LOW-LEVEL VARIANT DETECTION IN HUMAN MITOCHONDRIAL DNA USING THE ILLUMINA® MiSeq™ NEXT-GENERATION SEQUENCING (NGS) PLATFORM

A thesis presented to the faculty of the Graduate School of

Western Carolina University in partial fulfillment of the

requirements for the degree of Master of Sciences in Biology

By

Brandon Chase Smith

Director: Dr. Mark Wilson

Associate Professor and Director

Forensic Science Department

Committee Members: Dr. Indrani Bose, Biology

Dr. Jack Summers, Chemistry and Physics

April 2013

ACKNOWLEDGEMENTS

I would initially like to thank Dr. Mark Wilson, without whom this research would not have been possible. His expertise and immeasurable patience have been instrumental to my academic success, not only furthering my core understanding of forensic science but also honing my continued growth as an objective thinker. Thank you for trusting me with the keys to the instrumentation. A very special thanks also goes out to Brittania Bintz, M.S., and Erin Burnside, M.S., both of whom have been beyond helpful and are a continued source of inspiration for an aspiring, young scientist. In particular, Brittania Bintz has shown a phenomenal dedication to my continued growth, not only serving as my thesis reader but also, revising the many drafts it took to get me this far.

I would also like to thank the National Institute of Justice (#2010-DN-BX-K171) for providing the grant that funded this research opportunity.

I would also like to acknowledge Cydne Holt and the forensics applications team at Illumina®, whom not only helped us troubleshoot our methodology but also, in an outstanding measure of good faith, supplied the Forensic Science program with the kits needed to perform this research.

I also extend my thanks to my committee members, Dr. Indrani Bose and Dr. Jack Summers, whose insights have aided this thesis in various ways.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

%MIN – percent minor: the percentage of post-filtered reads that call the 2nd most prevalent base at a particular position
A, C, G, T – nucleotides: adenine, cytosine, guanine, thymine
BWA – Burrows-Wheeler Aligner
bp – base pair
CODIS – Combined DNA Index System
CCD – charged coupled device
CR – Control region of the mitochondrial genome
DAPI – 4,6-diamidino-2-phenylindole
DGGE – denaturing gradient gel electrophoresis
DNA – deoxyribonucleic acid
- **dsDNA** – double stranded DNA
- **mtDNA** – mitochondrial DNA
- **nucDNA** – nuclear DNA
- **ssDNA** – single stranded DNA

EV – expected variant/variation
GUI – graphic user interface
HV1 – hypervariable region 1
- **HV1a** – hypervariable region 1a
- **HV1b** – hypervariable region 1b

HV2 – hypervariable region 2
- **HV2a** – hypervariable region 2a
- **HV2b** – hypervariable region 2b

IGV – intergrative genomics viewer (Broad Institute)
IPC – internal positive control.
LED – light emitting diode
LOD – limit of detection
MS1 – mixture study 1
MS2 – mixture study 2
NGS – next-generation sequencing
NTP – nucleoside triphosphate
- **dNTP** – deoxynucleoside triphosphate
- **ddNTP** – dideoxynucleoside triphosphate

NumtS – nuclear pseudogene
PAL – pooled amplicon library
PEG – polyethylene glycol
PCR – polymerase chain reaction.
Pe – probability of error
PE – paired-end
Qs and Ks – questioned and known samples

qPCR – real-time or quantitative PCR
QV – quality value or q-score
Q30 – a Q-score of 30
RB – reagent blank
rCRS – revised Cambridge Reference Sequence
RFU – relative fluorescent unit
SAM – sequence alignment map
SNP – single nucleotide polymorphism
SPRI – solid phase reversible immobilization
SWGDAM – Scientific Working Group on DNA Analysis Methods
T#oDs – Total number of deviants: basecalls that oppose the reference call at a particular position
T#oRs – Total number of reads: pre-filtered reads that have aligned over a particular position
TCEP – tris(2-carboxyethyl)phosphine
TS1 – tissue study 1
TS2 – tissue study 2
UV – unexpected variant/variation
WGA – whole genome amplificaiton

LOW-LEVEL VARIANT DETECTION IN HUMAN MITOCHONDRIAL DNA USING THE ILLUMINA® MiSeq™ NEXT-GENERATION SEQUENCING (NGS) PLATFORM

Brandon Chase Smith, B.S.
Western Carolina University (August 2013)
Director: Dr. Mark Wilson

When challenged by difficult biological samples, the forensic analyst is far more likely to obtain useful data by sequencing the human mitochondrial DNA (mtDNA). Next-generation sequencing (NGS) technologies are currently being evaluated by the Forensic Science Program at Western Carolina University for their ability to reliably detect low-level variants in mixtures of mtDNA. The sequence profiles for twenty individuals were obtained by sequencing amplified DNA derived from the mitochondrial hypervariable (HV) regions using Sanger methods. Two-person mixtures were then constructed by mixing quantified templates, simulating heteroplasmy at discrete sites and in defined ratios. Libraries of unmixed samples, artificial mixtures, and instrument controls were prepared using Illumina® Nextera® XT and deep-sequenced on the Illumina® MiSeq™. Analysis of NGS data using a novel bioinformatics pipeline indicated that minor variants could be detected at the 5, 2, 1, and 0.5% levels of detection. Additional experiments which examined the occurrence of sequence variation in hair tissue demonstrates that a

considerable amount of sequence variation can exist between hairs and other tissues derived from a single donor.

CHAPTER ONE: BACKGROUND

Section 1.1: The biology of mitochondrial DNA and its relevance to forensic casework

There are hundreds of mitochondria within each nucleated cell of human body (Robin & Wong 1988). Visualizations of DAPI-stained mitochondrial DNA (mtDNA) in a human ovarian cancer cell line indicated a range of 1 to 15 copies of mtDNA per mitochondrion (Satoh & Kuroiwa 1991). This gives an estimated range of hundreds to thousands of copies of mtDNA per cell and an average of approximately five hundred copies of mtDNA molecules in cells comprising most somatic tissues (Satoh & Kuroiwa 1991). Since mtDNA is present in high copy number and possesses an exonuclease resistant, circular structure (see Figure 1), the likelihood of its recovery from biological evidence is significantly higher than its nuclear counterpart (DiZinno et al. 1999) and is the preferred molecule for analysis when extracting from challenging sample types which likely contain nuclear DNA that is either highly degraded or nonexistent. Such samples would include naturally shed hairs, aged teeth, and bone material.

The sequencing of mtDNA from evidentiary material is a valuable asset to the forensic scientist. Forensic analysts sequence amplicons, or amplified DNA, derived from the human mitochondrial control region (CR) to obtain useful data (Wilson et al. 1995a). The CR is a noncoding region of the mitochondrial genome and is of particular interest to forensic scientists because of the frequency of variants, including single nucleotide polymorphisms (SNPs) which occur there. These variants enable the analyst to identify a haplotype within the different mitochondrial lineages of the human

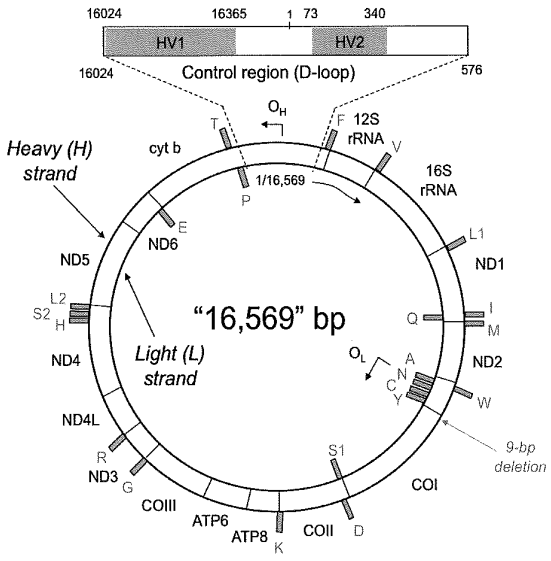

Figure 1 – Illustration of the human mitochondrial genome (Butler 2011). The hypervariable regions of the Control Region (CR) are indicated by the shaded regions entitled 'HV1' (16024-16365) and 'HV2' (73-340).

population (DiZinno et al. 1999). In forensic contexts, the utility of mtDNA sequence interpretation includes: the elimination of possible contributors of source material, establishing investigative leads in cases regarding missing persons, tracing maternal lineages among the human population, and aiding in the reassociation of a victim's remains after a mass disaster or armed conflicts. Forensic casework places a clear objective of obtaining meaningful results that meet the rigorous standards of quality assurance. The workflow used by crime labs in order to obtain mtDNA sequence data is presented in the grayed flowchart shown in Figure 2. Interpretations of mtDNA sequence data cannot however, be used to establish the absolute identity of its biological contributor, a consequence of its mode of inheritance.

The inheritance of mtDNA haplotypes are largely influenced by the maternal contribution to the developing zygote (Hutchinson et al. 1974). Explanations for this mode of inheritance include a stark numerical difference in the mtDNA copy numbers of oocytes relative to sperm cells (Chen et al. 1995, Hecht et al., 1984) and a ubiquitin-

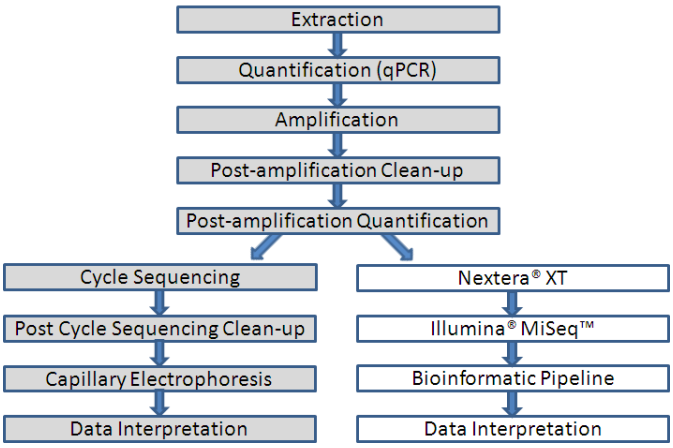

Figure 2 – Branched workflow of Sanger and next-generation sequencing methods for generating mitochondrial sequence data.

dependent proteolysis pathway that is implicated in the targeted degradation of paternally derived mitochondria and concomitant mtDNA (Sutovsky et al. 2003). Though exceptions to complete maternal transmission have been reported (Schwartz & Vissing 2002), mtDNA haplotypes are highly conserved from one maternal generation to the next and explains why the same mtDNA haplotype is observed across maternal siblings and other maternal relatives.

Heteroplasmy is the presence multiple mtDNA haplotypes within an individual and it has been observed in two distinct forms: as sequence (Gill et al. 1994, Ivanov et al. 1996, Wilson et al. 1997) and as length heteroplasmy (Pfeiffer et al. 2004). Sequence heteroplasmy can be observed when two or more nucleotides are present at a single base position. Length heteroplasmy can be observed when two or more different lengths of a homopolymeric C-stretch are present within the mtDNA sequence. Since the first documented occurrence of the phenomenon within a forensic context (Gill et al. 1994, Ivanov et al. 1996), heteroplasmy has merited numerous scientific inquiries. In an experiment to determine the interpretive limit for denaturing gradient-gel electrophoresis (DGGE) systems, the authors prepared mixtures of mtDNA for analysis (Tully et al.

2000). The threshold for mixture detection on DGGE was found to be approximately 1%. Using DGGE and amplicons derived from human mitochondrial hypervariable region 1 (HV1), the authors were able to detect low levels of heteroplasmy in 13.8% of its sample population of unrelated individuals (n=253), demonstrating that heteroplasmy is not uncommon. Additionally, the authors found that by using DGGE, heteroplasmy could be discerned at 16 different locations within HV1, that it occurred as both sequence and length variants, and that two donors exhibited triplasmy, a condition in which heteroplasmy occurs at two different locations within an individual. Given that mtDNA has shown higher mutation rates than nuclear DNA (Bogenhagen 1999), which is likely attributed to fewer DNA repair mechanisms and an inability of the mtDNA polymerase to proofread (Kunkel 1981), lends credence to the idea that trillions of mtDNA copies are not likely to share a single, uniform mtDNA sequence.

In humans, ovulation results in one or more mature oocytes are released from an ovarian follicle. If fertilized, the degree to which heteroplasmy occurs in these progenitor cells, whether high, low, or non-existent, will largely influence the degree of heteroplasmy observed in resulting progeny. Genetic bottlenecks such as these (Figure

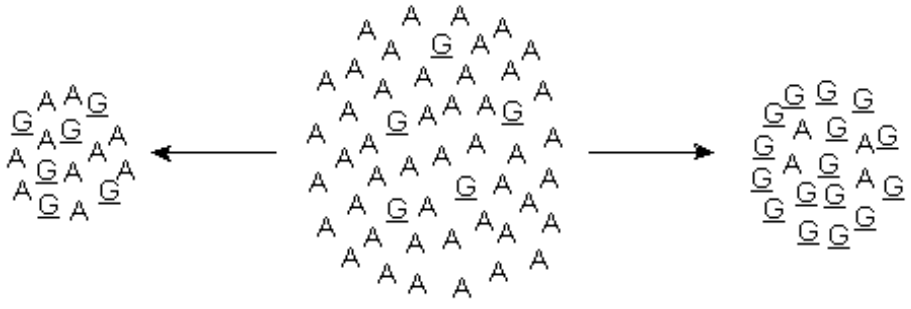

Figure 3 – Reduction of mtDNA haplotypes as a result of a genetic bottleneck (Arrows) in the parental (Center) that appears homoplastic for a particular base position. Minor variant in parental population is denoted as G. (Left) A moderate degree of allelic drift results in progeny that are heteroplasmic in appearance. (Right) A wide degree of allelic drift results in progeny that appears homoplastic for the minor variant, which has become the major component of a mixed population.

3), in which a large population is culled to a few representative types, can have a profound impact on observed population structure and may explain why single base differences have been observed across parent and progeny (Parsons et al. 1997). These base substitutions and the rate at which they approach fixation within a population of individuals, forms the basis for human mitochondrial evolution. While the implications of this are significant from an evolutionary perspective, of more relevance to the situations encountered during forensic casework is that a similar bottleneck event is expected to occur during the histogenesis of hair.

The growth and replacement of hair tissue is fairly complex. The cells that feed the emerging hair follicle represent a small, clonally propagating population (Linch et al. 2001). Unlike other tissues, hair emerges from the body; when these cells die, they are not reabsorbed. Therefore, the mtDNA extracted from hair shafts is expected to reflect the heteroplasmic content of the original population of cells which propagated from the follicle. Similar to the bottleneck event that occurs during oogenesis, the reduction of the mtDNA copies to a smaller representative population has the potential to cause a wide degree of sequence variation. This is believed to be the biological explanation for the wide degree of sequence variation observed across multiple hair samples collected from a single donor in comparison to other, more homogenous tissue types (Wilson et al. 1997), as well as in studies of mtDNA length variants observed in sub-cloned hair extracts of monozygotic twins (Pfeiffer et al. 2004). This research underscores the potential benefit of technologies that can more accurately resolve low-level mixtures of human mtDNA. The ability to resolve minor variations, or low-level mitochondrial haplotypes within a mixed population of mtDNA molecules, is dependent on the sequencing chemistry and

the resolution power of the process. Current methods of mtDNA sequencing utilize Sanger sequencing and capillary electrophoresis.

Section 1.2: Overview of chain-terminating sequencing and capillary electrophoresis

The dideoxy chain-terminating chemistry used by BigDye® Terminator v1.1 Cycle Sequencing Kit (Applied Biosystems 2002) and similar kits generate DNA fragments of varying lengths, each with a base specific fluorophore attached. The cyclical process of deoxynucleotide triphosphate (dNTP) incorporation into every position of the extending DNA chain is similar to PCR but instead of only incorporating dNTPs, a dideoxynucleotide triphosphate (ddNTP) may be incorporated by chance. The chemistry of ddNTP incorporation is demonstrated in Figure 4. These ddNTPs lack a hydroxyl group on the 3' carbon of the sugar moiety. Without this 3' hydroxyl group, the polymerase cannot induce the phosphodiester bond needed to link one nucleotide to the next, and extension of the DNA chain will cease. The terminating ddNTP at the 3'-end of these fragments has a covalently attached base specific fluorophore, which is an attached moiety capable of absorbing a photon, enabling fluorescence upon relaxation.

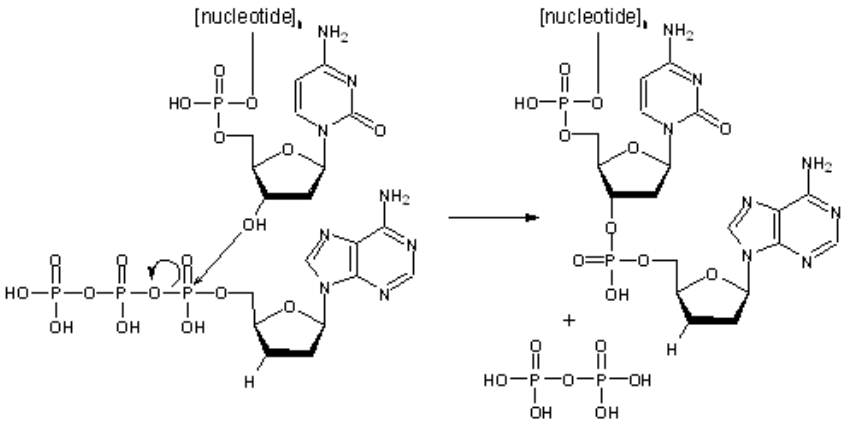

Figure 4 – During cycle sequencing, a ddNTP lacking the 3' hydroxyl group may be incorporated into the growing nucleotide chain. Without its hydroxyl moiety, the 3' carbon lacks the nucleophile needed to interact with the phosphate group of the next dNTP, effectively terminating DNA extension. Not shown: polymerase and the covalently attached fluorophore on the ddNTP.

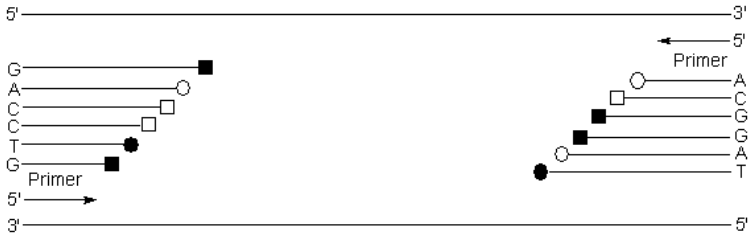

Figure 5 – Simplified illustration of Sanger sequencing products. Forward and reverse primers would be separated into different centrifuge tubes to keep forward and reverse amplicons from mixing. Circles and squares represent chain terminating ddNTPs. In this example, the forward sequence would be read 5'-GTCCAG-3' and the reverse sequence would be read as 5'-ACGGAT-3'.

The result of this sequencing chemistry is a collection of single stranded DNAs, of varying lengths, where each fragment is covalently bound to a base specific fluorophore of the last base incorporated. The products of this reaction are depicted in Figure 5.

After these fragments are cleaned of residual ddNTPs, vacuum centrifuged and resuspended in Hi-Di™ formamide, the cycle sequencing products are run on the Applied Biosystems® 3130*xl* Genetic Analyzer. This and other capillary electrophoresis instruments can separate single stranded DNA products by size (Applied Biosystems 2004). As the negatively charged DNAs of the sequencing reaction separate through a capillary array filled with a polymer matrix that acts as a sieving medium, fragments of increasing size pass through a detection window sequentially. There, the fragments are exposed to an argon laser which excites the base specific fluorophore of the terminating ddNTP and as the fluorophore relaxes, it emits a wavelength of light specific to the last base incorporated. This fluorescence is captured by a charged coupled device (CCD) camera. The interval at which light data is collected is then converted into an electropherogram peak, where the different wavelengths of light are represented by a four color scheme and their associated intensities are measured in relative fluorescence units (RFUs). Put succinctly, time and fluorescent emission data are used to determine the base-by-base sequence of the DNA template.

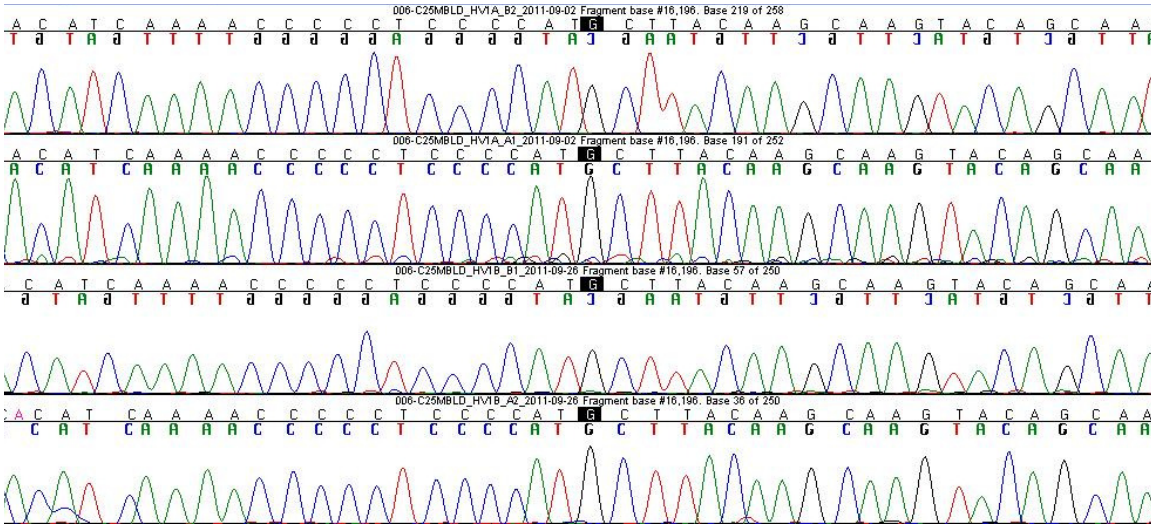

Figure 6: Chromatograms for sequence data obtained from HV1a and HV1b of donor 006-CM25Blood. These chromatograms were visualized using Sequencher v4.8, a sequence visualization software. The second chromatogram has a 'noisy' baseline.

While the use of chromatograms have been and, for the immediate future, will continue to provide useful information to forensic examiners, the peaks are however, not quantitative—peak height cannot be correlated to an exact number of DNA molecules that gave rise to it (Parker et al. 1995, Parker et al. 1996). This means that interpretation of potentially heteroplasmic samples is somewhat of an art, whereby forensic examiners must be trained to distinguish true mixtures above the baseline noise of the instrument. In instances where chromatograms have 'noisy' baselines, a low-level mixture is less likely to be called with confidence. Experimentally, the limit of resolution for capillary electrophoresis instruments in their ability to correctly call a mixed position by forensic examiners has been demonstrated to be approximately 10% (Wilson et al. 1995b), falling well above the reported sensitivities of other, non-sequencing forms of analyses (Underhill et al. 1999, Tully et al. 2000).

Section 1.3: Mitochondrial sequence interpretation in forensic casework

Sequence analysis of mtDNA in forensic casework involves comparing question samples (Qs), samples whose biological contributor(s) are unknown, to known samples (Ks), in which the contributor is known. Trained forensic examiners make direct observations of resulting mtDNA sequence data, which is in chromatogram format. Resultant sequences for Qs and Ks are directly compared to one another and the sequence differences are noted. Figure 7 gives an example of how Qs and Ks are compared. In

Figure 7 – Hypothetical sequence data for Qs and Ks. A) Q and K have the same base at every position. B) Q and K have a shared base at every position and a shared base at the underlined position. C) Q and K have a single base difference at the underlined position. D) Q and K have two bases that differ at the underlined positions.

example 7a, there is a common base at every position within the sequence window. In example 7b, the K indicates that both cytosine and thymine are observed at the highlighted position and no contamination is suspected. Since thymine is observed at the indicated position in both Q and K, the sequences are concordant. In example 7c, there is a single base difference between the Q and K, and in example 7d there are two base differences. Conceptually, the interpretations of the various insertions/deletions that occur around the C-stretch region of HV2 are no different. Consider the sequences presented in Figure 8. In example 8a, the Q and K have the same number of cytosine residues. In example 8b, the Q has a C-stretch type (C8TC6) that is shared with the K, which shows length variants (C8TC6/C9TC6). Example 8c has a single base

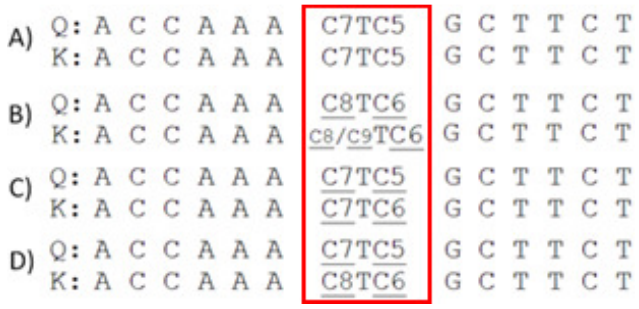

Figure 8 – Hypothetical sequence data for Qs and Ks. A) Q and K have the same base at every position. B) Q and K have a shared C-stretch haplotype. C) Q and K have a single have a single difference within the C-stretch region. D) Q and K have two differences within the C-stretch region.

difference and 8d has two base differences. To interpret observed sequence differences in routine casework, the Scientific Working Group on DNA Analysis Methods (SWGDAM 2003) suggests these guidelines be followed:

- Exclusion – if there are two or more nucleotide differences between the questioned and known samples, the sample can be excluded as originating from the same person or maternal lineage.
- Inconclusive – if there is one nucleotide difference between the questioned and known samples, the result will be inconclusive.
- Failure to Exclude – if the sequences from questioned and known samples under comparison have a common base at each position or a common length variant in the HV2 C-stretch, the samples cannot be exclude as originating from the same person or maternal lineage.

Assuming that no other differences occur outside the sequence windows that are presented in Figures 7 and 8, and using the SWGDAM guidelines for interpretation, the forensic examiner would not be able to exclude the known profile as a possible contributor for the questioned sample in examples 7a, 7b, 8a, and 8b. For the indicated examples, they would reach a conclusion of *failure to exclude*. For 7c and 8c, the conclusion would be *inconclusive* and for 7d and 8d, an *exclusion* would result.

A short-hand notation for describing sequence differences aligns questioned and known sequences against the revised Cambridge Reference Sequence (rCRS), a standard

reference sequence of human mtDNA (Anderson et al. 1981, Wilson et al. 1995a). Differences between the samples and the rCRS are described using the positions and the respective bases in which they differ, as demonstrated in Figure 9. These notations are used when documenting the observed differences between collected samples, when uploading reference sequences into the forensic mtDNA database, and when cross-referencing the rarity of an observed mtDNA haplotype against those stored in the database.

The common mtDNA sequence found in the Q/K samples is compared to a mtDNA population database. These databases allow the forensic scientist to derive a weight assessment of the association. Conveying this estimate of rarity involves counting the number of times a particular haplotype is observed within the DNA database among the different defined subsets of the human population and applying a 95% upper-bound confidence interval to this calculation (Wilson et al. 1993, Holland & Parsons 1999). As such, the strength of this assessment is based on the number of profiles within the database, the current estimates for which are approximately 15,000 profiles and growing (Parson & Dür 2007, Melton et al. 2012).

(a) mtDNA Sequences Aligned with rCRS (positions 16071–16140)

```
                16080       16100       16110       16120       16130       16140
rCRS  ACCGCTATGT ATTTCGTACA TTACTGCCAG CCACCATGAA TATTGTACGG TACCATAAAT
  Q   ACCGCTATGT ATCTCGTACA TTACTGCCAG CCACCATGAA TATTGTACAG TACCATAAAT
  K   ACCGCTATGT ATCTCGTACA TTACTGCCAG CCACCATGAA TATTGTACAG TACCATAAAT
```

(b) Reporting Format with Differences from rCRS

Sample Q	Sample K
16093C	16093C
16129A	16129A

Figure 9 – Comparison of hypothetical sequence alignments (Butler 2011). A) Question and known samples aligned beneath the reference sequence. B) Short-hand notation for reporting differences.

Section 1.4: The Illumina® MiSeq™

Next-generation sequencing (NGS) technologies have introduced new methods of DNA sequencing. While capillary electrophoresis instruments are designed to generate a sequence from a population of DNA fragments arising from a single amplicon, NGS technologies provides sequence data from millions of individual DNA molecules in parallel. Shared among all NGS platforms are (1) the immobilization of DNA molecules to a medium, separating them in space, and (2) the clonal amplification of the progenitor molecule to increase signal intensity. Beyond that, each platform has its own unique method of sequencing, and these emerging methods have dramatically reduced the time it takes to sequence DNA by increasing the amount of data obtained in each individual run, which also allows multiple donors to be sequenced in tandem.

With the advent of massively parallel sequencing technologies, it is possible to resolve and quantify mixtures of mtDNA (Bintz et al. 2013, Andréasson et al. 2006a) as well as broaden the sequencing breadth to areas outside the CR (Andréasson et al. 2006b). New extraction techniques that optimize the recovery of mtDNA from hair shafts (Burnside et al. 2013) and the development of whole genome amplification (WGA) techniques (Qiagen® 2011) may even make it feasible to routinely deep sequence the entire mtGenome from short hair fragments. These implications make next-generation sequencing technologies attractive to forensic science as their implementation could significantly impact the future of mtDNA analysis.

The Illumina® MiSeq™ is a sequencing-by-synthesis instrument currently housed in the Forensic Science DNA sequencing facility at Western Carolina University.

This platform utilizes a unique sequencing chemistry and a proprietary, optically transparent flow cell, covered in a dense lawn of oligonucleotide anchors (Bentley et al. 2008, Illumina® 2012a). Prepared single stranded template molecules hybridize to the flow cell bound oligonucleotides if they contain the complementary adapter sequence. These adapters are designed to be randomly incorporated into the DNA template to be sequenced during Nextera® XT processing, an enzymatic sample preparation step (Caruccio 2011, Illumina® 2012b).

Nextera® XT 'tagments' dsDNA by randomly fragmenting it into varying lengths and ligating oligonucleotide tails onto these fragments, which then serve as priming sites for a limited cycle PCR (Figure 10). This step incorporates the bidirectional indices and adapter sequences needed to bind template molecules to the flow cell. The bidirectional indices are analogous to a genetic barcode which can allow for the bioinformatic parsing of sequence data belonging to a particular individual or treatment. After PCR, a library

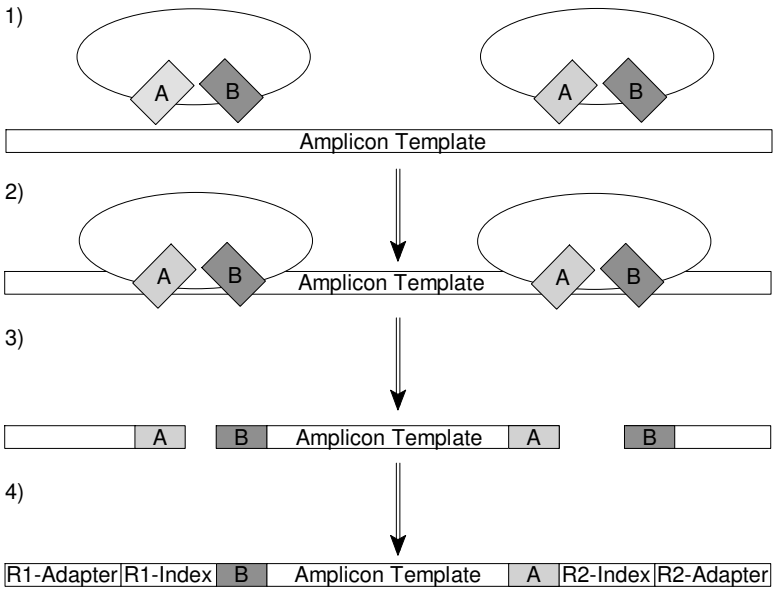

Figure 10 – Illustration of how Nextera® XT prepares template molecules for sequencing (Illumina® 2012). (1) Transposons are introduced to template molecules and (2) bind to templates at random locations. (3) The enzymatic fragmentation of DNA templates incorporates forward and reverse adapter sequences. (4) Adapter sequences are the targets of a limited cycle PCR which incorporates the flanking index and adapter sequences.

purification step using solid phase reversible immobilization (SPRI) beads in the presence of polyethylene glycol (PEG) and a salt buffer will effectively remove unincorporated PCR primers and short (<100 bp) fragments from solution without any need for gel electrophoresis size separation (DeAngelis et al. 1995). Removal of short fragments is critical because they can compete for space on the flow cell. Upon clean-up, a normalization step dilutes the libraries to a desired concentration. By the end of sample preparation, the pooled amplicon library contains sodium hydroxide which keeps the library single stranded. Samples are then loaded onto the MiSeq™ reagent cartridge for on-board cluster generation and subsequent deep sequencing.

When flowed through the flow cell, single stranded template molecules will hybridize to the lawn of primers bound to the flow cell. These individual single stranded templates are then clonally amplified via on-board cluster generation, which is depicted in Figure 11. Clonal amplification, sometimes referred to as 'bridge amplification,' begins with washing NaOH from the flow cell, lowering the pH of the environment so

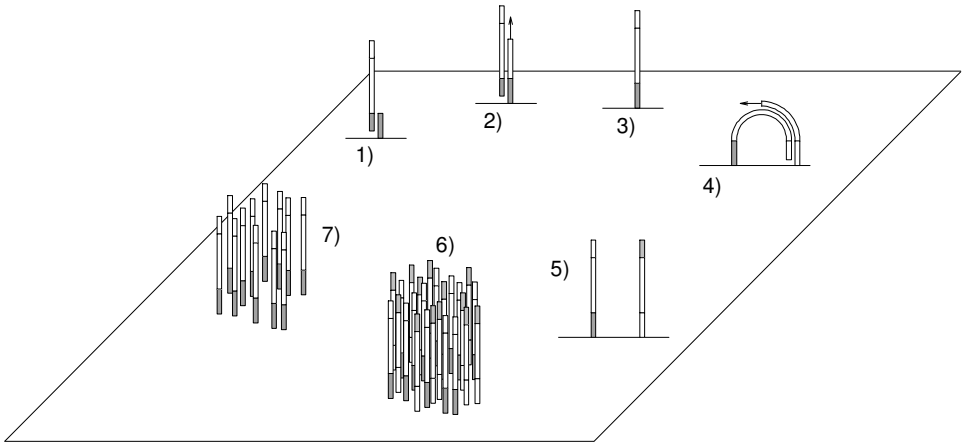

Figure 11: Summary of cluster generation (Bentley et al. 2008). 1) Single stranded template hybridizes to its complementary oligonucleotide anchor. 2) Polymerase and dNTPs are flowed over the flow cell to synthesize the template complement. 3) NaOH is flowed over the flow cell to denature DNA—only DNA complements anchored to the flow cell will remain. 4) NaOH is washed away, allowing DNA to hybridize to a complementary oligonucleotide anchor that is in close proximity forming a DNA bridge. Polymerase and dNTPs are flowed over the flow cell to synthesize to extend the bridge strand. 5) NaOH is flowed over the flow cell to denature the DNA bridge. 6) Repetitions of steps 4 and 5 will result in a bidirectional cluster of DNA derived from a single template molecule. 7) Enzymatic cleavage of an anchoring adapter sequence will leave products of a unidirectional flow.

that the complementary strand may be synthesized to the oligonucleotide anchor. The reverse adapter sequence is synthesized as well. A second denaturation wash with NaOH allows for the separation of the progenitor template molecules from their complements that are now covalently bound to the flow cell. Removal of NaOH will then allow the reverse adapter on the newly synthesized complement to hybridize to another oligonucleotide anchor, creating a DNA bridge. This bridge creates the primer-template complex, to which polymerase can be recruited for extension of the nucleotide chain. Repeated processes of denaturation, annealing, and extension will create clusters containing approximately two thousand clonal molecules.

Sequencing on the Illumina® MiSeq™ uses the proprietary TruSeq™ family of reagents. The TruSeq™ design utilizes specialized fluorescently labeled, chain terminating dNTPs. These reversible terminating dNTPs have a base-specific fluorescent moiety and 3'-O-azidomethyl blocking group, both of which are removed in the presence

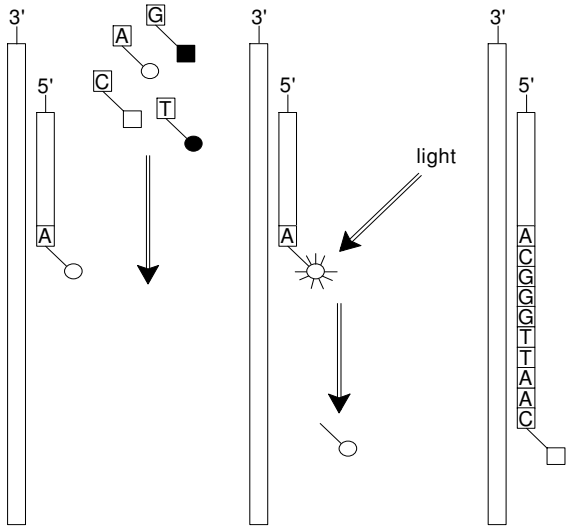

Figure 12: Sequencing using the TruSeq™ family of reagents (Bentley et al. 2008). (Left) Reversible terminating dNTPs are washed across the flow cell after the hybridization of a sequencing primer. Adenosine matches its complement and extends the sequencing chain by one base. (Middle) Clusters are excited by light causing the emission of light by the base specific fluorophore which is then recorded by the instrument. The fluorophore and 3'-blocking group are enzymatically cleaved, allowing the extension of the sequencing chain. (Right) Iterative cycles of base incorporation, excitation, recording, and deprotecting result in the bound template sequence.

of tris(2-carboxyethyl)phosphine (TCEP), not only flushing out the fluorescent dye of the previously incorporated dNTP but also enabling incorporation of the next nucleotide in sequence (Milton et al. 2004, Bentley et al. 2008). In the first step of sequencing, clusters are denatured and the reverse strands are eliminated by a cleavage reaction. This leaves bound templates of unidirectional flow. Sequencing is initiated, first, by hybridizing a sequencing primer specific to index 1 of the bound template molecules. Then, polymerase incorporates a single dNTP into the growing chain. Excess reagents are washed away and laser excitation of the newly integrated fluorophores causes emission of light at wavelengths specific to the base incorporated. This initial detection of light helps identify DNA clusters. The instrument gives each cluster a unique Cartesian coordinate based on this location within the flow cell. Base calling is achieved by calculating the signal intensities generated by each cluster. Once recorded, the fluorophore and the 3' blocking group on the chain terminating dNTP are chemically cleaved with TCEP and the 3' hydroxyl group is simultaneously regenerated, allowing the nucleotide chain to extend. Six additional cycles of single dNTP incorporation, washing excess dNTPs away, reading the signal intensities at specific Cartesian coordinates, and then cleaving the fluorophore allows the Illumina® MiSeq™ to record the forward index sequence of bound templates. Once the indexing cycle is completed, the indexing primer and associated reads are denatured and washed off the flow cell. A second sequencing primer is then flowed into the flow cell and hybridizes to template molecules at a position upstream of DNA template. Again, the cyclic repetition of dNTP incorporation, washing excess dNTPs away, recording signal intensities, and then cleaving the fluorophore allows the instrument to sequence up to 150 bases of each individual DNA cluster.

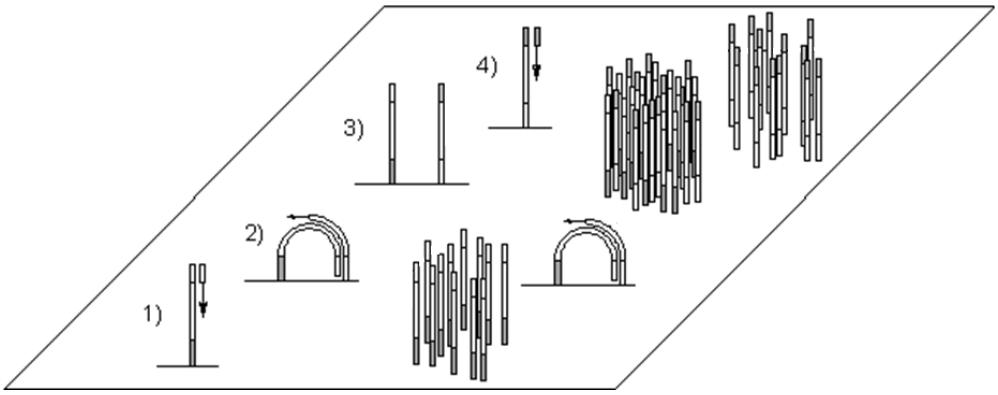

Figure 13: Summary of paired-end turn-around (Bentley et al. 2008). (Left) Single molecule representation of the event. (Right) Cluster representation of the event. 1) After 'Read 1' sequencing ends, NaOH is flowed across the flow cell to denature the sequencing primer and associated bases. 2) Removal of NaOH allows for the 'DNA bridge' to form. Extension is achieve one base at a time. When the sequence has extended to the reverse index, the index is sequenced and recorded. 3) After the reverse index is sequenced, the reverse strand is fully synthesized. Forward and reverse strands are made single stranded with NaOH. 4) The forward strand is enzymatically cleaved, leaving only the reverse strands. 'Read 2' sequencing primer hybridizes and the reverse strand is sequenced.

Paired-end turn-around is the process by which clusters sequenced during the first set are inverted by one cycle of bridge amplification, in effect, synthesizing the reverse strand. For greater clarity, Figure 13 presents a diagram of this event in both a single molecule and a cluster format. After the sequencing the forward read, the sequencing strand is denatured from anchored template molecules by flowing NaOH through the flow cell. Removal of NaOH allows anchored templates to hybridize to a nearby, complimentary anchor, creating the DNA bridge. The reverse complement is synthesized with the TruSeq™ reversible terminating dNTPs. This 'dark cycle' records no images. Instead, it extends the complement strand in well-controlled, incremental steps. This extends the newly synthesized chain to be adjacent to the reverse index. Once the chain reaches the appropriate length, the reverse index is sequenced and recorded. When a cluster has both its forward and reverse indices sequenced, it is given its identity. Upon sequencing the reverse index, unlabeled dNTPs and polymerase are flowed across the flow cell, fully synthesizing the complementary, reverse strand. After the reverse strand is synthesized, the forward reads are enzymatically cleaved and removed from the flow

cell, leaving only templates of a reverse directionality. A third primer is then hybridized to the template molecules and sequencing will commence in the reverse direction.

An optimized run on a v1 flow cell can achieve over five million clusters from which to derive sequence data. New iterations of flow cells can achieve 3 times that number and new kits can sequence 100 bases further in each direction. This depth of coverage is achieved by the paired-end sequencing of millions of DNA clusters derived from single template molecules, in parallel, coupled with the ability to parse individual treatments from a mixed library. This process allows a forensic analyst to scrutinize mtDNA sequence data from many individuals, at a degree of resolution that could not be previously achieved. However, before it can be determined that an observed minor variant is above threshold, it is required to necessary to establish the limits of detection of minor variants.

Section 1.5: Intentions of research

Given the limits of detection of traditional Sanger sequencing, subtle mixtures of mtDNA may not be detected. With the advent of next-generation sequencing technologies, like the Illumina® MiSeq™, not only will it be possible to detect minor variants with greater reliability from single and multi-source samples but it will also offer forensic science benefits that are beyond the scope of this project. Beyond resolving low-level variants, deep sequencing may also (1) sequence entire mtGenomes in order to identify SNPs which occur outside the CR, (2) substantially decrease the time it takes to interpret sequence profiles, and (3) optimize processing sizes in a manner that is both

time and cost efficient and subsequently, (4) may allow for the rapid generation of larger mtDNA databases to support forensic casework analyses.

This research seeks to evaluate the Illumina® MiSeq™ for mtDNA sequencing analysis. It will experimentally determine at what level of resolution this instrument detects minor variants in mixed samples of mtDNA. It will also discern the degree of sequence variation that can be observed, if any, across forensically relevant tissue types. Before performing these experiments, it will be necessary to obtain CR reference profiles for twenty individuals using the traditional methods of cycle sequencing and capillary electrophoresis on an Applied Biosystems® 3130*xl* Genetic Analyzer. With this Sanger data, candidates will be selected for two person mixtures. Using blood and buccal extracts, the four HV amplicons will be amplified using a high fidelity polymerase in order to minimize the occurrence of polymerase induced sequence misincorporation. After these amplicons are quantified, artificial mixtures will be prepared using two donors at four levels of resolution—5, 2, 1, and 0.5%—in order to simulate a biological mixture at discrete sites and in defined ratios. These samples will then be prepared for sequencing on the Illumina® MiSeq™ and an analysis of produced sequence data will

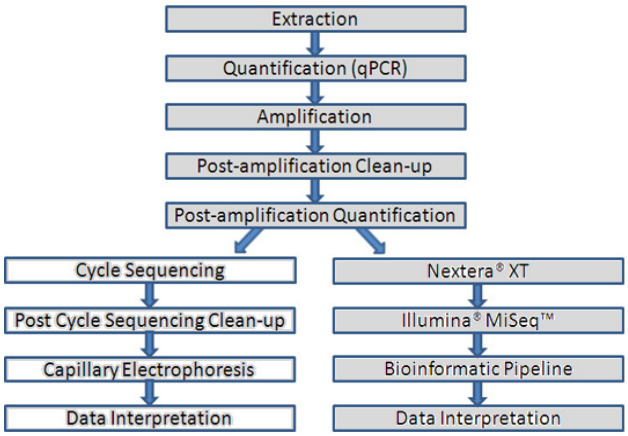

Figure 14 – Branched workflow of Sanger and next-generation sequencing methods for generating mitochondrial sequence data.

focus on the effects of alignment strategies, filtering regimes, potential sources of noise, setting statistical thresholds, and minor variant interpretation. This project will also perform an analysis of sequence data derived from hair samples by comparing CR sequence data produced from five different hair samples against the CR sequence data of a buccal extract from selected donors. Data analysis will focus on alignment strategies, filtering regimes, potential sources of noise, setting statistical thresholds for data interpretation, and minor variation.

CHAPTER TWO: BIOINFORMATICS

Section 2.1: Conclusions from the mixture experiment performed on the Illumina® Genome Analyzer II_x.

The major conclusions drawn from an early mixture experiment on the Illumina® $GAII_x$ were that (1) the targeted sequencing of amplicons, as opposed to fragmented genomes, hinders basecalling accuracy and that (2) the on-board variant calling software could not confidently call minor variants in mixtures of mtDNA. A fully balanced genome is one in which there is an equal distribution of the four bases at every cycle of sequencing (Illumina® 2012a). This is important for Illumina® based platforms because these instruments produce high resolution .TIF images in order to record base-specific fluorescence. At the time it was believed that modified primers, shown in Figure 15, which incorporated flanking adapters and the specific index sequence could bypass the need for a genomic shearing device. Instead, it created an unbalanced genome. For unbalanced genomes, like those produced during targeted amplification, sequencing begins at the same position across all template molecules. When combined with templates that have a high degree of sequence homology, the resulting .TIF images are

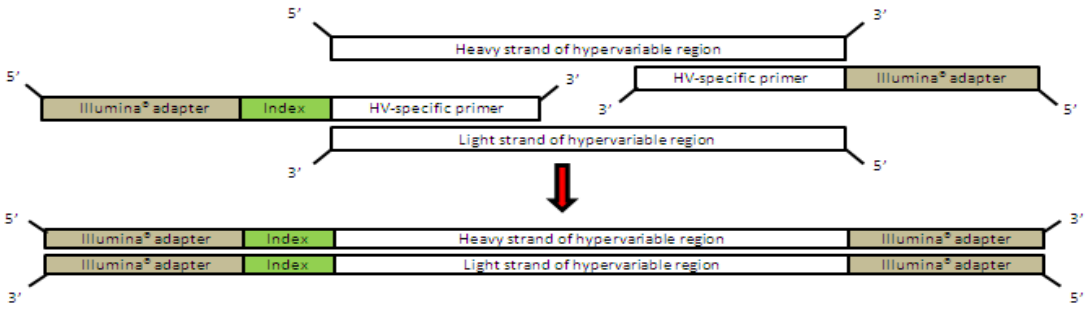

Figure 15: Modified primer design used to create templates ready for sequencing on the Illumina® $GAII_x$. The targeted deep-sequencing of amplicons prepared in this method generated unbalanced genomes.

saturated with the fluorescence of a single base at every cycle of sequencing. Illumina® instruments are not prepared for this and a subsequent decrease in basecalling confidence is observed. The development of Nextera® XT, which prepares templates by enzymatically fragmenting them at random locations, resolved the issue of library balance and greatly improved the quality of subsequent runs. Variant analysis however, necessitated the development of a novel bioinformatics pipeline.

Section 2.2: Quality scores and the development of a novel bioinformatics pipeline for paired-end datasets

Bioinformatics is a field of study dedicated to the analysis of data produced by molecular biology using computational science. It has coevolved and expanded in its utility with the recent advancements in biotechnology and offers many methods of analysis. The bioinformatics pipeline developed for this project is a series of steps that will analyze NGS datasets by demultiplexing mixed datasets, treating the data to preserve the best quality data, aligning the sequence data against the reference of interest, and may then offer some visualization of the reads against an aligned reference and report their respective basecalls and quality scores.

During sequencing on the Illumina® MiSeq™, each cluster is given a unique cluster identifier. If the instrument has confidence that it is recording data from a single cluster, the cluster 'passes filter.' Only clusters that pass filter record sequence data and corresponding base quality. The data is recorded in the .fastq file format, an example of which is presented in Figure 16. Individual .fastq files are separated by the instrument according to the two indices selected and defined by the user prior to sequencing. If a

```
@M00824:4:000000000-A1EA5:1:1101:19484:2375 1:N:0:6
CCTGTAGTACATAAAAACCCAATCCACATCAAAACCCCCTCCCTATGCTTACAAGCAAGTACAGCAATCAACCCTCAACTATCA
+
??AA?BBBDDDDEEDDGGGGGGIIIHIIIIIIIIHIHHHHHHHHIIIHHHHIHHIIHIHIIIHIHHIFCGHHIHIHIIIHHIIIII
```

Figure 16: An entry of the .fastq file type. (Top) Cluster identifier, stating its unique Cartesian coordinate. (Middle) The base sequence derived from sequencing the DNA cluster. (Bottom) ASCII code expressing respective base quality. For example, the ASCII code '?' = Q30, 'A' = Q32, and 'I' = Q40.

paired-end run is performed, sequence information belonging to a particular index combination is further sub-divied into forward and reverse .fastq files. Since a cluster is sequenced in both forward and reverse directions, the unique cluster identifier appears in the corresponding forward and reverse .fastq files. These entries also pair each base of sequence data with a single character ASCII symbol. These symbols are a method of conveying the quality scores (or Q-scores) of the *phred*-based quality scoring language of basecall accuracy. A Q-score is a shorthand notation of the probabilistic model used to assess bascalling error. Illumina® has adopted the *phred* scale for representing the base quality of sequence data. The *phred*-based algorithm used by most Sanger compatible software was designed to not only predict peaks for base calling but also to derive a statistic that estimates the probability of error (Ewing and Green 1998, Applied Biosystems 2002). This estimate of basecalling accuracy is expressed using the formula $QV = -10\log_{10}(Pe)$, where QV is the presented quality value (Q-score) and *Pe* is the probability of error. Given a QV of 10, the probability of a basecall being inaccurate is 1 in 10; given QV = 20, 1 in 100. Though differing from Sanger based calculations of *Pe*, Illumina® platforms still perform the log10-transformation of *Pe* to express base quality.

Quality scores can be used to filter and trim large datasets. The datasets that NGS platforms produce are large enough that the analyst can afford to remove sequences of poor quality, thus preserving only the best quality data and reducing the time it takes to

align files to the reference of interest. There are many different methods of filtering and trimming datasets, some are more suited for one experimental design as opposed to others. In particular, a large portion Illumina® sequence reads suffer from a noticeable drop in Q-scores as sequencing progresses. After paired-end turn around occurs and the second sequencing primer hybridizes to the reverse strands, the quality scores of the new sequencing strand start with high confidence and gradually drops in average Q-score as sequencing progresses. Therefore, in Illumina® datasets, the data of poorest quality in both forward and reverse sequence reads, is generally located on the 3'-end of the reads.

Initially, a bioinformatics pipeline was developed for single-read data produced by the Illumina® $GAII_x$ using Galaxy™, a free to use cloud-computing bioinformatics service that allows its users to build customizable analysis pipelines for data processing (Giardine et al. 2005, available at www.galaxyproject.org). This pipeline filtered datasets in various ways, visualized the effects of various filtering regimes on overall sequence quality, aligned filtered sequence data against the rCRS, generated data pileups and filtered those pileups by Q-score and coverage to perform variant analysis. While the results for this initial mixture study were promising (data not shown), demonstrating that minor variants could be detected in mixtures of human mtDNA down to the 5, 2, 1, and 0.5% levels of resolution, the generated data was derived from a sequencing run that experienced numerous errors and the data was therefore, abandoned. It did however prove, in principle, that the sequence data obtained from Illumina® platforms could be treated in a similar fashion to produce accurate variant calls. The initial bioinformatics pipeline was then modified to accommodate the effects of paired-end sequencing for data produced by the Illumina MiSeq™ (see Figure 17).

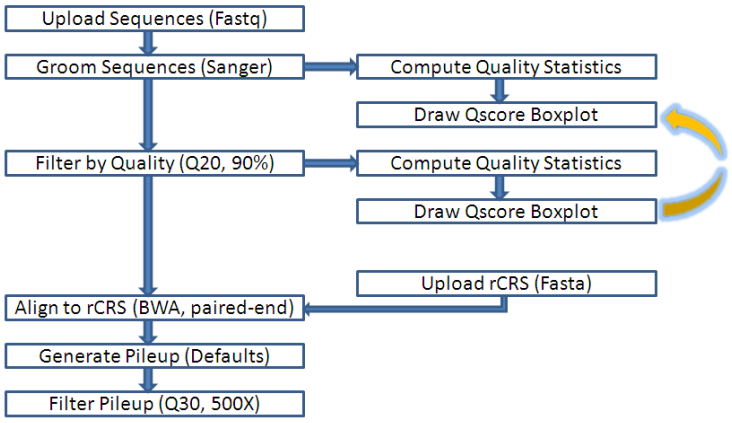

Figure 17 – An abridged version of the bioinformatic pipeline used to assess minor variants in the Illumina® MiSeq™ datasets. Alignment with BWA paired-end aligns the different sets of read data—forward and reverse—into a single alignment file.

Associated forward and reverse .fastq files are uploaded to the Galaxy cloud and groomed so that the data to could be filtered. Grooming converts the variety of *phred* scoring formats that exist among competing sequencers into a common quality scoring language, allowing the service to filter datasets from a variety of platforms (Blankenberg et al. 2010). Grooming MiSeq™ .fastq files converts the files to the 'fastqsanger' quality scoring format in which Q-scores have a range of Q0 to Q40. Both forward and reverse sequences reads were then filtered using the 'Filter by quality' function of the FASTX-toolkit in order to remove poor quality reads (Gordon & Hannon n.d.). This method of filtering retains reads which have ≥Q20 over 90% of the base composition. The rCRS was then uploaded to the workspace in a .fasta format (no associated Qscores) and used as a reference file for alignment of filtered reads. Filtered reads were then aligned to the rCRS using the default alignment parameters of the Burrows-Wheeler alignment (BWA) tool for paired-end data. The BWA alignment tool was selected as opposed to competing alignment algorithms because experimentally, it had been demonstrated that BWA can align millions of sequence reads with a higher percentage of confidently mapped reads (reads that have a lower probability of being mapped incorrectly) and lower error rates

(when confidently mapped reads are inappropriately aligned) than its competitors (Li & Durbin 2009). BWA alignment produces a file in SAM (Sequence alignment map) format. Following alignment, tabular datasets can be generated from SAM files using the 'Generate Pileup' function of SAMtools (Li et al. 2009). These data pileups present a subset of the base positions within the reference genome where sequence reads have aligned, showing the number of reads that have aligned over a particular position of the reference (coverage) and the overall base composition with respect to aligned reads. Data pileups could then be filtered by Q-score and coverage using the 'Filter Pileup' command of SAMtools in order to increase the confidence of variant calling (Li et al. 2009). Data outputs are then copied to an Excel Spreadsheet where rudimentary statistics are applied to each position. An example of data outputs are shown in Table 1.

CHROM	POS	REF	T#oRs	A CALLS	C CALLS	G CALLS	T CALLS	QARs	T#oDs	% MAJ	% MIN
rCRS	16129	G	6612	6424	0	35	0	6459	6424	99.46	0.54
rCRS	263	A	2841	2	0	2649	0	2651	2649	99.92	0.08
rCRS	73	A	2695	11	0	2417	0	2428	2417	99.55	0.45
rCRS	152	T	6433	0	272	0	6078	6350	272	95.72	4.28
rCRS	310	T	1544	0	114	0	631	745	114	84.70	15.30
rCRS	311	C	1541	0	1323	0	57	1380	57	95.87	4.13
rCRS	16189	T	7988	3	53	0	6728	6784	56	99.17	0.78
rCRS	16380	C	1906	31	1730	0	4	1765	35	98.02	1.76
rCRS	16386	T	1598	0	1	31	1469	1501	32	97.87	2.07
rCRS	16379	C	1914	0	1802	0	31	1833	31	98.31	1.69
rCRS	16383	A	1698	1641	0	0	31	1672	31	98.15	1.85
rCRS	16387	A	1166	1092	0	31	0	1123	31	97.24	2.76
rCRS	316	G	1409	0	26	824	0	850	26	96.94	3.06
rCRS	292	T	1623	25	0	0	1528	1553	25	98.39	1.61
rCRS	195	T	5977	0	24	0	5830	5854	24	99.59	0.41
rCRS	204	T	5668	0	22	0	4953	4975	22	99.56	0.44
rCRS	16293	A	7870	5716	9	12	0	5737	21	99.63	0.21

Table 1 – Example data output for donor 005-CF40Buccal (unmixed) using the Galaxy™ pipeline described in Figure 17 and sorted in descending order by total number of deviants. From left to right, the columns are as follows: CHROM = Reference genome, POS = position within reference, REF = reference call at respective position, T#oRs = total number of reads, A/C/G/T calls = individual bases called at that position, QARs = quality adjusted reads (number of reads retained after terminal filtering regime), T#oDs = total number of deviants (differences from the reference), %MAJ = percentage of basecalls that are the most prevalent basecall, %MIN = percentage of basecalls that are the second most prevalent basecalls. Purines and pyrimidines are shaded in different colors.

Gold = expected variants for the donor, Yellow = unexpected minor variant (≥1%), Blue = errors in alignment associated with C-stretch regions, Gray = misaligned reads within the sequence window, Dark blue = an insertion site at the end of a 6 base stretch of adenosine residues, Red = nuclear pseudogene (NumtS) that was discovered (Bintz et al. 2012) and now, is expected to cooamplify with the HV1b primer set. Unshaded positions (195 and 204) are positions in which no variant is expected.

CHAPTER THREE: MATERIALS AND METHODS

Section 3.1: Sample collection

Blood, buccal, and hair samples were collected from twenty donors according to the Human Subjects Institutional Review Board policies of Western Carolina University following informed consent.

Section 3.2: Sanger sequencing of the human mitochondrial HV region from the provided blood samples of twenty donors

Templates intended for downstream Sanger sequencing were derived from blood extracts stored on FTA® cards and were extracted using the Whatman FTA® Protocol BD09. An 'FTA® disc' was obtained by punching a hole in the blood stained portions of the storage card using the 1.2mm diameter Harris micro puncher and the disc was stored in a UV-treated microcentrifuge tube for no longer than a week, until the samples could be readied for PCR. No post extraction quantification was performed. Non-template controls and reagent blanks were prepared alongside samples. Prepared FTA® discs were gently coerced into their respective wells on a 96-well plate using a fresh pipet tip, taking care to avoid static charge that might alter the discs' trajectory. Amplification of the mtDNA hypervariable was performed by including 1.2 mm discs in 20 µl of PCR master mix. The PCR master mix was constructed in the following volumes: 5.00 µl of Bovine Serum Albumin (1.6 µg/µl), 2.00 µl of the light strand primer (10 µM), 2.00 of the heavy strand primer (10 µM), 2.50 µl of 10X LA PCR Buffer II (Mg^{2+} included), 4.00 of LA PCR dNTP mix (2.5 mM each), 0.25 µl of Takara LA Taq™ DNA polymerase

HV1a –	A1: (L 15997)	5’-CAC CAT TAG CAC CCA AAG CT-3’
	B2: (H 16237)	5’-GGC TTT GGA GTT GCA GTT GAT-3’
HV1b –	A2: (L 16159)	5’-TAC TTG ACC ACC TGT AGT AC-3’
	B1: (H 16391)	5’-GAG GAT GGT GGT CAA GGG AC-3’
HV2a –	C1: (L 048)	5’-CTC ACG GGA GCT CTC CAT GC-3’
	D2: (H 285)	5’-GGG GTT TGG TGG AAA TTT TTT G-3’
HV2b –	C2: (L 177)	5’-TTA TTT ATC GCA CCT ACG TTC AAT-3’
	D1: (H 409)	5’-CTG TTA AAA GTG CAT ACC GCC-3’

Table 2 – Light and heavy strand primers used to amplify the HV sub-regions of the human mtGenome.

(5U/µl). Light and heavy strand primer sequences are shown in Table 2. Thermal cycling was performed on Applied Biosystems GeneAmp® PCR System 9700 with the following thermal profile: 1 cycle of 95°C for 11 minutes, 32 cycles of 95°C for 10 seconds, followed by 60°C for 30 seconds, followed by 72°C for 30 seconds, 1 cycle of 15°C, and cycling ends with a 4°C hold. Amplicons are cleaned by adding 2 µl of ExoSAP to every 5 µl of amplified DNA product. Cleaned amplicons are quantified on the Agilent 2100 Bionanalyzer using P1000 reagent kit. Derived concentrations are used to calculate the dilution needed to obtain 20 ng of amplified product in a 7 µl volume and amplified products were diluted accordingly. Cycle sequencing was performed using half reactions of the BigDye v1.1 kit in the following stoichiometric volumes: 4.75 µl of diluted BigDye v1.1 sequencing mix, 1.75 µl of sequencing primer (see Table 5), and 3.50 µl of template DNA (approximately 10 ng). Dilutions of BigDye v1.1 sequencing mix are performed according to the number of reactions being processed in batch. Thermal cycling was performed of the Applied Biosystems GeneAmp® PCR System 9700 with the following thermal profile: 1 cycle of 96°C for 1 minute, 25 cycles of 96°C for 15 seconds, followed by 50°C for 1 second, followed by 60°C for 1 minute, and cycling ends with a 4°C hold. Sequenced products were then cleaned using Centri-Sep™ 96-well spin plates and according to protocol provided by Princeton Separations.

Cleaned products are resuspended in 10 µl of Hi-Di Formamide and loaded onto the Applied Biosystems® 3130*xl* Genetic Analyzer for sequencing. Sequence analysis was performed using the fragment analysis on the ABI Data Collection Software and POP-6. ABI Prism® Sequencing Analysis Software v5.0 performed basecalling using the KB basecaller and visualization of chromatogram data aligned against the rCRS was performed using Sequencher v4.8. Nucleotide differences from the aligned reference were recorded for each donor. After acquiring Sanger data for twenty donors, partner pairs were selected based on the maximum number of SNP differences within each of the four HV sub-regions. Analysis of Sanger data indicated 001-CF30 & 005-CF40 and 003-54M & 015-AM35 to be good candidates for deep sequencing mixture experiments.

Section 3.3: Performing the mixture study on the Illumina® MiSeq™ using amplicons derived from extracted buccal DNA

Extraction was performed on the buccal swabs of specified donors using the QIAmp mini spin kit's *Buccal swab spin protocol* and according to the volumes specified for 'cotton or DACRON' swabs. Extracts were quantified using Applied Biosystems Quantifiler™ kit and the ABI 7500 Real Time PCR System according to vendor guidelines (Applied Biosystems 2006). Quantified extracts were diluted to 1 ng/µl concentrations of nucDNA to dilute down mtDNA input into PCR reactions. Amplification of the HV sub-regions was performed in 25 µl volumes using the indicated primer pairs (Table 2) and the Roche® FastStart family of PCR reagents in the following stoichiometric volumes: 10.00 µl of template DNA (1 ng/µl nucDNA), 1.00 µl of the light strand primer (10 µM), 1.00 µl of the heavy strand primer (10 µM), 2.50 µl of 10X

Roche PCR reaction Buffer (with 18mM $MgCl_2$), 0.50 µl Promega dNTP mix (10 mM each), and 0.25 µl of Roche FastStart Enzyme (5U/µl). Thermal cycling was performed on Applied Biosystems GeneAmp® PCR System 9700 with the following thermal profile: 1 cycle of 95°C for 2 minutes, 32 cycles of 95°C for 30 seconds, followed by 60°C for 30 seconds, followed by 72°C for 30 seconds, 1 cycle of 72°C for 7 minutes, and cycling ends with a 4°C hold. Amplicons were cleaned by adding ExoSAP in a ratio of 2 µl of ExoSAP to every 5 µl of amplified DNA product. Cleaned amplicons are quantified on the Agilent 2100 Bioanalyzer using the P1000 reagent kit and in quintuplicate to control for instrument variation. A quantified average was used to dilute down amplicons to a 0.2 ng/µl concentration in a 200 µl volume. Lower total volumes of dilutions were used where appropriate. From each amplicon dilution, 50 µl was then alloquated into a single micocentrifuge tube, representing the HV library of a single reference. Mixtures were then constructed between partners using these reference libraries at the 5, 2, 1 and 0.5% of minor contributor. Reciprocal mixtures between donors were also prepared. From these libraries, 1 ng of input DNA was entered into Nextera® XT and given unique sample indices. A total of 23 samples were prepared using Nextera® XT representing: 4 unmixed references, 16 mixture experiments, 1 positive control (HL60), 1 non-template control, and 1 reagent blank. An instrument control genome, PhiX, was spiked into the pooled amplicon library (PAL) at 8 pM to compose 20% of the PAL's volume. This deviation from the outlined protocol occurred at step 6 during '*Library Pooling and MiSeq™ Sample Loading*.' This was performed out of a concern that Nextera® XT may not fragment small amplicons ($\leq$300 bp) at enough locations to properly balance the run and represents the only deviation protocol.

The PAL was then diluted to 1/25th of its concentration in accordance to the protocol's specifications and loaded onto the sequence cartridge. The sequence cartridge was then loaded onto the Illumina® MiSeq™ and the following addendums were made to the Sample Sheet: 'VariantCaller = somatic; VariantFrequencyFilterCutoff = 0.001.' The library was then sequenced in a 2x150, paired-end run. Resulting .fastq files were analyzed using the developed Galaxy™ bioinformatics pipeline.

Section 3.4: Performing the mixture study on the Illumina® MiSeq™ using amplicons derived from extracted blood DNA

For specified donors, DNA derived from blood samples stored on FTA® cards were extracted using the Whatman FTA Protocol BD09. No post extraction quantification was performed. Discs were transported to respective wells on at 96-well plate and amplification of the HV sub-regions was performed in 25 µl volumes using the indicated primer pairs (Table 2) and the Roche® FastStart family of PCR reagents in the stoichiometric volumes presented in the buccal mixture experiment. These amplicons were amplified, cleaned, quantified in quintuplicate, normalized, pooled, mixed, and taken through Nextera® XT as indicated in the previous section, creating a total of 23 prepared samples representing: 4 unmixed references, 16 mixture experiments, 1 positive control (HL60), 1 non-template control, and 1 reagent blank. PhiX, was spiked into the PAL at step 6 of '*Library Pooling and MiSeq™ Sample Loading*' and at 8 pM to compose 20% of the PAL's volume. The PAL was then diluted to 1/50th of its concentration, a deviation that is recommended by Illumina® if previous runs are observed to over cluster (which occurred). The diluted PAL was processed for

sequencing as described earlier and resultant .fastq files were analyzed using the Galaxy™ bioinformatics pipeline.

Section 3.5: Performing the tissue study on the Illumina® MiSeq™ using amplicons derived from hair DNA

Stereoscopic inspection of hair samples from specified donors observed the root-end of each hair. An inch from the root-end of the hair was removed and the next 2 cm of hair shaft was used for mtDNA extraction. Templates were derived from five different hair shafts from each donor and were extracted using the hybrid Qiagen digestion/ABI® PrepFiler™ purification methodology (Burnside et al. 2013). Quantification of hair extracts was performed using the mitochondrial DNA quantification assay described by Kavlick et al. (2011). Concentrations of hair DNA were diluted to 1750 copies/µl in order to input 17500 copies of mtDNA into each amplification reaction. Amplification of the HV sub-regions was performed in 25 µl volumes using the indicated primer pairs (Table 2) and the Roche® FastStart family of PCR reagents in the stoichiometric volumes presented in the buccal mixture experiment. Hair derived amplicons were amplified with the same thermal profile indicated in the previous sections with the exception of 36 cycles instead of 32. Amplicons derived from hair were purified, quantified, normalized, pooled, and taken through Nextera® XT as previously described, creating a total of 23 prepared samples representing 19 or 21 samples (depending on the number of hairs successfully amplified) representing: 2 buccal samples, 1 buccal non-template control, 1 buccal reagent blank, 1 buccal positive control (HL60), 8 or 10 hair samples, 2 hair non-template controls, 2 hair reagent blanks, and 2 hair positive controls

(HL60). PhiX was spiked in at 20% and the diluted PAL ($1/50^{th}$) was processed for sequencing as described earlier. Resultant .fastq files were analyzed using the Galaxy™ bioinformatics pipeline.

CHAPTER FOUR: RESULTS

Section 4.1: Cycle sequencing of the HV sub-regions twenty donors and assigning two-person mixtures

Using Sequencher v4.8, chromatogram data was aligned against the rCRS. Bases receiving poor base quality were trimmed from aligned data and descriptive SNPs were recorded. SNP positions were only recorded if the sequenced amplicon had at least double coverage (either from forward and reverse reads or overlapping coverage from the neighboring amplicon) over the SNP position. Amplicons whose sequence data were of poor quality were rerun on the Applied Biosystems® 3130*xl* Genetic Analyzer. Amplicons whose sequence data were uninterpretable were resequenced, purified, and run on the capillary electrophoresis instrument. Analysis of the chromatograms derived from the blood samples of twenty donors are displayed in Tables 6 and 7. An inspection

What Sequencher calls HV1 (16024-16365)
HV1A amplicon coverage (15978-16257)
HV1B amplicon coverage (16140-16410)

	rCRS	16069	16074	16093	16126	16129	16145	16172	16182	16183	16188.1	16189	16192	16193	16195	16221	16223	16224	16234	16239	16242	16248	16261	16270	16274	16291	16294	16304	16311	16319	16352	16356	16357	16359	16362	16390	16391
	rCRS	C	A	T	T	G	G	T	A	A	:	T	C	C	T	C	C	T	C	C	C	C	C	C	G	C	C	T	T	G	T	T	T	T	T	G	G
	001-CF30BLD	T		C	C																															A	
	002-CM32BLD		G			A	A										T																				A
	003-54MBLD	T			C									T	C	T					A									A			C				
!	004-23FBLD											C																				C			C		
	005-CF40BLD					A																															
*	006-C25MBLD																																	C			
	007-C21BLDF												T															C									
*	008-CM23BLD																																				
	009-CF30BLD																					T															
	010-CM50BLD																	C										C	C								
	011-AM24BLD					A		C																				C									
!*	012-MM28BLD					A					C	C														T											
	013-CM41BLD												T											T													
	014-CF31BLD																			T																	
	015-AM35BLD																T	C						T	A					A	C						
	016-CF24BLD				C																						T	C									
	017-**KNOWN**																																				
	018-CM26BLD																																				
!	019-UF24BLD					A			**C**	**C**		C											T														
	020-AF44BLD																T		T																C		

Table 3 – Cycle sequencing data for the descriptive SNPs found in HV1a and HV1b of the twenty donors. Underlined SNPs fall within a primer binding site. Bold SNPs are transversions. Donors earmarked with an "!" showed length variants in the HV1 C-stretch. Donors earmarked with an "*" showed length variants in the HV2 C-stretch.

		What Sequencher calls HV2 (73-340)																		
		HV2A amplicon coverage (29-306)																		
							HV2B amplicon coverage (154-429)													
	rCRS	73	140	146	150	152	185	195	199	204	207	228	249	250	263	295	309.1	309.2	315.1
	rCRS	A	C	T	C	T	G	T	T	T	G	G	A	T	A	C	:	:	:
	001-CF30BLD	G					A					A			G	T	C		C
	002-CM32BLD	G				C			C	C	A			C	G		C		C
	003-54MBLD	G			T	C									G	T			C
!	004-23FBLD		T												G				C
	005-CF40BLD	G													G		C		C
*	006-C25MBLD					C									G		C		C
	007-C21BLDF														G				C
*	008-CM23BLD								C						G		C	C	C
	009-CF30BLD	G													G		C		C
	010-CM50BLD	G							C						G				C
	011-AM24BLD	G											:		G		C		C
!*	012-MM28BLD														G		C		C
	013-CM41BLD	G													G				C
	014-CF31BLD			C											G				C
	015-AM35BLD	G								C					G		C		C
	016-CF24BLD	G				C		C							G				C
	017-**KNOWN**														G		C		C
	018-CM26BLD				T										G				C
!	019-UF24BLD	G													G		C		C
	020-AF44BLD	G						A							G		C		C

Table 4 – Cycle sequencing data for the descriptive SNPs found in HV2a and HV2b of the twenty donors. Underlined SNPs fall within a primer binding site. Bold SNPs are transversions. Donors earmarked with an "!" showed length variants in the HV1 C-stretch. Donors earmarked with an "*" showed length variants in the HV2 C-stretch.

of the SNP data revealed that donors 001-CF30 & 005-CF40 to be good candidates for deep-sequencing mixture experiments, as well as 003-54M & 015-AM30, based on the respective frequency of identifying SNPs that occurred in each HV sub-region.

	What Sequencher calls HV1 (16024-16365)															What Sequencher calls HV2 (73-340)										
	HV1A amplicon coverage (15978-16257)															HV2A amplicon coverage (29-306)										
		HV1B amplicon coverage (16140-16410)																	HV2B amplicon coverage (154-429)							
rCRS	15978	16069	16093	16126	16129	16193	16195	16221	16223	16224	16242	16270	16274	16319	16352	16357	16390	73	150	152	185	204	228	263	295	309.1	315.1
rCRS		C	T	T	G	C	T	C	C	T	C	C	G	G	T	T	G	A	C	T	G	T	G	A	C	:	:
001-CF30BLD		T	C	C													**A**	G			A		A	G	T	C	C
005-CF40BLD					A													G						G		C	C
003-54MBLD		T		C		T	C	T			**A**			A		C		G	T	C				G	T		C
015-AM35BLD									T	C		T	A	A	C			G				C		G		C	C

Table 5 – SNP differences across the four HV amplicons for donors 001-CF30, 005-CF40, 003-54M, and 015-AM35.

Section 4.2: Summary of the mixture study from buccal samples

Buccal swab extracts were quantified using the Quantifiler™ Kit according to vendor specifications (Applied Biosystems 2006). Concentrations of nucDNA were in

	Quantifiler (nucDNA ng/µl)		Bioanalyzer (mtDNA ng/µl)
001-CF30Buc	11.50	HV1a	15.69
		HV1b	11.97
		HV2a	2.67
		HV2b	6.03
003-54MBuc	2.77	HV1a	15.99
		HV1b	13.82
		HV2a	4.72
		HV2b	7.37
005-CF40Buc	6.76	HV1a	16.30
		HV1b	12.97
		HV2a	11.97
		HV2b	7.36
015-AM35Buc	1.80	HV1a	9.26
		HV1b	11.21
		HV2a	6.52
		HV2b	4.37

	Quantifiler (nucDNA ng/µl)		Bioanalyzer (mtDNA ng/µl)
NC-RB	0.00	HV1a	0
		HV1b	0
		HV2a	0
		HV2b	0
NC-NTC	n/a	HV1a	0
		HV1b	0
		HV2a	0
		HV2b	0
PC-HL60	n/a	HV1a	3.31
		HV1b	0.93
		HV2a	1.06
		HV2b	2.69

Table 6 – Post extraction quantification of buccal extracts using Quantifiler™ and averaged post-amplification quantification using the Agilent 2100 Bioanalyzer for amplicons derived from buccal tissue.

terms of ng/2µl. After quantification, extracts were diluted down to 1 ng of nucDNA per µl in order to dilute down the mitochondrial copy number of the extracts, which were expected to saturate the extract. This is common practice in forensic estimations of mtDNA inputs. After PCR, the sample and positive control amplicons were quantified in quintuplicate to reduce the effect of instrument variation on DNA inputs into Nextera® XT. The negative controls were quantified once. Results are shown below.

MiSeq™ run quality metrics and the number of clusters per index were determined using MSR v2.0. A summary of the run quality metrics and a total of the number of clusters that had usable data are shown in Table 7. Other run quality metrics (not shown) indicated that over clustering did occur and that cluster density achieved

	Clusters		Low %		High %
Raw	8,119 K	Phasing 1	0.186	PF	83.4
PF	6,768 K	Phasing 2	0.218	Align 1	76.1
Unaligned	110 K	PrePhasing 1	0.182	Align 2	74.7
Unindexed	1,441 K	PrePhasing 2	0.187	PE Resynth	96.2
Duplicate	0	MisMatch 1	1.306		
		MisMatch 2	1.343	Total:	5,326,863

Table 7 – Run quality metrics for the buccal mixture study. Raw = raw number of clusters recorded by the instrument, PF = number of clusters which passed the instruments chastity filter for confident base calling, Unaligned = number of clusters which pass filter but did not align to the reference, Unindexed = number of clusters which pass filter but have no indices, Phasing 1 & 2 = percentage of clusters in the forward (1) and reverse (2) sequence reads that fail to incorporate a dNTP, Prephasing 1 & 2 = percentage of clusters that incorporate more than one dNTP, MisMatch 1 & 2 = percentage of base mismatches to reference averaged over all cycles, Align 1 & 2 = percentage of clusters that align to the rCRS, PE Resynth = percentage of clusters that successfully recorded both forward and reverse sequence data.

approximately 1,100 K clusters/mm^2. The ideal maximum range for cluster density on the Illumina® MiSeq™ is 850 K clusters/mm^2. Over clustering notwithstanding, the run still achieved a high percentage of reads that passed filter (83.4%) with nominal rates of phasing (0.186%, 0.218%) and prephasing (0.182%, 0.187%). Phasing is the rate in which clusters fail to incorporate the next dNTP in the sequence, and thus, the derived sequence information lags behind by one base. Prephasing is the opposite and occurs when a cluster incorporates more than one dNTP. Future sequencing runs increased the dilution factor of the PAL during Nextera® XT to compensate for over clustering. Higher rates of base mismatches averaged over all cycles (1.306%, 1.343%) and the low average number of reads which align to the rCRS (76.1%, 74.7%) are intriguing. Paired-end resynthesis was successful and there were few duplicate clusters (clusters that share sequence basecalls and have the same start/ending positions). Raw .fastq outputs were obtained from the instrument and analyzed using the bioinformatics pipeline detailed in Figure 17. A summary of the run's analyzed data is presented in Appendix 1.

Section 4.3: Summary of the mixture study from blood samples

After PCR, amplicons derived from the blood of the four donors and positive control amplicons were quantified in quintuplicate. The negative controls were quantified once. Results are shown in Table 8. The dilution step during Nextera® XT was adjusted to comprise a 50 fold dilution of the PAL. Derived quality metrics indicated that the blood mixture study achieved a cluster density of approximately 1,126 K clusters/mm^2 which did not seem to greatly impact the percentage of reads that passed filter (84.4%, Table 9). After sequencing, raw .fastq outputs were obtained from the

		Bioanalyzer (mtDNA ng/µl)
001-CF30BLD	HV1a	24.05
	HV1b	21.65
	HV2a	16.53
	HV2b	7.33
003-54MBLD	HV1a	19.52
	HV1b	20.80
	HV2a	15.68
	HV2b	7.69
005-CF40BLD	HV1a	20.72
	HV1b	19.26
	HV2a	13.56
	HV2b	7.25
015-AM35BLD	HV1a	22.69
	HV1b	17.76
	HV2a	16.43
	HV2b	7.48

		Bioanalyzer (mtDNA ng/µl)
NC-RB	HV1a	0
	HV1b	0
	HV2a	0
	HV2b	0
NC-NTC	HV1a	0
	HV1b	0
	HV2a	0
	HV2b	0
PC-HL60	HV1a	15.63
	HV1b	14.94
	HV2a	11.61
	HV2b	6.48

Table 8 – Averaged post-amplification quantification for amplicons derived from blood tissue using the Agilent 2100 Bioanalyzer.

	Clusters		Low %		High %
Raw	8,480 K	Phasing 1	0.143	PF	84.4
PF	7,156 K	Phasing 2	0.214	Align 1	93.9
Unaligned	198 K	PrePhasing 1	0.179	Align 2	91.2
Unindexed	167 K	PrePhasing 2	0.193	PE Resynth	94.6
Duplicate	0	MisMatch 1	1.400		
		MisMatch 2	1.567	Total: 6,988,483	

Table 9 – Run quality metrics for the blood mixture study. See Table 7 for interpretation guidelines.

instrument and analyzed using the bioinformatics pipeline detailed in Figure 17. A summary of the run's analyzed data is presented in Appendix 2.

Section 4.4: Summary of sequencing results from hair shaft samples

Hair extracts that passed qPCR results were chosen for deep sequencing. Hair extracts that entered downstream PCR had to have at least 1,000 copies of mtDNA per µl, show no evidence of inhibition, and be at least 10 times greater in concentration than its corresponding reagent blank. qPCR of hair extracts consistently resulted in a standard curve that fell outside the recommended range (-3.38 $\pm$ 0.06) of the publication from which the assay was based. The slope of these standard curves ranged from -3.675 to -4.081 and may indicate a lower starting concentration of the secondary stock from which the standard dilution series was performed. Effectively, this would lead to an over-

		Bioanalyzer (mtDNA ng/µl)
001-CF30Buc	HV1a	16.25
	HV1b	3.86
	HV2a	2.37
	HV2b	19.25
003-54MBuc	HV1a	7.47
	HV1b	5.44
	HV2a	6.53
	HV2b	1.77
NC-RB_1 [x32]	HV1a	0
	HV1b	0
	HV2a	0
	HV2b	0
NC-H2O [x32]	HV1a	0
	HV1b	0
	HV2a	0
	HV2b	0
PC-HL60 [x32]	HV1a	8.92
	HV1b	2.84
	HV2a	5.79
	HV2b	2.17

		Bioanalyzer (mtDNA ng/µl)
001-CF30HairA	HV1a	11.67
	HV1b	11.55
	HV2a	6.52
	HV2b	3.04
001-CF30HairB	HV1a	8.19
	HV1b	9.99
	HV2a	6.53
	HV2b	3.10
001-CF30HairC	HV1a	7.63
	HV1b	9.56
	HV2a	7.68
	HV2b	4.22
001-CF30HairD	HV1a	11.08
	HV1b	11.64
	HV2a	9.4
	HV2b	3.19
001-CF30HairE	HV1a	9.54
	HV1b	12.1
	HV2a	7.51
	HV2b	3.16

		Bioanalyzer (mtDNA ng/µl)
003-54MHairA	HV1a	4.26
	HV1b	2.20
	HV2a	4.18
	HV2b	1.23
003-54MHairB	HV1a	12.60
	HV1b	9.49
	HV2a	9.11
	HV2b	4.01
003-54MHairC	HV1a	11.41
	HV1b	7.68
	HV2a	6.51
	HV2b	2.68
003-54MHairD	HV1a	7.69
	HV1b	12.54
	HV2a	6.03
	HV2b	2.15
003-54MHairE	HV1a	8.30
	HV1b	8.96
	HV2a	6.87
	HV2b	1.77

		Bioanalyzer (mtDNA ng/µl)
NC-RB_1_A-E [x36]	HV1a	0
	HV1b	0
	HV2a	0
	HV2b	0
NC-H2O_1 [x36]	HV1a	0
	HV1b	0
	HV2a	0
	HV2b	0
PC-HL60_1 [x36]	HV1a	11.14
	HV1b	8.16
	HV2a	10.83
	HV2b	5.41
NC-RB_3_A-E [x36]	HV1a	0
	HV1b	0
	HV2a	0
	HV2b	0
NC-H2O_3 [x36]	HV1a	0
	HV1b	0
	HV2a	0
	HV2b	0
PC-HL60_3 [x36]	HV1a	14.25
	HV1b	12.02
	HV2a	9.86
	HV2b	4.42

		Bioanalyzer (mtDNA ng/µl)
005-CF40Buc	HV1a	14.17
	HV1b	8.75
	HV2a	2.76
	HV2b	1.92
015-AM30Buc	HV1a	10.10
	HV1b	3.70
	HV2a	2.71
	HV2b	0.62
NC-RB_1,3, & 5 [x32]	HV1a	0
	HV1b	0
	HV2a	0
	HV2b	0
NC-RB_15 [x32]	HV1a	0
	HV1b	0
	HV2a	0
	HV2b	0
NC-H2O [x32]	HV1a	0
	HV1b	0
	HV2a	0
	HV2b	0
PC-HL60 [x32]	HV1a	8.92
	HV1b	2.84
	HV2a	5.79
	HV2b	2.17

		Bioanalyzer (mtDNA ng/µl)
005-CF40HairB	HV1a	8.88
	HV1b	7.24
	HV2a	9.47
	HV2b	2.80
005-CF40HairC	HV1a	4.40
	HV1b	1.11
	HV2a	7.80
	HV2b	2.99
005-CF40HairD	HV1a	7.60
	HV1b	11.50
	HV2a	11.87
	HV2b	3.60
005-CF40HairE	HV1a	9.52
	HV1b	10.84
	HV2a	10.82
	HV2b	3.05
NC-RB_5_A-E [x36]	HV1a	0
	HV1b	0
	HV2a	0
	HV2b	0
NC-H2O_1 [x36]	HV1a	0
	HV1b	0
	HV2a	0
	HV2b	0

		Bioanalyzer (mtDNA ng/µl)
015-AM30HairA	HV1a	4.23
	HV1b	4.69
	HV2a	8.10
	HV2b	1.88
015-AM30HairB	HV1a	3.59
	HV1b	1.78
	HV2a	7.71
	HV2b	2.16
015-AM30HairD	HV1a	6.12
	HV1b	5.28
	HV2a	8.99
	HV2b	2.77
015-AM30HairE	HV1a	5.33
	HV1b	1.96
	HV2a	7.30
	HV2b	2.87
NC-RB_15_A-E [x36]	HV1a	0
	HV1b	0
	HV2a	0
	HV2b	0
PC-HL60_1 [x36]	HV1a	7.74
	HV1b	7.59
	HV2a	10.04
	HV2b	3.24

Table 10 – Amplification quantification for amplicons derived from hair extracts using the Agilent 2100 Bioanalyzer. Quantification results are divided by the flow cells in which the amplicons would be sequenced. (Above) Amplicons sequenced during TS1. (Below) Amplicons sequenced during TS2. Hair extracts from donors 005-CF40 and 015-AM35 were amplified on the same PCR plate and therefore the same NC-H_2O and PC-HL60 controls.

estimation of quantified DNA product and therefore, the dilution of extracts prior to PCR underestimated total DNA input. This was deemed acceptable considering that the purpose of qPCR was to verify sufficient copy numbers within extracts and to screen for inhibition. After PCR, amplicons derived from hair extracts of the four donors, along with positive negative controls were quantified once (Tables 10). The deep sequencing

	Clusters
Raw	4,967 K
PF	4,557 K
Unaligned	148 K
Unindexed	139 K
Duplicate	0

	Low %
Phasing 1	0.146
Phasing 2	0.210
PrePhasing 1	0.167
PrePhasing 2	0.204
MisMatch 1	1.303
MisMatch 2	1.387

	High %
PF	91.7
Align 1	92.6
Align 2	91.8
PE Resynth	99.5

Hair: 3,555,781
Total: 4,417,859

	Clusters
Raw	10,872 K
PF	9,191 K
Unaligned	739 K
Unindexed	282 K
Duplicate	0

	Low %
Phasing 1	0.163
Phasing 2	0.212
PrePhasing 1	0.227
PrePhasing 2	0.246
MisMatch 1	0.958
MisMatch 2	1.232

	High %
PF	84.5
Align 1	87.9
Align 2	85.4
PE Resynth	99.8

Hair: 3,949,952
Total: 9,190,909

Table 11 – Run quality metrics for the hair tissue studies. (Left) Quality metrics for TS1. (Right) Quality metrics for TS2. See Table 7 for interpretation guidelines.

of hair samples derived from the four donors was divided across two different flow cells. Donor 001-CF30 and 003-54M and all corresponding samples were prepared on one flow cell (TS1), donors 005-CF40 and 015-AM35 were prepared on another (TS2). Both flow cells had additional mtDNAs that were included in the PAL but are not germane to this project. Presented quality metrics (Table 11) include these additional samples in their calculations. The 50 fold dilution of the PAL was maintained across both TS1 and TS2 and achieved 625 K clusters/mm^2 and 1,385 K clusters/mm^2 respectively. Raw .fastq outputs were obtained from the instrument and analyzed using the bioinformatics pipeline detailed in Figure 17. A summary of both runs' analyzed data are presented in Appendices 3 and 4.

CHAPTER FIVE: CONCLUSIONS

Section 5.1: Overview of the mixture experiments and minor variant detection

An analysis of the results derived from the mixture experiments conclusively demonstrate that minor variants can be detected in mixtures of human mtDNA at the stated levels of detection using the Illumina® MiSeq™ and the developed bioinformatics pipeline. Tables 12 and 13 are Galaxy™ datasets produced from the mixture study using blood as reference material and illustrate this observation in detail. The implications and caveats of this detection method are numerous. Firstly, this method of data sorting

CHROM	POS	REF	T#oRs	A CALLS	C CALLS	G CALLS	T CALLS	QARs	T#oDs	% MAJ	% MIN
rCRS	16126	T	8013	2	7578	0	6	7586	7580	99.89	0.08
rCRS	16093	T	7988	0	7568	0	49	7617	7568	99.36	0.64
rCRS	185	G	7950	7503	0	40	0	7543	7503	99.47	0.53
rCRS	16069	C	7999	0	3	0	7462	7465	7462	99.96	0.04
rCRS	228	G	4820	4604	0	18	0	4622	4604	99.61	0.39
rCRS	73	A	3496	2	0	3122	0	3124	3122	99.94	0.06
rCRS	263	A	1433	0	0	1212	0	1212	1212	100.00	0.00
rCRS	295	C	797	0	12	0	664	676	664	98.22	1.78
rCRS	16311	T	5859	0	32	0	5321	5353	32	99.40	0.60
rCRS	16355	C	2550	1	2211	0	29	2241	30	98.66	1.29
rCRS	16356	T	2514	0	30	0	2230	2260	30	98.67	1.33
rCRS	16218	C	6691	0	6136	0	26	6162	26	99.58	0.42
rCRS	204	T	6334	0	25	0	5033	5058	25	99.51	0.49
rCRS	16230	A	7997	7572	0	25	0	7597	25	99.67	0.33

CHROM	POS	REF	T#oRs	A CALLS	C CALLS	G CALLS	T CALLS	QARs	T#oDs	% MAJ	% MIN
rCRS	16129	G	7971	7834	0	3	2	7839	7836	99.94	0.04
rCRS	73	A	4475	3	0	4076	0	4079	4076	99.93	0.07
rCRS	263	A	1200	2	0	1091	0	1093	1091	99.82	0.18
rCRS	152	T	8012	0	63	0	7852	7915	63	99.20	0.80
rCRS	16356	T	2609	0	44	0	2342	2386	44	98.16	1.84
rCRS	16355	C	2647	0	2330	0	42	2372	42	98.23	1.77
rCRS	16230	A	7996	7607	0	40	0	7647	40	99.48	0.52
rCRS	16311	T	5892	0	37	0	5388	5425	37	99.32	0.68
rCRS	16218	C	6736	0	6282	0	33	6315	33	99.48	0.52
rCRS	16278	C	7996	0	7812	1	31	7844	32	99.59	0.40
rCRS	16249	T	7994	0	31	0	7609	7640	31	99.59	0.41
rCRS	16189	T	7978	9	17	0	6682	6708	26	99.61	0.25
rCRS	16274	G	7999	23	1	7524	1	7549	25	99.67	0.30
rCRS	16293	A	7909	6073	22	3	0	6098	25	99.59	0.36

Table 12 – The first 15 entries of data outputs for donors 001-CF30BLD (A) and 005-CF40BLD (I) using the Galaxy™ pipeline described in Figure 17 and sorted in descending order by total number of deviants. Expected variant 16390A was not recovered in donor 001-CF30BLD. Datasets have not been filtered of NumtS variants (red).

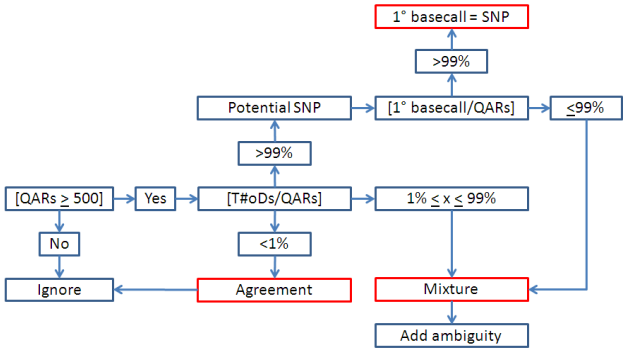

Figure 18 – A potential method of performing SNP and mixture detection in Galaxy™ outputs. The minimum post filtered coverage (500) for analysis and the interpretation threshold for DNA mixtures (1%) seen here are arbitrarily selected values.

implies that sequence interpretation can be performed using a series of computational 'If/Then' logic statements, like the one shown in Figure 18. These computations could calculate proportions of deviant basecalls, or basecalls that are in disagreement with the reference, with respect to the position in which they occur in post filtered sequence reads (QARs). An analysis of these proportions may yield one of three interpretations: (1) the position in question either contains a SNP, (2) the position in question falls below the threshold of mixture interpretation and therefore, agrees with the reference, or (3) the position in question is falls above the threshold of mixture interpretation and therefore, should be considered a site of base mixture. This method of data analysis would however, depend on two user defined variables: a minimum number of post-filtered sequence reads and a yet-to-be described interpretational threshold for interpreting mixed positions. In lieu of justified values, Galaxy™ outputs of MiSeq™ data were instead, visually characterized against the Sanger data presented in Table 5 and a color scheme was devised to demonstrate the recovery of expected variants and various NGS-related artifacts. An example of a visually characterized dataset is shown in Table 13.

Sites of expected variation (EV) based on the prepared mixtures are recovered by sorting datasets by %MIN in descending order. In Table 13, Gold represents an EV of

CHROM	POS	QARs	T#oDs	% MAJ	% MIN
rCRS	295	795	732	92.08	7.92
rCRS	228	3608	3380	93.68	6.32
rCRS	185	7491	7065	94.31	5.69
rCRS	16093	7676	7308	95.13	4.79
rCRS	16129	7605	317	95.83	4.17
rCRS	16126	7641	7327	95.88	4.11
rCRS	16069	7632	7335	96.11	3.89
rCRS	16355	950	17	98.21	1.68
rCRS	16356	939	14	98.51	1.49
rCRS	302	491	7	98.57	1.43
rCRS	310	372	5	98.66	1.34
rCRS	237	2522	19	99.25	0.75
rCRS	16020	4039	18	99.55	0.45
rCRS	40	469	2	99.57	0.43

CHROM	POS	QARs	T#oDs	% MAJ	% MIN
rCRS	568	518	35	93.24	6.76
rCRS	572	514	480	93.39	6.61
rCRS	573	519	33	93.64	6.36
rCRS	558	518	486	93.82	6.18
rCRS	295	1178	1150	97.62	2.38
rCRS	228	2452	2395	97.68	2.32
rCRS	185	7441	7307	98.20	1.80
rCRS	16093	7736	7624	98.55	1.45
rCRS	302	756	10	98.68	1.32
rCRS	16129	7584	76	99.00	1.00
rCRS	16126	7668	7598	99.05	0.91
rCRS	16069	7779	7711	99.13	0.87
rCRS	310	629	5	99.21	0.79
rCRS	16337	1362	7	99.49	0.51

CHROM	POS	QARs	T#oDs	% MAJ	% MIN
rCRS	295	988	938	94.94	5.06
rCRS	228	3296	3149	95.54	4.46
rCRS	185	7568	7307	96.55	3.45
rCRS	16093	7747	7535	97.26	2.74
rCRS	16356	420	11	97.38	2.62
rCRS	16355	424	10	97.64	2.36
rCRS	16129	7625	164	97.85	2.15
rCRS	16126	7683	7520	97.87	2.12
rCRS	16069	7733	7576	97.96	2.03
rCRS	310	542	11	97.97	2.03
rCRS	302	644	8	98.76	1.24
rCRS	237	2357	20	99.15	0.81
rCRS	16354	680	5	99.26	0.59
rCRS	16027	5331	27	99.49	0.51

CHROM	POS	QARs	T#oDs	% MAJ	% MIN
rCRS	302	471	10	97.88	2.12
rCRS	295	724	712	98.34	1.66
rCRS	310	362	5	98.62	1.38
rCRS	16093	7702	7607	98.77	1.23
rCRS	185	7525	7437	98.83	1.17
rCRS	228	4055	4012	98.94	1.06
rCRS	16356	1497	15	99.00	1.00
rCRS	16355	1458	13	99.11	0.89
rCRS	16366	759	6	99.21	0.79
rCRS	16129	7579	49	99.35	0.65
rCRS	237	3231	19	99.41	0.59
rCRS	16368	703	4	99.43	0.57
rCRS	16218	7055	39	99.45	0.54
rCRS	16069	7671	7630	99.47	0.53

Table 13 – The first 15 entries of data outputs for donors 001-CF30BLD and 005-CF40BLD in the blood mixture experiments, sorted by %MIN in descending order. (Top left) Donor 005-CF40BLD constitutes 5% of the library. (Bottom left) Donor 005-CF40BLD constitutes 2% of the library. (Top right) Donor 005-CF40BLD constitutes 1% of the library. (Bottom right) Donor 005-CF40BLD constitutes 0.5% of the library. Expected variant 16390A was not recovered in these samples.

the major contributor; green, an EV of the minor contributor. While these experiments show that variant positions are detectable at the four levels of resolution, sorting by %MIN also recovers several sites of unexpected variation (UV). Positions which showed either a high degree of variation within an unmixed donor or were consistently present in mixtures of a specific major contributor were highlighted in yellow, as these sites may reflect true biological variation within the donor and may bias averages of %MIN for UVs. Blue positions are areas within or are in close proximity to homopolymeric stretches of cytosine residues. In this instance, both donors have the same C-stretch polymorphisms (309.1C, 315.1C) and show no evidence of length heteroplasmy in

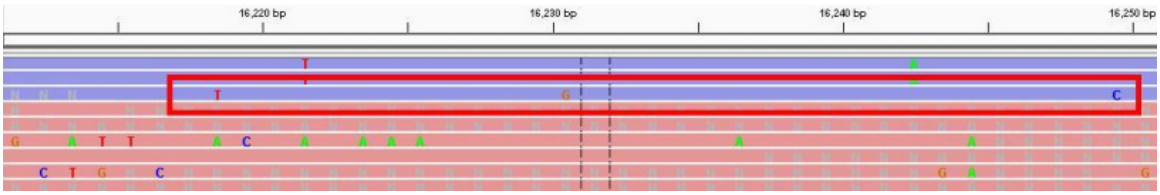

Figure 19 – Visualization of MiSeq™ derived SAM file using the Broad Integrative Genomics Viewer (IGV). Only variant sequences are reported. The highlighted sequence shows NumtS variants 16218T, 16230G, and 16249C.

Sanger data. While the variants that appear in Galaxy™ datasets may be real length variants, the frequencies of %MIN that occur at 302 and 310 reflect an inability of the analysis pipeline to align insertion variants in the format adopted for use in forensic science. Red positions are sites associated with nuclear pseudogenes (NumtS) and cannot be ruled out as a possible cause of variation. NumtS are ancient mitochondrial insertions into the nuclear genome and have been highly conserved through time (Zischler et al. 1995). In separate mixture experiments performed on the Roche® GS Junior, a consistent set of variants occurred at specific positions within aligned reads. A subsequent BLASTn search of the deviant sequence revealed them to be of NumtS origin and was later confirmed by Sanger sequencing (Bintz et al. 2013). Since the NumtS contains the primer binding sites for primers A2 and B1, NumtS are expected to coamplify with the HV1b primer set and have also been identified in the mixture experiments performed on the MiSeq™ (Figure 19). Dark red positions reflect BWA alignment errors, two examples of which are demonstrated in Figure 20. The positions presented in Table 13 (558, 568, 572, 573) are associated with a read which has aligned to the reference outside the PCR primer binding regions and are not likely to have been sequenced given the amplicon design of the project. Other alignment errors have been observed to occur within regions of properly aligned data and have artificially inflated calculations of %MIN at affected positions. A BLASTn search (Table 14) of the five

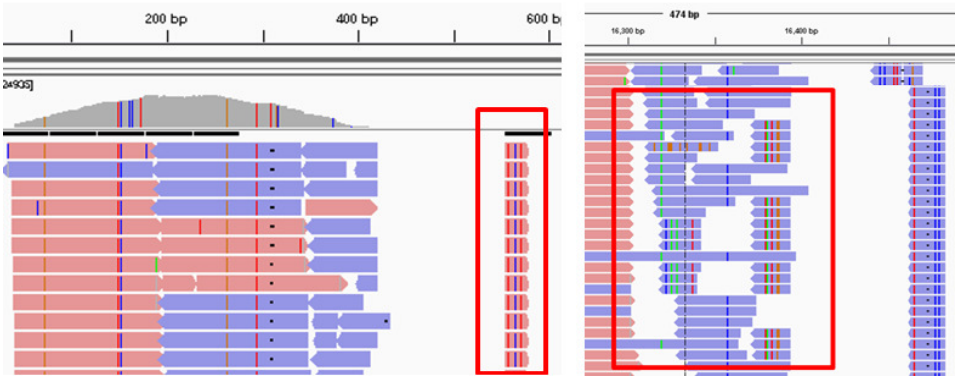

Figure 20 – Visualization of alignment errors due to BWA 'soft-clipping' using IGV. Within the red boxes are short reads that have numerous mutations. These are the reads that have aligned in err. (Left) Soft-clipped reads fall outside the sequence window and are less problematic. (Right) Soft-clipped reads are within the sequence window, skewing calculations of %MIN.

	Deviant Sequence	Misaligns to	BLASTn Results
	AACTCCAAAGCCACCCCTCAC	555-575	16246-16266
	GATGTGGATTGTTTTTTTATGTA	9942-9964	16157-16179 (R)
*	TGAACCCCCTCAG--ATAGGAGTCCC	13030-13055	16372-16395
	CTGG-AGCCGGAGCACCCTATGTCGAGTAT	13705-13732	98-127
	AATACTTGGGTGGTACCCAAA	16463-16482	16044-16064 (R)

Table 14 – Misaligned sequences in 001-CF30BLD. Red bases are sites interpreted as mutations. Dashes are interpreted as base deletions. Underlined bases are interpreted as base insertions. "R" are reverse complementary to the indicated sequence. The sequence demarcated with an asterisk is unique to donor 001-CF30BLD, likely due to the fact that this individual as a personal SNP at position 16390.

most prevalent misaligned sequences within 001-CF30BLD using the deviant sequence and ignoring deletion sites, revealed these sequences to be identical to other sites within the human mtGenome, specifically, at sites expected to have been sequenced given the amplicon design of this project. Discussions with bioinformaticists at Illumina® indicated 'soft-clipping' had occurred, a secondary alignment process inherent to BWA's paired-end alignment designed to recover poorly aligned reads, and was responsible for these errors. The cause of soft-clipping is discussed in section 5.2.

Sites with UVs that could be linked to different NumtS or soft-clipping were filtered from mixture datasets to produce Appendices 1 and 2. C-stretch variants were retained for the purposes of calculating the average coverage per base in both pre- and

	Recovery of Expected Variants (EV)	Positions showing unexpectedly high minor variation (≥1%) and associated frequencies
001-CF30Buc	**8 of 9**	**185 (1.45), 295 (2.77), 16093 (6.50)**
005-CF40Buc	3 of 3	152 (4.28)
003-54MBuc	13 of 13	**16242 (9.88)**
015-AM35Buc	9 of 9	none

	Recovery of Expected Variants (EV)	Positions showing unexpectedly high minor variation (≥1%) and associated frequencies
001-CF30BLD	**8 of 9**	**295 (1.78)**
005-CF40BLD	3 of 3	none
003-54MBLD	**12 of 13**	**16242 (1.84)**
015-AM35BLD	9 of 9	**16352 (4.13)**

Buccal	Recovery of Expected Variants (EV)	Suspected outliers and associated frequencies	Avg. % Minor	Std. Dev.
5.0% Donor 5	**7 of 8**	185 (5.59), 295 (8.64), 16093 (10.54)	3.99	0.62
2.0% Donor 5	**7 of 8**	185 (2.53), 295 (2.18), 16093 (7.80)	1.64	0.34
1.0% Donor 5	**7 of 8**	185 (2.23), 295 (2.52), 16093 (7.61)	1.10	0.31
0.5% Donor 5	**7 of 8**	185 (1.68), 295 (0.57), 16093 (7.03)	0.48	0.15
5.0% Donor 1	**7 of 8**	--	4.83	0.78
2.0% Donor 1	**7 of 8**	--	1.57	0.30
1.0% Donor 1	**7 of 8**	--	0.80	0.25
0.5% Donor 1	**7 of 8**	--	0.68	0.25
5.0% Donor 15	16 of 16	16242 (10.23)	4.55	0.85
2.0% Donor 15	16 of 16	16242 (2.87)	1.60	0.45
1.0% Donor 15	**14 of 16**	16242 (2.29)	1.03	0.30
0.5% Donor 15	**15 of 16**	16242 (1.55)	0.59	0.21
5.0% Donor 3	16 of 16	--	5.91	1.08
2.0% Donor 3	16 of 16	--	2.57	0.68
1.0% Donor 3	16 of 16	--	1.37	0.82
0.5% Donor 3	16 of 16	--	1.12	0.65

Blood	Recovery of Expected Variants (EV)	Suspected outliers and associated frequencies	Avg. % Minor	Std. Dev.
5.0% Donor 5	**7 of 8**	295 (7.92)	4.83	0.98
2.0% Donor 5	**7 of 8**	295 (5.06)	2.82	1.22
1.0% Donor 5	**7 of 8**	295 (2.38)	1.39	0.65
0.5% Donor 5	**7 of 8**	295 (1.66)	0.86	0.43
5.0% Donor 1	**7 of 8**	--	4.98	1.07
2.0% Donor 1	**7 of 8**	--	2.01	0.30
1.0% Donor 1	**7 of 8**	--	0.91	0.19
0.5% Donor 1	**6 of 8**	--	0.62	0.10
5.0% Donor 15	16 of 16	16242 (11.14)	5.15	0.86
2.0% Donor 15	**15 of 16**	16242 (3.78)	2.34	0.54
1.0% Donor 15	**14 of 16**	16242 (2.97)	1.29	0.52
0.5% Donor 15	16 of 16	16242 (2.12)	0.92	0.61
5.0% Donor 3	**n/a**	**n/a**	**n/a**	**n/a**
2.0% Donor 3	16 of 16	16352 (6.46)	2.82	0.81
1.0% Donor 3	16 of 16	16352 (9.19)	1.35	0.25
0.5% Donor 3	**14 of 16**	--	0.69	0.15

Table 15 – Calculations of average %MIN and standard deviation for expected variants (EVs). EVs that showed unusually high minor variation in unmixed donors were not used to calculate the average %MIN and the associated standard deviation in mixtures where that donor was the major contributor. These values of %MIN differ from those presented in Appendices 1 and 2 which include all sites of expected variation.

post-filtered reads as well as the length of sequence window. The sequence window is the number of positions which retained aligned data. C-stretch variants were not however, considered as sites of expected variation due to the inability of data pileups to format insertion sites correctly. Analysis of filtered datasets showed that a majority of EVs could be recovered in mixed and unmixed samples with few exceptions. Calculations of %MIN at sites of expected variation, using only the EVs that did not show unexpectedly high levels of variation in unmixed donors (Table 15), clustered these values around the expected levels of detection. An unweighted average (ignoring coverage) of these values corroborated this. Calculated standard deviations from these averages show that a relatively wide range of variation (~0.5%) among EVs can be achieved. These calculations however, should be taken with a degree of skepticism given the limitations of the bioinformatics pipeline, which is discussed in section 5.2.

Additionally, when data tables for lower levels of resolution (0.5-1%) are sorted by %MIN, the number of UVs recovered, or positions that show higher values of %MIN than sites of expected variation, increases. The recovery of UVs at lower levels of resolution are depicted in Appendices 1 and 2.

One EV in particular, 16390A in donor 001-CF30BLD, consistently dropped out in each associated mixture. This occurred in samples that achieved the largest number of positions sequenced (722 bp), the highest raw clusters PF (441 K), and in samples that achieved the highest average coverage per base (5133.60) following bioinformatic filtering. This indicates that these calculations may not be the most relevant data in the recovery of all EVs. Indeed, the proximity of this variant to the distal end of the HV1b amplicon (primer B1 primer binds to positions 16391-16410) is concerning since a technical note provided by the vender warns of an appreciable drop in coverage 50 bases in from each end of the amplicon (Illumina® 2012b). This drop in coverage is attributed to the inability of the transposons to correctly incorporate a second PCR primer site after fragmenting an amplicon near its distal end and is demonstrated in Figure 21 for greater clarity. Should fragments containing only 'A' or 'B' primer binding sites make it into the PAL, the fragments will not be able to perform bridge amplification. Increasing the size of the amplicons to be sequenced may recover these distal variants, but encumbers the successful amplification of mtDNA from the most challenging sample types—highly fragmented DNAs are amplified better with smaller designs. A more effective solution may be to incorporate the primer binding sequences used during Nextera® XT's short-interval PCR onto the 5'-ends of the HV primer sets. The result is illustrated in Figure 22. Of consequence, SPRI bead clean-up will recover shorter fragment sizes that are

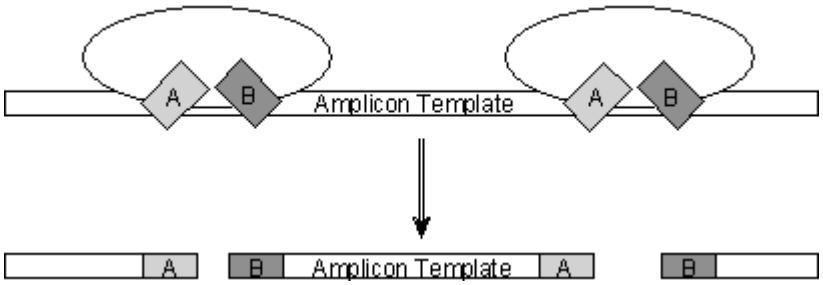

Figure 21 – Illustration of how Nextera® XT prepares template molecules for sequencing (Illumina® 2012b). The 'tagmentation' process results in fragments that either have the 'A' primer site, the 'B' primer site, or both 'A' and 'B' primer sites.

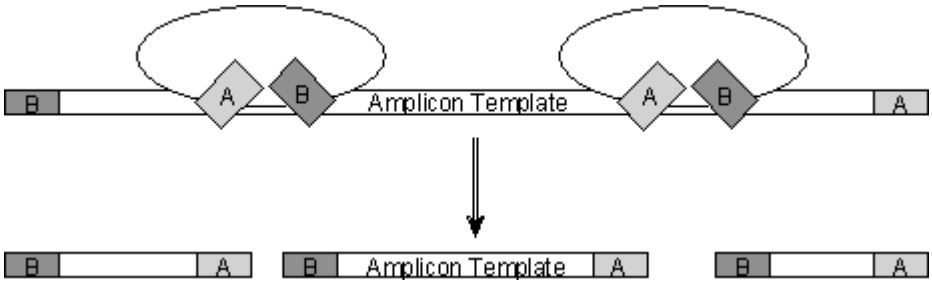

Figure 22 – The same illustration but with templates prepared using modified primers that incorporate the Nextera® XT PCR primer binding regions (Illumina® 2012b). Every product of the 'tagmentation' process should have both 'A' and 'B' primer sites.

capable of downstream bridge amplification and will likely compete for space on the flow cell. This may potentially decrease the overall yield of a sequencing run using the current method of sample preparation.

Section 5.2: Limitations of the Galaxy™ pipeline

Many bioinformatic algorithms are coded in Linux command prompt and many software packages (NextGENe®, CLC Bio) are marketed for expressed purpose of simplifying bioinformatics through tool-based GUIs. However, without a complete understanding of the Linux command prompt and the effects of toggling tool parameters, the user is limited by the capabilities of preselected features. In these experiments, BWA paired-end alignment was utilized for three reasons: (1) it had a published error rate, (2) it was one of two alignment tools Galaxy™ offered for Illumina® data, and (3) the service itself was free.

While Galaxy™ does allow the user to modify many parameters of BWA paired-end alignment, it does not currently allow the user to toggle-off 'soft-clipping.' When aligning paired-end data, BWA will perform Smith-Waterman alignment on unaligned sequences in an attempt to recover reads that have not mapped confidently by aligning the sequence to suboptimal hits (Li & Durbin 2009). This will occasionally clip and discard large portions of sequence reads, from anywhere between 100-130 bases, if the retained sequence has a higher mapping quality than the complete string. This may explain why when soft-clipping is suspected, it generally occurs in both a smaller proportion of reads as well as in a single direction—mapping quality for the one mate pair is calculated independently of the other, allowing one to map perfectly while the other does not. Of particular concern is the occurrence of unique soft-clips. The sequence demarcated by the asterisk in Table 14 is a misaligned sequence that does not occur in any other donor and should have aligned to 16372-16395. This misaligned sequence contains the 16390A EV of donor 001-CF30 and is alarming considering that a single SNP difference lowered the mapping quality enough to cause the aligner to misplace associated reads to a sub-optimal hit. Combined with low coverage, this may explain why the 16390A EV is not recovered in every experiment containing this donor. Misalignment due to poor mapping quality can be particularly problematic, especially for donors like 002-CM32. While samples from this donor were not deep sequenced in these experiments, this donor has three SNPs that are less than nine bases apart (199C, 204C, 207A). It stands to reason that the more SNPs a donor has within a sequence read, the greater the likelihood it will map erroneously using this method of analysis.

Soft-clipping is a systemic error and while it easy to visually locate misaligned sequences in small amplicon projects, projects that intend to sequence the entirety of the mtGenome will have difficulty isolating soft-clip induced noise if it does not disable this feature. Again, for the Linux savvy, this feature can be disabled in BWA command prompt by toggling "- s" so that it does not perform Smith-Waterman alignment on sub-optimal hits. This will make no attempt to recover unconfidently mapped reads. Alternatively, one could search the CIGAR strings of produced SAM files and filter all aligned reads containing "S" (S = soft-clip) and then, reupload the SAM file for further analysis. This alternative does represent a considerable time and technological investment, as one will likely need a computer capable of opening an Excel file with, on average, 70,000+ lines of text (each line representing an individual aligned read). This was not performed on produced datasets and instead, positions that were known to succumb to misaligned data were visually filtered from further analyses.

In Galaxy™, the 'Generate pileup' function of SAMtools has an approximate computational 8,000 line read cap over aligned positions and introduces bias into relevant calculations, such as the average coverage per base calculation shown in all attached appendices. This computational cap was set by the service provider to keep positions that have a phenomenal depth of coverage from chewing up too much bandwidth. Additionally, comparisons of multiple pileups generated from a single SAM file (data not shown) revealed that selected reads appear to be non-randomly selected. The tool may likely be selecting the first 8,000 reads that have aligned over a particular position and is an additional source of bias, but is not nearly as limiting as the cap itself.

Section 5.3: Reagent blank and negative control contamination

Reagent blanks (RBs) and PCR negative controls in experiments involving the use of buccal reference material produced a significant amount of NGS data (Appendices 1, 3, and 4). This is alarming considering that at no point during the sample preparation was contamination suspected. The Quantifiler™ qPCR assay used to quantify the nucDNA concentration of buccal extracts returned a negative result for the RBs that were prepared alongside reference material from donors 001-CF30, 003-54M, and 005-CF40. Unlike samples extracts, these RBs remained undiluted after post extraction quantification of nucDNA content. On two different dates, 10 µl of the prepared RB (RB-Buc-1, 3, 5) was input into the four respective PCR reactions, targeting the human CR—one reaction performed on one day was for the buccal mixture experiment, the other reaction was for the buccal reference in the tissue study experiment. Aside from the date in which these reactions were performed, nothing differed between PCR reactions. Post amplification quantification using a P1000 kit on the Agilent 2100 Bioanalyzer revealed that sample controls achieved a negative result (Table 6). It was not deemed necessary to quantify negative controls more than once. When samples were normalized to 0.2ng/µl concentrations and input into Nextera® XT, 5 µl of the undiluted RB was input into its respective reaction tube. When sample preparation was completed, pooled libraries were deep sequenced on the Illumina® MiSeq™. An analysis of the unexpected variants that appeared within the RB-Buc-1, 3, 5, linked several SNPs to donors that were prepared during a batch extraction. Specifically, these SNPs could be associated with references 002-CM32, 006-CM25, 019-UF24, and 020-AF44. At no point during the mixture experiments did reference material from these donors share hood-space with the prepared

reagent blank except during extraction, when references 001-006 and 017-020 were prepared in batch. Therefore, contamination could only have occurred during extraction and failed to achieve a positive result at multiple downstream quality control checks. In one instance, an RB that passed on the Agilent 2100 P1000 kit yielded upwards of 273 K clusters (Appendix 4).

At no point during the course of these experiments was the RB diluted at a proportion equivalent to what the samples experienced. Buccal references were diluted at two different intervals. Specifically, these samples were diluted after post-extraction quantification and after post-amplification quantification, resulting in ~5 and ~100 fold dilution factors, respectively. If reference material was contaminated at extraction, this contamination would likely represent a small proportion of that reference library and would also be diluted at the same proportions the library experienced. This may explain why SNPs observed in RB-Buc-1, 3, 5 cannot be identified in any of the NGS datasets that used buccal reference material (Appendices 1, 3, 4). Since the RBs remained undiluted, the samples used to produce these datasets did not experience the same conditions in which reference material was treated. Therefore, conclusions drawn from these RBs may not accurately portray the degree in which contamination could have occurred. Interestingly, and perhaps one of the greatest strengths of deep-sequencers, is that if contamination is suspected in NGS data, the user has the ability to look deep within sequence reads (provided the computational cap of data pileups has been removed) and to discern the degree in which the contamination is present.

Given the highly sensitive nature of mtDNA, contamination of the RBs is not entirely unexpected and interpretational guidelines for casework compensate for this (RB copy number cannot exceed 10% of the sample concentration following qPCR). However, mtDNA has not been previously scrutinized at this level of resolution. Indeed, the Agilent 2100 Bioanalyzer P1000 kit has a quantitative range of 0.1 ng/µl and normalizing experiment samples to 0.2 ng/µl prior to Nextera® XT dilutes sample inputs to concentrations twice the instrument's limit of detection (LOD). This may explain why RB-Buc-1, 3, 5 achieves a raw cluster count that is comparable to experimental samples. Agilent does produce high sensitivity kits (LOD = 5-500 pg/µl) which may warrant further investigation. However, the lack of contamination in RBs derived from blood and hair extracts points to the efficacy of the buccal extraction procedure. When NGS data for RB-Buc-1, 3, 5 is compared against RBs prepared using alternative extraction procedures (FTA® cards for blood, hybrid Qiagen digestion/Prepfiler™ BTA purification for hair), these alternative extraction procedures either yield no reads or uninformative reads when taken through the bioinformatics pipeline (Appendices 2, 3, and 4). Perhaps Qiagen's b*uccal swab spin protocol* for samples intended for deep-sequencing is less than optimal, considering the high concentration of mtDNA per swab and the multiple high-volume sample transfer steps. Remarkably, the RBs for the blood samples, even when undiluted after post-amplification quantification, produced less than 500 clusters from which to derive sequence data. This method of extract merits further investigation for samples intended for deep-sequencing.

Section 5.4: Assessments of noise and designing experimental controls

An experimental goal was to reliably detect minor variants above some measurable level of noise. Identifying patterns of noise is important, as accounting for its sources will likely affect relevant calculations of LOD. It was assumed that unexpected variation was noise. This noise is observed at low levels and appears mostly as transition differences (Table 16). There are many possible explanations for this: (1) the minor variation represents an instrument error, (2) the noise is PCR induced substitution, (3) the noise represents true heteroplasmy at extremely low levels, or (4) some combination thereof. A reexamination of how the instrument performed sequence analysis eliminated instrument error as a potential cause of variation. One of the major improvements that the MiSeq™ made over previous generations of Illumina® instruments is that it now excites fluorophores with a green and red LED. The red LED excites fluorophores for adenosine and cytosine; the green LED for guanine and thymine. This creates a filter and effectively separates the excitation of purines and pyrimidines.

CHROM	POS	REF	T#oRs	A CALLS	C CALLS	G CALLS	T CALLS	QARs	T#oDs	% MAJ	% MIN
rCRS	16126	T	8019	1	461	0	7135	7597	462	93.92	6.07
rCRS	16069	C	8005	0	7077	0	442	7519	442	94.12	5.88
rCRS	16129	G	7969	7360	0	443	0	7803	7360	94.32	5.68
rCRS	16093	T	7999	0	442	8	7368	7818	450	94.24	5.65
rCRS	185	G	7988	296	0	6857	0	7153	296	95.86	4.14
rCRS	228	G	3516	114	0	2952	1	3067	115	96.25	3.72
rCRS	295	C	1534	0	1403	0	54	1457	54	96.29	3.71
rCRS	310	T	1518	0	10	0	687	697	10	98.57	1.43
rCRS	302	A	1528	866	7	0	0	873	7	99.20	0.80
rCRS	152	T	8020	0	37	0	7863	7900	37	99.53	0.47
rCRS	204	T	5656	0	19	0	4107	4126	19	99.54	0.46
rCRS	340	C	1282	0	1137	0	4	1141	4	99.65	0.35
rCRS	53	G	3016	8	0	2717	0	2725	8	99.71	0.29
rCRS	16027	T	4590	0	13	0	4469	4482	13	99.71	0.29
rCRS	254	T	1475	0	4	0	1381	1385	4	99.71	0.29
rCRS	223	T	3999	0	9	0	3389	3398	9	99.74	0.26
rCRS	339	A	1300	1162	0	3	0	1165	3	99.74	0.26
rCRS	350	A	894	779	0	2	0	781	2	99.74	0.26
rCRS	16348	C	2181	0	2032	0	5	2037	5	99.75	0.25

Table 16 – The first twenty entries of Galaxy™ data for 5% Donor 001-CF30BLD::95% Donor 005-CF40BLD mixture experiment. Sites of unexpected variation have their minor basecalls boxed and appear as transitions.

It may be possible to statistically assess noise within Galaxy™ outputs. Early attempts to assess noise performed Monte Carlo sampling on confusion matrices that were generated from alignments of the PhiX control. This confusion matrix uses an aligned reference to tabulate the number of associated basecalls when a particular base is expected. An example of this is given in Table 17 and states that from all the independently aligned reads, 216 "C" calls were made when "A" was expected to be true. From this confusion matrix, an expected frequency matrix could be constructed. PhiX was spiked in to each of the experiments to constitute 20% of the library by volume. This was done initially out of a concern that the random fragmentation of small amplicons during Nextera® XT would create a library with low diversity. While this was not the case, PhiX sequence information could however be used to provide an instrument control—it is a synthetic piece of DNA that could hybridize to the flow cell and bridge amplify but is not amplified either during Nextera® XT processing or prior to it. Unindexed reads in run quality metrics (Tables 7, 9, 11) were the PhiX reads and their resultant sequence information was parsed by the instrument into an 'Undetermined'

	A	C	G	T	N	
A	334179	216	62	60	0	334517
C	265	300448	23	366	0	301102
G	64	25	325468	90	0	325647
T	143	363	61	448917	0	449484
N	0	0	0	0	0	
	334651	301052	325614	449433		

	A	C	G	T	N
A	0.99899	0.000646	0.000185	0.000179	0
C	0.00088	0.997828	7.64E-05	0.001216	0
G	0.000197	7.68E-05	0.99945	0.000276	0
T	0.000318	0.000808	0.000136	0.998739	0
N	0	0	0	0	0

	1000	1100	1200	1300	1400	1500	1600	1700	1800
A	20.79586	19.024403	17.846014	16.620779	16.795162	17.335217	17.199105	18.602134	21.658948
C	17.577261	17.223666	16.472385	15.602229	16.804868	17.085458	17.044491	18.674973	19.306813
G	22.054762	24.214134	21.801427	20.394611	19.671956	25.357076	23.768325	26.994567	25.672225
T	24.517635	23.310511	22.0759	20.312565	18.819179	17.638059	17.379383	16.229033	16.389636

Table 17 – Confusion matrices (top left), expected frequencies (top right), and simulated χ^2 values using Monte Carlo sampling of the PhiX control DNA (bottom). χ^2 values were simulated at 100x increments (coverage) up to 8,000 trials (not all simulations shown).

POS	REF	T#oRs	A CALLS	C CALLS	G CALLS	T CALLS	QARs	Expect (A)	Expect (C)	Expect (G)	Expect (T)	χ^2	p-value	est. χ^2 cut	est. pass
16126	T	3159	0	3030	0	5	3035	0.97	2.45	0.41	3031.17	3742675.21	0.00	17.54	TRUE
16093	T	2922	0	2517	0	175	2692	0.86	2.17	0.37	2688.60	2911382.73	0.00	17.38	TRUE
185	G	2452	2311	0	34	0	2345	0.46	0.18	2343.71	0.65	11586073.48	0.00	19.92	TRUE
16069	C	2019	0	2	0	1884	1886	1.66	1881.90	0.14	2.29	1546405.27	0.00	19.31	TRUE
16198	T	5802	0	10	0	5616	5626	1.79	4.54	0.76	5618.90	9.11	0.03	14.71	FALSE
189	A	2459	2205	1	8	0	2214	2211.76	1.43	0.41	0.40	140.92	0.00	18.19	TRUE
16233	A	5724	5317	0	9	0	5326	5320.62	3.44	0.99	0.96	69.44	0.00	14.28	TRUE
16278	C	5643	0	5523	0	9	5532	4.87	5519.98	0.42	6.72	6.06	0.11	16.11	FALSE
16274	G	5649	8	0	5206	0	5214	1.02	0.40	5211.13	1.44	49.33	0.00	17.17	TRUE
16291	C	5606	0	4971	0	8	4979	4.38	4968.19	0.38	6.05	5.39	0.15	14.76	FALSE
195	T	2221	0	7	0	2157	2164	0.69	1.75	0.29	2161.27	16.78	0.00	17.05	FALSE
16173	C	5200	1	4992	0	6	4999	4.40	4988.14	0.38	6.08	3.01	0.39	14.76	FALSE
16230	A	5696	5503	0	7	0	5510	5504.43	3.56	1.02	0.99	39.55	0.00	14.89	TRUE
16256	C	5656	0	5071	0	7	5078	4.47	5066.97	0.39	6.17	4.97	0.17	14.70	FALSE

Table 18 – Application of the Monte Carlo method to Galaxy™ outputs for 001-CF30Buc unmixed where the null hypothesis does not rule out instrument error for the distribution of observed bases. All sites that where estimated pass = 'True,' reject the null hypothesis—that something other than instrument error resulted in the observed base distribution. The simulated χ^2 value used at each position was fitted to along the χ^2 plot in Table 16 using coverage and reference call.

fastq. file. 'Undetermined' fastq. files were taken from the instrument and aligned to the PhiX genome using the same filtering criteria described above. All ambiguous bases ('N') were then trimmed from aligned reads and a confusion matrix (Table 17, top left) was constructed by examining all the aligned sequence reads and totaling the number of basecalls when a particular base was expected. This gives a distribution frequency of correctly called and miscalled bases (Table 17, top right). Monte Carlo sampling was performed to compare observed χ^2 values against simulated χ^2 values that were obtained through computational trials using the distribution frequencies for PhiX (Table 18). This method of analysis was suggested given the low expected values observed in Table 18 (Jocelyne Bruand, Illumina®, personal communication). Using the obtained distribution frequencies, Monte Carlo sampling was performed over a number of trails that incrementally increased by 100X coverage until it achieved 8,000 trials. Obtained χ^2 values were not demonstrably different from those obtained using 50,000 and 100,000 trials. When the calculated p-value for a particular position fell below an alpha-value of 0.01 and the calculated χ^2 was greater than its simulated value, then a value of 'TRUE' was given. This is interpreted as a rejection of the null hypothesis, that the test could not

eliminate instrument error (based on the PhiX distribution) as a potential explanation for the observed basecalls. When the null hypothesis is rejected at a p-value of 0.01, instrument error alone cannot explain the observed distribution of basecalls.

These statistical analyses were performed well before it was deduced that instrument error was likely not a factor in generating these calls. However, it does offer a methodology for analyzing future controls which may explain the predominance of transitions. PhiX is a synthetic genome and may not exhibit the same pattern of substitution observed in mtDNA. The positive control, HL60, comes from an immortalized, cancer cell-line and shows high variation at several positions across independent sequencing runs—runs which used amplified mtDNAs from separate PCR reactions that were performed under similar conditions. Far more desirable as a control would be a singular mtDNA haplotype. Methods of subcloning DNA isolated a single mtDNA haplotype in experiments intended for Sanger sequencing (Pfeiffer et al. 2003) but may still exhibit a low rate of substitution induced by the enzymatic amplification of DNA during processing and could alter the observed basecall distribution in NGS datasets. An alternative may be to synthetically prepare DNA fragments. These fragments would have a similar base composition to mtDNA. Additionally, the sum of these synthetic fragments could cover the entirety of the mtGenome. Whether developed in-house or commercially purchased, a control mtDNA that is properly balanced and capable of bridge amplification may be useful for measuring the rates of PCR induced substitution. If the deep-sequence results from an unamplified control DNA has a substantially different background pattern compared to an amplified counterpart, then it could be said that the observed background may be the result of PCR induced

substitution. If not, then the observed background in these experiments may be attributed to biological variation.

Section 5.5: Suggestions for establishing an interpretational threshold and optimizing depth of coverage

It is difficult to establish an interpretation threshold beyond an arbitrary value of 1% without accounting for all relevant NGS variables. Many variables are present which may affect assessments of observed noise. For example, there are many methods of data filtering based on quality metrics. The quality filtering scheme applied by this project ($\geq$Q20 over 90% of the base composition) was performed in order to prevent alignment of poorer quality sequences from competing with better quality reads for space in the capped data pileups. The removal of the computational cap from data pileups may merit an investigation of other methods of quality filtering and their respective effects on the sequence data. Also, filtering criteria may have some platform specific components to it. For example, Illumina® data tends to be of poorer quality closer to the 3'-end of the sequence read, which may require some trimming in order to generate reads with higher mapping quality.

Different alignment algorithms perform in different ways. For instance, the local alignment tool in CLC bio clips sequence entries into small fragments of approximately 20 bases in length and will remove poor quality basecalls from generated fragments. The software will then attempt to align each fragment to the reference using mismatch and gap penalties defined by the user. As penalties accrue, the software has less confidence in the alignment and will then seek to align the fragment elsewhere until an alignment

with the highest mapping quality is reached. This method of alignment is generally better at describing insertion sites around homopolymeric stretches than the analysis pipeline described in this project. However, it is not known how the local alignment algorithm may perform when several SNPs are in close proximity, and also, how reads that are NumtS in origin are handled.

Sequencing the primer is of no biological relevance and because of this, primer sequences are usually trimmed from NGS datasets. Primer trimming can be a challenge if overlapping amplicons constitute the DNA library. For example, a trimming scheme that is designed to remove basecalls associated with the B2 primer sequence would be appropriate for the HV1a amplicon, only if the amplicon was prepared and sequenced by itself. This method of trimming has no ability to differentiate the origin of a read. Therefore, if this method of trimming were applied to data resulting from a pool of prepared CR amplicons, it would also remove good quality reads from HV1b, whose breath of sequence data spans the B2 primer binding region of the other amplicon. This would also be true for the other primer binding regions which frame the overlapping CR amplicons (A2, C2, and D2).

Bidirectionality must also be considered. If a particular base is observed in both mate pairs (the forward and reverse read) of an individual cluster, there is generally greater confidence in that call. However, this greater confidence is not conveyed in the 'Generate Pileup' function of SAMtools, which ignores mate pairs. How this greater confidence should be accounted for has yet to be determined. If bidirectionality is desirable, additional library purification steps using SPRI beads may reduce the range of

library length to 100-200 bp and increase the degree of overlap between mate pairs. However, these additional steps may lower the overall yield of the library and will likely affect cluster density.

Sites associated with NumtS were visually filtered from datasets produced in this project. Developed bioinformatic pipelines should be able to identify reads that arise from NumtS, possibly by identifying reads which contain two or more NumtS variants. Once these reads are identified as NumtS, the read and its associated mate pair might be filtered from datasets during alignment. Whether this can be done effectively without discarding reads of mtDNA origin is unknown.

All aforementioned caveats of the bioinformatics design, along with minimum coverage and Q-score cutoff for basecalling, affects the post-filtered coverage and base composition of reported positions. This subsequently affects calculations of %MIN. Therefore, each variable should be considered if resultant calculations of %MIN are intended for the derivation of an interpretational limit of detection, above which, a mixture would be detected. If this calculation uses a computational average of %MIN across many positions, the post-filtered coverage should be considered as it will give additional weight to the calculation of %MIN. For example, a 2% variant that has 500X coverage may not be treated the same as 2% variant with 50,000X coverage.

Associating the coverage per base in post filtered data with the ability to identify all minor variants at the prescribed threshold may reveal an ideal cluster count per index needed for the detection of all minor variants. Once a targeted post-filtered coverage per base is achieved, then targeting of a defined number of clusters per index can be tested.

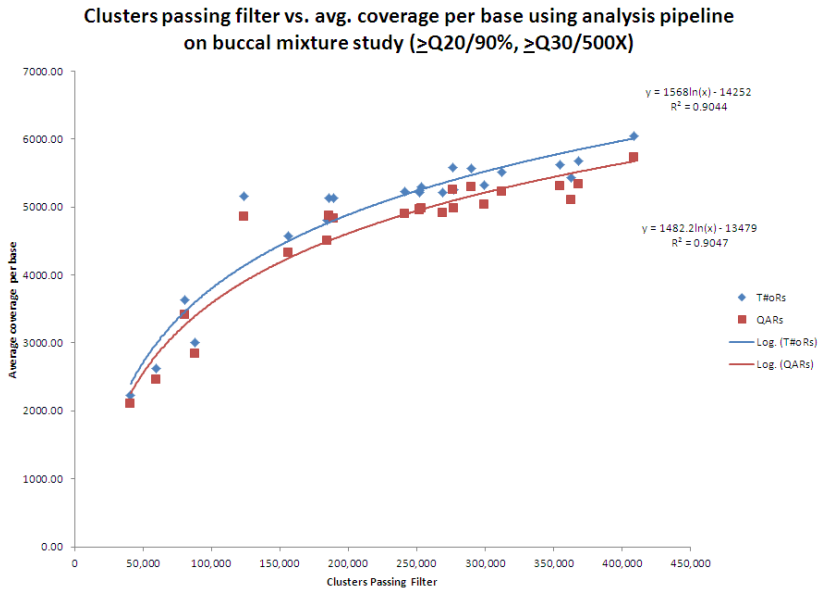

Figure 23 – Correlation of the average coverage per base in the buccal mixture experiment with respect to the number of clusters in which sequence data was derived.

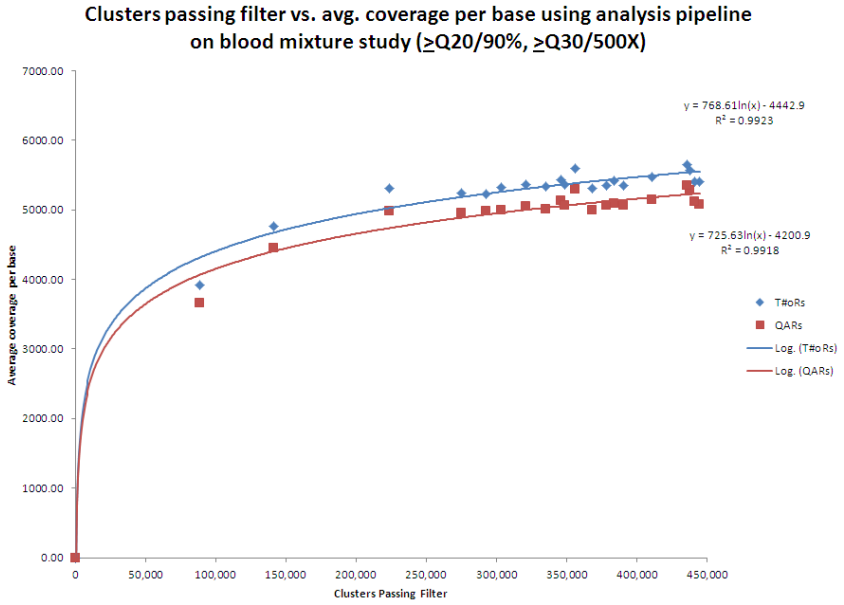

Figure 24 – Correlation of the average coverage per base in the blood mixture experiment with respect to the number of clusters in which sequence data was derived.

Depicted in Figure 23 and 24, are the average coverage per base calculations obtained from buccal and blood mixture experiments. The plateau of the logarithmic regressions is low due to the limit induced by the computational pileup cap.

Section 5.6: Overview of the tissue experiments and sequence variation

Deep sequence data derived from hair shaft amplicons shows a higher degree of sequence variation in Galaxy™ outputs and indicates that hairs may harbor a higher degree of sequence variation with respect to other tissue types. These findings are consistent with earlier studies of hair shafts using Sanger sequencing methodologies (Wilson et al 1997). In Table 19, the buccal swab from donor 001-CF30 shows two sites of variation greater than 1%. These positions show comparable levels of variation to the

CHROM	POS	QARs	T#ofDs	% MAJ	% MIN
rCRS	185	2968	2935	98.89	1.11
rCRS	228	1893	1886	99.63	0.37
rCRS	73	1103	1100	99.73	0.27
rCRS	263	752	751	99.87	0.13
rCRS	16126	653	653	100.00	0.00
rCRS	16390	613	613	100.00	0.00
rCRS	16093	564	524	92.91	7.09
rCRS	189	2769	16	99.42	0.58
rCRS	130	2583	10	99.61	0.23

CHROM	POS	QARs	T#ofDs	% MAJ	% MIN
rCRS	16126	7805	7802	99.95	0.04
rCRS	16069	7790	7790	100.00	0.00
rCRS	16093	7834	7656	97.70	2.27
rCRS	185	7628	7421	97.29	2.71
rCRS	73	4367	4364	99.93	0.07
rCRS	228	4032	4026	99.85	0.15
rCRS	295	1608	1596	99.25	0.75
rCRS	263	1497	1496	99.93	0.07
rCRS	16060	7791	51	99.35	0.65

CHROM	POS	QARs	T#ofDs	% MAJ	% MIN
rCRS	16126	7879	7877	99.96	0.03
rCRS	16069	7832	7831	99.99	0.01
rCRS	185	7523	7520	99.96	0.04
rCRS	16093	7881	7447	94.49	5.51
rCRS	73	6169	6167	99.97	0.03
rCRS	295	3522	3483	98.89	1.11
rCRS	228	3208	3201	99.78	0.22
rCRS	263	3171	3171	100.00	0.00
rCRS	16221	7846	151	98.08	1.92

CHROM	POS	QARs	T#ofDs	% MAJ	% MIN
rCRS	16126	7698	7692	99.92	0.08
rCRS	16093	6911	6908	99.96	0.04
rCRS	185	6509	6505	99.94	0.06
rCRS	16069	4982	4980	99.96	0.04
rCRS	228	4727	4721	99.87	0.13
rCRS	263	3096	3095	99.97	0.03
rCRS	73	2643	2637	99.77	0.23
rCRS	295	1950	1904	97.64	2.36
rCRS	16390	759	758	99.87	0.13

CHROM	POS	QARs	T#ofDs	% MAJ	% MIN
rCRS	185	5960	5960	100.00	0.00
rCRS	16126	5675	5675	100.00	0.00
rCRS	16093	5066	4950	97.71	2.29
rCRS	228	4519	4513	99.87	0.13
rCRS	16069	3509	3509	100.00	0.00
rCRS	263	2882	2882	99.97	0.03
rCRS	73	2163	2163	100.00	0.00
rCRS	295	1679	1626	96.84	3.16
rCRS	16390	814	813	99.75	0.12

CHROM	POS	QARs	T#ofDs	% MAJ	% MIN
rCRS	16069	7859	7858	99.99	0.01
rCRS	16126	7843	7841	99.96	0.03
rCRS	185	7584	7582	99.97	0.03
rCRS	73	4907	4894	99.74	0.26
rCRS	16093	7896	4457	56.45	43.55
rCRS	228	4353	4351	99.95	0.05
rCRS	295	3151	3137	99.56	0.44
rCRS	263	2829	2828	99.93	0.04
rCRS	16092	7888	194	97.54	2.46

Table 19 – First ten entries of Galaxy™ data for various samples from donor 001-CF30. (Top left) Buccal reference. (Center left) Hair A. (Bottom left) Hair B. (Top right) Hair C. (Center right) Hair D. (Bottom right) Hair E. Gold = expected variant, Yellow = expected variation with %MIN >1%, Brown = expected variation showing a large degree of variation, Gray = unexpected variant with %MIN >1%.

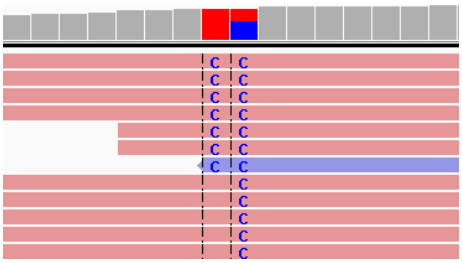

Figure 25 – Visualization of aligned reads for 001-CF30 Hair E using IGV. The position bound by brackets is 16092. All aligned reads have been sorted with respect to the base composition at 16092.

unmixed buccal reference in the first mixture experiment (Appendix 1). Across the five hair samples collected from this donor, these positions vary in the degree of observed %MIN. In four hairs (19a, b, d, e), the minor variant at position 185 is not observed. Interestingly, the minor variant at 16093 is near absent in one hair (19d, %MIN = 0.04) and shows a much larger degree of sequence variation in another (19e, %MIN = 43.55). Accompanied with the mixed sequence pattern at 16093 is the occurrence of a minor variant at 16092. Remarkably, when the minor variant at 16092 is visualized using IGV (Figure 25), in all cases observed thus far, the 16092 variant appears to be accompanied with the cytosine residue in the neighboring 16093 position.

Additional observations of Galaxy™ outputs for hair sequence data indicated that NumtS were not recovered in any of the hair samples. This is not unsurprising considering that the nuclear content of hair shafts are expected to be very low. Hair sequence data also retained more positions from which to derivate %MIN after bioinformatic filtering (described as 'sequence window' in appendices 1-4) than those of blood and buccal experiments. The explanation for this is unclear. It is similarly unclear as to why the sequence data derived from two of the hair shafts prepared from donor 001-CF30 were able to recover the distal 16390A variant (Appendix 3) which was

consistently absent in every buccal and blood experiment. When this variant was recovered in hair sequence data however, it achieved less than 1,000X coverage.

Galaxy™ datasets from hair samples also contained more positions with minor variants greater than 1% relative to buccal reference material. When the individual hairs are sorted by %MIN in descending order, several UVs are recovered (Appendices 3 and 4). Most of these UVs are not observed as variants within the buccal reference, so it is curious to see these in hair samples. These minor variants also appear to be scattered randomly throughout the range of obtained sequence data so it is unclear whether this is a result of natural variation within the sample tissue, the additional cycles of PCR, or some combination thereof. Without knowing the source of this variation, it is uncertain as to how a mixture detection threshold may need to account for this increased sequence variation.

CHAPTER SIX: REFERENCES

Anderson, S. Bankier, A. T., Barrel, B. G., de Bruijn, M. H., Coulson, A. R., Drouin, J., Eperon, I. C., Nierlich, D. P., Roe, B. A., Sanger, F., Schreier, P. H., Smith, A. J., Staden, R., Young, I. G. "Sequence and organization of the human mitochondrial genome." Nature, 290, 1981, pp. 457-465.

Andréasson, H., Nilsson, M., Styrman, H., Pettersson, U., Allen, M. "Forensic mitochondrial coding region analysis for increased discrimination using pyrosequencing technology." Forensic Science International: Genetics, 1, 2006, pp. 35-43.

Andréasson, H., Nilsson, M., Budowle, B., Frisk, S., Allen, M. "Quantification of mtDNA mixtures in forensic evidence material using pyrosequencing." International Journal of Legal Medicine, 120, 2006, pp. 383-390.

Applied Biosystems. "BigDye® Terminator v1.1 Cycle Sequencing Kit: Protocol." Foster City, CA. 2002.

Applied Biosystems. "Applied Biosystems 3130/3130*xl* Genetic Analyzers: Getting Started Guide." Foster City, CA. 2004."

Applied Biosystems. "Quantifiler® Kits, Quantifiler Human DNA Quantification Kit and Quantifiler Y Human Male DNA Quantification Kit: User's Manual." Foster City, CA. 2006.

Blankenberg D, Gordon A, Von Kuster G, Coraor N, Taylor J, Nekrutenko A; Galaxy Team. Manipulation of FASTQ data with Galaxy. Bioinformatics. 2010 Jul 15;26(14):1783-5.

Bendall, K. E., Macaulay, V. A., Baker, J. R., Sykes, B. C. "Heteroplasmic point mutations in the human mtDNA control region." American Journal of Human Genetics, 59, 1996, pp. 1276-1287.

Bentley, D. R. et al. "Accurate whole human genome sequencing using reversible terminator chemistry." Nature, 456, 2008, pp. 53-59.

Bintz, B., Dixon, G., Wilson, M. R. "Simultaneous Detection of human mitochondrial DNA and nuclear inserted mitochondrial-origin sequences (Numts) using forensic mtDNA amplification strategies and pyrosequencing technology." Journal of Forensic Sciences, 2013, in press.

Bogenhagen, D. F. "Repair of mtDNA in Vertebrates." The American Journal of Human Genetics, 64, 1999, pp 1276-1281.

Burnside, E. S., Bintz, B. J., Wilson, M. R. (2013) "Improved extraction efficiency of human mitochondrial DNA from hair shafts and its implication for sequencing of

the entire mtGenome from a single hair fragment." In: Proceedings of the American Acadmey of Forensic Sciences. Washington, D.C. http:// www.aafs.org/sites/default/files/pdf/ProceedingsWashingtonDC2013.pdf

Butler, J. M. *Mitochondrial DNA Analysis.* Forensic DNA Typing: Methodology. Elsevier Academic Press, Waltham, MA, 2011, pp. 405-456.

Caruccio, N. "Preparation of Next-Generation Sequencing libraries using Nextera™ technology: simultaneous DNA fragmentation and adaptor tagging by *in vitro* transposition." Methods in Molecular Biology, 733, 2011, pp. 241-255.

Chen, X., Prosser, R., Simonetti, S., Sadlock, J., Jagiello, G., Schon, E. A. "Rearranged mitochondrial genomes are present in human oocytes." American Journal of Human Genetics, 57, 1995, pp. 239-247.

DeAngelis, M. M., Wang, D. G., Hawkins, T. L. "Solid-phase reversible immobilization for the isolation of PCR products." Nucleic Acids Research, 23, 1995, pp. 4742-4743.

DiZinno, J. A., Wilson, M. R., Budowle, B., *Typing of DNA derived from hairs.* Forensic Examination of Hair. Taylor and Francis, London, 1999, pp. 155-173.

Ewing, B., Green, P. "Base-calling of automated sequencer traces using phred. II. Error probabilities." Genome Research, 8, 1998, pp. 186-194.

Giardine, B., Riemer, C., Hardison, R. C., Burhans, R., Elnitski, L., Shah, P., Zhang, Y., Blankenberg, D., Albert, I., Taylor, J., Miller, W., Kent, W. J., Nekrutenko, A. "Galaxy: a platform for interactive large-scale genome analysis." Genome Research, 15, 2005, pp. 1451-1455.

Gill, P., Ivanov, P. L., Kimpton, C., Piercy, R., Benson, N., Tully, G., Evett, I., Hagelberg, E., Sullivan, K. "Identification of the remains of the Romanov family by DNA analysis." Nature Genetics, 6, 1994, pp. 130-135.

Gordon A., Hannon GJ. "FASTX-Toolkit", FASTQ/A short-reads pre-processing tools (unpublished) http://hannonlab.cshl.edu/fastx_toolkit/

Hecht, N. B., Liem, H., Kleene, K. C., Distel, R. J., Ho, S. "Maternal inheritance of the mouse mitochondrial genome is not mediated by a loss or gross alternation of the paternal mitochondrial DNA or by methylation of the oocyte mitochondrial DNA." Developmental Biology, 102, 1984, pp. 452-461.

Holland, M. M., Parsons, T. J. "Mitochondrial DNA sequence analysis—validation and use for forensic casework." Forensic Science Review, 11, 1999, pp. 22-50.

Hutchison III, C. A., Newbold, J. E., Potter, S. S., Edgell, M. H. "Maternal inheritance of mammalian mitochondrial DNA." Letters to Nature, 251, 1974, pp. 536-538.

Illumina®. "MiSeq™ Reporter Quick Reference Guide." P#15038784, Rev. A. 2011.

Illumina®. "MiSeq™ System User Guide." P#15027617, Rev. C. 2012.

Illumina®. "Nextera XT Sample Preparation Guide." P# 15031942, Rev A. 2012.

Ivanov, P. L., Wadhams, M. J., Roby, R. K., Holland, M. M., Weedn, V. W., Parsons, T. J. "Mitochondrial DNA sequence heteroplasmy in the Grand Duke of Russia Georgij Romanov establishes the authenticity of the remains of Tsar Nicholas II." Nature Genetics, 12, 1996, pp. 417-420.

Kunkel, T. A., Loeb, L. A. "Fidelity of mammalian polymerases." Science, 213, 765, 1981.

Kavlick, M. F., Lawrence, H. S., Merritt, T., Fisher, C., Isenberg, A., Robertson, J. M., Budowle, B. "Quantification of human mitochondrial DNA using synthesized DNA standards." Journal of Forensic Sciences, 56, 2011, pp. 1457-1463.

Li H. and Durbin R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics, 25, 1754-60.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R; 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009 Aug 15;25(16):2078-9.

Linch, C. A., Whiting, D. A., Holland, M. M. "Human Histogenesis for the Mitochondrial DNA Forensic Scientist." Journal of Forensic Science, 46(4), 2001, pp. 844-853.

Melton, T., Holland, C., Holland, M. "Forensic mitochondrial DNA analysis: current practice and future potential." Forensic Science Review, 24, 2012, pp. 101-122.

Milton, J., Wu, X., Smith, M., Brennan, J., Barnes, C., Liu, X., Ruediger, S. "Modified Nucleotides." World Intellectual Property Organization, WO/2004/018497.

Parker, L. T., Deng, Q., Zakeri, H., Carlson, C., Nickerson, D. A., Kwok, P. Y. "Peak height variations in automated sequencing of PCR products using *Taq* dye-terminator chemistry." Biotechniques, 19, 1995, pp. 116-121.

Parker, L. T., Zakeri, H., Deng, Q., Spurgeon, S., Kwok, P. Y., Nickerson, D. A. "AmpliTaq DNA polymerase, FS dye-terminator sequencing: analysis of peak height patterns." Biotechniques, 21, 1996, pp. 694-699.

Parson, W., Dür, A. "EMPOP—A forensic mtDNA database." Forensic Science International: Genetics, 1, 2007, pp. 88-92.

Parsons, T. J., Muniec, D. S., Sullivan, K., Woodyatt, N., Aliston-Greiner, R., Wilson, M. R., Berry, D. L., Holland, K. A., Weedn, V. W., Gill, P., and Holland, M.

M. "A high observed substitution rate in the human mitochondrial DNA control region." Nature Genetics, 15, 1997, pp. 363-367.

Pfeiffer, H., Lutz-Bonengel, S., Pollak, S., Fimmers, R., Baur, M. P., Brinkmann, B. "Mitochondrial DNA control region diversity in hairs and body fluids of monozygotic triplets." International Journal of Legal Medicine, 118, 2004, pp. 71-74.

Princeton Seperataions, Inc. "Centri-Sep™ Columns." Adelphia, NJ. 1997.

Qiagen®. "QIAmp® DNA Mini and Blood Mini Handbook." 2010.

Qiagen®. "REPLI-g® Mitochondrial DNA Handbook." West Sussex, England. 2011.

Robin, E. D., Wong, R. "Mitochondrial DNA molecules and virtual number of mitochondria per cell in mammalian cells." Journal of Cell Physiology, 136, 1988, pp. 507-513.

Satoh, M. and Kuroiwa, T. *Experimental Cell Research*, 196, 1991, pp. 137-140

Schwartz, M. and Vissing, J. "Paternal inheritance of mitochondrial DNA." The New England Journal of Medicine, 347, 2002, pp. 576-580.

Scientific Working Group on DNA Analysis Methods (SWGDAM). *Guidelines for mitochondrial DNA (mtDNA) nucleotides sequence interpretation*. April 2003.

Sutovsky, P., McCauley, T. C., Sutovsky, M., Day, B. N. "Early degradation of paternal mitochondria in domestic pig (*Sus scrofa*) is prevented by selective proteasomal inhibitors lactacystin and MG132." Biology of Reproduction, 68, 2003, pp. 1793-1800.

Tully, L. A., Parsons, T. J., Seighner, R. J., Holland, M. M., Marino, M. A., Prenger, V. L. "A sensitive denaturing gradient-gel electrophoresis assay reveals a high frequency of heteroplasmy in hypervariable region 1 of the human mtDNA control region." The American Journal of Human Genetics, 67, 2000, pp. 432-443.

Underhill, P. A., Jin, L., Lin, A. A., Mehdi, S. Q., Jenkins, T., Vollrath, D., Davis, R. W., Cavalli-Sforza, L. L., Oefner, P. J. "Detection of numerous Y chromosome biallelic polymorphisms by denaturing high-performance liquid chromatography." Genome Research, 10, 1997, pp. 996-1005.

Whatman. "Preparing an FTA® Disc for DNA Analysis: Protocol BD09." 2009.

Wilson, M. R., Holland, M. M., Stoneking, M., DiZinno, J. A., Budowle, B. "Guidelines for the use of mitochondrial DNA sequencing in forensic science." Crime Laboratory Digest, 20, 1993, pp. 68-77.

Wilson, M. R., DiZinno, J. A., Polanskey, D., Replogle, J., Budowle, B. "Validation of mitochondrial DNA sequencing for forensic casework analysis." International Journal of Legal Medicine, 108, 1995, pp. 68-74.

Wilson, M. R., Polanskey, D., Butler, J., DiZinno, J. A., Replogle, J., Budowle, B. "Extraction, PCR amplification and Sequencing of mitochondrial DNA from human hair sahfts." Biotechniques, 18, 1995, pp. 662-669.

Wilson, M. R., Polanskey, D., Replogle, J., DiZinno, J. A., Budowle, B. "A family exhibiting heteroplasmy in the human mitochondrial DNA control region reveals both somatic mosaicism and pronounced segregation of mitotypes." Human Genetics, 100, 1997, pp. 167-171.

Zischler, H., Geisert, H., Von Haeseler, A., PAAbo, S. "A nuclear 'fossil' of the mitochondrial D-loop and the origin of modern humans." Nature, 378, 1995, pp. 489-492.

	Recovery of Expected Variants (EV)	Positions showing unexpectedly high minor variation (≥1%) and associated frequencies	Length of sequence window	Raw clusters PF	T#ofRs: Avg. coverage per base	QARs: Avg. coverage per base	QARs: Avg. coverage per EV	Avg. % Minor	Artifacts
001-CF30Buc	8 of 9*	185 (1.45), 295 (2.77), 16093 (6.50)	629 bp	59,405	2624.66	2465.78	1744.75	--	N
005-CF40Buc	3 of 3	152 (4.28)†	715 bp	156,001	4575.90	4332.82	4472.00	--	N, SC
003-54MBuc	13 of 13	16242 (9.88)‡	736 bp	185,674	5136.19	4870.41	5721.08	--	N, SC
015-AM35Buc	9 of 9	none	730 bp	184,493	4801.43	4510.05	4709.33	--	N, SC
RB-Buc	--	none	246 bp	87,885	3006.24	2844.06	--	--	none
NC-Water	--	none	217 bp	40,831	2224.82	2109.34	--	--	none
PC-HL60	9 of 9	295 (1.73), 16362 (1.15)	721 bp	124,033	5162.87	4858.85	5073.67	--	N, SC

	Recovery of Expected Variants (EV)	Recovery of unexpected variants (UV) and associated frequencies	Length of sequence window	Raw clusters PF	T#ofRs: Avg. coverage per base	QARs: Avg. coverage per base	QARs: Avg. coverage per EV	Avg. % Minor	Artifacts
5.0% Donor 5	7 of 8*	none	722 bp	189,392	5131.95	4830.94	6165.57	5.82	N, SC
2.0% Donor 5	7 of 8*	none	719 bp	276,787	5253.36	4980.89	5703.25	2.72	N
1.0% Donor 5	7 of 8*	none	721 bp	252,239	5216.63	4950.48	6338.14	2.39	N
0.5% Donor 5	7 of 8*	189 (0.71), 342 (0.42)	715 bp	269,290	5215.16	4917.13	6294.17	1.60	N
5.0% Donor 1	7 of 8*	152 (3.76)	715 bp	253,590	5298.85	4986.83	6381.86	4.83	N
2.0% Donor 1	7 of 8*	152 (3.70)	719 bp	251,804	5243.78	4962.34	5697.11	1.57	N
1.0% Donor 1	7 of 8*	152 (4.12), 385 (0.64)	721 bp	299,184	5320.73	5030.16	6409.71	0.80	N
0.5% Donor 1	7 of 8*	53 (0.47), 66 (0.68), 68 (0.40), 152 (4.18), 195 (0.52) 214 (0.64), 366 (0.45), 385 (1.09), 16360 (0.82), 16365 (0.42)	711 bp	363,193	5436.21	5097.95	6130.00	0.68	N
5.0% Donor 15	16 of 16	none	716 bp	80,797	3632.23	3409.62	4918.14	4.91	N, SC
2.0% Donor 15	16 of 16	385 (1.08)	718 bp	290,089	5571.34	5290.29	6421.29	1.68	N
1.0% Donor 15	14 of 16§	none	706 bp	408,890	6051.54	5729.03	7170.14	1.12	N
0.5% Donor 15	15 of 16‡	195 (0.38), 215 (0.39), 385 (0.65), 16020 (0.39), 16370 (0.46)	717 bp	276,648	5577.44	5252.73	6669.13	0.65	N
5.0% Donor 3	16 of 16	none	735 bp	241,253	5230.81	4906.27	6109.00	5.91	N
2.0% Donor 3	16 of 16	none	727 bp	312,132	5517.44	5224.29	6357.81	2.57	N, SC
1.0% Donor 3	16 of 16	195 (0.66)	726 bp	354,666	5625.49	5313.25	6343.75	1.37	N, SC
0.5% Donor 3	16 of 16	195 (0.71)	717 bp	368,587	5677.89	5333.93	6311.00	1.12	N, SC

Appendix 1 – Summary of Galaxy™ data from the buccal mixture experiment. Data was additionally filtered of NumtS induced variation and BWA alignment errors. SNPs that occurred over NumtS positions were retained. C-stretch variants were not treated as a site of expected variation. Unexpected variants are positions in which the calculated minor variant was higher than an expected variant. Artifact nomenclature: N = variation at NumtS positions, SC = alignment errors induced by BWA soft clipping reads.

* = expected variant 16390 not covered in the sequence window

† = unexpected minor variant where the minor call is in disagreement with the rCRS

‡ = primer binding variant

§ = expected variants 16352 and 16357 are not covered in the sequence window

ǂ = no variant reported at position 16352

	Recovery of Expected Variants (EV)	Positions showing unexpectedly high minor variation (≥1%) and associated frequencies	Length of sequence window	Raw clusters PF	T#ofRs: Avg. coverage per base	QARs: Avg. coverage per base	QARs: Avg. coverage per EV	Avg. % Minor	Artifacts
001-CF30BLD	8 of 9*	295 (1.78)	681 bp	223,416	5311.40	4977.30	4980.63	--	N
005-CF40BLD	3 of 3	none	706 bp	275,088	5240.54	4953.16	4337.00	--	N
003-54MBLD	12 of 13£	16242 (1.84)‡	700 bp	437,609	5571.11	5277.22	6265.33	--	N, SC
015-AM35BLD	9 of 9	16352 (4.13)	695 bp	444,915	5408.20	5077.99	5093.67	--	N, SC
RB	n/a	n/a	no reads	278	n/a	n/a	--	--	n/a
NC-Water	n/a	n/a	no reads	362	n/a	n/a	--	--	n/a
PC-HL60	9 of 9	295 (2.28)	720 bp	141,237	4761.90	4452.70	4706.33	--	N, SC

	Recovery of Expected Variants (EV)	Recovery of unexpected variants (UV) and associated frequencies	Length of sequence window	Raw clusters PF	T#ofRs: Avg. coverage per base	QARs: Avg. coverage per base	QARs: Avg. coverage per EV	Avg. % Minor	Artifacts
5.0% Donor 5	7 of 8*	--	686 bp	303,705	5318.19	4996.33	6064.00	5.27	N
2.0% Donor 5	7 of 8*	--	691 bp	390,255	5343.86	5060.35	6091.43	3.14	N
1.0% Donor 5	7 of 8*	--	688 bp	441,102	5407.79	5116.21	5976.86	1.53	N, SC
0.5% Donor 5	7 of 8*	237 (0.59), 16366 (0.79)	684 bp	368,255	5308.44	4994.79	6129.29	0.97	N
5.0% Donor 1	7 of 8*	--	698 bp	321,126	5366.40	5048.15	6059.14	4.98	N
2.0% Donor 1	7 of 8*	--	696 bp	345,905	5435.76	5133.60	6084.71	2.01	N
1.0% Donor 1	7 of 8*	152 (0.66)†	704 bp	348,937	5362.46	5068.11	6130.29	0.91	N
0.5% Donor 1	6 of 8**	152 (0.49)†	693 bp	383,486	5421.38	5089.11	6736.17	0.62	N
5.0% Donor 15	16 of 16‖	--	712 bp	88,332	3912.17	3653.41	5391.19	5.54	N
2.0% Donor 15	15 of 16£	--	703 bp	355,839	5590.79	5290.1	6518.60	2.24	N, SC
1.0% Donor 15	14 of 16§	--	702 bp	435,738	5651.28	5344.41	6936.79	1.41	N, SC
0.5% Donor 15	16 of 16	16002 (0.41), 16368 (1.75)	707 bp	334,869	5332.48	5015.74	6409.44	1.05	N, SC
5.0% Donor 3	n/a¶	n/a	722 bp	266,125	n/a	n/a	n/a	n/a	n/a
2.0% Donor 3	16 of 16	--	715 bp	292,500	5225.35	4977.23	6171.47	2.94	N, SC
1.0% Donor 3	16 of 16	--	708 bp	378,493	5355.37	5066.14	6224.25	1.84	N, SC
0.5% Donor 3	14 of 16§	382 (0.41), 16337 (0.71)	694 bp	410,911	5469.62	5141.9	6952.79	0.69	N, SC

Appendix 2 – Summary of Galaxy™ data from the blood mixture experiment. Data was additionally filtered in the same manner described in Appendix 1. SNPs that occurred over NumtS positions were retained. C-stretch variants were not treated as a site of expected variation. Additional nomenclature:

** = expected variant 16390 not covered in sequence window—no variant reported at expected variant 295

£ = expected variants 16357 not covered in sequence window

‖ = paired-end alignment failed—forward and reverse reads were aligned in single-read alignment and resultant BAM files were joined

¶ = no minor variants detected

	Recovery of Expected Variants (EV)	Positions showing unexpectedly high minor variation (≥1%) and associated frequencies	Length of sequence window	Raw clusters PF	T#ofRs: Avg. coverage per base	QARs: Avg. coverage per base	QARs: Avg. coverage per EV	Artifacts
001-CF30Buc	7 of 9**	185 (1.11), 16093 (7.09)	541 bp	32,413	2834.30	2685.41	1220.86	none
003-54MBuc	13 of 13	295 (1.58), 385 (1.52), 16242 (2.94)‡, 16357 (1.46)	721 bp	192,895	5206.20	5020.08	5742.38	N, SC
Buc-RB-1,3	n/a	n/a	625 bp	212,381	5867.39	5663.96	n/a	none
Buc-H2O-1,3	n/a	n/a	187 bp	51,437	1306.41	1190.08	n/a	none
Buc-PC-1,3 [x32]	9 of 9	295 (1.98)	726 bp	112,918	4822.40	4599.99	4963.00	N, SC
1-HairA	8 of 9*	295 (1.11), 16000 (1.05), 16093 (5.51), 16221 (1.92)	734 bp	370,787	5719.85	5497.38	5818.12	none
1-HairB	9 of 9	68 (1.82), 295 (3.16), 16093 (2.29)	744 bp	104,326	4171.57	4013.98	3585.22	none
1-HairC	8 of 9*	185 (2.71), 16093 (2.27)	719 bp	286,162	5415.45	5183.97	5320.13	none
1-HairD	9 of 9	295 (2.36)	751 bp	120,743	4500.64	4304.09	4363.89	none
1-HairE	8 of 9*	16092 (2.46)€, 16093 (43.55)€	730 bp	354,114	5663.98	5450.92	5801.75	none
1-Hair-RB	n/a	n/a	no reads	3,138	n/a	n/a	n/a	n/a
1-Hair-H2O	n/a	n/a	no reads	379	n/a	n/a	n/a	n/a
1-Hair-PC	9 of 9	295 (3.07)	715 bp	74,530	3897.34	3677.56	3666.89	N, SC
3-HairA	13 of 13	53 (1.15), 73 (1.29), 150 (1.16), 152 (1.15), 161 (1.15), 185 (1.06), 380 (1.25), 16000 (2.93), 16106 (1.57), 16142 (1.14) 16242 (2.04)‡, 16328 (1.37), 16337 (1.92), 16357 (4.05)	748 bp	434,816	5977.24	5765.21	6659.83	SC
3-HairB	13 of 13	16242 (1.18)‡, 16291 (1.74), 16384 (1.06)	742 bp	308,180	5575.08	5376.82	6101.61	SC
3-HairC	13 of 13	16242 (1.35)‡, 16384 (1.12)	757 bp	321,887	5705.30	5447.21	6384.31	SC
3-HairD	n/a	n/a	BWA failed	105,839	n/a	n/a	n/a	n/a
3-HairE	13 of 13	16242 (1.22)‡, 16357 (1.18)	744 bp	340,797	6498.69	6246.81	6779.31	SC
3-Hair-RB	n/a	n/a	no reads	31,985	n/a	n/a	n/a	n/a
3-Hair-H2O	n/a	n/a	no reads	6,989	n/a	n/a	n/a	n/a
3-Hair-PC [x36]	9 of 9	295 (1.87)	745 bp	89,065	4622.83	4382.1	4559.78	SC

Appendix 3 – Summary of Galaxy™ data from the first hair tissue experiment. Data was additionally filtered in the same manner described in Appendix 1. No SNPs occurred in either of the buccal references over NumtS positions. No NumtS variants were detected in hair data. C-stretch variants were not treated as a site of expected variation. Additional nomenclature:
€ = The unexpected variant that occurs at position 16092 only occurs if cytosine is present in position 16093.

	Recovery of Expected Variants (EV)	Positions showing unexpectedly high minor variation (≥1%) and associated frequencies	Length of sequence window	Raw clusters PF	T#ofRs: Avg. coverage per base	QARs: Avg. coverage per base	QARs: Avg. coverage per EV	Artifacts
005-CF40Buc	3 of 3	152 (5.35), 421 (1.89)	778 bp	62,388	6385.79	5638.11	6392.33	N
015-AM35Buc	9 of 9	195 (1.35), 421 (1.39), 16223 (2.52), 16224 (2.16) 16274 (1.09), 16319 (4.78), 16352 (10.40), 16402 (1.69)	781 bp	277,077	7308.65	6926.87	7399.67	
Buc-RB-5	n/a	n/a	769 bp	273,102	6463.49	6154.34	n/a	none
Buc-RB-15	n/a	n/a	244 bp	36,481	6868.35	6490.49	n/a	none
5-15-Buc-H2O	n/a	n/a	169 bp	3,075	941.94	898.14	n/a	none
5-15-Buc-PC	9 of 9	295 (3.44), 421 (2.56), 15987 (1.06)	788 bp	303,056	7342.04	6974.09	7140.33	N
5-HairB	n/a	hair could not be associated with any reference	802 bp	326,918	7323.64	6937.86	n/a	none
5-HairC	n/a	hair could not be associated with any reference	803 bp	239,296	7372.55	6564.14	n/a	none
5-HairD	n/a	hair could not be associated with any reference	803 bp	202,906	7267.17	6499.20	n/a	none
5-HairE	n/a	hair could not be associated with any reference	805 bp	371,104	7350.48	6969.88	n/a	none
5-Hair-RB	n/a	n/a	223 bp	5,206	2061.96	1960.49	n/a	
15-HairA	9 of 9	299 (1.03), 421 (3.06), 15987 (1.25), 16402 (2.58), 16403 (1.35)	793 bp	333,945	7377.46	6985.94	4095.56	none
15-HairB	9 of 9	35 (1.07), 152 (1.09), 276 (1.24), 299 (1.65), 421 (1.96), 16218 (1.38), 16273 (1.00)	791 bp	388,292	7396.40	6993.21	7371.11	none
15-HairD	9 of 9	35 (1.26), 189 (5.75), 251 (1.52), 299 (1.33), 16145 (1.21), 16260 (3.02), 16327 (6.94)	789 bp	410,425	7398.28	7005.32	7373.56	none
15-HairE	9 of 9	94 (1.76), 203 (1.44), 299 (1.23), 16352 (1.07)	789 bp	318,750	7353.81	6559.77	6975.00	none
15-Hair-RB	n/a	n/a	no reads	1,040	n/a	n/a	n/a	n/a
5-15-Hair-H2O	n/a	n/a	n/a	645	n/a	n/a	n/a	n/a
5-15-Hair-PC	9 of 9	295 (3.91), 16362 (1.08)	782 bp	396,246	7374.2	6990.93	7112.22	N

Appendix 4 – Summary of Galaxy™ data from the second hair tissue experiment. Data was additionally filtered in the same manner described in Appendix 1. SNPs that occurred in the buccal references over NumtS positions were retained. No NumtS variants were detected in hair data. C-stretch variants were not treated as a site of expected variation. Sequence data produced from the different hairs of donor 5 could not be matched to any reference.