Argueta, Wendy C. Comparison of Next Generation Sequencing Methodology on the Ion PGM™ System Performance versus that on the Sanger Sequencing Method for HV1 and HV2 Regions of mtDNA. Master of Science (Biomedical Sciences, Forensic Genetics). May 2015. 46 Pages, 7 tables, 7 figures, 41 references

Analysis of mitochondrial DNA in forensic applications has enabled the identification of a missing person through comparison with additional maternal relatives. Most forensic applications are based on sequencing of both hypervariable regions of the mtDNA. Sequencing of these regions has been commonly done using Sanger-type sequencing (STS) methodology, which is expensive, time-consuming and laborious. Next Generation Sequencing (NGS) technology, such as the Ion Torrent PGM™ System platform, overcomes most of these issues. In this study, samples from the Guatemalan population (n=40) were sequenced with both Ion Torrent PGM™ technology and STS methods. A high level of consistency (98%) was observed among data derived from both methods. Most of the discrepancies were point heteroplasmy, which were more easily detected by PGM™ technology. In terms of performance, the NGS method was shown to be quick, with high-throughput and more efficient compared to the traditional STS method. More accurate and reliable sequencing data were obtained from the Ion Torrent PGM™ method due to its high level of coverage. Sequencing data for all individuals, representing 19 different family groups, were obtained using the NGS technology. Sequence polymorphisms were detected in 55 positions, from which 26 were observed only in relatives belonging to the same family and were not observed for any other family group. In a forensic context, haplotype specific polymorphisms are the basis for identification and comparison between evidence and reference samples purposes. Haplotypes between maternally related individuals were consistent in 18 family groups.

Comparison of Next Generation Sequencing Methodology on the Ion PGM™ System Performance versus that on the Sanger Sequencing Method for HV1 and HV2 Regions of mtDNA

Wendy C. Argueta, B.S.

APPROVED:

_____

Major Professor

_____

Committee Member

_____

Committee Member

_____

University Member

_____

Arthur Eisenberg, Ph.D., Chair, Department of Molecular and Medical Genetics

_____

Meharvan Singh, Ph.D., Dean, Graduate School of Biomedical Sciences

COMPARISON OF NEXT GENERATION SEQUENCING METHODOLOGY ON THE ION PGM™ SYSTEM PERFORMANCE VERSUS THAT ON THE SANGER SEQUENCING METHOD FOR HV1 AND HV2 REGIONS OF mtDNA.

THESIS

Presented to the Graduate Council of the

Graduate School of Biomedical Sciences

University of North Texas

Health Science Center at Fort Worth

Partial Fulfillment of the Requirements

For the Degree of

MASTER OF SCIENCE

By

Wendy C. Argueta, B.S.

Fort Worth, TX

May 2015

ACKNOWLEDGEMENTS

I would like to thank Dr. Cristián Orrego for his motivation and help to start on the journey to get my Master´s degree at the University of North Texas Health Science Center. I am extremely grateful with my major professor, Dr. Arthur Eisenberg, for all his unconditional support, trust and guidance which helped me to achieve this step of my educational career.

I would also like to thank my committee, Dr. Michael Allen, Dr. Raghu Krishnamoorthy and Dr. Xiangrong Shi, for their suggestions and inputs to this project. I would especially like to thank Alessandra Alicea-Centeno, Jie Sun, Marc Sprouse and Elizabeth Mitchell for the guidance and advice dedicated to my training and laboratory work. To Linda LaRouse for her kindness and help to get everything I needed to carry out this project. To all my professors, I am thankful for all the knowledge and expertise imparted during this time. I would also like to thank my classmates for the support and experiences shared over the past two years, especially to Laura Guadian and Shantanu Shewale.

Lastly, I would like to express my gratitude to my family and dear friends for always believing in me and giving me the support and motivation to reach my goals.

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER I

INTRODUCTION

A. Background

Mitochondrial DNA (mtDNA) is an extrachromosomal genome located within the mitochondria. Its genome is separate and distinct from the nuclear genome [1]. Human mtDNA is a circular double-stranded molecule of 16,569 base pairs (bp) in length and it is histone free. It encodes for 13 polypeptides involved in the oxidative phosphorylation system, 2 ribosomal RNAs (rRNAs) and 22 transfer RNAs (tRNAs) (Figure 1). [1, 2]

Most of the mtDNA consists of coding DNA, with the exception of a non-coding region of approximately 1,100 bp long, which has mainly regulatory functions referred to as the control region [1]. Certain portions of the control region are highly polymorphic. [6]. Most of these sequence variations between individuals are found within two portions of the control region: hypervariable region 1 (HV1) and hypervariable region 2 (HV2). Typically, HV1 comprises positions 16,024 to 16,365 and HV2 encompasses positions 73 to 340 (Figure 1) [15]. The polymorphism in these regions allows the use of mtDNA in population, anthropology, evolution and forensic studies [16].

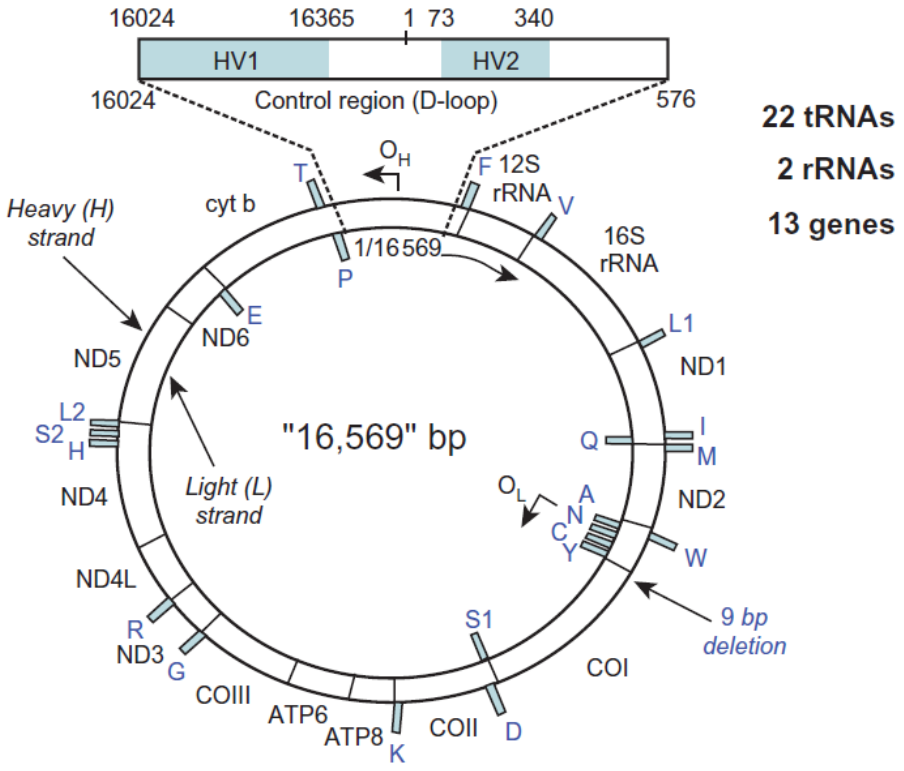

Figure 1 – Schematic illustration of the circular mtDNA genome. It has two different strands, the heavy (H) strand with a higher number of G residues than the light (L) strand. The total RNA (22 tRNAs and 2 rRNAs) and protein coding genes (13 genes) are abbreviated around the circular mtDNA. Most forensic analysis are carried out using HV1 and HV2 segments in the control region (shown at the top of the figure) [7].

Mitochondrial DNA has several properties making it a suitable molecule to be used in human population history, evolution and migration studies, as well as forensic analysis [5]. One of these properties is its high copy number in human cells [1]. Each somatic cell contains about 1,000 mitochondria, and each mitochondrion contains between 2 and 10 copies of mtDNA; therefore, hundreds to thousands of copies of mtDNA are present in each cell [2]. Along with this property, the fact that it is located outside the nucleus in the cell makes it is easier to obtain mtDNA than nuclear DNA. Additionally, its circular nature makes it less susceptible to exonucleases activity, making it less prone to degradation [7]. This is why mtDNA is the preferred molecule to use in

certain forensic applications such as the analysis of degraded samples, ancient DNA and identification of human remains [1].

Another unique property of mtDNA is its high mutation rate. It is several orders of magnitude higher than that of nuclear DNA, especially in both the hypervariable regions [6]. The higher mutation rate is associated with a low fidelity mtDNA polymerase, lack of repair mechanisms of mtDNA and its exposure to oxygen free radicals [16]. Because of the small size of each of these regions and their higher mutation rate, HV1 and HV2 are routinely typed for forensic testing purposes, such as human identity testing [2].

Unlike nuclear DNA, mtDNA is maternally inherited and is not subject to recombination events [3]. The mtDNA sequences between mother and child and among siblings must be identical unless mutations have occurred [6]. This characteristic allows mtDNA to be used to trace the maternal ancestry of a population [1] and comparisons of distant maternal relatives to evidence samples can be used to test a possible relationship or contributor of the sample [15].

Heteroplasmy is the condition where two or more mtDNA haplotypes, with different length or sequence, are present in a single individual [4]. Heteroplasmy can be found in biological material belonging to the same individual and/or inside maternal lineages [3]. The phenomenon of heteroplasmy can be due to maternal inheritance of heterogeneous mtDNA or by *de novo* generation of mtDNA during germ line development [16]. Nucleotide substitution caused by the majority of these events, where transitions (point mutation in which a pyrimidine is replaced by another pyrimidine or in which a purine is replaced by another purine) have shown to be more common than transversions (mutation in which a purine is replaced by a pyrimidine and vice versa) [7, 17]. The detection of heteroplasmy in the control region of the mtDNA is relevant in forensic analysis because it increases the power of discrimination of the haplotype [18].

Sequencing of mtDNA has become a commonly used technique for personal identification in forensic analysis [5]. During the 1990's a huge amount of data was generated from the sequence analysis of the HV1 segment of the control region. The recent development of high-throughput sequencing technology has allowed forensic sciences to sequence the whole mitochondrial genome [1]. Nevertheless, it is primarily the HV1 and HV2 regions of the mtDNA that are analyzed for forensic purposes [11, 19].

Sanger sequencing has been the gold standard sequencing technology for decades. This sequencing method is based on the use of dideoxynucleotides (ddNTPs) as inhibitors of the DNA polymerase activity during chain elongation due to the absence of a 3'- hydroxyl group. When a ddNTP is incorporated into the growing strand the chain cannot be extended further and termination occurs at the addition of that ddNTP [20]. Currently fluorescently labeled terminating nucleotides are used [21] allowing the detection of the different sequencing fragments by capillary electrophoresis (Figure 2) [22].

The general procedure for the generation of a mtDNA haplotype via Sanger sequencing involves the following steps: DNA extraction; DNA quantification and normalization; PCR amplification of control region [9]; yield gel run to determine if amplification of the desired fragment was achieved; post-PCR purification; cycle sequencing (nested or semi-nested PCR amplification of HV1 and HV2 regions) [8]; post-cycle sequencing purification; capillary electrophoresis; and data analysis [15]. If the data generated is of low quality it is necessary to go back and perform the cycle sequencing and downstream steps again; in some cases, it is even necessary to re-extract the sample. For this reason this method can become labor intensive [21] and expensive [11].

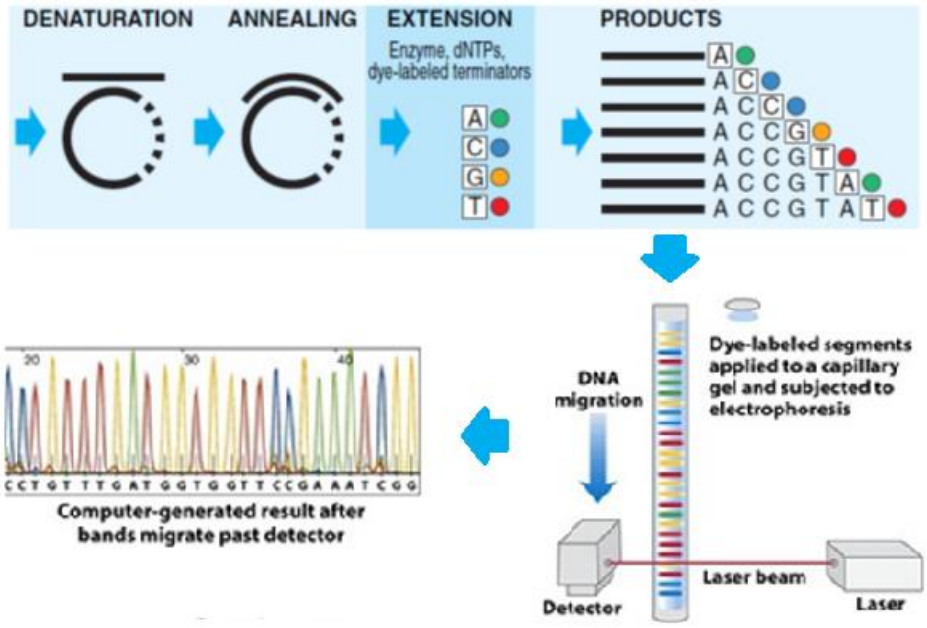

Figure 2 – Sanger dye terminator sequencing method. After DNA denaturation and annealing, the extension step is carried out in one reaction containing the polymerase, dNTPs and ddNTPs. Each ddNTP is tagged with a different fluorescent dye. The products obtained are injected into the Genetic Analyzer [22].

In recent years, next generation systems for DNA sequencing have been developed which provide major advantages over Sanger sequencing. The primary advantages include higher throughput and lower costs. Additionally, high-throughput platforms have a higher coverage giving a more reliable and accurate result compared to those obtained via Sanger Sequencing technology [23]. Coverage is one of the common measures of the amount of sequence data generated and it refers to the average number of times each base in the genome is sequenced [21].

Different platforms have been developed and each of them use different chemistries. For example, Ion PGM™ System is based on sequencing-by-synthesis where pH changes caused by the release of protons when nucleotides are incorporated into the growing strand complementary to the template DNA, are detected [10]. These chemical signals are converted into digital

information for further analysis [13]. Sequencing of mtDNA using the Ion Torrent PGM™ System involves the following general laboratory procedures: DNA extraction; PCR amplification of targeted DNA fragments; quantification of amplified product; PCR product purification; preparation of sequencing libraries; clean-up of amplified library; determination of amplified product yield; emulsion PCR; enrichment; sequencing using a chip; and data assembly and analysis [13, 24, 25].

The library sequencing construction involves different steps including: DNA fragmentation; adaptor ligation, size selection and PCR amplification of adaptor ligated DNA (Figure 3) [25]. During this phase different short indexing tags called "barcodes" are ligated to the amplified fragment of DNA from each sample to achieve proper differentiation between samples (Figure 3) [26, 27]. For this project, barcodes containing an indexed adapter with 10 nucleotides of unique sequence were used. In downstream steps, libraries for each sample are pooled together and then sequencing in parallel enabling large sample numbers to be sequenced simultaneously in a single run when using high- throughput sequencing instruments such as the Ion Torrent PGM™ Systems [28, 29]. The library construction process is critical to achieving the most genomic coverage and obtain a good quality sequencing data [29].

The optimal read length of the Ion PGM™ System ranges between 35 and 400 bp, which necessitates fragmentation and size selection steps when working with longer DNA fragments [30]. In this project a targeted PCR amplification of HV1 and HV2 regions for each samples was performed in order to avoid the addition of these steps to the NGS methodology. By performing targeted PCR amplification, DNA fragments of 432 bp and 399 bp were generated for HV1 and HV2, respectively (Figure 4).

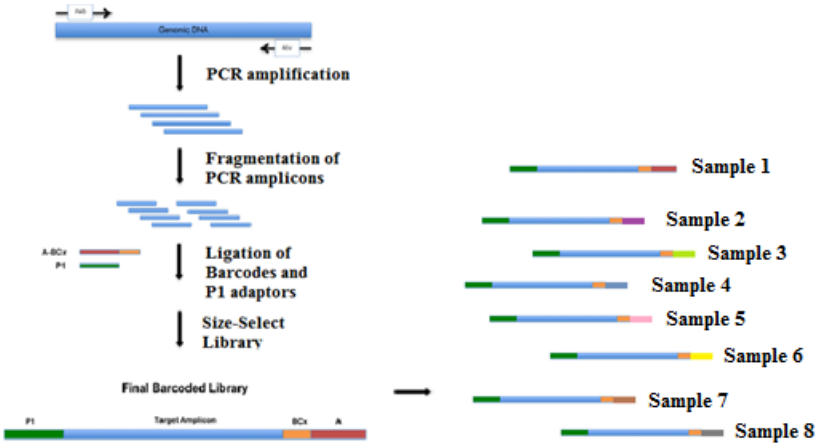

Figure 3 – Library construction process for Ion Torrent PGM^{TM} System. After PCR amplification, adaptors and individualizing barcodes are ligated to the targeted DNA fragments within each sample. Because each sample is fused to a specific barcode sequence, simultaneous sequencing of multiple samples can be achieved in one run [25].

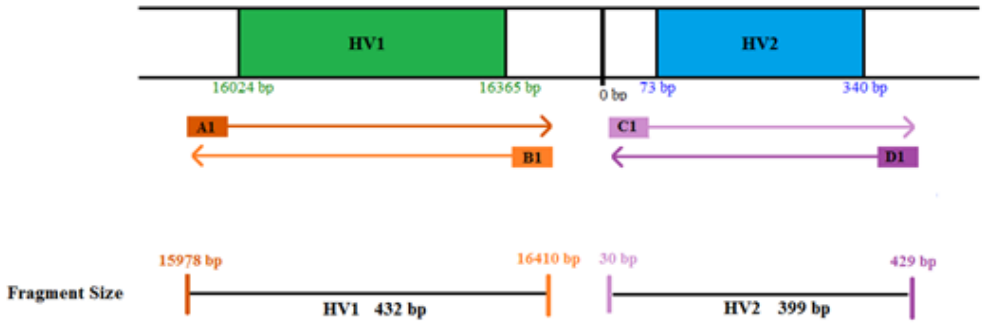

Figure 4 – Direct amplification of regions HV1 and HV2 separately. Annealing sites for two different sets of primers are shown for each region. The use of forward primer A1 and reverse primer B1 amplifies a fragment of 432 bp long for HV1. Amplification of region HV2 using forward primer C1 and reverse primer D1 generates a fragment of 399 bp long.

Once the library has been constructed, clonal amplification of each amplicon by a process called emulsion PCR is performed, by which singe-stranded DNA fragments are attached to the surface of beads called Ion Sphere Particles (ISPs). For this amplification step, primer coated beads with complementary sequences to the adaptor previously ligated to the DNA fragment are used [31, 32]. Water-oil droplets ideal for emulsion PCR are those containing one bead attached to one DNA strand. The droplets can then act as PCR microreactors. After the amplification step, multiple clonal copies of the single DNA template will be obtained [31]. Beads containing DNA are selected from those without DNA during enrichment step which helps to maximize the sequencing yield [24].

The enriched, template-positive ISPs are loaded into a chip and sequenced. The number of reads performed by the instrument will depend on the size of chip selected, ranging from $4x10^5$ to 5.5×10^6 reads. The run time also varies from 3 – 7 hours depending on the chip selected [13]. These characteristics have indicated that the Ion PGM™ System is an affordable, fast and high-quality sequencing platform, which involves simple sample preparation and data analysis.

In addition, the Next Generation Sequencing platforms are able to detect mitochondrial heteroplasmy due to their higher accuracy and sensitivity. This feature makes NGS technology suitable for use in forensic applications, especially where the detection heteroplasmy in a sample plays an important role [10].

Although Next Generation Sequencing technology appears to have an important future role in forensic studies, to date only few publications are available that describe the application of NGS technology to mitochondrial DNA testing in this context [14]. Sequencing of hypervariable regions HV1 and HV2 from control region of the mtDNA is an important tool for forensic identity testing [6].

This study will evaluate the performance of the Next Generation Sequence platform Ion PGM™ System in sequencing mitochondrial DNA hypervariable regions, HV1 and HV2 using reference samples from individuals from Guatemala. Previous efforts regarding the Guatemalan population genetic information have relied heavily on Short Tandem Repeats (STR) typing. There are no studies, or at least not published studies, describing genetic variations in the mtDNA from the Guatemalan population; and even less using high throughput sequencing technologies.

B. Hypothesis

Sequencing of both hypervariable regions (HV1 and HV2) of the mitochondrial DNA control region using Next Generation Sequencing Ion PGM™ System allows a higher sample throughput in a more accurate, sensitive, cost and time effective manner than Sanger-type sequencing method.

C. Specific Aims

1. To compare the reliability, sensitivity and accuracy of Next Generation Sequencing of the Ion PGM™ System platform against Sanger- type sequencing methods to sequence regions HV1 and HV2 of mtDNA.
2. To analyze the mtDNA sequences of HV1 and HV2 from close maternal relatives (mother-child, grandmother-child and sibling-sibling relationships) to determine if intergenerational differences are present.

CHAPTER II

MATERIALS AND METHODS

A. Samples:

A total of 40 blood reference samples were selected from questioned paternity cases where close maternal relatives were present. Blood spots in different matrices such as paper (regular or filter paper), Whatman® FTA® card (Sigma-Aldrich, St. Louis, MO) and Bode Buccal DNA Collector™ (Bode Technology, Lorton, VA) were used. These samples include 2 grandmother-child pairs; 16 mother-child pairs; and 1 case of one mother-three children. This study has been approved by the Institutional Review Board (IRB) at UNTHSC (Protocol #2012-170: Increasing the Efficiency of Mitochondrial DNA Processing of Reference Samples). The list of the samples is presented in Table 1.

Sanger-type sequencing method was performed using only one member of each family group in order to establish the haplotype for each family group. By contrast, NGS sequencing was carried out to obtain the haplotype for each sample in order to confirm the haplotype for each family group and detect any possible sequence polymorphism between relatives from each family group.

Along with the samples, a positive control sample was run. This was also a blood spot on a Whatman® FTA® card from a reference sample with a known haplotype.

Table 1 – Sample ID, relationship, matrix and family group for each sample.

No.	Family Group	Sample ID	Relationship	Matrix
1	1	12-0523.4a	Child	Paper
2		12-0523.4b	Mother	Paper
3	2	12-0523.7a	Child	Paper
4		12-0523.7b	Mother	Paper
5	3	12-0523.10a	Child	Paper
6		12-0523.10b	Grandmother	Paper
7	4	12-0523.12a	Child	Bode Buccal DNA Collector™
8		12-0523.12b	Mother	Bode Buccal DNA Collector™
9	5	12-0523.13a	Child	Paper
10		12-0523.13b	Mother	Paper
11	6	12-0523.14a	Child	Whatman® FTA® card
12		12-0523.14b	Mother	Whatman® FTA® card
13	7	12-0523.15a	Child	Paper
14		12-0523.15b	Child	Paper
15		12-0523.15c	Child	Paper
16		12-0523.15d	Mother	Paper
17	8	12-0523.16a	Child	Paper
18		12-0523.16b	Mother	Paper
19	9	12-0523.18a	Child	Paper
20		12-0523.18b	Grandmother	Paper
21	10	12-0523.20a	Child	Paper
22		12-0523.20b	Mother	Paper
23	11	12-0523.21a	Child	Paper
24		12-0523.21b	Mother	Paper
25	12	12-0523.22a	Child	Paper
26		12-0523.22b	Mother	Paper
27	13	12-0523.23a	Child	Paper
28		12-0523.23b	Mother	Paper
29	14	12-0523.29a	Child	Paper
30		12-0523.29b	Mother	Paper
31	15	12-0523.31a	Child	Paper
32		12-0523.31b	Mother	Paper
33	16	12-0523.34a	Child	Paper
34		12-0523.34b	Mother	Paper
35	17	12-0523.36a	Child	Paper
36		12-0523.36b	Mother	Paper
37	18	12-0523.43a	Child	Paper
38		12-0523.43b	Mother	Paper
39	19	12-0523.44a	Child	Paper
40		12-0523.44b	Mother	Paper

B. Direct PCR amplification

Samples were punched using the BSD600 DUET Punch System (Life Technologies, Foster City, CA), including the BSD600 software (Life Technologies, Foster City, CA). Each 1.2 mm punch was incubated in a tube containing the MitoReady Incubation Buffer (UNTHSC, Fort Worth, TX) for 40 minutes at 70 °C. Then, MitoReady Amplification Master Mix (UNTHSC, Fort Worth, TX) was added for amplification of samples using Applied Biosystems® GeneAmp® PCR System 9700 (Life Technologies, Foster City, CA) thermocycler. The thermal cycling parameters were: 95 °C for 11 minutes; 32 cycles of 95 °C for 10 seconds, 60 °C for 45 seconds and 72 °C for 1 minute; and 15 °C for 10 minutes.

For the Sanger sequencing method amplification of the whole control region was performed using A1 (forward) and D1 (reverse) primers, initially, generating a DNA amplicon of 1020 bp. Using this strategy most samples failed to amplify, so it was decided to amplify HV1 and HV2 regions separately. Two different sets of primers were used: A1 (forward) and B1 (reverse) for the amplification of HV1 region; and C1 (forward) and D1 (reverse) for the amplification of HV2 region. This same strategy was used to amplify samples sequenced using NGS platform. Table 2 shows the sequence, annealing sites, and amplicon size for each of the primers used.

Table 2 – Primers binding sites, sequence and amplicon size for each mtDNA region

Region	Primer Name	Direction	Sequence (5' to 3')	Start Point	End Point	Amplicon Size
HV1	A1	Forward	CACCATTAGCACCCAAAGCT	15978 bp	15997 bp	432 bp
	B1	Reverse	GAGGATGGTGGTCAAGGGAC	16391 bp	16410 bp	
HV2	C1	Forward	CTCACGGGAGCTCTCCATGC	30 bp	48 bp	399 bp
	D1	Reverse	CTGTTAAAAGTGCATACCGCCA	408 bp	429 bp	

C. Sequencing by Sanger-type Sequencing Method

1. Post-PCR mtDNA clean-up, cycle sequence and post-cycle Post-Amplification Yield Gel: In order to verify the presence of PCR mtDNA products a 2% DNA Typing Grade® Agarose (Life Technologies, Foster City, CA) yield gel was run. PCR products were separated by applying 120 V for 20 minutes; and further visualized using Ethidium Bromide staining and UV light. The Procedure Manual "Post-Amplification Yield Gel" was followed [33].

2. Post-amplification clean-up, cycle sequencing and post-sequencing clean-up: The Procedure Manual "Post-PCR mtDNA Processing" was followed [34]. Amplification products were purified by adding 8 µL of ExoSAP-IT® (Affymetrix, Santa Clara, CA) reagents and incubated at 37 °C for 15 minutes and 80 °C for 15 minutes in an Applied Biosystems® GeneAmp® PCR System 9700 equipment.

After post-PCR clean up, samples were cycle sequenced using BigDye® Terminator v1.1 (Life Technologies, Foster City, CA) chemistry. Four different sequencing reactions were prepared: one for each primer (A1, B1, C1 and D1) separately (Figure 5). Each reaction was prepared by adding 5.0 µL of BetterBuffer BigDye™ dilution buffer (Gel Company Inc., San Francisco, CA); 1.0 µL BigDye® Terminator v1.1 (Life Technologies, Foster City, CA); 1.5 µL of Primer at 3.3 µM (Invitrogen, Waltham, MA); 1.0 – 3.0 µL amplification product; and the required volume of water (Invitrogen, Waltham, MA) to reach a final reaction volume of 15 µL. Samples were incubated for 3 minutes at 96°C; 25 cycles of 15 seconds for 96 °C; 10 seconds at 50 °C; and 3 minutes at 60 °C using the Applied Biosystems® GeneAmp® PCR System 9700 thermocycler.

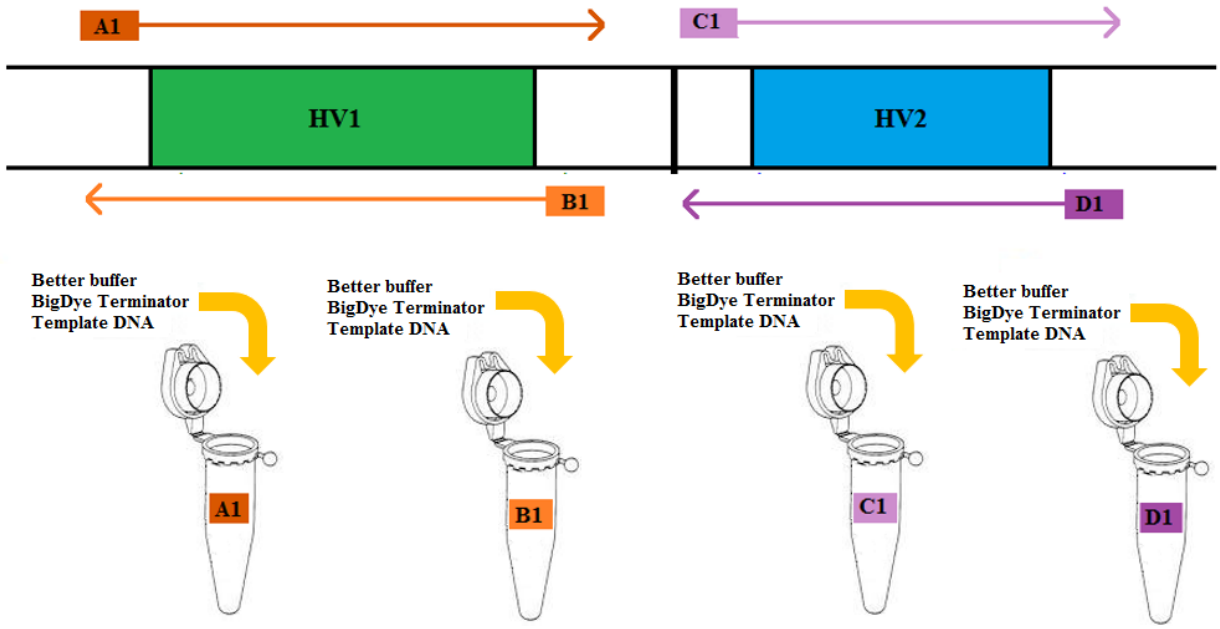

Figure 5 – Cycle sequencing reactions. Each complementary strand from each region must be cycle sequenced in separate reactions, one for each primer used. In this case, four different reactions were prepared.

A post-cycle sequencing clean up step was carried out with BigDye® XTerminator™ (Life Technologies , Foster City, CA) reagents. A volume of 22.5 µL SAM™ solution; 5.0 µL BigDye™ XTerminator solution; and 27.5 µL of water were added to each sample.

3. Capillary electrophoresis and data analysis: Detection was done by capillary electrophoresis using 3130*xl* Genetic Analyzer (Life Technologies, Foster City, CA) following Procedure Manual "Maintenance and Use of the 3130*xl* Genetic Analyzer" [35]. Analysis of data was carried out using MTexpert™ (Mitotech™, Sante Fe, NM) software. Sequencing data was aligned with the Revised Cambridge Reference Sequence (rCRS).

D. Sequencing using Ion PGM™ System

1. Quantification and Normalization of PCR Amplicons: DNA quantification was carried out using Qubit® dsDNA BR Assay Kit (Life Technologies, Foster City, CA) on a Qubit® 2.0 Fluorometer (Life Technologies, Foster City, CA) following manufacturer's instructions [36]. Samples were normalized into equal molar concentration for both regions.

2. Post-PCR Purification and Quantification: PCR amplicons were cleaned-up using Agencourt AMPure XP - PCR Purification Kit (Beckman Coulter, Beverly, MA) following manufacturer's instructions for a PCR reaction volume of 10 µL [37]. DNA quantification was carried out using Qubit® dsDNA BR Assay Kit following manufacturer's instructions [36].

3. Library preparation: NEBNext® Fast DNA Library Prep Set for Ion Torrent™ (New England BioLabs, Ipswich, MA) kit was used to construct the library. All steps described in manufacturer's protocol were followed, with the exception of size selection [25]. Several modifications were made to the protocol suggested by the manufacturer. For the "End Repair of DNA Protocol" the reaction was carried out using half of the volume of each reagent and adding 20 µL of DNA, to reach a final volume of 30 µL. The "Preparation of Adaptor Ligated DNA" step was carried out adding half of the volume of each reagent to each sample, to reach a total reaction volume of 50 µL. NEXTflex™ DNA Barcodes (BiooScientific, Austin, TX) were used instead of the ones suggested by the protocol. Barcode numbers from 23-63 were used to build the library for 41 samples total. For the "Cleanup of Adaptor Ligated DNA" step, half the volume of each reagent was used. Beads

were resuspended using the volume recommended by the manufacturer, a total of 20 µL was recovered. Downstream steps were carried out without any modification to the protocol recommended by the manufacturer.

4. Amplified Library Concentration verification and Library normalization: This step was carried out using Agilent High Sensitivity DNA Kit (Agilent Technologies, Santa Clara, CA) following manufacturer's instructions [38]. Samples were diluted (1 to 5 dilution) and measurements were made using an Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA). Based on the concentration indicated by the instrument, samples were normalized in order to reach a final concentration of about 26 pM. Equal volumes (3 µL) of the 41 libraries were combined for the next step.

5. Emulsion PCR and Enrichment of template-positive Ion Sphere™ Particles (ISPs): Emulsion PCR was conducted by using the Ion PGM™ Template OT2 400 Kit (Life Technologies, Foster City, CA) along with the Ion OneTouch™ 2 System (Life Technologies, Foster City, CA), following the recommended protocol [39]. Template-positive ISPs were enriched on the Ion OneTouch™ ES (Life Technologies, Foster City, CA) instruments. Protocol suggested by manufacturer was followed [39].

6. Sequencing and Data Analysis: DNA sequencing was performed using the Ion PGM™ Sequencing 400 Kit (Life Technologies, Foster City, CA) and the Ion PGM™ System (Life Technologies, Foster City, CA) instrument. Manufacturer's protocol was followed [40]. An Ion 316™ Chip (Life Technologies, Foster City, CA) was used for sequencing. NextGENe® Software (Softgenetics, State College, PA) was used for the analysis of the sequencing data. Sequencing data were aligned with the rCRS.

CHAPTER III

RESULTS AND DISCUSSION

A. Direct amplification performance to sequence blood type samples from the Guatemalan population:

Originally it was decided to amplify, by a direct amplification strategy, the mtDNA control region to be sequenced by Sanger-type sequencing (STS) method. By using primers A1 and D1, a PCR product of 1020 bp long was expected to be amplified. Verification of PCR products were done by visualization using an agarose gel. Because of the extended amplicon length, it was decided to performed a targeted direct amplification of HV1 and HV2 regions for all samples suitable to be sequenced using Ion PGM™ System. Direct amplification of each region will generate small DNA fragments which could be sequenced without the addition of fragmentation and size selection steps during library preparation. The amount of each PCR product for all samples was quantified in order to verify the success of the amplification step.

From all 19 samples selected to be sequenced by STS method, only 4 showed to have a bright signal (similar to the one shown by the positive control sample); 11 showed a less bright band; and 5 showed no band at all. As observed in Figure 6 all samples presented a smear, which was not observed in either the positive or negative controls run along with the samples. One possible reason that might have caused this phenomenon is the presence of degraded DNA in the samples. Nevertheless, samples with positive amplification showed a very clean and distinguishable band.

By using the proposed amplification strategy, the amplification success rate based on the yield gel results was 75%.

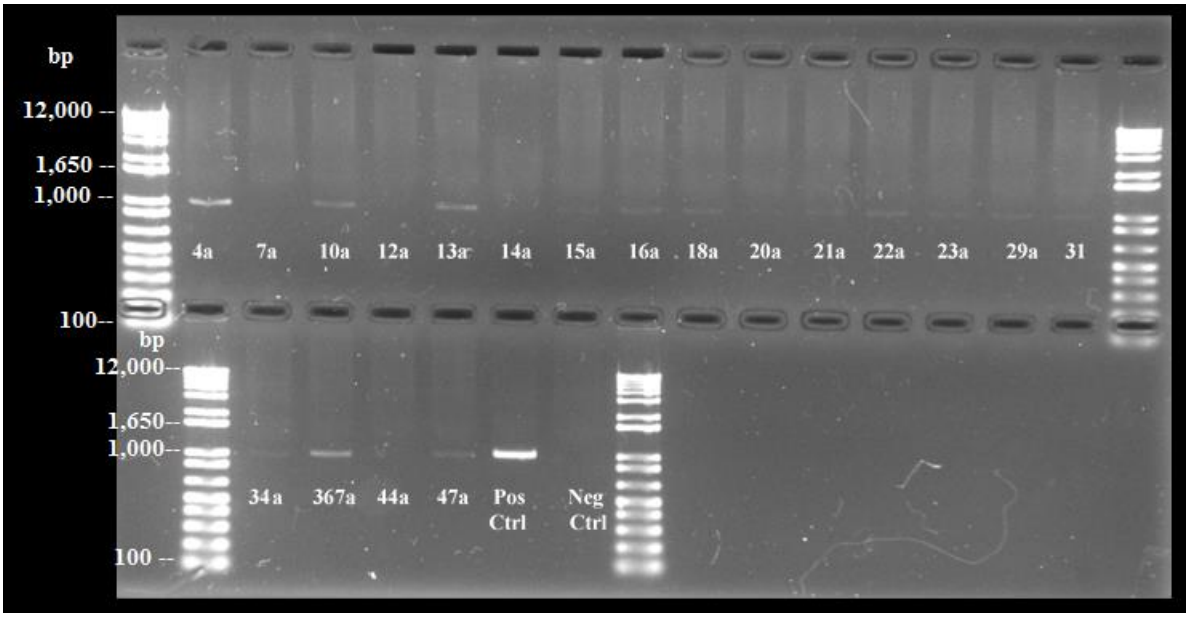

Figure 6 – Yield Gel to verify control region amplification (1020 bp long). A total of 19 samples, 1 positive control and 1 negative control were also amplified. Sample names were abbreviated for presentation purposes in this figure. Different band intensity can be observed between samples: 4 of them showed to have the same intensity as the positive control; 11 had a less bright band; and for 5 of them no band was observed.

Next, cycle sequencing was performed using A1, B1, C1 and D1 primers for these 19 samples. Complete sequencing data (for all primers used) was obtained for only 4 samples; partial data (for any of the primers used) was obtained for 5 samples; and for the other 10 samples no data were obtained. Most of the sequencing signal generated by this step was very low and it was not detected by the software. This first sequencing run had a 21% success rate. So even though the amplification yield showed to be high, the rate of obtaining good quality sequencing data proved to be low. All samples with partial or no sequencing data were processed a second time, yielding lesser success rate (14%). Both times, good quality data were obtained from the sequencing of

the positive control sample, proving there was not a problem with the method, instrumentation or reagents used.

Based on these results, it was decided to carry out targeted amplification of each region as it was originally proposed for the NGS phase. Primers A1 and B1 were used to amplify HV1 region and primers C1 and D1 were used to amplify HV2 region, generating fragments of 432 bp and 399 bp, respectively. By applying this strategy, possible DNA degradation was overcome. Targeted amplification was used for the remaining 13 samples, with partial or no sequencing data (Figure 7). Band for either PCR amplicons was observed for 12 samples and just one failed to show the presence of a band for both of them. This sample corresponds to family group 6, sample ID 12-0523-14a. In order to obtain the haplotype for this family group, mtDNA direct amplification was performed using sample 12-0523-14b; for which amplification of both regions was achieved (data not shown). Both samples of this family group were the only blood samples fixed in a Whatman® FTA® card.

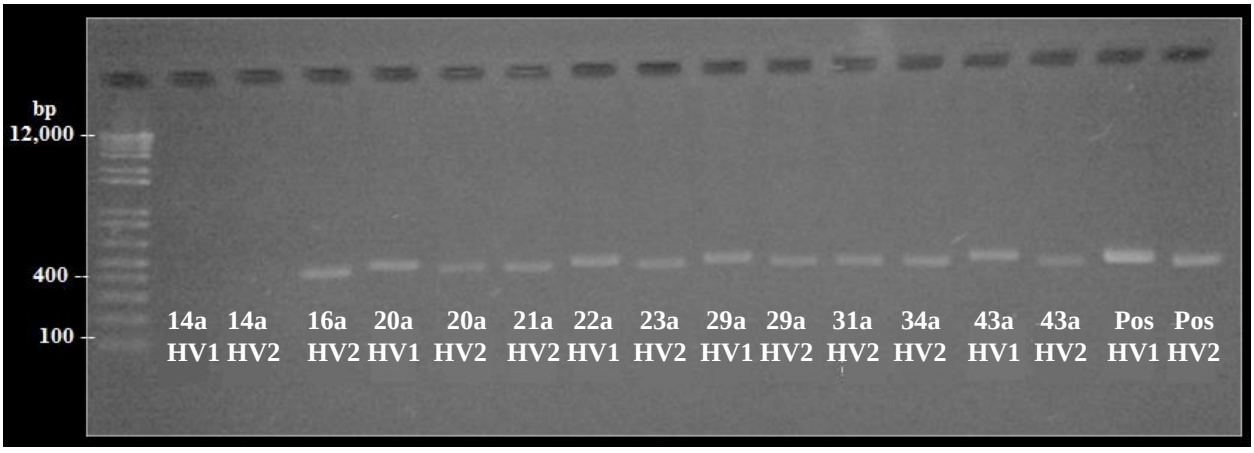

Figure 7 – Yield gel ran to verify the presence of HV1 (432 bp) and HV2 (399 bp) amplicons. This figure shows the electrophoretic results of 10 samples (data not shown for the other 3 samples) and the positive control; HV1 was amplified for 5 samples and HV2 was amplified for 9 of them. Samples ID were abbreviated for presentation purposes. As observed in this figure, no PCR amplification product was detected for sample 12-0523-14a (14a).

Separately, a gel was run to compare band intensity and quality from PCR products obtained by these two different amplification strategies. The HV1 and HV2 amplicons from two different samples were run along with the control region amplicon of 3 other samples. The intensity of the bands for the first set was higher than the one for the later ones (Appendix Figure 1), demonstrating a better quality and higher quantity of PCR products is obtained by using a targeted direct amplification for both hypervariable regions independently.

By conducting independent amplification of each hypervariable region, good quality data were obtained for 11 of those amplicons after cycle sequencing. The success rate for obtaining a good sequencing data was about 65%. Cycle sequencing steps were repeated for the other 9 amplicons; for 6 of them the signal was low, so a higher volume of PCR product was added to the cycle sequencing reaction. The other 3 amplicons generated a very high signal, so they were diluted before being sequenced. Good quality data were obtained for all of them on this second processing.

These results suggest that a direct amplification of HV1 and HV2 regions in separate reactions is more suitable to obtain good quality data using Sanger-type sequencing method for this type of samples, especially if degradation is suspected. Samples used in this project were collected in 2012 and most of them were fixed in regular filter paper which does not ensure the preservation the genetic material present in the sample. If degraded DNA is present the amplification of larger amplicons is more difficult in comparison to shorter ones. This could explain why the quality and intensity of the bands observed for HV1 and HV2 fragments were higher than those for the control region fragment.

As mentioned before, HV1 and HV2 regions were amplified independently for each sample to be sequenced by NGS method. After direct amplification was done, quantification of PCR products by Qubit® dsDNA BR Assay Kit was performed; The results are shown in Table 3.

Table 3 – PCR product quantification using Qubit® 2.0 Fluorometer after direct amplification of HV1 and HV2 regions to be sequenced by Ion PGM™ platform.

Family Group	Sample ID	DNA Concentration (pg/µL)	
		HV1	HV2
1	12-0523.4a	19,980	20,600
	12-0523.4b	22,400	18,900
2	12-0523.7a	25,800	18,000
	12-0523.7b	20,000	17,100
3	12-0523.10a	21,800	17,760
	12-0523.10b	18,020	15,420
4	12-0523.12a	27,400	19,540
	12-0523.12b	17,880	13,740
5	12-0523.13a	24,800	22,000
	12-0523.13b	24,000	18,100
6	12-0523.14a	6,840	12,580
	12-0523.14b	7,980	7,600
7	12-0523.15a	23,200	18,260
	12-0523.15b	22,600	17,140
	12-0523.15c	18,660	15,000
	12-0523.15d	23,000	16,360
8	12-0523.16a	25,200	17,940
	12-0523.16b	22,000	15,100
9	12-0523.18a	19,640	16,300
	12-0523.18b	19,820	15,400
10	12-0523.20a	14,760	13,620
	12-0523.20b	18,580	14,620
11	12-0523.21a	23,800	16,880
	12-0523.21b	25,400	19,120
12	12-0523.22a	22,800	19,540
	12-0523.22b	19,160	15,840
13	12-0523.23a	23,800	15,900
	12-0523.23b	23,200	15,660
14	12-0523.29a	23,200	18,180
	12-0523.29b	23,200	17,100
15	12-0523.31a	18,840	17,840
	12-0523.31b	21,000	18,060
16	12-0523.34a	26,800	16,760
	12-0523.34b	20,400	16,220
17	12-0523.36a	24,200	17,760
	12-0523.36b	22,200	17,140
18	12-0523.43a	16,420	17,440
	12-0523.43b	20,800	13,780
19	12-0523.44a	23,600	21,200
	12-0523.44b	22,400	22,400
	Positive Ctrl	29,400	15,900

DNA concentration for the HV1 fragments obtained from samples ranged from 6,840 pg/µL to 27,400 pg/µL; 68% of the samples had a concentration higher than 20,000 pg/µL. The concentration of the HV2 amplicons ranged from 7,600 pg/µL and 22,400 pg/µL; but only 10% of the samples had a concentration higher than 20,000 pg/µL. These results showed that amplification of HV2 generates less amount of copies of DNA than the amplification of HV1. In both cases, samples from family group 6 (12-0523-14a and 12-0523-14b) had the lower PCR product concentration values. These quantification results are concordant with the ones observed for these same samples during agarose gel electrophoresis, where no band was observed for one of these samples on multiple occasions. For these two samples, direct amplification of regions HV1 and HV2 was not as successful as for the rest of samples. These two samples were the only ones collected on a Whatman® FTA® card, so it is possible that this matrix affected the integrity of the DNA present in the blood sample.

For all 40 samples sequenced using Ion PGM™ Systems technologies, good quality data were obtained for 37 (93%) of them. Partial data or low bad quality data were obtained from two samples and no data were obtained for only one sample. Samples with partial or no data were not taken in consideration for any comparative determinations or conclusions reported in this study.

B. Performance comparison between Sanger-type Sequencing method and Ion PGM™ System method

Several performance characteristics were evaluate for both methods, including efficiency, costs and time (Table 4). As mentioned in the previous section, for the STS method, several samples needed to be processed more than one time in order to generate good quality data. Most samples were repeated because the signal generated was not high enough to be detected by the

software. Less frequently, a higher signal was obtained from which the software was not able to generate good quality data. In both instances the cycle sequencing step was repeated by either increasing or reducing the quantity volume of DNA added to each reaction. On average, cycle sequencing and downstream steps were done twice for about 45% of the samples. Because there is no step to verify the quantity of DNA being sequenced, the same yield percentage is assumed every time. By comparison, using the Ion PGM™ System platform only one sample (2.5%) failed to generate data in the first attempt and partial data were generated for two samples (5%). Repetition of processing steps increases costs and time to obtain sequencing data from samples when using the STS method carried out in this study.

Table 4 – Performance comparison of both methods in terms of yield, cost, time and efficiency.

	Sanger Sequencing	Ion PGM™ System
Number of samples processed	19	40
Efficiency per processing [a]	55%	93%
Cost per sample	$ 56.76	$ 68.23
Processing time consumed [b]	52 hours	38 hours
Data analysis time consumed	20 hours	14 hours
Number of reads	2	4810 [c]

[a] Percentage of samples with good quality per processing.
[b] These estimates include manpower time as well as incubation times and instrumentation run times.
[c] Average value calculated from the number of reads for all samples (n=40).

Even though the cost of reagents used by the Sanger sequencing method are lower than those used by the Ion PGM™ System method, its low yield into generating good quality data per run increases its cost per sample. Based on the calculations made, the cost per sample for the Ion

PGM™ System method is higher than the STS method by a total of $11.47. This calculation only includes the cost of reagents needed to perform either method.

In addition, time calculations for processing and data analysis were made for each method. The process used for the STS method showed was found to take longer than the one used by the Ion PGM™ System method. These values represent the time consumed to generate good quality data based on the number of samples processed by each method. Therefore, 19 samples were processed in 52 hours by the STS method. By comparison, twice the number of samples were processed with 14 hours less time by using the Ion PGM™ System method, making it a more suitable method for larger number of samples. This also shows that the STS method is more time-consuming, which would also increase its cost per reaction considering that more manpower time is required.

Additionally, the analysis of the data generated by the Ion PGM™ System technology takes 6 hours less than the analysis of the data generated by the STS method, even with double the amount of haplotypes. This difference relies in the coverage inherent to each method. The number of reads performed for each region by the STS method were 2 reads; while the average number of reads performed by the Ion PGM™ System method was 4,810 reads. By increasing the number of reads for each nucleotide the analysis of the data is easier and more reliable. By contrast, the analysis of the data generated by STS method by using MTexpert™ software requires manual evaluation and input from the analyst.

From the 20 samples sequenced by both methods, a total of 221 polymorphisms were detected by the STS method. Meanwhile, from the same samples, 227 variants were detected by the Ion PGM™ System method (Table 5). The variants of the haplotypes obtained from both methods

were defined in relation to the Revised Cambridge Reference Sample (rCRS). A total of six polymorphisms were detected only with the PGM platform.

Table 5 - Sequence polymorphisms within the variable regions of mtDNA detected by Sanger-type sequencing method and PGM platform.

	Sanger Sequencing	Ion PGM™ System
Total number of polymorphisms	**221**	**227**
Transitions	**173 (0)**	**177 (0)**
T > C	44 (0)	45 (0)
C > T	52 (0)	55 (0)
A > G	65 (0)	65 (0)
G > A	12 (0)	12 (0)
Transversions	**3 (0)**	**3 (0)**
A > C	3 (0)	3 (0)
Insertions	**33 (1)**	**14 (20)**
A	1 (0)	1 (0)
C	31 (1)	12 (20)
T	1 (0)	1 (0)
Deletions	**9 (0)**	**7 (2)**
A	7 (0)	5 (2)
C	2 (0)	2 (0)
Point Heteroplasmy	**2 (0)**	**3 (1)**
T>CT	1 (0)	1 (0)
C>CT	1 (0)	2 (1)

Numbers inside parenthesis indicate calls not detected by the software and were done manually by the analyst.

.

Both pyrimidine and purine transitions were observed 173 variants calls were reported by Sanger sequencing method and 177 by PGM technology. In all cases these transition events were not detected by STS method because there was no coverage at the position where the substitutions occurred for those specific samples. Three of these substitutions occurred at position 64 and one of them at position 16, 362 (Appendix Table 1). Both of which lay at the lower an upper binding sites of the primers used; for this reason no coverage of those positions was achieved for those samples using STS method. Both methods reached to detect the same number of transversions events, for a total of 3 substitutions.

The number of insertions detected by both methods was 34; for data obtained by STS method one insertions was called manually because the software failed to detect it. This insertion corresponded to a cytosine (C) at position 315.1, where a homopolymeric cytosine (poly-C) tract lies. It has been reported that the presence of these poly-C regions from position 303 to 315 in the mtDNA genome makes it difficult to be sequenced [41]. This could be the reason why the software was not able to detect the insertion of this nucleotide at this specific position for this sample. In comparison, the insertion of the same nucleotide at position 315.1 failed to be detected by the software used with the PGM data for 7 samples. Additionally, insertion of this same nucleotide at position 309.1 also failed to be detected by the software used with the PGM data for 13 samples (Appendix Table 1). Even though, the insertion of this nucleotide was not called by the software, the data were visually evaluated and the insertion was manually reported by the analyst (Appendix Figure 2). Despite the fact that sequencing software are tools that provide a lot of help during data analysis, manual evaluation is still required, especially in homopolymeric regions of the DNA.

Deletion of two nucleotides (adenine and cytosine) were detected for three different samples, for a total of 9 occurrences. All deletions were detected by the software for the STS method. For

the Ion PGM™ System method two of them were made manually. All of them corresponded to the deletion of an adenine at position 291 for two different samples (Appendix Table 1). Even though, the deletion of this nucleotide was not called by the software, the phenomenon was clear once data were visually evaluated (Appendix Figure 3). In this specific case, the deletion of this nucleotide happened in adjacent positions, which could be the reason why one of them was not detected by NGS technology.

Finally, two point heteroplasmy were detected among all samples, STS called a total of 2 incidences, while the PGM method detected 4. Point heteroplasmy (T >C) at position 16,093 for sample 12-0523-12a was detected by both methods. As can be observed in the Appendix section Figure 4, the signal generated by both nucleotides during Sanger sequencing generated peaks with similar heights and they were both detected and reported by the software. This point heteroplasmy was also detected by the PGM method. In addition, point heteroplasmy at position 16, 257 (C>CT) was reported for three samples. For samples 12-0523-7a and 12-0523-15a it was only detected by the NGS method. For the Sanger sequencing method only one clear signal for a cytosine was detected at this position (Appendix Figure 5). Based on the higher number of reads for this position and the C to T ratio reported by the software using the PGM technology, it was decided to report this point heteroplasmy for these two samples. For sample 12-0523-12a sequenced by STS method, heteroplasmy at position 16,257 was not detected by the software, probably because the signal generated from the cytosine was higher than the one generated from the thymine. Nevertheless, evaluation of the electropherogram generated shows the presence of both signals and CT heteroplasmy was manually called and reported at this position (Appendix Figure 6).

C. Sequence polymorphisms detected for HV1 and HV2 regions for Guatemalan individuals

Sequencing data from the 19 family groups evaluated, for HV1 and HV2 regions by both methods, were aligned with the rCRS and all differences were noted (Appendix Table 2 and Table 3). Sequencing data from these samples from the Guatemalan population presented differences at 55 positions, which were consistent for both sequencing methods. Among these positions, 41 of them presented nucleotide transition; one showed nucleotide transversion; 5 had single nucleotide insertions; 6 presented one nucleotide deletion; and 2 positions showed point heteroplasmy. The differences at these positions represented a total number of 227 sequence polymorphisms. A summary of the different type of polymorphisms are detailed in Table 6.

Table 6 – Sequence polymorphisms within HV1 and HV2 of mtDNA detected by both methods for a total of 19 family groups from the Guatemalan population.

Total number of polymorphisms	**227**
Transitions	**177**
T > C	45
C > T	55
A > G	65
G > A	12
Transversions	**3**
A > C	3
Insertions	**34**
A	1
C	32
T	1
Deletions	**9**
A	7
C	2
Point heteroplasmy	**4**
T>CT	1
C>CT	3

Nucleotide transitions represented 75% of these polymorphisms, where C > T and A > G changes were more common. Among the total polymorphism observed, 9% correspond to insertions, mostly due to cytosine insertions. Deletions represented 11%, of which adenine deletion was the most predominant. Point heteroplasmy, T > CT and C > CT represented 4% of the polymorphisms observed. Finally, the remaining polymorphism corresponded to a transversion, A > C, representing 1% of the polymorphisms detected.

The common polymorphisms observed in all Guatemalans individuals from all family groups tested were: transitions A > G at positions 73 and 263; and cytosine insertion at positions 315.1. Other common variation calls shared among family groups were transitions A > G at positions 153 and 235; transition C > T at positions 64, 16,111, 16,223 and 16,290; transition G > A at position 16,319; transition T > C at positions 146 and 16,362; and cytosine insertion at positions 309.1. Most importantly, less common differences were observed at 26 different positions (Table 7). These sequence polymorphisms, at the positions indicated, were observed only in relatives belonging to the same family and were not observed in any other family group.

Table 7 – Sequence polymorphisms observed exclusively at one family group.

Polymorphism		**Position(s)**						
Transition	C > T	150	16147	16187	16270	16292	16295	
	T > C	279	16094	16311	16342			
	A > G	189	200	272	291	16175	16216	16241
	G > A	16129	16346					
Insertion	A	16300.1						
	C	16362						
	T	308.1						
Deletion	C	308	309					
	A	16183						
Heteroplasmy	T > TC	16093						

D. Sequence polymorphisms detected for HV1 and HV2 among relatives by Ion PGM™ System method

All 40 samples, representing 19 family groups, were sequenced by PGM technology in order to evaluate any possible sequence polymorphism between relatives from each family group. Sequencing data was obtained for 37 samples (18 family groups); partial data was obtained for two samples (12-0523-14a and 12-0523-14b); and no data were obtained for one sample (12-0523-29a).

The haplotype obtained for each maternally related individual was the same for all family groups, with the exception of family group 1 (mother-child). Discrepancy between both samples belonging to this group were observed in two different positions, one at HV1 and one at HV2 region. The first discrepancy was observed at position 64. The rCRS presents a cytosine at this position. Sample 12-0523-4b (mother) sequenced using NGS technology presented point heteroplasmy of cytosine (85%) and thymine (15%). Sequencing data obtained from Sample 12-0523-4a (child), by both sequencing methods, only showed the presence of cytosine at this position. Similarly, at position 16,362, where a thymine is observed in the rCRS, point heteroplasmy between a cytosine (29%) and a thymine (71%) was detected just for sample 12-0523-4b. Sample 12-0523-4a was sequenced by both methods, PGM generated data only showed the presence of a thymine; no data were obtained for this position by STS method. Nevertheless, it is well known that heteroplasmy can even occur within a cell or different tissues from an individual, so it is possible to detect this phenomenon in one individual and not in the other so it was not considered as an intergenerational difference.

CHAPTER IV

CONCLUSIONS

In this study, Ion Torrent PGM™ System platform was used to sequence hypervariable regions from the human mitochondrial genome from samples corresponding to the Guatemalan population. Blood samples from 40 individuals belonging to 19 different family groups were sequenced using this technology. From each family group one sample was selected and sequenced using conventional Sanger-type sequencing protocols.

Direct amplification strategy to generate multiples copies of the mtDNA control region showed to have a low rate of amplification, leading to an inability to generate good quality data to be sequenced by the STS method. Many reasons could cause this problem, including, DNA degradation caused by time elapsed since samples were collected and this study was performed. Additionally, most of the samples (90%) were collected in regular paper which does not prevent DNA degradation by agents such as nucleases. Besides, the amplicon size corresponding to the control region was about 1020 pb, which is a relatively long DNA fragment. DNA degradation would totally affect the amplification of such long fragment. This issue was overcome by direct target amplification of HV1 and HV2 separately, generating fragments of 432 bp and 399 bp long, respectively. The rate of amplification and sequencing success was increased by using this strategy. Based on this, it is recommended to perform direct amplification of HV1 and HV2 regions separately when using blood samples collected and stored under similar conditions as the samples used in this study.

This same strategy was used for those samples sequenced by Ion PGM™ System method in order to reduce steps of the library preparation process, with the objective to reduce time and reagents used. For both methods, sample 12-0523-14a from family group 6 was found to generate the less amount of PCR product. Samples from this family group were the only ones collected in a Whatman® FTA® card; this fact supports the fact that the matrix used to collect the sample affects direct amplification of HV1 and HV2.

A 93% of success was achieved by the Ion PGM™ System method, on the first run, in comparison of a 45% obtained by STS method. In order to generate good quality data for the remaining samples, cycle sequencing steps needed to be repeated. Therefore, STS method showed to have a lower yield which increases the costs and time of the process.

The use of 40 different barcodes allowed the simultaneous sequencing of all sample by NGS strategy; meanwhile all 19 samples selected to be done by STS methods were done one by one. By sample barcoding, PGM technology noticeably out weights STS method in terms of throughput. Being a high-throughput technology, the Ion PGM™ System has a higher degree of coverage compared to that of the STS method. With higher coverage, more reliable and accurate sequencing data was obtained.

Data analysis from samples sequenced using both methods showed that more sequencing polymorphisms were detected using the Ion PGM™ System method. In general, the Ion PGM™ System method was able to detect more point heteroplasmy events. Nevertheless, it was not able to detect nucleotide insertions as well as the STS method. The Ion PGM™ System had some troubles to detect nucleotide insertions especially close to homopolymeric tracts. So, the evaluation of an expert becomes necessary to make the call of possible sequencing polymorphisms in this areas. Two tranversion events were detected by STS only at positions which were not covered by

NGS technology. It is know that transversions are not that common substitutions, so it is recommended to reprocessed this sample in order to confirm the transversions events observed in order to report them as true polymorphisms in this sample.

Samples tested from the Guatemalan population showed to differ in 55 different position in comparison to the rCRS, representing 227 different sequencing polymorphisms. Some were very common between individuals from different family groups. This might indicate that these variations could be shared between individuals from the population tested. In order to confirm this more samples would need to be sequenced in future studies. More importantly, this preliminary study, helped to detect other polymorphisms that were really specific to a family group. The observance of a distinguishable sequence polymorphism in a haplotype makes it unique, which is of great importance for forensic purposes. Relative rarity of mtDNA profiles obtained help in the comparison between a reference and evidence samples, as well as in identification of missing persons.

NGS data from all samples allowed the comparison of HV1 and HV2 sequences between maternally related individuals. For only one family group, differences between child (12-0523-4a) and mother (12-0523-4a) were detected and both corresponded to point heteroplasmic changes. For the remaining family groups, the same haplotype was obtained for all family members; even in both cases where the biological relationship was child-grandmother, showing that HV1 and HV2 regions of the mtDNA has not changed in at least two generations in this family groups. One case with more than one child was also evaluated, and all of them showed to have the same sequence for both regions.

APPENDIX

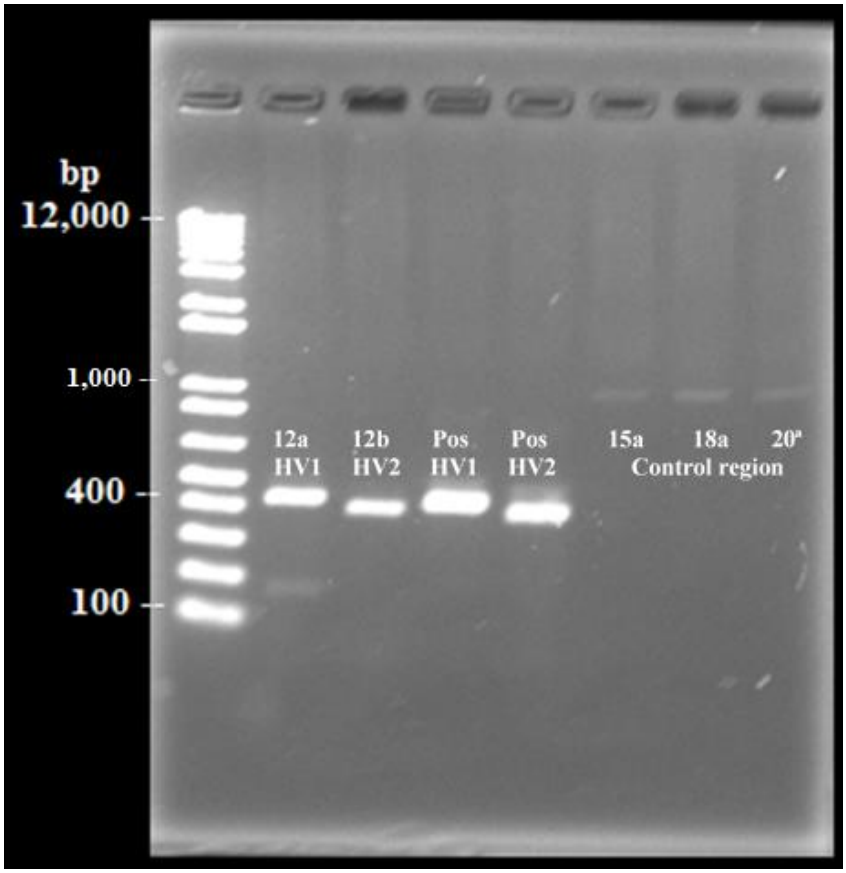

Figure 1 – Comparison of band intensities between targeted direct amplification of HV1 and HV2 (first four bands) and amplification of the control region (last three bands).

Table 1 – Samples for which polymorphism calls were inconsistent between methods

Sequence Polymorphism	Sample ID	Position	Method	Reason
T > C	12-0523-23a	16362	Sanger Sequencing	No coverage of this position.
C > T	12-0523-7a 12-0523-10a 12-0523-21a	64	Sanger Sequencing	No coverage of this position.
InsC	12-0523-10a	315.1	Sanger Sequencing	No coverage of this position.
	12-0523-12a 12-0523-18a 12-0523-21a 12-0523-29b 12-0523-31a 12-0523-36a	309.1	Ion PGM™ System	Not detected by software, calls were made manually.
	12-0523-10a 12-0523-15a 12-0523-16a 12-0523-20a 12-0523-23a 12-0523-34a 12-0523-44a	309.1 & 315.1	Ion PGM™ System	Not detected by software, calls were made manually.
DelA	12-0523-29b 12-0523-36b	291	Ion PGM™ System	Not detected by software, calls were made manually.
C > CT	12-0523-7a	16257	Sanger Sequencing	Clear and high signal for C only.
	12-0523-15a	16257	Sanger Sequencing	Not detected by software, calls were made manually.

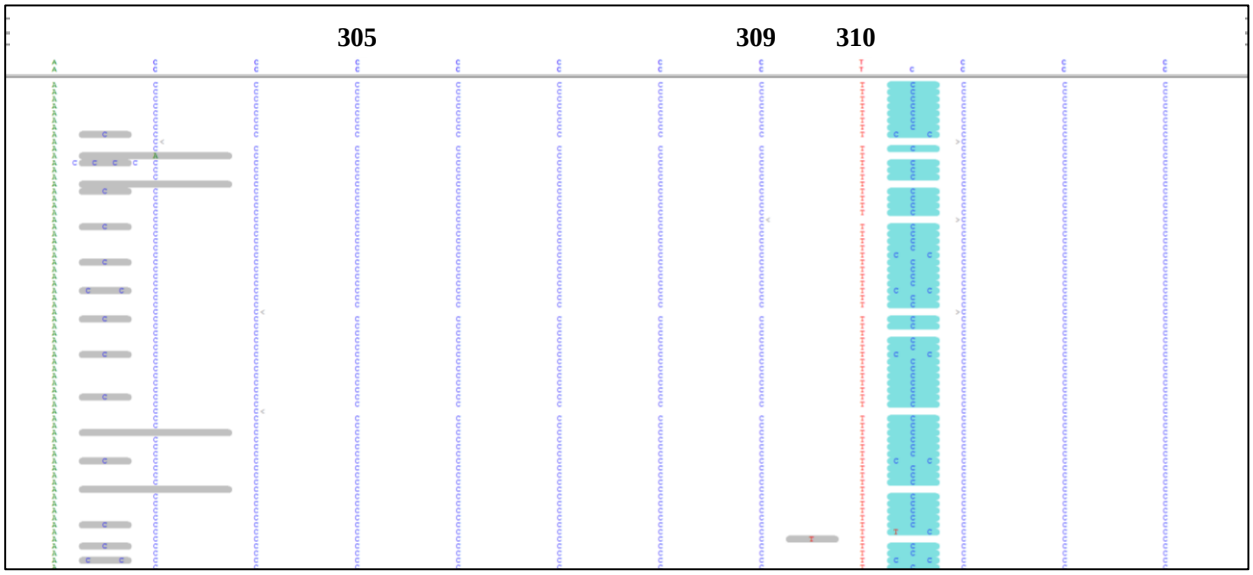

Figure 2 – Sequencing data (partial view) of the insertion of a cytosine at position 309.1 and 315.1 for sample 12-0523.18a. InsC 309.1 was done manually (in gray) and InsC 315.1 was detected and reported by the software (in turquoise). Failure to detect and report InsC at position 309.1 by the software was observed for several samples.

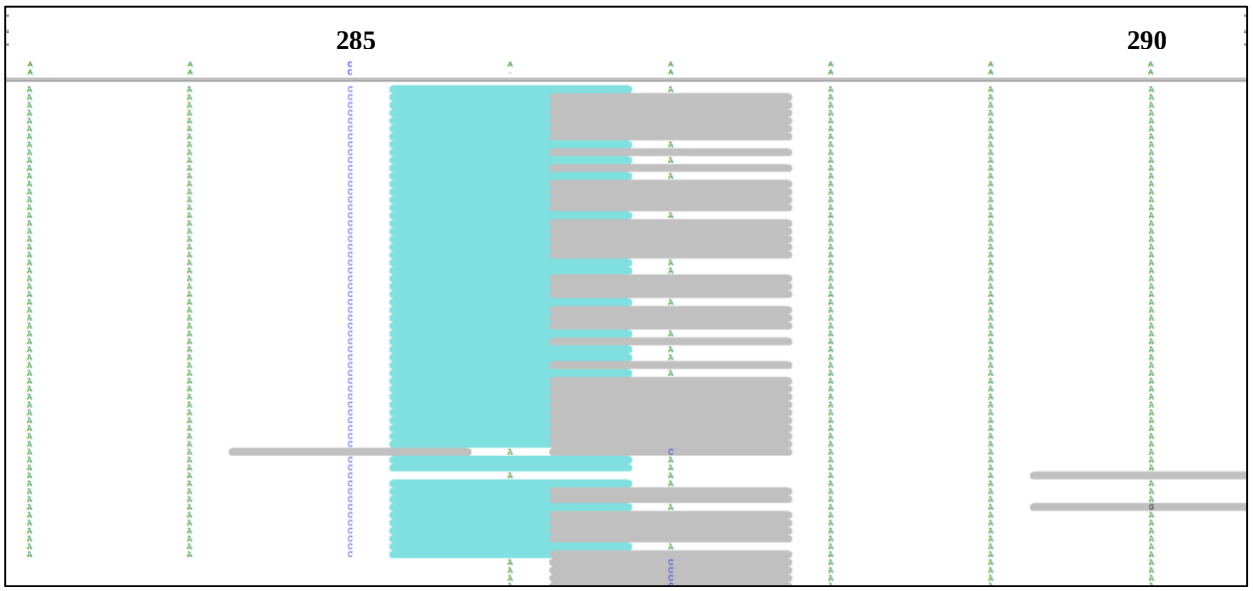

Figure 3 – Sequencing data (partial view) of the deletion of an adenine at position 290 and 291 for sample 12-0523-36a. DelA at position 290 was detected and reported by the software (in turquoise); delA at position 291 was done manually (in gray).

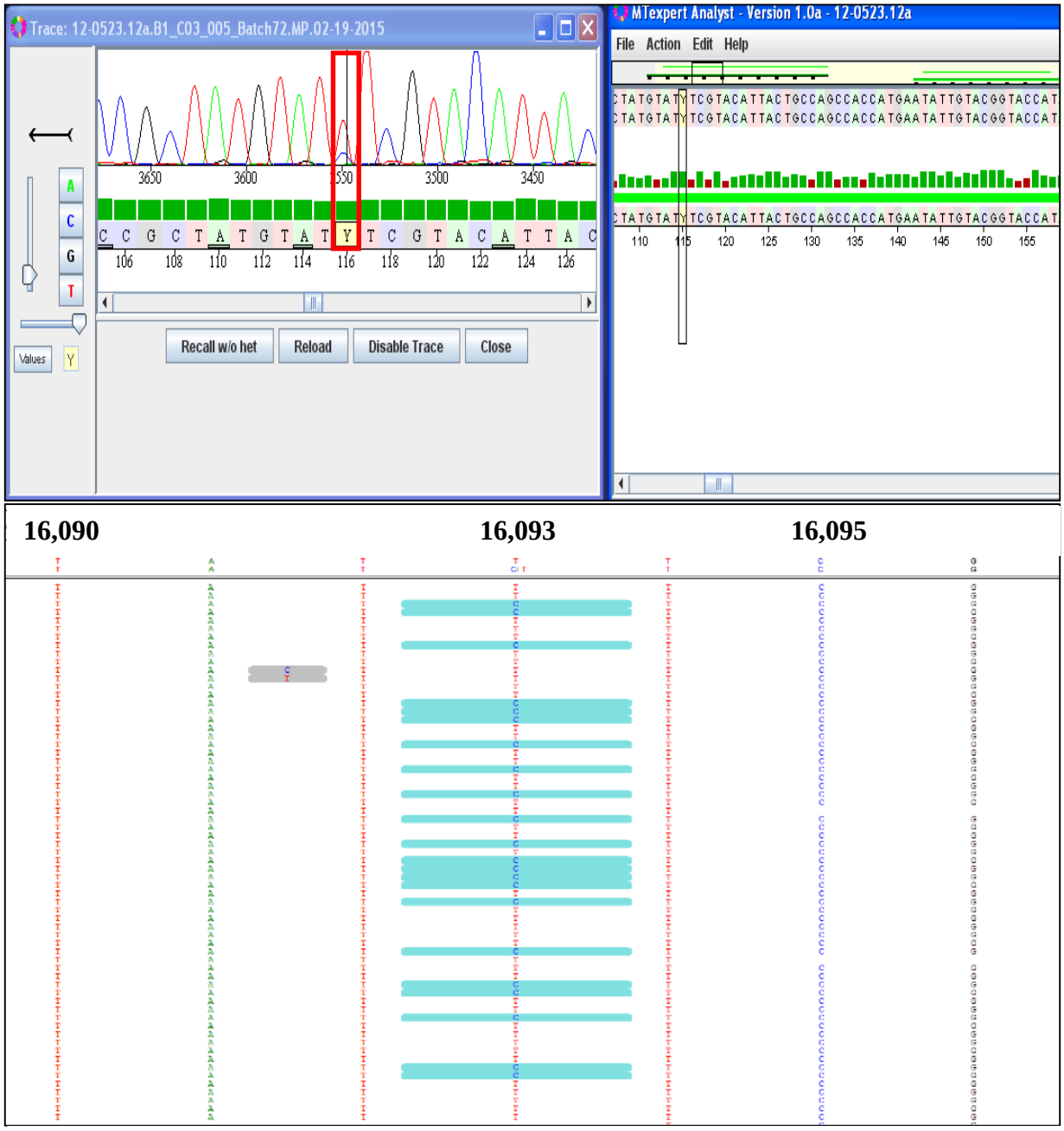

Figure 4 – Sequencing data for point heteroplasmy at position 16,093 for sample 12-0523-12a. Heteroplasmy CT was detected by both methods for this sample. In the top panel Sanger sequencing data is presented. Signal for both nucleotides (red and blue peaks) can be observed and both were detected and reported by the software (red box). In the bottom panel, NGS data are presented. Both nucleotides were detected (in turquoise) and reported by the software.

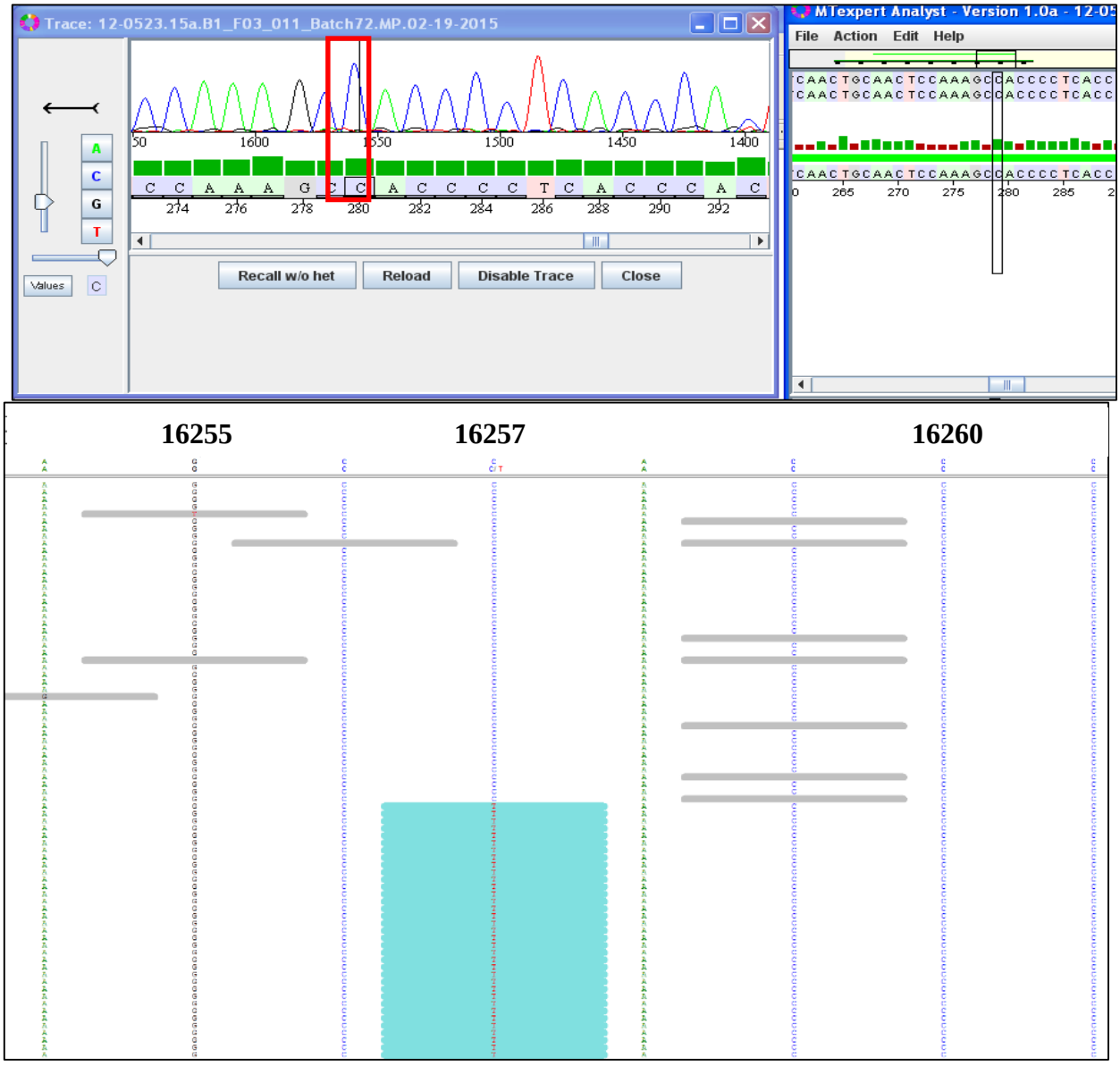

Figure 5 – Sequencing data at position 16, 257 for sample 12-0523-15a. By Sanger sequencing (top panel) a clear signal for only a cytosine (blue peak) was detected and reported by the software (red box). By Ion PGM™ method (bottom panel) sequencing data (partial view) show the presence of two different nucleotides at this position; CT heteroplasmy was detected by NGS technology at this position for this sample (in turquoise).

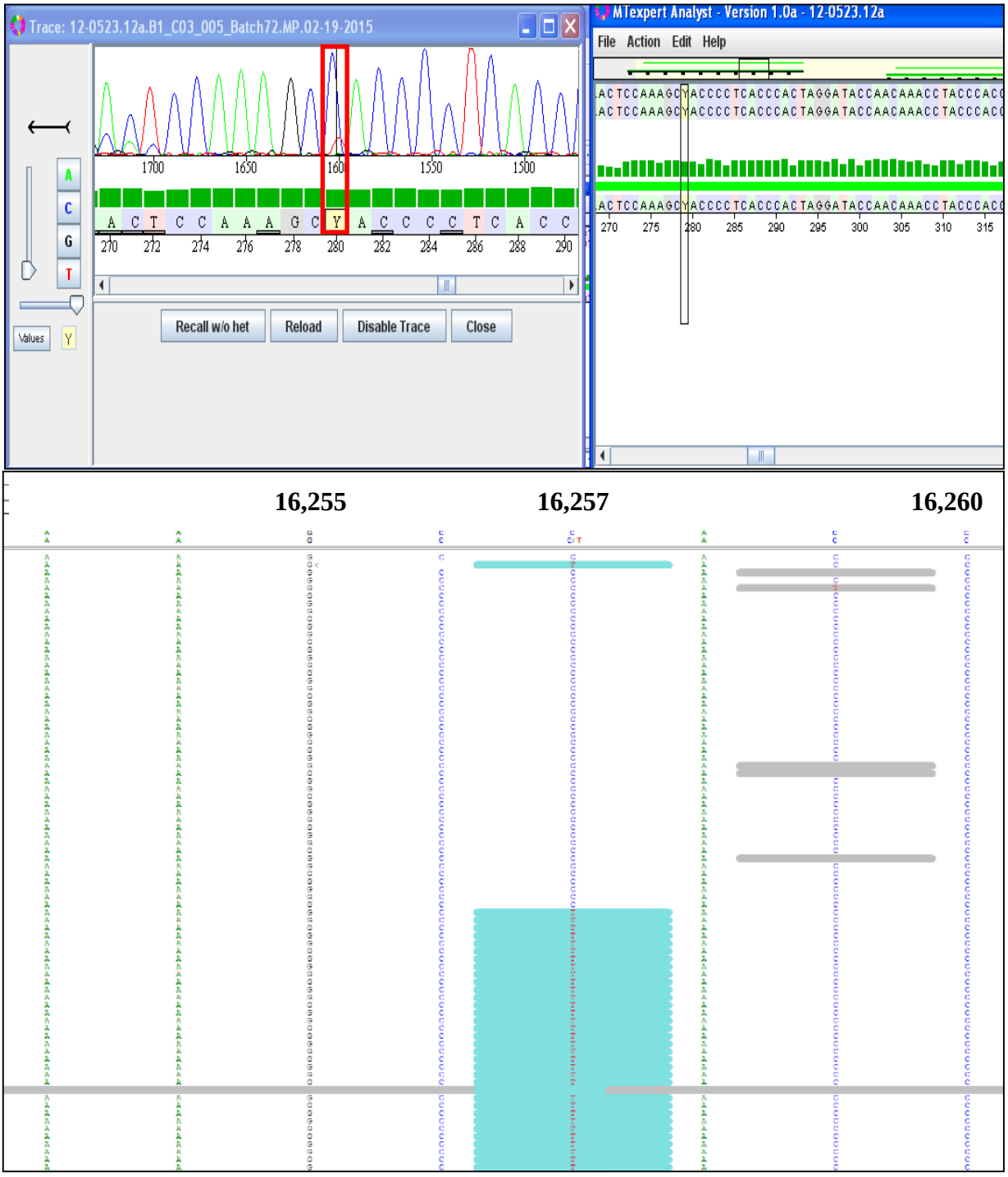

Figure 6 - Sequencing data at position 16,257 for sample 12-0523-12a. Heteroplasmy was not detected by the software used for Sanger Sequencing method, even though signal for a cytosine (blue peak) and a thymine (red peak) were present at that position. CT heteroplasmy call was made manually. The NGS method software was able to detect this heteroplasmy.

Table 2 – Sequence polymorphism in HV2 for each family group from the Guatemalan population detected by both methods

Sample ID	64 C>T	64 C>CT	73 A>G	146 T>C	150 C>T	152 T>C	153 A>G	189 A>G	194 C>T	195 T>C	200 A>G	235 A>G	249 delA	263 A>G	272 A>G	279 T>C	290 delA	291 delA	291 A>G	308 delC	308.1 insT	309 delC	309.1 insC	315.1 insC
12-0523.4a			1			1								1		1								1
12-0523.7a	1		1	1		1	1					1		1										1
12-0523.10a	1		1	1			1					1		1									1	1
12-0523.12a			1			1								1									1	1
12-0523.13a	1		1	1			1					1		1										1
12-0523.14b	1		1	1			1	1		1		1		1									1	1
12-0523.15a	1		1	1			1	1				1		1									1	1
12-0523.16a	1		1	1			1					1		1					1				1	1
12-0523.18a			1							1				1									1	1
12-0523.20a			1	1			1					1		1									1	1
12-0523.21a	1		1	1			1	1				1		1									1	1
12-0523.22a			1	1										1						1		1		1
12-0523.23a	1		1	1	1		1					1		1									1	1
12-0523.29b			1										1	1			1	1					1	1
12-0523.31a			1											1									1	1
12-0523.34a	1		1	1			1					1		1									1	1
12-0523.36a			1										1	1			1	1					1	1
12-0523.43a			1					1			1			1	1									1
12-0523.44a	1		1	1			1					1		1							1		1	1
Number of samples	10	0	19	12	1	3	11	4	0	2	1	11	2	19	1	1	2	2	1	1	1	1	14	19

Table 3 – Sequence polymorphism in HV1 for each family group from the Guatemalan population detected by both methods

Sample ID	16093 T>TC	16094 T>C	16111 C>T	16129 G>A	16147 C>T	16175 A>G	16183 delA	16183 A>C	16187 C>T	16189 T>C	16209 T>C	16216 A>G	16217 T>C	16223 C>T	16241 A>G	16257 C>CT	16270 C>T	16290 C>T	16291 C>T	16292 C>T	16295 C>T	16298 T>C	16300 A>G	16300 insA	16301 C>T	16311 T>C	16319 G>A	16325 T>C	16327 C>T	16342 T>C	16346 G>A	16362 T>C	16362 T>CT	16362 insC
12-0523.4a										1			1																					
12-0523.7a			1								1			1		1		1									1					1		
12-0523.10a			1											1				1									1							
12-0523.12a	1													1	1	1								1	1					1	1	1		
12-0523.13a			1											1				1									1							
12-0523.14b			1											1				1									1							
12-0523.15a			1											1		1		1	1				1				1					1		
12-0523.16a			1							1		1		1				1							1		1					1		1
12-0523.18a							1	1		1			1																					
12-0523.20a														1				1									1					1		
12-0523.21a			1			1								1				1					1				1					1		
12-0523.22a								1		1			1																					
12-0523.23a			1											1				1									1					1		
12-0523.29b														1								1						1	1					
12-0523.31a								1		1			1																					
12-0523.34a		1	1		1				1		1			1			1	1									1					1		
12-0523.36a														1								1						1	1					
12-0523.43a				1						1	1			1						1	1					1								
12-0523.44a			1											1				1									1					1		
Number of samples	1	1	10	1	1	1	1	3	1	6	3	1	4	15	1	3	1	11	1	1	1	2	2	1	2	1	11	2	2	1	1	9	0	1

REFERENCES

[1] Pakendorf, B., & Stoneking, M. Mitochondrial DNA and human evolution. *Annu. Rev. Genomics Hum. Genet.* 2005; 6: 165-183.

[2] Budowle, B., Allard, M. W., Wilson, M. R., & Chakraborty, R. Forensics and Mitochondrial DNA: Applications, Debates, and Foundations. *Annual review of genomics and human genetics* 2003; 4(1): 119-141.

[3] Turchi, C., Buscemi, L., Pesaresi, M., Di Saverio, M., Paoli, M., & Tagliabracci, A. Occurrence of heteroplasmy in related individuals. In *International Congress Series* 2003, Jan; 1239: 553-556.

[4] Hühne, H., & Brinkmann, B. Heteroplasmic substitutions in the mitochondrial DNA control region in mother and child samples. *International journal of legal medicine* 1998; 112(1): 27-30.

[5] Takayanagi, K., Asamura, H., Tsukada, K., Ota, M., & Fukushima, H. Investigation of mtDNA heteroplasmy discordance between mother and child. In *International Congress Series* 2004, April; 1261:380-382.

[6] Baasner, A., Schäfer, C., Junge, A., & Madea, B. Polymorphic sites in human mitochondrial DNA control region sequences: population data and maternal inheritance. *Forensic science international* 1998; 98(3): 169-178.

[7] Butler JM. Chapter 10: Mitochondrial DNA Analysis. In: Forensic DNA Typing. Massachusetts: Academic Press; 2005. p. 241-298.

[8] Date Chong, M., Calloway, C. D., Klein, S. B., Orrego, C., & Buoncristiani, M. R. Optimization of a duplex amplification and sequencing strategy for the HVI/HVII regions of human mitochondrial DNA for forensic casework. *Forensic science international* 2005; 154(2-3): 137-148.

[9] Melton, T., & Nelson, K. Forensic mitochondrial DNA analysis: two years of commercial casework experience in the United States. *Croatian medical journal* 2001; 42(3): 298-303.

[10] Yang, Y., Xie, B., & Yan, J. (2014). Application of Next-generation Sequencing Technology in Forensic Science. *Genomics, Proteomics & Bioinformatics*.

[11] Kinra, S. L. P. The use of mitochondrial DNA and short tandem repeat typing in the identification of air crash victims. *Ind J Aerospace Med* 2006; 50(1): 55.

[12] Weber-Lehmann, J., Schilling, E., Gradl, G., Richter, D. C., Wiehler, J., & Rolf, B. Finding the needle in the haystack: Differentiating "identical" twins in paternity testing and forensics by ultra-deep next generation sequencing. *Forensic Science International: Genetics* 2014; 9: 42-46.

[13] Life Technologies. Targeted sequencing solutions Ion Torrent™ [Internet] Life Technologies; 2014.

[14] Parson, W., Strobl, C., Huber, G., Zimmermann, B., Gomes, S. M., Souto, *et al*. Evaluation of next generation mtGenome sequencing using the Ion Torrent Personal Genome Machine (PGM). *Forensic Science International: Genetics* 2013; 7(5): 543-549.

[15] Melton, T., Holland, C., & Holland, M. Forensic Mitochondria DNA Analysis: Current Practice and Future Potential. *Forensic Science Review* 2012; 24(2); 101.

[16] Tzen, C. Y., Wu, T. Y., & Liu, H. F. Sequence polymorphism in the coding region of mitochondrial genome encompassing position 8389–8865. *Forensic science international* 2001; 120(3); 204-209.

[17] Morovvati, S., Morovvati, Z., & Ranjbar, R. Detecting Rare Triple Heteroplasmic Substitutions in the Mitochondrial DNA Control Region: A Potential Concern for Forensic DNA Studies. *Cell Journal (Yakhteh)* 2011; 13(2):103.

[18] Fendt, L., Zimmermann, B., Daniaux, M., & Parson, W. Sequencing strategy for the whole mitochondrial genome resulting in high quality sequences. *BMC genomics* 2009; 10(1): 139.

[19] Irwin, J. A., Parson, W., Coble, M. D., & Just, R. S. mtGenome reference population databases and the future of forensic mtDNA analysis. *Forensic Science International: Genetics* 2011; 5(3): 222-225.

[20] Sanger, F., Nicklen, S., & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences* 1977; 74(12): 5463-5467.

[21] Berglund, E. C., Kiialainen, A., & Syvänen, A. C. Next-generation sequencing technologies and applications for human genetic history and forensics. *Investig Genet* 2011; 2: 23.

[22] Applied Biosystems. DNA Sequencing by Capillary Electrophoresis Chemistry Guide. Applied Biosystems. Second Edition. 2009.

[23] Gunnarsdóttir, E. D., Li, M., Bauchet, M., Finstermeier, K., & Stoneking, M. High-throughput sequencing of complete human mtDNA genomes from the Philippines. *Genome research* 2011; 21(1): 1-11.

[24] Illumina. An Introduction to Next-Generation Sequencing Technology [Internet]. Illumina. 2013-2014.

[25] New England BioLabs. NEBNext® Fast DNA Library Prep Set for Ion Torrent Instruction Manual [Internet]. New England Biolabs. Version 4.1. 2013.

[26] Timmermans, M. J., Dodsworth, S., Culverwell, C. L., Bocak, L., Ahrens, D., Littlewood, D.T., *et al*. Why barcode? High-throughput multiplex sequencing of mitochondrial genomes for molecular systematics. *Nucleic acids research* 2010: 1-14.

[27] Head, S.R., Komofi, H. K., LaMere, S.A., Whisenant, T., Nieuwerburgh, F.V., Salomon, D.R., *et al*. Library construction for next-generation sequencing: Overview and challenges. *BioTechniques* 2014; 56: 61-77.

[28] Miyamoto, M., Motooka, D., Gotoh, K., Imai, T., Yoshitake, K., Goto, N, *et al.* Performance comparison of second-and third-generation sequencers using a bacterial genome with two chromosomes. *BMC genomics* 2014; 15(1): 699.

[29] van Dijk, E. L., Jaszczyszyn, Y., & Thermes, C. Library preparation methods for next-generation sequencing: Tone downt the bias. *Experimental Cell Research* 2014; 322: 12-20.

[30] Ion Personal Genome Machine® (PGM™) System [Internet]. Life Technologies; 2015 [cited March 3, 2015]. Available from: http://www.lifetechnologies.com/order/catalog/product/4462921

[31] Shendure, J. & Ji, H. Next-generation DNA sequencing. *Nature Biotechnology* 2008; 26(10): 1135-1145.

[32] Life Technologies. Ion Torrent Amplicon Sequencing Application Note [Internet]. Ion Torrent. 2011.

[33] Roby, R. Post-Amplification Yield Gel. Original. Fort Worth, TX: University of North Texas Center for Human Identification; 2012.

[34] Phillips, N., & Curtis, P. Post-PCR mtDNA Processing. 2nd rev. Fort Worth, Texas: University of North Center for Human Identification; 2011.

[35] Curtis, P., & Thomas, J. Maintenance and Use of the 3130*xl* Genetic Analyzer. 5th revision. Fort Worth, Texas: University of North Center for Human Identification; 2011.

[36] Invitrogen. Qubit® dsDNA BR Assay Kits [Internet]. Life Technologies. 2011.

[37] Beckman Coulter. AGENCOURT® AMPURE® XP Protocol [Internet]. Agencourt. 2013.

[38] Agilent Technologies. Agilent High Sensitivity DNA Kit Guide [Internet]. Agilent Technologies. 2013.

[39] Ion Torrent by Life Technologies. Ion PGM™ Template OT2 400 Kit Quick Reference [Internet]. Life Technologies. 2014.

[40] Ion Torrent by Life Technologies. Ion PGM™ Sequencing 400 Kit User Guide [Internet]. Life Technologies. 2013.

[41] Zhao, H., Shen, J., Medico, L., Platek, M., & Ambrosone, C.B. Length heteroplasmies in human mitochondrial DNA control regions and breast cancer risk. *Int J Mol Epidemiol Genet* 2010; 1(3): 184-192.