Document Title:	**Assessing Sample Preparation Methods for Emerging DNA Sequencing Technologies in Human Forensic mtDNA Analysis Applications**
Author(s):	**Brittania J. Bintz, M.S.**
Document Number:	**252006**
Date Received:	**November 2018**
Award Number:	**2013-DN-BX-K014**

The author(s) shown below used Federal funds provided by the U.S. Department of Justice and prepared the following report:

Document Title: **Assessing Sample Preparation Methods for Emerging DNA Sequencing Technologies in Human Forensic mtDNA Analysis Applications**

Author(s): **Brittania J. Bintz, M.S.**

Submission Date: **October 5th, 2016**

DUNS and EIN Numbers:

Recipient Organization (Name and Address):

Western Carolina University
140 Robinson Administration Building
Cullowhee, NC 28723

Recipient Identifying Number or Account Number:

Project/Grant Period (Start Date, End Date): 01/03/2014, 6/30/2017

Report Term or Frequency: Draft Final

Signature of Submitting Official:

Brittania J. Bintz, M.S.
Forensic Research Scientist
Western Carolina University
111 Memorial Drive, NSB #231
Cullowhee, NC 28723

DRAFT Final Technical Report

Assessing Sample Preparation Methods for Emerging DNA Sequencing Technologies in Human Forensic mtDNA Analysis Applications
Award #2013-DN-BX-K014
Brittania J. Bintz, M.S., Primary Investigator
Western Carolina University, Cullowhee, N.C. 28723
October 5th, 2016

Abstract

Human mitochondrial DNA 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) will reveal subtle mixtures that are currently not detected by forensic DNA laboratories. Hence, new DNA sequencing technologies have the promise of providing information in both of these dimensions and thereby expanding the utility of mtDNA analysis in forensic science.

The ultimate goal of our research effort is to continue to develop methods that enable generation of whole mt-genome DNA sequence information from compromised or limited DNA samples, thus greatly expanding the potential utility of this marker system. We have focused primarily on human hair shafts as a model for these challenging samples. However, we have also expanded our efforts to entomological samples, and dust bunnies, and calcified tissues including human cremated remains.

In order to accomplish this goal, we have developed enhanced DNA extraction techniques for hair shaft and calcified tissue samples. Additionally, in this effort we evaluated several enrichment strategies designed to increase the amount of mtDNA template sufficient for massively-parallel sequencing on the Illumina® MiSeq. These methods included whole genome amplification, probe capture enrichment using both RNA and DNA baits, and multiplexed PCR amplification. We found that the combination of the enhanced DNA extraction technique and multiplexed PCR amplification reactions around the mtGenome resulted in high-quality sequence information from highly compromised samples. Further developmental research and validation, based on our approach and data, will result in a significant enhancement over current forensic DNA typing procedures.

In a parallel study, we analyzed massively-parallel sequence data to determine a minimum frequency threshold above which differences from the rCRS would be considered true biological variation and not noise. The data was generated using synthetic oligonucleotides designed to contain sequences that match stretches of the human mtDNA hypervariable regions. These oligonucleotides were also designed so that they could be sequenced directly, without any additional preparation. Subsets of the oligonucleotides were also prepared for sequencing using the Nextera XT general workflow. Data generated from direct sequencing was compared to data generated from oligonucleotides that were prepared for sequencing to determine whether discrete steps in library preparation increased the amount of low-level noise or error in the data set. We

determined that a conservative frequency threshold of 5% in HV1 data would eliminate all noise. However, most error was observed at frequencies of 1-2% with only a few positions rising above this range. Data for HV2 oligonucleotides was similar, except error frequencies were slightly higher overall ranging from 1-3% for the majority of bases. However, very high error frequencies are observed in low-coverage samples and in areas associated with the c-stretch region. Applying a conservative static threshold to this region would result in a minimum frequency cutoff of 25%. We feel that this approach is not practical and would result in omission of important, analyzable data. A more appropriate method would be to experimentally determine dynamic frequency thresholds for each position within the mtGenome.

The expanded information available from deep mtDNA sequence analysis reveals that once this new technology is implemented into casework practice, interpretational changes in forensic mtDNA reflecting the amounts of information that are produced, are necessary. Massively-parallel sequencing offers a window into a level of variation that is currently under-appreciated in forensic casework.

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 (Andréasson, 2006; Holland, 2011; Bintz, 2014).

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 has attempted 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, ancient or highly degraded calcified tissues, remain a challenge to the forensic DNA typing community. A large 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, 2008; Benschop, 2012; Grisedale, 2012; Pfeifer, 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 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 (Parsons, 2001; Coble, 2004). 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, enable expansion of forensic practice to include the entire mt-genome (King, 2014; McElhoe, 2014; Mikkelsen, 2014; Parson, 2014; Peck, 2016).

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 working protocols to capture the entire mtGenome sequence at sufficient depth to identify and compare variants between forensic samples such as blood, buccal scrapes and hair. Importantly, we have demonstrated that whole mtGenome information may indeed be obtained from compromised human samples including hair shafts and calcified tissues. 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, multiplexed PCR amplification of targeted mtDNA regions, a simple

enzymatic library preparation method using a transposase mediated method, followed by NGS of the templates.

For reference samples, we targeted rapid and efficient NGS of the whole mt-genome molecule. We designed primers to generate two large, overlapping PCR fragments of approximately 9 kb and 11 kb, used the Nextera™ XT library preparation strategy, and then loaded the products onto the Illumina® MiSeq™ instrument. We also took this one step further and designed a direct amplification approach using the same primers to amplify DNA from blood or buccal cells deposited on treated FTA® paper. The results were impressive. In a few simple steps, we were able to generate high quality full mtGenome 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 and calcified tissues. 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. For hair shaft samples, highly efficient extraction was achieved only when hair shafts were chemically digested. The final optimized protocol for hair shaft extraction employs chemical digestion of hair using a Qiagen® reagent followed by magnetic bead-based clean-up with PrepFiler™ (Life Technologies, Inc.) solid phase DNA purification. This method is so effective that it has been adopted with great success by the FBI for use with casework. Next, we focused on evaluating several strategies designed to enrich for human mitochondrial DNA. We tested these methods using a set of samples that included buccal swabs, cremated human remains, hair shafts without follicular tags (modern and ancient), hair shafts, calcified tissues, and fly larvae obtained from deceased donors in various states of decomposition from the Western Carolina University Forensic Osteology Research Station (FOReSt). The same sample extracts were used for all enrichment treatments. Initially, we assessed the Sygnis® TruePrime™ Whole Genome Amplification (WGA) kit in which primers are synthesized in situ with the enzyme *Tth*PrimPol. Isothermal amplification then takes place to create large amounts of starting template for downstream applications. We found inconsistent levels of signal enhancement using this method, even with the use of robust, high-quality samples. Furthermore, the product does not seem to be compatible with enzymatic NGS library preparation methods, even when diluted up to 100-fold. We also evaluated human mtDNA enrichment by probe capture. First, two commercially produced custom probe capture assays were designed. The Agilent SureSelect™ assay includes the use of long RNA baits that are synthesized in an array format. Conversely, the IDT xGen® Lockdown includes use of individually synthesized DNA baits. Both assays require template input amounts that are not typically achievable from forensic samples. Probe capture products from both assays were sequenced on the Illumina®

MiSeq™. We found that the SureSelect™ assay results in high quality, analyzable NGS data from forensically samples. However, IDT xGen® Lockdown product caused a MiSeq™ run failure due to low concentration of captured library. Finally, we compared whole human mtGenome NGS data obtained using our optimized multiplexed PCR enrichment strategy to data generated using other enrichment methods. For most samples in the set, the multiplexed PCR method outperformed the other enrichment methods. This is promising since the PCR method is more cost-effective and less labor intensive than the other methods studied.

In general, it is relatively simple to generate high-throughput and deep coverage sequence data using the NGS methods described herein. However, analysis of large NGS data sets can be daunting. Primary data analysis typically occurs on the attendant instrument PC or server and includes image analysis and basecalling. Some preliminary filtering is often applied to the raw data during primary analysis to remove erroneous base calls related to instrumentation and sequencing chemistry. The resulting output includes a demultiplexed fastq file that can be further analyzed with secondary analysis pipelines that are custom designed or commercially available through second- and third-party vendors. During secondary analysis, additional quality filters are applied, data is aligned to a specified reference genome, and variants from the reference are called. Most software packages have different alignment and variant calling algorithms that may ultimately lead to different interpretations of the same data. In addition to informatics issues, the possibility of errors introduced into the library during sample preparation must also be considered. To investigate these issues, we designed an experiment in which synthetic oligonucleotides with sequences matching the rCRS hypervariable (HV) regions I and II of the human mtDNA genome were purchased from Life Technologies. Initially, Each oligonucleotide was designed to contain Illumina® sequencing primers, flow cell adapters and multiplexing indices on either end to enable direct sequencing without additional preparation. The oligonucleotides were also designed to contain restriction enzyme cut sites between the target sequence and Illumina® modifications. This design allowed for removal of Illumina modifications so the same sample could be prepared for sequencing using recommended library preparation strategies. Each synthetic oligonucleotide was sequenced a) directly with no additional preparation, b) after Illumina® Nextera® XT library preparation, and c) after triplicate PCR amplification with target specific primers followed by Nextera® XT library preparation. Primary analysis was performed on the Illumina attendant PC using Illumina® Real-Time Analysis (RTA) software. Secondary analysis was performed using CLC Genomics Workbench v8.0. Initially, fastq files were aligned to the rCRS using a proprietary alignment algorithm employed by CLC Genomics Workbench. Variant calling was then performed using the Basic Variant Detection algorithm with a 0.1% variant detection frequency threshold. We found no significant difference in error frequencies between treatments. Overall, the frequencies of unexpected variants were low, except in cases where coverage was low or in areas surrounding homopolymeric repeats. In HV1 data, we determined that a static frequency threshold of 5% could be applied above which, variant calls can be interpreted with high confidence. In HV2 data, this minimum frequency threshold climbs to 25% to avoid all erroneous calls. However, this is not a practical solution since quite a bit of high-quality data would also be eliminated using this approach. We recommend establishing a dynamic threshold that is dependent on position

within the genome and depth of coverage. Data was also analyzed using the Low Frequency Variant Detection algorithm that applies an error correction model to estimate sequence error rates. Furthermore, a statistical test is performed at each site to determine if the nucleotides observed in the reads at that site could be due simply to sequencing errors, or if they are significantly better explained by there being one (or more) alleles than the reference present in the sample at some unknown frequency. If the latter is the case, a variant corresponding to the significant allele will be called, with estimated frequency. No unexpected variants were called when this variant calling algorithm was used. While this may seem promising, caution should be used when analyzing data with this method since some biologically relevant data may be eliminated from the data set.

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 advancement in the area of forensic DNA analysis, and have implications well beyond human DNA, including microbial forensics and metagenomic analyses.

Introduction

Statement of the Problem

Although human mtDNA analysis is currently only performed in a small subset of forensic DNA laboratories, its utility in some forensic contexts is incontrovertible. Part of the reason for this limitation is that the informativeness of mtDNA is much less than that provided by forensic STR analysis and the interpretational complications that arise from heteroplasmy. 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) analysis on these samples. Recent court challenges to LCN analysis have highlighted these limitations. We believe that to be more useful in

forensic DNA laboratories, mtDNA analysis needs to be extended in two directions, the amount of sequence information analyzed, and the depth of sequence analyzed at each position of sequence. 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 obtainable for the forensic DNA community. It has been our long-term goal to develop methods and approaches to rapidly generate high-quality DNA sequence information from the whole mtDNA genome in support of forensic casework. An expansion of the population database used to support forensic casework is a critical component of this goal.

Literature Review

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. GlobalFiler™) 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 of limited or degraded template (Lindahl, 1972). 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 (Anderson, 1981). Mitochondrial DNA is a small, circular molecule of about 16,569 bp (Wolstenholme, 1992). 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, hypervariable regions (HV) HV1, HV2 and HV3. The HV1 (16024 to 16365), HV2 (73 to 438) and HV3 (438 to 574) positions are typically targeted for forensic identification purposes because of the high density of sequence variation (Tamura, 1993; Pesole, 1992, 1999; Wallace, 1999). Mitochondrial DNA has two additional unique features that make it particularly suitable for the analysis of biological remains, e.g. hair, calcified tissue, blood, and extremely limited or degraded DNA samples. First, mtDNA is inherited matrilineally (Giles, 1980). This mode of inheritance makes it a valuable genetic marker for 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 (Wilson, 1995; Holland, 1999). The 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 hundreds to thousands of copies of mtDNA molecules can be present per cell (Robin 1988).

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 (Calloway 2000; Sekiguchi 2004; Irwin 2009; Li 2010; Sosa 2012; Naue 2015). 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 (Allen, 1998; Budowle, 1999, 2003; Wilson, 2004). Accordingly, interpretational guidelines have been developed that are cognizant of these facts (Carracedo, 2000; SWGDAM, 2003; Parson, 2014). 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 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 Tool in Forensic DNA Casework

The possibilities offered by next generation sequencing (NGS) platforms are revolutionizing biotechnological laboratories. Over the past five to ten 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 2009; Metzker 2010; Verma 2017). 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), biases, 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 (Schwartz 2011; Nothnagel 2011; Dewey 2014; Shin 2016). 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. Sequencing errors may be associated with 1) sample preparation 2) sequencing chemistries and 3) bioinformatic processing of data. Regardless of their origin, these sequence artifacts must be removed before downstream analyses, otherwise they may lead to erroneous conclusions. In order to do this effectively, a systematic study must be performed to elucidate the cause of error, how different types of error appear in data pile-ups, and best quality filtering practices to exclude error from analyses. Many instrument-associated GUI-based software programs available for downstream analyses do not provide a flexible means for quality checking and filtering of NGS data before downstream processing. Additionally, a multitude of third-party are available, each

offering different algorithms for quality filtering, read mapping and variant calling that could potentially give rise to differences in data output and ultimately affect data interpretation. Therefore, it is advisable to assess the affects of quality filtering of sequencing data at the end-user level.

Sample Preparation

Several studies have been conducted to evaluate NGS methods for analysis of mtDNA from forensic samples (Parson, 2013; Templeton, 2013; McElhoe, 2014; Chaitanya, 2015). However, a systematic effort in the forensic community is needed to fully assess error associated with different NGS methods in order to assist with establishment of standardized methods. Focus should be placed on investigating error introduced during each discrete step of NGS sample preparation to identify areas in which improvements can be made so that error rates can be reduced when possible. Unfortunately, the need for these types of studies is not unique to forensic science (Chain, 2009; Gargis, 2012; Endrullat, 2016). The depth of analysis obtainable with NGS enables detection of variants well below the 10% threshold offered by Sanger sequencing. This is revolutionary for deconvolution of mixtures and identification of low-level heteroplasmy that could be used to increase the discriminatory power of mtDNA (Just, 2015; Kim, 2015). However, there is a paucity of information regarding methods used to select NGS variant frequency thresholds to exclude error and noise and include true biological variation. This threshold is likely dependent on many factors including sample preparation strategies, sequencing chemistry, motif surrounding the position of the basecall, depth of coverage, and bioinformatic processing of the data.

It is well known that base substitutions and INDELS are often introduced during the PCR process (Eckert, 1991; Batra, 2016). The rate at which this occurs is dependent on the proofreading capability and fidelity of the polymerase enzyme used. Thus, NGS vendors require use of high-fidelity enzymes for target enrichment PCR and limited cycle amplification employed during library preparation. However, some error is likely still introduced with these enzymes. Elucidating the rate at which these errors occur and whether the errors are position dependent could assist with establishment of variant frequency thresholds for each position within the mitochondrial genome.

Preparation of forensic samples that contain very small amounts of DNA that may also be degraded require a separate focus that is based on the sample metadata. These sample types are often insufficient to support traditional PCR amplification targeting the entire mtGenome, and hence may require a staged amplification approach that employs whole genome amplification (WGA) in the first step. While some work has been done to prove that WGA methods can increase DNA template in compromised samples in an unbiased manner, the work is not exhaustive and methods have not been evaluated with next-generation sequencing in a forensic context (Giardina, 2009; Tate, 2012; Maciejewska, 2014). Studies must be conducted to determine whether enrichment of template DNA from compromised samples using WGA introduces bias or elevated levels of base misincorporations that could convolute interpretation downstream.

Probe capture based enrichment of DNA from compromised samples is a newly emerging method that shows promise (Templeton, 2013; Gadipally, 2015; Wendt, 2016). Several kitted solutions have recently become available in which a custom set of baits is designed to capture a desired target. One such method, the SureSelect^{XT} target enrichment system available from Agilent, utilizes RNA baits to capture DNA after NGS library preparation. Conversely, the xGen® Lockdown® method from Integrated DNA Technologies® relies on hybridization of templates from an NGS library to DNA baits. Several studies have shown that these methods are sensitive and reliable when using highly concentrated, robust DNA namely from clinical samples (Brown, 2016; Garcia-Garcia, 2016).

Combining enrichment methods (PCR, WGA and/or probe capture) may be necessary to generate enough template DNA from forensic samples to enable NGS downstream. This type of strategy has been evaluated using high quality HapMap gDNA samples (ElSharawy, 2012). In this study, nanogram quantities of DNA (≥10 ng) were either PCR amplified using an emulsion method or subjected to WGA pre-amplification prior to emulsion PCR. Ultimately, microgram quantities of DNA were sequenced using SOLiD methods and data was analyzed using CLC Genomics Workbench with a 10% minimum allele frequency cutoff. The authors report high concordance in variants called between those samples that were pre-amplified with WGA and those that were not. Unfortunately, due to the high quality of the samples used in this study and the conservative analysis parameters used, the data cannot be extrapolated to forensic samples. Additional work is needed to evaluate these methods both individually and in combination for forensic use and to assess the reproducibility and quality of the data generated from samples prepared using the aforementioned methods (Nietsch, 2016).

In addition to enrichment, samples are often prepared for NGS by first fragmenting the DNA into a range of sizes that are compatible with the sequencing chemistry employed. After fragmentation, platform specific adapters are bound to the fragmented DNA, which allow the DNA to bind to the solid support on which sequencing takes place. In addition, barcoding indices are also incorporated to facilitate multiplexing of many samples in any given run. The barcodes are used to bioinformatically parse data from each sample when a run completes. There are several methods available to prepare libraries for sequencing. For examples, fragmentation of the DNA can be performed manually using focused acoustics, chemically using enzymatic methods or not at all if short amplicons are being sequenced. Each method of fragmentation requires different downstream processing for library preparation and likely results in different error rates in raw data.

Sequencing chemistry

Several commercially available NGS technologies exist including Illumina® Reversible Terminator Sequencing (Bentley, 2008), Ion Semiconductor Sequencing (Rothberg, 2011), Oxford Nanopore Sequencing (Iqbal, 2007), and Pacific Biosciences® single molecule real-time (SMRT) Sequencing (Harris, 2008; Eid, 2009). Each technology uses drastically different technology to sequence DNA. As a result, different errors may be observed when utilizing different sequencing methods. For example, Quail

et. al. evaluated all aforementioned platforms except the Oxford Nanopore method. Briefly, they sequenced 4 microbial genomes on each platform and compared the resulting data. Coverage was even in "GC rich, neutral and moderately AT rich" regions across all methods. However, in Ion Torrent™ data, entire areas with zero coverage were observed in AT rich genomes. Furthermore, more variants were detected in Ion Torrent™ data, however, the false positive rate was higher in these data as well. Variant calling in Pacific Biosciences data required higher depth of coverage than other methods. Finally, the authors report higher context specific errors in Ion Torrent™ and Illumina® data sets, but not in Pacific Biosciences data sets (Quail, 2012). Similar studies should be completed using human mtDNA with a known sequence.

Bioinformatic analysis

Several bioinformatics resources using different data processing algorithms have been developed for the processing of NGS data (Li 2009; McKenna 2010; Goecks 2010; Schmieder 2010; Merchant, 2016). However, there is still a need for the development of universal tools that conform to forensic standards. In order to develop these tools, a comprehensive analysis of the impact of sequencing artifacts, chemistry and instrument-dependent errors, utilization of different alignment and variant calling algorithms and modification of quality filtering options is needed. While some preliminary work has been done in this area, further assessment is needed to truly understand how each of these ultimately affect data output and interpretation. One recent study performed by Peck et. al. attempts to elucidate some of these issues. Further work is still needed.

NGS technologies are not the same. For instance, two NGS technologies, PacBio RS® (Pacific Biosciences) and Illumina® sequencing by synthesis technologies have equal or greater read lengths than Sanger sequencing (Margulies, 2005; Mardis, 2008; Glenn, 2011; Nothnagel, 2011; Carneiro, 2012; Loman 2012; Quail, 2012; Shin, 2016). In contrast, the Ion Torrent® generally yields 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 (Kidd, 2008, 2010; Dalloul, 2010; Gnerre, 2011). 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 (Mills, 2006, 2011; Korbel, 2007; Kidd, 2008, 2010; Alkan, 2009, 2011; Yoon, 2009; Antonacci, 2010, Medvedev, 2010; Teague, 2010; Handsaker, 2011; Schmeider, 2011; Huddleston, 2016; Jovel, 2016, Tarnecki, 2017).

Analysis of sequence data from an NGS run is commonly performed as files sequentially progress through a series of algorithms in the form of a pipeline. More generally, a pipeline is a specialized form of workflow management system designed to execute a series of computational or data manipulation steps. These steps include quality-filtering, alignment and mapping to a specified reference genome, and variant calling algorithms. 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. Some pipelines are preconfigured and automated allowing for limited manipulation by the user while others are fully customizable and parameter intensive.

From a forensic validation perspective, if data output can be affected by modification of a particular parameter, then the elements in the pipeline related to that parameter 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, 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.

Due to major advances in enzymology, DNA sequencing technology, bioinformatics and data processing, the potential now exists to overcome remaining limitations and greatly simplify analysis whole human mitochondrial genomes. These

improvements will offer a significant advancement in the field of forensic DNA typing. We propose to examine seven interrelated aspects of the entire analysis flow that are directly germane to forensic DNA typing: optimized DNA extraction methods; new development of nuclear and mtDNA quality assessment tools; whole genome amplification (WGA); optimization and application of multiplex PCR amplification of mtDNA from challenging sample types; direct sequencing of DNA extracts using probe capture technology, and rapid and efficient preparation of reference samples for NGS.

Deliverables

Development of Duplex mt/nuclear DNA Quantitation Tools

In forensic casework, multilocus short tandem repeat (STR) typing is often the preferred method of analysis due to its high power of discrimination. However, many evidentiary samples contain low amounts of DNA, or degraded DNA that is not suitable for STR typing. In these cases, mitochondrial DNA sequence analysis is typically performed. Determining which investigative approach is most suitable can be challenging, especially in cases where the sample or extract is limited. Here, we describe a powerful multiplex 5' nuclease DNA quantitation assay that enables simultaneous quantification of both human nuclear and mitochondrial DNA from a sample extract. This assay has been designed to work successfully on a real-time PCR instrument or a droplet digital™ PCR instrument with no modifications. This tool provides specific quantitative data that can be used to determine the most appropriate analytical workflow without consumption of additional sample or increase in labor compared to methods currently used in crime laboratories.

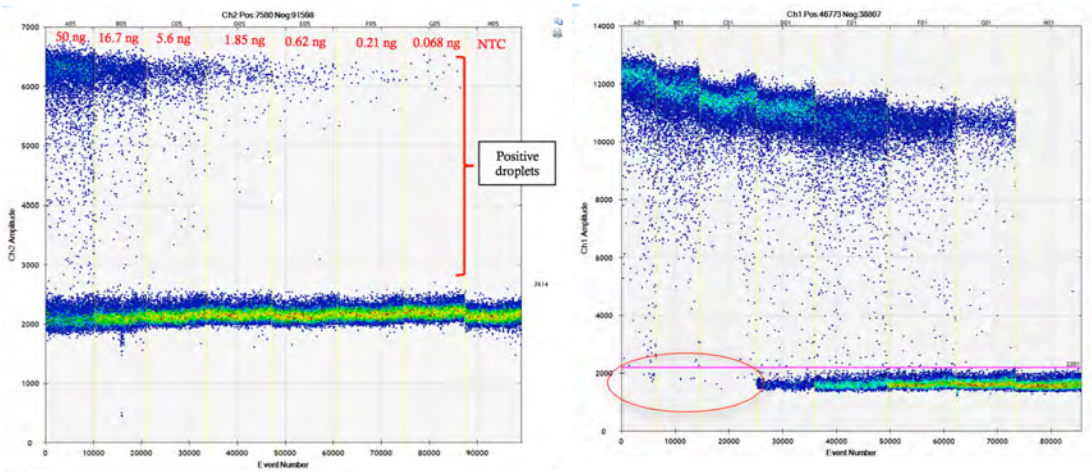
Droplet digital™ PCR (ddPCR™) is similar to quantitative PCR (qPCR) in that target specific primers and 5' nuclease probes are utilized for detection following an end-point PCR reaction. However, due to the nature of the method, no standard curve is needed for estimation of DNA concentration. With ddPCR™, a 20 µL aqueous PCR reaction is emulsified into 1 nL uniformly-sized droplets. Each droplet is then counted as fluorescence positive or negative and a Poisson correction is applied to estimate the starting copy number of DNA fragments in the sample.

The CODIS STR TH01 locus was first described as a target for a quantitative PCR assay by Swango et. al. in 2006. The target ranges in size from 170-190 bp (a mid-range length target compared to other CODIS loci) and can provide relevant information about the amplifiability of STR loci from a particular sample. The originally described assay employed a FAM™-labeled probe that had to be redesigned for multiplexing with an ND5 human specific mtDNA quantitation assay developed by Kavlick et. al., which also uses a FAM™-labeled probe (Kavlick, 2011). As a result, the reporter dye on the TH01 probe was changed to VIC™ for multiplexing capability in both qPCR and ddPCR™ assays with no additional modifications.

ddPCR™ assessment of ND5 and TH01 assays

Singleplex and multiplex assays with the TH01 primers and newly designed probe were run on a QX200™ ddPCR™ instrument (BioRad, Hercules, CA). A serial dilution was performed using 9947A control DNA (Promega, Madison, WI) with a starting concentration of 10 ng/µL for resulting ddPCR™ reaction inputs ranging from 50 ng – 68 pg. As a singleplex reaction, the assay appears to work very well (figure 1A). Estimated theoretical copy numbers calculated using input amounts were the same as those reported by the BioRad QuantaSoft™ software (BioRad, Hercules, CA). When multiplexed with the ND5 assay, the assay also performs well except when input concentrations of DNA are ≥16.7 ng. At these concentrations, droplets are saturated with ND5 targets (figure 2). It is likely that high numbers of ND5 target molecules per droplet are causing PCR inhibition by competition of the TH01 assay. However, this is not expected to pose an issue since forensically relevant samples rarely yield high concentrations of extracted DNA. If a sample is expected to yield a high concentration extract, input sample volume can be decreased from 5 µL to 1 µL per reaction.

Figures 1A and 1 B: *1D amplitude plot for TH01 (left) and ND5 (right) singleplex reactions.*



Figures 1A (left) and 1B (right). Ideal separation between fluorescence positive and negative droplets was observed for the singleplex TH01 assay at all concentrations ranging from 50 ng - 68 pg of input DNA. As a result, the assay appears to be robust and capable of quantifying DNA over a range suitable for forensic analyses. The ND5 assay also appears to be robust. However, droplet saturation is observed at high concentrations (red circle, figure 1B). This leads to inaccurate estimate of starting concentrations of DNA. This not observed with the TH01 assay since nuclear target copy number is substantially lower than mtDNA copy number in a sample.

Figure 2: *1D Amplitude plots for TH01 and ND5 multiplex reactions.*

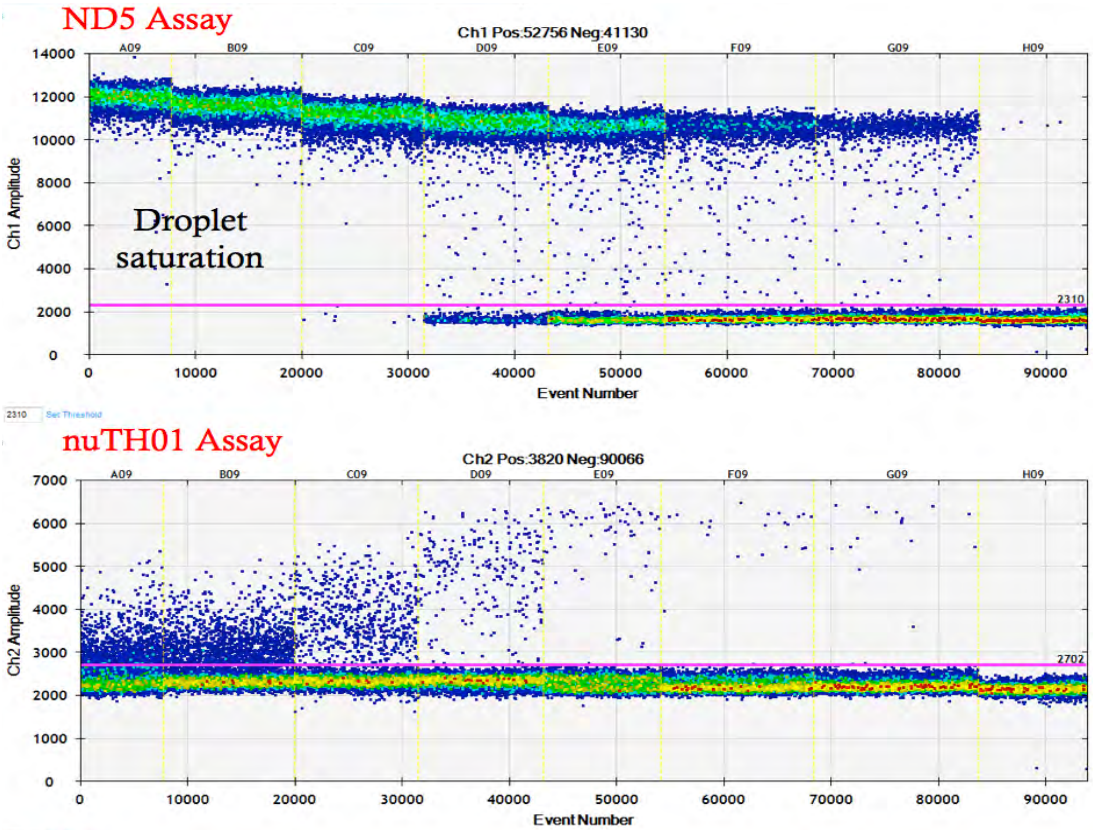


Figure 2: Droplet saturation is observed in ND5 data represented in columns A09 and B09 (above). No negative droplets are observed for these samples. Quantitation of these samples is not possible, since accurate quantitation relies on a Poisson algorithm requiring at least some negative droplets present. Very few negative droplets are present in the ND5 reaction containing 5.56 ng of input DNA (column C09). However, quantitation in this case was possible. Data from the same samples quantified using the TH01 assay was lower in multiplex reactions versus singleplex reactions due to inhibition by competition occurring at high concentrations of input DNA. Multiplex TH01 data from samples with lower concentrations of input DNA is unaffected.

Table 1: *ddPCR™ quantitation of ND5 and TH01 targets in both singleplex and multiplex reactions.*

Sample ID (Input)	ND5 Target Triplicate Average (copies /µL)		TH01 Target Triplicate Average (copies /µL)		Theoretical nuDNA copy number
	Singleplex	Multiplex	Singleplex	Multiplex	
50 ng	669,267	1,000,000	676.67	221.33	757.0
16.7 ng	669,667	1,000,000	211.33	159.67	253.0
5.56 ng	8,633	8,400	73.00	66.83	84.2
1.85 ng	2,857	3,003	25.57	25.20	28.0
0.62 ng	975	958	8.53	9.23	9.4
0.21 ng	337	319	2.20	2.93	3.2
0.068 ng	119	92	1.11	1.43	1.0
NTC	6.6	0.24	0.00	0.00	0

Table 1: Data for ND5 singleplex and multiplex reactions was reproducible for samples with DNA input ranging from 5.56 ng – 68 pg. Droplet saturation was observed in samples with 50 and 16.7 ng of input DNA. Quantitation calculations cannot be performed for these samples because the Poisson algorithm requires the presence of at least some negative droplets. As a result, concentrations are overestimated in these samples (capping at 1,000,000 copies/µL). Data for the TH01 target is also very good for qantifying DNA inputs ranging from 1.85 ng – 68 pg. For samples with higher inputs of DNA, concentrations are underestimated when using the multiplex assay, presumably due to PCR inhibition by competition occurring as a result of high numbers of mtDNA target per droplet. This is observed in reactions with 50 and 16.7 ng of input DNA, and to a lesser degree in the reaction containing 5.56 ng of input. However, DNA extracts from forensically relevant samples are unlikely to contain such high amounts of DNA. If a robust sample is expected to yield high amounts of DNA, ddPCR™ input volume can be reduced from 5 µL to 1 µL per reaction.

Characterization of NIST SRM 2372 component A for use as a qPCR standard

The DNA used for as a standard for qPCR was component A from NIST human quantitation standard SRM 2372 (NIST, Gaithersburg, MD). Component A is derived from a single male donor and is provided at a concentration of 57 ng/µL in 110 µL of low TE buffer. For use as a qPCR standard, the stock solution is diluted to 50 ng/µL and a 10X serial dilution is performed to obtain a total of 8 standards with concentrations ranging from 50 ng/µL – 5 fg/µL. Standards 2-8 were then quantified using the BioRad QX200™ ddPCR™ instrument and aforementioned ND5 assay to determine the mtDNA copy number.

Table 2: *ddPCR™ quantitation of mtDNA in NIST SRM 2372 component A*

	Copies of mtDNA/µL				
Standard	Replicate 1	Replicate 1	Replicate 1	Average	Standard Deviation
5 ng/µL	4,000,000	4,000,000	4,000,000	4,000,000	0
0.5 ng/µL	17,040	16,800	16,480	16,733.33	280.95
0.05 ng/µL	1,484	1,516	1,572	1,524.0	44.54
5 pg/µL	158	156.4	147.2	153.87	5.83
0.5 pg/µL	10	20.8	17.2	16.00	5.50
0.05 pg/µL	3.6	3.36	1.96	2.97	0.89
5 fg/µL	2.04	1.12	1.08	1.41	0.54
Negative Control	0.6	1.08	0.84	0.84	0.24

Table 2: Standards 1 and 2 (50 ng/µL, 5 ng/µL) were intentionally omitted from analysis since their concentrations are too high for accurate ddPCR™ quantitation leading to droplet saturation and failed estimation of copy numbers. When droplet saturation occurs, Quantasoft™ software displays a standard maximum value for sample concentration. For standards 3-6 (0.5 ng/µL – 0.5 pg/µL) reported absolute quantitation values are accurate and show an approximate 10-fold difference from sample-to-sample. These values can be extrapolated to standards 1 and 2. Standards 7 and 8 (0.05 pg/µL and 5 fg/µL respectively) did not differ significantly from a negative control. This information is used to include mtDNA copy numbers of standards into qPCR software.

Validation of singleplex and multiplex ND5 and TH01 qPCR assays

In addition to adopting a new standard for the qPCR assay, a new exogenous IPC developed by Mark Kavlick at the FBI was also evaluated (personal communication). Use of the newly designed IPC, which contains a NED™ reporter and a non-fluorescent quencher, circumvents use of the TAMRA quencher that is utilized in the TaqMan® exogenous IPC kit. TAMRA quenchers are limiting in multiplex experiments because unlike non-fluorgenic quenchers, use of TAMRA results in fluorescence emission, which may contribute to background signal or overlap with signal from other reporters used in the assay.

An experiment was conducted to verify that each independent assay (TH01, ND5, and IPC) worked well in qPCR singleplex format and that multiplexing the assays had no derogatory effect on assay efficiency. Standards 1-6 were used for the mtDNA portion of the assay, while standards 1-5 were used for the TH01 assay. HL60 (20 and 100 pg/µL) was also quantified to provide a point of comparison between singleplex and multiplex tests (ATCC, Manassas, VA). Final primer concentrations for ND5 and TH01 assays were 900 nM, with 250 nM final concentrations of probes. For the IPC assay, final primer concentrations were 300 nM and final probe concentration was 250 nM.

ND5 Data

Tables 3A-3C: *qPCR efficiency study – singleplex versus multiplex data for ND5 target*

	ND5 singleplex	ND5 multiplex
Slope	-3.391	-3.033
Y-intercept	38.469	35.928
R^2	1	0.998
Efficiency (%)	97.208	113.673

Table 3A: Comparison of line statistics obtained for singleplex and multiplex reactions. The slope and efficiency of the reaction increase substantially when assays are multiplexed. Inhibition is often considered an explanation when PCR efficiencies exceed 100%. It is possible that competitive inhibition is occurring in multiplexed reactions as a result of the increase in quantified targets per reaction.

	ND5 singleplex		ND5 multiplex	
Sample	C_T of Target	C_T difference	C_T of Target	C_T difference
50 ng/µL	17.467		16.831	
5 ng/µL	20.818	3.351	20.153	3.322
0.5 ng/µL	24.812	3.994	23.47	3.317
0.05 ng/µL	27.567	2.755	26.458	2.988
5 pg/µL	31.018	3.451	29.256	2.798
0.5 pg/µL	34.211	3.193	32	2.744
HL60 20 pg/µL	25.621		24.735	
HL60 100 pg/µL	22.97		22.30	
NTC	Undetermined		Undetermined	

Table 3B: Comparison of C_T values of standards and control samples in singleplex and multiplex reactions. In general, C_T values are lower for multiplex data. Since the standards were prepared via a 10-fold dilution, C_T differences between standards should be 3.32. The average C_T difference for standards assessed using the singleplex is 3.3488,

while the average C_T difference for standards assessed with the multiplex is 3.0338.

ND5 quantitation (copies/µL) – triplicate averages		
	Singleplex	Multiplex
HL60 20 pg/µL	6,171.062	4,909.907
HL60 100 pg/µL	37,262.43	31,175.00
NTC	Undetermined	2.306

Table 3C: Quantitation values of diluted HL60 samples. In general, reported copy numbers are lower for multiplex data.

TH01 Data

Tables 4A-4C: *qPCR efficiency study – singleplex versus multiplex data for TH01 target*

	TH01 singleplex	TH01 + IPC	TH01 full multiplex
Slope	-3.314	-3.273	-3.139
Y-intercept	29.496	29.554	26.737
R^2	0.998	1	0.989
Efficiency (%)	100.336	102.101	108.239

Table 4A: Comparison of line statistics obtained for singleplex and multiplex reactions. The slope and efficiency of each reaction increases as multiplexing becomes more complex.

	TH01 singleplex		TH01 + IPC		TH01 full multiplex	
Sample	C_T of Target	C_T difference	C_T of Target	C_T difference	C_T of Target	C_T difference
50 ng/µL	23.685		23.94		21.519	
5 ng/µL	27.211	3.526	27.273	3.333	24.26	2.741
0.5 ng/µL	30.664	3.453	30.587	3.314	27.646	3.386
0.05 ng/µL	34.096	3.432	33.903	3.316	30.085	2.439
5 pg/µL	36.813	2.717	36.939	3.036	34.535	4.45
HL60 20 pg/µL	33.076		32.91		28.69	
HL60 100 pg/µL	30.54		30.354		27.038	
NTC	Undetermined		Undetermined		Undetermined	

Table 4B: Comparison of C_T values of standards and control samples in singleplex and multiplex reactions. As with the ND5 assay, C_T values for the TH01 target decrease as multiplexing becomes more complex. Again, since the standards were prepared via a 10-fold dilution, C_T differences between standards should be 3.32. C_T differences are shown in the table above. The average C_T difference for standards assessed using the singleplex is 3.282, 3.249 for the TH01 + IPC multiplex, and 3.254 the full multiplex.

	TH01 quantitation (ng/µL) – triplicate averages		
	singleplex	+IPC only	full multiplex
HL60 20 pg/µL	0.083	0.095	0.241
HL60 100 pg/µL	0.49	0.571	0.803
NTC	Undetermined	Undetermined	Undetermined

Table 4C: Quantitation values of diluted HL60 samples. In general, reported copy numbers are higher for multiplex data.

The increase in efficiency for each assay in multiplex reactions suggests that competitive inhibition may be occurring. This is likely due to the difference in copy number of nuclear and mitochondrial targets, and early sequestration of reaction components by the mtDNA assay. There are also other discrepancies in the data that may

be a result of unbalanced target numbers, or reaction component interactions. It is probable that lower concentrations of primers and probes of the ND5 assay would ameliorate these issue. To test this theory, an experiment was performed in which concentrations of ND5 and IPC primers and probe were varied to determine the lowest amount of each that would result in favorable quantitation of targets target without affecting the performance of each assay.

Titration studies for the optimization of assay primer/probe concentrations

Initially, four reactions were run with varying concentrations of IPC primers and probe. Each reaction was multiplexed with the TH01 assay. Resulting data was compared to singleplex data, and appropriate concentrations of IPC primers and probe were selected for future use. The data is shown in tables 5A-5D.

IPC Optimization

Tables 5A-5D: *qPCR IPC optimization study*

	No IPC	300 nM primer/250 nM probe	150 nM primer/125 nM probe	100 nM primer/88.2 nM probe	50 nM primer/41.6 nM probe
Slope	-3.314	-3.166	-3.19	-3.189	-3.394
Y-intercept	29.496	29.553	29.616	29.666	29.358
R^2	0.998	0.999	0.998	0.998	0.998
Efficiency (%)	100.336	106.969	105.835	105.858	97.078

Table 5A: Comparison of standard curve line statistics across all reactions. Data that most closely resembles that obtained from a TH01 singleplex assay is obtained with IPC final primer concentrations of 50 nM and a final probe concentration of 41.6 nM.

Sample	300 nM primer/250 nM probe	150 nM primer/125 nM probe	100 nM primer/88.2 nM probe	50 nM primer/41.6 nM probe
50 ng/µL	27.938	27.65	26.727	26.874
5 ng/µL	26.935	26.749	26.01	25.647
0.5 ng/µL	27.036	26.874	25.063	26.128
0.05 ng/µL	26.051	25.969	25.508	25.394
5 pg/µL	26.523	26.779	26.12	25.83
HL60 20	26.939	26.896	26.277	26.218
HL60 100	27.584	27.607	26.924	26.917
NTC	27.676	27.917	27.262	27.117
Average	**27.085**	**27.055**	**26.236**	**26.266**

Table 5B: IPC C_T values obtained for all samples and controls across all treatments. The IPC C_T should not vary from sample-to-sample unless the reaction is not performing well or inhibition is occurring. An ANOVA was performed to statistically assess the similarity of the means of the data sets. A p-value of 0.01537 was obtained, which suggests that at least some of the treatments result in IPC C_Ts that are statistically different. This is not unexpected, since primer/probe concentrations can affect reaction kinetics.

Sample	No IPC	300 nM primer/250 nM probe	150 nM primer/125 nM probe	100 nM primer/88.2 nM probe	50 nM primer/41.6 nM probe
50 ng/µL	23.685	24.187	23.99	24.088	23.506
5 ng/µL	27.211	27.449	27.501	27.501	26.961
0.5 ng/µL	30.664	30.365	30.818	30.81	30.432
0.05 ng/µL	34.096	33.582	33.763	33.98	34.087
5 pg/µL	36.813	36.948	36.807	36.835	36.913
HL60 20	33.076	32.81	33.084	33.294	32.713
HL60 100	30.54	30.39	30.513	30.636	30.029
NTC	Undetermined	Undetermined	Undetermined	Undetermined	Undetermined

Table 5C: C_T values for TH01 target obtained for all samples and controls across all treatments. An ANOVA was performed to determine whether the means of the data sets differed. A p-value of 0.9988 was obtained indicating that varying concentrations of the IPC primers/probe has little to no affect on the C_T values of the TH01 assay.

	TH01 quantitation (ng/µL) – triplicate averages				
IPC Reagent Concentrations ➔	NA	300 nM primer/250 nM probe	150 nM primer/125 nM probe	100 nM primer/88.2 nM probe	50 nM primer/41.6 nM probe
HL60 20 pg/µL	0.083	0.106 ng/uL	0.082 ng/uL	0.073 ng/uL	0.103 ng/uL
HL60 100 pg/µL	0.49	0.604 ng/uL	0.523 ng/uL	0.497 ng/uL	0.635 ng/uL
NTC	Undetermined	Undetermined	Undetermined	Undetermined	Undetermined

Table 5D: Quantitative data for TH01 target of control samples. Target copy number differs significantly between treatments. However, the difference is mainly encountered when comparing singleplex data to multiplex data. Future assays will include use of the lowest concentrations of IPC primers/probe due to the small difference observed between singleplex and multiplex data.

ND5 Optimization

Tables 6A-6F: *qPCR ND5 optimization study*. Each assay was run in singleplex format with final primer concentrations of 900 nM and final probe concentrations of 250 nM. Each optimization experiment was a multiplex reaction with the TH01 assay (final concentrations of 900 nM per primer and 250 nM probe), and ND5 with indicated primer and probe concentrations.

	ND5 SP	900 nM primer/250 nM probe	600 nM primer/166.4 nM probe	300 nM primer/83.2 nM probe	180 nM primer/50 nM probe	120 nM primer/33.3 nM probe	60 nM primer/16.6 nM probe
Slope	-3.391	-3.1	-3.318	-3.422	-3.397	-3.436	-3.431
Y-intercept	38.469	37.385	37.758	37.117	35.58	35.744	36.07
R^2	1	0.992	0.999	1	0.999	1	1
Efficiency (%)	97.208	110.155	100.168	95.979	96.975	95.458	95.637

Table 6A: Human mtDNA standard curve line statistics for all ND5 primer/probe variations. Slope and efficiency values differ slightly between treatments. In all multiplex reactions except that containing final primer concentrations of 900 nM, efficiencies fall within an acceptable range of 90-100%. Y-intercepts are also similar with a small standard deviation between treatments. (Note: SP = singleplex)

	TH01 SP	900 nM primer/250 nM probe	600 nM primer/166.4 nM probe	300 nM primer/83.2 nM probe	180 nM primer/50 nM probe	120 nM primer/33.3 nM probe	60 nM primer/16.6 nM probe
Slope	-3.314	-3.22	-3.256	-3.37	-3.42	-3.536	-3.395
Y-intercept	29.496	28.718	28.818	28.698	28.712	28.208	28.639
R^2	0.998	0.993	0.991	0.998	0.999	0.996	0.998
Efficiency (%)	100.336	104.45	102.849	98.032	96.072	91.763	97.054

Table 6B: Human nuclear DNA standard curve line statistics for all ND5 primer/probe variations. This data is similar to the ND5 data presented in table 6A. Very little difference is observed between treatments.

Sample	ND5 SP	900 nM primer/250 nM probe	600 nM primer/166.4 nM probe	300 nM primer/83.2 nM probe	180 nM primer/50 nM probe	120 nM primer/33.3 nM probe	60 nM primer/16.6 nM probe
50 ng/µL	17.467	17.708	17.041	15.897	14.563	14.45	14.726
5 ng/µL	20.818	21.174	20.44	19.292	17.828	17.777	18.247
0.5 ng/µL	24.812	24.685	23.92	22.727	21.26	21.253	21.62
0.05 ng/µL	27.567	27.952	27.262	26.179	24.685	24.629	25.155
5 pg/µL	31.018	31.009	30.71	29.447	28.266	28.16	28.582
0.5 pg/µL	34.211	32.864	33.436	33.082	31.389	31.666	31.836
HL60 20	25.621	25.992	25.303	24.224	22.634	22.707	23.069
HL60 100	22.97	23.321	22.664	21.524	19.867	20.042	20.301
NTC	Undet	33.558	35.781	35.58	35.666	36.037	35.93

Table 6C: Sample C_T values for ND5 target. In general, a decrease in C_T value is observed as primer and probe concentrations decrease. This is not unexpected since primer and probe concentrations can have an impact on PCR reactions kinetics. ANOVA analysis shows that means of data sets are not significantly different (P-value of 0.79242, 95% confidence interval). However, comparison of each multiplex data set to the singleplex data set using a student t-test shows that significant differences do occur. Again, this is not unexpected since changes in primer and probe concentrations can affect reaction kinetics.

Sample	TH01 SP	900 nM primer/250 nM probe	600 nM primer/166.4 nM probe	300 nM primer/83.2 nM probe	180 nM primer/50 nM probe	120 nM primer/33.3 nM probe	60 nM primer/16.6 nM probe
50 ng/µL	23.685	23.054	23.19	22.916	22.958	22.249	22.846
5 ng/µL	27.211	26.492	26.523	26.323	26.285	25.696	26.304
0.5 ng/µL	30.664	29.811	29.951	29.917	29.697	29.197	29.748
0.05 ng/µL	34.096	33.436	33.196	33.091	33.132	32.953	32.868
5 pg/µL	36.813	35.669	36.13	36.396	36.634	36.331	36.538
HL60 20	33.076	32.044	32.433	32.115	32.15	31.351	32
HL60 100	30.54	29.136	29.464	29.315	29.296	28.613	29.218
NTC	Undet	Undet	Undet	Undet	Undet	Undet	Undet

Table 6D: Sample C_T values for TH01 target. In general, a slight decrease in C_T value is observed as primer and probe concentrations decrease. However, ANOVA analysis shows that means of data sets are not significantly different (P-value of 0.99902, 95% confidence interval). However, comparison of each multiplex data set to the singleplex data set using a student t-test shows that significant differences do occur. This is not unexpected since changes in primer and probe concentrations can affect reaction kinetics.

	ND5 quantitation (copy number/µL) – triplicate averages						
ND5 Reagent Concentrations ➔	ND5 SP	900 nM primer/250 nM probe	600 nM primer/166.4 nM probe	300 nM primer/83.2 nM probe	180 nM primer/50 nM probe	120 nM primer/33.3 nM probe	60 nM primer/16.6 nM probe
HL60 20 pg/µL	6171.062	4,728.75	5,679.63	5,857.04	6,475.62	6,236.91	6,164.32
HL60 100 pg/µL	37,262.43	34,392.50	34,182.80	36,024.10	42,279.88	37,183.82	39,468.26
NTC	Undet	17.356	3.968	2.849	1.309	0.938	1.103

Table 6E: Quantitative data for ND5 target of control samples. Target copy number differs significantly between treatments (ANOVA p-values of 0.003 for HL60 20 pg/µL and 0.008 for HL60 100 pg/ µL; 95% confidence interval). However, independent comparison of multiplex datasets against the singleplex dataset shows no significant difference is observed when the lowest concentrations of ND5 primers/probe are used (probability associated with a student's t-test with a two-tailed distribution: 0.987, 0.167 respectively). This indicates that use of the lowest concentrations of ND5 primers/probe is sufficient for accurate mtDNA quantitation.

	TH01 quantitation (ng/µL) – triplicate averages						
ND5 Reagent Concentrations ➔	TH01 SP	900 nM primer/250 nM probe	600 nM primer/166.4 nM probe	300 nM primer/83.2 nM probe	180 nM primer/50 nM probe	120 nM primer/33.3 nM probe	60 nM primer/16.6 nM probe
HL60 20 pg/µL	0.083	0.093	0.078	0.097	0.099	0.116	0.102
HL60 100 pg/µL	0.49	0.742	0.622	0.663	0.675	0.77	0.675
NTC	Undet	Undet	Undet	Undet	Undet	Undet	Undet

Table 6F: Quantitative data for TH01 target of control samples. Target copy number does differ significantly between treatments. Quantitation values obtained with singleplex reactions are lower than values obtained in all multiplex reactions. There is no statistically significant difference between data sets obtained with multiplex quantitation of HL60 20 pg/µL samples (ANOVA p-value of 0.1192). Data sets obtained with multiplex quantitation of HL60 100 pg/µL samples are significantly different (ANOVA p-value of 0.00299). However, these differences are not enough to affect the amount of DNA input into each reaction.

Validation of full multiplex qPCR reaction containing IPC, ND5, and TH01 assays

An experiment was conducted in which data generated from singleplex ND5 and TH01 assays (using modified primer/probe concentrations where applicable) was compared to data from a full multiplex experiment in which all assays were combined. Line statistics for both targets quantified in singleplex and multiplex format were the same. Quantitation values of HL60 positive controls were also found to be the same (table 7). This data shows that multiplexing independent assays using the experimentally derived primer/probe concentrations does not affect assay fidelity

Table 7: *Quantitation of human nuclear and mitochondrial DNA obtained from optimized singleplex and multiplex qPCR reactions.*

			Singleplex				Multiplex (with IPC)			
			Rep 1	Rep 2	Rep 3	Mean	Rep 1	Rep 2	Rep 3	Mean
mtDNA	HL60 20 pg/µL	copies/µL	22.45	22.36	22.33	22.38	22.83	22.85	22.86	22.85
		C_T	6642.17	7045.35	7188.50	6958.67	7062.64	6955.67	6752.66	6923.66
	HL60 100 pg/µL	copies/µL	19.41	19.61	19.61	19.55	20.20	20.13	20.18	20.17
		C_T	49694.43	43700.85	43556.36	45650.55	41417.85	43566.62	40733.05	41905.84
nucDNA	HL60 20 pg/µL	ng/µL	31.25	31.34	31.20	31.26	31.38	31.48	31.35	31.40
		C_T	0.11	0.11	0.12	0.11	0.09	0.13	0.14	0.12
	HL60 100 pg/µL	ng/µL	28.55	28.53	28.42	28.50	28.93	28.75	28.92	28.87
		C_T	0.66	0.68	0.72	0.69	0.68	0.77	0.68	0.71

Table 7: Statistical analysis of quantitative data for the human mtDNA target shows that there is no difference between means of technical replicates (student t-test: p-value of HL60 20 pg/µL = 0.900; p-value of HL60 100 pg/µL = 0.250). The same is true for human nucDNA quantitative data (student t-test: p-value of HL60 20 pg/µL = 0.731; p-value of HL60 100 pg/µL = 0.630). This data suggests that both assays can be multiplexed with an IPC without any apparent affect on assay performance.

Quantitation values obtained using the optimized multiplex were also compared to values obtained from the same samples using other well-established assays. Data is shown in tables 8A and 8B.

Tables 8A and 8B: *Comparison of quantitative data obtained using the optimized multiplex assay to other well established assays.*

Sample ID	Singleplex ND5	Multiplex ND5	Kavlick mtDNA qPCR Assay	ddPCR™
	copies/µL			
HL60 20	6,958.67	6,923.66	9,120	8,092
HL60 100	45,650.54	41,905.84	53,703	Theoretical = 40,460
NTC	0.39	1.01	29.627	0.88

Table 8A: Control samples were quantified in triplicate using singleplex and multiplex optimized qPCR assays. The same samples were also quantified using ddPCR™ as well as a human specific mtDNA qPCR assay that has been well established in our laboratory. Quantitative data varies depending on the assay used. qPCR assays rely on the accuracy of the starting concentration of a standard. It is likely that results vary because the optimized assay described herein employs use of a different standard than the existing method. Conversely, droplet digital™ PCR is an absolute quantitation method that does not include use of a standard curve. Data generated using this method is often highly accurate and precise, so is likely a truer representation of the quantitative value of the samples tested. Additional work is being done to resolve the differences.

Sample ID	Singleplex TH01	Multiplex TH01	Trio – small autosomal	Trio – large autosomal	Trio – Y
	ng/µL				
HL60 20	0.113	0.136	0.103	0.096	0
HL60 100	0.688	0.709	0.56	0.492	0
NTC	0	0	0	0	0

Table 8B: Control samples were quantified using singleplex and multiplex optimized qPCR assays. The same samples were also quantified using the Life Technologies Quantifiler® Trio kit. Quantitation values obtained using the optimized multiplex assay are slightly higher than values obtained using Quantifiler® Trio. Again, these differences could be a result of the difference standards used to generate standard curves. Additional work is being done to resolve these differences as well.

The optimized multiplex was also run on the BioRad QX200™ ddPCR™ instrument to verify that the new primer/probe concentrations were sufficient for digital quantitation. Data is shown in tables 9A and 9B.

Table 9A: *Singleplex and multiplex ddPCR™ quantitation of ND5 target*

ND5 Target						
	Singleplex			Multiplex		
	Average (copies/µL)	Standard deviation	Stock Concentration (copies/µL)	Average (copies/µL)	Standard deviation	Stock Concentration (copies/µL)
NTC	0.00	0.00	0.00	0.22	0.26	0.88
5 ng/µL	340100.00	571491.53	1360400.00	340166.67	571432.92	1360666.67
0.5 ng/µL	4026.67	25.17	16106.67	3840.00	120.00	15360.00
0.05 ng/µL	447.67	22.81	1790.67	379.67	12.06	1518.67
5 pg/µL	51.60	6.75	206.40	42.17	2.23	168.67
0.5 pg/µL	3.77	0.29	15.07	4.90	2.08	19.60
50 fg/ µL	0.51	0.17	2.05	1.17	0.74	4.69
5 fg/ µL	0.38	0.36	1.53	0.92	0.94	3.67
HL60 20	2463.333333	71.14	9853.33	2023.00	7.00	8092.00
HL60 100	1000000	0.00	4000000.00	4536.67	464.36	18146.67

Table 9A: Samples were quantified on the BioRad QX200™ ddPCR™ instrument with optimized singleplex and multiplex assays. Standard deviations of higher concentration technical replicates are low. Samples with very high target copy number (5 ng/µL and HL60 100 pg/µL) did not quantify due to an absence of negative droplets. In cases such as these, a maximum quantitative value of 1,000,000 copies/µL is assigned to the sample by the software. A single-sample student t-test was conducted which showed no significant difference between singleplex and multiplex quantitations (p value = 0.16).

Table 9B: *Singleplex and multiplex ddPCR™ quantitation of TH01 target*

TH01 Target						
	Singleplex			Multiplex		
	Average (copies/µL)	Standard deviation	Stock Concentration (copies/µL)	Average (copies/µL)	Standard deviation	Stock Concentration (copies/µL)
NTC	0.00	0.00	0.00	0.00	0.00	0.00
5 ng/µL	358.33	12.34	1433.33	329.00	9.54	1316.00
0.5 ng/µL	36.73	1.62	146.93	34.50	3.87	138.00
0.05 ng/µL	4.20	0.53	16.80	3.13	0.55	12.53
5 pg/µL	0.53	0.28	2.11	0.40	0.11	1.61
0.5 pg/µL	0.03	0.05	0.11	0.08	0.08	0.31
50 fg/ µL	0.06	0.05	0.23	0.06	0.10	0.23
5 fg/ µL	0.03	0.06	0.13	0.00	0.00	0.00
HL60 20	9.67	0.61	38.67	9.20	0.10	36.80
HL60 100	46.70	0.72	186.80	43.67	5.95	174.67

Table 9B: Samples were quantified on the BioRad QX200™ ddPCR™ instrument with optimized singleplex and multiplex assays. As expected, samples with low concentrations of input DNA (0.5 pg/µL - 5 fg/µL) did not appear very different from the NTC. Samples that were within the quantitative range of the instrument, yielded values that were similar to the theoretical nucDNA copy number calculated based on the reaction input. A single-sample student t-test was conducted which showed no significant difference between singleplex and multiplex quantitations (p-value = 0.24).

Conclusions

These data show that the optimized multiplex assay is robust and produces results that are highly correlated with those obtained for the same assays run in singleplex. Additionally, quantitative information from the multiplex yields results that align with data generated using other well-established methods. This assays offers analysts a tool that enables simultaneous quantitation of nuclear and mitochondrial DNA in a sample extract, which reduces the amount of extract, consumed. Additionally, use of a

streamlined quantitation assay has the potential to reduce analyst labor and per sample costs overall.

The commercial supplier of our oligos, IDT, has published a technical report describing an oligonucleotide stability study they conducted, which can be found here: http://www.idtdna.com/pages/docs/default-source/technical-reports/stability-guidance-external_final.pdf?sfvrsn=2. These results indicate that the quantitative standards will degrade over time, resulting in inaccurate quantitative estimates, as we have observed in our studies.

We require consistency across real-time runs to keep the ratios comparable to the efficiency of the extraction method, since most of our studies are comparative in nature. In other words, we are looking for the copies of mtDNA obtained when employing one extraction method versus another. In this case, the ratio of DNA obtained using one method versus another is important, rather than the exact copy number obtained. However, when we begin to examine the minimum copy number required to support successful amplification of our advanced multiplexing strategies using qPCR, the concentration of the standard becomes more important. Hence, we are transitioning into using ddPCR™ exclusively as a quantitative method in order to obviate the necessity of a stable quantitative standard. Since ddPCR™ uses absolute quantification without the necessity of a standard, we have chosen to emphasize this method in our quantitative assessments of mtDNA copy number when assessing different extraction and post-amplification-based methods.

ddPCR™ for QC analysis of Illumina® next-generation sequencing (NGS) libraries

NGS methods are quickly being adopted by the forensic community for analysis of precious evidentiary samples. These methods are capable of generating an unprecedented amount of data, particularly when analysis is performed using commercially available highly multiplexed panels designed to target hundreds of loci per amplification. Accurate qualitative and quantitative assessment of prepared NGS libraries is of paramount importance for obtaining maximum yield of high-quality data from a sequencing run. Many vendor recommended protocols suggest assessment of the final library using fluorometric methods, agarose gel or chip-based electrophoresis, or quantitative PCR (qPCR). However, these methods can be problematic because they often result in over/underestimation of library DNA concentrations, do not enable estimation of the size of DNA fragments in the library, which can lead to incompatible kit selection, and are not typically specific for those fragments that are NGS ready.

Previous literature has shown that droplet florescence intensity when using ddPCR™ is dependent upon the average length of the fragments being assessed.[3] Longer fragments tend to result in lower droplet fluorescence intensities than shorter fragments due to the kinetics and stoichiometry of the PCR reaction. As a result of this observed fluorescence intensity:fragment size correlation, Laurie et. al. have developed a ddPCR™ assay that allows for simultaneous quantitative and qualitative assessment of NGS libraries (Laurie, 2013). The assay specifically targets NGS platform specific library modifications (i.e. flow cell adapter sequences) to enable quantification of only those

fragments that are sequenceable. The assay also includes use of a series of size standards to facilitate estimation of average length of fragments in the prepared library. However, the standards are derived from a commercially available agarose gel electrophoresis ladder and reported preparation is time-consuming, labor intensive, and may give rise to low-level contamination evident in NGS data.

Here, we report an optimization of the aforementioned assay using synthetically prepared size standards. Initially, a series of oligonucleotides with known sizes ranging from 25 – 700 bp was designed to consist of PhiX DNA with Illumina® MiSeq™ sequencing primer flanking regions. The oligonucleotides were designed using PhiX to reduce possible run contamination from exogenous sources. PhiX is supplied for use as an MPS control, and data generated from any part of the PhiX genome is easily identified and bioinformatically filtered from raw data. The sequencing primer region serves as a primer binding site to allow for additional incorporation of barcoding indices and flow cell adapters into the synthetic oligonucleotide during a limited cycle PCR step. Final products are then normalized for reaction input of 10,000 copies to avoid fluorescence intensity variability due to copy number and not length. Average observed droplet fluorescence intensities range from 13,157 RFU ($^{+/-}$ 203.3) for the 25 bp standard to 3,860.4 RFU ($^{+/-}$352.3) for the 700 bp standard (table 10, figure 3). The standard series appears to be efficient at predicting the average size of Illumina® MiSeq™ libraries while avoiding quantification of adapter dimers and other artifacts often generated during library preparation. This increases first-pass Illumina® MiSeq™ run success.

Figure 3: *Size standard schematic.* Double stranded DNA fragments will sizes ranging from 25-700 bp were designed. Each fragment consisted of PhiX bacteriophage DNA flanked by Illumina® sequencing primer (SP) sites. Limited cycle PCR was then conducted to incorporate barcoding indices and adapters as shown below.

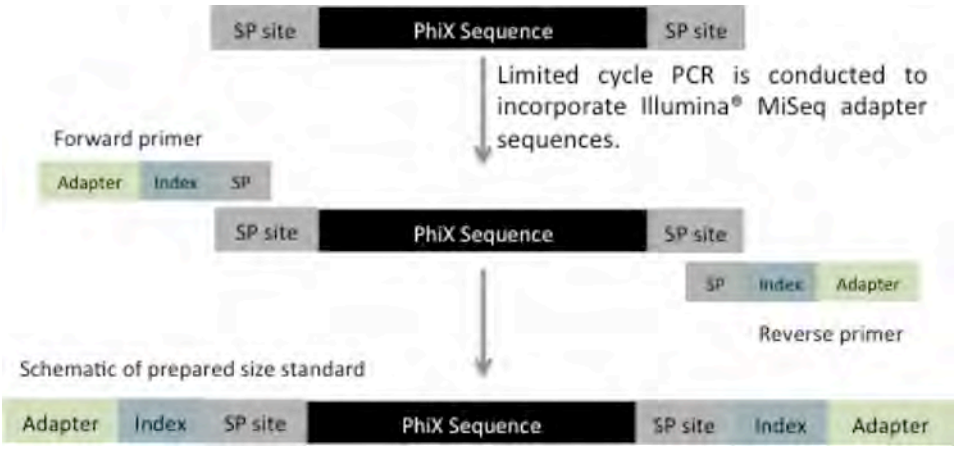


Figure 4: *Standard curve for fragment size estimation using ddPCR™*

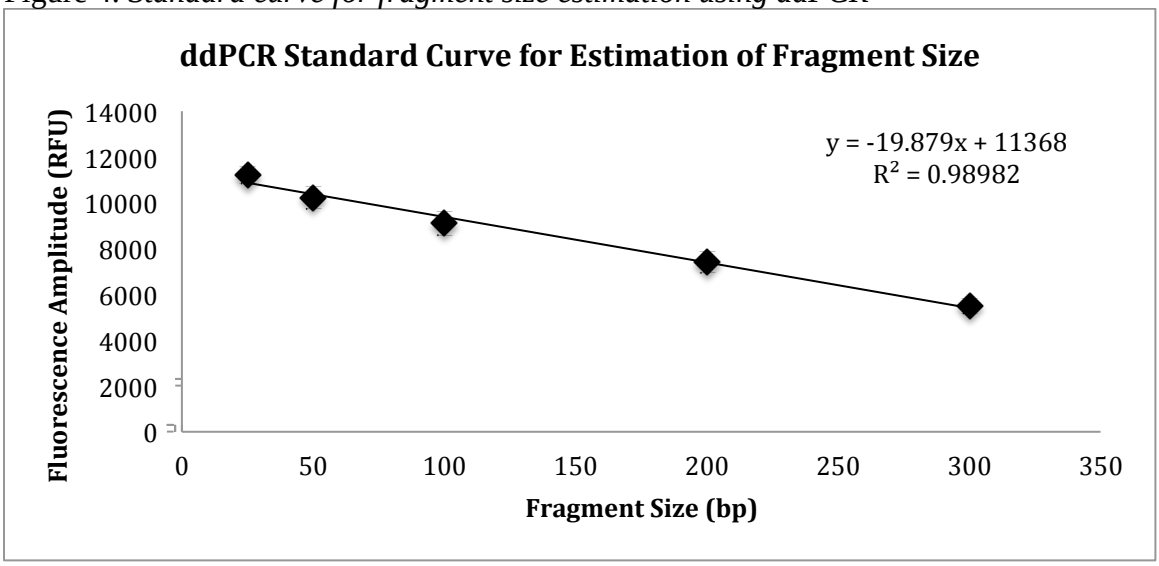


Figure 4: The figure shows a linear relationship between fragment size and fluorescence intensity value. Prepared standards were normalized to a concentration of 2,000 copies per microliter for a total ddPCR™ input of 10,000 copies. Standards were analyzed using a ddPCR™ assay designed with primers complementary to Illumina® adapter sequences and a probe complementary to the sequencing primer region. Primers and probe were designed by Laurie et. al. As expected, droplet intensities decrease as fragment size of standard increases. This trend does not appear to extend to longer fragments ranging from 500-700 bp in length. Additional work is needed to elucidate the reason for this.

Figure 5: *One dimensional amplitude plot of fluorescence positive and negative droplets of synthetic size standards.*

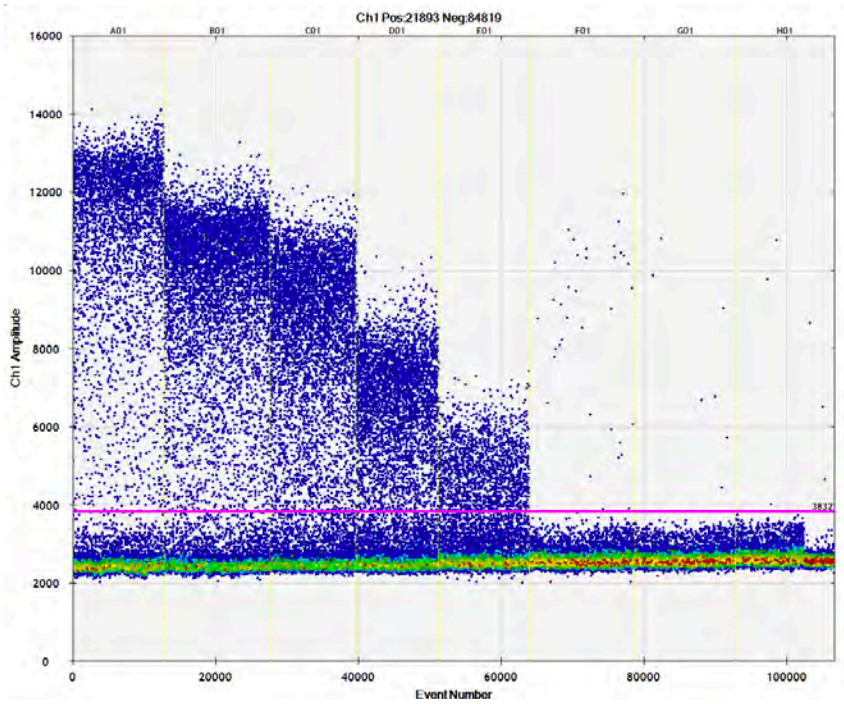


Figure 5: Standard size range from 25 bp (far left) to 700 bp (far right). Standards ranging from 25-300 bp produce obvious differences in fluorescence amplitude (y-axis), which is due to fragment size and not copy number variability since standards were normalized prior to analysis. Shorter fragments yield higher average fluorescence values due to lower overall consumption of dNTPs and reduced production of the PCR inhibitor pyrophosphate compared to longer fragments.

Control gDNA (2800M, Promega) was prepared for NGS using the Nextera® XT kit. Tagmentation was performed on eight replicates of the control sample, and resulting libraries were cleaned using Agencourt AmPure XP beads (Beckman Coulter, Indianapolis, IN). Double stranded libraries were quantified using both the aforementioned ddPCR™ assay and the Agilent 2100 Bioanalzyer. Average library size and quantification using each method is shown in table 10. Each library was then normalized using the information obtained with both quantitative methods, and each set of normalized libraries was sequenced on the Illumina® MiSeq using a 2 x 300 v3 run kit. Resulting data was analyzed using CLC Genomics Workbench v8. The number of sequences per library was compared for samples from each treatment (normalization using Bioanalzyer data or ddPCR™ data). This comparison is outlined in figure 6. Additionally, an analysis was performed to determine the average fragment size of each library. This data is shown in figures 7a and 7b.

Table 10: *Summary of quantitative and qualitative assessment of prepared MPS libraries.*

	Agilent 2100 Bioanalzyer Data		ddPCR™ Data	
Library ID	Average Size (bp)	Concentration (nM)	Average Size (bp)	Concentration (nM)
A	794	19.1	197	7.56
B	751	20.2	215	9.45
C	861	14.1	229	5.07
D	787	11.4	220	3.96
E	709	853.9	217	6.07
F	689	640.8	238	7.69
G	757	16.2	237	6.70
H	751	18.9	216	9.31

Table 10: Average library fragment size predicted when using the Agilent 2100 Bioanalzyer was significantly high than when using the optimized ddPCR™ assay. In addition, the average concentration of each library estimated using the Bioanalzyer was higher than when ddPCR™ was used for quantification. Quantitation information from each assay was used to normalize libraries for NGS on the Illumina® MiSeq.

Figure 6: *Total number of sequences generated per library.*

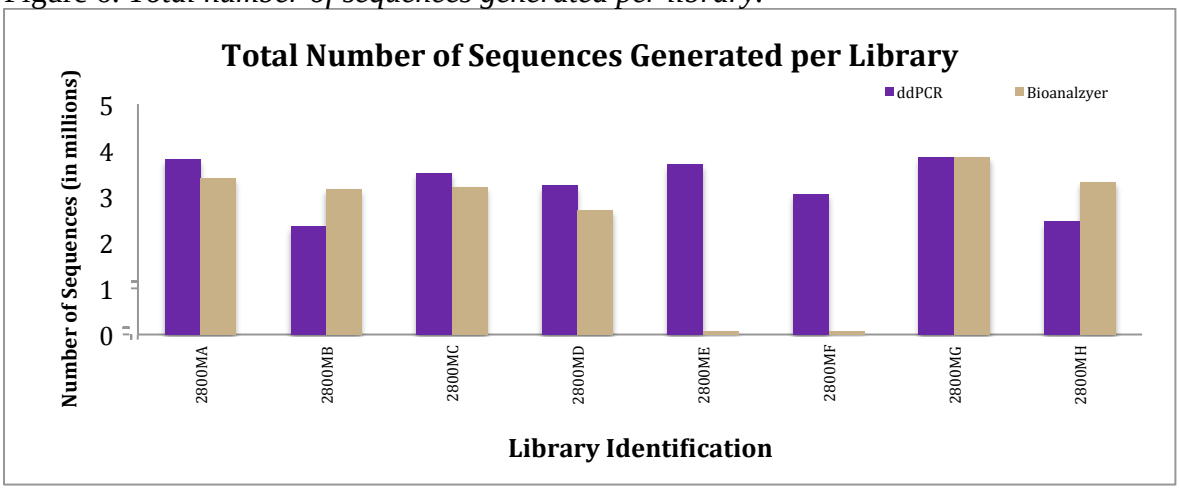
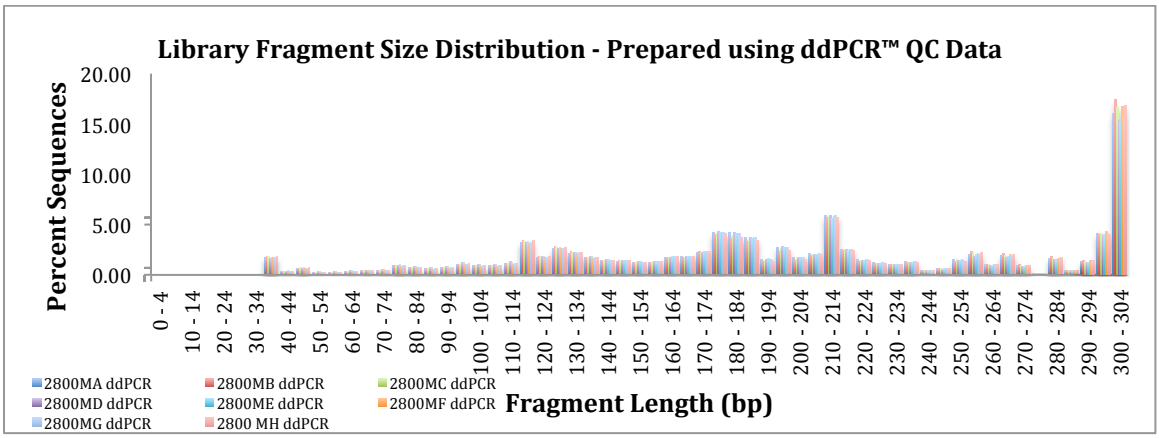


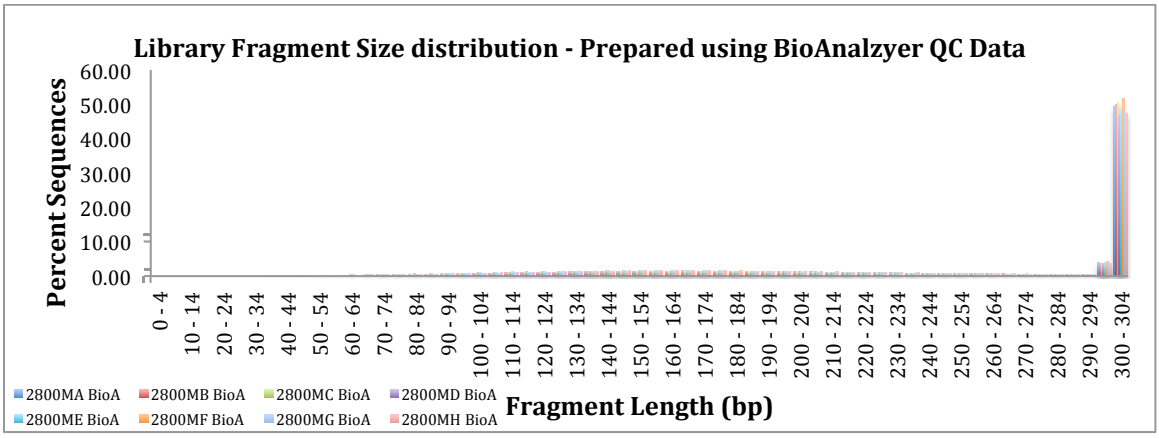
Figure 6: On average, the number of sequences generated per library prepared with ddPCR™ QC data exceeded the number of sequences generated for the same library prepared with BioAnalzyer data, except for libraries 2800MB and 2800MH. In samples 2800ME and 2800MF, very few sequencing reads were generated when the libraries were diluted using Agilent 2100 Bioanalzyer QC data. This suggests that ddPCR™ yields quantitation data more suitable for consistent library preparation.

Figures 7a (above) and 7b (below): *Fragment size distributions for libraries prepared using ddPCR™ and Bioanalzyer QC data respectively.*

7a



7b



Figues 7a and 7b: In general, fragment sizes reported for libraries prepared using ddPCR™ were smaller than those prepared using Bioanalyzer data. However, fragment sizes should not differ since the libraries were identical, differing only by dilution factor. It is possible that this is a result of MPS run quality. Further studies are warranted to elucidate the reason for these differences.

Conclusions

The optimized ddPCR™ method described is suitable for accurate quantitation of double stranded libraries prepared for sequencing on the Illumina® MiSeq. Representation of each multiplexed library is consistent when using ddPCR™ data for normalization. Libraries normalized using Agilent 2100 Bioanalzyer data were relatively evenly represented, however, two of the eight replicates yielded very low read counts leading to areas of zero coverage in the genome after alignment. Additionally, NGS data obtained from libraries normalized with ddPCR™ data showed a higher number of sequences per sample overall when compared to data from samples normalized using Bioanalzyer data. No conclusions can be made regarding the ability of the optimized ddPCR™ assay to estimate library fragment size as fragment size distributions for each data set were so different. Since libraries were identical differing only by the dilution factor used to normalize the samples prior to sequencing, fragment size distributions should be the same.

Long PCR (LPCR) Amplification

Primer design and PCR Amplification

We have developed a long PCR assay that employs a combination of a highly processive *Taq* polymerase and a proofreading enzyme with 3'-5' exonuclease activity. This enzyme combination has been used to generate amplicons of 25 kb and upwards (Goto, 2006). Two primer sets were designed to amplify the entire mtGenome in two reactions (table 11). The resulting amplicons overlap at the HV regions, in an attempt to increase sequence coverage in these areas.

Table 11: *Primer information for LPCR amplification of whole mtGenome*

Amplicon Size	rCRS 3' position	Primer ID	Primer Sequence
9,065 bp	9416	1F	5' AAA GCA CAT ACC AAG GCC AC 3'
	1873	1R	5' TTG GCT CTC CTT GCA AAG TT 3'
11,170 bp	9777	2F	5' TAT CCG CCA TCC CAT ACA TT 3'
	15214	2R	5' AAT GTT GAG CCG TAG ATG CC 3'

DNA from eight donors as well as negative and positive controls (1 ng HL60 DNA) was amplified using both primer sets. Approximately 200,000 copies of mtDNA template were added to PCR master mix containing 0.2 µM forward and reverse primers, 1X PCR buffer, 0.4 mM each dNTP, 0.05 U/µl enzyme blend, and sterile water to a total volume of 50 µL. DNA was amplified on an Applied Biosystems® Veriti® 96-Well thermal cycler as follows: 94°C for 1 min, 30 cycles of 94°C for 30 sec, 54°C for 15 sec, 68°C for 11 min, followed by 72°C for 10 min and a 4°C hold.

After amplification, the LPCR products were quantified using the Agilent Technologies® 2100 Bioanalyzer® with the Agilent Technologies® DNA 12000 Kit™ which quantifies DNA fragments of 100 - 12,000 bp in size (Agilent Technologies, Santa Clara, CA). Reactions were purified with the Zymo® Clean & Concentrator-5™ kit (Zymo Research, Irvine, CA) and requantified with the Agilent Technologies® DNA 12000 Kit™.

Sanger sequencing of LPCR product

Sanger sequencing reference data was successfully obtained for all eight donors with the Applied Biosystems™ mitoSEQr™ kit (Applied Biosystems™, Foster City, CA). For each of these donors, LPCR amplifications from buccal swabs extracts generated 6.6 ng/µl of PCR product on average (Table 12).

Table 12: *Efficiency of LPCR amplification on buccal extracts*

Donor	Copies of mtDNA in Buccal Swab Extract	LPCR Input (Copies of mtDNA)	Average LPCR Product (ng/µl)
001	16,998,840,000	226000	5.35
002	62,612,828	208709	6.78
003	33,251,937	443359	10.52
006	18,411,570,000	246000	7.23
009	5,940,112,500	198000	5.30
015	1,037,101,905	230467	8.66
020	148,382,018	197843	7.34
021	54,837,990,000	183000	5.42

Table 12: Average LPCR product is calculated as the average of the long and short amplicon per donor. A higher input was used for donor 003 because 1 ng of nuclear DNA was targeted for this amplification.

Note: DNA extraction was performed in three batches of two - three donors at a time, with a separate reagent blank created for each batch. No LPCR amplification was observed for any reagent blank.

Illumina® Nextera® XT library preparation and NGS of LPCR products

In an attempt to eliminate coverage bias due to the higher prevalence of the smaller amplicon in comparison to the larger amplicon, the 11.1 kb reactions were diluted to 200 pg/µl and the 9.1 kb reactions were diluted to 162 pg/µl with molecular biology gradewater. From each donor, 2.5 µl of each normalized long amplification product was pooled for Illumina® Nextera® XT (Illumina®, San Diego, CA) library preparation. All reagent blanks were pooled undiluted. Tagmentation was performed on an Applied Biosystems® Veriti® 96-Well thermal cycler (Thermo Fisher Scientific, Waltham, MA). Resulting fragmented libraries were assessed for quality and quantity using the Agilent Technologies 2100 Bioanalyzer and DNA 1000 Kit. Each sample was assigned a unique index combination for sample identification and data parsing post-run. Indices and required sequencing adapters were incorporated during a limited-cycle PCR amplification on the Veriti® thermal cycler. Prepared libraries were then purified with Agencourt® AMPure® XP beads (Beckman Coulter, Inc., Indianapolis, IN). Clean libraries were normalized with Nextera® XT magnetic beads. The normalized libraries were quantified with the Qubit® ssDNA Assay kit (Life Technologies™, Carlsbad, CA) and pooled to create the final library. Illumina® PhiX v3 sequencing control (Illumina®, San Diego, CA) was spiked into the library at a 20% v/v ratio. The library was then diluted 25 fold and was sequenced on the Illumina® MiSeq® in a 2x150 bp paired-end v2 run. Sequencing analysis was performed with Illumina® Sequence Analysis Viewer (SAV) 1.8, Illumina® MiSeq™ Reporter (MSR) 2.2 (Illumina®, San Diego, CA) and Integrative Genomics Viewer (IGV) 2.2 and 2.3 (Broad Institute, Boston, MA). The resulting NGS sequences were compared to those derived from the same donors using Sanger sequencing, and positions that did not show a common base in this comparison of treatments were marked as sequence differences.

Sequencing Results for LPCR Products

Illumina® MiSeq™ run quality metrics were in line with Illumina® guidelines. Whole mtGenome data was obtained for all 8 donors. Coverage depth across the genome was variable, as seen in Figure 4.

Figure 4: *Whole mtGenome coverage data for donor 002 obtained from MiSeq™ Reporter*

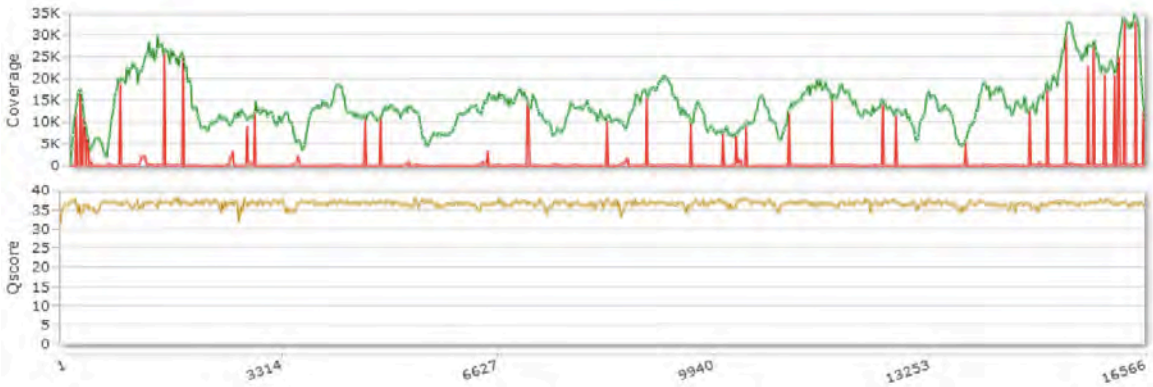


Figure 4: Top: read coverage across the mtGenome. Bottom: Quality scores. A score of Q30 or higher is considered desirable.

On average, the MSR analysis showed a sequence coverage of 13,072 reads across the whole mtGenome. The NGS data revealed a range of 11 to 41 variants from the rCRS outside of the HV regions, an average of 26 per donor. The median fragment length across all donors was 265 bp, which is consistent with the Agilent Technologies® 2100 Bioanalyzer® size distributions of the Illumina® Nextera® XT libraries.

Table 13: *Variants from the rCRS in NGS and Sanger sequencing data from donor 002*

Sanger			Illumina® MiSeq™					Sanger			Illumina® MiSeq™				
Pos	rCRS	Var	Pos	Type	Call	Freq	Depth	Pos	rCRS	Var	Pos	Type	Call	Freq	Depth
73	A	G	73	SNP	A->AG	100	11289	8,860	A	G	8860	SNP	A->AG	100	15077
152	T	C	152	SNP	T->TC	100	16694	9,548	G	A	9548	SNP	G->GA	100	9238
199	T	C	199	SNP	T->TC	100	10632	10,034	T	C	10034	SNP	T->TC	100	7260
204	T	C	204	SNP	T->TC	100	9558	10,238	T	C	10238	SNP	T->TC	100	6587
207	G	A	207	SNP	G->GA	100	9341	10,398	A	G	10398	SNP	A->AG	100	8828
250	T	C	250	SNP	T->TC	100	5959	11,065	A	G	11065	SNP	A->AG	100	11494
263	A	G	263	SNP	A->AG	100	4512	11,719	G	A	11719	SNP	G->GA	100	14204
309.1	:	C	302	Indel	-/C	91	1755	12,501	G	A	12501	SNP	G->GA	100	8863
315.1	:	C	310	Indel	-/C	100	2043	12,705	C	T	12705	SNP	C->CT	100	11234
573.1	:	C	567	Indel	---/CCC	49	1781	13,780	A	G	13780	SNP	A->AG	100	4520
750	A	G	750	SNP	A->AG	100	18207	14,766	C	T	14766	SNP	C->CT	100	12300
1,438	A	G	1438	SNP	A->AG	100	25567	15,043	G	A	15043	SNP	G->GA	100	16826
1,719	G	A	1719	SNP	G->GA	100	24450	15,326	A	G	15326	SNP	A->AG	100	28723
2,706	A	G	2706	SNP	A->AG	100	6461	15,673	A	G	15673	SNP	A->AG	83	26461
2,835	C	A	2835	SNP	C->CA	100	11764	15,758	A	G	15758	SNP	A->AG	100	26543
3,107	N	:	3106	Indel	N/-	94	10710	15,924	A	G	15924	SNP	A->AG	100	20390
4,529	A	T	4529	SNP	A->AT	100	10163	16,074	A	G	16074	SNP	A->AG	100	20066
4,769	A	G	4769	SNP	A->AG	100	11051	16,129	G	A	16129	SNP	G->GA	99	23467
7,028	C	T	7028	SNP	C->CT	99	13846	16,145	G	A	16145	SNP	G->GA	100	24327
7,055	A	T	7055	SNP	A->AT	100	12759	16,223	C	T	16223	SNP	C->CT	99	31446
8,251	G	A	8251	SNP	G->GA	100	9854	16,391	G	A	16391	SNP	G->GA	100	31781
8,843	T	T	*8843	SNP	T->TC	2	16098	16,519	T	C	16519	SNP	T->TC	100	11915

Table 13: Data was analyzed with MSR. Yellow: common base between Sanger and NGS analysis; Pink: low-level mixed position confirmed with Sanger sequencing; Blue: low-level mixed position in homopolymer region.

NGS has enhanced capability to enable deconvolution of sequence mixtures. For example, data from donor 001 shows a known low-level mixed position of approximately 8% at position 16,093. However, due to lower resolution achieved when using dye-terminator chemistry, these low-level variants might go undetected when using Sanger-type sequencing. Therefore, when comparing treatments (NGS and Sanger sequencing) any discrepancies may be the result of differences in limit of detection when using the

two methods. For example, in the NGS data from donor 002 shown in Table 13, an "A" was observed at position 15,673 with a frequency of approximately 17%. Upon revisiting the Sanger electropherograms for this donor, a mixed base at position 15,673 was observed that was overlooked during initial analysis. Ultimately, Sanger data was amended to include this finding. This observation underscores the value of decreasing the limit of detection of minor variants with the use of NGS, and also illustrates the potential use of NGS in mixture deconvolution in forensic casework.

It should be noted that some bioinformatics software packages have limitations in aligning small insertions and deletions (indels). These limitations may result in multiple variant calls for a single indel. Therefore, misalignments and small indels in NGS data were omitted from the analysis results in this study.

Conclusions

High concentrations of intact DNA was obtained from fresh buccal swabs. DNA concentrations were normalized for mtDNA input of 200,000 copies per amplification reaction. This input amount resulted in a yield of 6 ng/µl of LPCR product on average. llumina® Nextera® XT requires a total input of 1 ng, or 5 µl of 0.2 ng/µl sample. Thus, the described LPCR approach was successful in amplifying the entire mitochondrial genome from buccal swabs from all eight donors in this study for downstream NGS.

Occasionally, LPCR amplification failed when DNA was extracted from buccal cells on untreated cotton swabs that were dried and stored at room temperature, yet no such problems were encountered with fresh buccal swabs in this study. Perhaps ongoing microbial activity resulting in DNA degradation occurred in these stored swabs. It is recommended that DNA extraction be performed on fresh buccal swabs, or on swabs treated with antimicrobial compounds. Alternatively, buccal cells could be transferred to FTA® cards to prevent microbial degradation of DNA and to enable room temperature sample storage for extended periods of time (Whatman®, St. Louis, MO).

NGS data from LPCR products shows high depths of coverage across the entire mitochondrial genome, with an average depth of 13,000 reads across all donors. Based on the amplification design, double sequence coverage was expected for the areas between nucleotide positions 15,195 – 1,892, and for 9,397 – 9,777, as the LPCR primer sets overlap in these regions. An increase in coverage in these regions facilitates even deeper detection of low-level variants, particularly in the highly variable non-coding region. Higher coverage was observed in these regions, although in some instances the coverage may have been artificially lowered since a non-circularized genome was used for mapping due to software limitations.

NGS data derived from LPCR amplicons is concordant with donor reference sequences obtained using Sanger methods. However, low-level variants were detected in NGS data sets that were not originally detected in Sanger data. For instance, an 8% C to T transition in donor 001 at position 16,093 was observed in NGS data. This position is a known "hot spot" known to have a high mutation rate. In addition, sequencing the whole mtGenome enabled detection of up to 41 additional variants outside of the traditionally

sequenced HV region, which provides the forensic analyst with more genetic data for comparison ultimately increasing the discriminatory power of mtDNA.

Currently, indels may present a limitation in NGS analysis, as read mapping algorithms have been known to misalign reads containing indels. However, mapping algorithms have improved over the past years, and we will likely see more improvements in the future. Although small indels were ignored in our analyses, it was noted that a low-level mixed base at position 12,417, which is located in a region with eight adenines, is seen consistently at a frequency of approximately 4% in the NGS data from all donors, and across different runs. This may indicate that sequencing through this homopolymer using the Illumina® chemistry results in an artifact that appears as an indel in raw data. Alternatively, this may simply be the result of an alignment issue that happens consistently with the data analysis software. Although homopolymer regions are known to be heteroplasmic, clearly more research is needed to more fully understand this observation.

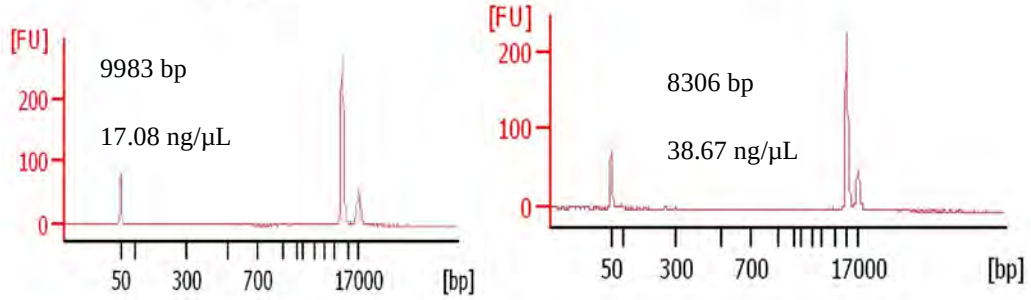
Preparation of Reference Samples for NGS using LPCR and Direct Amplification

We have optimized a direct amplification method for rapid databasing of whole mtGenome sequence data. Initially, buccal cells are transferred from a nylon FLOQSwab™ (Copan Diagnostics, Inc., Brescia, Italy) to a treated FTA® card for archival of samples. The protocol for purification of DNA on FTA® cards requires several washes for removal of PCR inhibitors. This lengthy process renders the rapid generation of full mtGenome sequence data tedious. In our method, the purification process has been omitted from the workflow, and amplification of the entire mtGenome is achieved in two PCR reactions, where amplicons of 9 and 11 kb overlapping at the HV region are generated. A 1.2 mm FTA® punch is added directly to PCR master mix containing an enhancer cocktail that reduces the effects of inhibitory compounds.

Amplification of buccal cells on FTA® paper following Whatman® protocol

Initially, buccal cells were transferred from Copan FLOQSwabs™ to treated FTA® classic cards. A Harris micropunch was used to obtain 1.2 mm punches from the FTA® cards, and punches were washed according to manufacturers protocol. Resulting purified DNA was amplified using primer pairs targeting 9 and 11 kb regions of the human mtGenome. Amplification products were assessed using the Agilent 2100 Bioanalyzer and DNA 12000 kit. This experiment was conducted to verify that our approach was suitable for amplifying DNA on FTA® paper. Sample Bioanalyzer data can be seen in figures 5A and 5B.

Figures 5A (left) and 5B (right): *Bioanalyzer data showing successful amplification of 9 and 11 kb mtDNA targets from DNA stored on treated FTA® paper.*



Figures 5A (left) and 5B (right): FTA® punches (1.2 mm) were purified according to Whatman®, and DNA was amplified. Figure 5B shows a peak corresponding to the 11 kb amplicon, and figure 5A shows a single peak corresponding to the 9 kb amplicon.

Direct amplification of buccal cells on FTA® paper

In a second experiment, we attempted to amplify DNA using the strategy described above, however, the FTA® punches containing DNA were not purified. Two positive control samples were included in this experiment to enable detection of possible PCR inhibition by chemicals present on the unpurified FTA® card. One positive control contained 2800M control DNA (0.1 ng) added directly to PCR master mix with no FTA® punch. A separate positive control contained 0.1 ng of 2800M DNA (Promega, Madison, WI) and an unpurified FTA® punch. Standard reagent blanks and NTCs were also included. Robust amplification was observed for both sets of primer pairs in positive control samples lacking unpurified FTA® punches (figures 17 and 18). No amplification was evident in positive control samples containing FTA® punches, or for buccal samples on FTA® punches (figures 19 and 20). This data suggests that PCR inhibition is occurring.

Figures 6A-6D: *Agilent 2100 Bioanalzyer data for direct amplification of DNA on unpurified FTA® cards.*

Figure 6A

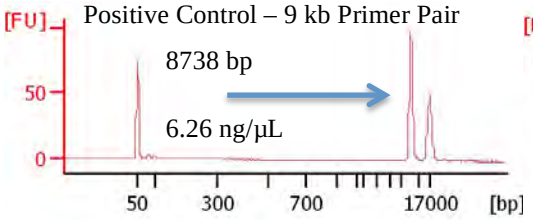


Figure 6B

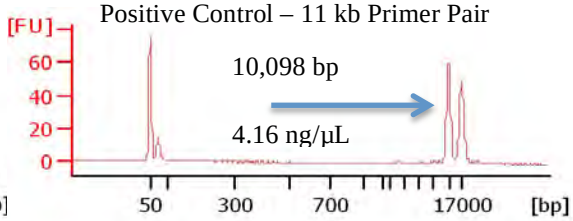


Figure 6C

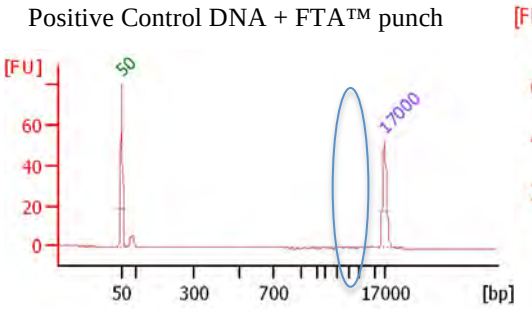
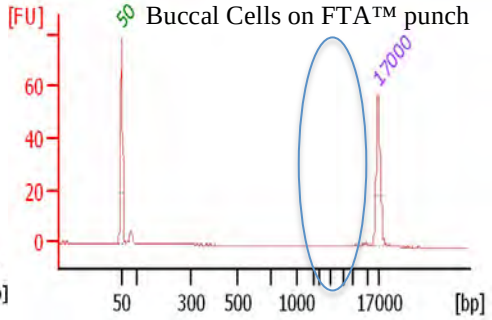


Figure 6D



Figures 6A-6D: Figures 6A and 6B show successful amplification of positive control DNA (0.1 ng 2800M DNA) in the absence of an FTA® punch. In figure 6C the same amount of control DNA was added to PCR reaction mix containing a neat 1.2 mm FTA® punch. No amplification is observed suggesting that chemicals on the FTA® punch are inhibiting amplification. The same result is seen in figure 6D where buccal cell DNA on an unpurified FTA® punch is not amplified.

Direct amplification including use of an enhancement cocktail to overcome inhibition of PCR by FTA® paper

A PCR enhancement cocktail that is designed to enable direct amplification of DNA while minimizing the inhibitory effects of chemicals on FTA® paper was utilized. The PCR enhancement reagent was added to PCR master mixes and DNA was amplified using the strategy described above. Assessment of amplification using the Agilent 2100 Bioanalyzer revealed that amplification was successful for all samples and positive controls. However, Bioanalyzer data also showed high levels of background noise and low amplification yields when using primer pairs for both 9 and 11 kb targets (figures 21 and 22). Amplification resulted in enough product for next-generation sequencing. Several other inhibition resistant engineered polymerase enzymes and PCR enhancement cocktails were also tested for their ability to increase amplification yields, and decrease background noise in Bioanalyzer traces. However, these enzymes and reagents did not result in yields as high as those obtained with the original combination.

Figures 7A and 7B: *Agilent 2100 Bioanalyzer traces showing peaks consistent with 9 and 11 kb amplicons from direct amplification of mtDNA using a PCR enhancement cocktail.*

Figure 7A

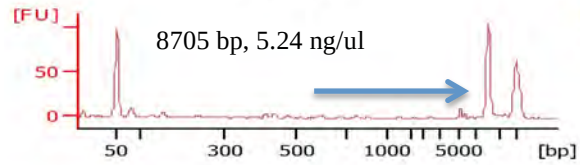
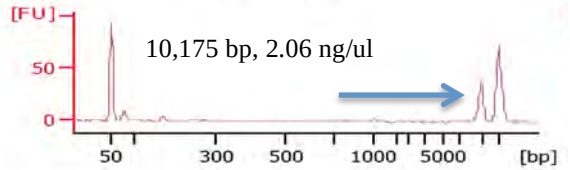


Figure 7B



Figures 7A and 7B: Direct amplification was performed on unpurified 1.2 mm FTA® card punches containing buccal cells transferred from FLOQswabs™. PCR enhancement cocktail was added to the PCR reaction mix to reduce inhibitory effects of chemicals on FTA® punches. Amplification yields were sufficient for downstream NGS whole

mtGenome sequencing.

Conclusions

We have developed a successful method to amplify the whole human mtGenome directly from buccal cells on treated FTA® punches using two overlapping PCR primer pairs that target 9 and 11 kb regions in two separate reactions. Use of a PCR enhancement cocktail reduces the effects of inhibitory chemicals introduced by treated FTA® paper. Yields from both reactions are suitable to support successful downstream NGS of the whole mtGenome.

Whole Genome Amplification (WGA)

Whole genome amplification (WGA) has been proposed as a promising method for increasing the template copy number of limited quantity DNA samples prior to traditional DNA profiling. Several methods have been developed for WGA of DNA including multiple displacement amplification (MDA) and PCR based techniques. While much of the focus of WGA research for forensic purposes has been in its ability to replicate nuclear DNA, WGA should be capable of copying nuclear and mitochondrial DNA in a representative fashion to produce large quantities of product for analysis. Therefore, we aim to investigate the ability of WGA to increase the sensitivity of downstream mtDNA analysis and also assess any sequence differences that may arise from the WGA process itself.

Evaluation of single cell WGA kits using robust samples

A study has been performed to assess the efficacy of two WGA kits – the REPLI-g Single Cell kit (QIAGEN, Valencia, CA) and the TruePrime™ Single Cell WGA kit (Sygnis™, Germany) – to amplify mtDNA from human bone samples. Commercial MDA kits like the REPLI-g Single Cell kit traditionally use a pool of random hexamer primers to prime the DNA template for isothermal amplification. However, the TruePrime™ WGA kit employs a DNA primase referred to as TthPrimPol in lieu of random hexamer primers to synthesize short DNA primers in situ that are complimentary to the DNA being amplified. Kits that use random hexamer primers have shown amplification bias of certain regions of the DNA. The ability to synthesize primers in situ could therefore assist with reducing amplification bias to provide more even and representative coverage of the entire mtDNA genome.

Buccal swabs and blood punches from FTA® paper were collected from two donors. DNA. Buccal swab DNA was extracted using the QIAGEN EZ1® DNA Investigator® kit (QIAGEN, Valencia, CA). DNA was extracted from FTA® blood punches following the Whatman™ protocol of the QIAamp® DNA Investigator Kit. DNA extracts were diluted five-fold to produce five dilutions. Neat DNA and corresponding dilutions then underwent WGA using the TruePrime™ single cell kit following a 6-hour incubation period. Mitochondrial DNA and nuclear DNA was quantified in neat DNA extracts and WGA amplified extracts using a mitochondrial DNA quantification assay[2] and Quantifiler™ Trio DNA Quantification kit (Thermo Fisher

Scientific, Waltham, MA) (tables 14-17). All qPCR reactions were performed in duplicate.

Table 14: *Quantification of mitochondrial DNA from buccal swabs pre- and post-WGA*

Sample Name	mtGenome copy # reaction input	mtGenome copy # reaction output (WGA yield)
003-neat	445,128.1	3,513,967.2
003-dil1	84,496.0	109,037.5
003-dil2	15,355.2	72,417.0
003-dil3	2,875.5	2,726.6
003-dil4	534.5	673.6
006-neat	273,697.9	196,669.2
006-dil1	50,590.5	37,049.0
006-dil2	9,631.6	7,401.3
006-dil3	1,658.6	1,767.6
006-dil4	360.4	471.4
RB	3.4	67.4

Table 14: Extracts from buccal swabs were serially diluted and amplified using the TruePrime™ single cell WGA kit. WGA appeared to work very well in cases where reaction inputs were high. However, as the input amount decreased, WGA fold differences decreased sometimes below the starting concentration.

Table 15: *Quantification of nuclear DNA from buccal swabs pre- and post-WGA*

	WGA input (ng)			WGA Yield (ng)		
	T.Small	T.Large	T.Y	T.Small	T.Large	T.Y
003-neat	4.22	0	0	1.69	0	0
003-dil1	0.86	0	0	0.01	0	0
003-dil2	0.13	0	0	0	0	0
003-dil3	0.03	0	0	0	0	0
003-dil4	0.01	0	0	0	0	0
RB	0	0	0	0	0	0
006-neat	3.36	0	0.0073	0.78	0	0
006-dil1	0.64	0	0.0017	0	0	0
006-dil2	0.10	0	0.0003	0	0	0
006-dil3	0.02	0	0	0	0	0
006-dil4	0.00	0	0	0	0	0

Table 15: DNA was extracted from buccal swabs using the QIAGEN EZ1® DNA Investigator® kit. The extract was then diluted five-fold to mimic concentrations often encountered in the crime laboratory

Table 16: *Quantification of mitochondrial DNA from blood samples pre- and post-WGA*

Sample Name	mtGenome copy # reaction input	mtGenome copy # reaction output (WGA yield)
003-neat	3308.1	3659.3
003-dil1	662.6	323.1
003-dil2	106.1	258.0
003-dil3	33.9	132.7
003-dil4	21.4	7.0
006-neat	2632.7	2514.6

006-dil1	544.9	627.2
006-dil2	114.0	75.4
006-dil3	27.4	103.3
006-dil4	8.0	50.8
RB	20.5	58.4

Table 16: Extracts from whole blood samples extracted from FTA® cards were serially diluted and amplified using the TruePrime™ single cell WGA kit. WGA did not appear to amplify mtDNA derived from whole blood samples very well.

Table 17: *Quantification of nuclear DNA from blood given in pre- and post-WGA*

	WGA input (ng)			WGA yield (ng)		
	T.Small	T.Large	T.Y	T.Small	T.Large	T.Y
003-neat	0.0155	0.0266	0	0	0	0
003-dil1	0.0034	0.0048	0	0	0	0
003-dil2	0.0004	0.0006	0	0	0	0
003-dil3	0.0001	0	0	0	0	0
003-dil4	0	0	0	0	0	0
RB	0	0	0	0	0	0
006-neat	0.0067	0.0127	0.0073	0	0	0
006-dil1	0.0022	0.0022	0.0017	0	0	0
006-dil2	0.0003	0.0001	0.0003	0	0	0
006-dil3	0	0	0	0	0	0
006-dil4	0	0	0	0	0	0

Table 16: DNA from whole blood samples extracted from FTA® cards was serially diluted and amplified using the TruePrime™ single cell WGA kit. No amplification of nuclear DNA was observed for small or large autosomal targets quantitated with the QuantiFiler™ Trio kit.

Evaluation of Sygnis® TruePrime™ WGA single cell kit with forensically relevant hair shaft samples

Hairs were collected from three separate donors. Each hair was examined microscopically and follicular tags, if present, were removed. DNA was then extracted using a lab developed solid-phase DNA extraction technique. DNA extracts underwent whole genome amplification using the TruePrime™ method with a 6-hour incubation period. Mitochondrial DNA and nuclear DNA was quantified in neat DNA extracts and WGA amplified extracts using a mitochondrial DNA quantification assay (Kavlick, 2011) and Quantifiler™ Trio DNA Quantification kit (tables 18 and 19). All qPCR reactions were performed in duplicate.

Table 18: *Quantification of mitochondrial DNA pre- and post-WGA*

Sample Name	mtGenome copy # reaction input	mtGenome copy # reaction output (WGA yield)
MPH_HS0620	7153.9	2238449600.0
JMM_HS0620	1559.6	2541.0
BB_HS0620	1743.3	1973.3
RB_0620	160.7	31.0

Table 18: DNA extracted from human hair shafts was amplified using the TruePrime™ WGA single cell kit. WGA products were quantified using a human mtDNA specific

qPCR assay. Amplification yields were inconsistent across samples obtained from three donors.

Table 19: *Quantification of nuclear DNA pre- and post-WGA*

	WGA input (ng)			WGA Yield (ng)		
	T.Small	T.Large	T.Y	T.Small	T.Large	T.Y
MPH_HS0620	0.003	0	0	0	0	0
JMM_HS0620	0	0	0	0	0	0
BB_HS0620	0	0	0	0	0	0
RB_0620	0	0	0	0	0	0

Table 19: DNA extracted from human hair shafts was amplified using the TruePrime™ WGA single cell kit. WGA products were quantified using QuantiFiler™ Trio qPCR kit No nuclear DNA amplification was evident. Results are not unexpected since hair shaft samples rarely yield nuclear DNA

Evaluation of Sygnis® TruePrime™ WGA and intra-donor hair shaft variation

Five hairs were collected from each of two donors and extracted using a lab developed solid-phase DNA extraction technique. DNA extracts underwent whole genome amplification via the TruePrime™ kit following a 6-hour incubation period. WGA reactions were performed triplicate on each hair. Mitochondrial DNA was quantified in neat DNA extracts and WGA amplified extracts using a mitochondrial DNA quantification assay (Table 20) (Kavlick, 2011) All qPCR reactions were performed in triplicate.

Table 20: *Quantification of mitochondrial DNA from hair given in copy number*

Sample Name	mtGenome copy # reaction input	mtGenome copy # reaction output (WGA yield) -1	mtGenome copy # reaction output (WGA yield) -2	mtGenome copy # reaction output (WGA yield) -3
MPH_HS1	9802.9	861867200.0	11872224000.0	3282048600.0
MPH_HS2	9682.1	3611244400.0	5298523600.0	5895456000.0
MPH_HS3	4591.1	4850.7	4526.4	64404.2
MPH_HS4	4626.5	4749.4	5608.3	9048.4
MPH_HS5	2588.6	2433.5	2224.4	2446.0
KG_HS1	2160.2	2143.3	2203.1	1897.2
KG_HS2	3854.1	3138.1	3255.8	3450.2
KG_HS3	7171.5	4966.7	5834.3	6113.7
KG_HS4	3134.9	2502.8	2859.6	2775.2
KG_HS5	5960.0	38493.0	32641068.8	4185.5
RB	6.4	247.3	631.2	516.3

Table 20: DNA extracted from human hair shafts was amplified using the TruePrime™ WGA single cell kit. WGA products were quantified using a human mtDNA specific qPCR assay. Extraction and amplification yields were inconsistent between hairs obtained from the same donor. Additionally, there seems to be no correlation between copy number input and WGA yield.

Evaluation of Sygnis® TruePrime™ and QIAGEN Repli-g WGA single cell kits with forensically relevant human calcified tissue samples

Two DNA extracts were each obtained from three human bones (a femur, rib and phalange) and mtDNA copy number was determined for each extract using the mtDNA-specific qPCR method developed by Kavlick et. al. Each extract was then amplified in triplicate using the REPLI-g Single Cell DNA kit and the TruePrime™ Single Cell WGA kit. The resulting WGA product was then quantified using the same mtDNA-specific qPCR assay. Each kit was also tested using the 10 ng/µl positive control DNA from the REPLI-g kit. Pre- and post-WGA mtDNA concentrations can be seen in table 17.

Both kits enabled amplification of mtDNA from the positive control sample, with the REPLI-g kit producing significantly more copies compared to the TruePrime™ kit. However, results show that both kits failed to amplify mtDNA from the bone extracts. Quantification results actually show fewer mtDNA copies after WGA than were put in the reaction at the start.

Table 21: *Mitochondrial DNA Copy Number pre- and post-WGA*

	Mitochondrial DNA Copy Number Pre- and Post-WGA				
Sample ID	WGA Input*	WGA starting concentration (copies/µL)	WGA Sample ID	Post-WGA TruePrime™+ (copies/µL)	Post-WGA REPLI-g+ (copies/µL)
Femur 1	5198.975	104	Femur 1-1	3089	2900
			Femur 1-2	3589	2358
			Femur 1-3	3305	2716
Femur 2	5019.575	100	Femur 2-1	3888	2514
			Femur 2-2	4413	2124
			Femur 2-3	4138	2548.5
Rib 1	1728.125	35	Rib 1-1	1068	318
			Rib 1-2	1200	634.5
			Rib 1-3	1029	525
Rib 2	1875.075	27	Rib 2-1	1143	458.5
			Rib 2-2	1198.5	337.5
			Rib 2-3	1361	508.5
Phalange 1	5801.8	116	Phalange 1-1	2823	2913.5
			Phalange 1-2	3746.5	5078
			Phalange 1-3	3292.5	2315
Phalange 2	5092.825	102	Phalange 2-1	2812.5	2596.5
			Phalange 2-2	3673	2619.5
			Phalange 2-3	2748	2074.5
Reagent Blank	4	0.08	Reagent Blank	18.5	64.5
Positive Control DNA	841470.625	16829	WGA Positive	27742234.5	180259900
TE Buffer – Negative Control	0.925	0.02	WGA Negative	48.5	196.5

Table 20: DNA extracted from human calcified tissues was amplified using the TruePrime™ WGA single cell kit. WGA products were quantified using a human mtDNA specific qPCR assay. In general, TruePrime™ WGA seemed to result in higher yields than REPLI-g.

* Total mtDNA copies in 2.5 µl DNA extract. Results represent the average mtDNA copy number from duplicate qPCR amplifications
+ Total mtDNA copies in the final 50 µl WGA reaction. Results represent the average mtDNA copy number from triplicate qPCR amplifications

Multiplex amplification of the whole human mitochondrial genome

The enrichment of mitochondrial DNA (mtDNA) typing over the last 25 years has distinguished it as a viable application in forensic casework. In forensically relevant samples, nuclear DNA is often limited due to the nature of the tissue or degraded due to exposure to environmental elements. Low quantity or poor quality nuclear DNA typically precludes the use of capillary electrophoresis to obtain reliable STR profiles. In these cases, mtDNA is more accessible due to its availability in multiple copies per cell, and confers an increased detection sensitivity compared to nuclear DNA (Robin, 1988; Wilson, 1993).

Historically, most forensic analyses of human mtDNA have focused on the hypervariable regions of the genome due to their elevated rates of mutation (Wilson, 1993). However, the hypervariable regions may not always provide adequate power of discrimination. In these cases, sequencing of the complete mtGenome may afford additional information necessary for identification. A viable method of amplifying the complete mtGenome for forensic samples will require an assay that is robust to low or degraded DNA input and yields quality sequence in a time and cost-efficient manner.

We have continued to optimize a multiplex PCR assay that amplifies small fragments around the mtGenome. This assay is suitable for studies in which samples may deliver degraded or limited DNA. Subsequent NGS provides complete mtGenome coverage in the majority of samples tested. So far we have tested this assay on telogen hairs, buccal swabs, calcified tissues, and commercially available DNA. To demonstrate the utility of this assay for forensically relevant samples, we also applied the technique to hairs isolated from dust bunnies.

Evaluation of whole mtGenome multiplex PCR with human hair shafts and calcified tissues

Hair roots were examined microscopically to ensure that they lacked follicular tissue. DNA was extracted from 2 cm portions of root or proximal root of hair shafts following Burnside et al. 2012 (Burnside, 2012). Calcified tissues (human ribs) were pulverized following the FBI Mitochondrial DNA Analysis Protocol and DNA was extracted using PrepFiler® BTA Forensic DNA Extraction kit. Mitochondrial genome copy number was quantitated using qPCR following an assay designed by Kavlick et al. that targets a 105 bp segment of the NADH dehydrogenase subunit 5 gene (Kavlick, 2011). To design each multiplex mtDNA PCR assay, we utilized 46 previously developed MitoSEQr™ (Applied Biosystems®, Foster City, CA) primers tiled around the mtGenome (figure 8). Three primer sets were redesigned with Primer-BLAST to enhance amplification efficiency (Ye, 2012). Primers were consolidated into three 10 µl reactions. PCR conditions are as follows: 1x FastStart™ High Fidelity Reaction Buffer, 1.8 mM $MgCl_2$ (Roche Diagnostics, Basel, Switzerland), 2.0 µg BSA, 200 µM dNTPs

(Thermo Fisher Scientific, Waltham, MA), and 1.0 U FastStart™ High Fidelity Enzyme Blend (Roche). Primer concentrations range from 40 – 350 nM (table 22). Each PCR reaction contained 1 µl of each DNA extract regardless of mtGenome copies/µl (table 23). Thermal cycling parameters were 2 minutes at 95°C followed by 36 cycles of 95°C for 30 s, 55°C for 30 s, 72°C for 1 min and final extension of 72°C for 7 min. Amplification success of each multiplex was evaluated on Agilent 2100 Bioanalyzer (figures 9A-9C). PCR products were prepared for sequencing using the Nextera® XT kit and NGS was performed on the Illumina® MiSeq® using 2 x 151 cycles. Read counts and coverage mapping were executed using CLC Genomics Workbench version 8 (figures 10A-10E 5, table 23). Additionally, sequences from different sample types (ie. hair and buccal) from the same reference donor were aligned and compared to previously generated Sanger sequence.

Figure 8: *Orientation of 46 primers around the mtGenome.*

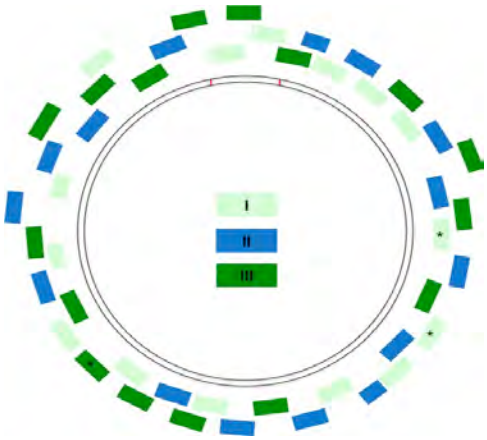


Figure 8: Primers were multiplexed into three reactions. Forty-three primers were modified from Applied Biosystems® MitoSEQr™ Kit; starred (*) primers were separately designed in Primer-BLAST.

For all samples tested, total DNA input into each multiplex PCR ranged between 1,245 – 195,057 mtGenome copies/µl (table 23). Mean coverage ranged from 350x – 71,161x. All root end hair shaft, buccal, and commercial control samples provided full genome coverage. Two proximal root samples from dust bunny hairs (DBHS2 and DBHS4) contained positions with zero coverage, however, these positions represented less than 0.03% of the entire mtGenome. Sequences from donor hair and buccal samples were aligned to one another and to previously generated Sanger sequences. No variants were detected.

Table 22: *Characterization of forty-six primer pairs multiplexed to amplify human mitochondrial genome*

Multi-plex	WCU RSA	Primer Sequence 5'- 3'	Size (bp)	Position in mtGenome (bp)	Conc. in PCR (nM)
I	23	F: GGTTGGTCAATTTCGTGCCAG R: CTGCTAAATCCACCTTCGACCCTTAAG	558	873 – 1431	200
	1	F: GCCCGTCACCCTCCTCAAGT R: GGGATAGAGGGTCTGTGGGC	593	1485 – 2078	300
	3	F: GCGTTCAAGCTCAACACCCA R: GCAGGTTTGGTAGTTTAGGACCTGTG	596	2201 – 2797	200
	36.02	F: CCCTCACCACTACAATCTTC R: GGGCCCGATAGCTTATTTAG	420	4013 - 4432	40
	46.01	F: CTCCACCTCAATCACACTAC R: GTGAGGTAAAATGGCTGAGT	533	5363 – 5895	300
	27	F: CAGCTCTAAGCCTCCTTATTCGAGC	542	5995 – 6537	300

		R: CTGTTAGTAGTATAGTGATGCCAGCAGCTAGG			
	39	F: CAATTGGCTTCCTAGGGTTTATCGTG R: GGGCATCCATATAGTCACTCCAGG	660	6739 – 7399	200
	29	F: GAAAATCTGTTCGCTTCATTCATTGCC R: GGTGGCGCTTCCAATTAGGTG	527	8533 – 9060	100
	31	F: CGAGTCTCCCTTCACCATTTCCG R: GGGTAAAAGGAGGGCAATTTCTAGATC	528	9752 – 10280	200
	8	F: CTAGTCTTTGCCGCCTGCGA R: GGGAAGGGAGCCTACTAGGGTGT	577	10659 – 11236	300
	33	F: CAAACTACGAACGCACTCACAGTCG R: GTCGTAAGCCTCTGTTGTCAGATTCAC	440	11754 – 12194	80
	34	F: CCTTCTTGCTCATCAGTTGATGATACG R: GCTTTGAAGAAGGCGTGGGTACAG	558	12788 – 13346	200
	13	F: GCCATCGCTGTAGTATATCCAAAGACA R: AGGCCTCGCCCGATGTGTAG	598	14453 – 15051	200
	44	F: GAAAAAGTCTTTAACTCCACCATTAGCACC R: GGGAACGTGTGGGCTATTTAGGCT	587	15961 – 16548	200
	22	F: CAGGTCTATCACCCTATTAACCACTCACG R: GGGTTGTATTGATGAGATTAGTAGTATGGGAG	490	6 – 496	200
II	21	F: CCCGTCCAGTGAGTCACCC R: CCCAGTTTGGGTCTTAGCTATTGTGTG	368	706 – 1074	200
	19	F: TGGCGGTGCTTCATATCCCTC R: CGCCAGGTTTCAATTTCTATCGC	596	1174 – 1770	200
	4	F: GCGGTACCCTAACCGTGCAA R: GGGAAGGCGCTGTGAAGTAGG	599	2571 – 3170	200
	6	F: CATACCCATGGCCAACCTCCT R: CGGTTGGTCTCTGCTAGTGTGGA	584	3306 – 3890	200
	25	F: CACCCCATCCTAAAGTAAGGTCAGC R: GTTTGGTTTAATCCACCTCAACTGCC	598	4389 – 4987	200
	26	F: CAGCTAAGCACCCTAATCAACTGGC R: GGCCTCCACTATAGCAGATGCG	567	5696 – 6263	200
	38	F: TGCCATAACCCAATACCAAACGC R: CTTCCGTGGAGTGTGGCGAG	467	6425 – 6892	40

Multi-plex	WCU RSA	Primer Sequence 5'- 3'	Size (bp)	Position in mtGenome (bp)	Conc. in PCR (nM)
II	45	F: CCCGATGCATACACCACATGAA R: CTAGGATGATGGCGGGCAGG	572	7233 – 7805	200
	28	F: CTACGGTCAATGCTCTGAAATCTGTG R: GTCATTGTTGGGTGGTGATTAGTCG	510	8161 – 8671	200
	17	F: ATTGGAAGCGCCACCCTAGC R: CAGGTGATTGATACTCCTGATGCGA	597	9046 – 9643	200
	32	F: CTTATGACTCCCTAAAGCCCATGTCG R: GTGATATTTGATCAGGAGAACGTGGTTAC	536	11398 – 11934	200
	10	F: TTACCACCCTCGTTAACCCTAACAAA R: CTGCTAGGAGGAGGCCTAGTAGTGG	599	12395 – 12994	200
	11	F: GCAGCAGTCTGCGCCCTTAC R: GCTGCCAGGCGTTTAATGGG	514	13198 – 13712	200
	12	F: CAGCCCTCGCTGTCACTTTCC R: GGATTGGTGCTGTGGGTGAAA	571	13802 – 14373	300
	15	F: GACAGTCCCACCCTCACACGA R: CGGATGCTACTTGTCCAATGATGG	555	15257 – 15812	200
III	2	F: AACTTTGCAAGGAGAGCCAAAGC R: GCATGCCTGTGTTGGGTTGA	568	1873 - 2441	200
	5	F: CCCTAGGGATACAGCGCATCCT R: GCGGTGATGTAGAGGGTGATGG	600	2927 – 3527	200

24	F: CCTCTAGCCTAGCCGTTTACTCAATCC R: GTGTATGAGTTGGTCGTAGCGGAATC	538	3629 – 4167	80
37	F: CTCTGAGTCCCAGAGGTACCCA R: AGGTAGGAGTAGCGTGGTAAGGGC	678	4805 – 5483	300
40	F: GAGCTTATCACCTTTCATGATCACGC R: GCTAAGTTAGCTTTACAGTGGGCTCTAG	674	7640 – 8314	200
7	F: CCTCCTCGGACTCCTGCCTC R: TGAGGAGCGTTATGGAGTGGAAG	561	8775 – 9336	60
30	F: CGATACGGGATAATCCTATTTATTACCTCAG R: TTATACTAAAAGAGTAAGACCCTCATCAATAGATGG	561	9444 – 10005	350
9	F: CCAACGCCACTTATCCAGCG R: TGTCGTAGGCAGATGGAGCTTG	596	10999 – 11595	200
41.01	F: TTGACTACCACAACTCAACG R: GGCCATATGTGTTGGAGATT	605	10124 – 10728	200
18	F: GGGCTCACTCACCCACCACAT R: TGGGTTGTTTGGGTTGTGGCT	553	12007 – 12560	80
42	F: CCACATCATCGAAACCGCAAAC R: GATGAGTGGGAAGAAGAAAGAGAGGAAG	609	13515 – 14124	200
20	F: ACGCCCATAATCATACAAAGCCC R: GGGAGGTCGATGAATGAGTGGT	587	14224 – 14811	200
14	F: CGCCTGCCTGATCCTCCAA R: GAAGGAAGAGAAGTAAGCCGAGGG	595	14860 – 15455	200
16	F: CTAGGAGGCGTCCTTGCCCT R: GGGTTTGATGTGGGTTGGGTT	577	15608 – 16185	200
43	F: CCCCCCATGCTTACAAGCAAGT R: CTGTGTGGAAAGCGGCTGTG	635	16188 – 275	200
35	F: TGGCCACAGCACTTAAACACATCTC R: CTATTGACTTGGGTTAATCGTGTGACC	606	321 – 927	200

Table 22: Primer sets are grouped in order of multiplex combination. RSA is the Resequencing Amplicon number designated by WCU. Primer sequences, expected amplicon size, position in human mtGenome, concentration of each forward and reverse primer are given for each RSA.

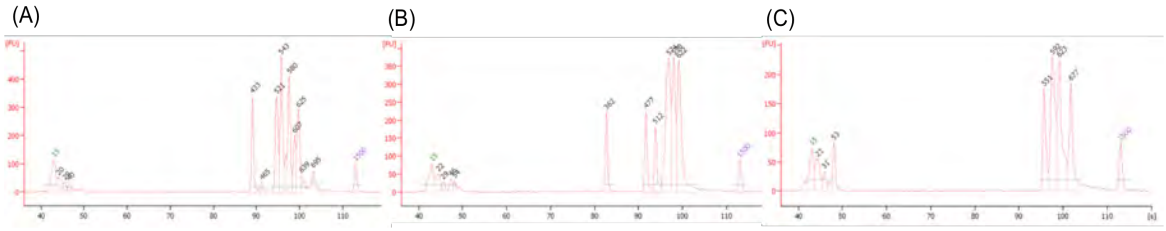
Table 23: *DNA input per multiplex PCR* reaction

Sample	Source	DNA input (mtGenome copies)	Mean coverage
Hair Shaft			
RHS1-root	Reference	12911	20794
RHS1-proximal root	Reference	3179	13705
DBHS1-root	Dust bunny	5362	10588
DBHS1-proximal root	Dust bunny	3193	71161
DBHS2-root	Dust bunny	8592	8295
DBHS2-proximal root	Dust bunny	2268	350
DBHS4-root	Dust bunny	2172	5443
DBHS4-proximal root	Dust bunny	1245	1079
DBHS3-root	Dust bunny	4083	10802
DBHS3-proximal root	Dust bunny	1520	527
Buccal			
RBS1	Reference	Not quantified	5630
Bone			

CS7114-320	Rib	6019	11530
CS7114-322	Rib	4171	13809
Commercially available			
HL-60		195057	61173

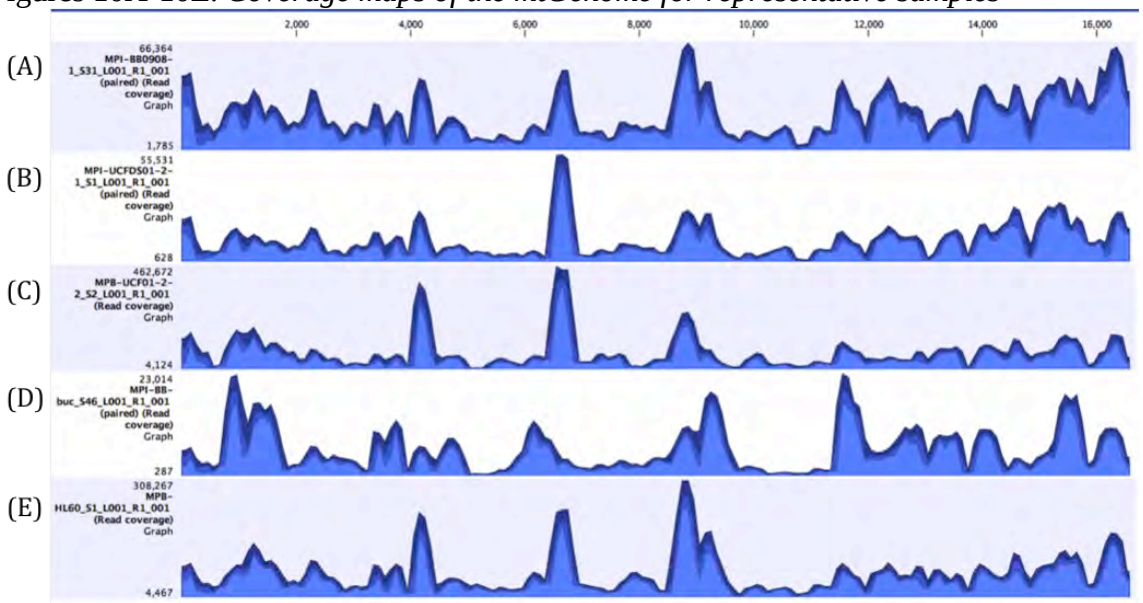
Table 23: The above table shows that higher numbers of mtDNA copies put into a PCR reaction correlates to high mean coverage values in NGS data. However, sequencing of samples with lower inputs still resulted in NGS data with mean coverage values high enough to obtain enough coverage across the genome to call variants from the rCRS with high certainty.

Figures 9A-9C: *Bioanalyzer results for multiplex amplification of a hair shaft*



Figures 9A-9C: (A) Multiplex I amplifies fifteen targets ranging from 430 – 690 bp in length. (B) Multiplex II amplifies fifteen targets ranging from 370 – 630 bp. (C) Multiplex III amplifies sixteen targets ranging from 550 – 680 bp. We had some difficulty individually evaluating amplification success of each amplicon due to overlap in size among fragments within each multiplex reaction.

Figures 10A-10E. *Coverage maps of the mtGenome for representative samples*



Figures 10A-10E: Read tracks for multiplexes I, II, and III were mapped to the rCRS reference genome using a global alignment algorithm in CLC Genomics Workbench version 8 (QIAGEN, Valencia, CA). Coverage maps are shown for (A) DNA extracted from a 2 cm root portion of hair shaft from reference donor 1 (B) DNA extracted from a 2 cm root portion of hair shaft isolated from dust bunny (C) DNA extracted from a 2 cm

proximal root portion of hair shaft isolated from dust bunny (D) DNA extracted from buccal swab from reference donor 1 (E) Commercially available DNA, HL60.

Evaluation of whole mtGenome multiplex PCR with highly compromised samples including cremated remains and single whole cells

To further assess the utility of this multiplex assay, two additional sample types were processed: ashes from human cremated remains and single whole cells. In an attempt to further evaluate the consistency of the performance of the multiplex assay, we also tested two additional human bone samples from a femur and an additional hair isolated from a dust bunny. Ashes were processed using the PrepFiler® BTA Forensic DNA Extraction (Thermo Fisher Scientific, Waltham, MA) following the protocol for calcified tissues. Bone powder from femur samples was prepared following the FBI Mitochondrial DNA Analysis Protocol and DNA was extracted using PrepFiler® BTA Forensic DNA Extraction kit. DNA was extracted from 2 cm portions of root or proximal root regions of hair shafts following Burnside et al. 2012. Single whole cells were placed directly into each multiplex reaction. Mitochondrial genome copy number was quantitated for each DNA extract using the qPCR assay described above.[2] Mitochondrial genome copy numbers are given in Table 24.

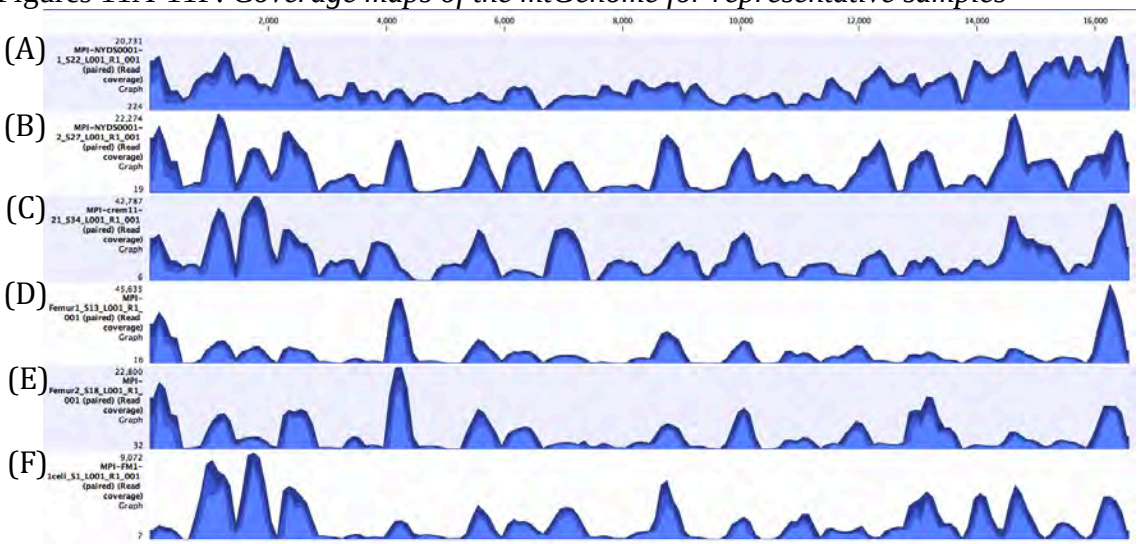
Multiplex PCR reactions were conducted as described above and amplification success was evaluated with the Agilent 2100 Bioanalyzer. PCR products were prepared for sequencing using the Nextera® XT kit and NGS was performed on the Illumina® MiSeq® using 2 x 151 cycles. Read counts and coverage mapping were executed using CLC Genomics Workbench version 8.5.1 and are given in Table 24. Coverage maps are shown in figures 11A-11F.

Table 24: *DNA input into PCR and mean coverage following NGS*

Sample	Source	DNA input (mtGenome copies)	Mean coverage
Hair Shaft			
DBHS5-root	Dust bunny	4882	6177
DBHS5-proximal root	Dust bunny	3244	5356
Ashes			
Crem_1121	Cremated remains	93	10,506
Bone			
Femur1	Femur	5124	4753
Femur2	Femur	6659	3251
Single whole cell			
FM1_1cell	Single cell	N/A	1481

Table 24: Mean coverage values obtained from each sample were high enough to enable variant calling from the rCRS with high confidence. These data show that the multiplex PCR assay is robust and is suitable for samples that may be highly compromised. Combined with the sensitivity of NGS, this method is very promising for generating whole mtGenome sequence data from forensically relevant samples.

Figures 11A-11F: *Coverage maps of the mtGenome for representative samples*



Figures 11A-11F: Read tracks for multiplexes I, II, and III were mapped to the rCRS reference genome using a global alignment in CLC Genomics Workbench version 8.5.1. Coverage maps are shown for (A) DNA extracted from a 2 cm portion of root from hair haft isolated from a dust bunny (B) DNA extracted from 2 cm portion of proximal root region of hair shaft isolated from dust bunny (C) DNA extracted from ashes from human cremated remains (D) DNA extracted from human femur (E) DNA extracted from human femur (F) direct amplification of single whole cell.

Conclusions

The multiplex PCR approach described has proven to be success for amplification of DNA extracted from compromised samples including hair shafts, calcified tissues, and single whole cells. Amplification using this method generally results in sufficient coverage across the entire mtGenome to call variants with high-levels of confidence.

Modified Human Whole mtGenome Multiplex Amplification and Next Generation Sequencing

Amplification of DNA from human hair shafts with modified multiplex PCR assay

PCR primers described for multiplex amplification of the whole mtGenome were redesigned to include Illumina® sequencing primer modifications on their 5' ends. This design obviates the need for the fragmentation step of library preparation and may be more suitable for mixture deconvolution. Initially, multiplex III (12 primer pairs) was chosen in order to test the viability of the modified primer design. DNA was extracted from three 2 cm hair shaft fragments using the protocol described by Burnside et al. 2012.[7] The hair shaft samples included a) a 2 cm root portion of a darkly pigmented, thick hair shaft (sample ID = MaH) b) a 2 cm root portion of a color treated hair shaft (sample ID = KeG) c) a 2 cm end portion of a hair stored moist in a Ziploc™ bag for one year (sample ID = KyG). Upon microscopic analysis, none of the samples were observed

to contain soft tissue adhered to the root. A human specific qPCR assay was used to quantify DNA in the extracts.[2] Each extract was quantified in triplicate. Results are shown in Table 25.

Table 25: *Quantitative PCR results for hair shaft extracts*

Sample ID	Average of Triplicate Quants (copies/2 µL)	PCR Input (copies mtDNA per reaction)
MaH	9,801	4,900
KeG	4,144	2,072
KyG	2,801	1,400
Reagent Blank	9.87	4.94
Non-Template Control	undetected	NA

Each sample was amplified with a modified primer set in singleplex, and with a set of pooled primers (equimolar concentrations of each) in a multiplex format. HL60 DNA was also amplified as a positive control (input = 10 pg). Amplification was conducted with 1.0 µL of extract in a 10 µL reaction containing 1.0 µL of Roche FastStart™ High Fidelity PCR 10X buffer, 0.16 µg/µL Bovine Serum Albumin (BSA), 200 µM dNTPs, 1.0 µM primers (either forward and reverse singleplex primers or total concentration for multiplex primers) and 0.5 U Roche FastStart™ High Fidelity enzyme. PCR was conducted with the following conditions: 2 minute at 95°C followed by 36 cycles of 30 seconds at 95°C, 30 seconds at 55°C, and 1 minutes at 70°C with an infinite 4°C hold. Resulting amplification products were assessed using the Agilent 2100 Bioanalyzer. Bioanalyzer results are shown in Table 26.

Table 26: *Agilent 2100 Bioanalyzer results for modified multiplex primer amplification*

Primer ID	HL60	MaH	KeG	KyG
	PCR Yield (ng/µL)			
26	18.73	5.74	0.48	0.49
32	22.4	9.42	0.9	1.54
38	18.01	10.54	4.15	2.49
10	22.56	9.65	1.94	0.47
15	22.41	8.3	2.54	1.56
28	18.54	5.88	0.86	0.86
11	17.49	8.03	1.15	0.91
4	24.96	8.31	0.91	0.99
25	20.41	10.2	1.86	0.81
19	22.72	9.24	2.09	0.43
17	21.87	7.83	2.68	1.03
12	19.18	7.87	1.12	0.77
Multiplex	14.14	13.82	2.95	1.73

NGS of samples amplified with modified multiplex III PCR primers

Amplified samples were then prepared for NGS on the Illumina® MiSeq®. All sample were diluted to 0.04 ng/µL with molecular biology grade water. The samples were then integrated into the Nextera® XT library preparation workflow at the PCR amplification step. The vendor recommended protocol was then followed from this point on. Prepared libraries were sequenced on the Illumina® MiSeq® with a v3 600 cycle run kit. Data was analyzed using CLC Genomics Workbench software v8. For all

samples and positive controls, all 12 amplicons were represented in NGS coverage plots. Average coverage values are included in Table 27.

Figure 12A-12D: *Coverage plots for multiplex amplification of hair extracts using modified primers*

(A)

(B)

(C)

(D)

Figures 12A-12D: NGS reads obtained for all samples amplified with modified multiplex III were mapped to the rCRS reference genome using a global alignment algorithm in CLC Genomics Workbench version 8.5.1. Coverage maps are shown for (A) DNA extracted from HL60 commercial control (B) DNA extracted from 2 cm portion of root region of hair shaft isolated from donor MaH (C) DNA extracted from 2 cm portion of root region of hair shaft isolated from donor KeG (D) DNA extracted from 2 cm portion of root region of hair shaft isolated from donor KyG.

Table 27: *Average coverage values for whole mtDNA modified multiplex amplification of human hair shafts*

Primer ID	HL60	MaH	KeG	KyG
	Average Coverage			
26	6,879	8,742	9,853	7,824
32	37,406	35,551	29,318	24,884
38	149,894	165,771	237,199	168,213
10	5,859	7,518	3,887	5,960
15	19,957	20,875	23,167	10,125
28	3,642	6,266	5,709	3,625
11	14,058	21,952	23,068	26,267
4	5,554	5,832	4,370	7,996
25	12,099	14,336	11,683	7,660
19	8,400	8,289	3,468	3,652
17	12,036	10,794	8,736	9,709
12	6,380	7,747	4,531	5,549

Conclusions

The modified multiplex PCR approach is suitable for amplification of DNA extracted from compromised samples. Modifying of the primers to contain 5' regions that are complementary to Illumina® sequencing read primers obviates the need for fragmentation prior to library preparation. This approach is desirable for processing samples that may already contain highly fragmented DNA.

Synthetic oligonucleotide sequencing and Illumina® MiSeq® error rate estimation

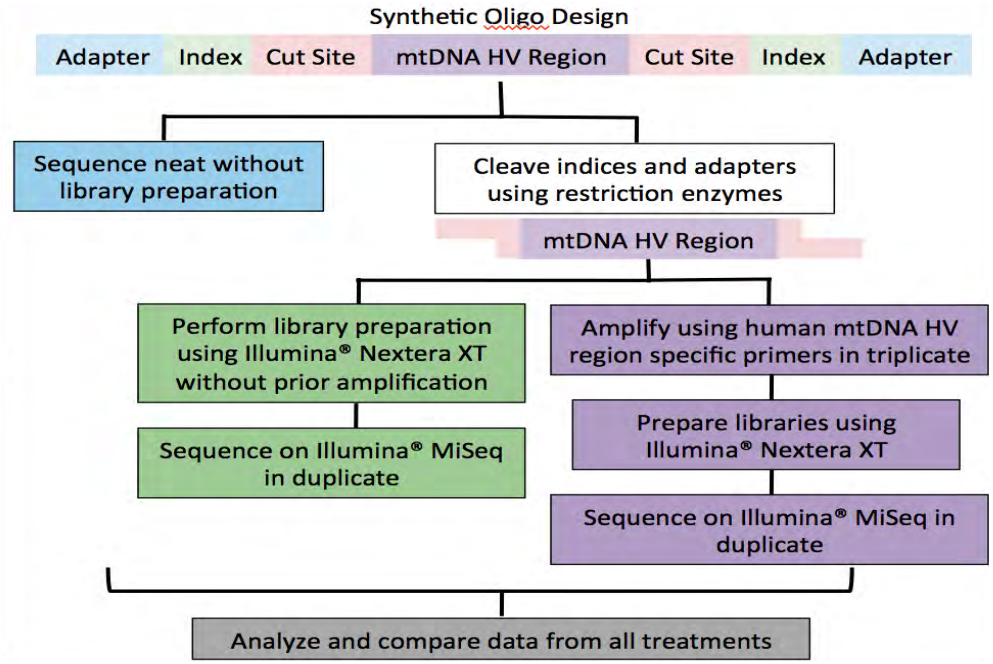
NGS methods are proving to be particularly well-suited for mitochondrial DNA analysis, and may provide forensic analysts with a powerful tool that enables deconvolution of mtDNA mixtures, or accurate quantitation of low-level heteroplasmy. However, some effort remains in validating the systems for such analyses. Several NGS platforms are commercially available, each with a unique library preparation strategy and sequencing chemistry that may give rise to method-specific errors. Furthermore, since many alignment and variant calling algorithms are available, there is limited consistency in the use of data analysis methods employed. Finally, no studies have been performed to determine what depth of coverage is required to confidently call a true biological low-level variant above the level of method-generated noise.

NGS of synthetic oligonucleotides

Here, we describe a study that aims to identify error rates associated with each step in the Illumina® MiSeq® NGS workflow. Initially, synthetic oligonucleotides with sequences matching the rCRS hypervariable (HV) regions I and II of the human mtDNA genome were purchased from Life Technologies. Each oligonucleotide was designed to contain Illumina® sequencing primers, flow cell adapters and multiplexing indices on either end to enable direct sequencing without additional preparation. The oligonucleotides were also designed to contain restriction enzyme cut sites between the

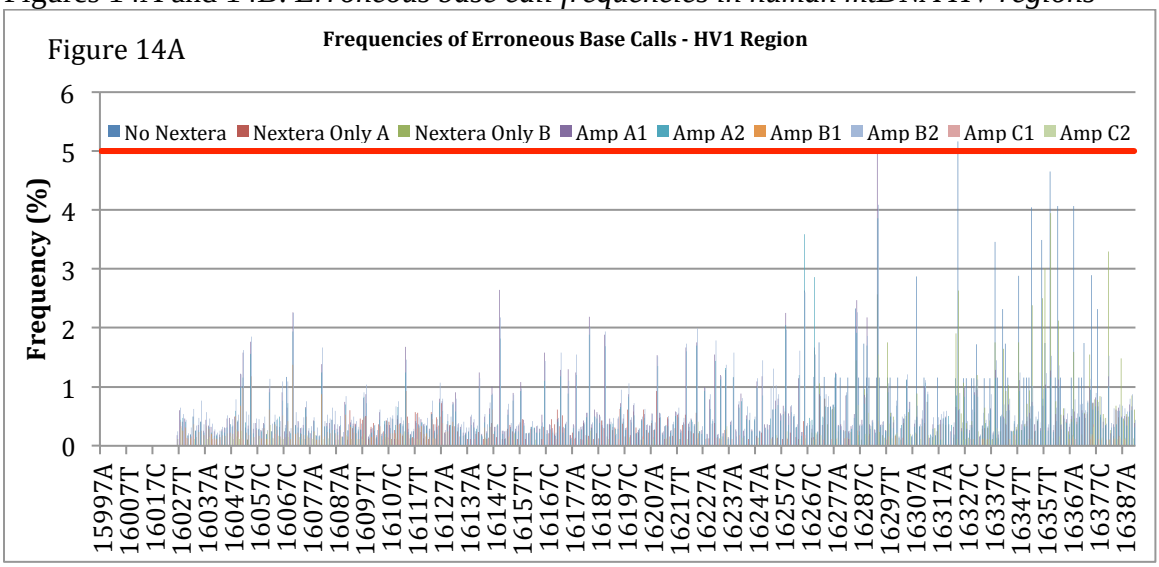
target sequence and Illumina® modifications. This design allowed for removal of Illumina modifications so the same sample could be prepared for sequencing using recommended library preparation strategies. Each synthetic oligonucleotide was sequenced a) directly with no additional preparation, b) after Illumina® Nextera® XT library preparation, and c) after triplicate PCR amplification with target specific primers followed by Nextera® XT library preparation. Samples prepared with treatments B and C were sequenced in duplicate to enable assessment of intra-run variation (figure 13).

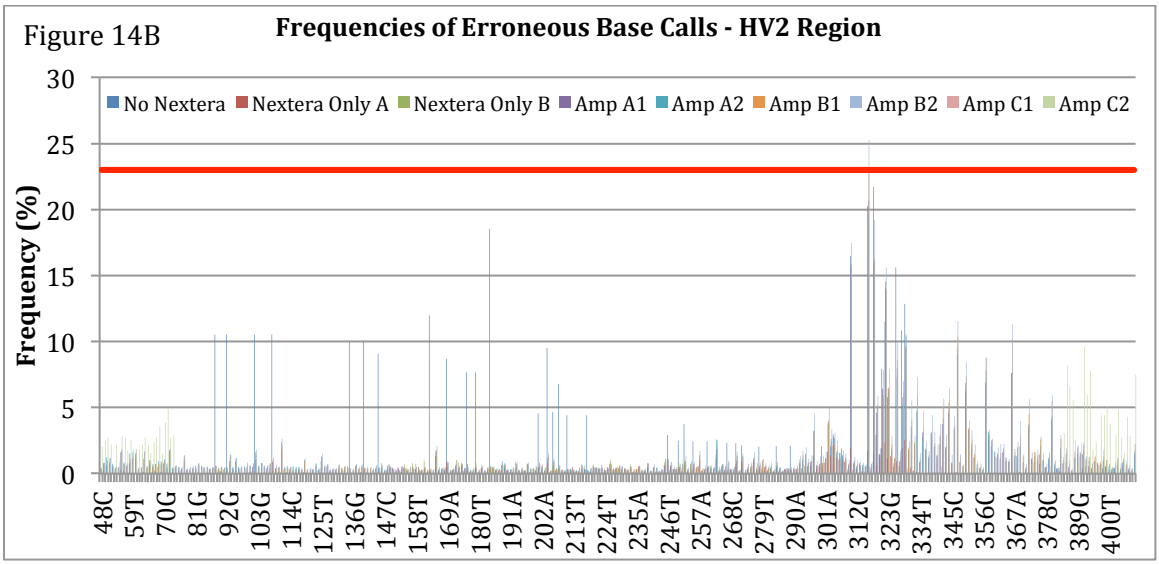
Figure 13: *Experimental design for synthetic oligonucleotide sequencing to assess NGS error associated with discrete steps of the Illumina MiSeq workflow*



Sequences were generated on the Illumina® MiSeq® with a v2 300 cycle run kit. Resulting sequence data was aligned to the rCRS. Variant calling was performed with CLC Genomics Workbench software v8.0 using both the Basic Variant Detection and Low Frequency Variant Detection algorithms with a frequency threshold of 0.1%. Error rates obtained from all sample treatments were compared to identify differences at each step in the library preparation workflow. Ultimately, this experimentation sets the groundwork for validation of the Illumina® MiSeq® NGS system for mtDNA analysis in forensic casework.

Figures 14A and 14B: *Erroneous base call frequencies in human mtDNA HV regions*





Figures 14A and 14B: Data was analyzed using CLC Genomics Workbench v8.0. Initially, data was aligned to the rCRS reference genome (NC_012920) using the proprietary heuristic-based GxWb5.5 algorithm. Variant calling was performed with Basic Variant Detection using a minimum variant frequency of 0.1% and ploidy setting of 1. Basic Variant Detection calls a maximum number of variants rapidly without applying error-model estimation. All other parameters were unmodified. Analysis settings are provided upon request. Frequencies were similar across all treatments except in cases where average coverage was low. Higher variant frequencies are observed toward the end of the targeted region, or in regions surrounding homopolymeric stretches, which may be an artifact of oligonucleotide synthesis or sequencing chemistry. Nearly all frequencies of erroneous base calls fell below 5% in HV1 data suggesting that an appropriate minimum frequency setting for these amplicons is ≥5% to avoid calling low level errors (figure 14A). HV2 data contains higher error frequencies, particularly in low coverage data sets and homopolymeric stretches (figure 14B). While a frequency threshold of ≥5% would be appropriate for the majority of the HV2 targeted region, calls

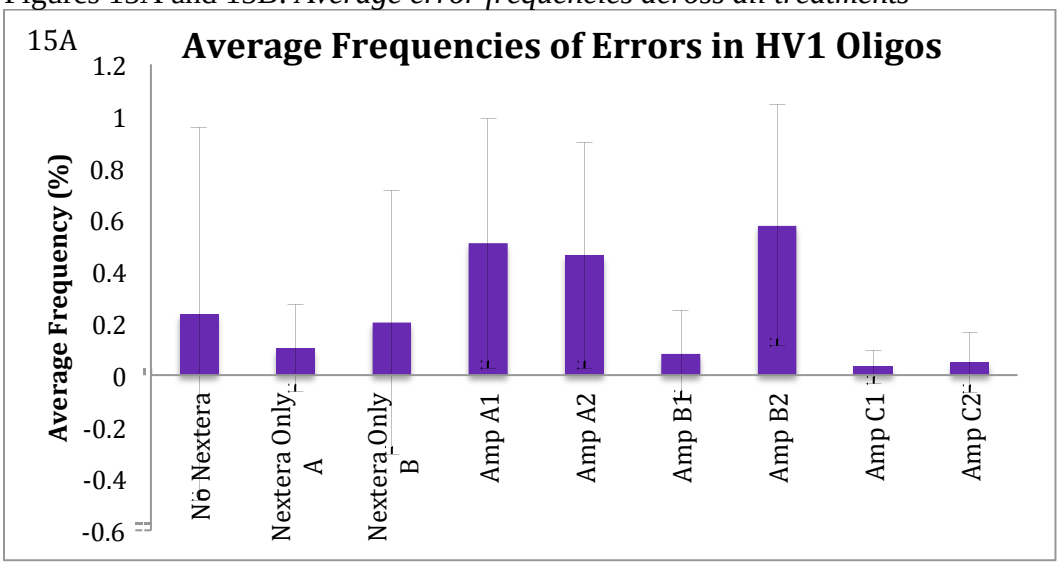
at certain positions may need to be cautiously interpreted using a higher threshold. The most ideal approach would include setting an independent frequency threshold for each position within the targeted region. When average frequencies are calculated per position, ranges of 0-3.97 +/- 1 standard deviation, or 0-5.86 +/- 2 standard deviations for HV1 data. In HV2 data, frequency ranges of 0.19-23.61 +/- 1 standard deviation or 0.28-34.82 +/-2 standard deviations are observed.

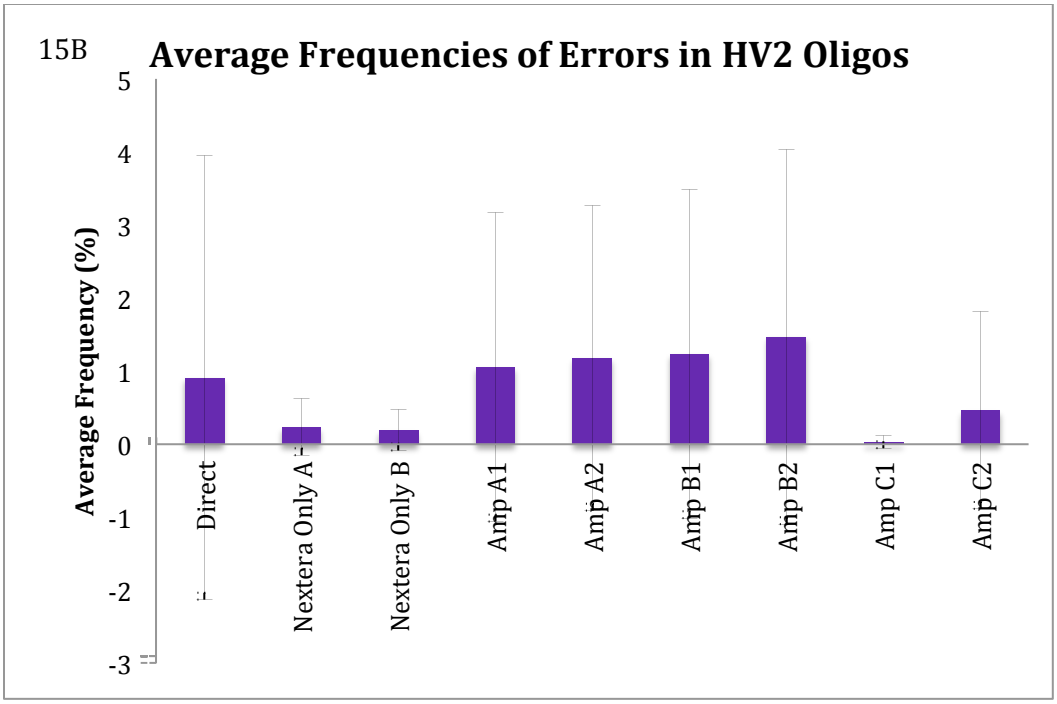
Table 28: Average coverage and error quality of NGS data

	HV1		HV2	
Library ID	Average Coverage	Average Error Quality	Average Coverage	Average Error Quality
No Nextera	60	15.63	329	18.5
Nextera Only A	24,595	17.97	11,729	17.05
Nextera Only B	34,020	16.03	11,457	16.92
Amplification A1	29,168	29.14	22,392	24.5
Amplification A2	51,621	28.04	21,470	25.73
Amplification B1	54,825	29.15	19,352	27.81
Amplification B2	58,703	29.14	52,880	27.29
Amplification C1	62,618	32.9	11,861	26.85
Amplification C2	68,683	25.9	1,187	17.24

Table 28: This table outlines the average coverage and error quality (Q-score) for each experimental treatment. Coverage was low for synthetic oligos that were sequenced directly without prior amplification. All samples were sequenced using a 2 x 151 cycle paired-end run kit. As a result, oligos sequenced directly (> 300 bp in length with no fragmentation) showed very low coverage in the center of the target region. It should be noted that MPS error quality is dependent on depth of coverage, with a maximum of 40 (error probability of 1 in 10,000).

Figures 15A and 15B: *Average error frequencies across all treatments*





Figures 15A and 15B: Average frequencies were calculated for erroneous basecalls derived from each data set. Low coverage data sets (oligos sequenced directly) had higher standard deviations due to the high error frequency observed at certain positions within the targeted region. In general, average error frequencies were <1.1% in HV1 data and <5% in HV2 data. HV2 data has higher overall error frequency presumably due to the c-stretch spanning positions 303-315 (error frequencies increase substantially in this region with highest frequencies observed at positions 310 and 316). This may be a result of sequencing chemistry or synthesis of the oligonucleotide. Again, these observations argue that position dependent thresholds be developed for each targeted region sequenced.

Conclusions

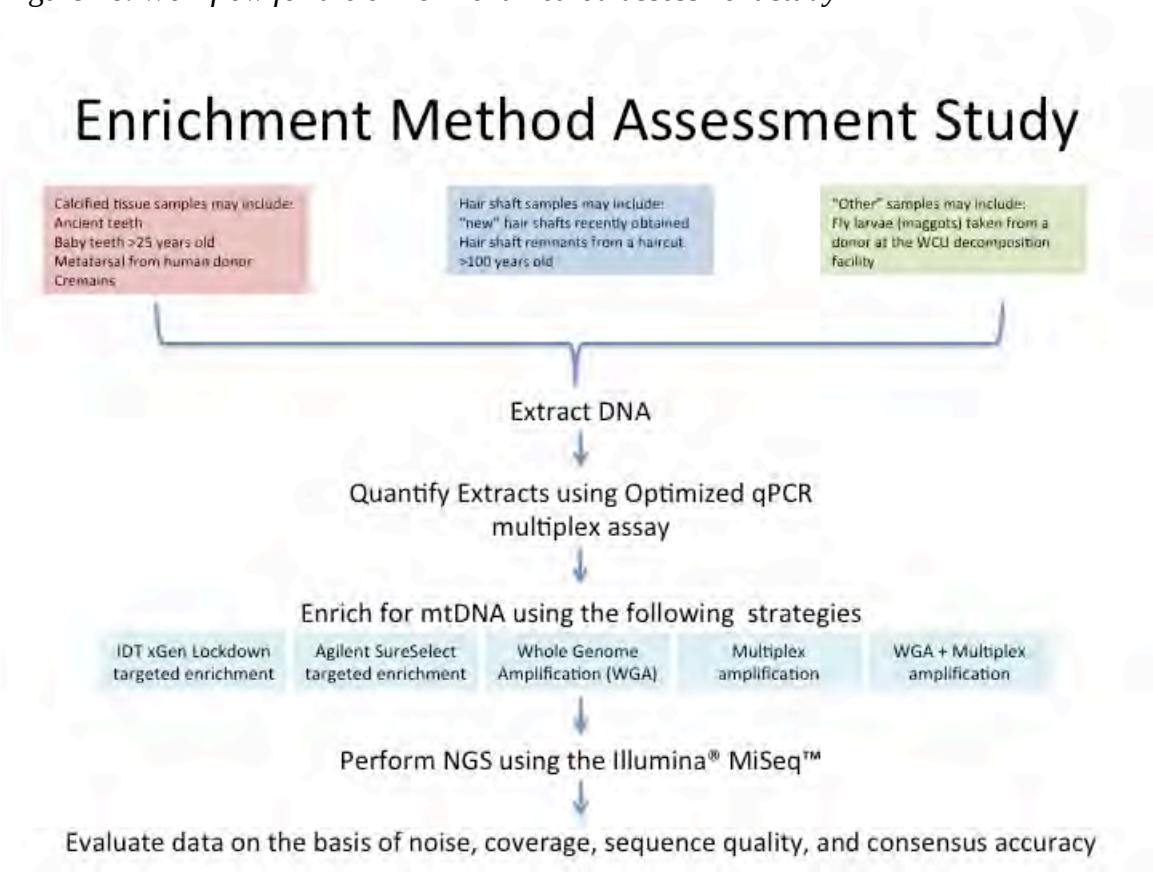
Overall, error frequencies in Illumina® NGS data sets generated using synthetic oligos with known sequences were low. Elevated per-position error rates were observed in data sets where coverage was low. Additionally, corresponding Q-scores were low in these data sets. Excluding low coverage data sets, maximum frequencies of error calls were <5% for HV1 oligos and <10% for HV2 oligos except in areas surrounding homopolymeric C-stretches. A universal threshold could be applied to data that includes calculating average frequency across all positions +/- 1 or 2 standard deviations. A more appropriate strategy would include establishing an independent threshold for each position within the targeted region. All data sets were also analyzed using CLC Genomics Workbench with the Low Frequency Variant detection algorithm using a 0.1% frequency threshold and a required significance of 1.0%. This method applies an error correction model to the data to remove erroneous base-calls. When using this approach, no differences from the rCRS were called in any data set. While this method may increase the certainty that base-calls represent true biological variation, it may result in exclusion of significant low-level variants including heteroplasmy or calls arising from a

low-level secondary contributor. Caution should be used when interpreting data analyzed with this algorithm.

Human mtDNA Enrichment

We worked closely with two competing vendors to design probe capture assays that target the whole human mitochondrial genome. Integrated DNA Technologies offers an assay called the xGen® Lockdown® Panel that is prepared using independently synthesized DNA oligonucleotide 5'-biotinylated baits (IDT®). Once synthesized, each bait is individually assessed for quality (length, sequence, etc.) using mass spectrometry. Alternatively, Agilent Technologies offers a similar target enrichment assay in which RNA baits are synthesized on a microarray. Quality control of the finalized assay is performed on the population of baits as a whole. We have chosen to evaluate both assays for whole mtGenome enrichment since each differs synthesis and structure of the probe capture baits, and in the per sample cost.

Figure 16: *Workflow for the enrichment method assessment study*



Extraction and quantitation of DNA from enrichment study samples

Initially, DNA was extracted from a series of compromised forensically relevant samples (table 29). Hair samples were microscopically examined to verify that no follicular tag was present. Portions of each hair shaft were isolated for extraction (2 cm

fragments were obtained from samples HS1, HS2, and OH; 1.5 cm fragments were used from donors FDH1 and FDH2). Hairs were cleaned thoroughly and were batch extracted in triplicate using a hair protocol developed in-house.[7] Triplicate extracts were combined to create a master sample with a large enough volume so the same extract could be used for all enrichment strategies. A reagent blank was also extracted alongside hair samples. Calcified tissue samples were also batch extracted. Initially, samples AT, BT, and PH were pulverized using the SPEX 6770 freezer/mill® (SPEX Sample Prep®, Metuchen, NJ) with polycarbonate coated components to prevent metal contamination of the powdered samples. Cremated remains were not pulverized using the SPEX mill as they were already ash or brittle enough to pulverize manually. Triplicate aliquots (50 mg) of each pulverized sample were placed in UV irradiated microcentrifuge tubes. Samples were extracted using the PrepFiler® BTA Forensic DNA Extraction kit. A reagent blank was also extracted alongside calcified tissue samples. Soft tissue samples were batch extracted. Triplicate fly larvae extractions were prepared by weighing ~50 mg of larvae and performing manual homogenization in tissue lysis buffer using a disposable glass matched mortar and pestle set. Homogenates were incubated at 56°C for 30 minutes for full digestion of residual tissue. Fly larvae homogenates, and triplicates of buccal swabs 1 and 2 were extracted using the QIAGEN EZ-1® DNA Tissue kit and extraction robot (QIAGEN, Valencia, CA). A reagent blank was also extracted alongside soft tissue samples. Prior to quantitation, triplicate extracts were combined to create a master sample with a large enough volume so the same extract could be used for all enrichment strategies. All samples were quantified using the nuclear/mitochondrial DNA multiplex qPCR assay described herein. Each master extract was quantified in triplicate (table 30).

Table 29: *Enrichment study sample information*

Sample ID	Sample Description	Storage conditions
HS1	2 cm hair shaft, no follicular tag	Freshly obtained from donor
HS2	2 cm hair shaft, no follicular tag	Freshly obtained from donor
OH	2 cm portion of haircut remnants, > 100 years old	Unknown, provided in Ziploc™ bag
FDH1	2 cm portion of beard hair obtained from deceased donor	Beard hair obtained from a deceased male donor that was placed outdoors at the WCU human decomposition facility
FDH2	2 cm portion of head hair obtained from deceased donor	Head hair obtained from a deceased female donor that was placed outdoors at the WCU human decomposition facility
AT	Tooth sample	Tooth sample unearthed from an unmarked grave in a local family burial plot. Suspected to be from the early 1800s
BT	Baby tooth sample	~23 years old. Storage conditions unknown
PH	Human phalanx	Obtained from a deceased donor that was placed outdoors at our human decomposition facility
CRA	Cremated human remains	Ash portion of sample was used
CRB	Cremated human remains	Large bone fragment was used
FL	Fly larvae	Recovered from a deceased donor that was placed outdoors at our human decomposition facility. Stored in absolute ethanol at -20°C for ~4 years after collection.

BUC1	Buccal swab	Fresh buccal swab obtained from same donor that provided sample HS1. Serves as a control.
BUC1	Buccal swab	Fresh buccal swab obtained from same donor that provided sample HS2. Serves as a control.
HL60	Purified cell line DNA	Low concentration positive control (100 pg/µL)

Table 30A and 30B: *qPCR nuclear (30A) and mitochondrial (30B) DNA quantitation values of enrichment study sample extracts*

Sample ID	Rep 1	Rep 2	Rep 3	Average	Standard Deviation
	ng/µL				
HS1	0	0.001	NA	0.00	0.00
HS2	0	0	NA	0.00	0.00
OH	0	0	0	0.00	0.00
FDH1	0	0	0	0.00	0.00
FDH2	0	0	0	0.00	0.00
H RB	0	0	NA	0.00	0.00
AT	0	0.001	0.001	0.00	0.00
BT	3.471	3.973	4.225	3.89	0.38
CRA	0	0	0	0.00	0.00
CRB	0	0	0	0.00	0.00
PH	0.032	0.022	0.016	0.02	0.01
CT RB	0	0	0	0.00	0.00
FL	0	0	NA	0.00	0.00
BUC1	14.574	15.476	NA	15.03	0.64
BUC2	34.14	32.263	34.771	33.72	1.30
T RB	0	0	0	0.00	0.00
HL60 20	0.116	0.105	0.076	0.10	0.02
NTC	0	0	0	0.00	0.00

Table 30A: Nuclear DNA quantification values for enrichment study samples were very low overall. This is not unexpected since the majority of these samples are either compromised or contain low amounts of DNA. Sample BT (baby tooth) yielded enough DNA for successful traditional STR typing. Samples BUC1 and BUC2 also yielded high concentrations of nuclear DNA. This is not unexpected since these robust samples are included for control purposes. The average nuclear DNA concentration of HL60 was exactly as expected (100 pg/µL). All reagent blanks and non-template controls had undetectable levels of nuclear DNA.

Sample ID	Rep 1	Rep 2	Rep 3	Average	Standard Deviation
	copies/µL				
HS1	1890	1850	NA	1870.00	28.28
HS2	484.79	563.22	NA	524.01	55.46
OH	145.76	137.95	135.01	139.57	5.56
FDH1	71.34	73.59	74.75	73.23	1.73
FDH2	408.2	381.27	393.73	394.40	13.48
H RB	2.77	3.88	NA	3.33	0.78
AT	357.91	329.95	317.61	335.16	20.65
BT	164000	190000	178000	177333.33	13012.81
CRA	4.33	4.35	4.7	4.46	0.21

CRB	4.22	6.96	7.59	6.26	1.79
PH	3230	3360	3400	3330.00	88.88
CT RB	5.3	4.92	2.37	4.20	1.59
FL	752000	791000	NA	771500.00	27577.16
BUC1	966000	986000	NA	976000.00	14142.14
BUC2	1780000	1770000	1700000	1750000.00	43588.99
T RB	2.78	3.94	3.79	3.50	0.63
HL60 20	6566.048	7344.857	5877.102	6596.00	734.34
NTC	4.77	9.87	3.65	6.10	3.32

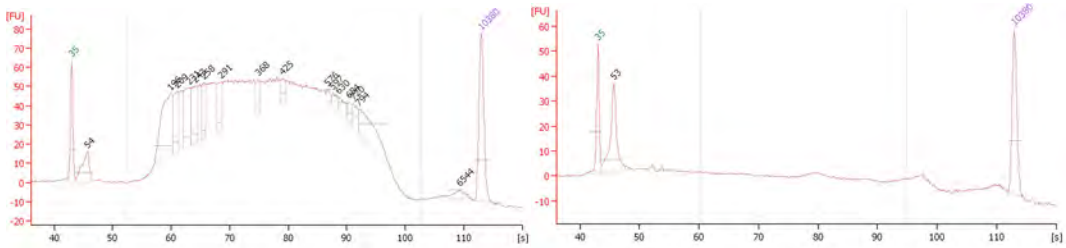
Table 30B: All enrichment study sample extracts contain enough DNA for successful downstream PCR amplification (minimum of 100 copies/µL) except FDH1, CRA, and CRB (highlighted in red). However, these samples will be prepared using all enrichment methods regardless. It is possible, given the nature of the enrichment methods and sensitivity of NGS that analyzable sequence data will be obtained for these samples. All reagent blanks and non-template controls had very low quantities of mitochondrial DNA (highlighted in green).

Enrichment strategy 1: IDT xGen® Lockdown® Target Capture

Neat DNA extracts were enriched for human mtDNA using the IDT xGen® Lockdown® custom target capture panel. Initially, sequencing ready libraries were prepared using the Illumina® Nextera® XT library preparation kit. All samples were processed using the vendor recommended protocol up to a final purification step with Agencourt AMPure XP beads. This method involves enzymatic fragmentation and simultaneous tagging (tagmentation) of sample DNA with adapters complementary to Illumina® sequencing read primers. Limited cycle PCR then enables addition of barcoding indices and flow cell adapters to the DNA. Barcoding indices facilitate bioinformatic parsing of raw data generated for each sample sequenced concurrently on a single NGS run. Flow cell adapters help anchor DNA to a solid support on which sequencing takes place. Bead-based normalization was not performed. Prepared libraries were assessed using the Agilent 2100 Bioanalyzer with the DNA High Sensitivity kit. Bioanalzyer results showed that library preparation was successful for several enrichment study samples including FDH1, FDH2, AT, BT, FL, BUC1, and BUC2 (illustrative data shown in figures 17A and 17B).

Figures 17A and 17B: *Bioanalzyer electropherograms illustrating successful (17A) and failed (17B) library preparation*

17A – FDH1 17B - PH



Figures 17A and 17B: Successful library preparation is evidenced by a broad peak in the electropherogram showing a wide distribution of fragment sizes typically ranging from

100-1000 basepairs (figure 17A). This peak is absent when concentrations fall below the 5 pg/µL limit of detection of the Bioanalzyer and DNA High Sensitivity kit or when DNA is high-molecular weight and not tagmented (figure 17B).

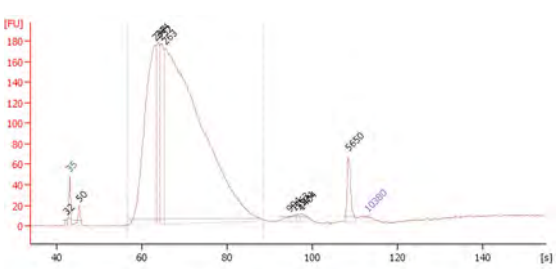
Regardless of apparent library preparation success, 19 µL of each Nextera® XT tagmented library was pooled. The entire volume of the resulting pooled library was combined with 5 µg of Cot-1 DNA and 1 µL each of xGen® Nextera® XT blocking oligos. It should be noted that recommended input for the probe capture assay is 500 ng. Significantly less than the recommended amount was added. The entire volume of pooled library was evaporated using a vacuum concentrator. The dried library was reconstituted with 8.5 µL of 2X xGen® hybridization buffer, 2.7 µL xGen® hybridization buffer enhancer and 1.8 µL of molecular biology grade water and the solution was incubated at room temperature for 10 minutes. Hybridization and capture were performed according to the vendor protocol with no modifications.

Enrichment strategy 2: Agilent Technologies SureSelect^{XT} Target Capture

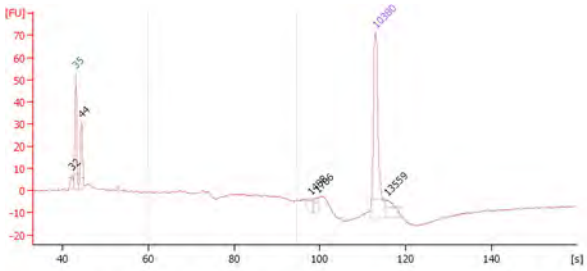
Neat DNA extracts were enriched for human mtDNA using the Agilent Technologies SureSelect^{XT} Target Capture kit. Initially, DNA was enzymatically fragmented using NEBNext® dsDNA Fragmentase® (New England Biolabs®, Inc., Ipswich, MA). Digestion reactions were prepared by combining 4 µL of 10X reaction buffer, 0.4 µL 100X BSA and 33.6 µL of each extract. Samples were incubated on ice for 5 minutes and 2 µL dsDNA Fragmentase® was added. The samples were allowed to incubate for 20 minutes at 37°C. SureSelect^{XT} library preparation was performed with no modifications starting with repairing the ends of the fragmented samples. Following amplification of post-capture libraries, each sample was assessed for successful probe capture using the Agilent 2100 Bioanalyzer and DNA High Sensitivity kit (figures 18A and 18B).

Figures 18A and 18B: *Bioanalzyer electropherograms illustrating successful (17A) and failed (17B) SureSelect^{XT} library preparation*

18A – FL



18B – FDH1



Figures 18A and 18B: Successful SureSelect^{XT} probe capture is evidence by a broad peak in the electropherogram showing a wide distribution of fragment sizes typically ranging from 100-1000 basepairs (figure 18A). This peak is absent when concentrations fall below the 5 pg/µL limit of detection of the Bioanalzyer and DNA High Sensitivity kit or when DNA is high-molecular weight and not fragmented (figure 18B).

Enrichment strategy 3: Amplification with Sygnis® TruePrime™ Single Cell WGA Kit

All enrichment sample extracts were amplified using the TruePrime™ single cell WGA kit according to manufacturers recommendations. A buffer (L2) was prepared by combining 2.5 µL of molecular biology grade water with 22.5 µL of TruePrime™ buffer L1 per sample. DNA extract (2.5 µL) was combined with buffer L2 (2.5 µL) and the resulting solution was incubated at room temperature for 3 minutes. Neutralization buffer (2.5 µL) was added and WGA was performed by adding 42.5 µL of PCR master mix to each sample. The master mix was prepared according to the TruePrime™ user manual. Reactions were incubated for 6 hours at 30°C. WGA products were quantified using the multiplex qPCR assay designed in-house (table 31).

Table 31: *qPCR mtDNA quantitation values obtained for enrichment samples following WGA*

Sample ID	WGA Input	WGA yield Rep 1	WGA yield Rep 2	WGA yield Rep 3	Average	Standard deviation
	(copies/µL)					
HS1	94	1,559	1,571	2,063	1,565	9.01
HS2	26	Inhibition/qPCR Fail			NA	NA
OH	7	Inhibition/qPCR Fail			NA	NA
FDH1	4	Inhibition/qPCR Fail			NA	NA
FDH2	20	Inhibition/qPCR Fail			NA	NA
H RB	0.2	Inhibition/qPCR Fail			NA	NA
AT	17	Inhibition/qPCR Fail			NA	NA
BT	8,867	7,574	6,061	7,366	7,000	820.2
CRA	0.2	NA	NA	NA	NA	NA
CRB	0.3	Inhibition/qPCR Fail			NA	NA
PH	167	66	59	68	64	4.63
CT RB	0.2	C_T but no quant			NA	NA
FL	38,575	1,932,652	2,113,340	1,968,072	2,022,996	127,765.7
BUC1	48,800	3,972,826	4,053,949	4,024,702	4,013,387	57,363
BUC2	87,500	6,150,423	8,532,578	8,034,134	7,572,378	1,256,415.2
T RB	0.2	Inhibition/qPCR Fail			NA	NA
HL60 20	330	2,090,181	2,042,477	1,849,235	1,993,964	127,588.6
NTC	0.3	49,044			49,044	NA
qPCR NTC	NA	1			1.29	NA

Table 31: Robust samples with high extract concentrations resulted in high WGA yields. Compromised samples did not appear to amplify successfully with WGA. However, IPC DNA also failed to amplify during qPCR of these samples. It is possible that competitive inhibition is occurring during qPCR of these samples because the WGA product concentration is so high. However, 10X and 100X dilutions of these sample yielded similar trends in quantitation data.

To ascertain that residual primers synthesized in-situ during WGA were not affecting the reaction kinetics of qPCR, WGA products were incubated at 95°C for 5 minutes and snap cooled on ice for 2 minutes for denaturation of unincorporated primers and template DNA. Denatured samples were cleaned using AMPure XP beads to remove fragments <100 bp in length. Cleaned samples were requantified using qPCR. No change was observed in qPCR data (not shown).

WGA products were normalized to 0.2 ng/µL. In cases in which the post-WGA concentration was below 0.2 ng/µL or undetectable, no dilutions were performed and samples were sequenced neat. WGA products were enzymatically fragmented using NEBNext® dsDNA fragmentase®. Resulting fragmented products were end-repaired, an A-overhang was added, and Nextera® XT sequencing primer adapters were ligated to both ends of the fragments. Adapter ligated libraries were then further prepared for NGS using the Nextera® XT library preparation kit starting with the limited cycle PCR step and moving forward with no additional modifications to the vendor recommended protocol.

Enrichment strategy 4: Amplification of whole mtGenome using multiplex PCR

Neat sample extracts were amplified using the whole mtGenome multiplex PCR assay described herein. Amplification products were quantified using the Agilent 2100 Bioanalzyer and DNA 1000 kit. Amplification yields are listed in table 32.

Table 32: *Amplification yields for multiplex amplification of enrichment study sample extracts*

Sample ID	Total multiplex amplification yield (ng/µL)		
	Multiplex I	Multiplex II	Multiplex III
HS1	39.86	59.67	19.83
HS2	13.06	13.61	0.86
OH	1.06	0	0
FDH1	1.18	1.93	0
FDH2	13.39	7.39	1.10
HRB	0	0	0
AT	0	0	0
BT	42.1	35.17	12.03
CRA	0	0	0
CRB	0	0	0
PH	8.49	1.58	0
CT RB	0	0	0
FL	40.33	46.04	23.18
BUCI	33.24	97.33	20.04
BUC2	0	33.77	21.94
T RB	0	0	0
NTC	0	0	0
HL60	0	0	23.19

Table 32: In general, amplification yields were sufficient for NGS library preparation for all samples except those highlighted in red (OH, AT, CRA, and CRB). Regardless of the lack of apparent amplification, sequencing will be performed on these samples since the sensitivity of NGS may result in low-coverage data. In general, multiplex III yields are lowest overall. This is not unexpected since this particular reaction typically performs less efficiently than multiplex reactions I and II. No amplification was evident for reagent blanks and non-template controls (highlighted in green).

Amplification products were normalized to 0.2 ng/µL. In cases in which the post-amplification concentration was below 0.2 ng/µL, no dilutions were performed and

samples were sequenced neat. NGS libraries were prepared using the Nextera® XT library preparation kit with no modifications to the vendor supplied protocol.

Enrichment strategy 5: Amplification of whole mtGenome using WGA and multiplex PCR

WGA products (described in section 9.4) were amplified using the multiplex PCR strategy described herein. Resulting PCR products were assessed on the Agilent 2100 Bioanalzyer DNA 1000 kit. No amplification was evident for any sample except BUC2 (5.56 ng/µL) and HL60 (3.23 ng/µL) when coupling WGA with multiplex PCR amplification. Further research is needed to determine why WGA yields are inconsistent and unpredictable.

Next-generation sequencing of enriched libraries

Prepared libraries generated using each enrichment strategy (except enrichment strategy 5) were sequenced on the Illumina® MiSeq® using a 2 x 151 cycle paired-end approach with v2 reagents. Resulting data from all enrichment libraries was compared to determine which approach 1) yields analyzable data for compromised samples containing degraded and/or low template DNA 2) results in lowest instance of error or noise 3) provides highest consistency in haplotype assignment.

Libraries prepared using the IDT xGen® Lockdown® target capture method were sequenced in an independent MiSeq® run with no other libraries. Initial cluster counts of 947 K/mm^2 were slightly below the recommended range of 1000-1200 K/mm^2. The percentage of clusters passing filter was very low (18.12%). No fastq files were produced for analysis. The same library was resequenced on the same day to eliminate the possibility that run failure was a result of instrumentation error. In this run, the cluster count of 1202 K/mm^2 was slightly higher than the recommended range. However, the percentage of clusters passing filter fell to 0% and no data was generated for analysis.

Data Analysis

Analyzable data was obtained for libraries prepared using the Agilent Technologies SureSelect^{XT} target capture method and the multiplex PCR strategy developed in-house. Pooled libraries prepared using each method were sequenced in independent NGS runs on the Illumina® MiSeq® with no other libraries. Reads were mapped to the rCRS using CLC Genomics Workbench v8.0 and variant calling was performed using the Low Frequency Variant caller with a 10% required significance level, and a 10% minimum frequency threshold setting. Average depths of coverage obtained for sample libraries prepared with each enrichment method were compared to determine which method, if any, yields higher average coverage overall. These data suggest that multiplex amplification is a more effective enrichment method, as 5 samples prepared using SureSelect^{XT} yielded higher average depth of coverage versus 8 samples prepared using multiplex PCR (figure 19). It should be noted that samples BUC1 and HS1 were obtained from same donor as were samples BUC2 and HS2. Higher depth of coverage was achieved for samples BUC1 and HS1 using multiplex PCR amplification.

Conversely, higher depth of coverage was achieved for samples BUC1 and HS1 using SureSelect^{XT} enrichment. These data may suggest that successful enrichment is sequence dependent, however, additional work is needed to verify this.

Data was further analyzed to determine the percentage of the human mitochondrial genome with coverage of zero obtained using each method. In general, the SureSelect^{XT} method resulted in full coverage of the whole genome with minimal gaps (figure 20), except with highly compromised samples such as cremated human remains, though analysis of these samples was also not possible using multiplex amplification. Furthermore, depth of coverage was more even and consistent across the genome for SureSelect^{XT} libraries (example shown in figure 21a), though some samples (BUC1, FDH2, BT and HL60) did perform as well as or better with multiplex amplification (example shown in figure 21b). However, data obtained from these same samples enriched using SureSelect^{XT} was generally of high quality and even depth of coverage overall, except HL60, which resulted in data that was unanalyzable.

To further evaluate each enrichment strategy, quality statistics were assessed for each sample library. FASTQ files were imported into Galaxy (Goecks, 2010) and Illumina® quality scores were converted to Sanger-type PHRED scores using the FASTQ Groomer (Blankenburg, 2010). Quality score boxplots were constructed using the FASTX-toolkit developed by Assaf Gordon (figures 22a-d). Similar to PHRED scores computed for Sanger sequencing data, an NGS quality score is a prediction of the probability of an error in base calling. Base calls with a maximum q-score of 40 are associated with an error probability of 1 in 10,000 while a q-score of 10 represents an error probability of 1 in 10. For Illumina® data, quality predictor values are used to derive the q-score of each base call. These values include parameters such as depth of coverage, cluster intensity, and signal-to-noise ratios to name a few. Quality scores less than 20 are typically considered poor for NGS data often leading to increased levels of false-positive variant calls. Q-score data assessed for this experiment appears to be correlated with average depth of coverage achieved for each library, which is not unexpected. In the example given in figures 22a and b, average coverage for sample OH was 496.8 and 2,323.51 for multiplex amplified and SureSelect^{XT} libraries respectively. Q-score distributions for the SureSelect^{XT} library were significantly higher than those obtained using multiplex PCR enrichment of the same sample. On the other hand, average depth of coverage for sample BT was 23,915.47 for the multiplex amplified library (22c) and 12,056.76 for the SureSelect^{XT} library (22d). In this example, q-score distributions are higher for the multiplex amplified library than for the library prepared using SureSelect^{XT}. However, in general, q-scores and depths of coverage for this sample are acceptable for both enrichment treatments.

Resulting rCRS variants with frequencies >70% for each sample were uploaded into HaploGrep2 to identify the haplogroup of each donor and to assess concordance between treatments (Kloss-Brandstätter, 2011; Weissensteiner, 2016) (table 33). In general, haplogroup concordance was observed across all samples originating from a specified donor. However, several incongruities were detected in high frequency variant calls between some of these samples. For example, several expected variants belonging to the designated haplogroup for sample BUC2SS (U5a2c3) were marked as missing. Upon

further analysis, these variants were identified at high frequencies in pile-up data (example shown in figure 23). The data was then re-analyzed using both additional variant calling algorithms in CLC Genomics Workbench (Fixed Ploidy and Basic Variant Detection) with quality filtering parameters that were similar to those used with the Low Frequency Variant Detector. The questioned variants were all called when using the alternative variant callers. The Low Frequency Variant caller has a built-in proprietary error correction model that aids in the removal of "sequencing errors". However, we have shown that this often results in removal of true biological variants from the data set that are high coverage, high frequency and have equal forward and reverse read balance.

Figure 19: *Average depth of coverage of multiplex amplified versus Agilent Technologies SureSelect^{XT} libraries.*

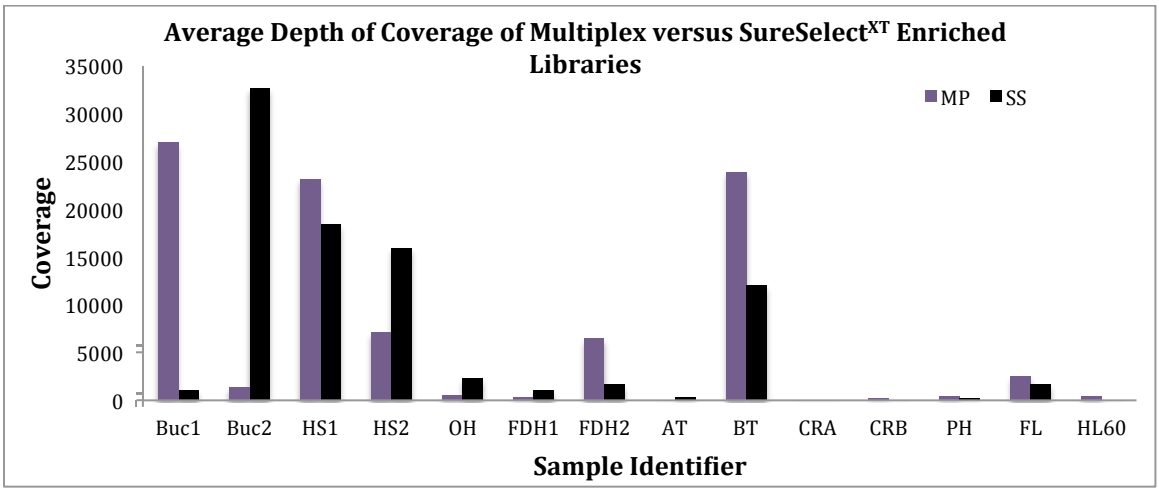


Figure 19: The average depth of coverage for each sample was assessed and compared to determine which enrichment strategy, if any, gives rise to higher depths of coverage. This data suggests that multiplex amplification is a more efficient enrichment strategy than SureSelect^{XT} probe capture. However, additional data analysis is needed to make this conclusion.

Figure 20: *Total number of positions with zero coverage in the human mitochondrial genome in libraries prepared using multiplex targeted amplification or Agilent Technologies SureSelect^{XT} probe capture enrichment.*

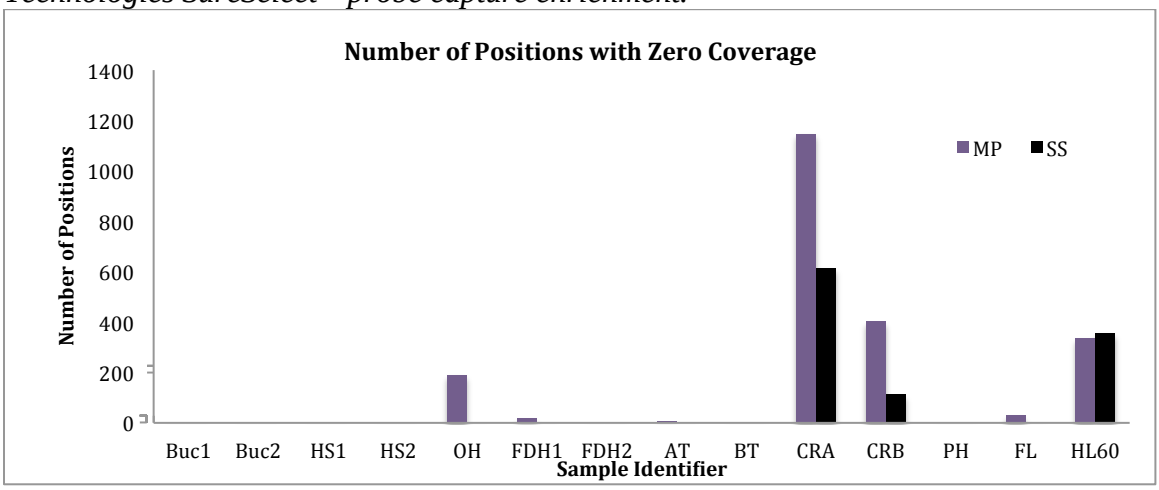
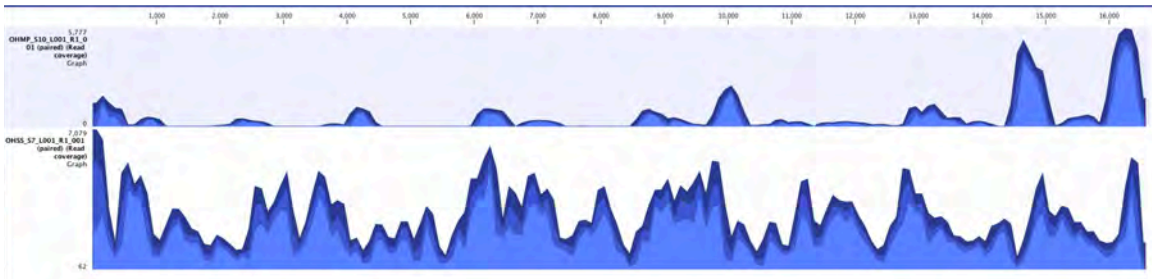


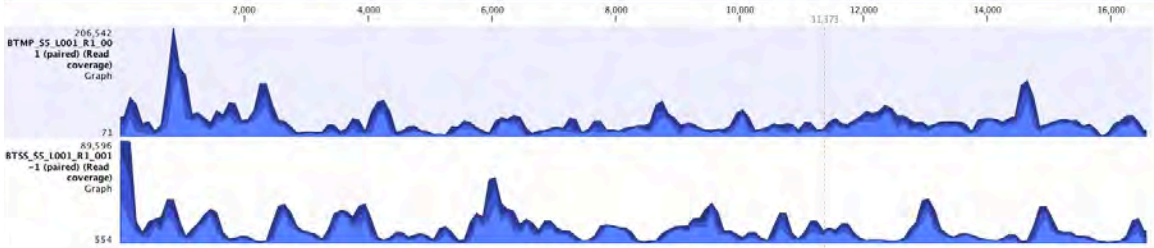
Figure 20: Several samples had full genome coverage using both enrichment strategies (BUC1, BUC2, HS1, HS2, FDH2, BT and PH). Of the remaining samples only 1 (HL60) exhibited a higher number of uncovered regions in the genome when using SureSelect^{XT} versus multiplex amplification. Conversely, 6 samples (OH, FDH1, AT, CRA, CRB and FL) had higher numbers of uncovered positions when prepared with multiplex PCR amplification. This data suggests that SureSelect^{XT} is a superior enrichment method for achieving full coverage of the targeted region of compromised sample types versus multiplex amplification.

Figures 21a and b: *Coverage maps for samples OH (21a) and BT (21b).*

21a



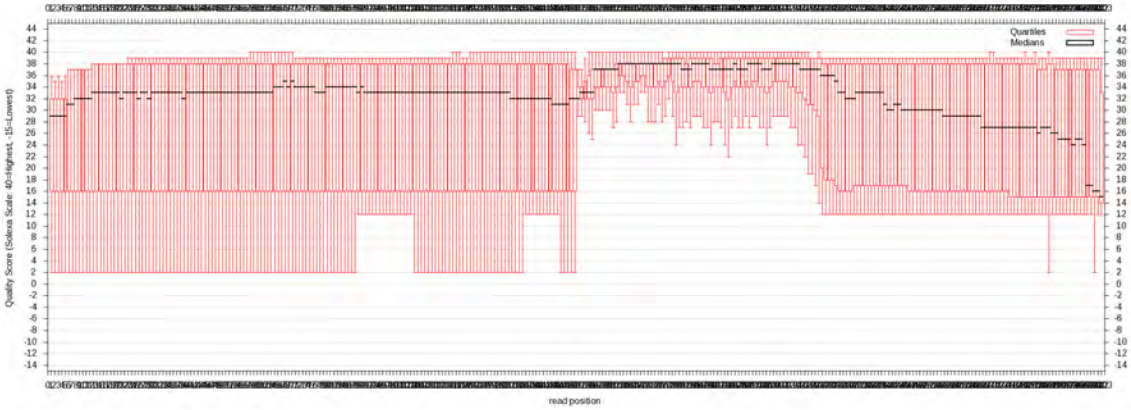
21b



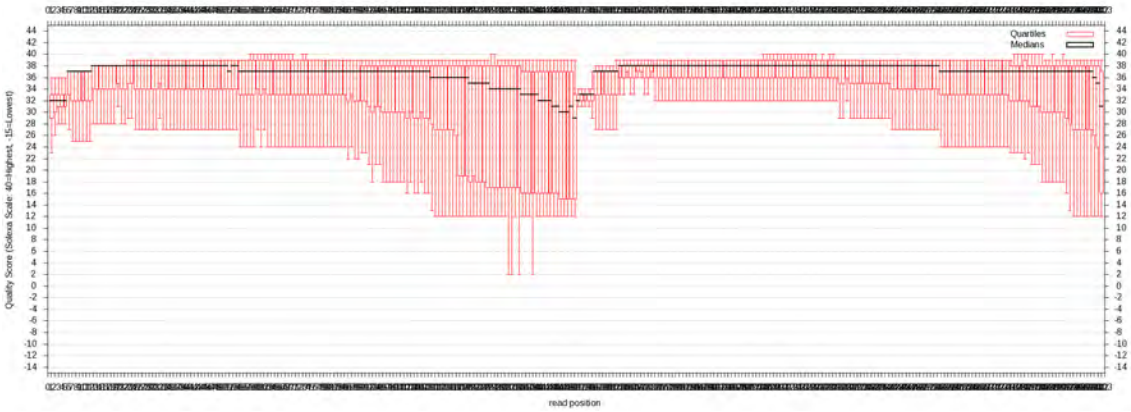
Figures 21a and b: For a majority of samples, coverage across the human mitochondrial genome was more consistent when using SureSelect^{XT} for library preparation. Figure 21a shows coverage for sample OH when using multiplex amplification (top image) versus SureSelect^{XT} (bottom image). In this instance, several regions of the genome are either not covered at all, or have very low coverage when multiplex amplification is used. SureSelect^{XT} libraries yield relatively even full coverage of the whole genome for this particular sample. Figure 21b shows coverage maps for sample BT in which libraries were prepared with multiplex amplification (top image) and SureSelect^{XT} (bottom image). In this example, average depth of coverage is higher overall for the multiplex amplified library. However, both libraries yielded relatively even coverage across the genome with no areas of zero coverage.

Figures 22a-d: *Quality score boxplots for samples OH (a and b) and BT (c and d).*

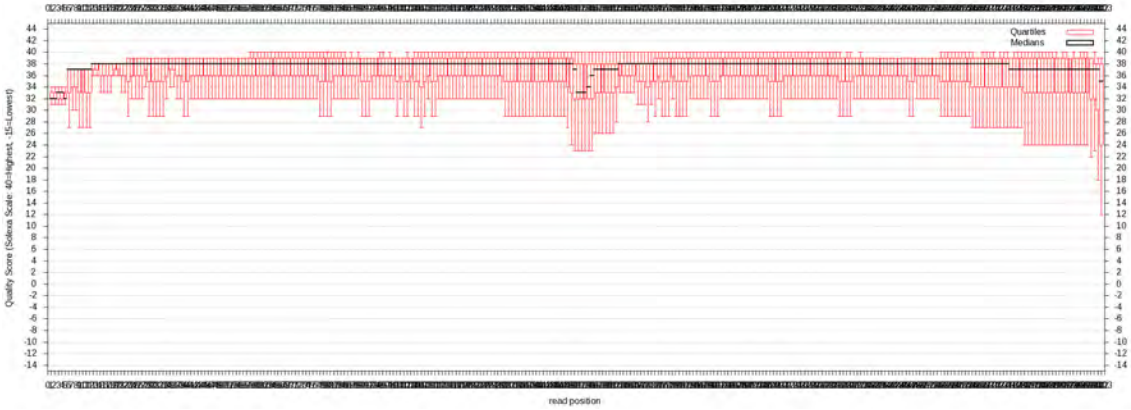
22a – Library OH prepared using multiplex amplification.



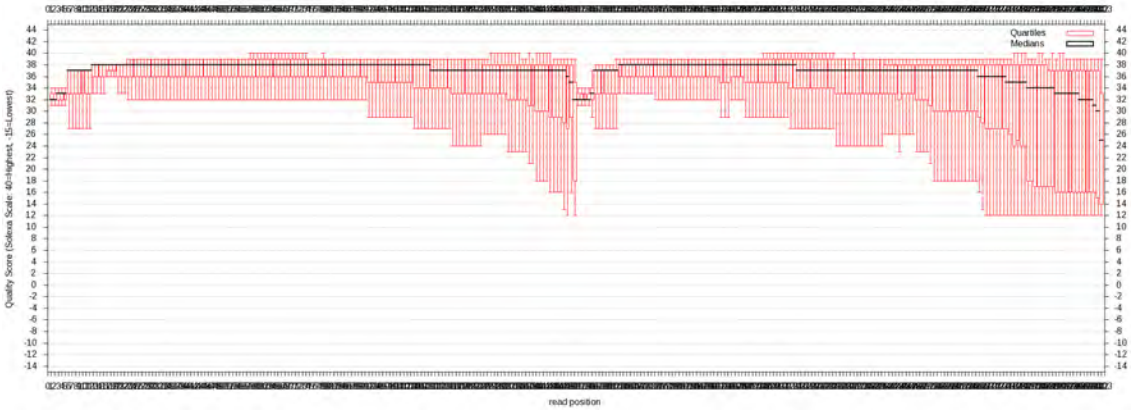
22b – Library OH prepared using SureSelect^{XT} probe capture.



22c – Library BT prepared using multiplex amplification.



22d – Library BT prepared using SureSelect^{XT} probe capture.



Figures 22a-d: The figures above show quality score distributions across read positions for each specified library. A 2 x 151 cycle paired-end sequencing approach was used where 150 bp of each molecule is sequenced in one direction, the molecule is turned around and is sequenced for 150 bp in the opposite direction. In the boxplots above, quality scores are given for all 300 cycles of a given run with paired-end turnaround and index reads in the center of each plot. Generally speaking, the quality of base calls will decrease towards the end of a read. Figures 22a and b were derived from sample OH libraries prepared using multiplex PCR amplification and SureSelect^{XT} probe capture enrichment respectively. Data quality for the multiplex amplified library is inferior to data obtained from the SureSelect^{XT} library. Boxplots 22c and d were derived from sample BT libraries prepared using multiplex PCR amplification and SureSelect^{XT} probe capture enrichment respectively. In this case, the data quality is higher for the multiplex amplified library. These differences appear to be directly related to the average depth of coverage obtained for each library.

Table 33: *Haplogroup assignments for all samples prepared for NGS using SureSelect^{XT} and multiplex PCR amplification.*

	SureSelect^{XT} Samples			Multiplex Amplified Samples		
Sample	Haplogroup	Quality	Unexpected Mutations or missing SNPs?	Haplogroup	Quality	Unexpected Mutations or missing SNPs?
BUC1	V12	100%	8520 local private mutation	V12	100%	8520 local private mutation
BUC2	U5a2c3	82.36%	Missing 2706 (present in pile-up, not called), 3197 (present in pile-up, not called), 10619 (present in pile-up, not called), 12372 (present in pile-up, not called), 14766 (present in pile-up, not called), 16526 (present in pile-up, not called); 15299, 16223 local private mutations; 3107d hotspot	U5a2c3a	92.30%	Missing 14793; 15299, 16223 local private mutations
HS1	V12	100.00 %	8520 local private mutation	V12	100%	8520 local private mutation
HS2	U5a2c3a	94.91%	Missing 14793 (present in pile-up, not called); 15299, 16223 local private mutations	U5a2c3a	94.91%	Missing 14793 (present in pile-up, not called); 15299, 16223 local private mutations.
OH	H2a1 + 146	93.02%	12795A, local private mutation	H2a1 + 146	100%	No
FDH1	H1a1	100.00 %	3107d, 16519 hotspots; 16209C present, but not	H1a	96.77%	3107d hotspot; 16209 local private mutation; coverage of 4 at 16519

			called			
FDH2	H3	95.19%	3918 - local private mutation; 3107d, 16519 hotspots	H3	95.19%	3918 - local private mutation; 3107d, 16519 hotspots
AT	T2b13	96.92%	Missing 709 (present in pile-up, not called), 4216 (coverage of 33 at 4216); 3705 local private mutation; 16519 hotspot	U5a'b	60.5%	Missing several variants. Low coverage overall. Possible contamination from donor 1.
BT	U5b1c2	94.10%	Missing 150 (present in pile-up, not called), 5656 (present in pile-up, not called), 16192; 9110 local private mutation; 146C present in pile-up, not called	U5b1c2	93.71%	Missing 750 (present at a frequency of 55%), 16192, 16311 (present in pile-up, not called); 146, 9110 local private mutations
CRA			Not enough data for analysis			Not enough data for analysis
CRB			Not enough data for analysis	U5a'b	93.72%	720d global private mutation; 15299 local private mutation. Possible contamination from donor 1.
PH	U5b1b1g1a	98.68%	Missing 150 (present in pile-up, not called), 16192; 3107d hotspot; 189G and 199C present in pile-up, not called	U5b1b1g1a	92.65%	Missing 16192; 189 and 199 local private mutations
FL	H2a2a2	76.34%	263 local private mutation; 3107d hotspot	H2a2a2	66.67%	263, 11017, 16172 local private mutations; 3017d hotspot
HL60			Not enough data for analysis	J2b1a1a	74.46%	Many calls missing due to no coverage

Table 33: Variants from the rCRS were obtained for each sample. Variants were uploaded into HaploGrep2 and haplogroups were identified. Quality values are defined by HaploGrep based on how well each set of variants matches the particular haplogroup identified for the specified donor. In most cases, haplogroup assignment is concordant between samples prepared with different enrichment strategies. In some instances, one enrichment method yields a more highly resolved haplogroup than the other. This is typically a result of an increase in coverage across the genome. There are several discrepancies in variants called between sample treatments. Most of these discrepancies are a result of the variant calling algorithm used and can actually be identified in pile-up data. Reanalysis of the data using a different variant caller typically resolves differences. However, additional anomalies often appear in data after reanalysis. For example, in sample FDH1 SureSelect^{XT}, variant 16209C is not called when the Low Frequency Variant Detector is used even though it can be clearly seen in a majority of reads in pile-up data. When the data is reanalyzed with the Fixed Ploidy Variant Detector, 16209C is called but variant 4769G, which is called with the Low Frequency Variant Detector, drops out.

Table 34: *Differences in data output in sample BUC2 prepared with SureSelect^{XT} when analyzed using different variant detection algorithms.*

		Basic Variant Detection				Fixed Ploidy Variant Detection				Low Frequency Variant Detection			
Position	rCRS	Variant	Coverage	Frequency (%)	Read balance	Variant	Coverage	Frequency (%)	Read balance	Variant	Coverage	Frequency (%)	Read balance
73	A	G	10188	98.32	0.37	G	10188	98.32	0.37	G	83303	98.29	0.5
263	A	G	4009	98.8	0.26	G	4009	98.8	0.26	G	15997	96.04	0.35
750	A	G	10866	93.97	0.48	G	10866	93.97	0.48	G	65048	95.98	0.5
1438	A	G	11355	98.64	0.48	G	11355	98.64	0.48	G	56434	96.58	0.5
2706	A	G	7736	91.53	0.4	G	7736	91.53	0.4				
3107	N					-	8486	97.18	0.41	-	39324	93.15	0.44
3197	T	C	4900	98.08	0.41	C	4900	98.08	0.41				
4769	A	G	4675	97.54	0.47	G	4675	97.54	0.47	G	23997	95.62	0.49
7028	C	T	16926	98.22	0.49	T	16926	98.22	0.49	T	76963	97.35	0.48
8860	A	G	8626	99.14	0.45	G	8626	99.14	0.45	G	41442	97.69	0.48
9477	G	A	5489	93.68	0.49	A	5489	93.68	0.49	A	28560	91.68	0.49
10619	C	T	12670	97.25	0.44	T	12670	97.25	0.44				
10709	A	C	14393	97.62	0.43	C	14393	97.62	0.43				
11465	T	C	6207	97.34	0.5	C	6207	97.34	0.5	C	32170	96.32	0.48
11467	A	G	6198	97.92	0.5	G	6198	97.92	0.5	G	32221	96.95	0.49
11719	G	A	10340	95.81	0.47	A	10340	95.81	0.47	A	46297	95.54	0.48
12308	A	G	4648	97.18	0.43	G	4648	97.18	0.43	G	24108	95.02	0.45
12372	G	A	4220	97.11	0.46	A	4220	97.11	0.46				
13617	T	C	4667	95.44	0.43	C	4667	95.44	0.43	C	24855	94.24	0.45
14766	C	T	3034	94.79	0.46	T	3034	94.79	0.46				
14793	A	G	3749	96.13	0.42	G	3749	96.13	0.42	G	19056	94.07	0.44
15299	T	C	7976	97.59	0.43	C	7976	97.59	0.43	C	45543	96.02	0.45
15326	A	G	7286	99.09	0.45	G	7286	99.09	0.45	G	42947	97.04	0.47
16223	C	T	7755	96.91	0.47	T	7755	96.91	0.47	T	28252	95.02	0.48
16256	C	T	7977	95.39	0.47	T	7977	95.39	0.47	T	30278	94.88	0.48
16270	C	T	8597	96.1	0.44	T	8597	96.1	0.44	T	32265	94.88	0.46
16526	G	A	2535	97.12	0.13	A	2535	97.12	0.13				

Table 34: FASTQ files were uploaded into CLC Genomics Workbench v8.0. Data was analyzed using three different variant calling algorithms with similar filtering parameters and data outputs were compared. Variants called using the Fixed Ploidy and Basic Variant Detection options were almost identical with the exception that 3107d is called with the Fixed Ploidy algorithm and not with the Basic Variant Detector (highlighted in yellow). Seven true biological variants are omitted from the data set when the Low Frequency Variant Detector (highlighted in red) is used.

Figure 23: *Pile-up data for sample BUC2SS showing a majority of G residues at position 2706.*

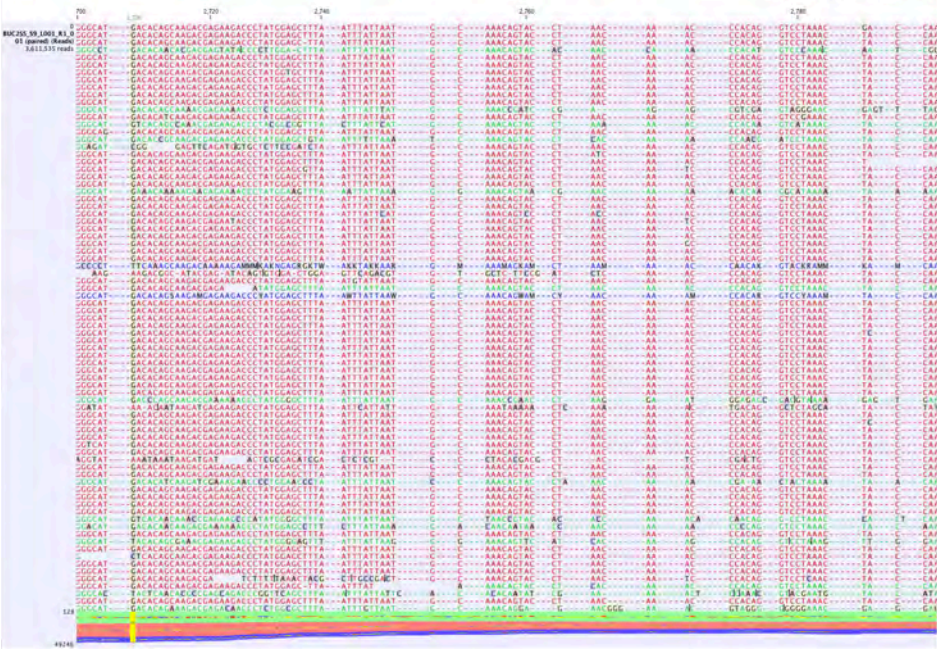


Figure 23: The pile-up data above shows an obvious difference from the rCRS in a majority of reads at position 2706 in BUC2SS data. This particular variant is called when using the Fixed Ploidy and Basic Variant Detection algorithms. However, it is omitted from the output when the Low Frequency Variant Detector is used.

Conclusions

Several enrichment strategies were compared for the ability to enable analysis of human mitochondrial DNA from highly compromised sample types. Two methods, IDT xGen® Lockdown® target capture and Sygnis® TruePrime™ whole genome amplification methods did not work well in our hands. Further experimentation may elucidate reasons why. Multiplex PCR amplification of the whole human mtGenome and Agilent SureSelect^{XT} target enrichment strategies worked well with a myriad of sample types. Average coverage across the genome was comparable for both methods, however consistency in coverage was higher in SureSelect^{XT} data overall. Additionally, the number of positions with coverage of zero was lower in SureSelect^{XT} data. In general, data quality seemed to correlate with average coverage. Haplogroup assignments were concordant between samples originating from the sample donor prepared using each enrichment method. Some anomalies were observed in variants called, but in most cases these issues were linked back to variant calling algorithms used to generate the data. Caution should be used when assessing variant tables generated with any of the variant calling options in CLC Genomics Workbench as major differences are observed in data outputs when using these different methods. None of the enrichment strategies enabled analysis of human cremated remains. Multiplex PCR amplification lead to analyzable data with low input, high molecular weight HL60 DNA while SureSelect^{XT} did not. Conversely, SureSelect^{XT} enabled analysis of DNA from a compromised tooth sample unearthed from a clay burial site after an estimated 200 years. Data obtained for this sample when using multiplex PCR amplification was likely a result of contamination from donor 1. Based on these findings, no recommendations can be made as to whether one enrichment method outperforms the other. While data obtained with SureSelect^{XT} is slightly better in many cases than that obtained with multiplex PCR amplification (due mainly to higher coverage attained), the cost and labor associated with the kit is prohibitive for many crime laboratories. It may be most appropriate for crime laboratories to employ multiplex PCR amplification for whole genome analysis of human mitochondrial DNA from forensic samples since variants called and haplogroups assigned are concordant between samples prepared using both strategies.

Dissemination of findings

The following grant-related presentations have been given:

<u>68th Annual Meeting of the American Academy of Forensic Sciences (AAFS) 2016, Las Vegas, NV</u>
Poster: *Assessment of low-level error in massively-parallel sequencing (MPS) data sets generated using the Illumina® MiSeq® platform and synthesized human mitochondrial DNA oligonucleotides.* B.J. Bintz and M.R. Wilson.

Federal Bureau of Investigation, 2015, Quantico, VA
Invited Talk: *Development of a multiplex Droplet Digital™ PCR (ddPCR™) assay for simultaneous absolute quantitation of human nuclear and mitochondrial DNA.* B.J. Bintz.

26th International Symposium on Human Identification 2015, Grapevine, TX
Poster: *Assessment of low-level error in massively-parallel sequencing (MPS) data sets generated using the Illumina® MiSeq® platform and synthesized human mitochondrial DNA oligonucleotides.* B.J. Bintz and M.R. Wilson.

Poster: *Amplification of whole mitochondrial genome from challenging samples via multiplex PCR assay.* M.P. Hickman, E.S. Burnside, B.J. Bintz, K.S. Grisedale, N. Petraco, E.K. Hanson, J. Ballantyne, and M.R. Wilson.

Poster: *Use of massively parallel sequencing (MPS) to assist with deconvolution of STR mixture profiles.* K.S. Grisedale, B.J. Bintz, and M.R. Wilson.

Defense Forensic Science Center, 2015, Atlanta, GA
Talk: *Ongoing Research in the Forensic Science Program at WCU.* B.J. Bintz and M.R. Wilson.

67th Annual Meeting of the American Academy of Forensic Sciences (AAFS) 2015, Orlando, FL
Talk: *Development of a multiplex quantitative PCR (qPCR) assay for simultaneous quantification of human nuclear and mitochondrial DNA from forensically relevant samples.* B.J. Bintz and M.R. Wilson.

Poster: *Optimization of a method for the extraction of DNA from human skeletal remains.* Presented by S. Deaton, B.J. Bintz, and M.R. Wilson.

25th International Symposium on Human Identification 2014, Phoenix, AZ
Poster: *Development of a multiplex quantitative PCR (qPCR) assay for simultaneous quantification of human nuclear and mitochondrial DNA.* B.J. Bintz and M.R. Wilson.

Poster: *Optimization of a method for the extraction of DNA from human skeletal remains.* S. Deaton, B.J. Bintz, and M.R. Wilson.

66th Annual Meeting of the American Academy of Forensic Sciences (AAFS) 2014, Seattle, WA
Poster: *An evaluation of next-generation sequencing (NGS) instrumentation and commercially available bioinformatics software tools for forensic mitochondrial DNA analysis.* B.J. Bintz, E.S. Burnside, K. Kiesler, K. Gettings, P.M. Vallone, and M.R. Wilson.

Our optimized DNA extraction method has successfully been transferred to the FBI Laboratory in Quantico, VA, where the mitochondrial DNA Unit has incorporated it into casework.

Dr. Wilson provided a keynote address at the 9th International Conference on Forensic Inference and Statistics in Leiden, Netherlands on August 21, 2014. The presentation outlined the goals and some of the results of this project.

Brittania Bintz and Maureen Hickman organized a Next-Generation Sequencing Workshop that was held at Western Carolina University entitled *Tackling Big Data: Next-Generation Sequencing from Sample Prep to Data Analysis*. Invited speakers included local scientists, and representatives from Illumina® and Life Technologies. Ms. Hickman presented NIJ funded research in a talk entitled *Amplification of the whole mitochondrial genome from challenging samples via multiplex PCR assay*.

Pending presentations include:

69th Annual Meeting of the American Academy of Forensic Sciences (AAFS) 2017, New Orleans, LA

Poster: *Optimization of a droplet digital™ PCR (ddPCR™) assay for quantitative and qualitative analysis of Illumina® Miseq® massively-parallel sequencing (MPS) libraries.* B.J. Bintz.

Talk: *Ashes to ashes: Analysis of enhanced methods for genetic identification of human cremated remains.* K.S. Grisedale.

Participants & Other Collaborating Organizations

What individuals have worked on the project?

Name: Brittania Bintz
Project Role: Research Scientist, Forensic Science Program; Principle Investigator
Contribution to Project: Ms. Bintz has performed work in the area of modified and primer design, improved DNA extraction from hair shaft, PCR protocol development, development of quantitation assays (both qPCR and ddPCR™), DNA extraction and quantitation, amplification efficiency comparison, and operation of the NGS instruments including comprehensive assessment of error in MiSeq™ data sets.

Name: Mark Wilson
Project Role: Principle Investigator
Contribution to Project: Mark Wilson performed administrative duties and organization of grant-related research until he left WCU in October of 2015.

Name: Kelly Grisedale, Ph.D.
Project Role: Associate Professor, Forensic Science Program, Biology Dept.
Contribution to Project: Dr. Grisedale has performed work in the area of DNA extraction from bones, STR mixture deconvolution using NGS, DNA extraction and quantitation, amplification strategies of low-level DNA samples.

Name: Maureen Peters-Hickman, M.S.

Project Role: Research Assistant, Forensic Science Program
Contribution to Project: Ms. Peters-Hickman has performed work in the areas of multiplex amplification design, DNA extraction from bones, amplification strategies from low-level DNA samples, and operation of the NGS instruments.

What other organizations have been involved as partners?

Illumina, Inc.
9885 Towne Centre Drive
San Diego, CA 92121 USA

As detailed in the project proposal, Illumina, Inc. is collaborating with WCU in the design of experiments that will reveal the potential of the Illumina instrument in generating NGS deep sequencing data. They have also graciously agreed to provide sequencing services in support of the project and have loaned a MiSeq™ DNA Sequencing instrument and reagents to WCU in support of ongoing collaborative efforts.

Have other collaborators or contacts been involved?

The work under this NIJ grant has lead to a collaboration between WCU and Jack Ballantyne's group at UCF. We have worked with their group to generate whole mitochondrial genome data from dust bunnies and small collections of cells.

We have also collaborated with scientists at the FBI Laboratory (Mark Kavlick), the National Institute of Standards and Technology (Peter Vallone, Kevin Keisler, Katherine Gettings), CLC-Bio, Incorporated, Pennsylvania State University (Mitch Holland, Jen McElhoe), and Mitotyping Technologies (Terry Melton).

Impact

Products:

None

What is the impact on the development of the principal discipline(s) of the project?

We have shown the feasibility of newly emerging NGS methods on typical forensic DNA typing samples. We have also improved the ability to extract DNA from hair shaft, and also have begun to apply these principles to bone material. Our research results have increased the chance of using whole mt-genome analysis on challenging casework samples, significantly expanding the capabilities of the forensic DNA community. We have also developed a versatile quantitative assay that will ultimately enable simultaneous assessment of extracted nuclear and mitochondrial DNA quantity and quality.

What is the impact on other disciplines?

There is a potential for a positive impact in many areas of forensic DNA typing, including an expansion of the utility of human mtDNA in forensic casework, with the adoption of whole mt-genome analysis. There is also the potential for positive impact in the area of bioinformatics. New programs, or modifications of existing programs, may need to be developed so that minor DNA variant detection can be simplified in a user-friendly manner. Currently, the NGS analysis pipeline includes a variety of separate scripts written for a variety of purposes. Our project requires the seamless integration of many scripts into a pipeline. The development of such a tool may be useful in other disciplines within the broader disciplines of molecular biology and evolution.

What is the impact on the development of human resources?

Nothing to report.

What is the impact on physical, institutional, and information resources that form infrastructure?

The WCU Forensic Science Program has been awarded a grant from The North Carolina Biotechnology Center and has acquired a Thermo Fisher 3500xl 24-capillary DNA sequencer. This acquisition has enabled the Forensic Science Program to establish a DNA Sequencing Core Facility on the campus of WCU to provide a multitude of DNA sequencing services to university laboratories and other institutions in the surrounding areas. Additionally, the acquisition of the BioRad QX200 ddPCR™ instrument through funding provided in this grant has greatly increased the capabilities of the laboratory. The instrument enables absolute quantitation of nucleic acids using TaqMan™ probe chemistry or intercalating dye chemistry without the use of a standard curve. We will continue to consult with internal and external collaborators on how best to implement this technology into their workflows. We will also use it to assess quality and quantity of massively-parallel sequencing libraries.

What is the impact on technology transfer?

The results of this project may serve as a modification or replacement of current standard operating procedures with crime laboratories conducting human mitochondrial DNA (mtDNA) sequencing in criminal and civil casework applications.

References

Executive Summary

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

Anderson S, Bankier AT, Barrell BG, de Bruijn MH, Coulson AR, Drouin J, Eperon IC, Nierlich DP, Roe BA, Sanger F, Schreier PH, Amith AJ, Staden R, Young IG. Sequence and organization of the human mitochondrial genome. Nature. 1981; 209:457-465

Andréasson, H, Nilsson, M, Styrman, H, Pettersson, U, Allen, M. Forensic mitochondrial coding region analysis for increased discrimination using pyrosequencing technology. Forensic Sci Int Genet. 2006; 1:35-43

Andréasson H, Nilsson M, Budowle B, Frisk S, Allen M. Quantification of mtDNA mixtures in forensic evidence material using pyrosequencing. Int J Legal Med. 2006; 6:383-390.

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

Bintz BJ, Dixon GB, Wilson MR. Simultaneous detection of human mictochondrial DNA and nuclear-inserted mitochondrial-origin sequences (NumtS) using forensic mtDNA amplification strategies and pyrosequencing technology. J Forensic Sci. 2014; (4):1064-1073.

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

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

Calloway CD, Reynolds RL, Herrin Jr GL, Anderson WW. The frequency of heteroplasmy in the HVII region of mtDNA differs across tissue types and changes with age. Am J Hum Genet. 2000; 66:1384-1397.

Carracedo A, Bär W, Lincoln P, Mayr P, Morling N, Olaisen B Schneider P, Budowle B, Brinkmann B, Gill P, Holland M, Tully G, Wilson MR. DNA commission of the International Society for Forensic Genetics: guidelines for mitochondrial DNA typing. Forensic Sci Int. 2000; 110:79-85.

Coble MD, Just RS, O'Callaghan JE, Letmanyi IH, Peterson CT, *et. al.* Single Nucleotide Polymorphisms over the entire mtDNA Genome that increase the Forensic Power of mtDNA Testing in Caucasians. Int j Legal Med. 2004; 3:137-146.

Dewey FE, Grove ME, Pan CP, Goldstein PA, Berstein JA, Chaib H, *et. al.* Clinical interpretation and implications of whole-genome sequencing. JAMA. 2014; 311:1035-1044

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; (4):318-28.

Giles RE, Blanc H, Cann HM, Wallace DC. Maternal inheritance of human mitochondrial DNA. Genet. 1980; 77:6715-6719

Grisedale KS, van Daal A. Comparison of STR profiling from low template DNA extracts with and without the consensus profiling method. Investig Genet. 2012; (1):14.

Holland MM, Parson TJ. Mitochondrial DNA sequence analysis – validation and use for forensic casework. Forensic Sci Rev.1999; 11:29

Holland MM, McQuillan MR, O'Hanlon KA. Second generation sequencing allows for mtDNA mixture deconvolution and high resolution detection of heteroplasmy. Croat Med J. 2011; 3:299-313.

Irwin JA, Saunier JL, Niederstätter H, Strouss KM, Sturk KA, Diegoli TM, *et. al.* Investigation of heteroplasmy in the human mitochondrial DNA control region: a synthesis of observations from more than 5000 global population samples. J Mol Evol. 2009; 68:516-527

King JL, LaRue BL, Novroski N, Stoljarova M, Seo SB, Zeng X, *et. al.* High-quality and high-throughput massively-parallel sequencing of the human mitochondrial genome using the Illumina MiSeq. Forensic Sci Int Genet. 2014; 12:128-135.

Li M, Schonburg A, Schafer M, Schroeder R, Nasidze I, Stoneking M. Detecting heteroplasmy from high-throughput sequencing of complete human mitochondrial DNA genomes. Am J Hum Genet. 2010; 87:237-249.

Lindahl T, Nyberg B. Rate of depurination of native deoxyribonucleic acid. Biochem. 1972; 11:3610-3618.

Lindahl T, Andersson A. Rate of chain breakage at apurinic sites in double-stranded deoxyribonucleic acid. Biochem. 1972; 11:3618-3623.

McElhoe JA, Hollan MM, Makova KD, Su MS, Paul IM, Baker CH, Faith SA, Young B. Development and assessment of an optimized next-generation DNA sequencing approach for the mtGenome using the Illumina MiSeq. Forensic Sci Int Genet. 2014; 13:20-29.

Mikkelsen M, Hansen RF, Hansen AJ, Morling N. Massively parallel pyrosequencing methodology of the mitochondrial genome in forensic genetics. Forensic Sci Int Genet. 2014; 12:30-37.

Monson KL, Miller KWP, Wilson MR, DiZinno JA, 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

Naue J, Horer S, Sanger T, Strobl C, Hatzer-Grubwieser P, Parson W, et. al. Evidence for frequent and tissue-specific heteroplasmy in human mitochondrial DNA. Mitochondrion. 2015; 20:82-94

Parsons TJ, Coble MD. Increasing the forensic discrimination of mitochondrial DNA testing through analysis of the entire mitochondrial DNA genome. Croatian Medical J. 2001; 3: 304-309.

Parson W, Gusmao L, Hares DR, Irwin JA, Mayr WR, Morling N, *et. al.* DNA commission for the International Society for Forensic Genetics: revised and extended guidelines for mitochondrial DNA typing. For Sci Int Genet. 2014; 13:134-142

Parson W, Strobl C, Huber G, Zimmerman B, Gomes SM, Souto L, *et. al.* Evaluation of next-generation mtGenome sequencing using the Ion Torrent Personal Genome Machine (PGM). Forensic Sci Int Genet. 2013; 7:543-549.

Peck MA, Brandhagen MD, Marshall C, Diegoli TM, Irwin JA, Sturk KA. Concordance and reproducibility of a next generation mtGenome sequencing method for high-quality samples using the Illumina MiSeq. Forensic Sci Int Genet. 2016; 24:103-111.

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

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

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

Robin ED, Wong R. Mitochondrial DNA molecules and virtual number of mitochondria per cell in mammalian cells. J Cell Physiol. 1988; 136:507-5013.

Sekiguchi K, Sato H, Kasai K. Mitochondrial DNA heteroplasmy among hairs from single individuals. J Forensic Sci. 2004; 49:986-991.

Sosa MX, Sivakumar IK, Maragh S, Veermachaneni V, Hariharan R, Parulekar M, *et. al.* Next-generation sequencing of human mitochondrial reference genomes uncovers high heteroplasmy frequency. PLoS Comput Biol. 2012; 10:e1002737

SWGDAM, Guidelines for mitochondrial (mtDNA) nucleotide sequence interpretation. Forensic Sci. 2003

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

Vigilant L. An Evaluation of Techniques for the Extraction and Amplification of DNA from

Naturally Shed Hairs. Biol Chem. 1999; 380:1329-1331.

Wallace DC, Brown MD, Lott MT. Mitochondrial DNA variation in human evolution and disease. Gene 1999; 238:211-230.

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

Wilson MR, Polanskey D, Butler JM, DiZinno JA, Replogle J, Budowle, B. Extraction, PCR Amplification, and Sequencing of Mitochondrial DNA from Human Hair Shafts. BioTechniques. 1995; 4:662-669.

Wilson MR, Allard MW. Phylogenetic and mitochondrial DNA analysis in the forensic sciences. Forensic Science Reviews. 2004;16:37-62.

Wolstenholme DR. Animal mitochondrial-DNA: structure and evolution. Int Rev Cytol. 1992; 141:173-216.39.

New Developments in Cancer Diagnostics and Human Mitochondrial DNA Variation

Calabria I, Pedrola L, Berlanga P, Aparisi MJ, Sánchez-Izquierdo D, Cañete A, Cervera J, Millán JM, Castel V. The new challenge in oncology: Next-generation sequencing and its application in precision medicine. An Pedriatr. 2016; 85:273.e1-273.e7.

Cha YJ, Koo JS. Next-generation sequencing in thyroid cancer. J Transl Med 2016; 14:322-332.

Coller HA. High frequency of homoplasmic mitochondrial DNA mutations in human tumors can be explained without selection. Nature Genet. 2001; 28:147-150.

Greaves LC. Quantification of mitochondrial DNA mutation load. Aging Cell. 2009; 8:566 - 572.

He Y, Wu J, Dressman DC, Iacobuzio-Donahue C, Markowitz SD, Velculescu VE, Diaz Jr LA, Kinzler KW, Vogelstein B, Papadopoulos N. Heteroplasmic mitochondrial DNA mutations in normal and tumour cells. Nature. 2010; 464:610-614.

Legros F, Malka F, Frachon P, Lombes A, Rojo M. Organization and dynamics of human mitochondrial DNA. J Cell Sci. 2004; 117:2653-2662.

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; 286:774-779.

Santos C. Frequency and pattern of heteroplasmy in the control region of human mitochondrial DNA. J Mol Evol. 2008; 67:191-200.

Sabour L, Sabour M, Ghorbian S. Clinical application of next-generation sequencing in

cancer diagnosis. Patho Oncol Res. 2016; 1-10

Sekiguchi, K, Kasai, K, Levin, BC. 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; 2:401-414.

Wallace DC, Chalkia D. Mitochondrial DNA Genetics and the heteroplasmy conundrum in evolution and disease. Cold Spring Harbor Perspect Biol. 2013

White HE. Accurate detection and quantitation of heteroplasmic mitochondrial point mutations by pyrosequencing. Genet Test. 2005; 9:190-199.

Wu H, Wang C, Wu S. Single-cell sequencing for drug discovery and drug development. Curr Top Med Chem. 2016; e-pub

Zsurka G. Recombination of mitochondrial DNA in skeletal muscle of individuals with multiple mitochondrial DNA heteroplasmy. Nature Genet. 2005; 37:873-877

Next Generation Sequencing Applications

Alkan C, Kidd JM, Marques-Bonet T, Aksay G, Antonacci F, *et. al.* Personalized copy number and segmental duplication maps using next-generation sequencing. Nat Genet. 2009; 10:1061-1067.

Alkan C, Coe BP, Eichler EE. Genome structural variation discovery and genotyping. Nat Rev Genet. 2011; 5:363-376.

Alkan C, Sajjadian S, Eichler EE. Limitations of next-generation genome sequence assembly. Nat Methods. 2011; 1:61-65.

Antonacci F, Kidd JM, Marques-Bonet T, Taegue B, Ventura M, *et. al.* A large and complex structural polymorphism at 16p12.1 underlies microdeletion disease risk. Nat Genet. 2010; 9:745-750.

Batra VK, Beard WA, Pedersen LC, Wilson SH. Structures of DNA polymerase mispaired DNA termini transitioning to pre-catalytic complexes support an induced fit fidelity mechanism. Structure. 2016; 11:1855-1856.

Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, *et. al.* Accurate whole genome sequencing using reversible terminator chemistry. Nature. 2008; 7218:53-59.

Brown JR, Roy S, Yara Romera E, Shah D, Williams R, Breuer J. Norovirus whole-genome sequencing by SureSelect target enrichment: a robust and sensitive method. J Clin Microbiol. 2016; 10:2530-2537.

Carneiro MO, Russ C, Ross MG, Gabriel SB, Nusbaum C, DePristo MA. Pacific biosciences sequencing technology for genotyping and variation discovery in human data. BMC Genomics. 2012; 13:375.

Chain PSG, Grafham DV, Fulton RS, FitzGerald MG, Hostetler J, *et. al.* Genome project standards in a new era of sequencing. Science. 2009; 5950:236-237.

Chaitanya L, Ralf, A, van Oven M, Kupiec T, Chang J, *et. al.* Simultaneous whole mitochondrial genome sequencing with short overlapping amplicons suitable for degraded DNA using the Ion Torrent Personal Genome Machine. Hum Mutat. 2015; 12:1236-1247.

Dalloul RA, Long JA, Zimin AV, Aslam L, Beal K, *et. al.* Multi-platform next-generation sequencing of the domestic turkey (Meleagris gallopavo): genome assembly and analysis. PLoS Biol. 2010; 9.

Dewey F, Ashley E, Quartermous T. Interpreting whole-genome sequencing – reply. JAMA. 2014; 3:296-297

Eckert KA, Kunkel TA. DNA polymerase fidelity and the polymerase chain reaction. PCR Methods Appl. 1991; 1:17-24.

Eid J, Fehr A, Gray J, Luong K, Lyle J, *et. al.* Real-time DNA sequencing from single polymerase molecules. Science. 2009; 5910:133-138.

ElSharawy A, Warner J, Olson J, Forster M, Schilhabel MB, Link DR, *et. al.* Accurate variant detection across non-amplified and whole genome amplified DNA using targeted next-generation sequencing. BMC Genomics. 2012; 13:500.

Endrullat C, Glökler J, Franke P, Frohme M. Standardization and quality management in next-generation sequencing. Appl Transl Genom. 2016; 10:2-9.

Gadipally SR, Sarkar A, Nandieneni MR. Selective enrichment of STRs for applications in forensic human identification. Electrophoresis. 2015; 15:1768-1774.

Garcia-Garcia G, Baux D, Faugère V, Moclyn M, Koenig M, Claustres M, Roux AF. Assessment of the latest NGS enrichment capture methods in clinical context. Sci Rep. 2016: (6):20948.

Gargis AS, Kalman L, Berry MW, Bick DP, Hambuch T, *et. al.* Assuring the quality of next-generation sequencing in clinical laboratory practice. Nat Biotechnology. 2012; 11:1033-1036.

Giardina E, Pietrangeli I, Martone C, Zampaiit S, Marsala P, *et. al*. Whole genome amplification and real-time PCR in forensic casework. BMC Genomics. 2009; 10:159.

Glenn TC. Field guide to next-generation DNA sequencers. Mol Ecol Resour. 2011; 5:759-769.

Gnerre S, Maccallum I, Przybylski D, Ribeiro FJ, Burton JN, *et. al.* High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc Natl Acad Sci USA 2011; 4:1513-1518.

Goecks J, Nekrutenko A, Taylor J, *et. al.* Galaxy: a comprehensive approach for supporting

accessible, reproducible, and transparent computational research in the life sciences. Genome Biol. 2010; 8:R86.

Harris TD, Buzby PR, Babcock H, Beer E, Bowers J, *et. al.* Single-molecule DNA sequencing of a viral genome. Science. 2008; 320:106-109.

Handsaker RE, Korn JM, Namesh J, McCarroll SA. Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. Nat Genet. 2011; 3:269-276.

Huddleston J, Chaisson MJ, Meltz Steinberg K, Warren W, Hoekzema K, Gordon DS, *et. al.* Discovery and genotyping of structural variation from long-read haploid genome sequence data. Genome Res. 2016; Epub.

Iqbal SM, Akin D, Bashir R. Solid-state nanopore channels with DNA selectivity. Nat Nanotechnology. 2007; 4:243-248.

Jovel J, Patterson J, Wang W, Hotte N, O'Keefe S, *et. al.* Characterization of the gut microbiome using 16s or shotgun metagenomics. Front Microbiol. 2016; 7:459.

Just RS, Irwin JA, Parson W. Mitochondrial DNA heteroplasmy in the emerging field of massively-parallel sequencing. Forensic Sci Int Genet. 2015; 18:131-139.

Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, Graves T, *et. al.* Mapping and sequencing of structural variation from eight human genomes. Nature. 2008; 7191:56-64.

Kidd JM, Sampas N, Antonacci F, Graves T, Fulton R, *et. al.* Characterization of missing human genome sequences and copy-number polymorphic insertions. Nat Methods. 2010; 5:365-371.

Kim H, Erlich HA, Calloway CD. Analysis of mixtures using next-generation sequencing of mitochondrial DNA hypervariable regions. Croat Med J. 2015; 3:208-217.

Korbel JO, Urban AE, Affourtit JP, Godwin B, Grubert F, Simons JF, *et. al.* Paired-end mapping reveals extensive structural variation in the human genome. Science 2007; 5849: 420-426.

Loman NJ, Misra RV, Dallman TJ, Constantinidou, Gharbia SE, Wain J, Pallen MJ. Performance comparison of benchtop high-throughput sequencing platforms. Nat Biotechnol. 2012; 5:434-439.

Li H, Handsaker B, Wysoker A, Fennel T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. Genome project data processing: the sequence alignment/map format and SAMtools. Bioinformatics. 2009; 16:2078-2079.

Maciejewska A, Jakubowska J, Pawlowski R. Different whole-genome amplification methods as a pre-amplification tool in Y-chromosome loci analysis. Am J Forensic Med Pathol. 2014; 2:140-144.

Mardis ER. Next-generation DNA sequencing methods. Annu Rev Genomics Hum Genet.

2008; 9:387-402.

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, *et. al.* Genome sequencing in microfabricated high-density picolitre reactors. Nature. 2005; 7057:376-380.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research. 2010; 20:1297-303

Medvedev P, Fiume M, Dzamba M, Smith T, Brudno M. Detecting copy number variation with mated short reads. Genome Res. 2010; 11:1613-1622.

Merchant N, Lyons E, Goff S, Vaughn M, Ware D, Micklos, Antin P. The iPlant Collaborative: Cyberinfrastructure for enabling data to discovery for the life sciences. PLoS Biol. 2016; 14:e1002342

Metzker ML. Sequencing technologies – the next generation. Nat Rev Genet. 2010; 11:31-46

Mills RE, Luttig CT, Larkins CE, Beauchamp A, Tsui C, Pittard WS, Devine SE. An initial map of insertion and deletion (INDEL) variation in the human genome. Genome Res. 2006; 19:1182-1190.

Mills RE, Walter K, Stewart C, Handsaker RE, Chen K, Alkan C, *et. al.* Mapping copy number variation by population-scale genome sequencing. Nature. 2011; 7332:59-65.

Nietsch R, Haas J, Lai A, Oehler D, Mester S, Frese KS, *et. al.* The role of quality control in targeted next-generation library preparation. Genomics Proteomics Bioinformatics. 2016; 4:200-206.

Nothnagel M, Herrmann A, Wolf A, Schreiber, Platzer M, Seibert R, Krawczak M, Hampe J. Technology-specific error signatures in the 1000 genomes project data. Hum Genet. 2011; 130:505-516.

Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, Bertoni A, Swerdlow HP, Gu Y. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC Genomics. 2012; 13:341.

Rokas A, Abbot P. Harnessing genomics for evolutionary insights. Trends Ecol Evol. 2009; 4:192-200.

Rothberg JM, Hinz W, Rearick TM, Schultz J, Mileski W, *et. al.* An integrated semiconductor device enabling non-optical genome sequencing. Nature. 2011; 7356:348-352.

Schmieder R, Edwards R. Quality control and preprocessing of metagenomic datasets. Bioinformatics. 2011; 6:863-864.

Schwartz S, Oren R, Ast G. Detection and removal of biases in the analysis of next-generation sequencing reads. PLoS One. 2011; 6:1.

Shin S, Park J. Characterization of sequence specific errors in various next-generation sequencing systems. Mol Biosyst. 2016; 12:914-922.

Tarnecki AM, Burgos FA, Ray CL, Arias CR. Fish intestinal microbiome: Diversity and symbiosis unraveled by metagenomics. J Appl Microbiol. 2017; Epub

Tate CM, Nuñez AN, Goldstein CA, Gomes I, Robertson JM, Kavlick MF, Budowle B. Evaluation of circular DNA substrates for whole genome amplification prior to forensic analysis. Forensic Sci Int Genet. 2012; 2:185-190.

Teague B, Waterman MS, Goldstein S, Potamousis K, Zhou S, Reslewic S, *et. al.* High-resolution human genome structure by single-molecule analysis. Proc Natl Acad Sci U S A 2010; 24:10848-10853.

Templeton JEL, Brotherton PM, Llamas B, Soubrier J, Haak W, Cooper A, Austin JJ. DNA capture and next-generation sequencing can recover whole mitochondrial genomes from highly degraded samples for human identification. Investig Genet. 2013; 4:26.

Verma M, Kushrestha S, Puri A. Genome Sequencing. 2017; 1525:3-33.

Wendt FR, Warshauer DH, Zeng X, Churchill JD, Novroski NM, Song B, King JL, LaRue BL, Budowle B. Massively parallel sequencing of 68 insertion/deletion markers identifies novel microhaplotypes for utility in human identity testing. Forensic Sci Int Genet. 2016; 25:198-209.

Wendt FR, Zeng X, Churchill JD, King JL, Budowle B. Analysis of shirt tandem repeat and single nucleotide polymorphism loci from single-source samples using a custom HaloPlex target enrichment system panel. Am J Forensic Pathol. 2016; 2:99-107.

Yoon S, Xuan, Markarov V, Ye K, Sebat J. Sensitive and accurate detection of copy number variants using read depth of coverage. Genome Res. 2009; 9:1586-1592.

Deliverables

Blankenburg D, Gordon A, Von Kuster G, Coraor N, Taylor J, Nekrutenko A; Galaxy Team. Manipulation of FASTQ data with Galaxy. Bioinformatics. 2010; 14:1783-1785.

Burnside ES, Bintz BJ, Wilson MR. Improved extraction efficiency of human mitochondrial DNA from hair shafts and its implications for sequencing of the entire mtGenome from a single hair fragment. 2012.

Goto K, Nishino I, Hayashi, YK. Rapid and accurate diagnosis of facioscapulohumeral muscular dystrophy. Neuromuscular Disorders. 2006; 4:256-261.

Kavlick MF, Lawrence HS, Merritt RT, Fisher C, Isenberg A, Robertson JM, Budowle, B. Quantification of human mitochondrial DNA using synthesized DNA standards. J Forensic Sci 2011; 56:1457-63.

Kloss-Brandstätter A, Pacher D, Schönherr S, Weissensteiner H, Binna R, Specht G,

Kronenberg F. HaploGrep: a fast and reliable algorithm for automatic classification of mitochondrial DNA haplogroups. Hum Mutat. 2011; 1:25-32.

Laurie MT, Bertout JA, Taylor SD, Burton JN, Shendure JA, Bielas, JH. Simultaneous digital quantification and fluorescence-based size characterization of massively-parallel sequencing libraries. Biotechniques. 2013; 1:29-38.

Robin ED, Wong R. Mitochondrial DNA molecules and virtual number of mitochondria per cell in mammalian cells. Journal of cellular physiology. 1998; 3:507-513.

Swango KL, Timken MD, Chong MD, Buoncristiani MR. A quantitative PCR assay for the assessment of DNA degradation in forensic samples. Forensic Sci Int. 2006; 158:14-26.

Weissensteiner H, Pacher D, Kloss-Brandstätter A, Forer L, Specht G, Bandelt HJ, Kronenberg F, Salas A, Schönherr S. HaploGrep2: mitochondrial haplogroup classification in the era of high-throughput sequencing. Nucleic Acids Res. 2016; 44:W58-W63.

Wilson MR, Stoneking M, Holland MM, DiZinno JA, Budowle B. Guidelines for the use of mitochondrial DNA sequencing in forensic science. Crime Lab Digest. 1993; 4:68-77.

Ye J, Coulouris G, Zaretskaya I, Cutcutache I, Rozen S, Madden, TL. Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. BMC Bioinformatics. 2012; 1:134.