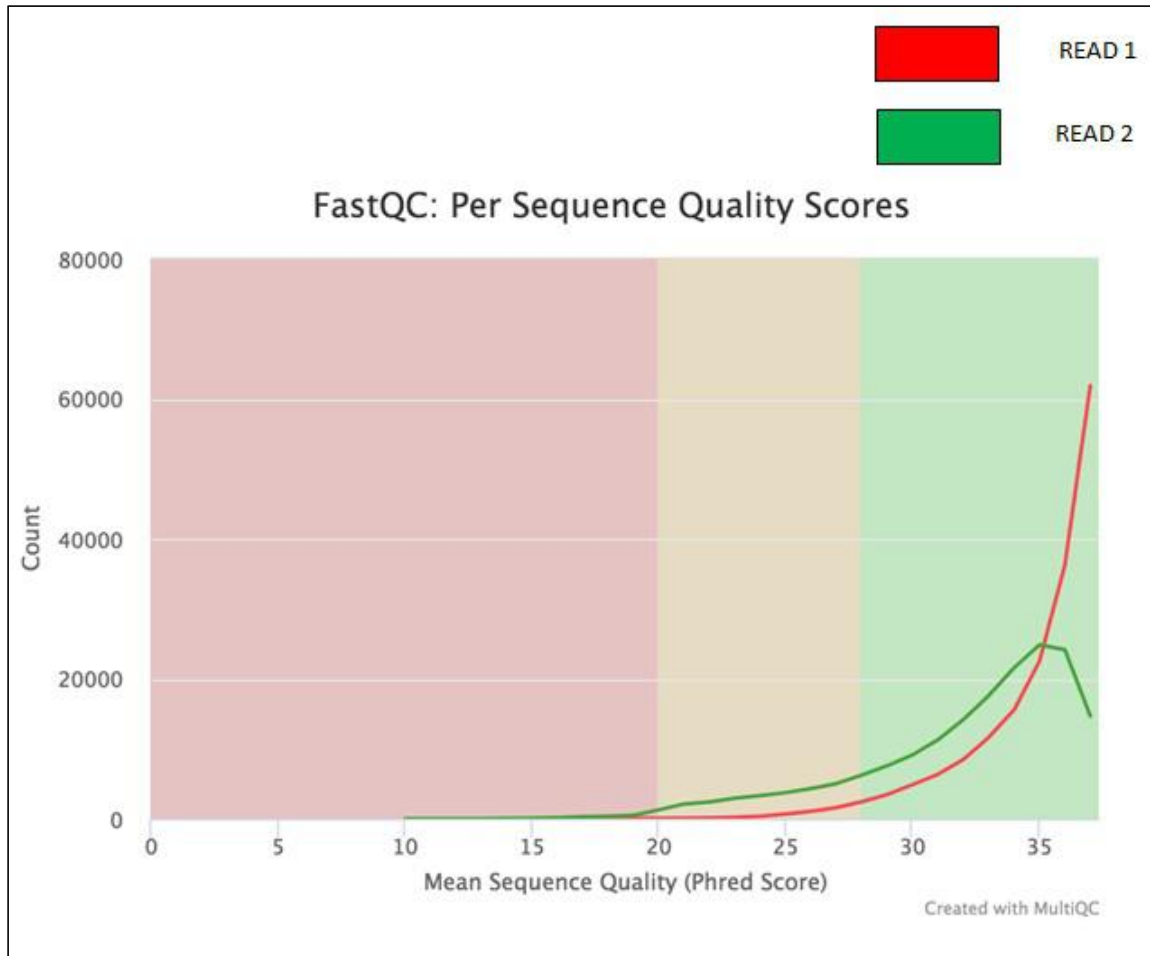
85.39% of reads mapped onto rCRS indicating the data was generated from the mitochondrial DNA (Figure 17).

The y-axis on the graph shows the quality scores (Figure 13). The higher the score the better the base call. The background of the graph divides the y axis into very good quality calls (green), calls of reasonable quality (orange), and calls of poor quality (red). The quality of calls on most platforms will degrade as the run progresses, so it is common to see base calls falling into the orange area towards the end of read.



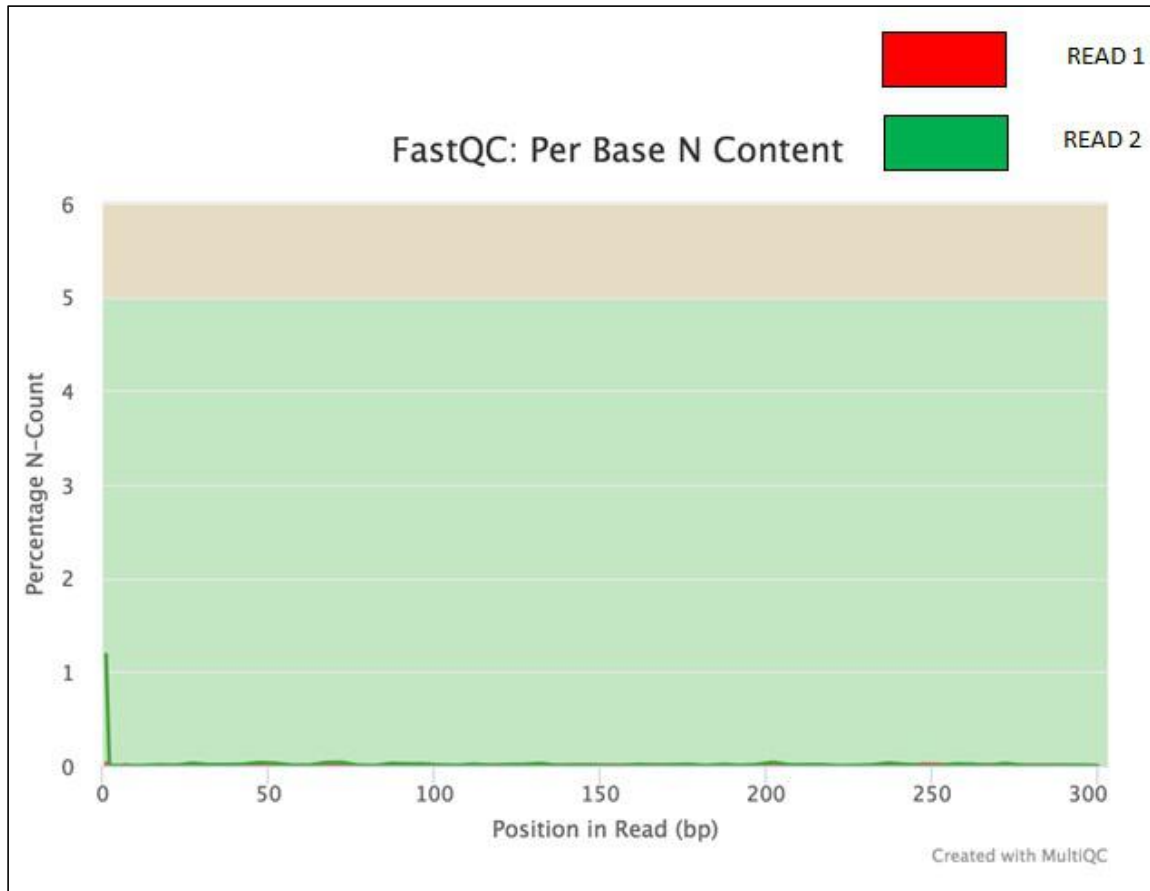
**Figure 13.** Mean quality score distribution across the read length.

Both Read 1 and Read 2 are within the green region indicating minimum base call accuracy of 99.9%.



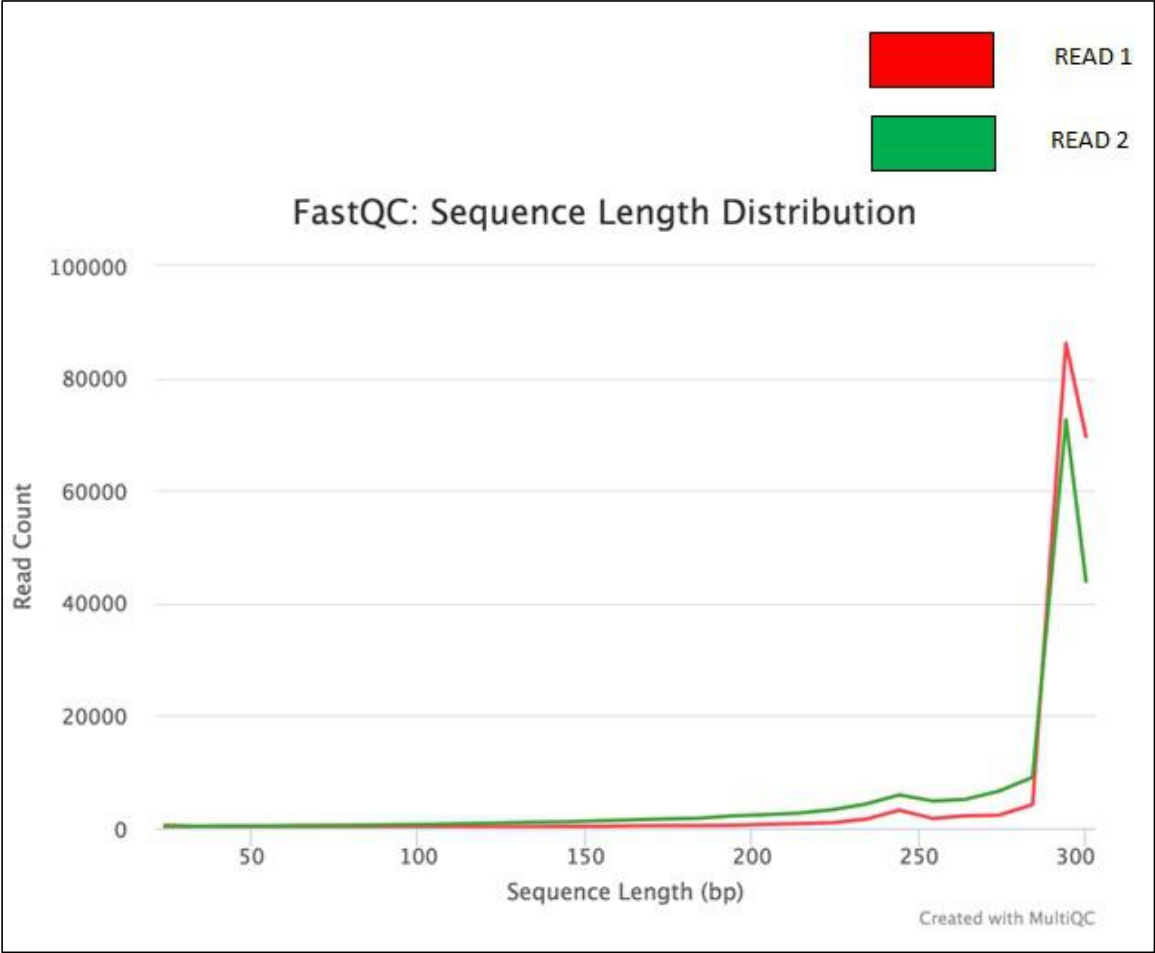
**Figure 14.** Quality score distribution of overall sequences.

The per sequence quality score report allows us to see if a subset of our sequences have universally low quality values. It is often the case that a subset of sequences will have universally poor quality, often because they are poorly imaged (on the edge of the field of view etc), however these should represent only a small percentage of the total sequences.



**Figure 15.** N content across all bases.

If a sequencer is unable to make a base call with sufficient confidence then it will normally substitute an N rather than a conventional base call. This graph plots out the percentage of base calls at each position for which an N was called. It's not unusual to see a very low proportion of Ns appearing in a sequence, especially nearer the end of a sequence. However, if this proportion rises above a few percent it suggests that the analysis pipeline was unable to interpret the data well enough to make valid base calls.



**Figure 16.** Distribution of sequence lengths of all sequences.

After adapter removal and quality trimming the sequence length distribution was checked. The data shows majority of the reads are in the range of 280-300.



**Figure 17.** Screenshot of IGV showing reads mapped to rCRS. The histogram shows the depth of covered region.

## **What Did I learn**

Genetic/molecular diagnosis or screening is a fast growing field. The diseases associated with mitochondria can be identified quickly by their inheritance patterns and sequencing the mtDNA. Capillary electrophoresis (CE) sequencing requires time and relatively more amount of DNA when compared to Next Generation Sequencing which is able to identify heteroplasmy levels as low as 1%. NGS also has the capability of identifying the large insertion or deletions in mtDNA.

The scope of the current study was to standardize the laboratory protocol for sequencing human mitochondrial DNA using Illumina Next Generation Sequencing platform. NGS has a limitation when it comes to the read length, hence we need to generate a library which should be < 1000 bp in length. To achieve this the mtDNA was amplified to generate two large amplicons (9.1 Kb and 11.2 Kb) and these amplicons were fragmented to form 300-500 bp DNA fragments. These fragments were used as input for standard Illumina library preparation kit.

The library preparation procedure included, repairing the physically damaged DNA and ligating the adaptors and enriching the adapter ligated DNA fragments by PCR. Every step of experiment has quality checks, starting from DNA to Library preparation.

The experiment can be broadly divided into wet lab and dry lab. The procedure up to sequencing the library was conducted in a typical molecular biology laboratory but the data quality check and analysis was conducted in a bioinformatics facility. The NGS application requires versatility when it comes to assay development. The assay needs to be tested and validates multiple times before offering it as a product.

The current work has showed promising results; however, it needs to be further optimized to get rid of the nonspecific amplification in MTL-1 primer. The non specific amplification do not affect the data quality however it takes space in flow cell while clustering. The read that are generated by these non specific amplicons are of no use.

In conclusion, this work had helped me understand the overall next generation sequencing procedure and its applications. In addition, I was able to have a glimpse of bioinformatic tools used for NGS data analysis.

## References

1. van der Giezen M, Tovar J. Degenerate mitochondria. *EMBO Rep.* 2005;6:525–30.
2. Andrews RM, Kubacka I, Chinnery PF, et al. Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nat Genet.* 1999;23:147.
3. Schon EA, DiMauro S, Hirano M. Human mitochondrial DNA: roles of inherited and somatic mutations. *Nat Rev Genet.* 2012;13:878.
4. Schapira AH. Mitochondrial diseases. *Lancet.* 2012;379:1825–34.
5. Koopman WJ, Distelmaier F, Smeitink JA, et al. OXPHOS mutations and neurodegeneration. *EMBO J.* 2013;3.
6. Andrews RM, Kubacka I, Chinnery PF, et al. Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nat Genet.* 1999;23:147.
7. Temperley R, Richter R, Dennerlein S, et al. Hungry codons promote frame shifting in human mitochondrial ribosomes. *Science.* 2010;327:301.
8. Metzker ML. Sequencing technologies the next generation. *Nat Rev Genet* 2010;11:31–46.
9. Smith DR. Not seeing the genomes for the DNA. *Brief Funct Genomics* 2012;11:289–90.
10. King J, LaRue B, Novroski N, et al. High-quality and high-throughput massively parallel sequencing of the human mitochondrial genome using the Illumina MiSeq. *Forensic Sci Int Genet* 2014;12:128–35.
11. Seo SB, Zeng X, King JL, et al. Underlying data for sequencing the mitochondrial genome with the massively parallel sequencing platform Ion Torrent™ PGM™. *BMC Genomics* 2015.
12. Hert DG, Fredlake CP, Barron AE. Advantages and limitations of next generation sequencing technologies: a comparison of electrophoresis and non-electrophoresis methods. *Electrophoresis*, 2008;29:4618–26.
13. Anita Kloss-Brandstätter, Hansi Weissensteiner, Gertraud Erhart, Georg Schäfer, Lukas Forer, Sebastian Schönherr, Dominic Pacher, Christof Seifarth, Andrea Stöckl, Liane Fendt, Irma Sottas, Helmut Klocker, Christian W. Huck, Michael Rasse, Florian Kronenberg, Frank R. Kloss;

Validation of Next Generation Sequencing of Entire Mitochondrial Genomes and the Diversity of Mitochondrial DNA Mutations in Oral Squamous Cell Carcinoma , August 11, 2015.

14. Maitra ,et al., 2004. Advantages and limitations of next generation sequencing technologies:

15. Lott M.T Leipzig, J.N Derbeneva, O Xie H.M., Chalkia, D., Sarmady, M., Procaccio V and Wallace, D.C. 2013. mtDNA variation and analysis using MITOMAP and MITOMASTER. *Current Protocols in Bioinformatics*1(123):1.23.1-26. PMID: 25489354

14. Abaci et al 2014; Craigen et al 2013 Underlying data for sequencing the mitochondrial genome with the massively parallel sequencing platform.

15. Sosa, et al., 2012. Human mitochondrial DNA: roles of inherited and somatic mutations.

16. Nakazato T Ohta T Bono H. Experimental design-based functional mining and characterization of high-throughput sequencing data in the sequence read archive. *PLoS One*. 2013; 8(10) :e77910.

17. 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/>

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