**DRAFT & Final Technical Report**

# Assessing Deep Sequencing Technology for Human Forensic Mitochondrial DNA Analysis

**2010-DN-BX-K171**
**Mark R. Wilson, Ph.D., Primary Investigator**
**Western Carolina University, Cullowhee, NC 28723**
**January 26, 2013**

## Abstract

Human mitochondrial DNA (mtDNA) analysis, in a forensic setting, is currently limited in both breadth (the amount of sequence data obtained) and depth (the ability to detect minor variants arising from mutations but present at very low levels). Using emerging technologies, an extension of the breadth of sequence data obtained can easily extend to the entirety of the human mtDNA genome. Extension in the complementary dimension (depth) as reported in this study, has revealed that subtle mixtures are present in forensic DNA samples that are not currently detected by current forensic mtDNA analysis.

The ultimate goal of our research effort is to generate whole mt-genome DNA sequence information from limited DNA samples, thus greatly expanding the potential utility of this marker system. We have chosen human hair shaft as a model for these challenging forensic samples. In order to accomplish this goal, we have developed enhanced DNA extraction techniques, performed whole genome amplification of the DNA extracts, employed multiplexed PCR amplification reactions to generate sufficient template mtDNA for NGS applications, and directly sequenced the samples on the Illumina® MiSeq™ instrument. Importantly, we have identified and employed a direct sample preparation step, Nextera®, that easily performs DNA library preparation from our enhanced extraction and amplification steps, thus rendering our goal within sight.

In this effort, we evaluated two newly emerging methods of DNA sequence analysis to obtain massively parallel mtDNA sequence information (deep sequencing) from hair, buccal, and blood samples. The expanded information available from deep mtDNA sequence analysis revealed that once this new technology is implemented into casework practice, interpretational changes in forensic mtDNA analysis, reflecting the amounts of information that are produced, are necessary. Deep sequencing offers a window into a level of variation that is currently under-appreciated in forensic casework. We have also revealed that the general level of sequence heteroplasmy present in hair shaft samples, as compared to blood and buccal samples, is heightened, but not to a level that would seriously call into question the utility of mtDNA sequencing of hair shaft samples in a forensic context.

We demonstrated that careful design of fusion PCR primers supports the creation of amplified targets ready for deep sequencing on both the Roche GS-Junior™ and Illumina® platforms. However, it became clear that this approach was not only unnecessary, but also required much more time and effort than other emerging methods that we explored. We found that using the Nextera-XT™ kit from Illumina, Inc. allowed us to directly process any double-stranded DNA, including amplicons, for deep sequencing in a much simpler and cost-effective manner.

Careful quantitative analysis of amplified DNA molecules allowed us to generate mixtures of DNA templates from different individuals in defined proportions. The deep sequencing results revealed that generally, we could detect a minor variant within a mixture at a 1% level or lower. We noticed that the well-characterized issue of homopolymeric stretches is indeed problematic when performing pyrosequencing reactions. We found that when we directly compared the pyrosequencing results to the results obtained using the Illumina®-based chemistry and instrumentation, the ease of identifying the minor variants was significantly enhanced.

During the analysis of our deep sequencing results on the Roche GS-Junior™ platform, we noticed that a set of persistent minor DNA sequence variants was present in both our blood and buccal DNA extracts. We performed a series of experiments to confirm that these are in fact nuclear insertions of human mitochondrial DNA fragments. These results were submitted for publication in the *Journal of Forensic* Sciences and are currently in the revision review stage.

The combination of an enhanced DNA extraction technique, whole genome amplification of the DNA extracts, multiplexed PCR amplification reactions around the mtDNA genome, and direct sequencing of the DNA samples on the Illumina® MiSeq™ instrument resulted in mtDNA sequence information from hair shafts that matches that found in blood and buccal extracts from the same donors. Further developmental research and validation, based on our approach and data, will result in a significant enhancement over current forensic DNA typing procedures.

**Table of Contents**

**Executive Summary**

DNA sequencing has an important and expanding role in forensic practice, both for non-human and human-based analyses. The newly emerging, often called 'next generation' DNA sequencing platforms (NGS) offer high throughput capabilities and data redundancy that ensure that high quality DNA sequencing can be a tremendous benefit to forensic science. While the forensic utility of NGS in microbial and non-human forensics is also of paramount importance, on the human side, mitochondrial DNA (mtDNA) is the obvious target of interest for these technologies.

Forensic mitochondrial DNA analysis remains a niche procedure that is practiced in a few, specialized laboratories. Although the reason(s) for this limited applicability are many, one particular limitation to forensic mtDNA analysis is the perceived inability to reliably interpret mtDNA mixtures. While there is some validity to this viewpoint as mtDNA is currently practiced, with the advent of NGS analysis, mixture deconvolution in all areas of DNA typing, including both STRs and mtDNA, is likely to be re-conceived.

There are two major advantages of the expanded amount of data offered by NGS to human mtDNA casework. These advantages can be understood as two complimentary dimensions, sequence length and combined read depth. Length refers to the amount of DNA sequence information captured for a case analysis, and depth is the degree to which the sequence is interrogated in order to identify minor variants present within a sequence.

Our analyses revealed that there are many potential sources of variation within mtDNA sequences obtained from a questioned sample or a reference sample. These sources generally fall into five categories, background noise, low-level short-lived mutational variants subject to loss via genetic drift, low-level relatively stable heteroplasmic mutations that may be either sequence or length-based, the co-amplification of nuclear pseudogenes, and fixed changes resulting from mutational events (polymorphisms). Further validation work will attempt to more fully understand the nature of these variants and lead to full implementation of these technologies into forensic casework.

Forensic samples that, by their nature, contain very little DNA, such as hair shafts, partial fingerprints, old or highly degraded bones, remain a challenge to the forensic DNA typing community. Alarge amount of effort has been placed on attempting to obtain STR profiles from these kinds of samples, the reasoning being that STR typing results are much more informative than mitochondrial DNA, and hence even a partial result would have more discriminating power than a full mtDNA analysis. However, STR analysis on these samples remains highly controversial, mainly because of the difficulty of reliably interpreting low-copy number DNA results, and the myriad of different, and sometimes conflicting, approaches that forensic practitioners have advanced in this area. (Forster, et.al. 2008; Grisedale and van Daal, 2012; Pfeifer, et.al. 2012; Benschop, et.al. 2012).

Mixture deconvolution rests on unambiguously, or at least with some statistical power, identifying the individual components of a mixture as individual entities, identifying their characteristics, so that the total number, characteristics, and relative contribution of each component of the mixture can be ascertained. Once this is accomplished, then forensic comparisons can be made between these components and reference samples.

Deep sequencing results within NGS offer hundreds or thousands of individual sequencing reactions that provide a level of information that allows for this mixture deconvolution. Ultimately, this is based on counting the number of independent runs comprising the mixture. Accordingly, the evidential sum of a particular evidentiary sample contains an added characteristic, namely, a complex collection of components that can now be considered both individually and collectively. Our results, although preliminary, show in fact that this level of mixture deconvolution is obtainable with NGS. Hence, upon full adoption of NGS in casework, mtDNA can be an analysis target for samples that may be mixed, greatly expanding its utility in the field.

Current forensic practice is to focus on the D-loop, or control region, of the human mtDNA genome. While this region contains the most population variability in the molecule, it is a small portion of the entire mt-genome.Hence, it would be desirable to expand the forensic analysis of mtDNA to the entire genome. Historically, however, this has been difficult due to the sheer amount of sequence data that would have to be generated and compared in a forensic case. Hence, forensic practitioners have continued to limit their analyses to the control region. NGS methods, however, combined with enhanced DNA extraction techniques and the possibility of pre-amplification using whole genome amplification, offer the possibility of expanding the forensic practice to include the entire mt-genome.

Expanded sequencing depth arising from next generation sequencing applications promise to offer very important advantages to forensic science. The ability to detect a minor component of mixed templates using the current Sanger method is currently about 10% on average. The inability to detect the minor components of mixtures below this threshold has led forensic analysts to interpret one base pair differences between samples as inconclusive. A method that can reach below this threshold and capture the presence of low abundance components of mixtures could significantly assist in the forensic interpretation of mtDNA sequencing results, especially in revealing common low level mixtures in both questioned and reference samples. NGS methods can also provide this advantage.

Through this project, we have developed preliminary working protocols to capture the entire mt-genome sequence at sufficient depth to identify and compare 1% variants between forensic samples such as blood, buccal scrapes and hair. Importantly, we have demonstrated that whole mt-genome information may indeed be obtained in the near future from hair shaft DNA extracts. In order to accomplish the goal of obtaining whole mtDNA genome information from hair shaft material, we employed enzymatic pre-amplification steps known as whole genome amplification, multi-plexed PCR amplification of targeted mtDNA regions, a simple enzymatic library preparation method using a transposase/integrase, followed by direct NGS of the templates.

For reference samples, we targeted rapid and efficient NGS of the whole mt-genome molecule. We employed two large, overlapping PCR fragments of approximately 8 kb and 10 kb, used the Nextera™ transposase/integrase processing step, and then loaded the products onto the Illumina MiSeq™ instrument. The results were impressive. In a few simple steps, we were able to generate high quality full mt-genome sequences from reference samples and could easily observe minor variants present in the sequence. These results have startling implications for forensic casework – namely, that whole mtDNA genome data from reference samples for comparison purposes can easily be generated using this NGS-based approach. Further, the construction of a large-scale population database to support mtDNA casework is simplified as a matter of generating large PCR amplicons, followed by simple enzymatic sample processing and then direct loading onto the NGS instrument. Using the 96 currently available indices from Illumina®, large population databases to support forensic casework that consist of deep sequence coverage can be attained relatively easily.

Much more challenging are limited forensic DNA samples, such as those from hair shaft. In this case, we had to perform experiments to increase the efficiency of each step in the process. Starting with the extraction step, we evaluated and tested a number of revisions to DNA extraction, settling on a combined Qiagen® reagent and PrepFiler™ (Life Technologies, Inc.) solid phase DNA capture method. This protocol enhanced our extraction efficiency, rendering downstream procedures more efficient as well. Next, we focused on a whole genome amplification (WGA) strategy that employed the Repli-G™ amplification kit from Qiagen®. We found slight but reproducible levels of signal enhancement using this method.

Because the human mtDNA genome has been extensively studied, conserved regions with limited levels of variable bases are known. These regions were used to place PCR primers in positions that would support amplification of the target. We are still in the process of optimizing the multiplexed reactions to cover the entire genome, but have demonstrated that the multiplex products are also easily prepared for NGS using the same Nextera™ transposase/integrase processing step that we employed for the reference samples. The mtDNA sequence obtained from the multiplexed amplicons originating from known hair donors matches the known reference sequence of these donors, providing a crucial insight into the eventual validation of this method.

A critical component of our research goal is to comparatively evaluate the ability of two leading NGS platforms, the Roche GS-Junior™ pyrosequencer and the Illumina MiSeq™ instrument, to reliably detect sequence variants arising from a DNA template. In order to accomplish this goal, we systematically prepared a number of mixed DNA samples from known donors in defined ratios and then subjected the mixtures to DNA sequencing on these platforms. Although both platforms performed well, the Illumina data proved to be much cleaner and hence easier to interpret. We show through these results that this is most likely due to the well-known limitation of pyrosequencing chemistry when encountering homopolymeric regions of the template. By performing carefully designed mixture studies using two common and popular NGS platforms, we have shown that detecting minor DNA sequence variants at approximately a 1% threshold is now obtainable, but depends on a number of quality and filtering parameters that must be understood and then properly employed. Consequently, we address a number of issues related to data processing, such as quality filtering steps and an analysis of how the depth of coverage affects the ability to detect a minor variant in a mixture, hoping that in so doing we can remove any hidden processing steps that may affect the outcome of a forensic comparison.

We prepared mixtures of DNA templates at ratios of 5%, 2%, 1%, and 0.5%. The donor mitotypes were previously characterized through DNA sequencing (Sanger sequencing) using current forensic protocols. In this way, we were able to predict where minor variants in the mixed DNA sequence were likely to be found. While the pyrosequencing reactions did not ultimately fare as well as the Illumina chemistry in our mixture

experiments, these results did lead us todetect, in some instances, the presence of nuclear DNA-encoded mtDNA insertions, known as Numts (pronounced 'new-mites').

As mentioned above, in addition to expected minor variants arising from deliberate mixtures, a subset of read-clustered, unexpected variants was also detected in reads originating from nuclear DNA-rich blood and buccal cell samples. Our analysis showed that these variants are amplified products from a nuclear pseudogene, an ancient nuclear insertion of mitochondrial DNA that contains our control region primer-binding sites. These NumtS sequences, to our knowledge, have never been co-detected in mitochondrial sample preparations using Sanger sequencing, further supporting the assertion that the sensitivity of low-level variant detection reported herein is a vast improvement over current methodologies. Additionally, the data show slight differences in the amounts at which the unexpected variants are detected in whole blood, buccal, and hair samples. Because amplification of the NumtS using either an mtDNA control region or a NumtS specific primer set requires a relatively high amount of input template, this difference is likely due mostly to the total amount of DNA present, and the ratio of nuclear to mitochondrial DNA present in the sample

From a forensic perspective, the presence or absence of the NumtS sequence in a sample should accordingly vary with a host of factors, including the amount of sample DNA amplified and the individual's genotype with respect to the NumtS itself, and hence not be expected *a priori* to be present in each analysis. Consequently, until a fuller characterization of the genetic variation within each NumtS is elucidated, we recommend that at this time no direct comparative utility be made of the NumtS insertion itself, but the presence of these inserted elements can enhance the ease of interpreting casework-derived data. In other words, knowing that these sequences may be present as low-level variants using NGS typing procedures is beneficial, since they can be properly considered as elements that are not derived from extraneous contamination nor necessarily present in each sample under comparison.

We have developed a protocol to more efficiently extract mtDNA from hair shaft samples. Our goal was to maximize the mtDNA extracted from two centimeters of hair shaft so that more mt-genome sequence information may be obtained from a challenging sample, leading to a higher discriminatory power of mtDNA analysis. The optimized extraction method shows, on average, a fourteen-fold increase in the amount of mtDNA recovered when compared with traditional manual grinding and organic extraction methods. The optimized method is also less time-consuming, and fewer hands-on steps and tube transfers are required, which reduces the risk of contamination. PCR inhibitors are successfully removed as indicated from qPCR internal controls and subsequent amplification success. The protocol is robust, and has been shown to be effective in multiple analysts' hands.

In furtherance of our goal of generating whole mt-genome information from challenging forensic samples, we reasoned that a multiplexed amplification scheme, combined with the optimized DNA extraction protocol, could potentially yield large amounts of mtDNA sequence data from hair shafts. Using published primer sequences obtained from the Life Technologies/Applied Biosystems, Inc. MitoSeqR™ kit, we designed a series of human mt-specific primers to span the entire genome. We have shown that indeed the extracted DNA from hair shaft can serve as a robust template for multiplexed amplification. When we then processed this amplified material with newly developed tagmentation technology from Illumina Nextera XT®, a library preparation technique designed for Illumina® sequencing platforms, the resulting DNA sequence matched the expected reference type obtained separately from the same donors.

In a separate set of experiments designed to deal with robust samples that are commonly used as reference templates, such as blood and buccal samples, we were able to generate whole mt-genome sequencing data using a long PCR amplification technique with two overlapping primer sets that cover the entire mt-genome. When processed with Illumina® Nextera XT®, we found it relatively simple to generate high-throughput and deep coverage sequence data from these samples. The DNA sequencing data can be quickly obtained from the instrument software and the mtDNA sequence of the amplified sample is then

obtained. These data can be further analyzed with second-party online freeware using a custom analysis pipeline, and subsequently viewed in a genome browser. We have investigated a number of different software packages that will allow the forensic science community to easily navigate through the large amount of data presented by these instruments and generate timely reports that will enhance casework productivity.

We have determined that the informatics issues related to these technologies are substantial. There are many secondary analysis software packages available that allow the analyst to view and interpret NGS data. These packages have an impressive array of capabilities, however, many of these capabilities do not pertain to forensic analysis, and many are hidden from the view of the user. With some commercial software packages, the analyst has the ability to adjust the quality-filtering parameters, and re-queue the data for analysis. A built-in variant comparison tool present in many packages allows multiple files to be pulled into the software and directly compared. Additionally, the analyst has the option to view histogram reports, which show distribution of coverage across the length of the reference sequence, showing the starting point of both forward and reverse reads, and the average read length. These capabilities allow for a rapid and straightforward assessment of the impact that changes to analytical parameters can have on data interpretation.

Further work is warranted in a number of areas related to NGS sequencing in support of forensic casework, including further protocol development, quality-filtering, software package evaluation, advanced mixture studies, validation, and rapid population database creation of the whole mt-genome to support casework analyses. We believe a well coordinated effort in this area will result in a significant advance in the area of forensic DNA analysis, and have implications well beyond human DNA, including microbial forensics and metagenomic analyses.

**Introduction and Statement of the Problem**

The major strength of human mtDNA analysis in forensics is its sensitivity. Alternatively, the major weaknesses are its lack of informativeness (as compared to STR analysis) and the interpretational complications that arise from heteroplasmy. Therefore, in the context of forensic practice, human mitochondrial DNA analysis is currently limited to the kinds of samples that will not routinely work with STR analysis. In order to address both of these limitations, additional information needs to be gleaned from the mtDNA molecule. Two distinct dimensions of information available from mtDNA are the amount of sequence data that is analyzed, and the extent to which that information is analyzed. That is, additional information can be obtained by increasing the amount of sequence data obtained as well as interrogating each base pair in order to determine whether or not minor variants arising from mutations are present at very low levels.

Using emerging technologies, an extension of the breadth of sequence data obtained can now extend to the entirety of the human mtDNA genome. The current focus of forensic laboratories is the control region, consisting of about 1123 base pairs, or a portion thereof. Extension in this dimension would require that amplification reactions are designed and validated to capture all the variation within the genome. Once generated, the amplified material would require sequence analysis of significantly more data compared to what is currently analyzed. Although this can be accomplished with existing DNA sequencing methodologies, the time and effort required to generate and analyze these data is problematic. Because they were designed to generate large amounts of data, newly emerging deep sequencing technologies can provide a remedy to this issue.

With respect to information in the complementary dimension (depth) the design of newly emerging methods are also well suited. In effect, using these methods, each molecule generated during the amplification phase is independently sequenced and reported as a separate, individual item of information. This means that if the original template is mixed, the components of the mixture will each be independently analyzed and presented to the analyst.

Polymorphism and mutation are two-edged swords in forensic mtDNA analysis. Polymorphisms arise from mutations, and hence are advantageous to those seeking informative loci for typing purposes, such as forensic examiners. However, a high mutation rate at a locus also raises the probability of observing mutations in the process of segregation. The proper scientific approach to such a situation is to carefully study the mechanism(s) of segregation within individuals, collect and record as many instances as possible, and then develop interpretational guidelines based on these observations. Match criteria can then be stated in terms that support the underlying observations. In addition, mixture studies should also be employed that create the mixed template that is then studied. The emerging interpretational guidelines will, of course, be dependent on the particular technology employed. If the ability to detect variants within a mixture changes, then the interpretational approach should also change.

Due to the activity of fast changing sites (hot spots) within the mtDNA genome, subtle sequence variants are often observed between cells or tissues within an individual. This observation is called heteroplasmy. Rather than being viewed as an anomaly, heteroplasmy is actually a principle of mitochondrial DNA genetics. If we desire to use such a locus for forensic purposes, our conception of what constitutes a match should be widened to consider the possibility of observing heteroplasmy in casework. Accordingly, interpretational guidelines may require expansion to reflect the additional sensitivity of NGS methods.

Sequencing depth and accuracy arising from next generation sequencing applications have a very important advantage that is specific to mtDNA analysis. The ability to detect the minor component of mixtures using the Sanger method is currently about 10% on average. The inability to detect the minor components of mixtures below this threshold has lead forensic analysts to interpret one base pair differences between samples as inconclusive. A method that can reach below this threshold and capture the presence of low abundance components of mixtures could greatly assist in the forensic interpretation of mtDNA sequencing results, especially in revealing common low level mixtures in both questioned and reference samples. Because of deep sequencing's ability to now detect low-level mixtures at greatly reduced levels, it is imperative that the forensic community now begin to implement such improvements and potentially revise of our current interpretational approach to casework comparisons.

**Review of Relevant Literature**

The detection of genetic variation at the DNA level that underlies DNA profiling for individual identification has been developed during the last two decades. Today, numerous PCR-based DNA typing tests are in use for identification purpose in the analysis of biological evidence samples. PCR-based DNA typing kits targeting the nuclear genome, (e.g. Profiler$^{TM}$ and Identifiler$^{TM}$) are particularly useful for individual identification because of their sensitivity and high discrimination power. However, in some cases the analysis of genomic DNA fails because the limited or degraded template. In these cases, polymorphisms within the mitochondrial genome can serve as a useful target.

Mitochondrial DNA (mtDNA) found in the organelle, is haploid in nature. The complete DNA sequence of the human mitochondrial genome was determined in 1981, and hundreds of sequences have since been determined. Mitochondrial DNA is a small, circular molecule of about 16,569 bp. The control region (or D-loop region) of mtDNA is an approximate 1123 bp region of noncoding DNA that contains one origin of replication and both origins of transcription as well as additional transcription and replication control elements. Mitochondrial DNA is highly polymorphic with the majority of the sequence variability concentrated in the control region, specifically, hyper variable regions (HVR) HVR-I, HVR-II and HVR-III. The HVR-I (16024 to 16365), HVR-II (73 to 438) and HVR-III (438 to 574) positions are typically targeted for forensic identification purposes because of the high density of sequence variation.

Mitochondrial DNA has two additional unique features that make it particularly suitable for the analysis of biological remains, e.g. hair, bone, blood, teeth and extremely limited or degraded DNA samples. First, mtDNA is inherited matrilineally. This mode of inheritance makes it a valuable genetic marker for the investigation and identification of missing person cases because the subject's mother and siblings, as well as the mother's siblings (uncles and aunts) will all carry the same mtDNA sequence as that of the subject in question. Consequently, samples from maternally related individuals can be used as reference samples for the missing person. Second unique feature of the mitochondrial genome is that it is present in high copy number. Alleles of the nuclear genes typed by the existing PCR-based tests are present in only one (spermatozoa and ova) or two copies per cell, whereas mtDNA sequences can be present hundred thousand times per cell.

Due to the presence of sites with high mutation rates within the mtDNA genome, subtle sequence variants are often observed between cells or tissues within an individual. This observation is called heteroplasmy. Operationally defined, heteroplasmy is the presence of more than a single mtDNA sequence within an individual's body or within a sample obtained from an individual. Rather than being viewed as an anomaly, heteroplasmy is actually a principle of mitochondrial DNA genetics. In order to use a highly changing locus for forensic purposes, our conception of what constitutes a match has been widened to consider the possibility of observing mixtures arising from heteroplasmy in case work. Accordingly, interpretational guidelines have been developed that are cognizant of these facts.

A wealth of recent publications have revealed not only the patterns of human mtDNA variation within and between tissues, but have also shown that cancer cells harbor a set of unusual mtDNA variants that have been the subject of intense study as potential cancer diagnostic targets (see section in Bibliography entitled "New Developments in Cancer Diagnostics and Human Mitochondrial DNA Variation"). These studies are beginning to reveal patterns in the cellular and tissue segregation of mtDNA variants. Although extremely interesting from a basic scientific perspective, the forensic relevance is limited to the question of how the forensic analyst is to properly interpret patterns of variation reveled in those sample types commonly investigated in forensic casework, such as bones, hairs, buccal scrapes, and blood samples.

A particularly relevant article that has recently appeared in this regard is He *et. al.*, Nature advance online publication 3 March 2010 | doi:10.1038/nature088022010. Using deep sequencing methods, these investigators found widespread heteroplasmy in normal human cells. Many of these low-level heteroplasmic sites were located at positions of known polymorphisms in the mtDNA genome. For example, sites 16,126; 60; 72; 94; 189 and 228 in the control region exhibited heteroplasmy at levels between 1.5 – 5% compared to the dominant type. Most of these sites have been observed as heteroplasmic in forensic casework, but at these levels the mixed nature of the profile would be missed using current technology. Other, more complex mixtures were noted in a variety of different sites but were restricted to cancer cells. Although not unexpected, these results confirm that individuals comprise a complex mixture of related mitochondrial genotypes rather than a single genotype. The authors point out that thus an individual, and perhaps even a single cell, does not have a single mtDNA genotype. Instead, tissues have a mixture of genotypes, a few of which may be maternally inherited and the remaining ones the result of somatic mutations.

Although these authors do not appear to have reviewed the amount of previous work that has gone into forensic assessment of both sequence and length heteroplasmy in human mtDNA, they suggest caution in excluding identity on the basis of a single or small number of mismatched base pairs when the tissue in evidence is not the same as the reference tissue of the suspect. Based on these published results, Forensic magazine, in the March 12, 2010 issue, made the following statement: "This new revelation is sure to lead to a reevaluation of forensic uses of mitochondrial DNA in identifying suspects, with the study recommending that only samples from the same tissue be compared." This is a misrepresentation of what the study actually said. As noted above, the authors suggest caution in interpretations of exclusion based on a single or small number of apparent differences between a questioned sample and a known sample, especially when they derive from different tissues.

Regardless of the misrepresentation, the forensic community should take note of these findings. In order to stay ahead of this issue scientifically, it is crucial that the forensic community evaluate deep sequencing methods for patterns of variation that can only be revealed by these newly emerging methods. Moreover, it is crucial that these studies be conducted in a manner that is consistent with current casework, for example, by using existing forensic protocols and focusing on those types of samples that are commonly encountered in casework. For instance, in this study we will carefully evaluate patterns seen in hair evidence compared with blood and buccal known reference samples.

**Next-Generation DNA Sequencing as a Potential Tool in Forensic DNA Casework**

The possibilities offered by next generation sequencing (NGS) platforms are revolutionizing biotechnological laboratories. Over the past five to seven years, large-scale sequencing has been realized by the development of several so-called next-generation sequencing (NGS) technologies. These technologies provide an unprecedented tool for numerous biological applications (Mardis 2008; Rokas and Abbot 2009; Wang, Gerstein *et. al.* 2009). Although each chemistry and accompanying instrument varies, the output from an NGS run can exceed several gigabases of sequence data. These technologies are increasingly used for various nucleic acid sequencing-related applications. Several potential artifacts, including read errors (base calling errors and small insertions/deletions), poor quality reads and primer or adaptor contamination can occur in the NGS data, which can impose significant impact on the downstream sequence processing/analysis. For forensic applications, full validation of NGS requires a thorough understanding of these potential sources of interpretational error. However, such potential errors must be viewed within the context of the meaning of error in casework applications, and placed into the wider perspective of assessing the potential of actually mistyping a sample when reasonable and validated interpretational procedures are in place.

High quality data is very important for various downstream analyses, such as sequence assembly, single nucleotide polymorphisms identification and gene expression studies. Therefore, these sequence artifacts need to be removed before downstream analyses, otherwise they may lead to erroneous conclusions. Many of the programs available for downstream analyses do not provide a flexible means for quality checking and filtering of NGS data before downstream processing. Hence it is advisable to assess the affects of quality filtering of sequencing data at the end-user level.

A few bioinformatics pipelines with different features have been developed for the quality control of NGS data (Martinez-Alcantara, Ballesteros *et. al.* 2009; Blankenberg, Gordon *et. al.* 2010; Cox, Peterson *et. al.* 2010; Schmieder, Lim *et. al.* 2010; Schmieder and Edwards 2011). Many of these are specific for a particular sequencing platform and hence reflect the known limitation(s) of these platforms. Hence, there is still a need for the development of better quality filtering tools with additional/better features. At present, changing the analysis parameters and comparing the final outputs is not straightforward in some applications, and hence thoroughly understanding the affects of identifying the critical parameters and how they may impact an interpretation needs further development.

NGS technologies are not the same. For instance, two NGS technologies, PacBio RS® (Pacific Biosciences) and the GS FLX Titanium® (Roche) have equal or greater read lengths than Sanger sequencing (Li, Victoria *et. al.* 2010; Carneiro, Russ *et. al.* 2012). In contrast, the Ion Torrent®, SOLiD® (Life Technologies) and Illumina-based NGS technologies generally yield shorter read lengths when compared to Sanger sequencing. Despite these differences, these technologies have greatly facilitated genome sequencing for both prokaryotic and eukaryotic genomes. Along with the development of highly parallel and robotic chemistries, this advance was possible due to a concomitant development of software that allows for the *de novo* assembly of draft genomes from large numbers of short reads. In addition, NGS is used in metagenomics studies for the detection of sequence variations within individual genomes, e.g., single-

nucleotide polymorphisms (SNPs), insertions/deletions (indels), or structural variants (DePristo, Banks *et. al.* 2011).

The analysis of the sequence data from NGS runs is commonly referred to as a pipeline because it involves a series of related sequential steps of analysis. More generally, a bioinformatics workflow management system (pipeline) is a specialized form of workflow management system designed to execute a series of computational or data manipulation steps. Many different kinds of workflow systems exist, and analysis pipelines have been created for scientists from many different disciplines. Almost all of these systems are presented in an abstract representation of how a computation proceeds in the form of a directed graph, where each node represents a task to be executed and edges represent either data flow or execution dependencies between different tasks (see Figure 20 for an example). The systems typically provide a visual front-end allowing the user to build and modify complex applications with little or no programming expertise.

From a forensic validation perspective, if the quality of the analysis may be affected by changing the parameter in question, then these elements in the pipeline should be tested and understood. For instance, if finding rare variants in a mixture is the goal, all the relevant parameters within the pipeline that can significantly alter the final output file and potentially lead to the identification or misidentification of the variant should be validated for the stated purpose. It may be desirable to conduct a coordinated analysis of the data by deliberately altering a number of these parameters and observing the effect(s) on the final result. This will give statistical rigor to the interpretation as well as indicate which parameters are important variables.

There are many steps in the analysis pipeline that contain parameters that can be adjusted that will affect the final set of sequence data collection. It should be noted that in the context of forensic investigation, the ideal would be to employ a specific analysis pipeline based on best practices identified through validation, but to always retain the raw sequence reads in case other analyses using modified parameters are warranted. In this way, nothing is lost from the original run, and the interpretation can benefit from using all of the data, albeit in slightly different forms. For instance, the choice of which reads to retain in an analysis and which reads to discard may significantly impact the final interpretation of the comparison, and hence retaining, as well as trimming, reads is an important consideration that warrants careful consideration.

As has been the case with earlier technologies, forensic validation of NGS data utilization would benefit from the development of a standard set of run conditions and analyses. This allows multiple users to compare the performance of a protocol in their laboratory to others in the same field. Further, the adoption of a common template (e.g. a commonly used human cell-line control) that could be adopted and used for testing of all platforms would be advantageous. The National Institute for Standards and Technology (NIST) currently provides some templates for this purpose. Results from the analyses of these templates could then be used to directly compare different NGS platforms, chemistries and software upgrades (Glenn, 2011). For instance, in their comparison of different versions of the Ion Torrent chip technologies, moving from the 314 to the 316 version, some investigators (Loman, Misra *et. al.* 2012) created an assembly from a sample which they had used in earlier analyses of the original Ion Torrent 314 chip. They found that the newer chip resulted in an assembly of this same genome that contained fewer than 400 contigs, whereas the original analysis returned over 3,000. As their purpose was high quality assembly, this template served as an important quality control standard.

**Research Objectives**

The objective of this research effort was to evaluate a number of aspects of a new forensic DNA analysis method focused on emerging, powerful DNA analysis techniques. Specifically, we sought to investigate the following features of next generation DNA sequencing (NGS) in the context of forensic science:

1. Sample Preparation Methods –We evaluated the use of fusion primers to direct library preparation methods. We also explored an enzymatic preparation method, Nextera and Nextera XT, and compared the ease of use with that of fusion primers. We further developed a long PCR amplification strategy for reference samples and showed that the enzymatic sample preparation strategy worked well on these sample types.
2. Whole Genome Amplification Methods – We conducted a limited evaluation of some commercial WGA kits for sample pre-amplification prior to targeted PCR.
3. Chemistries - Directly compared two popular NGS DNA sequencing chemistries. We directly compared the quality of DNA sequencing results from the Roche GS-Junior and the Illumina platforms.
4. Limits of Detection – We investigated the limits of detecting low-level variants using the procedure that we determined to have the optimal performance characteristics of those that we investigated. Using carefully prepared mixtures, the Illumina-based chemistry was so evaluated.
5. Expanded mtDNA genome coverage - employed the use of multiplex amplification strategies to target whole mt-genome data.
6. Software - Evaluated a number of NGS-related software packages and algorithms.
7. Quality Metrics - Conducted substantial research on quality metrics and variables.

Although it was not part of the original research plan, we also confirmed the power of NGS to detect genetic variants arising from nuclear pseudogenes.

**Research Proposal and Methodologies**

Using Next Generation Sequencing technologies, we proposed to examine amplified mtDNA from separate human hair extracts and compare these results to reference samples consisting of both buccal and blood samples from individual donors. The amplicons were deep sequenced using both the Roche/454 pyrosequencing and Illumina bridge-sequencing technologies. These studies provided a direct comparison between these two technologies on the same forensic samples, and also indicated the amount of sequence heteroplasmy present in hair samples that are not generally observed using current methods.

In the initial approach, DNA templates from different individuals were amplified using specially-designed addressing primers. Addressing primers, also called barcoding primers, are comprised of three regions: a 3' target-complementary sequence that binds to the target, a downstream recognition sequence on the 5' end of the addressing primer, and a smaller 'barcoded' sequence that falls between the target-specific sequence and the downstream recognition sequence. Like other PCR protocols, the entire addressing primers are incorporated into the PCR product. This incorporation allows for further downstream processing steps, and also allows the amplified target to bind to complementary sequences in downstream-specific deep sequencing steps.

A barcoded sequence (also called a multiplex identifier, or MID) is a short DNA sequence located just upstream (3') of the general recognition sequence. Hence, when the amplified target is sequenced, this barcode sequence is read out as DNA sequence during the downstream sequencing run. In this way, the user can co-analyze a mixture of many different targets generated from different individuals within the same run, drastically saving reagent costs and time. We proposed to examine the utility of this approach within the scope of forensic DNA analysis.

Amplification was assessed using an Agilent 2100 Bioanalyzer™. Each amplicon was quantitated and then normalized prior to the sequence run. This important normalization step ensures that each region of the molecule will generate approximately the same number of deep sequencing fragments. Following the sequence run, parsing of the sequences was accomplished using the addressing information present within the modified primers. We used addressing primers to parse out the individuals from the combined run. In this

manner, sequence data from individuals was collated from a large data set arising from a single deep sequencing run.

For the Illumina-based sequence generation, the protocol to generate the templates for sequencing was slightly different. We employed the Illumina® Nextera® transposase (also called tagmentation) approach to these templates. The details of this experiment are described below.

We generated a series of mixtures of two distinct templates at the 99.5 / 0.5%, 99% / 1%, the 98% / 2% level, and the 95 % / 5% levels. These mixtures were sequenced using both the Roche/454 GS-Junior™ pyrosequencing method and the Illumina® sequencing-by-synthesis methods. We performed detailed comparisons on the ability of each method to detect the minor component at each position of difference. These analyses included the detection of both single nucleotide polymorphisms (SNPs) and homomeric length variants, but given the complexity of indel assignment using pyrosequencing and the limited utility of length variants in forensic mtDNA casework, we chose to focus our analysis on SNP variation.

**Roche/454 Pyrosequencing**

We have purchased a Roche/454 GS-Junior™ DNA sequencer. This instrument is a parallel pyrosequencing system capable of sequencing roughly 35 million bases of raw DNA per 10-hour run. The system relies on sequence analysis of adapter-ligated DNA fragments to small DNA-capture beads in a water-in-oil emulsion. We performed data analysis using bioinformatics tools that are supplied by Roche and are designed to support highly parallel resequencing applications. Polymorphisms were assessed by automated comparison to the known Revised Cambridge Reference Sequence (rCRS), which was loaded into the program as a reference text file.

**Illumina-based Sequencing**

We have partnered with Illumina, Inc., 9885 Towne Centre Drive, San Diego, CA to evaluate the Illumina® sequencing technology for forensic applications. We directly compared the results from the Illumina® runs to that obtained from the Roche/454 GS-Junior™ instrument. While we initially purchased an Illumina® 2e instrument that was subsequently upgraded to a 2x, the bulk of our analyses were conducted on a MiSeq® instrument provided by Illumina, Inc. to foster future research collaborations.

Illumina® sequencing technology relies on the attachment of randomly fragmented genomic DNA to a planar, optically transparent surface. The attached DNA fragments are extended and bridge-amplified to create a sequencing flow cell with hundreds of millions of clusters, each containing ~1,000 copies of the same template. These templates are sequenced using a four-color DNA sequencing-by-synthesis technology that employs reversible terminators with removable fluorescent dyes. This approach ensures high accuracy and may eliminate sequence-context specific errors such as homopolymers and related sequences found in a few locations within the human mtDNA genome.

**Hair Samples and Control region Analyses**

In order to assess the level of heteroplasmic mixtures in hair samples, we evaluated deep sequencing on hair samples taken from different parts of the head from IRB-approved donors. Initially, we employed the current FBI hair extraction protocol to this effort. However, we also examined many other DNA extraction procedures to possibly improve the extraction efficiency of DNA from hair shafts. In our opinion, the forensic DNA community could greatly benefit from the regular employment of a challenging DNA source, such as hair shaft, in furtherance of improved protocols for low-level sample types. Hence we regularly employed this standardized sample approach to our efforts.

We analyzed the DNA from the amplified hair extracts using deep sequencing and compared these results to the same amplified targets in blood and buccal samples. Control region primers were used so that the portion of the mtDNA genome utilized in forensic casework was the same as in current practice. Standard forensic protocols for buccal and blood samples were also used.

Using capabilities currently in our laboratory, studies were designed that enabled us to identify the amount of amplified DNA to use in the deep sequencing reactions. We routinely utilized the Agilent 2100 Bioanalyzer™ (equipment used in many forensic DNA laboratories) to quantify the amount of labeled or unlabeled amplicon obtained from each PCR reaction. A series of dilutions of these templates were the subjected to deep sequencing, and the depth of coverage was noted and correlated to the Agilent quantification results. This allowed us to predict the number of data points obtained from both the Roche/454 GS-Junior™ and Illumina® platforms from a given amount of amplified DNA as quantitatively assessed using the Agilent instrument. These results were crucial to an assessment of the ability of deep sequencing techniques to reliably detect minor variants from forensic mtDNA targets.

**Whole Genome Amplification (WGA)**

Over the past few years several methods have been developed for the Whole Genome Amplification (WGA) of DNA. Through the use of these protocols, it is now possible to increase the amount of template DNA in extracts containing a small amount of DNA for use in downstream PCR analyses. For a WGA approach to work, it must be able to faithfully amplify the entire DNA sample without introducing errors into the DNA sequence that may confuse the analysis. The best WGA methods in this regard are based upon the multiple displacement amplification (MDA) chemistry (Dean, et.al. 2002).

MDA relies on priming the genomic DNA with exonuclease-resistant random hexamers and the use of phi-29 DNA polymerase (Dean, et.al. 2002). Phi-29 DNA polymerase is a highly processive, strand displacing polymerase and has a very low error rate estimated at about 1 in $10^6$-$10^7$ nucleotides (Esteban, et. al, 1993). This can be contrasted to an estimated error rate of 3 in $10^4$ for native *Taq* DNA polymerase (Eckert, et.al. 1991) and 1.6 in $10^6$ for Pfu polymerase, enzymes that are commonly used in PCR.

A number of investigators have evaluated WGA in the context of forensic DNA analysis. Generally, the results have been mixed using a variety of approaches, with some studies indicating limited success (Ballantyne, et.al. 2007; Maciejewska, et.al. 2013; Lee, et.al. 2012), while others showed no improvement in typing success (Barber and Foran, 2006) while still others have showed internally mixed results (Sun, et.al. 2005).

In our approach, we utilized two valuable tools to assess the success at each step of the process. First, we have adopted a human mtDNA-specific real time PCR amplification assay (Kavlick, et al) that allows us to carefully estimate the number of copies of human mtDNA that are present in any sample. We have fully implemented this assay into our analysis stream, and use it routinely to quantitate the efficiency of our extraction and amplification protocols. As a post-extraction tool, the assay shows us the relative efficiency of different DNA extraction methods. We have found that the current FBI hair extraction protocol generates between 2,000 and 150,000 copies of human mtDNA, depending on the nature of the extracted material itself. Once extracted, the same real-time PCR assay can also be used after WGA pre-amplification of the DNA extracts. We are evaluating three separate WGA pre-amplification kits for this purpose.

The second critical tool at our disposal is the Agilent™ 2100 Bioanalyzer™. This instrument gives both qualitative and quantitative information to the user on the nature of the DNA present in a sample. We have recently demonstrated that WGA pre-amplified material appears as an amorphous distribution of DNA fragments as expected. The Agilent™ 2100 Bioanalyzer™ provides us with the ability to assess whether or

not the targeted PCR reactions have generated the expected sized fragments and the amount of amplified DNA present following the amplification reaction.

We have further shown that these WGA pre-amplified fragments, when processed using the Nextera™ processing protocol, reduces these fragments into a smaller subset of sizes, as expected, which are then ready for downstream DNA sequencing on the respective platforms available. The requisite recognition sequences that support the appropriate platform and also identify the particular sample must be incorporated during the Nextera™ enzymatic reactions. We have demonstrated that this can be accomplished by careful bioinformatic design of the transposase/oligonucleotide complexes.

**Whole mtDNA genome analyses**

The use of the entire mtDNA genome in forensic analyses will serve to increase the power of forensic mitochondrial DNA analysis and hence the applicability of mtDNA sequence analysis to forensic science. Moreover, the addition of a few samples where we conducted whole genome analyses allowed us to extend our evaluation from the control region to the whole mtDNA genome. This also allowed us to evaluate the relative cost of deep sequencing the whole mt-genome and identify potential technical and practical issues associated with this approach.

On a subset of reference samples, we investigated the use of Illumina®-based DNA sequencing on whole-genome mtDNA targets. This allowed us to assess the level of heteroplasmy present throughout the entire genome, similar to the analysis presented in the recent He *et. al.* paper in *Nature* (see Bibliography). This portion of the project added valuable information to the growing body of knowledge supporting the prospect of extending forensic mtDNA analysis to the whole genome, a promise now achievable by deep sequencing techniques.

In one approach to our goal of generating whole mt-genome information from challenging forensic samples, we reasoned that a multiplexed amplification scheme could potentially be beneficial. Using the published primer sequenced obtained from the Life Technologies/Applied Biosystems, Inc. MitoSeqR™ kit, we designed a series of human mt-specific primers to span the entire genome. These primers lacked the M13 tails that are included in the MitoSeqR™ kit.

We evaluated combinations of primer sets in single and multiplexed amplification reactions that collectively covered the entire mtDNA genome. For this effort, we focused on the Illumina® technology because it allowed us to employ the Nextera® tagmentation reaction, greatly simplifying sample preparation. Again, the efficacies of these amplification reactions were assessed using an Agilent 2100 Bioanalyzer™.

Utilizing the Applied Biosystems, Inc. (ABI) 3130xl instrument in our laboratory, our mtDNA sequencing results were compared to the whole genome sequencing results obtained using commercially available kits and conventional Sanger chemistry. Using the results obtained from the commercially-available ABI mitoSEQr™ kit, we were able to perform whole genome sequence analysis of the reference samples that accompany the forensic DNA extracts. In this way we generated information using traditional Sanger methodology from which to compare the deep sequencing results. Accordingly, carefully constructed mixtures that simulate heteroplasmic samples were evaluated using traditional Sanger sequencing techniques as well. In some cases we obtained Sanger data from these same mixtures in order to directly show the limit of detection of the Sanger approach and compare it to the results from the same DNA templates using NGS-based studies.

In the commonly employed protocol for sample preparation, fragment libraries are generated by shearing larger amplicons, followed by blunt-end ligation of linkers containing specific recognition sequences. The method requires the generation of large amplicons, typically in the range of over 3 kb. Although we

have successfully amplified the entire mtDNA genome in a single PCR reaction, we have shown that these large amplicons can be processed using Nextera® chemistry and directly sequenced on the Illumina® platform. Parallel Sanger-based sequencing of these large amplicons generated in our laboratory confirmed the target-specificity of the assays. These results have startling implications for forensic casework – namely, that whole mtDNA genome data from reference samples for comparison purposes can now be generated using this NGS-based approach. Further, the construction of a large-scale population database to support mtDNA casework is simplified as a matter of generating large PCR amplicons, followed by a simple enzymatic sample processing step and then direct loading onto the NGS instrument. Using the 96 currently available indices from Illumina®, large population databases consisting of deep coverage can be obtained relatively easily.

**NGS Analysis Stream and Bioinformatics**

The analysis of NGS data begins with raw, unprocessed files arising from the instrument run. In the case of the Roche 454 GS-Junior™ instrument, these files are in the .sff file format. The Illumina® instrument outputs different kinds of files, in this case .csa files (or their equivalent). In most current applications, a file converter is used to convert these instrument files into what are called .fastq files. Individual FASTQ (pronounced FASTQ) files (often called 'reads') contain four distinct segments: a unique identifier, meta-data information associated with the run and the read, the DNA sequence itself, and finally a quality indicator for each nucleotide in the DNA sequence. The means by which the individual quality scores are assigned to each base are distinct for each chemistry and instrument, and in some cases are not intuitive to the end-user.

Once the FASTQ files are obtained, a number of filtering and statistical parameters can be applied. The user can, at this stage, decide whether he or she wishes to use all of the data, or a sub-set of the data. This is called quality filtering. In most cases, after quality-filtering the FASTQ files, a sub-set of the reads is retained for further use. Usually, filtering includes removing sequences of poor quality (or containing poor quality segments), insufficient length, or any other parameter deemed important to the particular application. For instance, a user may wish to only use the highest quality 1,000 filtered reads, and hence a command to retain only this many reads might be invoked. More often, a threshold quality score is applied, such as Q25, and those reads that fall under this threshold are discarded.

The next step is to separate the reads into categories. This is often termed 'parsing' or 'demultiplexing.' In a very popular application, users incorporate a specific sequence, called a Multiplex Identifier (MID), or barcode, which is usually about six base pairs in length, into the DNA sequence adjacent to the segment that binds to the flowcell or bead, depending on the chemistry employed. In one particular approach, this barcode sequence can be located internally within the complex amplification primer (sometimes called a fusion primer since it contains a fusion of sequences with distinct purposes). The MID sequence provides an identifying code that has a particular meaning to the user. For instance, the MID can identify an individual within a large DNA sequencing experiment containing hundreds of different amplification targets, or it can identify a part of a complex experiment, for instance, a particular mixture of DNA templates.

Another filtering target might be a template-specific sequence arising from a conserved region within an amplicon. For example, in human mtDNA amplicons, known regions of conserved sequence within the amplified region can be used to filter complex runs into its constituent amplicons. For instance, a known, highly conserved segment of ten base pairs within HV1a can be used to separate the HV1a amplicons from the others. After performing all of the requisite filtering, at this stage the user has a series of quality and sequence-filtered reads, each now separated as a series of related text files, each arising from a single amplification reaction.

Following quality filtering, the next step is to align the reads to a reference sequence. In the case of human mtDNA, the revised Cambridge Reference Sequence (rCRS) is used for this purpose (Anderson, *et. al.*, 1981). There are a variety of alignment programs capable of performing this task. The resulting alignment file also can be evaluated statistically, and these statistics will assist the examiner in evaluating the quality of the run as exemplified in the alignment to the reference. Fortunately, in human mtDNA analysis, the reference sequence is almost identical to most of the read sequences and there are no complex repeats, and hence this step is relatively straightforward. However, some complications can arise with small gaps in the alignments, especially with 454 data, as discussed below.

The alignment of a series of individual reads is often called a contig, or a scaffold. The use of a reference sequence in this step greatly facilitates the alignment. In the case of amplified DNA, in most cases there will be a single contig for each amplification target, for instance, HV1a, HV1b, HV2a, and HV2b are common targets in human forensic mtDNA testing, and hence each region would contain its own contig showing the particular alignment to a specific region of the rCRS.

Because the goal of amplicon resequencing in this project is to evaluate the potential for mixed bases at particular positions, the alignment must be evaluated at each position in order to determine whether or not, and to what extent, a particular base position is mixed. An algorithm called pileup counts up the total number of character states (A,C,G,T, gap) at each position within an alignment and provides a report at each position. Pileup is designed to work on .bam alignment files (which are binary files), and hence any other kind of alignment file must generally be converted to a .bam file to perform the pileup analysis.

The output of the pileup analysis is currently relatively crude. A GUI-based viewer that would allow the user to graphically evaluate the pileup results, including all the information underlying the analysis, would be beneficial. One idea is to color code the consensus sequence by the degree of homogeneity in the runs comprising the base call. For example, those bases with greater than 99.99% homogeneity would be black, 99.9% red, 99% green, etc. The coloring scheme would allow for easy visual scanning through the consensus sequence for colored bases and identify those that warrant a more careful investigation. Further, a tool that directly compared the color distributions between two samples, for instance a known and questioned item from a case, would help the examiner quickly compare the sequences and properly interpret the comparison.

## Quality Control of NGS data

### Noise

The sheer number of reads that are produced using NGS methods inevitably results in some incongruencies within the large data set. Understanding the general level of these inconsistencies should lead the investigator to apply filters that remove them without compromising the final analysis. For instance, a partial sequence arising from a primer-dimer might be retained and aligned to the reference. A full analysis of the exact cause of each incongruency is not necessary, however. Rather, validation studies can be designed and implemented that seek to identify these variants in an effort to design filters that can effectively remove them from further consideration.

A much more important consideration is the level of background noise and how one goes about identifying it as noise. This is addressed empirically, by looking carefully at the pileup results at each position for any patterns that may emerge at specific positions of the sequence. There may be, for instance, positions that characteristically show elevated background compared to others. This may be a function of the chemistry employed, or may be related to the quality of the run itself, or may result from other unknown processes. The presence of these low-level variants should not, in the end, affect the ability of a forensic examiner to properly interpret a comparison that arises from NGS. For instance, it is well known that current

Sanger sequencing technologies show the presence of background fluorescent noise near the baseline of the sequence. In almost all instances, this noise is ignored, because the examiner recognizes that the sequence is readable above the nosie and hence it is ignored. Further, validation studies have shown that a proper interpretive conclusion can be obtained in the presence of this noise. Implementation of NGS into forensic casework analyses will require a distinctive approach to this issue that reflects the nature of the data, but the general approach is unchanged.

**Chemistry-Related Variation**

There are a number of validation issues related to the particular chemistry or platform used. A full understanding of the sources of these kinds of variants would involve a careful comparison between platforms arising from the same DNA templates. For instance, the Illumina sequencing-by-synthesis chemistry algorithms attempt to correct for out-of-phase incorporation of nucleotides (called phasing and pre-phasing) by applying custom filters which collect related information from known seeded templates within each run.

Another well-studied limitation of pyrosequencing chemistry is that it is difficult to reliably interpret homoploymeric sequences longer than approximately 7 base pairs in length. This is a well-known and well-studied issue (Kunin, Engelbrektson *et. al.* 2010). The signal intensity distribution broadens with the length of the homopolymer, resulting in some cases with an ambiguous base call (Margulies, Egholm *et. al.* 2005; Kunin, Engelbrektson *et. al.* 2010), which may lead to a frame-shift affecting the downstream base calling. Orthogonal analyses, such as Sanger sequencing or the use of another NGS platform (e.g. MiSeq®) can be used to identify the nature of this limitation.

There have been a number of published studies in this area. For instance, investigations have shown that the Ion Torrent® instrument did not generate reads at all for homopolymer tracts above 14 in length, and could not predict the correct number of bases in homopolymers greater than 8 bases long. Conversely, as expected, very few errors were observed following short homopolymer stretches in the MiSeq® data due to the single base incorporation/deprotection chemistry. Additionally, they observed strand-specific errors in the PGM data but were unable to associate these errors with any obvious sequence motif. (Quail, Smith *et. al.* 2012). This is not unexpected due to the similarity in the sequencing-by-synthesis chemistry employed in both the pyrosequencing and Ion Torrent® approaches.

A separate comparison (Loman, Misra *et. al.* 2012) of different NGS platforms also showed a significant strand-bias with respect to the accuracy of calling short stretches of homopolymeric tracts. In this case, manual inspection of assembly alignments revealed that many of the falsely called indels associated with short homopolymeric tracts demonstrated strand bias, with the correct call in either the forward or reverse reads and the erroneous sequences associated with the opposite strand. Such asymmetry is unexpected given that both strands contain a homomeric run of bases, and hence may be due to the orientation of the reference sequence as necessarily one of the two strand possibilities.

One approach to this issue is to use the filtering capabilities of the software to remove all indels and instead focus strictly on SNPs. Owing to the sheer number of SNP variants found in whole genome sequencing, this approach has been utilized with some success, obviating the need to also include indels in the comparative analysis of outbreak strains of *Salmonella* species (Allard, Luo *et. al.* 2012). However, in some cases the length slippage due to homomeric incorporation ambiguity may manifest itself as a potential SNP downstream of the homomer, and hence further filtering or careful manual inspection may be necessary to ensure that only high quality reads are retained.

**Quality Scores**

Every DNA sequencing chemistry includes metrics that assess the quality of the output data, both as an overall score and an individual score for each base. The FASTQ file format provides a simple extension to the FASTA format, which is a simple concatenation of metadata followed by the DNA sequence, and includes a simple numeric quality score with each base position following the sequence from the read. (Cock, Fields *et. al.* 2010). FastQ has a number of variants, arising from different ways of calculating the probability that a base has been called in error, including differing ways of encoding that probability in ASCII text. In all cases, this format uses one character per base position, and arises from some assessment of the probability of a miscalled base at each position. The assumptions that go into the probability estimate are unique to the chemistry and instrument, and in many cases the exact algorithm applied is not expressly available. This issue may need further development as the validation of NGS into casework progresses.

**PHRED values**

Quality scores were originally derived from the PHRED program that applied to DNA sequence trace files from traditional Sanger-based fluorescent dye-based reads (Ewing and Green, 1998). These estimates were originally applied to results of careful validation studies on known DNA templates. The PHRED program assigns quality scores to each base, according to the following formula:

Q PHRED = -10 log10 (Pe)

where Pe is the probability of erroneously calling a base. PHRED puts all of these quality scores into another file called QUAL (which has a header line as in a FASTA file, followed by whitespace-separated integers. The lower the integer, the higher the probability that the base has been called incorrectly.

Below is the probability of incorrect base call accuracy

PHRED score of 10 = 1 in 10 90 %
PHRED score of 20 = 1 in 100 99 %
PHRED score of 30 = 1 in 1000 99.9 %
PHRED score of 40 = 1 in 10000 99.99 %
PHRED score of 50 = 1 in 100000 99.999 %

While scores of higher than 50 in raw reads are rare, with post-processing (such as read mapping or assembly), scores of as high as 90 are possible.

Emerging quality assurance methods recommend the use of an internal quality calibrants that are included within each NGS run. These approaches offer an expanded assessment of quality than simple PHRED scoring. One such method, called duplicate read inferred sequencing error estimation (DRISEE) provides error estimates for each position within the read as well as global error estimates that can be used to find samples with relatively high sequencing error that can be factored into further downstream processing. The DRISEE error estimate is obtained by analyzing sets of artifactual duplicate reads (ADRs), a known by-product of the Illumina® and Roche®-based sequencing platforms (Keegan, Trimble *et. al.* 2012).

**FASTQ Files and Quality Control**

There are a variety of tools that provide overall statistics for a run that include an assessment of the quality scores that arise therein. As noted above, each chemistry and platform uses a different internal method of calculating quality scores. Some of these quality assessment tools provide an output in the form of columns that show the minimum, maximum, mean, median and first and third quartile quality scores, as well as an overall count of each type of base found for that column. These tools are especially useful for determining at which position a run should be trimmed or filtered out so that only high quality sequence is

retained in the further NGS analysis stream. Other analysis tools provide some simple composition statistics for the input file, including filename, filetype, encoding (which FastQ format used), total number of sequences, sequence length and % GC present (Martinez-Alcantara, Ballesteros *et. al.* 2009).

**Quality Distributions**

The range of quality values over all bases at each position can be shown using bioinformatics tools available with each instrument, while other tools allow the user to examine whether or not there is a subset of sequences in the run that have low scores. These low-scoring sequences should represent only a small number of the total reads, and may be filtered out of subsequent analyses in an effort to retain only the highest quality reads. The concern is that if a subset of bases is systematically miscalled this may represent a bias in the particular NGS platform. Hence, known biases can be filtered out as low probability variables once they are discovered.

**Base proportions**

Some analysis tools show the proportion of each base at each position in the read. In a library prepared using methods that don't favor a particular sequence over any other, one would expect that the base frequencies are approximately equal at all positions over all the sequences. If base composition biases are observed, this usually indicates an overrepresented sequence present in the library. A bias that is consistent across all bases either indicates that the original library was sequence-biased, which may occur in amplicon sequencing, or that there was an issue that arose during the preparation of the library for sequencing.

**GC Content**

Observation of the GC content of each position in the run is also available with some tools. In most cases, there should be minimal difference between positions in the genome, and the overall GC content should reflect the GC content of the genome under study. Deviations across all positions could indicate an over-represented sequence. Again, this would be expected from targeted sequencing strategies, such as amplicon sequencing. As in the case of base proportion analysis, sample preparation may also factor in. GC content can also be assessed across each sequence and compared to a modeled GC content plot. In a random library, the plot should look normal and the mean should correspond to the overall GC of the genome under study. A non-normal distribution may indicate a contaminated library or a biased subset.

**Ambiguous Base Calls and Read Lengths**

In almost all sequencing applications, ambiguous calls more commonly appear towards the end of a run due to a decrease in overall sequencing quality. Tools exist that can plot the percentage of base calls that were Ns (i.e., a base call could not be made with certainty) at every position. This information can assist the user in identifying the position from which to filter the reads and retain only the highest quality data. Similarly, length distribution of all the read lengths found provides critical quality control information that can be compared to expected read lengths from similar runs.

**Sequence Duplication and Overrepresentation**

It may be desirable to count the number of times any sequence appears in the dataset and plot the relative number of sequences with different degrees of duplication. In a diverse library most sequences will occur only once, and hence a low level of duplication may indicate a very high level of coverage of the target

genome. Conversely, a high level of duplication is likely to indicate some kind of enrichment bias as would be expected to occur with amplicon sequencing following PCR enrichment.

A list of all the sequences that make up more than a specified percentage of the total can also be obtained with the use of the appropriate tools. In such instances the user should know what constitutes an abnormally high percentage given the context of the experiment. A normal high-throughput library will contain a diverse set of sequences, and hence finding that a single sequence is over-represented either means that it has some biologically significance, indicates that the library may be contaminated, or may not be as diverse as expected.

Overrepresentation of specific oligonucleotide sequences of a particular length (k), commonly referred to as k-mers, may also be assessed. These tools count the enrichment of every k-mer within the sequence library. Based on the nucleotide content of the library as a whole, these tools then calculate an expected level at which the k-mer should have been observed, and uses the actual count to calculate an observed/expected ratio for that k-mer. Graphing tools can show the top number of hits to indicate patterns of enrichment of k-mers across the length of the reads. This can indicate a general enrichment, or reveal a pattern of bias at different points over the read length.

**Trimming**

It has been observed that in general, the quality of few bases at the end(s) of reads is substantially lower as compared to other bases. Often quality tools are unable to filter the reads containing such low-quality bases due to their overall high quality score. However, it may not be advisable to discard the entire read due to lower quality of only few bases at the ends. It may be beneficial in such instances to only trim these particular bases prior to any downstream analysis.

Trimming tools can trim both 3' and 5' ends from each read in a dataset. For fixed-length reads such as Illumina® and SOLiD® data, the base offsets are often defined by the absolute number of bases removed from either end, whereas for variable length reads like 454, the number of bases to be trimmed off the end is defined by the percentage of the entire length.

**Removing individual reads from analysis**

It is possible that the user may desire to remove some reads that contain one or more bases of low quality within the read. This is done using filtering tools that allow the user to filter out reads that have some number of low quality bases and return only those reads of the highest quality. For example, to remove a subset of reads where the quality score of some of the bases was less than 20, this parameter can be specified and executed using the appropriate tool, which also allows the user to select how many bases per read within a string or as a percentage must be above a specified threshold to avoid being discarded.

**The Use of Index Sequences**

In the multiplexed sequencing method, DNA libraries are "tagged" with a unique identifier, (also called an index, or barcode, or MID), during sample preparation. Multiple samples are then pooled and sequenced together in one run. The user can then filter away the index reads in order to analyze only those positions that report the targeted sequence of interest.

To identify samples after pooling, each sample is uniquely tagged with a sequence index during the sample preparation protocol. The sequenced index is used in the analysis pipeline after the run is complete in order to parse the runs into bins that are user-defined treatments, different individuals, different amplicons, DNA strands, or any other kind of difference that is designed into the experiment.

**Alignment to a Reference**

Once the sequence reads have been quality filtered, it is then necessary to then align them in order to identify specific variants of interest to the investigation. In almost all cases, formerly sequenced and agreed upon references are used as standards for this purpose. The computational complexity of aligning to a reference is much less challenging than creating alignments of reads without such references, so called *de novo* alignments, which are compiled from overlapping portions of the reads themselves in an ever-expanding scaffold of the target sequence. For the vast majority of forensic cases, alignment to a previously-established reference is expected.

After alignment, another quality filter is often conducted such as minimum allele scoring which tests whether or not all members of the group have a SNP in that position. Forensic analysis generally errs on the conservative side by not considering possible error and bias through this approach risks losing some real variants.

**Errors and Error Detection**

Because the read lengths of the short-read sequencers such as the Illumina MiSeq® can now equal or exceed 100 bases from each end of the template molecule, some groups are reporting that these data can now be used for de novo assemblies [e.g. Li *et. al.* 2009 (Paszkiewicz and Studholme 2010), (Feldmeyer, Wheat *et. al.* 2011), especially when these short reads are supplemented with mate-paired reads (Gnerre, Maccallum *et. al.* 2011) and/or data from one of the longer-read platforms (e.g. Dalloul, Long *et. al.* 2010) in what is being called a hybrid assembly process. (Glenn, 2011; Bashir, Klammer *et. al.* 2012). These efforts attempt to overcome the inherent error rates of the new methods by combining the strengths of each while at the same time minimizing the effects of their respective weaknesses.

Using previously sequenced templates as controls, a recent study compared the error rates between three commonly used platforms, and observed error rates of below 0.4 % for the Illumina® platforms, 1.78 % for Ion Torrent® and 13% for PacBio® sequencing. (Quail, Smith *et. al.* 2012). Although the error rates of single-molecule reads are high, single-molecule sequencing instruments such as the PacBio RS® can generate long reads with the potential to improve the assembly process. Further, correction algorithms have been developed that use short, high-fidelity sequences to correct the errors in single-molecule reads. As the short read lengths increase, much higher efficiencies are predicted to accrue from these algorithms.

Increased accuracy and read length in NGS would facilitate the discovery of new variants also bring higher accuracy to these methods. Longer reads and inserts are needed to increase the specificity in read mapping, obviating the issue of reads being shorter than the repeats themselves. In many laboratories, long reads from traditional platforms are combined with short reads from NGS outputs to form large contigs that span over large regions that may include repeats. Recent studies have combined so-called third generation sequencing methods, those that sequence individual templates without a previous amplification step, such as bridge PCR or emulsion PCR, with the shorter reads arising from such methods, to obtain long contiguous assemblies. In one such study, second and third-generation sequencing data were used to assemble the two-chromosome genome of a recent Haitian cholera outbreak strain into two nearly finished contigs at >99.9% accuracy. This study looked at separate control assemblies on experimental and simulated data to correct several errors in contigs assembled from the short-read data alone (Hasan, Choi *et. al.* 2012).

In order to overcome the limitations of single-molecule sequencing approaches and widen it's applicability, another group developed an approach that utilizes short, high-accuracy reads to correct the error inherent in long, single-molecule reads obtained from the PacBio® instrument. In this case the researchers used a 'corrected reads' algorithm that initially maps short-read sequences to individual long-read sequences

and corrects them by computing a highly accurate hybrid consensus sequence based on all the data, thus improving the read accuracy in this case from as low as 80% to over 99.9% (Koren, Schatz *et. al.* 2012). It therefore appears that an emerging practice is to combine the strengths and weaknesses of each kind of platform and chemistry by combining data from different kinds of technologies into a single analysis in an effort to minimize the impact of the weaknesses of each while also building on their respective strengths.

By virtue of its long read lengths, the PacBio® platform should, however, have advantages in de novo sequencing and may also benefit the identification of alternative splicing of variants across long amplicons. Also, the potential for direct detection of epigenetic modifications has been demonstrated (Flusberg, Webster *et. al.* 2010). Some authors (e.g. Loman, Misra *et. al.* 2012), however, have noted that the DNA-input requirements of PacBio® can be prohibitory. While the Illumina® and PGM® library preparation methods can be performed with far less DNA; the sample-input requirements and amplification-free library preparation methods of the PacBio® render it potentially unsuitable for quantitative applications and those that require significant prior enrichment of the template (Choi, Scholl *et. al.* 2009) (Johnson, Mortazavi *et. al.* 2007).

**Results – Section 1 - The use of fusion primers to support NGS.**

In order to bind to the physical matrix of either the magnetic bead (e.g. Roche™) or the flowcell (e.g. Illumina®), oligonucleotide sequences must be incorporated into the product to be sequenced. In genome sequencing applications, until very recently this was done by shearing the template and incorporating the recognition sequences via blunt-end ligation. In amplicon sequencing, this can be accomplished by using modified PCR primers. The recognition sequences are placed on the 5' end of the primer, and the product of amplification is then ready for direct sequencing.

We have shown that this approach works well with the current set of mtDNA-specific primer sequences (see Table 6). We designed fusion primers that contain not only the template-specific portion but also the other sequences necessary for use in this context. The primers were approximately 84 base pairs in length. These primers did support amplification of human mtDNA targets, and the resulting amplicons were readily sequenced (data not shown).


*Figure 1 – Diagram of the modified primers used to generate the PCR product for our amplicon libraries.*

Figure 2 shows the amplification results from these experiments using the Agilent 2100 Bioanalyzer™. As can be seen, amplification products of the expected size are obtained. These products are of the expected size if the longer fusion primer sequence is included.

*Figure 2: Sample Agilent 2100 Bioanalyzer Data for Multiplexing Strategy Using AmpliTaq® Gold.*

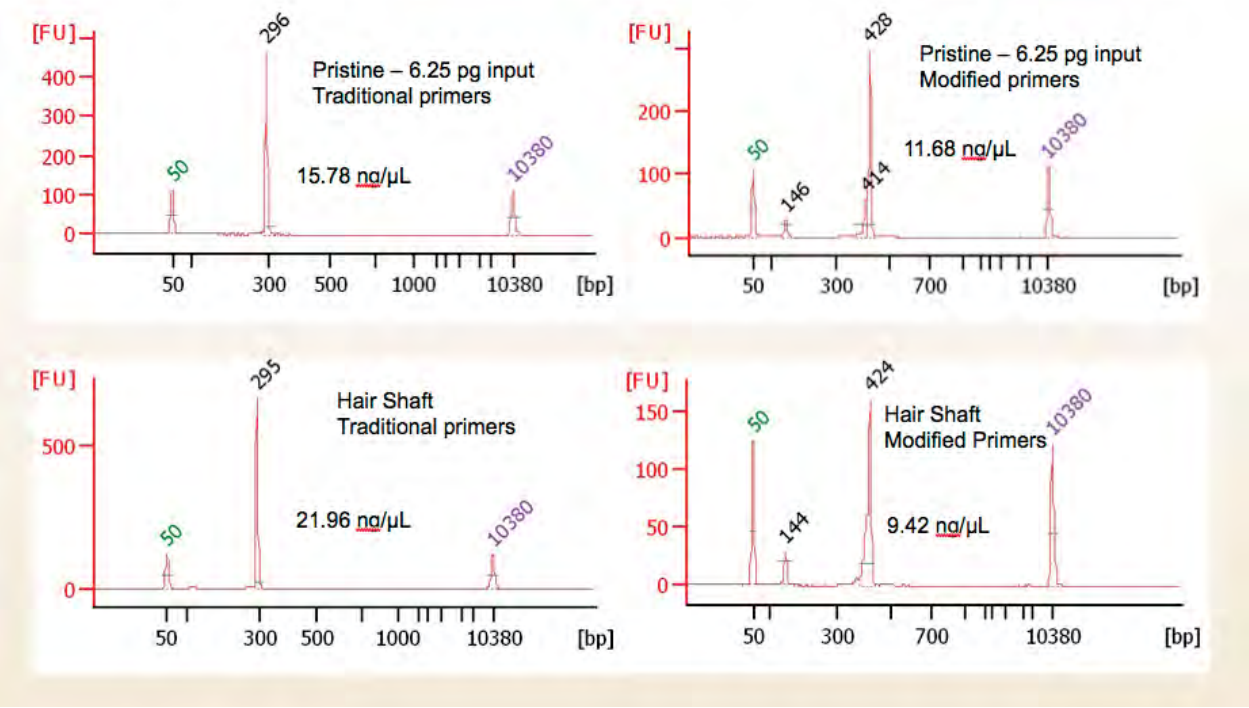



*Figure 2 – PCR results using traditional primers (left column) on diluted pristine DNA and hair shaft samples, as well as the longer fusion primers on the same DNA templates. Amplicons of expected size from the human mtDNA control region are obtained in both instances.*

We reasoned that this approach to NGS sample preparation would require separate amplification reactions sufficient to cover the entire mtDNA genome. This would require the crime laboratory to keep track of a large number of separate amplification reactions, burdening the process and potentially limiting its usefulness. Therefore, we began to investigate other methods of sample preparation that may be simpler and more useful. We eventually settled on a transposition-based approach called Nextera, a process now owned by Illumina. Using this approach, all double-stranded DNA, including PCR products, can directly be prepared for NGS in a single integrated step. The issue of sample preparation then became much simpler, and also did not necessarily require dedicated PCR products as the template for NGS. If we could generate sufficient double-stranded DNA in the preparation step, then direct processing by the transposase would render these fragments directly amenable to sequencing.

**Identification of Nuclear Pseudogene Inserts in PCR products**

The Roche GS-Junior™ instrument can generate thousands of independent sequencing reads (generally ~70,000) per sequencing run. It is thought that the proportions of these reads will corresponds directly to the proportions of the amplicons from which they are generated. Thus, by quantifying the proportion of reads that show an SNP, or variant, we anticipated to be able to infer the proportion of the variant within an amplicon library. To test this method, we used the Roche Jr. platform to sequence mixtures of known proportions of PCR product from two previously sequenced donors. We analyzed the sequence data using Roche Amplicon Variant Analyzer (AVA) software, and compared the results against the known sequence variants between the two donors, and the known proportions of the mixtures.

The capability to accurately detect and quantify variants within a mixture of amplicons will benefit challenging analyses of mitochondrial DNA samples, especially those involving heteroplasmy (the presence of two or more mitochondrial haplotypes within a single individual, tissue, cell, or mitochondrion). The

purpose of this experiment was to demonstrate our ability to detect and quantify minor variants from amplicon libraries using massively parallel sequencing on the 454 Roche Jr. platform. While performing this experiment, we noticed a persistent group of minor variants that warranted further investigation. A search of the NCBI nucleotide database revealed that the sequence that we identified as the unexpected variant has been previously described as a nuclear insertion of a mitochondrial DNA fragment (NumtS)[1-3]. We performed a series of experiments to confirm that these variants are in fact mostly likely nuclear inserted elements. We discuss the interpretational ramifications of these findings in the context of forensic science.

**Library Preparation for Pyrosequencing Reactions**

To generate the PCR product for our mixtures, we amplified segments of the hypervariable region (HV1a and HV1b) of the mitochondrial genome using FTA card blood samples from two donors. Reference sequences for the donors were previously obtained using the same FTA cards and Sanger sequencing using an Applied Biosystems™ 3130xl Genetic Analyzer. To maximize the efficiency of each sequencing run on the Roche Jr., we multiplex sequenced multiple amplicon libraries on a single Pico Titer plate (PTP) (Figure 1). For the experiment's first sequencing run (hereafter refered to as run A) we multiplex sequenced twelve distinct amplicon libraries (Table 2). In the second, (run B) we included fourteen amplicon libraries (Table 1). In order to resolve which sequencing reads are connected to each amplicon library, short (~20 nucleotides) spans of sequence that include multiplex identifiers (MIDs) are incorporated into the amplicons during PCR using 5' modified primers (Figure 2). The MID sequences are sequenced along with the amplicon during the sequencing run, and act as a 'sequence barcodes', indicating the provenance of each sequencing read. We used the same MID for both the forward and reverse primer for each PCR, so that the amplicons of each library were marked with the same MID at both ends.

To create a single mixed amplicon library, we amplified a segment of the HV region from the two donors separately, but with the same modified primer set bearing the same MID. We quantified the PCR product from each reaction, then mixed them according to set ratios. The result was a mixture of amplicons from different donors, each of which bears the same MID sequence at both ends. To ensure accurate quantification of our mixtures, we quantified our PCR products with five replicates per sample on the Agilent 2100 Bioanalyzer™ DNA 1000 kit (Cat.# 5067-1504), using the replicates' means for our mixture calculations. Negative controls were quantified once using the instrument's high sensitivity kit (Cat.# 5067-4626). For run B, we included MID 13 and 14 as non-template controls (NTCs) in the sequencing run. We used a blank hole punch from each donor's FTA card for the NTCs. The primer sets for MIDs 13 and 14 were diluted separately from those for MIDs 1-12. All negative controls rendered negative results with the high sensitivity kit. Using our quantification data from the Bioanalyzer™, we diluted all PCR products to $10^9$ molecules/µL before creating our mixtures. After mixing, we pooled the all amplicon libraries together by adding equal volumes of each. We diluted the pool to $2e^6$ molecules/µl for immobilization to capture beads using emPCR. From here we followed the manufacturer's protocol for amplicon sequencing on the Roche GS Junior™.

**Data Analysis**

We analyzed our the sequence reads from our HV1b libraries from each run using the Roche 454 Amplicon Variant Analyzer software (Figure 4 A and B). We used the software to 1) sort the reads by the MID sequences detected at the beginning or end of each read (demultiplexing), 2) align the reads against reference sequences to generate a table of putative variants and their percentage in each library, and 3) generate bar graphs illustrating the position, nucleotide identity, and percentage of the variants detected within a given library (Figure 3). The magnitude of each variant is given as a percent. This value indicates the number of reads that sequenced with a nucleotide different from that of the designated reference sequence at

each position. The software assigns each read to a library based on the identity of the MID sequence(s) detected within the read.

The summary statistics for each of the two sequencing runs are shown in Table 3. Table 4 shows the nucleotide positions within HV1b that differ between the two donors. The positions that differ between the donors were our expected variants. With the exception of positions 16223 and 16224, all expected variants were detected in all of the mixtures. In both runs, the variants at 16223 and 16224 were not detected for any library when donor 3's sequence was used as the reference. With donor 15's sequence as a reference, the software detected all expected variants for all mixed libraries (see attached spreadsheet for tables of variants). In addition to the expected variants, a large number of unexpected variants were detected. These variants averaged 0.800% (SD = 0.261) among all libraries in Run A, and 0.903% (SD = 0.445) for Run B. These variants appeared consistently among each library with the exception of those designated by MIDs 13 and 14. The unexpected variants were also consistent between runs, with the exception of variants at positions 16173, 16174, and 16175, which appear only in run B. Seventeen reads were detected for the non-template control library represented by MID 13. The variants detected for this library did not match those detected in the other twelve. From these preliminary results, we concluded that the method can detect variants in amplicon libraries with proportions as low as 0.5 percent, albeit not without some caveats. The method can quantify the variants within 2% of expected values (Mean difference between experimental and expected percentage values for expected variants was 1.065, (SD = 0.601) (data from Run B aligned against reference 15)..

**Unexpected Variants**

Although the method successfully captured the variants that we expected in our mixed amplicon libraries, we also detected a number of unexpected variants. The possible sources of these variants are 1) artifact noise from the sequencing method or 2) exogenous contamination of the libraries. As the same variants appear consistently between libraries and between runs, we consider noise from the sequencing method an unlikely explanation. The unexpected variants are also common SNPs from the region, further supporting the explanation of exogenous contamination. These particular unexpected variants were distinct from those identified as originating from nuclear pseudogene inserts. The detection of this contamination speaks to the sensitivity of the method, as the Agilent 2100 Bioanalyzer™ high-sensitivity kit did not detect the presence of peaks in our negative controls. We also note that the minor variants detected for MIDs 1-12 were not detected for MID13. It is therefore possible that the varaints shared between the first twelve MIDs was introduced during their dilution, as MIDs 13 and 14 were diluted separately. This would explain the consistency of the unexpected variants between the two Runs and between the different amplicon libraries.

The method appears sufficiently sensitive to detect extremely low-level minor variants. We anticipate the greatest challenge being the differentiation between variants that were truly innate to the source sample, and those that arise from trace exogenous contamination or noise. Further analysis is needed to assess the precision and accuracy with which the method can identify and quantify variants. However, our results indicate that within a reasonable margin of error determined statistically, the method can quantify variants within an amplicon library.

| MID | Mixture Ratio | Donor(s) | HV region |
|---|---|---|---|
| 1 | 100 | 3 | hv1a |
| 2 | 100 | 3 | hv1b |
| 3 | 95:5 | 3:15 | hv1b |
| 4 | 98:2 | 3:15 | hv1b |
| 5 | 99:1 | 3:15 | hv1b |
| 6 | 99.5:0.5 | 3:15 | hv1b |
| 7 | 5:95 | 3:15 | hv1b |
| 8 | 2:98 | 3:15 | hv1b |
| 9 | 1:99 | 3:15 | hv1b |
| 10 | 0.5:99.5 | 3:15 | hv1b |
| 11 | 100 | 15 | hv1a |
| 12 | 100 | 15 | hv1b |
| 13 | NTC | NTC | hv1b |
| 14 | NTC | NTC | hv1a |

*Table 1: Description of the libraries included for run A, indicating the Roche standard multiplex identifier used to generate the library's amplicons, the ratio of the mixture, the donors from the DNA template came from, and the segment of the hypervariable region that was amplified.*

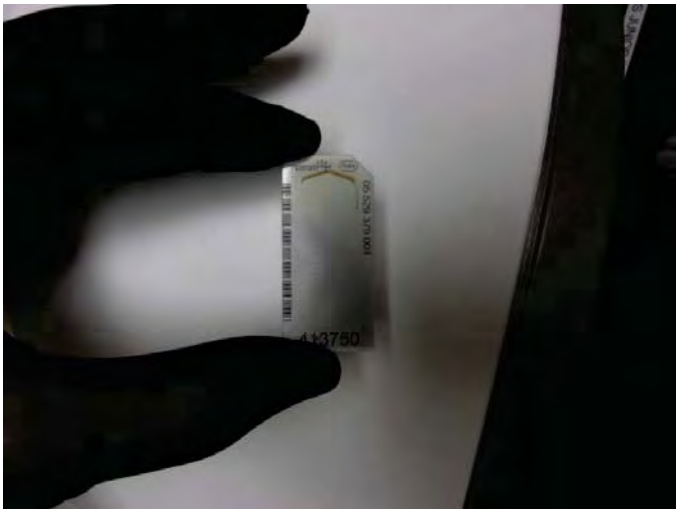


*Figure 3: The Pico Titer Plate (PTP) is the platform on which the DNA capture beads are immobilized in the Roche GS Junior™, and the location of the sequencing reactions and data collection. It houses a honeycomb of wells sized to hold a single DNA capture bead. Each sequencing run consumes a non-reusuable PTP.*

| MID | Mixture Ratio | Donor(s) | HV region |
|---|---|---|---|
| 1 | 100 | 3 | hv1b |
| 2 | 100 | 15 | hv1b |
| 3 | 90:10 | 3:15 | hv1b |
| 4 | 95:5 | 3:15 | hv1b |
| 5 | 99:1 | 3:15 | hv1b |
| 6 | 99.5:0.5 | 3:15 | hv1b |
| 7 | 10:90 | 3:15 | hv1b |
| 8 | 5:95 | 3:15 | hv1b |
| 9 | 99:1 | 3:15 | hv1b |
| 10 | 0.5:99.5 | 3:15 | hv1b |
| 11 | 100 | 3 | hv1a |
| 12 | 100 | 15 | hv1a |

*Table 2: Description of the libraries included for run B, indicating the Roche standard multiplex identifier used to generate the library's amplicons, the ratio of the mixture preparation, the donors from whom the DNA template came from, and the segment of the hypervariable region that was amplified. This general mixture scheme was employed throughout this study.*

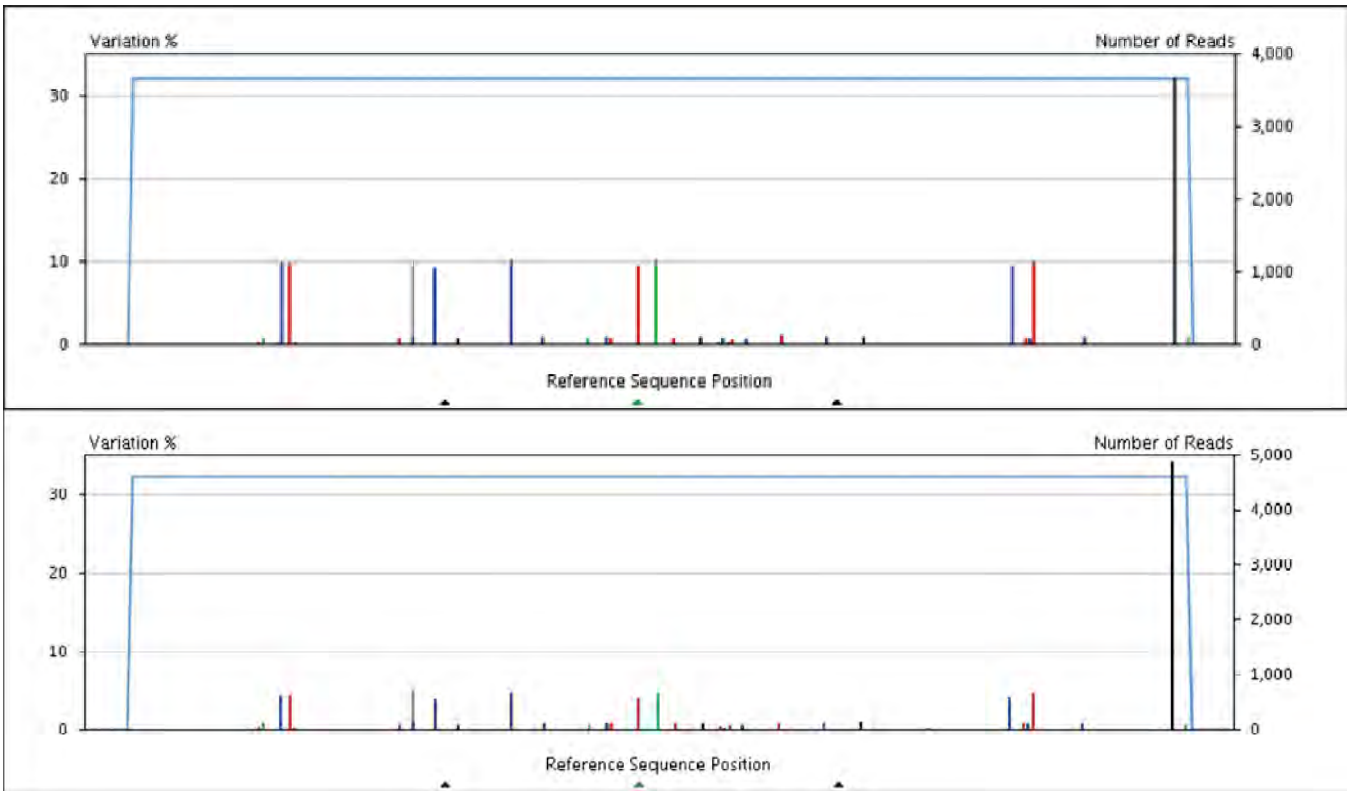



*Figure 4 (A and B): Examples of graphs exported from the AVA software. The top graph shows the variants detected for the MID 3 library for Run A referenced against the sequence for donor 3. The bottom shows the variants for the MID 4 library from Run A. Each bar represents the position of a nucleotide that varies from the reference sequence. The color of a bar indicates the identity of the nucleotide, and its height represents the proportion of reads that show the variant.*

| | Run A | Run B |
|---|---|---|
| Raw Wells | 101,994 | 214,129 |
| Key Pass Wells | 87,279 | 200,235 |
| Passed Filter Wells | 45,023 | 97,836 |
| Total Bases | 13,173,819 | 28,571,581 |
| Length Average | 292.6 | 292.04 |
| Length Std Deviation | | |
| Longest Reads Length | 680 | 799 |
| Shortest Reads Length | 53 | 45 |
| Median Reads Length | 291 | 291 |

*Table 3: Summary of results from Runs A and B. Passed Filter wells are those that passed all quality filters used by the data processing software to ensure a yield of only high quality reads. Passed-filter wells represents the total number of useable reads sequenced in the run.*

| Position | Donor 3 | Donor 15 |
|---|---|---|
| 16193 | T | C |
| 16195 | C | T |
| 16221 | T | C |
| 16223 | C | T |
| 16224 | T | C |
| 16242 | A | C |
| 16270 | C | T |
| 16274 | G | A |
| 16352 | T | C |
| 16357 | C | T |

*Table 4: Nucleotide positions within HV1b that differ between the two donors. These were our expected variants for our mixed libraries.*

These results lead to a full investigation of the possible origin of the unexpected variants as reported in the next section.

**Results - Section 2 – Sample Preparation for Mixture Study**

Forensically relevant samples including buccal swabs, whole blood on Whatman® FTA® cards and hairs were collected from twenty donors according to the Human Subjects Institutional Review Board policies implemented at Western Carolina University and following informed consent. Reference sequence data for the mtDNA control region was obtained for all donors using Sanger methods. Pairs of donors exhibiting the highest amount of sequence variation within the control region were chosen for mixture studies using pyrosequencing to elucidate the minor variant limit of detection of the Roche GS Junior™ instrument.

**Obtaining Donor Reference Sequence Data**

DNA from bloodstain card punches (1.2 mm) was purified using FTA® purification reagent (GE Healthcare, UK) following manufacturers protocol[16]. The mtDNA hypervariable region was amplified in 4 distinct amplification reactions as follows: purified template DNA on 1.2 mm FTA punches in a reaction

mixture containing 5 U of AmpliTaq Gold DNA polymerase (Applied Biosystems, Foster City, CA), 1x GeneAmp PCR Buffer (Applied Biosystems, Foster City, CA), 160 ng/μL BSA (Thermo Fisher Scientific, Rockford, IL), 200 μM each dATP, dTTP, dCTP, dGTP from PCR grade nucleotide mix (Promega Corporation, Madison, WI), 600 nM forward primer, and 600 nM reverse primer. Primer sequences are shown in Table 6. Reaction mixtures were amplified on a GeneAmp® PCR System 9700 (Applied Biosystems, Foster City, CA) with an initial 11 minute hold at 95ºC, followed by 32 cycles comprised of a 15 second denaturation at 95ºC, a 30 second annealing step at 56ºC, and a 45 second extension at 72ºC with a final hold at 4ºC. Following amplification, unincorporated dNTPs and primers were enzymatically removed from each sample with ExoSAP-IT® (Affymetrix, Santa Clara, CA)[18]. Resulting amplification products were analyzed using the Agilent 2100 Bioanalyzer™ with the DNA 1000 kit (Agilent Technologies, Inc., Waldbronn, Germany), and were normalized to 1 ng/uL using $TE^{-4}$ buffer (Teknova, Hollister, CA). Samples (5.0 ng) were cycle-sequenced using the BigDye® Terminator v1.1 (Applied Biosystems, Foster City, CA) kit according to manufacturers instructions[18], and fragments were separated on a 3130xl Genetic Analyzer (Applied Biosystems, Foster City, CA) and analyzed with Sequencher® 5.0 software (Gene Codes Corporation, Ann Arbor, MI).

| | | |
|---|---|---|
| HV1a | A1 (L15997) | CAC CAT TAG CAC CCA AAG CT |
| | B2 (H16237) | GGC TTT GGA GTT GCA GTT GAT |
| HV1b | A2 (L16259) | TAC TTG ACC ACC TGT AGT AC |
| | B1 (H16391) | GAG GAT GGT GGT CAA GGG AC |
| HV2a | C1 (L 048) | CTC ACG GGA GCT CTC CAT GC |
| | D2 (H 285) | GGG GTT TGG TGG AAA TTT TTT G |
| HV2b | C2 (L 177) | TTA TTT ATC GCA CCT ACG TTC AAT |
| | D1 (H 409) | CTG TTA AAA GTG CAT ACC GCC |

*TABLE 5- Mitochondrial DNA control region primer sequences. Primer IDs show strand represented (light versus heavy) and rCRS position of the 3' base of the primer sequence.*

**Sample Preparation and Pyrosequencing – Mixture Study**

To determine the minor variant limit of detection of the Roche GS Junior™ pyrosequencing instrument, a mixture study was designed in which mixtures were prepared from pairs of donors with maximum variability between their known mtDNA control region sequences. The mtDNA control region was amplified in four independent PCR reactions from whole blood samples stored on Whatman® FTA® classic cards (GE Healthcare, UK) from four donors (donors 001-CF30, 003-CM54, 005-CF40, and 015-AM30). To reduce the occurrence of polymerase-induced base-misincorporations, the Roche FastStart High-fidelity PCR system (Roche Applied Science, Indianapolis, IN) was used for DNA amplification as follows: purified template DNA on 1.2 mm FTA punches in a reaction mixture containing 1.25 U of Roche FastStart High Fidelity enzyme blend, 1X FastStart High Fidelity reaction buffer with 1.8 mm $MgCl_2$, 200 μM each dATP, dTTP, dCTP, dGTP from PCR grade nucleotide mix (Promega Corporation, Madison, WI), 400 nM forward primer and 400 nM reverse primer. Forward and reverse fusion primers (Fig. 1) were designed in which 454 specific adaptor sequences, and multiplex identifiers were included immediately 5' of an mtDNA template specific primer sequence (Table 6), enabling next-generation sequencing (NGS) sample preparation using PCR amplification. Reaction mixtures were amplified on a GeneAmp® PCR System 9700 (Applied Biosystems, Foster City, CA) with an initial 2 minute hold at 95ºC, followed by 32 cycles comprised of a 30

second denaturation at 95ºC, a 30 second annealing step at 60ºC, and a 30 second extension at 72ºC with a final 7 minute extension at 72ºC and a long term 4ºC hold. Resulting PCR products were purified with Agencourt® AMPure® XP beads (Beckman Coulter, Indianapolis, IN) for removal of unincorporated primers, and dNTP's. Purified amplicons were quantitated in quintuplicate using the Agilent 2100 Bioanalyzer™, and the concentrations were averaged. The amplified products were normalized to 1 ng/µL and mixed in defined ratios of 10, 5, 2, 1 and 0.5%. Single donor samples and prepared mixtures were then pooled at equimolar concentrations for multiplexed sequencing on the GS-Junior™ instrument.

Individual single-stranded template molecules from the pooled library were clonally amplified on the surface of paramagnetic DNA capture beads using the Roche GS Junior™ Titanium Lib-A kit (Roche Applied Science, Indianapolis, IN) for emulsion PCR (emPCR). This kit contains two sets of beads, each coated with oligonucleotides complementary to adapters corresponding to either the sense or antisense PCR products to allow for bidirectional sequencing reads. Enrichment for beads with clonally amplified sequencing template was performed, and resulting beads were deposited into independent microwells of a PicoTiter™ Plate (PTP) device. Pyrosequencing and raw image collection were carried out for 200 cycles. Pooled libraries were deep-sequenced across three independent pyrosequencing runs with a target of 5,000x coverage per sample to capture minor variants at 1% or lower.

**Library Preparation – Tissue Comparison Study**

DNA from three forensically relevant tissue types including hair, blood and buccal cells from donor 001-CF30 was deep-sequenced to determine if minor sequence differences exist between different tissue types originating from the same individual. DNA was extracted from five hair shafts from different regions of the scalp of donor 001-CF30 using the Qiagen DNA Investigator Kit (Qiagen, Germany), and from buccal swabs using the Qiagen DNA mini kit (Qiagen, Germany) with no modifications to the vendor recommended protocols. Extracts were quantified using a human mtDNA specific 5'-nuclease real-time PCR assay[19]. DNA was amplified using the Roche FastStart PCR System (Roche Diagnostics, Indianapolis, IN) and fusion primers for 454 library preparation as described above, with 10 µL of template added per reaction from buccal and hair extracts. Additionally, DNA was purified from whole blood samples stored on Whatman® FTA® classic cards (GE Healthcare, UK), and amplified directly using the Roche FastStart PCR System. DNA extracts were assigned unique MID's for post-run sample parsing by tissue type. Resulting PCR products were purified using Agencourt® AMPure® XP beads, were quantitated using an Agilent 2100 Bioanalyzer™ and the DNA 1000 kit (Agilent, Waldbronn, Germany), and were normalized to a concentration of 1 ng/µL using $TE^{-4}$ pH 8.0 (Teknova, Hollister, CA). Normalized samples were then pooled, and clonally amplified using emPCR as described in the section titled *Sample Preparation and Pyrosequencing – Mixture Study*. A 200-cycle sequencing run was performed.

**Roche GS Junior 454 Data Analysis**

Roche Amplicon Variant Analyzer (AVA) software v2.7 was used for identification and quantification of minor variants using default analysis parameters. This software is capable of parsing reads according to the multiplex identifier (MID) detected, aligning reads against a specified reference sequence, and generating a list of putative variants and their occurrence frequency within each library. All libraries were aligned against the rCRS[21]. Putative variants were called where ≥20 reads differed at a given position from the reference sequence, and were quantified as a proportion of reads from a given library sharing the same MID. In addition to standard quality filters applied to the data set by the AVA software, minor variants were only further considered "real" if they appeared bidirectionally within the library. Roche AVA software has a "bidirectional support" option to assist the analyst with assessment of variants representative of true biological events. This option is only useful if variants lie in a region of the amplicon that is covered by both forward and reverse reads. If the variant in question is found in both forward and reverse reads at similar frequencies it is more likely to represent a true variant. It does not appear that the "bidirectional support"

option removes variants with large bidirectional discrepancies from the data set, however, such variants are flagged in the data table. Analysts have the option of viewing frequencies for variants in forward and reverse reads independently.

**PCR Confirmation of NumtS**

DNA was extracted from buccal swabs from twenty donors using the QIAamp® DNA mini kit (Qiagen, Valencia, CA). Extracted nuclear DNA was quantified with real-time PCR using the Quantifiler™ Human DNA Quantification Kit (Applied Biosystems, Foster City, CA) and 7500 real-time PCR system (Applied Biosystems, Foster City, CA) according to the manufacturer's instructions21. Extracts were normalized to a concentration of 1 ng/μL with TE-4 buffer, pH 8.0 (Teknova, Hollister, CA). Two control samples commonly encountered in forensic laboratories were also included in the sample set (9947A and HL60). A nuclear specific primer set (Integrated DNA Technologies, Coralville, IA) designed by Thomas *et. al.*3 to flank the region containing the NumtS insertion was used to confirm the presence of the NumtS in all twenty donors. Samples were amplified using the Roche FastStart High-Fidelity PCR System (Roche Diagnostics, Indianapolis, IN). Initially, input quantities ranging from 5.8 – 14.2 ng were amplified. The resulting yields of PCR products were low in most cases (0.04 – 0.95 ng/μL) and it was thought that amplification failure of the NumtS insertion negative peak might be occurring. As a result, extracted DNA samples were then re-amplified with increased input amounts of DNA as follows: 10 μL of purified buccal extract (ranging from 24.0 – 147.0 ng input) in a reaction mixture containing 2.5 U of Roche FastStart High Fidelity enzyme blend, 1X FastStart High Fidelity reaction buffer, 3.0 mm MgCl2 (Applied Biosystems, Foster City, CA), 4% DMSO, 400 μM each dATP, dTTP, dCTP, dGTP from PCR grade nucleotide mix (Promega Corporation, Madison, WI), 400 nM forward primer and 400 nM reverse primer. Primer sequences are shown in Table 6. Reaction mixtures were amplified on a GeneAmp® PCR System 9700 (Applied Biosystems, Foster City, CA) with an initial 2 minute hold at 95ºC, followed by 32 cycles comprised of a 30 second denaturation at 95ºC, a 30 second annealing step at 60ºC, and a 30 second extension at 72ºC with a final 7 minute extension at 72ºC and a long term 4ºC hold. The resulting amplification products were analyzed using the Agilent 2100 Bioanalyzer™ and DNA 1000 kit (Agilent Technologies, Waldbronn, Germany). Individuals that were found to be heterozygous for the NumtS insertion were defined as those who presented two peaks with sizes of approximately 155 bp and 711 bp. Homozygous individuals included those who showed a single peak at approximately 155 bp (no NumtS on either chromosome) or 711 bp (NumtS insertion on both chromosomes).

| | |
|---|---|
| F 5' – AGTCTTGCTTATTACAATGATGG – 3' | 49,840,047 – 49,840,069 |
| R 5' – ACAAAGTCCAGGTTTCTAACAG – 3' | 49,840,174 – 49,840,195 |

*TABLE 6- Primer sequences for NumtS-specific amplification (from Thomas et. al.[3]) Amplicon size without NumtS insertion = 155 bases Amplicon size with NumtS insertion = 711 bases*

**Sanger Sequencing of NumtS**

Initially, HV1b amplification primers were used to sequence the NumtS insertion amplicons. PCR products (10 ng) from all twenty donors were sequenced in forward and reverse directions using the BigDye® Terminator v1.1 Cycle-Sequencing kit (Applied Biosystems, Foster City, CA) according to manufacturers instructions[18]. Fragments were separated on a 3130xl Genetic Analyzer (Applied Biosystems, Foster City, CA) and analyzed with Sequencher® 5.0 software (Gene Codes Corporation, Ann Arbor, MI). To show that nuclear DNA was being sequenced exclusively, amplified samples were then Sanger sequenced as described above using the nuclear specific primers originally used to amplify the NumtS insertion region (Table 6). At the point in the sequence where the insertion is expected, the resulting sequence data from

heterozygous individuals was out-of-phase, resulting from the co-amplification and co-sequencing of inserted and non-inserted alleles. To confirm the mixed nature of these templates, PCR products were run on a 1.2% agarose gel at 100V for 20 minutes and the resulting bands were excised from the gel to isolate NumtS negative and positive products for further sequence analysis. DNA purification was accomplished with the Qiaquick® Gel Extraction kit (Qiagen, Germany) according to manufacturers instructions[22]. Purified amplicons were quantified using the Agilent 2100 Bioanalyzer™ and the DNA 1000 kit, and the amplicons were Sanger sequenced independently as described above with nuclear DNA-specific primers.

### Detection of Minor Variants

#### Mixture Study – HV1a Data

Mixtures of HV1a amplicons were prepared in defined ratios to determine the limit of detection of minor sequence variants using the Roche GS Junior™ 454 pyrosequencing instrument as described above. Single donor amplicons or prepared mixtures were bioinformatically parsed using a sample dependent multiplex identifier sequence (MID), and showed an average depth of coverage of 1,950 reads. The frequencies of variants reported by the Roche AVA software for single donors and prepared mixtures are shown in Table 7. In addition to the expected variants arising from the minor contributor, several unexpected variants with frequencies below 1.5% were detected. However, all HV1a unexpected variants appeared in either a forward or reverse read (but not both) and as a result could be excluded from the data set by applying a bidirectional read variant confirmation quality filter. This quality filter is built into the AVA software and can be applied to any data set. As shown in Table 9, the remaining unexpected variants are completely excluded from the data set after applying the bidirectional read quality filter.

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **16069: C/T** | **99.95** | **0** | **0.09** | **1.1** | **1.85** | **3.89** | **95.04** | **98.3** | **99.43** | **99.35** |
| **16093: T/C** | **97.97** | **0** | **0.09** | **1.15** | **1.8** | **3.95** | **93.78** | **96.78** | **97.86** | **97.63** |
| 16124: T/G | 0 | 0.69 | 0.67 | 0.62 | 0.31 | 0 | 0.9 | 0.6 | 0.61 | 1.28 |
| 16125: G/A | 0 | 0.69 | 0.71 | 0.62 | 0.31 | 0 | 0.99 | 0.6 | 0.65 | 1.31 |
| **16126: T/C** | **99.09** | **0** | **0.26** | **1.1** | **1.8** | **3.96** | **94.1** | **97.06** | **98.08** | **97.62** |
| **16129: G/A** | **0** | **97.51** | **98.28** | **97.53** | **96.92** | **94.59** | **4.49** | **1.56** | **0.28** | **0.7** |
| 16130: G/T | 0 | 0 | 0 | 0 | 0 | 0.45 | 0.51 | 0.24 | 0.33 | 0 |
| 16131: T/G | 0 | 0 | 0 | 0 | 0 | 0.45 | 0.51 | 0.24 | 0.33 | 0 |
| 16132-16138: DEL(7) | 0 | 0 | 0 | 0 | 0 | 0.15 | 0.56 | 0.24 | 0.37 | 0.52 |
| 16132-16139: DEL(8) | 0 | 0 | 0 | 0 | 0 | 0.15 | 0.56 | 0.24 | 0.37 | 0 |
| 16134-16135: CA/-- | 0 | 0.69 | 0.62 | 0.57 | 0.31 | 0.15 | 1.45 | 0.84 | 0.98 | 1.8 |
| 16137-16139: AAA/--- | 0 | 0.69 | 0.62 | 0.57 | 0.31 | 0.15 | 1.45 | 0.84 | 0.98 | 1.28 |
| 16139: A/G | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.7 |
| 16236: C/T | 0 | 0 | 0 | 1.52 | 0 | 0 | 0 | 0 | 0 | 0 |
| **Total Coverage** | **2,094** | **1,163** | **2,272** | **2,117** | **1,953** | **1,988** | **2,350** | **1,679** | **2,153** | **1,726** |

*TABLE 7 - Roche GS Junior™ 454 Pyrosequencing Mixture Experiment: HV1a Variants detected using Roche Amplicon Variant Analyzer Software.*

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **16069: C/T** | **99.95** | **0** | **0.09** | **1.1** | **1.85** | **3.89** | **95.04** | **98.3** | **99.43** | **99.35** |
| **16093: T/C** | **97.97** | **0** | **0.09** | **1.15** | **1.8** | **3.95** | **93.78** | **96.78** | **97.86** | **97.63** |
| **16126: T/C** | **99.09** | **0** | **0.26** | **1.1** | **1.8** | **3.96** | **94.1** | **97.06** | **98.08** | **97.62** |
| **16129: G/A** | **0** | **97.51** | **98.28** | **97.53** | **96.92** | **94.59** | **4.49** | **1.56** | **0.28** | **0.7** |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Total Coverage** | **2,094** | **1,163** | **2,272** | **2,117** | **1,953** | **1,988** | **2,350** | **1,679** | **2,153** | **1,726** |

*TABLE 8- HV1a expected variants remaining after bidirectional read variant confirmation. Variants highlighted in gray represent expected variants originally obtained for each sample donor using Sanger sequencing. Mixtures were prepared in defined ratios by combining quantified HV1a amplification products from donors 001-CF30 and 005-CF40. Data is presented as the frequency of each variant detected per MID.*

**Mixture Study – HV1b Data**

Roche AVA software successfully parsed all HV1b samples based on the MID sequence, and allowed for detection of all expected variants with an average read depth of approximately 8,158 per sample. In addition to the expected variants arising from the mixture, a subset of unexpected variants was detected with an average frequency of 0.87% of the total library data set. In contrast to the HV1a data set, bidirectional filtering of unexpected variants did not remove these variants from the HV1b data set (Table 9). These unexpected variants were reproduced with similar frequencies in an additional NGS run with an average depth of coverage of 1,875 per library. None of the unexpected variants were detected at positions spanning the region shared between the overlapping HV1a and HV1b amplicons, and all negative controls lacked any analyzable sequence. Furthermore, using the AVA Global Alignment Consensus viewing option, all unexpected HV1b variants were linked as they were consistently detected as a group within a small subset of the same reads.

A Blast-N search[23] of the rCRS HV1b sequence with unexpected variants against the NCBI nucleotide database revealed that the detected minor unexpected DNA sequence is identical to a previously reported nuclear insertion of mitochondrial DNA found on the short arm of chromosome 11 (NCBI accession #HE613849.1)[2]. This insert contains the HV1b primer binding sites, consistent with our finding of a co-amplification event (see Figure 3).

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 16189: T/A | 0.87 | 1.47 | 1.20 | 1.11 | 0.92 | 1.22 | 0.57 | 0.74 | 0.43 | 0.74 |
| **16193: C/T** | **97.91** | **0** | **0.92** | **1.14** | **2.18** | **6.74** | **94.40** | **96.34** | **95.87** | **97.47** |
| **16195: T/C** | **99.07** | **0** | **0.93** | **1.23** | **2.17** | **6.88** | **95.36** | **97.56** | **98.54** | **98.74** |
| 16218: C/T | 1.03 | 1.48 | 1.41 | 1.21 | 1.07 | 1.18 | 0.44 | 0.70 | 0.51 | 0.63 |
| **16221: C/T** | **98.50** | **0** | **0.93** | **1.15** | **2.15** | **6.82** | **94.89** | **96.79** | **97.86** | **98.00** |
| **16223:T/C** | **1.54** | **99.72** | **98.93** | **98.69** | **97.91** | **95.55** | **8.02** | **3.24** | **2.43** | **2.36** |
| **16224:C/T** | **1.50** | **99.69** | **98.88** | **98.67** | **97.92** | **95.49** | **7.96** | **3.21** | **2.43** | **2.40** |
| 16230: A/G | 1.05 | 1.54 | 1.49 | 1.29 | 1.13 | 1.34 | 0.47 | 0.76 | 0.55 | 0.73 |
| **16242: C/A** | **98.74** | **0** | **0.92** | **1.14** | **2.15** | **6.82** | **95.17** | **97.32** | **98.32** | **98.46** |
| 16249: T/C | 1.08 | 1.64 | 1.57 | 1.26 | 1.13 | 1.40 | 0.52 | 0.72 | 0.47 | 0.73 |
| 16259:C/A | 1.02 | 1.59 | 1.52 | 1.24 | 1.11 | 1.39 | 0.46 | 0.70 | 0.49 | 0.69 |
| 16263: T/C | 1.00 | 1.64 | 1.53 | 1.31 | 1.14 | 1.44 | 0.49 | 0.72 | 0.49 | 0.71 |
| 16264: C/T | 1.02 | 1.60 | 1.65 | 1.28 | 1.11 | 1.48 | 0.49 | 0.73 | 0.58 | 0.70 |
| **16270: T/C** | **1.68** | **99.90** | **99.40** | **99.01** | **98.22** | **95.84** | **8.25** | **3.27** | **2.52** | **2.57** |
| **16274: G/A** | **0** | **98.80** | **98.50** | **98.44** | **97.36** | **95.28** | **6.88** | **2.36** | **1.31** | **1.01** |
| 16278: C/T | 1.05 | 1.62 | 1.68 | 1.34 | 1.23 | 1.44 | 0.50 | 0.78 | 0.55 | 0.73 |
| 16284: A/G | 1.03 | 1.59 | 1.53 | 1.32 | 1.15 | 1.31 | 0.58 | 0.76 | 0.53 | 0.73 |
| 16288: T/C | 0.54 | 1.06 | 1.01 | 0.83 | 0.63 | 0.89 | 0.33 | 0.43 | 0.06 | 0.54 |
| 16290: C/T | 0.52 | 0.92 | 1.08 | 0.84 | 0.68 | 0.95 | 0.40 | 0.41 | 0.06 | 0.49 |
| 16293: A/C | 0.60 | 1.12 | 1.27 | 0.85 | 0.77 | 1.00 | 0.31 | 0.46 | 0.06 | 0.51 |
| 16301: C/T | 1.09 | 1.60 | 1.58 | 1.27 | 1.21 | 1.33 | 0.46 | 0.76 | 0.53 | 0.68 |
| 16311: T/C | 1.03 | 1.64 | 1.63 | 1.39 | 1.24 | 1.44 | 0.51 | 0.84 | 0.66 | 0.80 |
| 16319: A/G | 1.03 | 1.62 | 1.41 | 1.34 | 1.24 | 1.38 | 0.53 | 0.82 | 0.49 | 0.78 |
| **16352: C/T** | **95.95** | **1.64** | **2.35** | **2.33** | **3.27** | **7.86** | **91.24** | **93.57** | **97.72** | **95.13** |
| 16355: C/T | 0.99 | 1.65 | 1.67 | 1.50 | 1.35 | 1.48 | 0.51 | 0.76 | 0.51 | 0.69 |
| 16356: T/C | 1.03 | 1.67 | 1.54 | 1.37 | 1.30 | 1.43 | 0.50 | 0.76 | 0.60 | 0.73 |
| **16357: T/C** | **0.12** | **98.83** | **98.65** | **98.50** | **97.48** | **95.25** | **7.00** | **2.30** | **1.42** | **1.10** |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 16368: T/C | 1.08 | 1.56 | 1.53 | 1.38 | 1.26 | 1.38 | 0.53 | 0.95 | 0.53 | 0.83 |
| 16390: G/A | 1.02 | 1.56 | 1.72 | 1.25 | 1.26 | 1.48 | 0.55 | 0.80 | 0.62 | 0.67 |
| **Total Coverage** | **6,693** | **6,609** | **9,206** | **9,064** | **8,439** | **8,343** | **8,431** | **8,337** | **4,869** | **11,589** |

*TABLE 9- Roche GS Junior™ 454 Pyrosequencing Mixture Experiment: HV1b Variants detected using Amplicon Variant Analyzer Software. Mixtures were prepared in defined ratios by combining quantified HV1b amplification products from donors 003-CM54 and 015-AM30. Data is shown as the frequency of each variant detected per MID. Highlighted variants are those expected based on the Sanger reference data. Little difference is seen after applying a bidirectional read variant confirmation quality filter.*

**Tissue Comparison Study – HV1a Data**

The AVA software successfully parsed the read library into sub-libraries based on the MID sequence, and allowed for detection of all expected variants across all tissue types. No unexpected variants were detected in whole blood-derived DNA samples, and a single low-level unexpected variant was detected at rCRS position 16199 in the buccal sample from donor 001-CF30. Several unexpected low-level variants were detected across five hair samples from donor 001-CF30. A number of these variants were detected with inconsistency among hair samples, but exhibited similar frequencies above the minimum read detection threshold of 0.26% in both forward and reverse reads, suggesting that they warrant further study. It is possible that these variants are low-level intra-individual differences not previously characterized using Sanger methods. It should be noted that the minimum read detection threshold of 0.26% is a default setting within the AVA software that can easily be modified by the user. Additional research is required to determine the most appropriate value for inclusion of variants that represent true biological variation at low-level frequencies above an identified level of noise.

Varying levels of heteroplasmy were detected across all tissue types at position 16093, a known mutational hotspot[25]. Sanger reference data from whole-blood samples for donor 001-CF30 shows a C at this site with no evidence of heteroplasmy (data not shown). Pyrosequencing data shows a range of variant frequencies from <25% to >99% for the reported transition, with similar frequencies across all tissue-specific forward and reverse reads. It is possible that these variants reveal differences in the level of heteroplasmy between tissues and between samples of the same tissue originating from heteroplasmic mixtures within a single individual. Further studies are ongoing to elucidate the properties of these variants from other sources of variation.

A proportion of the unexpected variants appear as short indels, which cluster around homopolymeric stretches of 3 or more identical nucleotides. As expected, homopolymeric-associated variants tended to have large frequency disparities between forward and reverse reads, with some detected 100% in one direction and 0% in the other direction. This class of variants is expected due to the reported inability of pyrosequencing chemistry to sequence through homopolymeric regions accurately[22-23]. Fortunately, many of these variants can easily be filtered from the data set by applying the bidirectional read quality filter.

Other uncharacterized variants were also detected with an occurrence frequency the default minimum read detection threshold of ≤0.26%. These variants potentially arise as a result of PCR induced base-misincorporations, 454 sequencing chemistry artifacts, or random sampling effects. A default minimum read percentage of 0.26% is a default setting within the AVA software used to filter any variants with frequencies below this threshold. Further analysis is currently being conducted in our laboratory to elucidate the cause(s) of these variants and the appropriate use of such filters.

**Tissue Comparison Study – HV1b Data**

As expected, the same set of NumtS variants detected in HV1b mixture data was also detected in the HV1b amplicons derived from nuclear DNA-rich blood and buccal extracts in the tissue comparison study. All expected NumtS variants were detected consistently in whole blood-derived DNA extracts, however, NumtS variants were only detected in buccal tissue samples after applying modified AVA analysis parameters where the minimum read percentage was changed from the default value of 0.26% to 0% in order to capture all true biological variation (NumtS) with frequencies below 0.26%. It is likely that this discrepancy is due to the amount of template DNA originally amplified (5-20 ng of input DNA from whole blood on FTA® cards versus 1 ng of input DNA from buccal extracts). Supporting this idea is the observation that relatively high input nuclear DNA template concentrations were required for NumtS amplification and subsequent Sanger sequencing (24 – 147 ng of input template DNA).

Four SNP variants were detected at frequencies between 1% and 2% from hair samples at positions consistent with the NumtS insertion, even after employing the lower stringency AVA analysis parameters. However, not all of the NumtS variants appear. This is expected due to the overall low amount of nuclear DNA in hair shaft extracts, and hence the appearance of a subset of these variants from hair samples may be expected due to stochastic sampling effects.

In addition to the expected variants, several unexpected variants were detected inconsistently across tissues in the HV1b data set. Many of these variants could be filtered from the data set by removing those variants with read direction imbalance from the data set, or by increasing the minimum read percentage back to the default setting of 0.26% with the expectation that a majority of NumtS associated variants would also be filtered from the data set. However, if necessary, the presence of the NumtS variants can be confirmed by viewing data in the Global Alignment Consensus tab of the AVA software to verify that they are read-clustered.

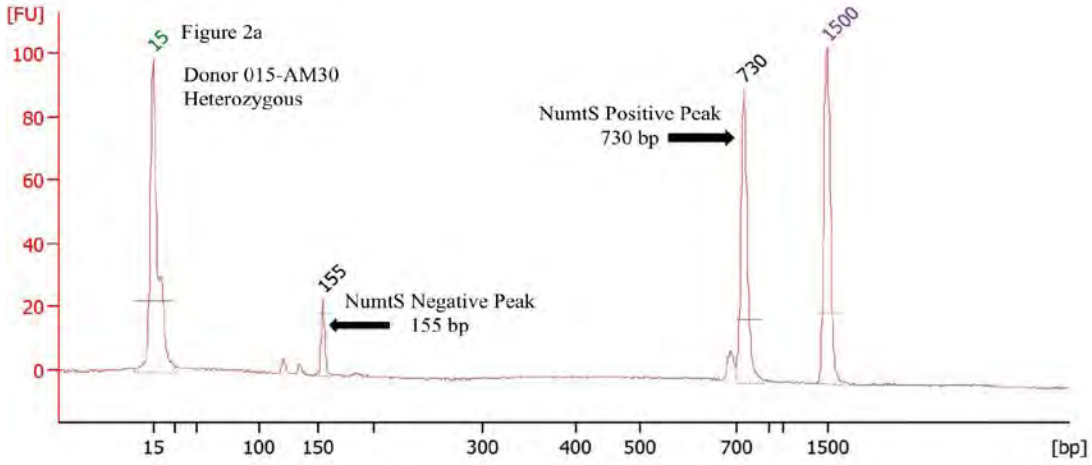
**NumtS – specific amplification and sequencing**

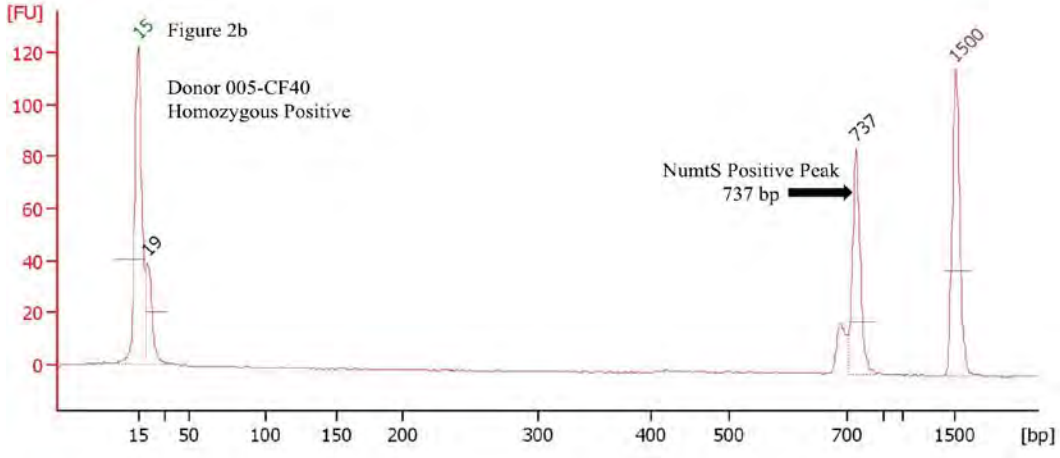
Of the twenty-two donor and control samples amplified, Agilent 2100 Bioanalyzer™ amplicon analysis revealed that thirteen donors were heterozygous for the NumtS insertion, eight were homozygous positive, and one was homozygous negative (Table 10). Those donors whose sample extracts were used for the deep-sequencing mixture analysis and tissue comparison studies (001-CM30, 003-CM54 and 015-AM30B) all possessed the NumtS insertion on at least one chromosome. This evidence further supports the conclusion that the set of variants obtained with HV1b deep-sequencing is due to the expected co-amplification of the NumtS nuclear insertion and mtDNA using this particular primer set, previously reported to be part of the inserted NumtS sequence.

| | | |
|---|---|---|
| 001-CF30 | | X |
| 002-CM32 | X | |
| 003-CM54 | X | |
| 004-CF23 | | X |
| 005-CF40 | | X |
| 006-CM25 | X | |
| 007-CF21 | | X |
| 008-CM23 | X | |
| 009-CF30 | X | |
| 010-CM30 | X | |
| 011-AM24 | X | |
| 012-MM28 | | X |
| 013-CM41 | X | |

| | | | |
|---|---|---|---|
| 014-CF31 | X | | |
| 015-AM30 | X | | |
| 016-CF24 | X | | |
| 017-MM40 | X | | |
| 018-CM26 | | | X |
| 019-UF24 | | X | |
| 020-AF44 | X | | |
| 9947A | | X | |
| HL60 | | X | |

*TABLE 10 - NumtS insertion Data for donors 001-020 and control samples 9947A and HL60, showing the results as either heterozygous or homozygous positive or negative.*



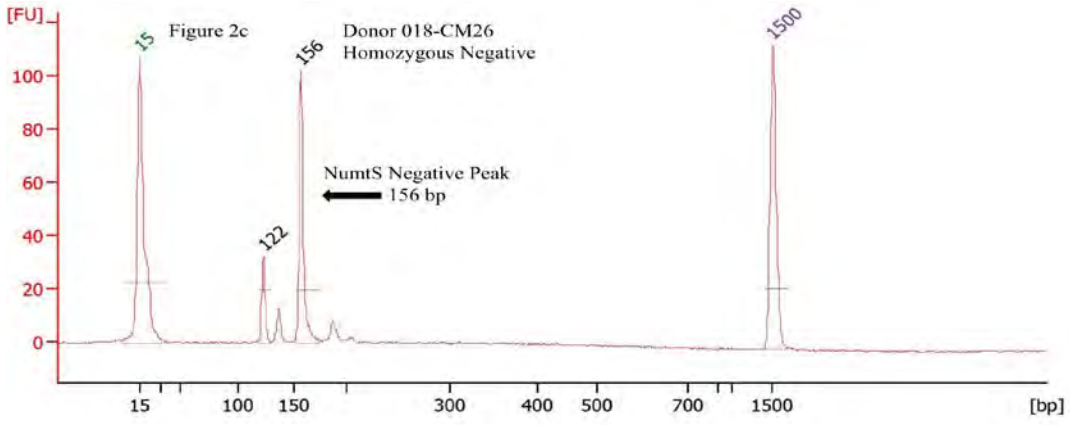



*FIGURE 5 (A-C) - Agilent 2100 Bioanalyzer™ data for NumtS amplification from donors 005-CF40, 015-AM30 and 018-CM26 respectively. Each trace shows an upper and lower molecular weight marker (15 and 1500 bp). Figure 5a shows the amplification products from donor 015-AM30, a heterozygous individual for the NumtS insertion. In this trace, a peak is detected at 155 bp (NumtS negative allele), and a peak at 737 bp (NumtS positive allele). Figure 5b shows the amplification product from a homozygous positive donor 005-CF40, a single peak at 737 bp. Figure 5c shows the amplification results from a homozygous negative donor 018-CM26, resulting in an amplification product at 156 base pairs.*

The only sequence variants observed in this data set are those expected from the reported sequence of the NumtS insertion and also found within the minor unexpected variants from HV1b amplicons. Additionally, no mtDNA-specific polymorphisms were detected in any of the donor sequences amplified with the nuclear-specific primer set, confirming that the source of the sequence data was the NumtS insertion rather than mtDNA.

Dideoxy terminator sequence data obtained using the mtDNA HV1b specific primers aligns to the rCRS for all samples and controls except in the case of individual 018-26M. As seen in Table 10, this individual is homozygous-negative for the NumtS insertion and thus no sequence data aligning to the rCRS is expected to result. The sequence data for homozygous positive individuals obtained using NumtS specific primers (see Table 7) is identical to the NumtS sequence reported by Lang *et. al.*[2]. Individuals heterozygous for the insertion show out-of-phase sequence data after regions of sequence similarity between insertion-positive and insertion-negative amplicons (see Figure 8). This pattern was observed for all thirteen heterozygous donors. Sequence data for homozygous individuals did not exhibit regions in the chromatogram that appeared out-of-phase. Alignment of the 5' and 3' ends of both fragments in the NumtS negative and positive samples reveals that these regions are of nuclear origin, and not part of a mitochondrial DNA insertion as reported previously (Figures 6 and 7)[2].

TTTTCTTTTGTTGATTGAGCAGCATTCCATTGTGGGAAAATACCAAATGCATGGAGAGCTCCCGT
GAGTGGTTAATAGGGTGATAGACCTGTGATCCATCGTGATGTCTTATTTAAGGGGAACGTGTGG
GCTATTTAGACTTTATGGCCCTGAAGTAGGAACCAGATGTTGGATACAGTTCACTTTAGCTACCC
CCAAGTGTTATGGGCCCGGAGCGAGGAAAGTAGCACTCTTGTGCGGGATATTGATTTCACGGAG
GATGGTGGCCAAGGGACTCCTATCTGAGGGGGGTCATCCGTGGGGACGAGAGAGGATTTGACT
GTAATGTGCTATGTACGGTAAATGGCTTTATGTGCTATGTACTATTAAGGGGGGATGGGTCTGTT
GATATTCTAGTGGGTAGGGGTTGGCTTTGGGGTTGCAGTTGATGTGTGACAGTTGAGGGTTAAT
TGCTGTACTTGCTTGTAAGCATGGGGTGGGGGTTTTGATGTGGATTGGGTTTTTATGTACTACAG
GTGGTCAAGTATTTATGGTACTGTACAATATTCATGGTGGCTGGCAGTAATGTACGAAATACTA
TGGATTGTTTATTCACTCTTCTGTTAGAAACC

*FIG. 6: - NumtS sequence reported by Thomas et. al.[3] Highlighted regions at the 5' and 3' ends are nuclear specific, and appear in both NumtS positive and NumtS negative fragments amplified with nuclear specific primer sets.*

Sanger sequencing data obtained from Donor #18 amplified material is shown in Figure 7 as a DNA sequence. As expected, the inserted sequence is missing. Figure 8 shows the traces from heterozygous donors. As expected, the sequence data is out-of-phase due to the presence of the insert as a heterozygote and hence the template is of mixed-length.

TTGTATGTATTAGTTTTTTTCTTTTGTTGATTGAGCAGCATTCCATTGTGGGATTATACTATGGAT
TGTTTATTCACTCTTCTGTTAGAAACCTGGACTTTGTA

FIG. 7: - *NumtS negative sequence showing flanking DNA without the inserted sequence.*

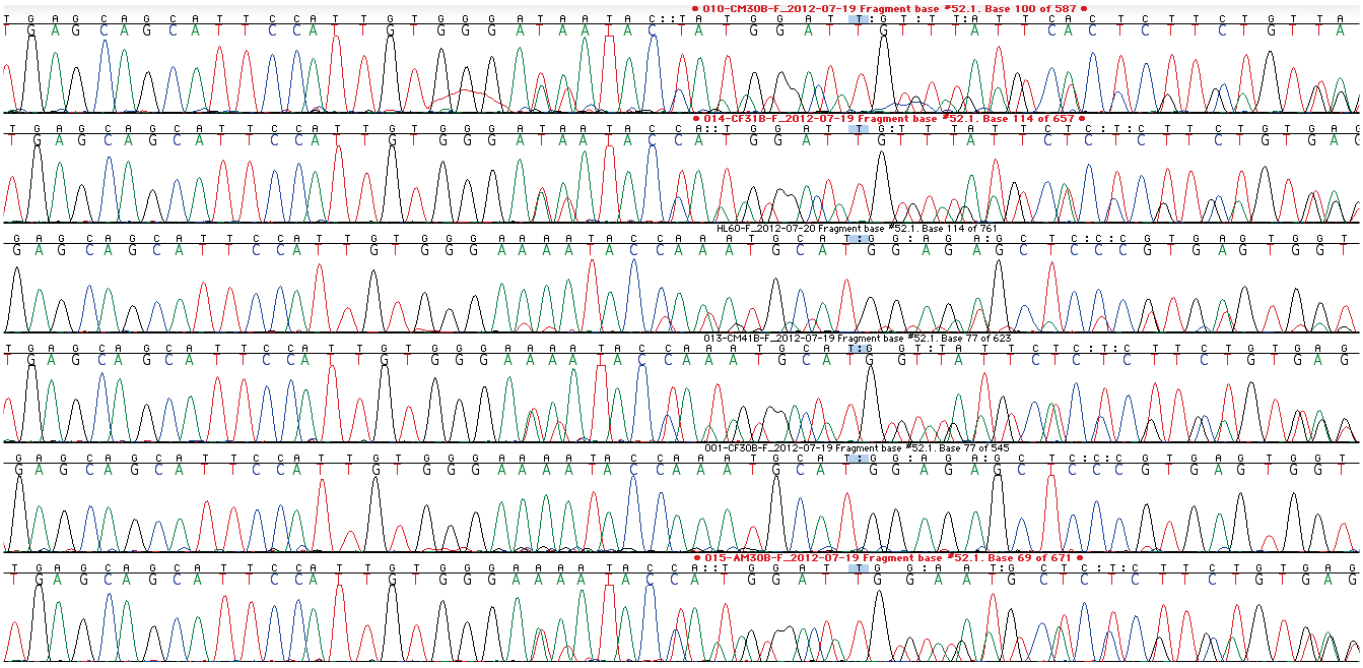


*FIG. 8: - Sanger sequence data for NumtS heterozygous donors. Resulting sequences were aligned against the full NumtS insertion. Data is out-of-phase in dissimilar areas between NumtS positive and NumtS negative sequences. Sample 001-CF30 (fifth sequence from the top) is homozygous positive for the insertion, with no out-of-phase data seen.*

**Discussion – NumtS Identification Experiments**

Targeted deep-sequencing of the human mtDNA hypervariable region allows for an increase in the depth of mtDNA analysis and hence enhances the ability to detect minor variants present in the sequence. The goal of forensic validation studies is to fully understand all the potential sources of variation, including those that arise from true genetic differences in the sample and those that are artifacts from the technique or chemistry, such that a proper interpretation can be made in a forensic case.

We have shown that minor SNP variants can be reproducibly and accurately detected at a level of 1% or lower in multiple deep-sequencing runs using the Roche GS Junior™ 454 pyrosequencing instrument. In addition to expected minor variants, a subset of read-clustered, unexpected variants was also detected in HV1b reads originating from nuclear DNA-rich blood and buccal cell samples. These data show that these variants are amplified products from a nuclear pseudogene, an ancient nuclear insertion of mitochondrial

DNA that contains HV1b control region primer-binding sites. These NumtS sequences, to our knowledge, have never been co-detected in mitochondrial sample preparations using Sanger sequencing, further supporting the assertion that the sensitivity of low-level variant detection reported herein is a vast improvement over current methodologies. Additionally, the data show slight differences in the amounts at which the unexpected variants are detected in whole blood, buccal, and hair samples. Because amplification of the NumtS using either an mtDNA HV1b or NumtS specific primer set requires a relatively high amount of input template, this difference is likely due mostly to the tissue-type-dependent amounts of the NumtS template present in the particular sample.

From a forensic perspective, the presence or absence of the NumtS sequence in a sample should accordingly vary with a host of factors, including the amount of sample DNA amplified and the individual's genotype with respect to the NumtS itself, and hence not be expected a priori to be present in each analysis. Consequently, Until a fuller characterization of the genetic variation within each NumtS is elucidated, we recommend that at this time no direct comparative utility be made of the NumtS insertion itself.

As expected, deep-sequencing of HV1b from hair shaft samples does not give rise to the read-clustered, unexpected variants observed in blood and buccal tissue. We attribute this observation to the known scarcity of nuclear DNA present in hair shaft DNA extracts. However, additional inconsistent, unexpected variants were detected in all control region amplicons from hair samples that were also absent in blood and buccal tissue from the same individual. Some of these low-level variants may represent low-level intra-individual variation, or low-level heteroplasmy. Alternatively, variants may, in some cases, arise as a result of 454 pyrosequencing-specific chemistry, and may be omitted from the data set with increased stringency quality filter settings. Additional research is needed to elucidate the cause(s) of these low-level variants.

Low-level heteroplasmy is readily detected in samples using NGS, because individual template molecules are interrogated independently of one another. This is observed at position 16093 in HV1a data from the tissue comparison study. As a result, crime laboratories wishing to implement next-generation sequencing into their mtDNA analysis workflow should reevaluate interpretational criteria accordingly. It should be noted in this regard that using mtDNA interpretation guidelines in place in forensic laboratories, which include consideration of potential variation, all of the blood, buccal and hair samples amplified and sequenced in this study from a single individual would be interpreted as an inclusion, since they all share the same predominant DNA sequence. Hence we anticipate no significant interpretational complications due to the co-detection of NumtS sequences in forensic casework. Further, mtDNA evidentiary samples often lack or contain very low amounts of nuclear DNA, and hence are not expected to reveal any NumtS specific variants. In either case, the NumtS insertion sequence appears to be highly conserved in most cases, and hence the ability to detect the NumtS insertions can be included in further validation studies of NGS involving mtDNA. Additionally, if warranted in a particular context and sample-type, we have shown that confirmation of NumtS co-detection can be accomplished with nuclear DNA-specific amplification of a suspected NumtS insertion.

It has been reported that much of the mtDNA genome has been inserted as small fragments into various positions within the nuclear genome, creating a mitochondrial pseudo-genome[25]. Based on these findings, it is probable that expansion of mtDNA analysis beyond the control region using NGS will result in further co-detection of additional NumtS specific variants, particularly from nuclear DNA-rich reference samples. Several publications have reported incorrect characterization of co-amplified NumtS as heteroplasmy [25]. Careful characterization of NumtS that are co-detected with whole mtDNA genome data will result in better identification and resolution of true mtDNA heteroplasmic mixtures, and hence provide a further conceptual foundation to continue to correctly interpret mtDNA comparisons in the future.

## Results – Section 3 - Improving the DNA Extraction Efficiency from Hair Shaft Samples

We have optimized a protocol, described in detail below, to extract mtDNA from hair shaft. Our goal was to maximize the mtDNA extracted from two centimeters of hair shaft so that more mt-genome sequence information may be obtained from a challenging sample, leading to a higher discriminatory power of mtDNA analysis. The optimized extraction method shows, on average, a fourteen-fold increase in mtDNA concentration when compared with traditional manual grinding and organic extraction methods. The optimized method is also less time-consuming, and fewer hands-on steps and tube transfers are required, which reduces the risk of contamination. PCR inhibitors are successfully removed as indicated from qPCR studies and subsequent amplification. The protocol is robust, and has been shown to be effective in multiple analyst's hands.

### Collection and Cleaning of Hair Samples

Hair samples were obtained from donors and examined microscopically for the presence of root material. When necessary, root material was discarded and two centimeters of hair shaft were carefully measured and cut. Hair fragments were subjected to cleaning via sonication in 5% Terg-a-zyme for 20 minutes, followed by a brief rinse in ethanol, then molecular biology grade water.

### Extraction Protocol Comparison Studies

Eleven hair fragments were subjected to a traditional manual grinding / organic extraction method as described by Wilson *et. al.* 1995, and the Federal Bureau of Investigation mtDNA Analysis Protocol. Eleven hair fragments were processed using the newly developed extraction method, which employs a combination of Qiagen® QIAamp® DNA Investigator and Applied Biosystems® PrepFiler® Forensic DNA Extraction kit-based methods. In this method, the cleaned hair fragment is placed in a digestion solution of 300 µl Qiagen® Buffer ATL, 20 µl 0.6 U/µl Proteinase-K, and 20 µl 1M DTT. Digestion is generally complete (no visible fragments) after one hour incubation at 56°C, 900 rpm on a thermal shaker. Qiagen® Buffer AL (300 µl) is then added to the digestion solution and incubated at 70°C, 900 rpm for 10 minutes. The digested sample is then subjected to DNA extraction according to the PrepFiler® Forensic DNA Extraction Kit protocol beginning with the addition of 15 µl PrepFiler® Magnetic Particles.

### Real-Time qPCR mtDNA Quantitation from Hair Shaft DNA Extracts

All purified hair extracts were quantified using a custom real-time quantitative PCR (qPCR) assay specific for human mtDNA (Kavlick *et. al.* 2011). Results are summarized in Table 12.

| **Organic Extraction** | | **Optimized Extraction** | |
|---|---|---|---|
| Hair Sample | Total mtDNA extracted (copies) | Hair Sample | Total mtDNA extracted (copies) |
| 1 | 2,850 | 1 | 433,890 |
| 2 | 4,080 | 2 | 367,860 |

| | | | |
|---|---|---|---|
| 3 | 23,520 | 3 | 304,080 |
| 4 | 12,180 | 4 | 435,210 |
| 5 | 41,100 | 5 | 270,330 |
| 6 | 19,650 | 6 | 801,570 |
| 7 | 43,530 | 7 | 293,070 |
| 8 | 7,170 | 8 | 103,110 |
| 9 | 36,300 | 9 | 425,610 |
| 10 | 40,770 | 10 | 214,410 |
| 11 | 43,770 | 11 | 180,480 |
| **Average** | **24,993** | **Average** | **348,147** |

*Table 11. Comparison of mtDNA extracted from eleven hairs using organic extraction and eleven hairs using an optimized extraction method.*

We reasoned that an increase in extraction efficiency, as shown using the modified DNA extraction protocol, would have positive effects on all the downstream steps in the analysis scheme. This, in fact, proved to be the case. We subjected the extracts to PCR amplification using both the traditional control region amplification primer sets, as well as an expanded set designed to amplify more of the human mt-genome. Our approach was to develop a multiplex amplification scheme that uses non-overlapping primer sets spread across the entire mt-genome. We also were curious as to whether or not a pre-PCR non-specific amplification step, such as whole genome amplification, could increase the success rate and lead to our ultimate goal of achieving whole mt-genome data from hair shafts.

Table 13 shows the results of a direct comparison study performed with the traditional grinding/PCIA extraction method against the improved Qiagen/PrepFiler method.

| **Sample ID** | **PCIA Extraction (mtDNA copies/2 μL)** | **Qiagen Extraction (mtDNA copies/2 μL)** |
|---|---|---|
| Hair 22 | 1,360 | |
| Hair 93 | 1,459 | |
| Hair 411 | 1,482 | |
| Hair EB1 | | 4,715 |
| Hair EB2 | | 4,685 |
| Hair EB3 | | 2,648 |

*Table 12. DNA Extraction Efficiency Study: Organic Extraction with manual grinding versus Qiagen QIAamp DNA Investigator Kit using chemical digestion. Data was generated using a human mtDNA-specific real-time PCR assay as described herein. DNA extracted from hairs using the organic extraction method were eluted in a volume of 35 μL, while Qiagen-extracted DNA was eluted in a volume of 50 μL. DNA from hair 22 was extracted by one analyst; DNA from hairs 93 and 411 was extracted by a different analyst. We still see a high degree of consistency in resulting concentrations using the PCIA method. Hairs 22, EB1, EB2 and EB3 were obtained from the same donor. DNA from all of these hairs were extracted by the same analyst. All hairs were carefully measured and 2 cm fragments were used for analysis.*

**The Application of Whole Genome Amplification to Extracted DNA Samples**

With this improvement in DNA extraction efficiency, we have also seen some limited promise in the ability to use whole genome amplification (WGA) to further increase the analyzable amount of mtDNA from challenging forensic samples for downstream applications.

Hair extracts were also subjected to whole genome amplification (WGA) using the Qiagen® RepliG® Mini Kit, a multiple displacement amplification (MDA) based kit according to the manufacturer's recommendations. Efforts to pre-amplify mtDNA using the Qiagen® RepliG® Mini Kit were largely unsuccessful when using extracts processed with manual grinding and organic extraction. However, a 2-fold to 41-fold increase in mtDNA concentration has been achieved when using extracts processed with the optimized method (Table 14).

| **Organic Extraction** | | | | **Optimized Extraction** | | | |
|---|---|---|---|---|---|---|---|
| **Sample ID** | **WGA Input (copies / µl)** | **Post-WGA Output (copies / µl)** | **Fold Increase** | **Sample ID** | **WGA Input (copies / µl)** | **Post-WGA Output (copies / µl)** | **Fold Increase** |
| **Hair 1** | 69 | 105 | 1.5 | **Hair 1** | 493 | 2,438 | 4.9 |
| **Hair 2** | 132 | 238 | 1.8 | **Hair 2** | 940 | 38,971 | 41.5 |
| **Hair 3** | 68 | 146 | 2.1 | **Hair 3** | 424 | 875 | 2.0 |
| **Hair 4** | 74 | 169 | 2.3 | **Hair 4** | 531 | 15,065 | 28.4 |
| **Hair 5** | 73 | 105 | 1.4 | **Hair 5** | 1336 | 6,945 | 5.2 |
| **Positive Control DNA (HL60)** | 94 | 732,981 | 3,909 | **Positive Control DNA (HL60)** | 368 | 2,390,000 | 3,249 |
| **Reagent Blank** | 0.18 | 4.3 | 2.1 | **Reagent Blank** | 0.32 | 0 | 0 |
| **Negative Control** | 0 | 4 | 4 | **Negative Control** | 0 | 1.4 | 1.4 |

*Table 13. Comparison of whole genome amplification studies using two DNA extraction methods The optimized DNA extraction method resulted in the recovery of higher quantities of DNA in all instances and also generally better supported WGA as well.*

As shown in Tables 14 and 15, hair shaft DNA extracts do not support the robust whole genome amplification levels that are observed with pristine DNA samples. This is to be expected by the compromised nature of the DNA found in many forensic samples. However, we did observe a slight increase in the amount of template following WGA, leading us to posit that perhaps sufficient template exists following WGA to support the next step of the process, namely multiplexed amplification of the entire mt-genome.

| Sample ID | WGA Input (copies mtDNA/µL) | Post-WGA (non-specific) (copies/µL) | Fold Increase Non-specific kit | Post-WGA (mtDNA) (copies/µL) | Fold Increase mtDNA kit |
|---|---|---|---|---|---|
| Hair 1 | 236 | 3,771 | 16 | 193 | -1.2 |
| Hair 2 | 234 | 491 | 2 | 199 | -1.2 |

| | | | | | |
|---|---|---|---|---|---|
| Hair 3 | 132 | 238 | 2 | 115 | -1.2 |
| Reagent Blank | 2.3 | 13.73 | 6.5 | 6.7 | 3 |
| HL60 + Control | 2,134 | 3,749,429 | 1,757 | 71,120,408 | 40,478 |
| Negative Control | 0 | 4 | 4 | 7.6 | 7.6 |

*Table 14. WGA Results from Hair Extracts using both Qiagen Repli-g® Non-Specific MDA Kit, and Qiagen Repli-g® mtDNA MDA Kit as recommended by the kit manufacturers.*

**Results – Section 4 Generating Rapid Whole mt-Genome Information from Reference Samples**

We are able to generate whole mtGenome sequencing data from robust DNA samples, such as buccal swabs, using a long PCR amplification technique with two overlapping primer sets that cover the entire genome. When tagmented with Illumina Nextera XT®, a library preparation technique designed for Illumina® sequencing platforms, it is simple to generate high-throughput sequencing data from these samples. The process is simple and fast. When combined with Illumina Nextera XT®, samples are rapidly tagmented, normalized and ready for sequencing. Sequencing data can be quickly obtained from the instrument software and is sufficient to determine the presence of expected variants, however these data can be further analyzed with second-party online freeware using a custom analysis pipeline, and subsequently viewed in a genome browser.

Buccal samples have been obtained from eight individual donors, according to IRB standards. DNA was extracted from the obtained buccal swabs using the Qiagen DNA extraction kit. Mitochondrial DNA present in the extracts was amplified by long PCR and then tagmented with the Illumina® Nextera XT® DNA Sample Preparation Kit. The resulting samples were then sequenced using the Illumina MiSeq® instrument. Reference data from the whole mtDNA genome for comparison purposes was obtained with the Life Technologies, Inc., Applied Biosystems® mitoSEQr™ kit according to the manufacturer's recommendations.

**Long PCR**

MtDNA present in the buccal cell extracts was quantified using the mtDNA real time PCR protocol (Kavlick *et. al.* 2011) and 200,000 copies of mtDNA were used as input for a long PCR using the TaKaRa LA Taq™DNA Polymerase mix. TaKaRa™consists of a mixture of a *Taq* polymerase and a proofreading polymerase with 3′-5′ exonuclease activity. This system has a reported fidelity that is 6.5 times higher than conventional *Taq* polymerase, and is routinely used to generate amplicons of 20 kb, with less frequent amplification up to 48 kb. Two primer sets in two separate reactions will be used to generate two amplicons of 9065 bp and 11170 bp, overlapping at the control region. The primers for short amplicon were: (F1) 5'aaagcacataccaaggccac 3' and (R1) 5' ttggctctccttgcaaagtt 3'. The primers used for the long amplicon were (F2) 5' tat ccg cca tcc cat aca tt 3' and (R2) 5' aat gtt gag ccg tag atg cc 3'. The short and long amplicons were amplified in separate reactions and later pooled for tagmentation and sequencing. The 50 ul PCR reactions were prepared as shown in Tables 16 and 17.

| **Reagent** | **Quantity** |
|---|---|
| Template | 200,000 copies of mtDNA in 5 µl |
| Fw primer | 1 µl of 10 mM stock (0.2 uM) |
| Rv primer | 1 µl 0f 10 mM stock (0.2 uM) |
| TaKaRa DNA Polymerase | 0.5 µl (2.5U) |

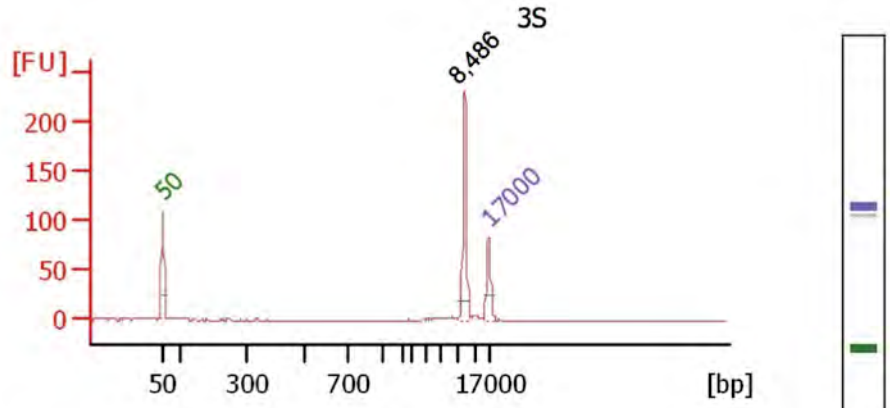
| | |
|---|---|
| dNTPs mix | 8 µl (0.4 mM) |
| Water (MBG) | 29.5 ul |

*Table 17: PCR Reaction Conditions for Long PCR. The volume is 50 ul.*

| | |
|---|---|
| 94°C 1 min | |
| 94°C 30 sec | 30 cycles |
| 54°C 15 sec | |
| 68°C 11 min | |
| 72°C 10 min | |

*Table 18: Thermal cycling parameters for Long PCR on the ABI Veriti® thermal cycler.*

The Agilent 2100 Bioanalyzer™ trace shown in Fig. 9 show an example of the successful long PCR amplification of a buccal sample. Long PCR amplification usually generates 5-12 ng of DNA per amplicon with 200,000 copies of input mtDNA.




**Overall Results for sample 6 : 3S**

Number of peaks found: 1

**Peak table for sample 6 : 3S**

| Peak | Size [bp] | Conc. [ng/µl] | Molarity [nmol/l] | Observations |
|---|---|---|---|---|
| 1 | 50 | 8.30 | 251.5 | Lower Marker |
| 2 | 8,486 | 9.77 | 1.7 | |
| 3 | 17,000 | 4.20 | 0.4 | Upper Marker |

*Figure 9: Bioanalyzer™ trace of long amplification products. The amplification product is shown as the peak labeled 8,486. Standard peaks at 50 and 17000 base pairs are shown in green and purple, respectively.*

In order to follow the progress of the tagmentation reaction, we utilize the Agilent 2100 Bioanalyzer™ to analyze the PCR products both before and after tagmentation as shown in Figures 10 and 11. Following the process of tagmentation, the large amplicon peaks disappear, and a large, amorphous distribution of fragments appears, indicating that the amplicons have been successfully processed for direct NGS on the Illumina® platform.

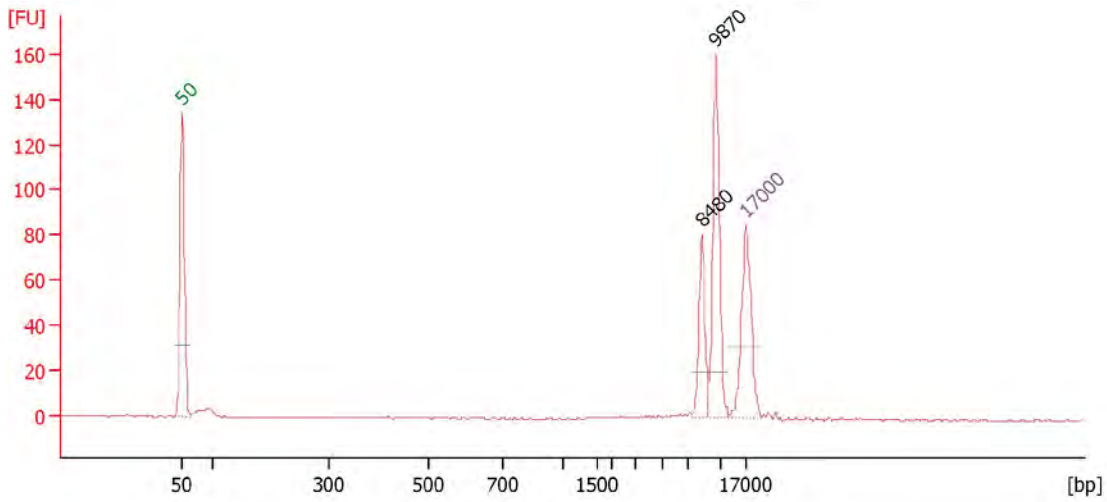



*Figure 10. Agilent™ 2100 Bioanalyzer™ traces demonstrating long PCR performed with two primer sets, resulting in two large, overlapping amplicons that span the entire human mtDNA genome.*

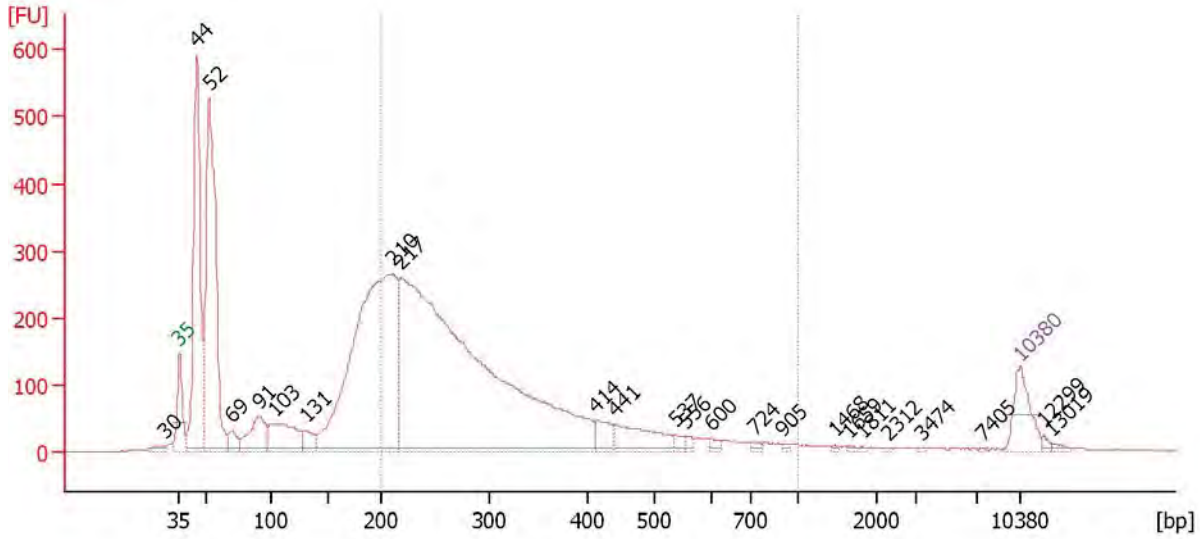



*Figure 11. Monitoring the fragmentation of two large amplicons following treatment with Nextera™ transposome complex. The majority of the fragments are 200 – 400 bp in size. The sizing standards are 35 base pairs (green) and 10380 base pairs (purple).*

**Rapid library preparation with Illumina Nextera XT® and sequencing**

The two long amplicons were pooled together, normalized to ensure the same number of short and long molecules were present in the sample, and approximately 1 ng of input of this pooled product was processed with Illumina® Nextera® XT protocol. In addition to the eight buccal samples and their controls, three NIST Human Mitochondrial DNA Standards and ten NIST Mixture Standards were also diluted to 0.2 ng/µl according to their stated quantity by NIST, and 1 ng of each sample was processed. A different index sequence was used for each sample, totaling 26 samples. Each sample, and a PhiX control, was quantified with the Qubit® ssDNA Assay Kit, according to the manufacturer's recommendations. All samples were pooled together, with a 20% PhiX spike-in, and run on the Illumina® MiSeq®.

Sequencing data from donor 002 was obtained from Illumina®'s MiSeq Reporter software (Table

17). All expected variants are presented in the data (shown in yellow). Some misalignments around regions with insertions or deletion cause incorrect variant calls (shown in orange) but these can be easily detected due to lower quality and filtering scores as well as a drop in coverage, and subsequently ignored if necessary. One variant, 10398, that was observed in the NGS data was found in an area not covered in our Sanger reference sequence (shown in bold).

| Position | Variant Type | Call | Frequency | Depth |
|---|---|---|---|---|
| 73 | SNP | A->AG | 1 | 11289 |
| 152 | SNP | T->TC | 1 | 16694 |
| 199 | SNP | T->TC | 1 | 10632 |
| 204 | SNP | T->TC | 1 | 9558 |
| 207 | SNP | G->GA | 1 | 9341 |
| 250 | SNP | T->TC | 1 | 5959 |
| 263 | SNP | A->AG | 1 | 4512 |
| 302 | Indel | -/C | 0.91 | 1755 |
| 310 | Indel | -/C | 1 | 2043 |
| 310 | SNP | T->TC | 0.6 | 1652 |
| 567 | Indel | ---/CCC | 0.49 | 1781 |
| 567 | SNP | A->AC | 0.32 | 1781 |
| 567 | Indel | --/CC | 0.19 | 1781 |
| 750 | SNP | A->AG | 1 | 18207 |
| 1438 | SNP | A->AG | 1 | 25567 |
| 1719 | SNP | G->GA | 1 | 24450 |
| 2706 | SNP | A->AG | 1 | 6461 |
| 2835 | SNP | C->CA | 1 | 11764 |
| 3106 | Indel | N/- | 0.94 | 10710 |
| 4529 | SNP | A->AT | 1 | 10163 |
| 4769 | SNP | A->AG | 1 | 11051 |
| 7028 | SNP | C->CT | 0.99 | 13846 |
| 7055 | SNP | A->AT | 1 | 12759 |
| 8251 | SNP | G->GA | 1 | 9854 |
| 8860 | SNP | A->AG | 1 | 15077 |
| 9548 | SNP | G->GA | 1 | 9238 |
| 10034 | SNP | T->TC | 1 | 7260 |
| 10238 | SNP | T->TC | 1 | 6587 |
| **10398** | **SNP** | **A->AG** | **1** | **8828** |
| 11065 | SNP | A->AG | 1 | 11494 |
| 11719 | SNP | G->GA | 1 | 14204 |
| 12501 | SNP | G->GA | 1 | 8863 |
| 12705 | SNP | C->CT | 1 | 11234 |
| 13780 | SNP | A->AG | 1 | 4520 |
| 14766 | SNP | C->CT | 1 | 12300 |
| 15043 | SNP | G->GA | 1 | 16826 |
| 15326 | SNP | A->AG | 1 | 28723 |
| 15673 | SNP | A->AG | 0.83 | 26461 |
| 15758 | SNP | A->AG | 1 | 26543 |
| 15924 | SNP | A->AG | 1 | 20390 |
| 16074 | SNP | A->AG | 1 | 20066 |
| 16129 | SNP | G->GA | 0.99 | 23467 |
| 16145 | SNP | G->GA | 1 | 24327 |

| | | | | |
|---|---|---|---|---|
| 16223 | SNP | C->CT | 0.99 | 31446 |
| 16391 | SNP | G->GA | 1 | 31781 |
| 16519 | SNP | T->TC | 1 | 11915 |

*Table 17: Sequencing data from donor 002, obtained from Illumina®'s MiSeq Reporter. The frequency value for variants is rounded using this software, potentially losing valuable information on potential mixed positions.*

One particular position of interest, 15673, shows what appears to be a low-level mixture at position 15673. As shown in Table 19, there are 1355 reads that contain an 'A' at this position (17.9% of the total). The electropherogram from the Sanger sequence from this donor is shown in Figure 12. Here a slight mixture can be discerned at 15673, but not as clearly as depicted in the NGS data. What is intriguing is the presence of other low-level variants, all less than 1%, shown for instance at positions 207, 11719, 15043, 8251, etc. Two control region variants that define major clades, 16129 and 16223, are also variable at low levels, in the case of 16129, the minor variant is present at about 1.2%. Whether or not these low-level variants are in fact true biological, heteroplasmic variation is left for future experimental efforts.

| Chrom | Pos | Ref | Total reads | A | C | G | T | Quality adjusted reads | Total variants | % Variant |
|---|---|---|---|---|---|---|---|---|---|---|
| rCRS | 207 | G | 2894 | 2822 | 0 | 7 | 0 | 2829 | 2822 | 99.75256 |
| rCRS | 11719 | G | 7862 | 6946 | 0 | 20 | 1 | 6967 | 6947 | 99.71293 |
| rCRS | 15043 | G | 7730 | 7227 | 0 | 22 | 0 | 7249 | 7227 | 99.69651 |
| rCRS | 8251 | G | 6880 | 6185 | 1 | 19 | 1 | 6206 | 6187 | 99.69384 |
| rCRS | 16391 | G | 7847 | 7470 | 0 | 24 | 0 | 7494 | 7470 | 99.67974 |
| rCRS | 250 | T | 2566 | 0 | 2094 | 0 | 7 | 2101 | 2094 | 99.66683 |
| rCRS | 204 | T | 2864 | 0 | 2793 | 2 | 10 | 2805 | 2795 | 99.64349 |
| rCRS | 7028 | C | 7488 | 3 | 29 | 0 | 7139 | 7171 | 7142 | 99.59559 |
| rCRS | 16223 | C | 7970 | 7 | 50 | 0 | 7335 | 7392 | 7342 | 99.32359 |
| rCRS | 16129 | G | 7369 | 7106 | 1 | 86 | 2 | 7195 | 7109 | 98.80473 |
| rCRS | 15673 | A | 7907 | 1355 | 0 | 6211 | 2 | 7568 | 6213 | 82.09567 |

*Table 18. List of variants sorted by percent major base from donor 002 (100% omitted). The A calls at position 15673 are highlighted in red. These variants may rise above a threshold for detection, the existence of which requires further experimental development.*


*Figure 12. Electropherogram results from Sanger sequencing of donor 002. Position 15673 appears to harbor both A and G bases with G predominating. This subtle variation is much more pronounced when observed using NGS technologies as shown above in Table 18.*

whole mt-genome data can be visualized with a number of software packages. The Integrated Genome Viewer (IGV), a software package developed by the Broad Institute, it can be observed that sequencing generates a depth of approximately 15,000 reads throughout the entire genome (Fig. 13).

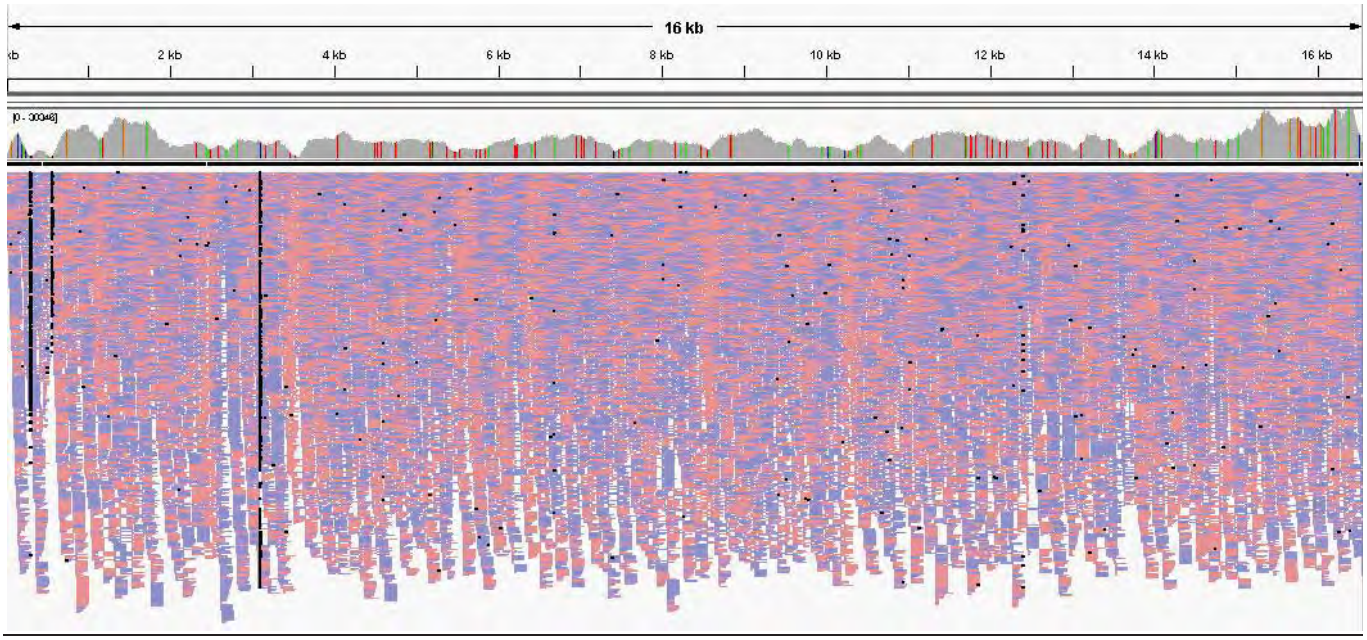



*Figure 13: Whole genome sequencing data from donor 002 as viewed in the Integrative Genome Viewer 2.1.30. Individual reads are condensed and graphically depicted to show the depth of coverage across the region of interest.*

**Results – Section 5 DNA Sequencing of NIST Standards**

We obtained two sets of Standard Reference Material (SRM) prepared by the National Institute of Standards & Technology (NIST) for sequencing on the Illumina® MiSeq™. SRM 2392 consists of three samples: two DNA extracts from the highly characterized lymphoblastoid cell lines CR and GM09447A, and one sample that contains the mitochondrial HV1 sequence amplified from CHR. SRM 2934 is a mtDNA heteroplasmy detection standard series, in which a 285 bp region from CHR and GM09447A is mixed at 10 known ratios. In this region, both cell lines differ from the rCRS by one base (T/C) at the same nucleotide position. Thus, the mixtures simulate different degrees of heteroplasmy. Both SRMs are used as internal controls to evaluate sequencing efficiency and accuracy. In this study, the mixture standards were fragmented with Illumina® Nextera® XT without any prior amplification, and run on the Illumina® MiSeq™.

The ten NIST Mixture Standards (Table 19) all showed very high coverage, with an average depth of approximately 400,000 reads. The expected variant frequencies that we obtained were close to those reported by NIST by the Illumina® MiSeq™ Reporter frequency output as shown in Table 21.

| | Expected | Called | Frequency | Depth |
|---|---|---|---|---|
| NISTMix1 | T | C->CT | 0.99 | 274,063 |
| NISTMix2 | C | - | - | - |
| NISTMix3 | 0.5 T | C->CT | 0.52 | 468,301 |
| NISTMix4 | 0.4 T | C->CT | 0.41 | 467,779 |
| NISTMix5 | 0.3 T | C->CT | 0.31 | 566,795 |
| NISTMix6 | 0.2 T | C->CT | 0.22 | 381,094 |
| NISTMix7 | 0.1 T | C->CT | 0.11 | 480,589 |
| NISTMix8 | 0.05 T | C->CT | 0.06 | 319,786 |
| NISTMix9 | 0.025 T | C->CT | 0.04 | 39,130 |
| NISTMix10 | 0.01 T | C->CT | 0.02 | 727,996 |

*Table 19. NIST Mixture Standards and the minor variant detection results obtained with the Illumina® MiSeq®.*

**Results – Section 6 NGS Chemistry Comparison Study**

An NGS instrumentation comparison study was completed to elucidate sources of noise associated with sample preparation, sequencing chemistry, between the Roche GS-Junior and the MiSeq platforms. Three hairs were obtained from different areas of the scalp of donor 014-CF31. Hairs were observed using a stereomicroscope. Follicular tags, if present, were removed and discarded, and adjacent 2 cm hair shafts were obtained for analysis. DNA was extracted from the hair shafts using an optimized extraction protocol developed in our laboratory. The extracts were quantitated in duplicate using qPCR, and the HV1b region of mtDNA (rCRS positions 16,159-16,391) was amplified in triplicate using both Roche Fusion Primers, and unmodified primers developed by the FBI. The purpose was to compare the DNA sequencing results from amplicons containing the longer Roche Fusion Primers to those lacking this extraneous sequence, such as those resulting from the use of the standard primers found in the FBI Laboratory protocols.

Four independent NGS runs were performed. Initially, the 345 bp amplicons generated using Roche Fusion primers with pyrosequencing adapters and multiplex identifiers were deep-sequenced using the Roche GS Junior™ in two independent runs. To enable direct comparison of NGS platforms, and to determine the tagmentation efficiency of the Nextera™ library preparation on amplicons of various sizes, the same 345 bp

amplicons, and the 270 bp amplicons generated using the unmodified primers were deep-sequenced using the Illumina® MiSeq™ in two independent runs. Figure 14 shows a flowchart describing the overall workflow for this study.

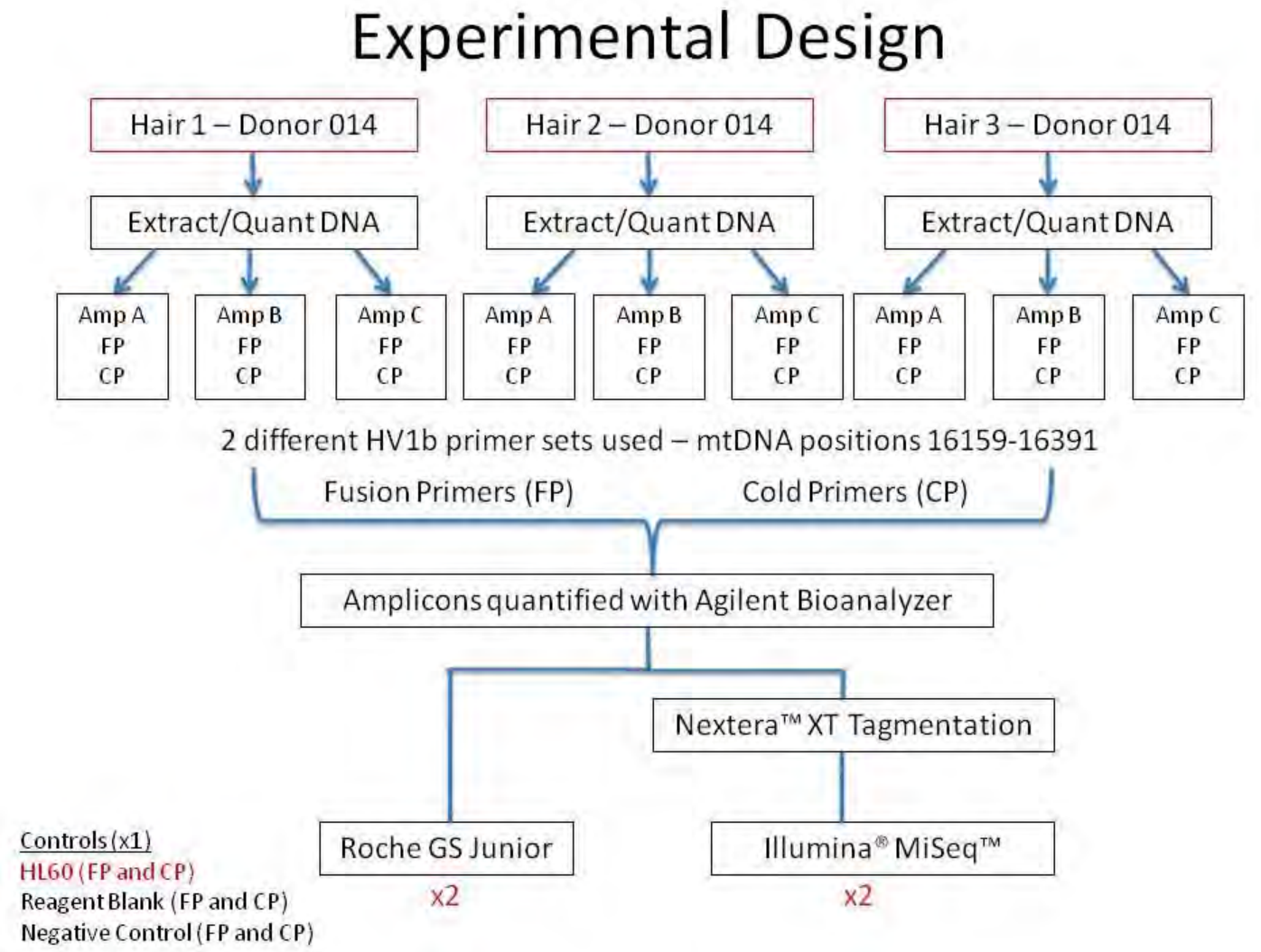



*Figure 14 – Next-Generation Sequencing Instrumentation Study – Experimental Design. This figure shows the overall workflow for the experiment described herein. Three hairs were obtained from a single donor. DNA was extracted from 2 cm hair shafts and amplified in triplicate using two different primer sets. The samples amplified using Roche fusion primers with pyrosequencing adapters and indices were run twice in independent runs on both the Roche GS Junior™ and the Illumina® MiSeq™. The samples generated using the unmodified primers were run twice on the Illumina® MiSeq™ only. HL60 positive controls, negative controls and reagent blanks were included in all 4 runs.*

| | |
|---|---|
| 014A-1 | 014A-1 |
| 014A-2 | 014A-2 |
| 014A-3 | 014A-3 |
| 014B-4 | 014B-4 |
| 014B-5 | 014B-5 |
| 014B-6 | 014B-6 |
| 014C-7 | 014C-7 |

| | |
|---|---|
| 014C-8 | 014C-8 |
| 014C-9 | 014C-9 |
| RB-10 | RB-10 |
| NEG11 | NEG11 |
| HL60-12 | HL60-12 |
| | 014A-A |
| | 014A-B |
| | 014A-C |
| | 014B-A |
| | 014B-B |
| | 014B-C |
| | 014C-A |
| | 014C-B |
| | 014C-C |
| | RB |
| | Negative |
| | HL60 – Positive |

*Table 20: A list of libraries deep-sequenced using the Roche GS Junior™ and Illumina® MiSeq™ NGS platforms. Libraries were generated using two different primer sets. For a direct comparison of NGS platforms, the same libraries generated using Roche Fusion primers were deep-sequenced on the Roche GS Junior™ and Illumina® MiSeq™. Libraries generated with unmodified primers were also run on the Illumina® MiSeq™ so that the degree of Nextera™ tagmentation in relation to amplicon length could be assessed.*

**DNA Extraction and Amplification**

Three hairs were obtained from donor 014-CF31. Hairs were viewed microscopically for the presence of a follicular tag. Cellular material was evident on all hairs. The roots were removed and discarded, and the adjacent 2 cm hair shaft was used for mtDNA analysis from each hair. DNA was extracted from the hairs using the optimized protocol developed in our laboratory, with a final elution volume of 70 µL. In addition to hair samples, an HL60 positive control, a reagent blank and a negative control were included. HL60 is a human immortalized cell line containing both nuclear and mitochondrial DNA. The resulting DNA extracts were then quantified in duplicate using a human mtDNA specific 5' exonuclease real-time PCR assay developed by Kavlick *et. al.* All line statistics reported for the standard curve fell within the acceptable range as reported by Kavlick *et. al.* The average copy number per µL was calculated for duplicates. Real-time PCR results can be found in table 21.

| | |
|---|---|
| 014A (Hair 1) | 8,500 |
| 014B (Hair 2) | 15,400 |
| 014C (Hair 3) | 4,880 |
| RB | undetected |
| HL60 Positive Control | 2,460 |

*Slope = -3.37, y-intercept = 41.31, $R^2$ = 0.998

*Table 21: Donor 014-CF31 - 2 cm hair shaft mtDNA quantification results. A human mtDNA specific real-time PCR assay was used to quantify DNA extracted from hairs obtained from donor 014-CF30. All hairs were quantified in duplicate. Values reported in this table are averages of the duplicate runs.*

Extracted DNA from each hair was amplified in triplicate using both Roche Fusion primers and unmodified primers specific for the HV1b region of mtDNA. The Roche FastStart High-fidelity PCR System™ (Roche Diagnostics, Indianapolis, IN) consists of a blend of FastStart™ Taq DNA polymerase and a thermostable chemically modified proofreading protein. This system was chosen for amplification of all samples to reduce the incidence of polymerase induced base misincorporations. Samples were amplified as follows: 10 µL of hair shaft DNA extract in a reaction mixture containing 2.5 U of Roche FastStart High

Fidelity enzyme blend, 1X FastStart High Fidelity reaction buffer, 1.8 mm $MgCl_2$, 4% DMSO, 200 μM each dATP, dTTP, dCTP, dGTP from PCR grade nucleotide mix (Promega Corporation, Madison, WI), 400 nM forward primer and 400 nM reverse primer. Primer sequences are shown in Table 22. Reaction mixtures were amplified on a GeneAmp® PCR System 9700 (Applied Biosystems, Foster City, CA) with an initial 2 minute hold at 95ºC, followed by 36 cycles comprised of a 30 second denaturation at 95ºC, a 30 second annealing step at 60ºC, and a 30 second extension at 72ºC with a final 7 minute extension at 72ºC and a long term 4ºC hold. Resulting amplification products were analyzed using the Agilent 2100 Bioanalyzer™ and DNA 1000 kit (Agilent Technologies, Clara, CA). Table 24 shows the concentrations of amplicons obtained. All amplified samples were diluted two-fold prior to library preparation so that the sample volumes were sufficient for all 4 planned NGS runs.

| | |
|---|---|
| A2 (16159L): 5' – CGT ATC GCC TCC CTC GCG CCA *TCA GXX XXX XXX XX***T ACT TGA CCA CCT GTA GTA C** – 3' | A2 (16159L): 5' – TAC TTG ACC ACC TGT AGT AC – 3' |
| B1 (16391H): 5' – CTA TGC GCC TTG CCA GCC CGC *TCA GXX XXX XXX XX***G AGG ATG GTG GTC AAG GGA C** – 3' | B1 (16391H): 5' – GAG GAT GGT GGT CAA GGG AC – 3' |

*Table 22: mtDNA HV1b Primer sequences used for DNA library amplification. The numbers denote the location of the 3' base of the primer with respect to the rCRS. *Underlined regions in Roche Fusion primer sequences represent adaptors required for pyrosequencing, while italicized regions are 4-base tags that serve as Roche GS Junior™ sequencing controls. The regions marked with X placeholders are reserved for multiplexing indices that vary depending on the source of the library. The bold regions are template specific, and are identical to the unmodified primer sequences.*

| | | | | |
|---|---|---|---|---|
| | | | | |
| 014A-1 (Hair 1, amp 1) | 1 | 28.88 | 014A-A (Hair 1, amp 1) | 22.67 |
| 014A-2 (Hair 1, amp 2) | 2 | 28.07 | 014A-B (Hair 1, amp 2) | 24.49 |
| 014A-3 (Hair 1, amp 3) | 3 | 24.00 | 014A-C (Hair 1, amp 3) | 20.75 |
| 014B-4 (Hair 2, amp 1) | 4 | 30.31 | 014B-A (Hair 2, amp 1) | 25.50 |
| 014B-5 (Hair 2, amp 2) | 5 | 28.21 | 014B-B (Hair 2, amp 2) | 24.52 |
| 014B-6 (Hair 2, amp 3) | 6 | 28.17 | 014B-C (Hair 2, amp 3) | 24.13 |
| 014C-7 (Hair 3, amp 1) | 7 | 3.32 | 014C-A (Hair 3, amp 1) | 13.40 |
| 014C-8 (Hair 3, amp 2) | 8 | 3.84 | 014C-B (Hair 3, amp 2) | 12.15 |
| 014C-9 (Hair 3, amp 3) | 9 | 4.63 | 014C-C (Hair 3, amp 3) | 11.15 |
| RB-10 | 10 | 0 | RB | 0 |
| Negative-11 | 11 | 0 | Negative Control | 0 |
| HL60-12 | 12 | [17.12 | HL60 Positive Control | 24.37 |

*Table 23: Agilent 2100 Bioanalyzer™ reported amplicon concentrations. A unique multiplex identifier (MID) included on the 5' end of Roche fusion primers was assigned to samples to enable sequencing on the Roche GS Junior™, and subsequent sample-dependent data parsing.*

**Roche GS Junior™ Library Preparation**

Nine total sample libraries (three hair extracts each amplified in triplicate) and three controls were deep-sequenced on the Roche GS Junior™. According to Roche, sequencing 14 or fewer libraries per run will result in an overall depth of coverage of 5,000X or greater. This total depth results in an estimated 1% minor-variant fold coverage of 50X, enabling accurate and reliable detection over background noise.

The Roche Fusion DNA libraries were purified using Agencourt® AMPure® beads according to the Roche GS Junior™ amplicon library preparation method manual. Purified libraries were re-quantified prior to each respective run using the Agilent 2100 Bioanalyzer™ (results seen in Table 25).

| | | |
|---|---|---|
| 014A-1 (Hair 1, amp 1) | 13.63 | 11.98 |
| 014A-2 (Hair 1, amp 2) | 16.11 | 10.46 |
| 014A-3 (Hair 1, amp 3) | 13.43 | 13.96 |
| 014B-4 (Hair 2, amp 1) | 13.46 | 11.53 |
| 014B-5 (Hair 2, amp 2) | 15.90 | 10.29 |
| 014B-6 (Hair 2, amp 3) | 14.71 | 9.99 |
| 014C-7 (Hair 3, amp 1) | 1.61 | 1.75 |
| 014C-8 (Hair 3, amp 2) | 2.06 | 2.04 |
| 014C-9 (Hair 3, amp 3) | 2.63 | 2.65 |
| RB-10 | 0 | 0 |
| Negative-11 | 0 | 0 |
| HL60-12 | 9.54 | 8.28 |

*Table 24: Purified library concentrations in ng/µL obtained using the Agilent 2100 Bioanalyzer™. All libraries were quantified prior to dilution and pooling for sequencing using the Roche GS Junior™.*

| | | | | |
|---|---|---|---|---|
| 014A-1 | 13.63 | 36,221,995,612 | 2.8 | 97.2 |
| 014A-2 | 16.11 | 42,812,644,850 | 2.3 | 97.7 |
| 014A-3 | 13.43 | 35,690,491,641 | 2.8 | 97.2 |
| 014B-4 | 13.46 | 35,770,217,237 | 2.8 | 97.2 |
| 014B-5 | 15.9 | 42,254,565,681 | 2.4 | 97.6 |
| 014B-6 | 14.7 | 39,092,117,054 | 2.6 | 97.4 |
| 014C-7 | 16.61 | 4,278,606,965 | 23.4 | 76.6 |
| 014C-8 | 2.06 | 5,474,490,900 | 18.3 | 81.7 |
| 014C-9 | 2.63 | 6,989,277,216 | 14.3 | 85.7 |
| RB10 | 0 | 0 | 100 | 0 |
| NEG11 | 0 | 0 | 100 | 0 |
| HL60-12 | 9.54 | 25,352,739,409 | 3.9 | 96.1 |

| | | | | |
|---|---|---|---|---|
| 014A-1 | 11.98 | 32,316,012,972 | 3.1 | 96.9 |
| 014A-2 | 10.46 | 28,215,817,670 | 3.5 | 96.5 |
| 014A-3 | 13.96 | 37,657,056,853 | 2.7 | 97.3 |
| 014B-4 | 11.53 | 31,102,139,363 | 3.2 | 96.8 |
| 014B-5 | 10.29 | 27,757,243,196 | 3.6 | 96.4 |
| 014B-6 | 9.99 | 26,947,994,123 | 3.7 | 96.3 |
| 014C-7 | 1.75 | 4,720,619,591 | 21.2 | 78.8 |
| 014C-8 | 2.04 | 5,502,893,695 | 18.2 | 81.8 |
| 014C-9 | 2.65 | 67,148,366,809 | 14.0 | 86.0 |
| RB10 | 0 | 0 | 100 | 0 |
| NEG11 | 0 | 0 | 100 | 0 |
| HL60-12 | 8.28 | 22,335,274,408 | 4.5 | 95.5 |

$$molecules/uL = \frac{[DNA(g/uL)]x(6.022x10^{23}molecules/mole)}{(6.022x10^{23}molecules/mole)x(1.096x10^{-21}g/bp)x345bp/molecule}$$

*Table 25 A and B: Dilution Strategy for Roche GS Junior™ Runs 1 (A) and 2 (B). The instrument protocol requires that libraries be normalized to a concentration of 1.0 x $10^9$ molecules/µL prior to pooling. The equation used to calculate number of molecules/µL from DNA concentration in ng/µL is included below table 25B. The tables below outline the dilutions performed for normalization of each library to 1.0 x $10^9$ molecules/µL.*

Equal volumes of each library were pooled following normalization. The pooled amplicon library was diluted 100 fold to a final concentration of 1.0 x $10^7$ molecules/µL, and was then subjected to clonal amplification using emulsion PCR (emPCR) and the Lib-A kit according to the Roche emPCR Amplification Method Manual. One-half of a molecule per DNA capture bead was targeted for both Roche NGS runs to reach the desired bead enrichment percentage. This target is a deviation from the 2 molecules per DNA capture bead recommended in the manual. In our experience, targeting 2 molecules per bead results in an excess of DNA capture beads.

**Illumina® MiSeq™ Library Preparation**

Illumina® strongly recommends that Nextera™ tagmentation be performed on amplicons >300 bp in length. However, our control-region amplification primers give rise to amplicons below this recommended size (270 bp). To test the efficiency of Nextera™ tagmentation on amplicons above and below the recommended size, all amplified libraries (those generated using both Roche fusion primers and unmodified primers) were sequenced on the Illumina® MiSeq™. The libraries were requantified using the Agilent 2100 Bioanalyzer™ (Table 26) prior to preparation for sequencing on the Illumina® MiSeq™. All libraries were diluted to a final concentration of 0.2 ng/µL, and were prepared for sequencing using the Illumina® Nextera® XT Tagmentation kit, and a dual library indexing strategy was used according to the Nextera® XT DNA Sample Preparation Guide. The resulting tagmented, and indexed libraries were normalized and pooled. These tagmented libraries were used for both MiSeq™ NGS runs. Prior to each run, the libraries were pooled and 8.0 pM, PhiX was added at a 20% volume/volume ratio as a highly characterized control.

| | | | | | | |
|---|---|---|---|---|---|---|
| 014A-1 | Fusion | N701 TAAGGCGA | S501 TAGATCGC | 39.55 | 1.3 | 248.7 |
| 014A-2 | Fusion | N701 TAAGGCGA | S502 CTCTCTAT | 34.67 | 1.4 | 248.6 |
| 014A-3 | Fusion | N701 TAAGGCGA | S503 TATCCTCT | 26.69 | 1.9 | 248.1 |
| 014B-4 | Fusion | N702 CGTACTAG | S501 TAGATCGC | 31.70 | 1.6 | 248.4 |
| 014B-5 | Fusion | N702 CGTACTAG | S502 CTCTCTAT | 18.33 | 2.7 | 247.3 |
| 014B-6 | Fusion | N702 CGTACTAG | S503 TATCCTCT | 24.07 | 2.1 | 247.9 |
| 014C-7 | Fusion | N703 AGGCAGAA | S501 TAGATCGC | 3.23 | 15.5 | 234.5 |
| 014C-8 | Fusion | N703 AGGCAGAA | S502 CTCTCTAT | 3.44 | 14.5 | 235.5 |
| 014C-9 | Fusion | N703 AGGCAGAA | S503 TATCCTCT | 2.06 | 24.3 | 225.7 |
| RB10 | Fusion | N704 TCCTGAGC | S501 TAGATCGC | 0.95 | NA | NA |
| NEG11 | Fusion | N704 TCCTGAGC | S502 CTCTCTAT | 0 | NA | NA |
| HL60-12 | Fusion | N704 TCCTGAGC | S503 TATCCTCT | 9.22 | 5.4 | 244.6 |
| 014A-A | Unmodified | N705 GGACTCCT | S501 TAGATCGC | 22.51 | 2.2 | 247.8 |
| 014A-B | Unmodified | N705 GGACTCCT | S502 CTCTCTAT | 36.35 | 1.3 | 248.7 |
| 014A-C | Unmodified | N705 GGACTCCT | S503 TATCCTCT | 26.98 | 1.9 | 248.1 |
| 014B-A | Unmodified | N706 TAGGCATG | S501 TAGATCGC | 33.57 | 1.5 | 248.5 |
| 014B-B | Unmodified | N706 TAGGCATG | S502 CTCTCTAT | 16.45 | 3.0 | 247 |
| 014B-C | Unmodified | N706 TAGGCATG | S503 TATCCTCT | 33.12 | 1.4 | 248.6 |
| 014C-A | Unmodified | N701 TAAGGCGA | S504 AGAGTAGA | 23.20 | 2.2 | 247.8 |
| 014C-B | Unmodified | N702 CGTACTAG | S504 AGAGTAGA | 25.20 | 2.0 | 248 |
| 014C-C | Unmodified | N703 AGGCAGAA | S504 AGAGTAGA | 5.44 | 9.2 | 248.8 |
| RB | Unmodified | N704 TCCTGAGC | S504 AGAGTAGA | 0 | NA | NA |
| NEG | Unmodified | N705 GGACTCCT | S504 AGAGTAGA | 0 | NA | NA |
| HL60 | Unmodified | N706 TAGGCATG | S504 AGAGTAGA | 27.41 | 1.8 | 248.2 |

*Table 26: Illumina® MiSeq™ library preparation. Library index assignments and normalization strategies for Illumina® MiSeq™ runs 1 and 2 are shown in the table. Indices are incorporated into libraries during Nextera™ tagmentation limited-cycle PCR.*

**Data Analysis - Roche Amplicon Variant Analyzer Software**

All Roche GS Junior™-derived data was analyzed using the Linux-based GS Run Processor, and Roche Amplicon Variant Analyzer (AVA) software via a graphic user interface (GUI). Initially, the GS Run Processor performs data processing in which captured images from raw wells are converted to .pif files, and background noise is normalized. This data is then subjected to a signal-processing algorithm where filtering, correction, and trimming is performed prior to base-calling and quality scoring. This pipeline ultimately results in .sff (standard flowgram format) files. The .sff files are then uploaded into the AVA software for analysis. Primer sequences are trimmed from each read, and reads are demultiplexed into library specific bins based on user defined multiplex identifier (MID) sequences. The parsed reads are then aligned to a reference. Aligned reads were viewed either individually or as consensus sequences of error-corrected, collapsed high depth reads. This application enables more rapid data analysis with a marked reduction in noise. Although this option is useful, it can lead to omission of true variants from the data set.

Variants from the reference are reported in tabular format with calculated frequencies represented as a percentage of the total number of reads. The variant frequencies can be viewed as an average percentage calculated from forward and reverse reads, or the data table can be expanded to show frequencies from forward and reverse reads independently. The software offers a further filtering option entitled "bidirectional support." If a variant is found in equal proportions in both forward and reverse reads, it is likely a true variant. However, bidirectionality is difficult to achieve if the variant is close to the ends of the amplicon where base quality declines, and likelihood of bidirectional coverage is reduced. Adjusting the analysis parameters of AVA software is limited, especially for individuals without experience using a Linux command line interface.

**Illumina® MiSeq™ Reporter Software**

MiSeq® Reporter is an on-instrument secondary analysis software package that accompanies the Illumina® MiSeq™. Initially, basecalls and PHRED quality scores are generated on the instrument during primary data analysis. Demultiplexing, FASTQ file generation, alignment to a reference, and variant calling are conducted by the MiSeq® Reporter software during secondary analysis. Prior to starting a run, the user populates a "sample sheet" template in which a reference genome, where secondary analysis applications are specified. For mtDNA amplicons or whole genome data, the resequencing analysis workflow is selected, and index sequences are specified. This workflow is designed for small genomes, and employs the Burrows-Wheller Alignment (BWA) method of alignment, which allows for 3' trimming of low quality data, and removal of adapter sequences. Variant calling parameters are also set within the sample sheet, and include filtering of single-stranded variants, establishing a minimum variant coverage depth, a minimum variant Q-score, and a minimum variant frequency (amongst other parameters). Default settings were used for all parameters except the variant frequency filter cutoff, which was changed from 20% to 1% in order to capture the presence of the low-level variants.

BWA typically uses a soft clipping method during alignment. This method generates short reads from sequences that have been almost entirely trimmed. MiSeq™ Reporter has a short read masking setting that removes short reads that could confound a downstream alignment to a reference. Customization of analysis parameters is limited within MiSeq™ Reporter to those parameters listed as modifiable within the sample sheet.

### SoftGenetics NextGENe® Software

NextGENe® is an NGS data analysis product from SoftGenetics, Inc. The software is an all-inclusive, secondary NGS analysis software package with a user-friendly, Windows®-based graphic user interface (GUI) designed for biologists with little to no Linux command-line experience. The package runs on a Windows® operating system, and can be used for analysis of data generated by Roche, Illumina® and Applied Biosystems® NGS instruments. As a result, this software package enables direct comparison of data generated with multiple NGS platforms with little software-induced bias. All Illumina® MiSeq™ data, and HL60 data from the Roche GS J™ was analyzed using this tool.

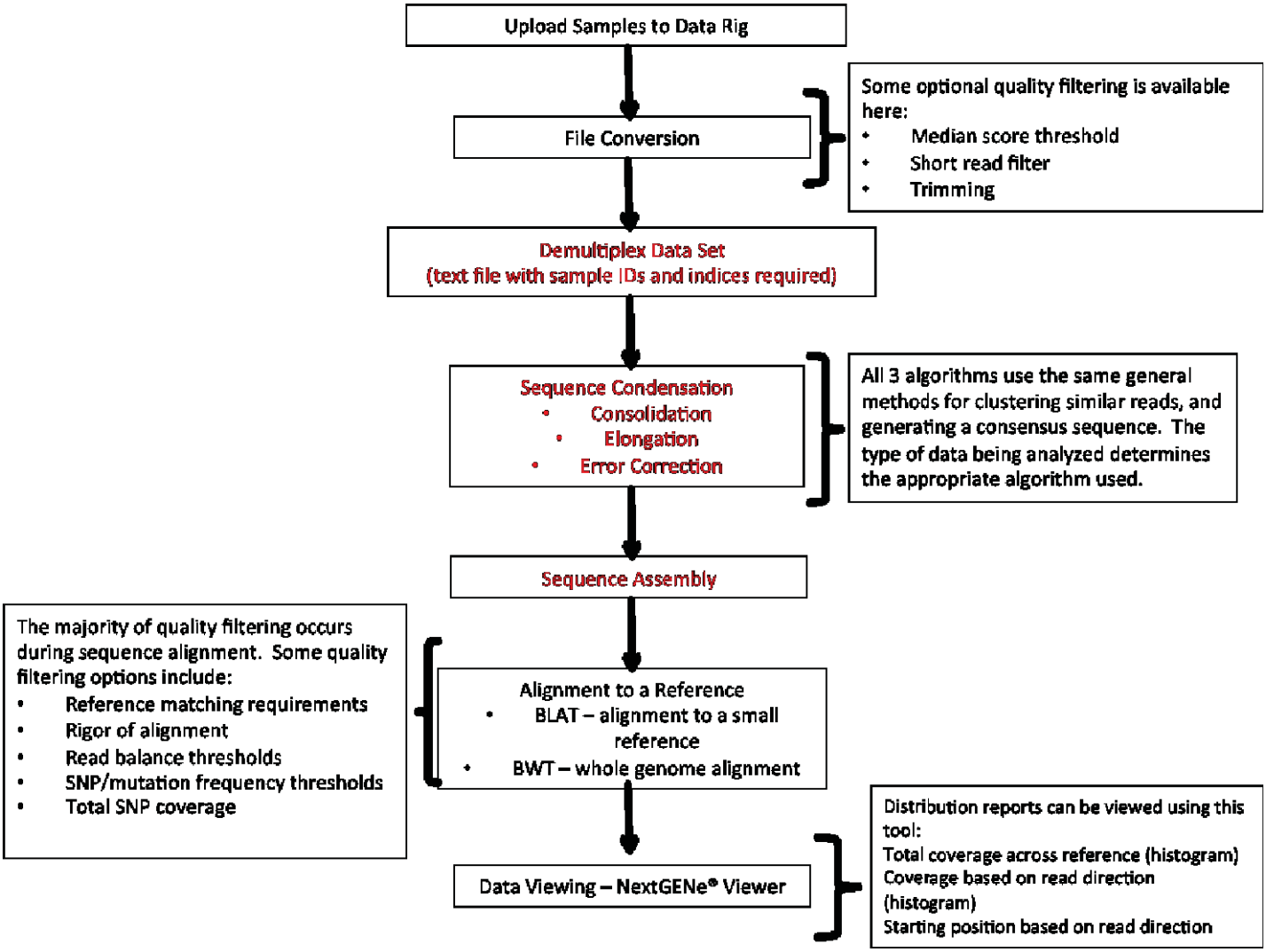



*Figure 15: SoftGenetics NextGENe® data analysis pipeline. SoftGenetics NextGENe® software is user-friendly, fully customizable, and quality filtering is transparent. The pipeline described in this figure is generalized. All parameters in red are optional and can be easily modified by the user.*

### NextGENe® File Conversion

Data files were initially converted to FASTA files using the NextGENe® file conversion application. This application enables conversion of different file types generated with Roche, Illumina® and Applied Biosystems® NGS instrumentation. During file conversion, the user has the option to customize several quality filtering parameters. These parameters enable filtering of reads with low NextGENe® quality scores,

shorts reads below a specified size, and also allows for dynamic or static read trimming. No file conversion quality filtering was performed for this study.

### NextGENe® Demultiplexing

Roche GS Junior™-converted data was parsed into sample dependent files using the built-in NextGENe® demultiplexer. Recall that demultiplexing occurs during early secondary data analysis on the Illumina® MiSeq™ instrument. As a result, this tool was not used to parse the Illumina® data.

### NextGENe® Sequence Condensation and Assembly

Once the FASTA files are generated, the user has the option to perform sequence condensation and assembly prior to alignment. Sequence condensation is a tool within the NextGENe® software package that uses depth of coverage to correct for base-calling errors. The recommended methods of condensation differ depending on the instrument used to generate the data.

The tool is designed to correct sequence reads with instrument dependent base-calling errors using three different algorithms, including error correction, elongation, and consolidation. The algorithm employed is dependent upon the instrument used to generate the data, as well as the depth of coverage of the data set. Consolidation is used when the data set has a high depth of coverage, and is not applicable for 454 data. This algorithm combines overlapping reads, and a consensus sequence is used to represent those reads that were combined. After applying the algorithm, low-frequency variants are removed from the data set. The error correction is designed explicitly for 454 data, and has a built-in low-frequency homopolymer (consisting of 3 bases or more) error correction algorithm. In this case, the data are not "condensed" as the overall read count is preserved. We did not use the elongation algorithm, as this is designed for low coverage, mate-paired data.

A "consolidation method", in which similar reads are merged into consensus sequences, is suggested for data with a high depth of coverage. An advanced settings option within the sequence condensation tool enables the user to set additional thresholds to apply to the data set. Some of these settings include minimum read length for condensation, forward and reverse read balance ratio, and removal of low quality ends when quality scores dip below a specified threshold. The "error correction" method is the only condensation method available for pyrosequencing data. This algorithm allows for the correction of homopolymer read errors, without consolidating like reads into consensus sequences. Very few advanced settings can be modified when using this method. Since amplicons were sequenced, sequence assembly, which creates large contigs from short overlapping reads, was not used.

### NextGENe® Sequence Alignment

NextGEN*e*® alignment can be accomplished using a BLAT-like algorithm for genomic regions or small genomes <250 Mbp, or BWT (Burrows-Wheeler Transform) for mapping reads to whole large genomes. The BLAT-like option was selected for this particular study, since the mtDNA reference genome is 16,569 bp in length. Alignment variant-calling thresholds are easily modified using NextGEN*e*® software. The matching requirement option was set in which a minimum of 12 bases and 85% of the read sequence must match the reference in order for the read to be aligned. The mutation percentage was set to 0.1%. All variants above this set threshold are called and displayed in a summary data table. The remainder of the

alignment parameters were kept at default values. These parameters include a threshold for total read coverage of a SNP, and balance ratios for removing errors that are unidirectional.

The computational cap of the NextGENe® software is 65,535 reads. Since analyzed reads are not selected at random, bidirectional balance is skewed if the depth of coverage exceeds this value. In such cases the reads are selected in the order of generation.

### HL60 NextGENe® Data Analysis – Illumina® MiSeq™ Data

HL60 MiSeq™ data was analyzed using 2 different NextGENe® methods. The positive control data was selected for this comprehensive analysis, since low-level NumtS variants were expected. The data analysis for the 2 methods differed in the use sequence condensation for error correction.

### HL60 Illumina® MiSeq™ Data Analysis, NextGENe® - Method 1

FASTQ files were converted to FASTA files using no quality filtering during file conversion, and no sequence condensation was performed.

### HL60 Illumina® MiSeq™ Data Analysis, NextGENe® - Method 2

Sequence condensation was performed for error correction. The consolidation algorithm was chosen due to the high depth of coverage of the data set. When this algorithm is employed, consensus sequences are generated that represent groups of similar individual reads. This reduces the size of the data set, and enables rapid variant identification.

### HL60 NextGENe® Data Analysis - Roche GS Junior™ Data

HL60 GS Junior™ data was analyzed using 2 different NextGENe® methods. Sequence condensation was applied in one instance, and not in the other. The data was subjected to sequence condensation using the error correction algorithm. This is the only method available for GS Junior™ data, because it has a built-in method to correct for errors associated with homopolymeric stretches.

### Roche GS Junior™ – AVA Analysis

The average depth of coverage across all libraries reported by the AVA software differed considerably from run 1 (5,700X) to run 2 (2,200X). However, in all cases the expected variants from all of the hair samples and the positive control were accurately detected using the AVA software with default parameters. Additional, unexpected variants were also detected at various locations throughout HV1b in all samples. No discernable pattern was detected, because these unexpected variants were not reproducible in different hairs, amplifications, or runs. In addition, a large proportion of the unexpected variants possessed high frequencies in one read direction, with the same variant completely absent in the opposite read (showing a frequency of 0). Nearly all unexpected variants are associated with homopolymers of 2 or more consecutive, identical nucleotides. This is not unexpected, since noise associated with homopolymeric regions is a well-known artifact of pyrosequencing. NumtS variants were expected in the HL60 data, since this sample contains a both nuclear and mtDNA. However, no NumtS variants were detected.

### Illumina® MiSeq™ - MiSeq™ Reporter Analysis

MiSeq Reporter reports numbers of clusters passing filter for each parsed sample library. A depth of coverage is also given for variants reported. The numbers of clusters passing filter differed depending on the

library analyzed. The average depth of coverage for MiSeq™ runs 1 and 2 ranged from 90,000 – 200,000. All expected variants were detected for hairs and the HL60 positive control in all cases. Unexpected variants were also detected, however, all of these were in positions outside of HV1b (16,159-16,391), and were attributed to noise. Reagent blanks and negative controls showed high numbers of clusters passing filter, but only a small number of the reads passed met filtering thresholds and aligned to the reference sequence. The unmodified primer set negative control showed a 16239T variant with 1,427X coverage in run 1 and 811X coverage in run 2. This variant matches that of the hair donor, so cross contamination is the likely explanation. The unmodified primer set reagent blank showed a 16267T variant with 21,486X average coverage across both runs. This is attributed to spurious contamination since this variant does not match the donor of the hair samples, the analyst preparing the samples for NGS, or any other individuals with laboratory access. All other RB and negative controls possess variants with <840X coverage. Since NGS is a very sensitive technique, low-level contamination is unavoidable when performing mtDNA analysis.

There were no unexpected variants associated with NumtS in any hair shaft libraries detected by MiSeq™ Reporter, which is expected since hair shafts have little to no nuclear DNA present. However, NumtS were not detected in HL60 data. This is attributed to the high frequency threshold set for MiSeq™ Reporter data analysis.

### Illumina® MiSeq™ Data – NextGENe® Analysis

The average depth of coverage obtained when using NextGEN*e*® is 65,535X (the computational cap of the software) for read 1 and <50,000X for read 2. . These numbers are lower, on average than those presented by the MiSeq™ Reporter software. This is attributed to the differences in quality filtering and alignment algorithms between the two software packages. All expected variants were detected for hairs and the HL60 positive control when using the NextGEN*e*® software. Read balance, in general, appears poor for read 1 because the coverage achieved for this run exceeded the capacity of the software. However, read balance is more evenly distributed in the center of the target sequence in run 2 data, however, it becomes skewed at the distal ends. This is explained by the fact that coverage declines towards the 3' end of a read.

The number of variants within the reportable region of the target sequence is significantly higher when using NextGEN*e*® than when using MiSeq™ Reporter. The frequency threshold was set at 0.1% within the NextGEN*e*® software versus 1% within the MiSeq™ Reporter. This not only explains the greater number of unexpected variant calls, but also the fact that NumtS variants are detected in HL60 data when using the NextGEN*e*® software.

### HL60 Positive Control, Roche GS Junior™ – NextGENe® Analysis

HL60 data generated using the Roche GS Junior™ was analyzed with SoftGenetics NextGEN*e*® software with and without using an error correction method designed to reduce the incidence of errors associated with homopolymers. Reported coverage was similar to that reported by the Roche AVA software. However, in all cases a greater number of total reported variants was detected with the NextGEN*e*® software, because the variant detection threshold was significantly lower than in the AVA software. Some of the variants detected were at positions consistent with the NumtS, but no correlation could be made because association of the variants to defined reads was not possible. It did appear however, that the error correction sequence condensation algorithm was effective at correcting homopolymer read errors. This is evidenced at expected variant position 16193T where a frequency of 90.57% was observed in run 1 data with no sequence condensation, and at 99.61% when sequence consolidation was applied. This data is shown in Table 27.

| | | | | | |
|---|---|---|---|---|---|
| | | Run 1 | Run 2 | Run 1 | Run 2 |
| Maximum Coverage | | ~5,100 | ~2,100 | ~5,100 | ~2,100 |
| Reported read length | | 278-289 | 280-290 | 279-286 | 277-282 |
| Total number of | | 128 | 90 | 78 | 64 |

| | | | | | |
|---|---|---|---|---|---|
| Variants in mutation report | | | | | |
| # of Variants outside of amplicon range | | 46 | 32 | 30 | 34 |
| # of Reportable Variants | | 128-46=82 | 90-32=58 | 78-30=48 | 64-34=30 |
| Frequency of Expected Variants Called | 16193T<br>16278T<br>16362C | 90.57%<br>99.66%<br>92.23% | 97.28%<br>99.80%<br>98.46% | 99.61%<br>99.92%<br>98.35% | 99.7%<br>99.95%<br>99.95% |
| NumtS positions called | | 8/19 | 6/19 | 5/19 | 5/19 |
| Manual resolution of SNP donors | | No | No | No | No |
| NumtS Frequencies | Average<br>St. Dev. | 0.14<br>0.12 | 0.39<br>0.40 | 0.095<br>0.07 | 0.12<br>0.13 |

*TABLE 27: Summary data table for HL60. NextGENe® software was used for analysis of HL60 data with and without applying a homopolymeric read error correction algorithm. The table shows several parameters including total number of variants reported, number of variants lying outside of the HV1b region (16159-16391), and number of reportable variants. The table also outlines the number of variants that are associated with positions consistent with the NumtS insertion in this region, and whether or not the variants are manually resolvable.*

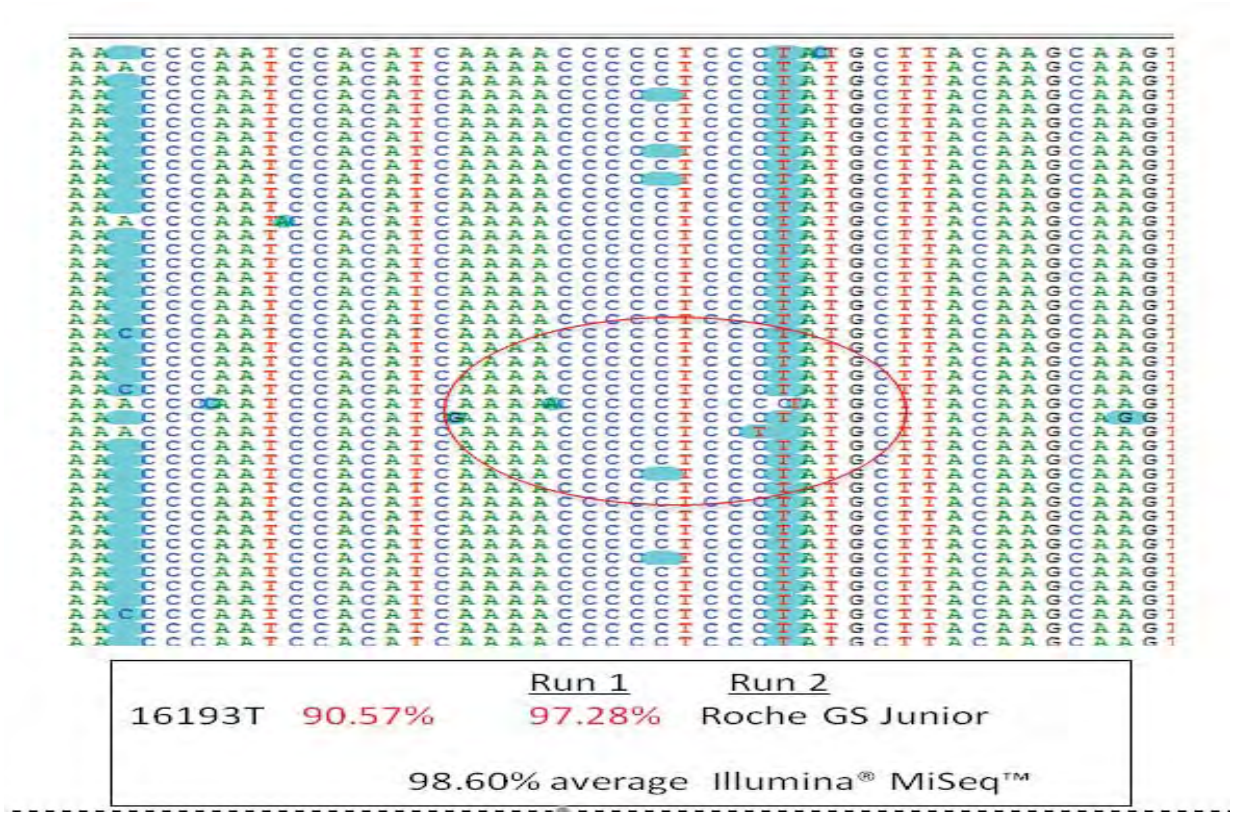



*Figure 16: Roche GS Junior™ homopolymer read error seen in HL60 data with NextGENe® viewer. The figure below shows an instance of homopolymer read error in the mtDNA HV1 C-stretch.*

**HL60 Positive Control, Illumina® MiSeq™ – NextGENe® Analysis**

HL60 data generated using the Illumina® MiSeq™ was analyzed with SoftGenetics NextGENe® software. Coverage of expected variants was similar to coverage obtained when using the Illumina® MiSeq™ Reporter software, except in the case where sequence condensation was used. However, a greater number of total reported variants was detected with the NextGENe® software, because the variant detection threshold was significantly lower than in the MiSeq® Reporter software. This lower variant frequency threshold (0.1%) has allowed for the detection of NumtS associated variants. These variants are easily

associated to one donor when using the NextGENe® viewer, because they appear clustered within discrete reads

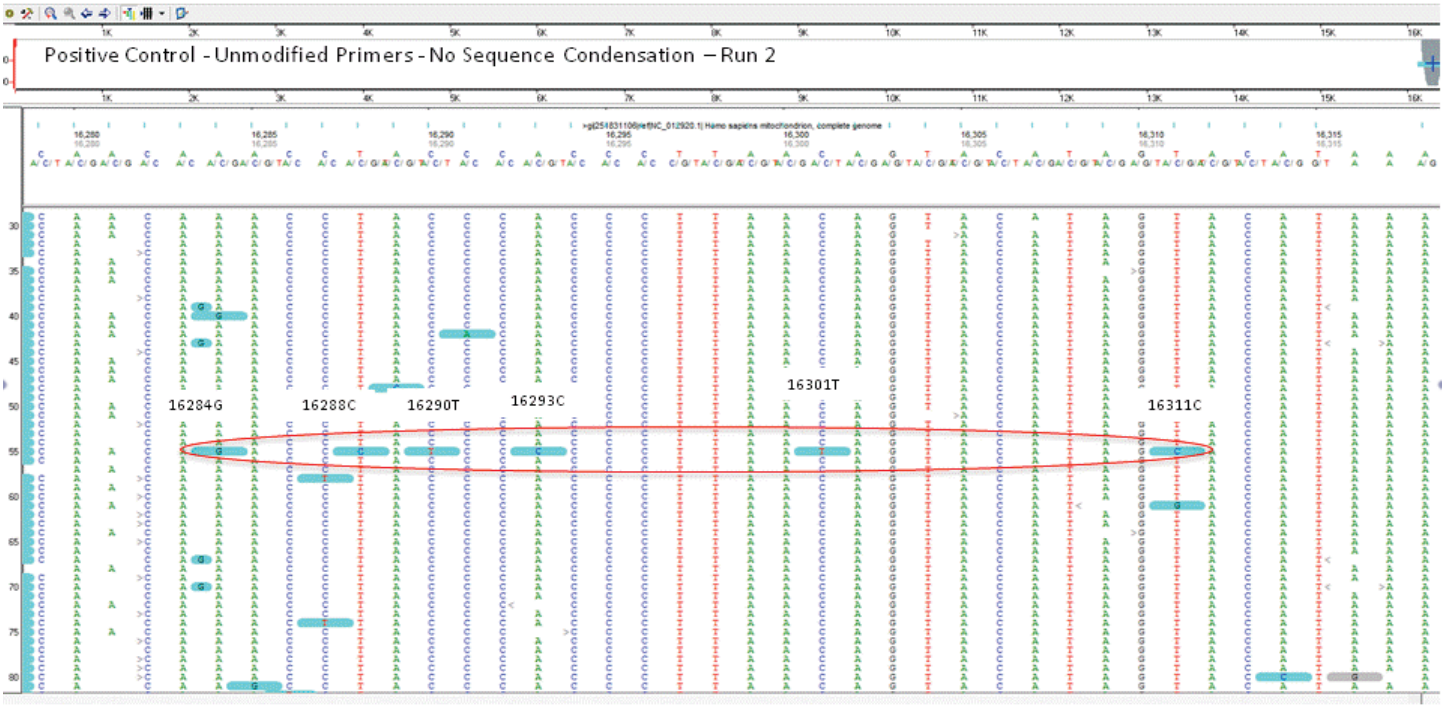
Run 2 data shows a range of read lengths. The same Nextera™ prepared library was used for Illumina® MiSeq™ runs 1 and 2, which means that degradation was likely to have occurred by run 2 as a result of extended storage, and multiple freeze-thaw cycles. However, this does not appear to be the explanation. Degradation would ultimately give rise to fragments with adapters on a single end. These fragments would not be capable of bridge amplification, and thus, would not be sequenced. Additional studies are required to elucidate the cause of this issue.

Sequence condensation, used in analysis method 2, results in skewed frequencies of both expected and unexpected variants. Since one benefit of NGS in forensic DNA analysis is the ability to quantify variants, this error-correction method may not be desirable. However, since the overall depth is markedly reduced following condensation (~66,000 reads → 630 reads), manual resolution of SNP donors is much easier. Therefore, it is recommended that this tool be used only for simplified data viewing and not for variant quantification purposes. FIGURE 17B shows the clustering of NumtS variants within consensus reads generated with sequence condensation.

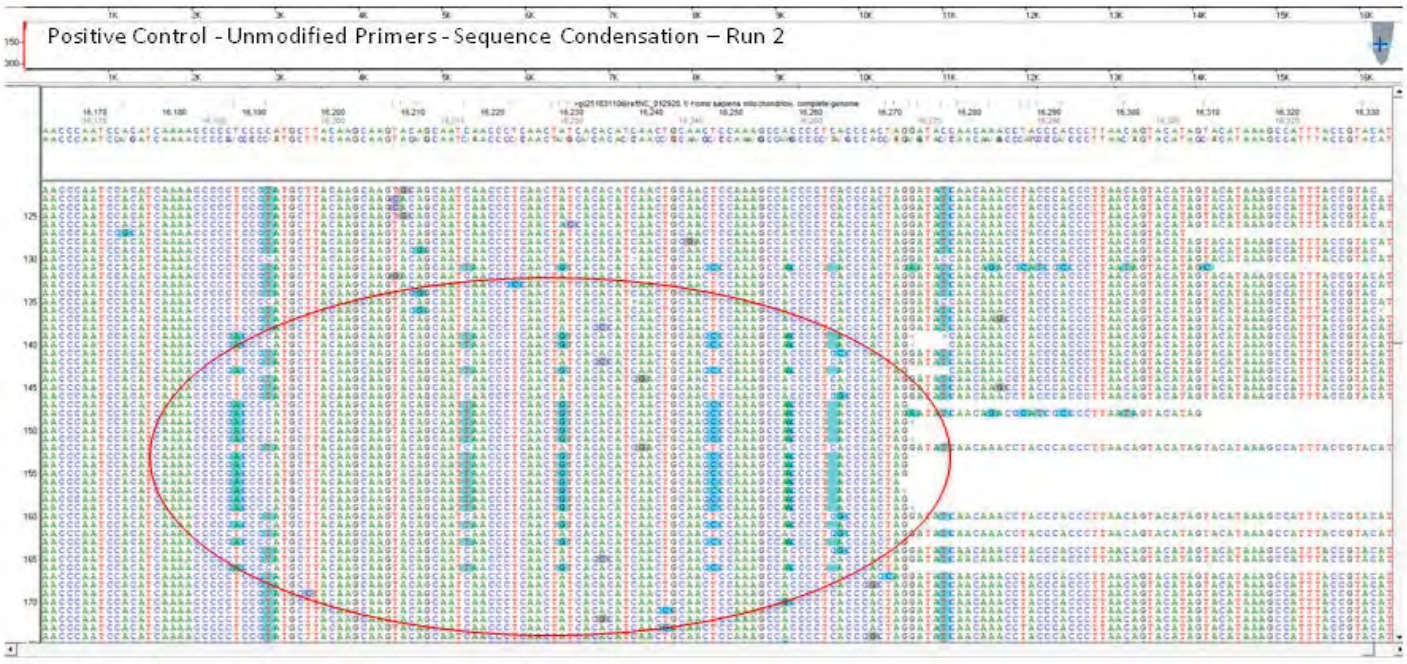
| **Parameter** | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | |
| Maximum Coverage | | ~58,000 | ~65,535 | ~46,000 | ~65,535 | ~320 | ~630 | ~530 | ~330 |
| Reported max read length | | 151 | 35-151 | 151 | 35-151 | 80-260 | 85-280 | 140-290 | 60-255 |
| Total # of variants in mutation report | | 223 | 261 | 196 | 230 | 47 | 109 | 103 | 46 |
| # of Variants outside amplicon range | | 81 | 81 | 36 | 41 | 1 | 42 | 48 | 13 |
| # of Reportable variants | | 142 | 180 | 160 | 189 | 46 | 67 | 55 | 34 |
| Frequency of Expected variants | 16193T<br>16278T<br>16362C | 98.38%<br>97.99%<br>98.92% | 99.10%<br>98.56%<br>99.70% | 97.39%<br>99.55%<br>98.08% | 98.19%<br>99.04%<br>99.35% | 85.06%<br>98.01%<br>93.16% | 92.39%<br>98.91%<br>94.19% | 97.93%<br>98.86%<br>98.60% | 85.17%<br>97.24%<br>94.78% |
| NumtS Positions | | 16/19 | 18/19 | 19/19 | 19/19 | 17/19 | 12/19 | 7/19 | 18/19 |
| Manual Resolution of SNP donors? | | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| NumtS frequency | Average<br>StDev | 0.27<br>0.35 | 0.45<br>0.63 | 0.31<br>0.24 | 0.71<br>0.48 | 6.53<br>3.83 | 5.72<br>1.80 | 1.23<br>0.35 | 14.15<br>6.29 |

*65,535 is the computation cap of NextGENe® software.

*Table 28 – Comparison of Method 1 and Method 2, with and without sequence consolidation. While the number of reads is greatly diminished after consolidation, the percentage of NumtS-derived variants is substantially increased.*




NextGENe™ Sequence Alignment Viewer – Non-condensed data

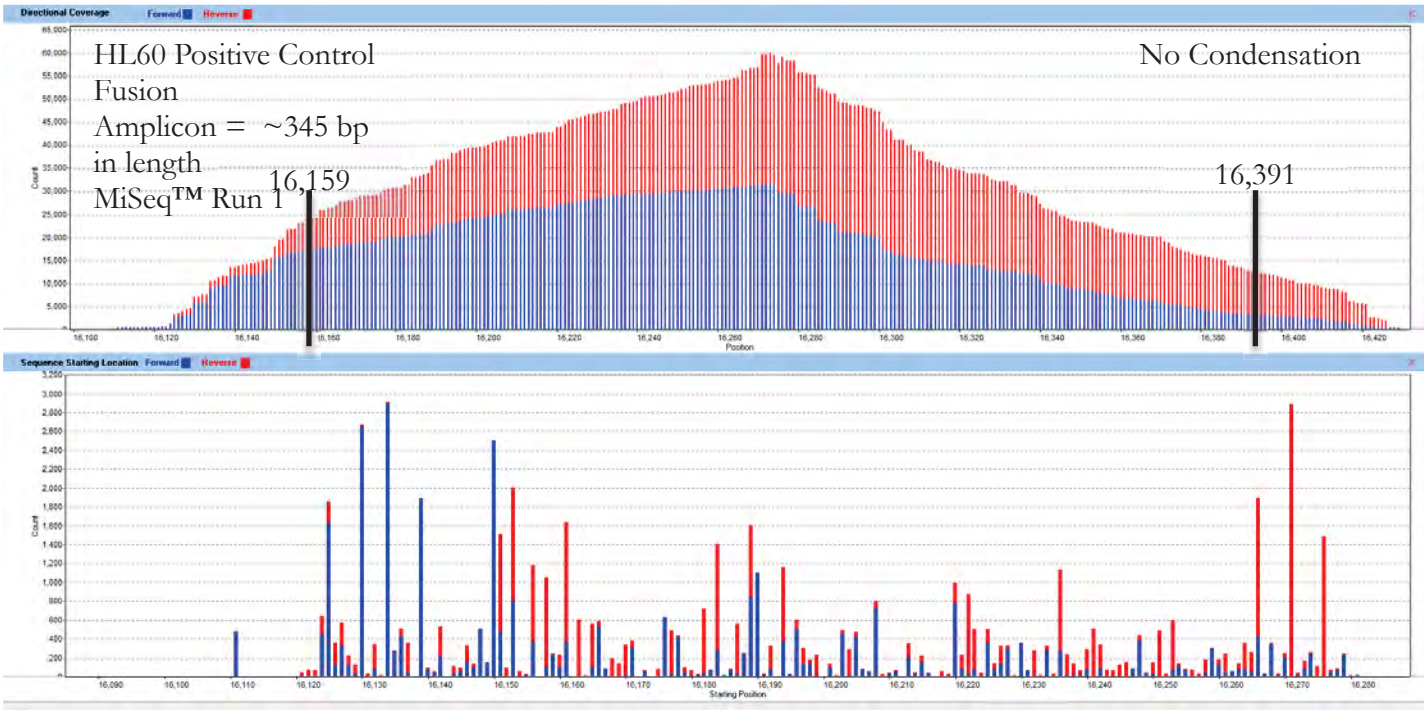



NextGENe™ Sequence Alignment Viewer – Condensed data

*Figure 17 (A and B): Visualization of HL60 Variants using NextGENe® Sequence Alignment Viewer. A.) Illumina® MiSeq™ data generated without applying sequence condensation prior to alignment. The circled read contains 6 out of 19 NumtS associated variants, indicating that it originated from the nuclear DNA of a single individual. However, this read was difficult to locate within the data set due to the high depth of coverage. B.) Sequence condensation was used to reduce the number of reads appearing in the NextGENe® Sequence Alignment Viewer. As a result, mixture detection and variant donor association is much more rapid.*

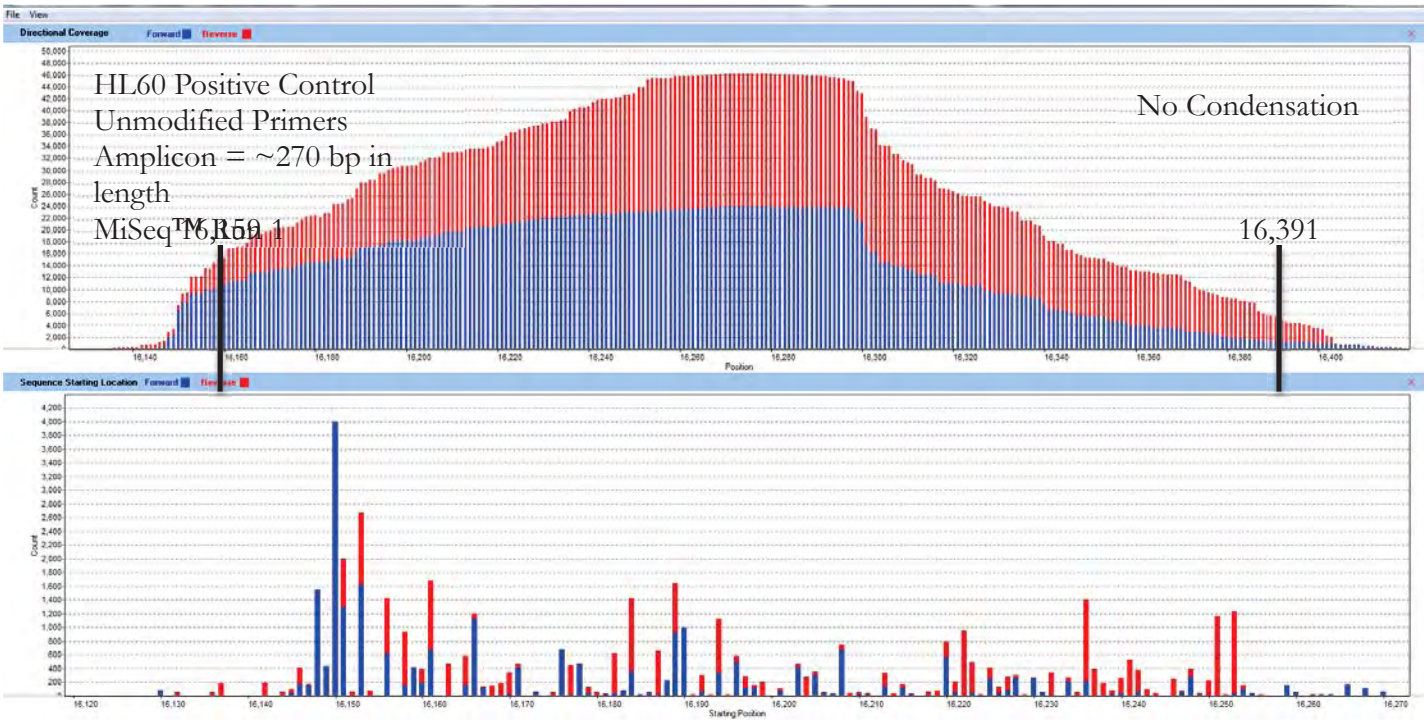
NextGEN*e*® Sequence Alignment Viewer includes an application that allows for the generation of reports that show distribution of coverage across the length of the amplicon, and distribution of sequence starting position of forward and reverse reads. These reports enabled assessment of Nextera™ tagmentation efficiency with respect to the 345 bp and 270 bp amplicons. Though coverage decreases at the 3' ends of forward and reverse reads, several thousand-fold coverage is still achieved at the terminal ends of both amplicons (Figure 18A and 18B). Low-level NumtS variants are called in positions near the terminal ends, with higher overall frequencies observed in 270 bp amplicons, suggesting no significant loss of data in shorter amplicons. This is also evident in HL60 data whose 3 expected variants fall 34 bases from the 3' end of the forward primer, 29 bases from the 3' end of the reverse primer and in the center of the amplicon. However, more studies are needed to determine if tagmentation results in data loss at the ends of amplicons when fold coverage is much lower.

A.




*Created with NextGEN*e*™ viewer

B.




*Created with NextGEN*e*™ viewer

*FIGURE 18 (A and B) – Distribution reports for HL60 Illumina® MiSeq™ data. A.) Coverage across the length of the 345 bp amplicon, based on read direction. B.) Coverage across the length of the 270 bp amplicon. As expected, coverage decreases at the 3' ends of forward and reverse reads for both amplicons. However, appreciable coverage is maintained, and no significant loss of data is observed in either case. When looking at the graph showing read starting position the pattern of tagmentation appears to show bias. The majority of reads seem to start at the 5' ends of forward and reverse reads.*

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | FP_Run1 | FP_run2 | CP_Run1 | CP_Run2 | FP_Run1 | FP_run2 | CP_Run1 | CP_Run2 |
| 16189A | 0 | 0.97 | 21.49 | 12.7 | 0.16 | 0.32 | 0.3 | 0.87 |
| 16218T | 5.86 | 0 | 23.02 | 12.54 | 0.09 | 0.16 | 0.19 | 0.65 |
| 16230G | 6.16 | 1.59 | 21.09 | 11.11 | 0.12 | 0.22 | 0.22 | 0.75 |
| 16249C | 7.25 | 1.74 | 27.22 | 14.11 | 0.39 | 1 | 0.72 | 1.65 |
| 16259A | 0 | 0 | 9.82 | 5.31 | 0.18 | 0.28 | 0.19 | 0.38 |
| 16263C | 0 | 0.96 | 4.61 | 3.34 | 0.29 | 0.66 | 0.32 | 0.81 |
| 16264T | 0 | 0 | 5.84 | 3.33 | 0 | 0.07 | 0.08 | 0.28 |
| 16278T | 98.91 | 98.86 | 93.79 | 97.24 | 97.99 | 99.57 | 99.55 | 99.04 |
| 16284G | 7.1 | 0 | 13.43 | 6.03 | 0.16 | 0.22 | 0.19 | 0.44 |
| 16288C | 7.19 | 0 | 14.17 | 5.45 | 0.16 | 0.51 | 0.27 | 0.53 |
| 16290T | 6.07 | 0 | 13.49 | 5.45 | 0.1 | 0.14 | 0.22 | 0.36 |
| 16293C | 6.01 | 1.13 | 13.49 | 5.41 | 1.53 | 2.71 | 1.15 | 2.1 |
| 16301T | 5.38 | 0 | 15.33 | 3.97 | 0 | 0.21 | 0.24 | 0.85 |
| 16311C | 6.14 | 0.99 | 14.17 | 0 | 0.19 | 0.33 | 0.27 | 0.8 |
| 16319G | 0.76 | 0 | 0 | 4.41 | 0.13 | 0.18 | 0.14 | 0.17 |
| 16355T | 0 | 0 | 9.23 | 4.44 | 0 | 0 | 0.33 | 0.68 |
| 16356C | 0 | 0 | 9.23 | 5.43 | 0.27 | 0.25 | 0.31 | 0.78 |
| 16368C | 5.01 | 0 | 10.71 | 1.54 | 0.17 | 0.19 | 0.34 | 0.45 |
| 16390A | 0 | 0 | 0 | 0 | 0.13 | 0.13 | 0.14 | 0.24 |
| Freq Avg | 5.72 | 1.23 | 14.15 | 6.53 | 0.27 | 0.45 | 0.31 | 0.71 |
| StDev | 1.8 | 0.35 | 6.29 | 3.83 | 0.35 | 0.63 | 0.24 | 0.48 |

*Table 29: Frequencies of NumtS associated variants in HL60. Data was generated with and without sequence consolidation. In general, NumtS frequencies are slightly higher in amplicons generated using unmodified primers. No loss of data is observed in positions close to the proximal ends of the amplicons. Sequence condensation results in skewed frequencies across all NumtS positions. Position 16278 is an expected HL60 variant, therefore the frequency at this position is expected to be 100%.*

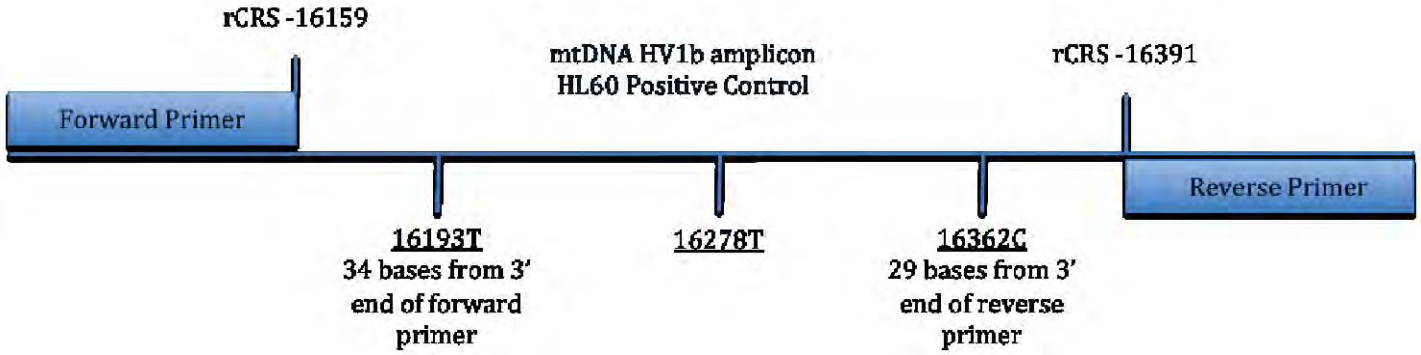



*Figure 19: Schematic illustration showing the positions of known HL60 variants with respect to mtDNA template specific primers. Two variants lie in close proximity to the distal ends of the amplicons, while one variants is found in the center. This is useful in assessing the degree of Nextera™ tagmentation and potential loss of data in these regions.*

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | |
| | | | | | | | | | |
| Expected variants | 16193T | 100% | 56,090 | 100% | 42,633 | 99% | 73,557 | 99% | 50,600 |
| | 16278T | 100% | 89,599 | 100% | 70,337 | 100% | 103,517 | 100% | 69,852 |
| | 16362C | 100% | 52,527 | 100% | 42,283 | 100% | 68,562 | 100% | 49,055 |
| | | | | | | | | | |
| Expected variants | 16193T | 99.28% | 55,039 | 99.50% | 44,797 | 98.60% | 50,396 | 98.25% | 53,326 |
| | 16278T | 97.91% | 65,535 | 99.50% | 61,449 | 98.90% | 65,535 | 99.12% | 65,535 |
| | 16362C | 99.55% | 58,082 | 99.73% | 47,577 | 99.22% | 49,936 | 99.39% | 49,959 |

*Table 30: Expected HL60 variant frequencies and fold coverages for Illumina® MiSeq™ runs 1 and 2 reported by SoftGenetics NextGENe®, and MiSeq™ Reporter software. Though a decrease in coverage is observed at positions 16193, and 16362 the variant frequencies remain the same across both runs. There is no significant difference between in Nextera™ tagmentation efficiency between amplicons.*

Due to the homopolymer associated read errors, time required for library preparation, and per sample cost, we do not recommend the Roche GS-Junior™ for use in forensic mtDNA analysis. Conversely, the Illumina® MiSeq™ reversible terminator chemistry does not lend itself to read errors associated with homopolymer regions, and hence is better suited for detecting minor variants at deep sequence coverages. In addition, exponentially more samples can be multiplexed per run than on the Roche GS Junior™ decreasing the per run cost substantially. Furthermore, Illumina® enzymatic tagmentation library preparation with the commercially available Nextera™ kit has simplified library preparation.

The secondary analysis software packages described in this paper (MiSeq™ Reporter and SoftGenetics NextGEN*e*®) are both excellent options for the analysis of Illumina® NGS data. The additional cost of the NextGEN*e*® software may be difficult for forensic laboratories to absorb. However, the software package offers several applications that the Illumina® Miseq™ Reporter does not. With NextGEN*e*® software, the analyst has the ability to easily adjust quality-filtering parameters, and requeue the data for analysis. A built-in variant comparison tool allows multiple files to be pulled into the software and directly compared. This makes for rapid, and simple assessment of the impact that changes to analytical parameters can have on data interpretation. The NextGEN*e*® Sequence Alignment Viewer enables the analyst to view a distribution of

basecalls across reads for resolution of donor dependent SNPs. This is exceptionally helpful in cases where a mixture may be present. Additionally, the analyst has the option to view histogram reports, which show distribution of coverage across the length of the reference, sequencing starting point of forward and reverse reads, and average read length.

**Results – Section 7 Low Level Variant Detection Experiments on the Illumina® MiSeq™**

A series of mixture experiments were performed on the Illumina® MiSeq™—a smaller Illumina® platform that performs the same sequencing-by-synthesis chemistry but at a smaller yield. The Illumina® MiSeq™ not only combines cluster generation and the previously unmentioned paired-end module, but is also cheaper to operate and features a kit based sample preparation strategy that is far more user friendly and less daunting than the operation of the Illumina® $GAII_x$. The paired-end module allows for sequencing of clusters from both sides of template molecules, recovering sequence data that could otherwise not be obtained with a single-read sequencing strategy and increasing the quality of reads that obtain redundant coverage. While the estimated yield of the Illumina® MiSeq™ is smaller than the Illumina® $GAII_x$, it could still yield the high degree of coverage needed to optimally call minor variants: an estimated 19,500X coverage per amplicon for our project. Values of estimated coverage will increase as improvements are made in supplied reagent kits and flowcells, allowing a far greater degree of library optimization in future projects.

Templates were prepared from buccal and blood extracts of four donors using unmodified HV1a, HV1b, HV2a, HV2b primer sets and the Roche® FastStart family of PCR reagents. Amplicons were then cleaned, quantified in quintuplicate, normalized, and pooled by donor. Using these pooled samples, mixtures were then constructed at the four levels of detection and in reciprocal fashion. Each unmixed sample, mixture, and experimental control was taken through Nextera® XT: an enzymatic sample preparation kit used to prepare nucleic acids for sequencing on the Illumina® MiSeq™. Nextera® XT uses modified transposons that fragment template molecules at random and ligate primer sequences to the flanking ends of cut templates. A short-interval PCR incorporates the adapters needed to hybridize template molecules to the flowcell and also incorporates a sample specific index. Subsequent steps clean PCR products, concentrate DNA products of an acceptable length, dilutes products to an appropriate concentration, and then pools all samples into a single library which may then be loaded onto the delivery cartridge. As a precautionary measure, 8 pM PhiX was spiked into the pooled amplicon library to constitute 20% of the library by volume prior to loading the library onto the delivery cartridge. Once loaded onto the Illumina® MiSeq™, the cartridge will deliver DNA products to the flowcell and the DNA will then be sequenced.

Results from the Illumina® MiSeq™ answered many questions that were posed at the onset of the mixture experiments. The generated FASTQ files are demultiplexed by the instrument and were binned into two files representing each direction of paired-end sequencing. Files were then uploaded to the Galaxy™ cloud (www.galaxyproject.org) where the analysis pipeline was modified to reflect the effects of paired end sequencing. The bioinformatics pipeline is summarized in Figure 20:

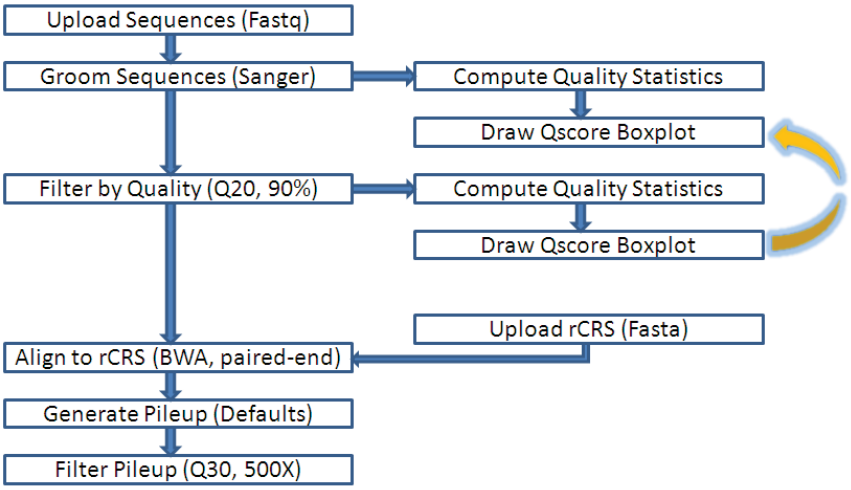



*Figure 20 – The bioinformatics analysis pipeline. Drawing quality score boxplots allows visualization of base quality before and after filtering, if desired. Alignment with BWA paired-end aligns the different sets of read data—forward and reverse—into a single alignment file.*

The output files were then imported into an Microsoft Excel™ spreadsheet where simple statistics and sorting were applied to each base. Examples of alignment data are presented in Table 31. From left to right, the columns are as follows: the reference chromosome (rCRS), the base position within the reference, the reference call at that position, the total number of reads that align over that position, the specific number of A, C, G, and T calls at that position, the quality adjusted reads or the number of reads remaining after the last quality filter, and the total number of variants, which are the number of base-calls that disagree with the reference at that position. On the other side of the divider are applied statistics. Respectively, they are percent major, or the percentage of base calls from the quality adjusted reads that are identified as coming from the major contributor, and the percent minor, which is the percentage of reads that represent the second highest frequency base called. Sorting the data by total number of variants gives a good indication of the variants obtained from the questioned sample to the reference sequence.

| CHROM | POS | REF | T#oRs | A CALLS | C CALLS | G CALLS | T CALLS | QARs | T#oDs | % MAJ | % MIN |
|---|---|---|---|---|---|---|---|---|---|---|---|
| chrM | 39 | C | 516 | 1 | 478 | 1 | 0 | 480 | 2 | 99.58 | 0.21 |
| chrM | 40 | T | 751 | 0 | 0 | 1 | 703 | 704 | 1 | 99.86 | 0.14 |
| chrM | 41 | C | 2771 | 0 | 2670 | 0 | 0 | 2670 | 0 | 100.00 | 0.00 |
| chrM | 42 | T | 2845 | 0 | 1 | 0 | 2753 | 2754 | 1 | 99.96 | 0.04 |
| chrM | 43 | C | 2970 | 1 | 2890 | 0 | 0 | 2891 | 1 | 99.97 | 0.03 |
| chrM | 44 | C | 3116 | 0 | 2997 | 0 | 0 | 2997 | 0 | 100.00 | 0.00 |
| chrM | 45 | A | 3135 | 2984 | 0 | 0 | 1 | 2985 | 1 | 99.97 | 0.03 |
| chrM | 46 | T | 3216 | 0 | 0 | 0 | 3111 | 3111 | 0 | 100.00 | 0.00 |

*Table 31 – The first eight base positions and corresponding base-calls from a DNA sample obtained from donor 003-54M using the Galaxy™ bioinformatics pipeline. Purines are shown in light blue, pyrimidines are tan.*

| CHROM | POS | REF | T#oRs | A CALLS | C CALLS | G CALLS | T CALLS | QARs | T#oDs | % MAJ | % MIN |
|---|---|---|---|---|---|---|---|---|---|---|---|
| chrM | 152 | T | 8013 | 0 | 7883 | 0 | 1 | 7884 | 7883 | 99.99 | 0.01 |
| chrM | 150 | C | 8012 | 0 | 9 | 0 | 7847 | 7856 | 7847 | 99.89 | 0.11 |
| chrM | 16069 | C | 8007 | 0 | 5 | 0 | 7749 | 7754 | 7749 | 99.94 | 0.06 |
| chrM | 16126 | T | 8012 | 0 | 7702 | 1 | 8 | 7711 | 7703 | 99.88 | 0.10 |
| chrM | 16242 | C | 7986 | 7644 | 143 | 1 | 1 | 7789 | 7646 | 98.14 | 1.84 |
| chrM | 16221 | C | 7984 | 0 | 32 | 0 | 7471 | 7503 | 7471 | 99.57 | 0.43 |
| chrM | 16195 | T | 7645 | 0 | 7120 | 0 | 13 | 7133 | 7120 | 99.82 | 0.18 |
| chrM | 16193 | C | 7476 | 0 | 19 | 1 | 6953 | 6973 | 6954 | 99.71 | 0.27 |
| chrM | 73 | A | 7161 | 3 | 0 | 6574 | 0 | 6577 | 6574 | 99.95 | 0.05 |
| chrM | 16319 | G | 3830 | 3536 | 0 | 24 | 0 | 3560 | 3536 | 99.33 | 0.67 |
| chrM | 295 | C | 2490 | 0 | 18 | 0 | 2224 | 2242 | 2224 | 99.20 | 0.80 |
| chrM | 263 | A | 2414 | 0 | 0 | 2202 | 0 | 2202 | 2202 | 100.00 | 0.00 |

*Table 32 – Control region variants from a blood-extracted DNA sample from donor 003-54M. Highlighted bases are the expected variants that were observed in Sanger sequencing data. The yellow highlight indicates a position in which the minor variant is above a 1% threshold.*

Since these data are derived from a single donor, the results indicate positions that may experience higher levels of variation in our analysis of mixtures. Sorting mixture data by 'percent minor' brings the positions in which we expect to observe variations from the reference to the top of the data set. An example for a 5% mixture between two donors that has been sorted by 'percent minor' is presented in Table 34. This data sorting function is very helpful, and indicates that the instrument is able to discern minor variations in mixed mitochondrial samples, like those we expect to find in heteroplasmic individuals, down to the 5% level of detection. The reciprocal mixture experiment for these donors and the analogous 5% mixture experiments performed using buccal extracts as source material corroborate these findings. In fact, the Illumina® MiSeq™

| CHROM | POS | REF | T#oRs | A CALLS | C CALLS | G CALLS | T CALLS | QARs | T#oDs | % MAJ | % MIN |
|---|---|---|---|---|---|---|---|---|---|---|---|
| chrM | 16242 | C | 7982 | 6664 | 871 | 4 | 0 | 7539 | 6668 | 88.39 | 11.55 |
| chrM | 16126 | T | 4884 | 0 | 4311 | 0 | 296 | 4607 | 4311 | 93.57 | 6.43 |
| chrM | 16193 | C | 7824 | 0 | 460 | 0 | 6852 | 7312 | 6852 | 93.71 | 6.29 |
| chrM | 16195 | T | 7873 | 0 | 7036 | 0 | 448 | 7484 | 7036 | 94.01 | 5.99 |
| chrM | 16221 | C | 7671 | 0 | 413 | 0 | 6598 | 7011 | 6598 | 94.11 | 5.89 |
| chrM | 204 | T | 5043 | 0 | 235 | 0 | 4040 | 4275 | 235 | 94.50 | 5.50 |
| chrM | 16069 | C | 3416 | 0 | 175 | 0 | 3024 | 3199 | 3024 | 94.53 | 5.47 |
| chrM | 16224 | T | 7812 | 0 | 383 | 0 | 6766 | 7149 | 383 | 94.64 | 5.36 |
| chrM | 295 | C | 1214 | 0 | 53 | 0 | 1002 | 1055 | 1002 | 94.98 | 5.02 |
| chrM | 16223 | C | 7826 | 1 | 6927 | 0 | 365 | 7293 | 366 | 94.98 | 5.00 |
| chrM | 152 | T | 5499 | 0 | 5097 | 0 | 263 | 5360 | 5097 | 95.09 | 4.91 |
| chrM | 16274 | G | 7979 | 339 | 0 | 6661 | 0 | 7000 | 339 | 95.16 | 4.84 |
| chrM | 150 | C | 5499 | 0 | 251 | 0 | 5069 | 5320 | 5069 | 95.28 | 4.72 |
| chrM | 16357 | T | 2467 | 1 | 2207 | 0 | 109 | 2317 | 2208 | 95.25 | 4.70 |
| chrM | 16270 | C | 7990 | 0 | 7362 | 0 | 345 | 7707 | 345 | 95.52 | 4.48 |
| chrM | 16352 | T | 2707 | 0 | 103 | 0 | 2398 | 2501 | 103 | 95.88 | 4.12 |
| chrM | 310 | T | 1173 | 0 | 13 | 0 | 554 | 567 | 13 | 97.71 | 2.29 |
| chrM | 316 | G | 1108 | 1 | 10 | 784 | 0 | 795 | 11 | 98.62 | 1.26 |
| chrM | 309 | C | 1180 | 0 | 653 | 0 | 5 | 658 | 5 | 99.24 | 0.76 |
| chrM | 16319 | G | 3712 | 3146 | 1 | 17 | 0 | 3164 | 3147 | 99.43 | 0.54 |
| chrM | 16356 | T | 2487 | 0 | 10 | 0 | 1898 | 1908 | 10 | 99.48 | 0.52 |
| chrM | 16368 | T | 2136 | 0 | 10 | 0 | 2030 | 2040 | 10 | 99.51 | 0.49 |

*Table 33 – Control region variants in a 5% mixture of donors 003-54MBlood (major) and 015-AM30Blood (minor). Gold positions are the expected variants for the major contributor. Green positions are the expected variants for the minor contributor. Yellow indicates a position in which the major contributor exhibited the minor variant above a 1% threshold. Purple indicates a shared variant. All expected variants have a percent minor that clusters around 5%.*

instrument is able to recover minor variants down to the 2, 1, and 0.5% level of detection. It is worth noting, however, that interpretations of true mitochondrial variants at these low levels of resolution are less obvious as the noise becomes more pronounced. In Table 33, the unhighlighted positions 309, 310, and 316 show increased levels of variation. These positions are in close proximity to the HV2 C-stretch, and hence may reveal an inherent weakness of the alignment algorithm in dealing with complex indels. However, these length variants have been well characterized as arising from mixed populations of length variants in the samples themselves and hence should not be confused with as an alignment issue.

Also showing increased variation are positions 16,356 and 16,368. These positions are two of the nineteen NumtS sequence variants (see Section 2 above) that have been discovered to co-amplify with the HV1b primer sets and because of this, are expected to show increased levels of variation in all future NGS studies of the human mitochondrial control region when the source material is likely to contain nuclear DNA. One possible approach that would avoid the co-amplification of the nuclear inserts would be to re-design the HV1b primer set to target the mtDNA exclusively.

While it is possible to explain some of these variants, other sources of noise may be more elusive. Consider Table 34, where two donors have been mixed to reflect 0.5% variants at expected positions. Most of the

| CHROM | POS | REF | T#oRs | A CALLS | C CALLS | G CALLS | T CALLS | QARs | T#oDs | % MAJ | % MIN |
|---|---|---|---|---|---|---|---|---|---|---|---|
| chrM | 302 | A | 839 | 461 | 10 | 0 | 0 | 471 | 10 | 97.88 | 2.12 |
| chrM | 295 | C | 840 | 0 | 12 | 0 | 712 | 724 | 712 | 98.34 | 1.66 |
| chrM | 310 | T | 830 | 0 | 5 | 0 | 357 | 362 | 5 | 98.62 | 1.38 |
| chrM | 16093 | T | 7990 | 0 | 7607 | 0 | 95 | 7702 | 7607 | 98.77 | 1.23 |
| chrM | 185 | G | 7941 | 7437 | 0 | 88 | 0 | 7525 | 7437 | 98.83 | 1.17 |
| chrM | 228 | G | 4216 | 4012 | 0 | 43 | 0 | 4055 | 4012 | 98.94 | 1.06 |
| chrM | 16356 | T | 1693 | 0 | 15 | 0 | 1482 | 1497 | 15 | 99.00 | 1.00 |
| chrM | 16355 | C | 1725 | 0 | 1445 | 0 | 13 | 1458 | 13 | 99.11 | 0.89 |
| chrM | 16366 | C | 819 | 0 | 753 | 0 | 6 | 759 | 6 | 99.21 | 0.79 |
| chrM | 16129 | G | 7993 | 49 | 0 | 7530 | 0 | 7579 | 49 | 99.35 | 0.65 |
| chrM | 237 | A | 3280 | 3212 | 0 | 19 | 0 | 3231 | 19 | 99.41 | 0.59 |
| chrM | 16368 | T | 768 | 0 | 4 | 0 | 699 | 703 | 4 | 99.43 | 0.57 |
| chrM | 16218 | C | 7560 | 1 | 7016 | 0 | 38 | 7055 | 39 | 99.45 | 0.54 |
| chrM | 16069 | C | 8000 | 0 | 41 | 0 | 7630 | 7671 | 7630 | 99.47 | 0.53 |
| chrM | 16126 | T | 8009 | 0 | 7609 | 0 | 40 | 7649 | 7609 | 99.48 | 0.52 |

*Table 34 – Control region variants in a 0.5% mixture of blood-derived DNA samples from donors 001-CF30 (major) and 005-CF40 (minor). Gold positions are the expected variants for the major contributor. Green positions are the expected variants for the minor contributor.*

unexpected variants can be explained by reasons discussed previously. However, positions 237 and 16,366 are neither in proximity to a C-stretch, nor are they associated with a NumtS sequence variant. These sites of increased variation may either be the result of instrument or sequencing error, PCR misincorporation, or represent true biological variation present at low levels. Regardless of the source, without defined interpretational thresholds it is currently challenging to resolve biological variation from noise at or below the 1% level of detection. A sound statistical approach toward the interpretation of noise will be necessary for the development of casework interpretational criteria. We are currently working to develop a statistical interpretation process for noise detection in the development of reasonable interpretation criteria.

Unexpected variation also shows a distinct transitional bias as opposed to transversions and is demonstrated by the outlined cells in Table 35. This non-random observation may be explained by two opposing hypotheses or some combination thereof: 1) that either the underlying variation is representative of true biological variation, in which a sub-population of cells contributes to the minor mitotype observed at a distinct position or 2) that the underlying variation is not biological in origin and is likely the result of polymerase induced error or sequencing error. Polymerase induced error is plausible considering the many cycles of PCR the templates are subjected to in order to prepare templates for sequencing on the Illumina® MiSeq™ (PCR, Nextera® XT PCR, and on-board cluster generation). It is also worth noting that instrument reading error

| CHROM | POS | REF | T#oRs | A CALLS | C CALLS | G CALLS | T CALLS | QARs | T#oDs | % MAJ | % MIN |
|---|---|---|---|---|---|---|---|---|---|---|---|
| chrM | 175 | A | 8021 | 7617 | 0 | 2 | 0 | 7619 | 2 | 99.97 | 0.03 |
| chrM | 176 | A | 8022 | 7658 | 0 | 2 | 0 | 7660 | 2 | 99.97 | 0.03 |
| chrM | 177 | T | 8022 | 0 | 0 | 0 | 7775 | 7775 | 0 | 100.00 | 0.00 |
| chrM | 178 | A | 8015 | 7755 | 0 | 5 | 3 | 7763 | 8 | 99.90 | 0.06 |
| chrM | 179 | T | 8015 | 0 | 2 | 0 | 7809 | 7811 | 2 | 99.97 | 0.03 |
| chrM | 180 | T | 8005 | 0 | 1 | 0 | 7866 | 7867 | 1 | 99.99 | 0.01 |
| chrM | 181 | A | 8003 | 7870 | 0 | 0 | 0 | 7870 | 0 | 100.00 | 0.00 |
| chrM | 182 | C | 8003 | 0 | 7779 | 0 | 6 | 7785 | 6 | 99.92 | 0.08 |
| chrM | 183 | A | 7984 | 7716 | 0 | 4 | 0 | 7720 | 4 | 99.95 | 0.05 |
| chrM | 184 | G | 7982 | 0 | 0 | 7532 | 0 | 7532 | 0 | 100.00 | 0.00 |
| chrM | 185 | G | 7951 | 7504 | 0 | 40 | 0 | 7544 | 7504 | 99.47 | 0.53 |
| chrM | 186 | C | 7945 | 1 | 7606 | 0 | 2 | 7609 | 3 | 99.96 | 0.03 |
| chrM | 187 | G | 7912 | 2 | 0 | 6752 | 0 | 6754 | 2 | 99.97 | 0.03 |
| chrM | 188 | A | 7912 | 6813 | 0 | 4 | 0 | 6817 | 4 | 99.94 | 0.06 |
| chrM | 189 | A | 7884 | 6846 | 0 | 17 | 0 | 6863 | 17 | 99.75 | 0.25 |

*Table 35 – Low-level variation for unsorted sequence data from a buccal-derived DNA sample from donor 001-CF30 shows a distinct transitional bias. Outlined cells represent the second most prevalent basecall at positions with deviant calls from the reference.*

does not seem likely, considering that the Illumina® MiSeq™ utilizes dual color channels to separate, in time, the fluorescent activity of bases A and C from bases G and T; minimizing purine-to-purine and pyrimidine-to-pyrimidine crosstalk. If the underlying variation is representative of true biological variation, then the deep sequencing of hair shafts that are comprised from small, clonally dividing population of cells may yield a mitotype with a larger degree of sequence variation with respect to more homogenous tissue types, such as blood and buccal tissues. Alternatively, if the underlying variation is the result of polymerase misincorporation, then the deep sequencing of a control genome that has sustained various cycles of PCR may reveal a discrepancy in the observed variation at discrete locations within the genome. We are testing these hypotheses.

## Conclusions

Using a combination of procedures and experimental design, we have introduced a potential new forensic DNA analysis method focused on emerging, powerful DNA analysis techniques. The results of this project have exceeded our expectations. Not only have we obtained the expected results and satisfied the goals of our study, but, more broadly, we have demonstrated the vast utility of these methods. We moved from the use of fusion primers to direct library preparation methods, developed limited whole genome amplification procedures, directly compared two popular NGS chemistries, confirmed the deep sequencing limitations of one chemistry method, demonstrated the power of NGS to detect previously unidentified genetic variants, employed the use of multiplex amplification strategies to target whole mt-genome data, developed a rapid library preparation method for reference sequences, evaluated a number of software packages and algorithms, conducted substantial research on quality metrics and variables, and determined the limits of detecting low-level variants using the procedure that we determined to have the optimal performance characteristics of those that we investigated.

## Implications for Criminal Justice policy and practice

Although human mtDNA analysis is currently only performed in a small subset of forensic DNA laboratories, its utility in some important casework contexts is incontrovertible. Part of the reason for the current limitation on its practice is that the informativeness of mtDNA is much less than that provided by forensic STR analysis. However, because of random stochastic effects that occur with low-level DNA samples, STR analysis can become problematic with some sample types. Many laboratories attempt to overcome the inherent limitations of STR analysis by performing low copy number (LCN) STR analyses on these kind of samples.

Recent legal challenges to LCN analysis have highlighted the limitations of this approach. We believe that to be more useful in forensic DNA laboratories, the inherent advantages of mtDNA analysis need to be more fully leveraged. This expansion should be extended in two directions, by the amount of sequence information analyzed, and by the depth of sequence analyzed at each position of comparison. The reason for the amount of sequence data is obvious, as more sequencing information means that the probability of exclusion is enhanced. The reason for the depth requirement has always been appreciated, but until recently no reliable and commercially viable methods have been available for detecting this level of DNA sequence variation. Newly emerging technologies, such as deep sequencing, and the eventual decrease in costs associated with them, makes this goal now obtainable.

## Implications for further research

### Streamlining Protocols and Whole Genome Amplification

It is necessary to examine an alternative, streamlined, two-step method of directly preparing sufficient DNA from forensic samples for forensic mtDNA analysis of the entire mt-genome. This would involve coupling whole genome amplification (WGA) techniques, using a cocktail of mt-specific primers covering the entire mt-genome from both strands, to direct library preparation using the Nextera protocol. In this approach, the oligos will bind to their template sequences at conserved regions and be extended by a DNA polymerase to generate longer fragments for subsequent cleavage and processing during in the Nextera library preparation step. The oligonucleotides should be designed to target known conserved regions of the human mt-genome from both the light and heavy strands of the molecule. Following the first step, the accumulation of mt-specific template during the WGA reaction should be monitored by the use of quantitative real-time PCR using a human-specific probe and primer set, such as that described in our work.

At the point where sufficient DNA has accumulated as a result of the initial WGA-based reaction, then the second, targeted portion of the approach will take place. We envision that the first WGA step will result in a series of fragments with significant complementarity such that double-stranded products will be present (see Figure 21, top). These newly synthesized fragments will serve as the template for the subsequent enzymatic accumulation of NGS-targeted fragments using a transposase approach such as that employed in the Nextera kit. The double-stranded products will be those utilized in the second step, and need not be blunt-ended but rather may contain free, single stranded ends that should not interfere with the second step (see Figure 9, bottom). However, we will also investigate creating blunt-end double stranded fragments at this stage and comparing the results to the process without this step. Figure 21 shows a graphic depiction of generalized bi-directional whole genome amplification of the two mt-genome strands of the circular molecule and also the linear stylized view of these products.

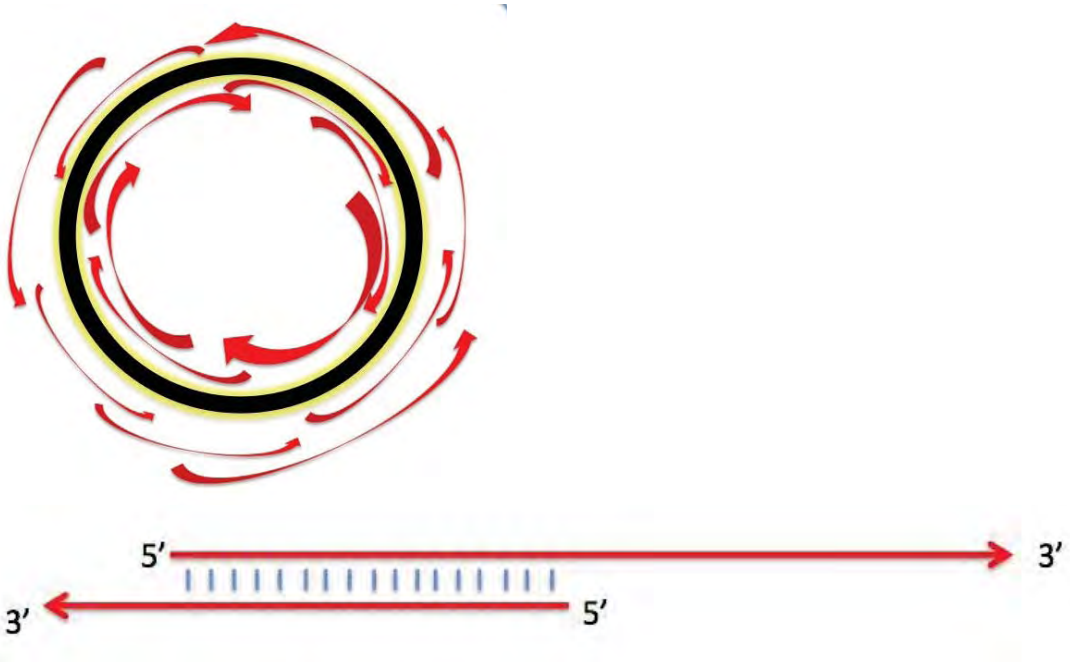



*Figure 21 – Stylized depiction of the expected products from an mtDNA-based whole genome amplification reaction. Top – circular view, bottom – linear view. These partially double stranded products should serve as templates for limited whole genome amplification, providing sufficient template for whole mt-genome analysis.*

**Mixture Devonvolution**

Using past technologies, such as Sanger sequencing of mtDNA amplicons, the amount of signal arising from each component of a mixture is not retained with any concomitant quantitative information. Consequently, the relative amount of each component of the mixture was difficult or impossible to ascertain. In some cases the dynamic range of the instrument did not allow the user to identify minor components that fell beneath the threshold level of detection.

Conversely, deep sequencing results within NGS offer hundreds or thousands of individual sequencing reactions that provide a level of information that allows mixture deconvolution. This is based ultimately on counting the number of independent runs comprising the mixture. Accordingly, the evidential sum of a particular evidentiary sample contains an added characteristic, namely, a complex collection of components that can now be considered both individually and collectively.

Suppose a mixture of individual sequences is comprised of four distinct components: 50% of which are contributed by the major component; 30% by a dominant minor component; 15% by a second minor component; and 5% by a third component (50/30/15/5). Following sample preparation, current deep sequencing methods will generate sequence information from thousands of individual runs, each derived from a single starting template molecule. Given the appropriate confidence limits (which should be confirmed with validation studies conducted using each technique) the number of fragments containing sequence data from the major component should be at or near 50% of the total, those showing the sequence from the dominant minor about 30% of the total, those from the second minor component about 15%, and finally the last minor component, 5%. Provided that sufficient data is obtained using these newly emerging techniques, the mixture can be de-convoluted as a mixture of four different sequence variants in the approximate ratio of 50/30/15/5 by simply counting the numbers of each differing sequence component in the mixture. Hence, provided that the components differ in proportions such that their statistical confidence intervals do not overlap, this approach allows the investigative team to identify the sequence of each distinct component of the mixture. Hence, deep sequencing potentially provides a level of information not attainable with earlier technologies.

**Future Protocol Development**

We anticipate that the implementation of amplification strategies that will allow forensic laboratories to assess whether or not deep sequencing runs are necessary or not in any given casework situation. In many cases, at least initially, they may not be necessary. However, provided that no sensitivity is lost by the use of NGS protocols, the user would have the option of using deep sequencing runs to generate further information should the particular case warrant it. Accordingly, directly measuring the efficacy and cost efficiency of NGS methods compared to existing methods is important.

NGS technologies also offer the potential of rapidly generating population databases to support forensic casework analyses. We have addressed this issue and described a robust and simple library preparation method for the Illumina platform. However, we believe that this activity should be coordinated and conducted by a number of participating laboratories in order to generate a large, high quality population database.

While we have shown that DNA extraction efficiency can be enhanced using components of various solid-phase capture methods in novel ways, further work in this area is warranted. One obstacle to these experimental approaches is that many commercially available kits contain components that cannot be easily separated and evaluated in new combinations. We surmise that the particular buffers used in the extraction process provide the requisite ionic strength and chaotropic salts that serve to enhance the binding of DNA to the silica particles, resulting in improved extraction efficiencies.

Full forensic validation of NGS will require substantial additional effort. While we have initiated this process, full implementation of NGS in forensic casework requires, among other things, a commitment to ongoing developmental research as well as training current examiners in the new techniques and instuments. Rather than fund these large transitional phases piecemeal, it would be beneficial for NIJ and other funding agencies to develop a plan for continued funding with well-defined milestones and goals.

## References

### General and Forensic Use of MtDNA References


Allard M.W., Miller K., Wilson M.R., Monson K.L., Budowle B. (2002) Characterization of the Caucasian haplogroups present in the SWGDAM forensic mtDNA data set for 1771 human control region sequences. J. Forensic Sciences. 47: 1215-1223.

Allen M., Engström A.S., Meyers S., Handt O., Saldeen T., von Haeseler A., Pääbo S., and Gyllensten U. (1998) Mitochondrial DNA sequencing of shed hairs and saliva on robbery caps: sensitivity and matching probabilities. J. Forensic Sci. 43:453 – 464.

Anderson S., Bankier A.T., Barrell 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. (1981) Sequence and organization of the human mitochondrial genome. Nature. 290:457–65.

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.

Barber AL, Foran DR. (2006) The utility of whole genome amplification for typing compromised forensic samples. J Forensic Sci. 2006 Nov;51(6):1344-9.

Ballantyne KN, van Oorschot RA, Mitchell RJ. (2007) Comparison of two whole genome amplification methods for STR genotyping of LCN and degraded DNA samples. Forensic Sci Int. 2007 Feb 14;166(1):35-41.

Benschop CC, van der Beek CP, Meiland HC, van Gorp AG, Westen AA, Sijen T. (2011) Low template STR typing: effect of replicate number and consensus method on genotyping reliability and DNA database search results. Forensic Sci Int Genet. 2011 Aug;5(4):316-28.

Brandstätter A., Parsons T.J., Niederstätter H., Parson W. (2003) Rapid Screening of mtDNA Coding Region SNPs for the Identification of Caucasian Haplogroups. Int. J. Legal Med. 2003 Oct;117(5):291-8.

Budowle B., DiZinno J., Wilson M. (1999) Interpretation guidelines for mitochondrial DNA sequencing. in: Proccedings of the Tenth International Symposium on Human Identification. Promega Corporation, Madison, WI. http://www.promega.com/ussvmp10proc/default.html

Budowle B., Allard M.W., Wilson M.R., Chakraborty R. (2003) Forensics and mitochondrial DNA: applications, debates, and foundations. Ann. Rev. Genomics and Human Genet. 4:119-41.

Coble M.D., Hamm R.S., O'Callaghan J.E., Letmanyi I.H., Peterson C.T., and Parsons T.J. (in press) Single Nucleotide Polymorphisms over the entire mtDNA Genome that Increase the Forensic Power of mtDNA Testing in Caucasians.

De Benedictis G, Carrieri G, Garasto S, Rose G, Varcasia O, Bonafe M, Franceschi C, and Jazwinski SM (2000) Does a retrograde response in human aging and longevity exist? Exp Gerontol 35: 795.

Finnila S., Lehtonen M.S., and Majamaa K., (2001) Phylogenetic Network for European mtDNA. Am. J. Hum. Genet. 68:1475-84. 36

Forster L, Thomson J, Kutranov S., Direct comparison of post-28-cycle PCR purification and modified capillary electrophoresis methods with the 34-cycle "low copy number" (LCN) method for analysis of trace forensic DNA samples. Forensic Sci Int Genet. (2008) Sep;2(4):318-28.
Grisedale KS, van Daal A. Comparison of STR profiling from low template DNA extracts with and without the consensus profiling method. (2012). Investig Genet. 2012 Jul 2;3(1):14. doi: 10.1186/2041-2223-3-14.

Helgason A., Hickey E., Goodacre S., Bosnes V., Stefansson K., Ward R., Sykes B. (2001) mtDNA and the Islands of the North Atlantic: Estimating the Proportions of
Norse and Gaelic Ancestry. Am. J. Hum. Genet. 68:723-37.

Lee JC, Tsai LC, Lai PY, Lee CC, Lin CY, Huang TY, Linacre A, Hsieh HM. (2012) Evaluating the performance of whole genome amplification for use in low template DNA typing. 2012. Med Sci Law. 2012 Oct;52(4):223-8. doi: 10.1258/msl.2012.011126.

Maciejewska A, Jakubowska J, Pawłowski R. (2013) Whole genome amplification of degraded and nondegraded DNA for forensic purposes. Int J Legal Med. 2013 Mar. 127(2):309-19.

Monson K. L., Miller K. W. P., Wilson M. R., DiZinno J. A., and Budowle B. (2002) The mtDNA population database: An integrated software and database resource, *Forensic Science Communications* [Online]. Available: www.fbi.gov/hq/lab/fsc/backissu/april2002/miller1.htm

Parsons T.J. and Coble M.D. (2001) Increasing the forensic discrimination of mitochondrial DNA testing through analysis of the entire mitochondrial DNA genome. *Croatian Medical J.* 42(3): 304-309.

Pesole G., Sbisà E., Preparata G., and Saccone C. (1992) The evolution of the mitochondrial D-loop region and the origin of modern man. Mol. Biol. Evol. 9(4): 587-598.

Pesole G., Gissi C., De Chirico A., and Saccone C. (1999) Nucleotide substitution rate of mammalian mitochondrial genomes. J. Mol. Evol. 48(4):427-434.

Pfeifer CM, Klein-Unseld R, Klintschar M, Wiegand P. Comparison of different interpretation strategies for low template DNA mixtures (2012). Forensic Sci Int Genet. 2012 Dec;6(6):716-22.

Sun G, Kaushal R, Pal P, Wolujewicz M, Smelser D, Cheng H, Lu M, Chakraborty R, Jin L, Deka R. (2005) Whole-genome amplification: relative efficiencies of the current methods. Leg Med (Tokyo). 2005 Oct;7(5):279-86.

Tamura K., Nei M. (1993) Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol. Biol. Evol.10(3):512-26.

Torroni A, Achilli A, Macaulay V, Richards M, Bandelt HJ. Harvesting the fruit of the human mtDNA tree. Trends Genet. 2006 Jun;22(6):339-45.

Vigilant L. (1999) An Evaluation of Techniques for the Extraction and Amplification of DNA from Naturally Shed Hairs. Biol. Chem. 380:1329-1331.

Vallone PM, Just RS, Coble MD, Butler JM, Parsons TJ. (2003) A multiplex allele specific primer extension assay for 11 forensically informative SNPs distributed throughout the mitochondrial genome. Int J Legal Med. 2004 Jun;118(3):147-57.

Wallace DC, Brown MD, Lott MT (1999) Mitochondrial DNA variation in human evolution and disease. Gene 238: 211,

Wilson M.R., DiZinno J.A., Polanskey D., Replogle J., and Budowle B. (1995) Validation of mitochondrial DNA sequencing for forensic casework analysis. Int. J. of Legal Medicine. 108:68-74.

Wilson, M.R., Polanskey, D., Butler, J.M., DiZinno, J.A., Replogle, J., and Budowle, B. (1995) "Extraction, PCR Amplification, and Sequencing of Mitochondrial DNA from Human Hair Shafts" BioTechniques: 18 (4): 662-669.

Wilson M. and Allard M.W. 2004 Phylogenetic and mitochondrial DNA analysis in the forensic sciences. Forensic Science Reviews 16:37-62.

Wolstenholme D. R., (1992) Animal mitochondrial-DNA: structure and evolution. Int. Rev. Cytol. 141:173-216.39

**Pyrosequencing References**

Hoffmann C, Minkah N, Leipzig J, Wang G, Arens MQ, Tebas P, Bushman FD. (2007) DNA bar coding and pyrosequencing to identify rare HIV drug resistance mutations. Nucleic Acids Res. 2007;35(13):e91.

Weber AP, Weber KL, Carr K, Wilkerson C, Ohlrogge JB. (2007) Sampling the Arabidopsis transcriptome with massively parallel pyrosequencing.
Plant Physiol. 2007 May;144(1):32-42.

Farrell CL, Crimm H, Meeh P, Croshaw R, Barbar TD, Vandersteenhoven JJ, Butler W, Buckhaults P. (2008) Somatic mutations to CSMD1 in colorectal adenocarcinomas.Cancer Biol Ther. 2008 Jan 22;7(4).

Novaes E, Drost DR, Farmerie WG, Pappas GJ Jr, Grattapaglia D, Sederoff RR, Kirst M. (2008) High-throughput gene and SNP discovery in Eucalyptus grandis, an uncharacterized genome.BMC Genomics. 2008 Jun 30;9(1):312.

Briggs AW, Stenzel U, Johnson PL, Green RE, Kelso J, Prüfer K, Meyer M, Krause J, Ronan MT, Lachmann M, Pääbo S. (2007) Patterns of damage in genomic DNA sequences from a Neandertal.Proc Natl Acad Sci U S A. 2007 Sep 11;104(37):14616-21.

Gilbert MT, Binladen J, Miller W, Wiuf C, Willerslev E, Poinar H, Carlson JE, Leebens-Mack JH, Schuster SC. (2007) Recharacterization of ancient DNA miscoding lesions: insights in the era of sequencing-by-synthesis. Nucleic Acids Res. 2007;35(1):1-10.

Gilbert MT, Tomsho LP, Rendulic S, Packard M, Drautz DI, Sher A, Tikhonov A, Dalén L, Kuznetsova T, Kosintsev P, Campos PF, Higham T, Collins MJ, Wilson AS, Shidlovskiy F, Buigues B, Ericson PG, Germonpré M, Götherström A, Iacumin P, Nikolaev V, Nowak-Kemp M, Willerslev E, Knight JR, Irzyk GP,

Perbost CS, Fredrikson KM, Harkins TT, Sheridan S, Miller W, Schuster SC. (2007) Whole-genome shotgun sequencing of mitochondria from ancient hair shafts. Science 28;317(5846):1927-30.

Krause J, Dear PH, Pollack JL, Slatkin M, Spriggs H, Barnes I, Lister AM, Ebersberger I, Pääbo S, Hofreiter M. (2006) Multiplex amplification of the mammoth mitochondrial genome and the evolution of Elephantidae. Nature. 2006 Feb 9;439(7077):724-7.

Ronaghi M, Uhlén M, Nyrén P. (1998) A sequencing method based on real-time pyrophosphate. Science 281: 363.

Ronaghi M, Karamohamed S, Pettersson B, Uhlén M, Nyrén P. (1996) Real-time DNA sequencing using detection of pyrophosphate release. Analytical Biochemistry 242: 84.

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer ML, Jarvie TP, Jirage KB, Kim JB, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu P, Begley RF, Rothberg JM. (2005). Genome sequencing in microfabricated high-density picolitre reactors. Nature 437:376-380.

Sogin ML, Morrison HG, Huber JA, Mark Welch D, Huse SM, Neal PR, Arrieta JM, Herndl GJ. (2006) Microbial diversity in the deep sea and the underexplored "rare biosphere". Proc Natl Acad Sci USA 103:12115-12120.

**Human mtDNA Whole Genome references**

Palanichamy,M.G., Sun,C., Agrawal,S., Bandelt,H.J., Kong,Q.P.,Khan,F., Wang,C.Y., Chaudhuri,T.K., Palla,V. and Zhang,Y.P. Phylogeny of mitochondrial DNA macrohaplogroup N in India, based on complete sequencing: implications for the peopling of South Asia. Am. J. Hum. Genet. 75 (6), 966-978 (2004)

Friedlaender,J.S., Friedlaender,F.R., Hodgson,J.A., Stoltz,M., Koki,G., Horvat,G., Zhadanov,S., Schurr,T.G. and Merriwether,D.A. Melanesian mtDNA Complexity
(er) PLoS ONE 2, E248 (2007)

Coble,M.D., Just,R.S., O'Callaghan,J.E., Letmanyi,I.H., Peterson,C.T., Irwin,J.A. and Parsons,T.J. Single nucleotide polymorphisms over the entire mtDNA genome that increase the power of forensic testing in Caucasians. Int. J. Legal Med. 118 (3), 137-146 (2004).

Roostalu, U., Kutuev,I., Loogvali,E.L., Metspalu,E., Tambets,K., Reidla,M., Khusnutdinova,E., Usanga,E., Kivisild,T. and Villems,R. Origin and expansion of haplogroup H, the dominant human mitochondrial DNA lineage in west eurasia: the near eastern and Caucasian perspective. Mol. Biol. Evol. 24 (2), 436-448 (2007).

Olivieri, A., Achilli, A., Pala, M., Battaglia, V., Fornarino, S., Al-Zahery, N., Scozzari,R., Cruciani,F., Behar,D.M., Dugoujon, J.M., Coudray,C., Santachiara-Benerecetti, A.S., Semino, O., Bandelt, H.J. and Torroni, A. The mtDNA Legacy of the Levantine Early Upper Palaeolithic in Africa. Science 314 (5806), 1767-1770 (2006)

Hill, C., Soares, P., Mormina, M., Macaulay, V., Meehan, W., Blackburn, J., Clarke,D., Raja,J.M., Ismail,P., Bulbeck,D.,Oppenheimer,S. and Richards,M. Phylogeography and Ethnogenesis of Aboriginal Southeast Asians (er) Mol. Biol. Evol. 23 (12), 2480-2491 (2006)

Hinttala, R., Smeets, R., Moilanen, J.S., Ugalde, C., Uusimaa, J., Smeitink, J.A.M. and Majamaa, K. Analysis of mitochondrial DNA sequences in children with isolated or combined oxidative phosphorylation system deficiency. J. Med. Genet. 43 (11), 881-886 (2006)

Kivisild, T., Shen, P., Wall, D.P., Do, B., Sung, R., Davis, K.K., Passarino, G., Underhill, P.A., Scharfe, C., Torroni, A., Scozzari, R., Modiano, D., Coppa, A., de Knjiff, P., Feldman, M.W., Cavalli-Sforza, L.L. and Oefner,P.J. The role of selection in the evolution of human mitochondrial genomes. Genetics 172 (1), 373-387 (2006)

Kong,Q.-P., Yao,Y.-G., Sun,C., Bandelt,H.-J., Zhu,C.-L. and Zhang,Y.-P.
Phylogeny of east Asian mitochondrial DNA lineages inferred from complete sequences. Am. J. Hum. Genet. 73 (3), 671-676 (2003)

Maca-Meyer, N., Gonzalez, A.M., Pestano, J., Flores, C., Larruga, J.M. and Cabrera, V.M.
Mitochondrial DNA transit between West Asia and North Africa inferred from U6 phylogeography BMC Genet. 4 (1), 15 (2003)

Ingman, M. and Gyllensten, U. Mitochondrial genome variation and evolutionary history of Australian and new guinean aborigines Genome Res. 13 (7), 1600-1606 (2003)

Fraumene, C., Belle, E.M., Castri, L., Sanna, S., Mancosu, G., Cosso, M., Marras, F., Barbujani, G., Pirastu, M. and Angius, A. High Resolution Analysis and Phylogenetic Network Construction Using Complete mtDNA Sequences in Sardinian Genetic Isolates Mol. Biol. Evol. 23 (11), 2101-2111 (2006)

van Holst Pellekaan, S.M., Ingman, M., Roberts-Thomson, J. and Harding, R.M.
Mitochondrial genomics identifies major haplogroups in Aboriginal Australians Am. J. Phys. Anthropol. 131 (2), 282-294 (2006)

Pierson, M.J., Martinez-Arias, R., Holland, B.R., Gemmell, N.J., Hurles, M.E. and Penny, D.
Deciphering Past Human Population Movements in Oceania: Provably Optimal Trees of 127 mtDNA GenomesMol. Biol. Evol. 23 (10), 1966-1975 (2006)

Ingman, M. and Gyllensten, U. A Recent Genetic Link between Sami and the Volga-Ural Region of Russia. Eur. J. Hum. Genet. (2006) In press

Annunen-Rasila, J., Finnila, S., Mykkanen, K., Moilanen, J.S.,Veijola, J., Poyhonen, M., Viitanen, M., Kalimo, H. and Majamaa, K. Mitochondrial DNA sequence variation and mutation rate in patients with CADASIL Neurogenetics 7 (3), 185-194 (2006)

Moilanen, J.S., Finnila,S. and Majamaa,K. Lineage-Specific Selection in Human mtDNA: Lack of Polymorphisms in a Segment of MTND5 Gene in Haplogroup J Mol. Biol. Evol. 20 (12), 2132-2142 (2003)

Finnila,S., Lehtonen, M.S. and Majamaa, K. Phylogenetic network for European mtDNA Am. J. Hum. Genet. 68 (6), 1475-1484 (2001)

Phan,V.C., Nong,V.H., Nguyen,B.N., Tran,T.M.N., Le,T.B.T., Do,Q.H. Nguyen,N.L., Bui,T.H., Pham,D.M., Tran,T.T., Tong,Q.M.,Nguyen,T.T., Nguyen,D.T., Le,T.T.H., Nguyen,D.C., Le,Q.H., Dang,D.H., Quyen,D.T., Van,D.H., Trinh,V.B., Le,B.Q. and Nguyen,D.B. Submitted (30-JUN-2006) Protein Biochemistry, Institute of Biotechnology (IBT), Vietnamese Academy of Science and Technology (VAST), 18 Hoang Quoc Viet, Hanoi, Hanoi 10000, Vietnam

Thangaraj,K., Chaubey,G., Singh,V.K., Vanniarajan,A., Thanseem,I., Reddy,A.G. and Singh,L. In situ origin of deep rooting lineages of mitochondrial Macrohaplogroup M in India (er) BMC Genomics 7 (1), 151 (2006)

Torroni,A., Achilli,A., Macaulay,V., Richards,M. and Bandelt,H.J. Harvesting the fruit of the human mtDNA tree. Trends Genet. 22 (6), 339-345 (2006)

Gonzalez,A.M., Garcia,O., Larruga,J.M. and Cabrera,V.M.
The mitochondrial lineage U8a reveals a Paleolithic settlement in the Basque country (er) BMC Genomics 7 (1), 124 (2006)

Thangaraj,K., Chaubey,G., Kivisild,T., Reddy,A.G., Singh,V.K., Rasalkar,A.A. and Singh,L. Reconstructing the origin of Andaman Islanders. Science 308 (5724), 996 (2005)

Behar,D.M., Metspalu,E., Kivisild,T., Achilli,A., Hadid,Y.,
Tzur,S., Pereira,L., Amorim,A., Quintana-Murci,L., Majamaa,K., Herrnstadt,C., Howell,N., Balanovsky,O., Kutuev,I., Pshenichnov,A., Gurwitz,D., Bonne-Tamir,B., Torroni,A., Villems,R. and Skorecki,K. The matrilineal ancestry of ashkenazi jewry: portrait of a recent founder event. Am. J. Hum. Genet. 78 (3), 487-497 (2006)

Sun, C., Kong, Q.P., Palanichamy, M.G., Agrawal,S., Bandelt, H.J., Yao, Y.G., Khan, F., Zhu, C.L., Chaudhuri, T.K. and Zhang, Y.P. The Dazzling Array of Basal Branches in the mtDNA Macrohaplogroup M from India as Inferred from Complete Genomes Mol. Biol. Evol. 23 (3), 683-690 (2006)

Rajkumar,R., Banerjee,J., Gunturi,H.B., Trivedi,R. and Kashyap,V.K. Phylogeny and antiquity of M macrohaplogroup inferred from complete mtDNA sequence of Indian specific lineages(er) BMC Evol. Biol. 5 (1), 26 (2005)

Macaulay,V., Hill,C., Achilli,A., Rengo,C., Clarke,D., Meehan,W., Blackburn,J., Semino,O., Scozzari,R., Cruciani,F., Taha,A., Shaari,N.K., Raja,J.M., Ismail,P., Zainuddin,Z., Goodwin,W., Bulbeck,D., Bandelt,H.J., Oppenheimer,S., Torroni,A. and Richards,M. Single, rapid coastal settlement of Asia revealed by analysis of complete mitochondrial genomes. Science 308 (5724), 1034-1036 (2005)

Merriwether,D.A., Hodgson,J.A., Friedlaender,F.R., Allaby,R., Cerchio,S., Koki,G. and Friedlaender,J.S. Ancient mitochondrial M haplogroups identified in the Southwest Pacific. Proc. Natl. Acad. Sci. U.S.A. 102 (37), 13034-13039 (2005)

Trejaut,J.A., Kivisild,T., Loo,J.H., Lee,C.L., He,C.L., Hsu,C.J., Lee,Z.Y. and Lin,M. Traces of archaic mitochondrial lineages persist in Austronesian-speaking Formosan populations. PLoS Biol. 3 (8), E247 (2005)

Bandelt,H.J., Achilli,A., Kong, Q.P., Salas,A., Lutz-Bonengel,S., Sun,C., Zhang,Y.P., Torroni,A. and Yao,Y.G. Low 'penetrance' of phylogenetic knowledge in mitochondrial disease studies. Biochem. Biophys. Res. Commun. 333 (1), 122-130 (2005)

Starikovskaya,E.B., Sukernik,R.I., Derbeneva,O.A., Volodko,N.V., Ruiz-Pesini,E., Torroni,A., Brown,M.D., Lott,M.T., Hosseini,S.H., Huoponen,K. and Wallace,D.C.
Mitochondrial DNA diversity in indigenous populations of the southern extent of Siberia, and the origins of Native American Haplogroups. Ann. Hum. Genet. 69 (PT 1), 67-89 (2005)

Marzuki,S., Noer,A.S., Lertrit,P., Thyagarajan,D., Kapsa,R., Utthanaphol,P. and Byrne,E.

Normal variants of human mitochondrial DNA and translation products: the building of a reference database. Hum. Genet. 88 (2), 139-145 (1991)

Achilli,A., Rengo,C., Battaglia,V., Pala,M., Olivieri,A., Fornarino,S., Magri,C., Scozzari,R., Babudri,N., Santachiara-Benerecetti,A.S., Bandelt,H.J., Semino,O. and Torroni,A. Saami and berbers--an unexpected mitochondrial DNA link. Am. J. Hum. Genet. 76 (5), 883-886 (2005)

Achilli,A., Rengo,C., Magri,C., Battaglia,V., Olivieri,A., Scozzari,R., Cruciani,F., Zeviani,M., Briem,E., Carelli,V., Moral,P., Dugoujon,J.M., Roostalu,U., Loogvali,E.L.,Kivisild,T., Bandelt,H.J., Richards,M., Villems,R., Santachiara-Benerecetti,A.S., Semino,O. and Torroni,A. The molecular dissection of mtDNA haplogroup H confirms that the Franco-Cantabrian glacial refuge was a major source for the European gene pool. Am. J. Hum. Genet. 75 (5), 910-918 (2004)

Mishmar,D., Ruiz-Pesini,E., Golik,P., Macaulay,V., Clark,A.G., Hosseini,S., Brandon,M., Easley,K., Chen,E., Brown,M.D., Sukernik,R.I., Olckers,A. and Wallace,D.C. Natural selection shaped regional mtDNA variation in humans. Proc. Natl. Acad. Sci. U.S.A. 100 (1), 171-176 (2003)

Ingman,M., Kaessmann,H., Paabo,S. and Gyllensten,U. Mitochondrial genome variation and the origin of modern humans Nature 408 (6813), 708-713 (2000).

**New Developments in Cancer Diagnostics and Human Mitochondrial DNA Variation**

He, Yipying, Jian Wu, Devin C. Dressman, Christine Iacobuzio-Donahue, Sanford D. Markowitz, Victor E. Velculescu, Luis A. Diaz Jr, Kenneth W. Kinzler, Bert Vogelstein & Nickolas Papadopoulos (2010) Heteroplasmic mitochondrial DNA mutations in normal and tumour cells. *Nature* 464, 610-614 (25 March 2010) doi:10.1038/nature08802 http://www.nature.com/nature/journal/v464/n7288/pdf/nature08802

Legros, F., Malka, F., Frachon, P., Lombes, A. Rojo, M.Organization and dynamics of human mitochondrial DNA. *J. Cell Sci.* (2004) vol. 117, pp. 2653-2662

White, H. E. Accurate detection and quantitation of heteroplasmic mitochondrial point mutations by pyrosequencing. *Genet. Test.* (2005) vol. 9, pp. 190-199 10.1089/gte.2005.9.190 http://dx.doi.org/10.1089/gte.2005.9.190

Santos, C. Frequency and pattern of heteroplasmy in the control region of human mitochondrial DNA. *J. Mol. Evol.* (2008) vol. 67, pp. 191 – 200. 10.1007/s00239-008-9138-9 http://dx.doi.org/10.1007/s00239-008-9138-9

Greaves, L. C., Quantification of mitochondrial DNA mutation load. *Aging Cell* (2009) vol. 8, pp. 566 -572 10.1111/j.1474-9726.2009.00505.x http://dx.doi.org/10.1111/j.1474-9726.2009.00505.x

Michikawa, Y., Mazzucchelli, F., Bresolin, N., Scarlato, G.
Attardi, G. Aging-dependent large accumulation of point mutations in the human mtDNA control region for replication. *Science* (1999) vol. 286, pp. 774-779 10.1126/science.286.5440.774 http://dx.doi.org/10.1126/science.286.5440.774

Zsurka, G. Recombination of mitochondrial DNA in skeletal muscle of individuals with multiple mitochondrial DNA heteroplasmy. *Nature Genet.* (2005) vol. 37, pp. 873-877. 10.1038/ng1606 http://dx.doi.org/10.1038/ng1606

Coller, H. A. High frequency of homoplasmic mitochondrial DNA mutations in human tumors can be explained without selection. *Nature Genet.* (2001) vol. 28, pp. 147-150

Sekiguchi, K., Kasai, K., Levin, B. C. Inter-and intragenerational transmission of a human mitochondrial DNA heteroplasmy among 13 maternally-related individuals and differences between and within tissues in two family members. *Mitochondrion* (2003) vol. 2, pp. 401-414.
10.1016/S1567-7249(03)00028-X http://dx.doi.org/10.1016/S1567-7249(03)00028-X

Sekiguchi, K., Sato, H., Kasai, K. Mitochondrial DNA heteroplasmy among hairs from single individuals. *J. Forensic Sci*. (2004) vol. 49, pp. 986-991
10.1520/JFS2003216
http://dx.doi.org/10.1520/JFS2003216

**NumtS Identification Study (As Numbered in Text)**

1. Zischler H, Geisert H, von Haeseler A, *et. al.* A nuclear "fossil" of the mitochondrial D-loop and the origin of modern humans. Nature 1995;378:489-492.

2. Lang M, Sazzini M, Calabrese FM, Simone D, Boattini A, Romeo G, Luiselli D, Attimonelli M, Gasparre, G. Polymorphic NumtS trace Human Population Relationships. Human Genetics 2012;131:757-771.

3. Thomas T, Zischler H, Pääbo S, Stoneking, M. Novel mitochondrial DNA insertion polymorphism and its usefulness for human population studies. 1996:

4. Metzker ML. Sequencing technologies – the next generation. Nature Reviews Genetics 2010;11:31-46.

5. Mardis ER. The impact of next generation sequencing on genetics. Trends in Genetics 2008;24:133-141.

6. Huse SM, Huber JA, Morrison HG, Sogin ML, Mark Welch D. Accuracy and quality of massively parallel DNA pyrosequencing. Genome Biology 2007;8(7).

7. Gilles A, Meglécz E, Pech N, Ferreira S, Malausa T, Martin JF. Accuracy and Quality Assessment of 454 GS-FLX Titanium Pyrosequencing. BMC Genomics 2011;12:245-256.

8. Budowle B, Wilson MR, DiZinno JA, Stauffer C, Fasano MA, Holland MM, *et. al.* Mitochondrial DNA regions HVI and HVII population data. Forensic Sci Int 1999;103:23–35.

9. Budowle B, Allard MW, Wilson MR, Chakraborty R. Forensics and mitochondrial DNA: applications, debates, and foundations. Annu Rev Genomics Hum Genet 2003;4:119–41.

10. Carracedo A, Bär W, Lincoln P, Mayr W, Morling N, Olaisen B, *et. al.* DNA commission of the international society for forensic genetics: guidelines for mitochondrial DNA typing. Forensic Sci Int 2000;110:79–85.

11. Wilson MR, DiZinno JA, Polanskey D, Replogle J, Budowle B. Validation of mitochondrial DNA sequencing for forensic casework analysis. Int J Legal Med 1995;108:68–74.

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

13. Wilson MR, DiZinno JA, Polanskey D, Budowle, B. Assessing Heteroplasmy in the Control Region of Human Mitochondrial DNA. Proceedings of the 50th Anniversary Meeting of the American Academy of Forensic Sciences, San Francisco, California, 1998, p.62, printed by McCormick-Armstrong, Colorado Springs, CO.

14. Wooley JC, Ye Y. Metagenomics: Facts and artifacts and computational challenges. J Comput Sci Technol 2009;25:71-81

15. Salas A, Lareu MV, Carracedo A. Heteroplasmy in mtDNA and the weight of evidence in forensic mtDNA analysis: a case report. Int J Leg Med 2001;114:186-190.

16. http://www.kirkhousetrust.org/Docs/FTAProtocolBD08.pdf

17. http://media.affymetrix.com:80/support/technical/usb/brief_proto/78200B.pdf

18. http://tools.invitrogen.com/content/sfs/manuals/cms_041330.pdf

19. Kavlick MF, Lawrence HS, Merritt T, Fisher C, Isenberg A, Robertson JM, Budowle, B. Quantification of Human Mitochondrial DNA Using Synthesized DNA Standards. J Forensic Sci 2011;56:1457-1463.

20. Anderson S, Bankier AT, Barrell BG, de Bruijn MH, Coulson AR, Drouin J, Nierlich DP, Roe BA, Sanger F, Schreier PH, Smith AJ, Staden R, Young IG. Sequence and organization of the human mitochondrial genome. Nature 1981;290:457-465.

21. http://www3.appliedbiosystems.com/cms/groups/applied_markets_support/documents/generaldocuments/cms_041395.pdf

22. www.qiagen.com/literature/render.aspx?id=201083

23. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Bio 1990;215:403-410.

24. Stoneking, M. Hypervariable sites in the mtDNA Control region are mutational hotspots. Cell Press 2000;67:1029-1032.

25. Parr RL, Maki J, Reguly B, Dakubo GC, Aguirre A, Wittock R, Robinson K, Jakupciak JP, Thayer RE. The pseudo-mitochondrial genome influences mistakes in heteroplamsy interpretation. BMC Genomics 2006;7:185-198

**Quality Issues in Next Generation Sequencing Applications**

Alkan, C., B. P. Coe, *et. al.* (2011). "Genome structural variation discovery and genotyping." Nat Rev Genet 12(5): 363-376.

Alkan, C., J. M. Kidd, *et. al.* (2009). "Personalized copy number and segmental duplication maps using next-generation sequencing." Nat Genet 41(10): 1061-1067.

Alkan, C., S. Sajjadian, *et. al.* (2011). "Limitations of next-generation genome sequence assembly." Nat Methods 8(1): 61-65.

Allard, M. W., Y. Luo, *et. al.* (2012). "High resolution clustering of Salmonella enterica serovar Montevideo strains using a Next-generation Sequencing approach." BMC Genomics 13: 32.

Antonacci, F., J. M. Kidd, *et. al.* (2010). "A large and complex structural polymorphism at 16p12.1 underlies microdeletion disease risk." Nat Genet 42(9): 745-750.

Bailey, J. A., Z. Gu, *et. al.* (2002). "Recent segmental duplications in the human genome." Science 297(5583): 1003-1007.

Bakker, H. C., A. I. Switt, *et. al.* (2011). "A whole-genome single nucleotide polymorphism-based approach to trace and identify outbreaks linked to a common Salmonella enterica subsp. enterica serovar Montevideo pulsed-field gel electrophoresis type." Appl Environ Microbiol 77(24): 8648-8655.

Bashir, A., A. A. Klammer, *et. al.* (2012). "A hybrid approach for the automated finishing of bacterial genomes." Nat Biotechnol.

Blankenberg, D., A. Gordon, *et. al.* (2010). "Manipulation of FASTQ data with Galaxy." Bioinformatics 26(14): 1783-1785.

Carneiro, M. O., C. Russ, *et. al.* (2012). "Pacific biosciences sequencing technology for genotyping and variation discovery in human data." BMC Genomics 13: 375.

Chen, K., J. W. Wallis, *et. al.* (2009). "BreakDancer: an algorithm for high-resolution mapping of genomic structural variation." Nat Methods 6(9): 677-681.

Choi, M., U. I. Scholl, *et. al.* (2009). "Genetic diagnosis by whole exome capture and massively parallel DNA sequencing." Proc Natl Acad Sci U S A 106(45): 19096-19101.

Cock, P. J., C. J. Fields, *et. al.* (2010). "The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants." Nucleic Acids Res 38(6): 1767-1771.

Cox, M. P., D. A. Peterson, *et. al.* (2010). "SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data." BMC Bioinformatics 11: 485.

Currie, B. J., A. Haslem, *et. al.* (2009). "Identification of melioidosis outbreak by multilocus variable number tandem repeat analysis." Emerg Infect Dis 15(2): 169-174.

MLVA-4 can establish or refute that a clonal outbreak of melioidosis has occurred within 8 hours of receipt of bacterial strains.

Dale, J., E. P. Price, *et. al.* (2011). "Epidemiological tracking and population assignment of the non-clonal bacterium, Burkholderia pseudomallei." PLoS Negl Trop Dis 5(12): e1381.

Dalloul, R. A., J. A. Long, *et. al.* (2010). "Multi-platform next-generation sequencing of the domestic turkey (Meleagris gallopavo): genome assembly and analysis." PLoS Biol 8(9).

Davis, M. A., D. D. Hancock, *et. al.* (2003). "Evaluation of pulsed-field gel electrophoresis as a tool for determining the degree of genetic relatedness between strains of Escherichia coli O157:H7." J Clin Microbiol 41(5): 1843-1849.

Dean FB, Hosono S, Fang L, Wu X, Faruqi AF, Bray-Ward P, Sun Z, Zong Q, Du Y, Du J, Driscoll M, Song W, Kingsmore SF, Egholm M, Lasken RS. (2002). Comprehensive human genome amplification using multiple displacement amplification. *Proc. Natl. Acad. Sci. USA.* Apr 16;99(8):5261-6.

DePristo, M. A., E. Banks, *et. al.* (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 43(5): 491-498.

Eckert K.A., Kunkel T.A. (1991) DNA polymerase fidelity and the polymerase chain reaction. *PCR Methods Appl*. Aug;1(1):17-24.

Escarmis, C., E. Lazaro, *et. al.* (2006). "Population bottlenecks in quasispecies dynamics." Curr. Topics Microbiol. Immunol. 299: 141-170.

Esteban J.A, Salas M., Blanco L. (1993) Fidelity of phi 29 DNA polymerase: Comparison between protein-primed initiation and DNA polymerization. *J. Biol. Chem.* Feb 5;268(4):2719-26.

Ewing, B. and P. Green (1998). "Base-calling of automated sequencer traces using phred. II. Error probabilities." Genome Res 8(3): 186-194.

Feldmeyer, B., C. W. Wheat, *et. al.* (2011). "Short read Illumina data for the de novo assembly of a non-model snail species transcriptome (Radix balthica, Basommatophora, Pulmonata), and a comparison of assembler performance." BMC Genomics 12: 317.

Flusberg, B. A., D. R. Webster, *et. al.* (2010). "Direct detection of DNA methylation during single-molecule, real-time sequencing." Nat Methods 7(6): 461-465.

Fricke, W. F., M. K. Mammel, *et. al.* (2011). "Comparative genomics of 28 Salmonella enterica isolates: evidence for CRISPR-mediated adaptive sublineage evolution." J Bacteriol 193(14): 3556-3568.

Gilles, A., E. Meglecz, *et. al.* (2011). "Accuracy and quality assessment of 454 GS-FLX Titanium pyrosequencing." BMC Genomics 12: 245.

Glenn, T. C. (2011). "Field guide to next-generation DNA sequencers." Mol Ecol Resour 11(5): 759-769.

Gnerre, S., I. Maccallum, *et. al.* (2011). "High-quality draft assemblies of mammalian genomes from massively parallel sequence data." Proc Natl Acad Sci U S A 108(4): 1513-1518.

Handsaker, R. E., J. M. Korn, *et. al.* (2011). "Discovery and genotyping of genome structural polymorphism by sequencing on a population scale." Nat Genet 43(3): 269-276.

Hasan, N. A., S. Y. Choi, *et. al.* (2012). "Genomic diversity of 2010 Haitian cholera outbreak strains." Proc Natl Acad Sci U S A 109(29): E2010-2017.

Hormozdiari, F., C. Alkan, *et. al.* (2009). "Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes." Genome Res 19(7): 1270-1278.

Hormozdiari, F., I. Hajirasouliha, *et. al.* (2010). "Next-generation VariationHunter: combinatorial algorithms for transposon insertion discovery." Bioinformatics 26(12): i350-357.

Jacobsen, A., R. S. Hendriksen, *et. al.* (2011). "The Salmonella enterica pan-genome." Microb Ecol 62(3): 487-504.

Johnson, D. S., A. Mortazavi, *et. al.* (2007). "Genome-wide mapping of in vivo protein-DNA interactions." Science 316(5830): 1497-1502.

Keegan, K. P., W. L. Trimble, *et. al.* (2012). "A platform-independent method for detecting errors in metagenomic sequencing data: DRISEE." PLoS Comput Biol 8(6): e1002541.

Kidd, J. M., G. M. Cooper, *et. al.* (2008). "Mapping and sequencing of structural variation from eight human genomes." Nature 453(7191): 56-64.

Kidd, J. M., N. Sampas, *et. al.* (2010). "Characterization of missing human genome sequences and copy-number polymorphic insertions." Nat Methods 7(5): 365-371.

Korbel, J. O., A. E. Urban, *et. al.* (2007). "Paired-end mapping reveals extensive structural variation in the human genome." Science 318(5849): 420-426.

Koren, S., M. C. Schatz, *et. al.* (2012). "Hybrid error correction and de novo assembly of single-molecule sequencing reads." Nat Biotechnol.

Kunin, V., A. Engelbrektson, *et. al.* (2010). "Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates." Environ Microbiol 12(1): 118-123.

Lee, S., F. Hormozdiari, *et. al.* (2009). "MoDIL: detecting small indels from clone-end sequencing with mixtures of distributions." Nat Methods 6(7): 473-474.

Li, L., J. G. Victoria, *et. al.* (2010). "Bat guano virome: predominance of dietary viruses from insects and plants plus novel mammalian viruses." J Virol 84(14): 6955-6965.

Lienau, E. K., E. Strain, *et. al.* (2011). "Identification of a salmonellosis outbreak by means of molecular sequencing." N Engl J Med 364(10): 981-982.

Loman, N. J., R. V. Misra, *et. al.* (2012). "Performance comparison of benchtop high-throughput sequencing platforms." Nat Biotechnol 30(5): 434-439.

Mardis, E. R. (2008). "Next-generation DNA sequencing methods." Annu Rev Genomics Hum Genet 9: 387-402.

Margulies, M., M. Egholm, *et. al.* (2005). "Genome sequencing in microfabricated high-density picolitre reactors." Nature 437(7057): 376-380.

Martinez-Alcantara, A., E. Ballesteros, *et. al.* (2009). "PIQA: pipeline for Illumina G1 genome analyzer data quality assessment." Bioinformatics 25(18): 2438-2439.

Medvedev, P., M. Fiume, *et. al.* (2010). "Detecting copy number variation with mated short reads." Genome Res 20(11): 1613-1622.

Mills, R. E., C. T. Luttig, *et. al.* (2006). "An initial map of insertion and deletion (INDEL) variation in the human genome." Genome Res 16(9): 1182-1190.

Mills, R. E., K. Walter, *et. al.* (2011). "Mapping copy number variation by population-scale genome sequencing." Nature 470(7332): 59-65.

Novitsky, V., R. Wang, *et. al.* (2011). "Transmission of single and multiple viral variants in primary HIV-1 subtype C infection." PLoS One 6(2): e16714.

Paszkiewicz, K. and D. J. Studholme (2010). "De novo assembly of short sequence reads." Brief Bioinform 11(5): 457-472.

Prosperi, M. C., L. Prosperi, *et. al.* (2011). "Combinatorial analysis and algorithms for quasispecies reconstruction using next-generation sequencing." BMC Bioinformatics 12: 5.

Quail, M. A., M. Smith, *et. al.* (2012). "A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers." BMC Genomics 13: 341.

Rokas, A. and P. Abbot (2009). "Harnessing genomics for evolutionary insights." Trends Ecol Evol 24(4): 192-200.

Rozera, G., I. Abbate, *et. al.* (2009). "Massively parallel pyrosequencing highlights minority variants in the HIV-1 env quasispecies deriving from lymphomonocyte sub-populations." Retrovirology 6: 15.

Rozera, G., I. Abbate, *et. al.* (2012). "Ultra-deep sequencing reveals hidden HIV-1 minority lineages and shifts of viral population between the main cellular reservoirs of the infection after therapy interruption." J Med Virol 84(6): 839-844.

Schmieder, R. and R. Edwards (2011). "Quality control and preprocessing of metagenomic datasets." Bioinformatics 27(6): 863-864.

Schmieder, R., Y. W. Lim, *et. al.* (2010). "TagCleaner: Identification and removal of tag sequences from genomic and metagenomic datasets." BMC Bioinformatics 11: 341.
interface facilitates export functionality for subsequent data processing, and is available at http://edwards.sdsu.edu/tagcleaner.

Schwartz, D. C., X. Li, *et. al.* (1993). "Ordered restriction maps of Saccharomyces cerevisiae chromosomes constructed by optical mapping." Science 262(5130): 110-114.

Shukla, S. K., M. Pantrangi, *et. al.* (2012). "Comparative Whole Genome Mapping to Determine Staphylococcus aureus Genome size, Virulence Motifs and Clonality." J Clin Microbiol.

Teague, B., M. S. Waterman, *et. al.* (2010). "High-resolution human genome structure by single-molecule analysis." Proc Natl Acad Sci U S A 107(24): 10848-10853.

Tuzun, E., A. J. Sharp, *et. al.* (2005). "Fine-scale structural variation of the human genome." Nat Genet 37(7): 727-732.

Wang, Z., M. Gerstein, *et. al.* (2009). "RNA-Seq: a revolutionary tool for transcriptomics." Nat Rev Genet 10(1): 57-63.

Xi, M., J. Zheng, *et. al.* (2008). "An enhanced discriminatory pulsed-field gel electrophoresis scheme for subtyping Salmonella serotypes Heidelberg, Kentucky, SaintPaul, and Hadar." J Food Prot 71(10): 2067-2072.

Ye, K., M. H. Schulz, *et. al.* (2009). "Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads." Bioinformatics 25(21): 2865-2871.

Yoon, S., Z. Xuan, *et. al.* (2009). "Sensitive and accurate detection of copy number variants using read depth of coverage." Genome Res 19(9): 1586-1592.

Zheng, J., C. E. Keys, *et. al.* (2011). "Simultaneous analysis of multiple enzymes increases accuracy of pulsed-field gel electrophoresis in assigning genetic relationships among homogeneous Salmonella strains." J Clin Microbiol 49(1): 85-94.

Zheng, J., C. E. Keys, *et. al.* (2007). "Enhanced subtyping scheme for Salmonella enteritidis." Emerg Infect Dis 13(12): 1932-1935.

**Dissemination of Research Findings**

**1. Citation for each publication that resulted from this funded grant project.**

Bintz, B., Dixon, G., Wilson, M.R. 2012. *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 (submitted 9/2012).

**2. Citations for each presentation that resulted from this funded project.**
**** presenting author***

Brittania J. Bintz*, Groves B. Dixon, Mark R. Wilson. 2012. *Detection of low-level artifacts associated with homopolymeric regions in human mitochondrial DNA (mtDNA) using massively parallel 454 pyrosequencing.* Poster presentation at the National Institute of Justice Conference, 2012, Arlington, Virginia

Brittania J. Bintz*, Brandon C. Smith, Groves B. Dixon, Mark R. Wilson 2012. *Development of a Novel Human Mitochondrial DNA (mtDNA) Amplification Method for use with Illumina Next-Generation Sequencing Instrumentation*. Molecules in the Mountains Conference, 2012, Cullowhee, NC

Brittania J. Bintz*, Brandon C. Smith, Hilde Stawski, Erin S. Burnside, Mark R. Wilson. 2012. *Development of a Novel Human Mitochondrial DNA (mtDNA) Amplification Method for use with Illumina Next-Generation Sequencing Instrumentation*. American Academy of Forensic Sciences (AAFS) Conference, 2012, Atlanta, GA

Brittania J. Bintz*, Brandon C. Smith, Mark R. Wilson. 2011. *Next-Generation Sequencing of Human Mitochondrial DNA Amplicons – Technical and Interpretational Issues.* Illumina, Inc. NGS Workshop, November, 2011, Cullowhee, NC

Mark R. Wilson. 2011. *Next-Generation Sequencing of Human Mitochondrial DNA Amplicons – Technical and Interpretational Issues.* Illumina, Inc. NGS Workshop, November, 2011, Cullowhee, NC

Brittania J. Bintz*, Brandon C. Smith, Groves B. Dixon, Erin S. Burnside, Mark R. Wilson. 2012. *Next-Generation Sequencing of Human Mitochondrial DNA Amplicons – Technical and Interpretational Issues.* Poster presentation at the International Symposium on Human Identification (ISHI) 22, 2011, National

Harbor, MD

Brittania J. Bintz*, Brandon C. Smith, Mark R. Wilson. 2011. *Development of a Novel Human Mitochondrial DNA (mtDNA) Amplification Method for use with Illumina Next-Generation Sequencing Instrumentation*. Poster presentation at the National Institute of Justice Conference, 2011, Arlington, VA

Groves B. Dixon*, Brittania J. Bintz, Mark R. Wilson. 2012. *A single Library Preparation Method for Both Sanger Sequencing and Quantification of Minor Variants by Pyrosequencing.* American Academy of Forensic Sciences (AAFS) Conference, 2012, Atlanta, GA

Brandon Smith*, Brittania J. Bintz, Erin S. Burnside, Mark R. Wilson. 2012. *Low-Level Variant Detection in mitochondrial DNA using the Illumina® GA IIx Next-Generation Sequencing (NGS) Platform.* Poster presentation at the American Academy of Forensic Sciences (AAFS) Conference, 2012, Atlanta, GA

Hilde Stawski*, Brittania J. Bintz, Erin S. Burnside, Mark R. Wilson. 2012. *Preparing massively parallel sequencing libraries of human mitochondrial DNA using Illumina® Nextera® technology.* Poster presentation at the American Academy of Forensic Sciences (AAFS) Conference, 2012, Atlanta, GA

Groves B. Dixon*, Brittania J. Bintz, Mark R. Wilson. 2012. *A Single Library Preparation Method for Both Sanger Sequencing and Quantification of Minor Variants by Pyrosequencing.* Poster presentation at the International Symposium on Human Identification (ISHI) 22, 2011, National Harbor, MD

Groves B. Dixon. 2012. *An Evaluation of the Roche GS Junior™ Pyrosequencer to Detect Minor Variants in Mixed Mitochondrial DNA Amplicons.* UNC Charlotte Workshop in Next-Generation Sequencing, May, 2012, Charlotte, NC

Mark R. Wilson. 2012. *Detecting Minor Variants in Mixed Mitochondrial DNA Amplicons.* UNC Charlotte Workshop in Next-Generation Sequencing, May, 2012, Charlotte, NC

Mark R. Wilson. 2012. *Assessing Deep DNA Sequencing Technologies for Human Forensic mtDNA Analysis, Green Mountain DNA Conference, Burlington, Vt., Aug. 1, 2012.*

Further information on presentations can be found here:

http://thereporter.wcu.edu/2012/11/forensic-science-program-faculy-staff-and-students-present-at-regional-national-and-international-conferences/