Document Title:	Resolution of DNA Mixtures and Analysis of Degraded DNA Using the 454 DNA Sequencing Technology
Author(s):	Cassandra D. Calloway, Ph.D., Hanna Kim, Henry Erlich
Document No.:	249100
Date Received:	August 2015
Award Number:	2010-DN-BX-K141

Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

Resolution of DNA Mixtures and Analysis of Degraded DNA

Using the 454 DNA Sequencing Technology

Grant #:2010-DN-BX-K141

Principal Investigator

Cassandra D. Calloway, PhD

Authors

Cassandra D. Calloway, PhD, Hanna Kim, and Henry Erlich

I. Abstract

The massively parallel and clonal nature of next generation sequencing (NGS) technologies has the potential to revolutionize the forensics DNA field. NGS is capable of analyzing sequence polymorphisms (e.g. mtDNA, SNPs) as well as length polymorphisms (e.g. STRs) on the same platform. Moreover, sequencing STR markers will detect any sequence polymorphisms in the STR loci, thus increasing the discrimination potential of these widely used markers. The high throughput nature of NGS makes it technically feasible and cost-effective to sequence the entire mtDNA genome rather than just the HIVI/HVII regions, thereby greatly increasing the discrimination potential of mtDNA analysis. In addition, the sensitivity of NGS systems facilitates the analysis of forensics specimens with limiting or degraded DNA. The clonal nature of NGS provides a powerful and quantitative means of deconvoluting mixtures as the contribution of each component in a mixture is quantified by simply counting the number of sequence reads. We proposed to develop methods for analyzing mixed, limited and degraded samples using the 454 NGS technology for deep sequencing mtDNA and STR markers.

A duplex PCR assay targeting the mtDNA HVI/HVII regions was successfully developed using eight sets of 454 MID tagged fusion primers in a combinatorial approach for deep sequencing 64 samples in parallel. This assay was shown to be highly sensitive for sequencing limited DNA amounts (~100 mtDNA copies) and detecting mixtures with low level variants (~1%) as well as "complex" mixtures (≥3 contributors).

We have also successfully designed and showed proof-of-concept for a solution phase sequence capture and NGS assay for targeted enrichment and deep sequencing of the entire mitochondrial genome for increased discrimination power. Using this NimbleGen SeqCap NGS assay, 100%

sequence coverage of the mitochondrial genome with an ~80% on target rate was achieved. The sequence read coverage achieved with probe capture was more evenly distributed than the coverage reported for PCR enrichment of the mtDNA genome. The input DNA amount was successfully lowered to the forensically relevant level of 100 pg and a 10% and 5% mixture was resolved. Moreover, a DNA fragmentation method using mechanical shearing (Covaris) was optimized and shown to be DNA quantity and quality independent, essential for preparation of highly degraded or limited samples often encountered in forensic cases. Proof-of-concept for using the optimized fragmentation method for analysis of degraded samples was established with artificially degraded samples.

Also, NGS assays targeting the CODIS STR loci were developed using 454 mini STR fusion primers in a multiplex PCR as well as an approach using a universal 454 primer set for amplification of a mini STR multiplex and analyzed using modified commercial software. Results show successful amplification and sequencing of 50 pg of DNA. Additionally, proof-of-concept for sequencing mtDNA and STR markers in a single 454 NGS run was shown.

We also collaboratively modified softwares for mtDNA and STR next generation sequence alignment and analyses. The commercially available NextGENe software was customized by SoftGenetics for the analysis of data generated for the HVI/HVII and the STR assays which included modifications to the barcode sorting tool, STR sequence alignment algorithm, and addition of forensic specific reports for mixture and STR lengths analysis. We also developed a flexible front-end for running and analyzing the results from Mapping Iterative Assembler (MIA) for analysis of mtDNA NGS data and demonstrated the ability of MIA to detect a minor component in a 10% mixture in collaboration with Dr. Richard Green. The development of customized software for forensic applications is critical to the successful implementation of NGS in a forensic laboratory setting.

Overall, we successfully demonstrated the use of NGS for analysis of challenging forensic samples.

II. Table of Contents

III. Executive Summary

Degraded and mixed DNA samples are often encountered in forensic cases and pose interpretation challenges. Alternative markers such as nuclear bi-allelic SNPs and mtDNA are often used to analyze limited and/or degraded DNA. However, there are some limitations to these approaches. Nuclear bi-allelic SNP markers do not allow for efficient detection of mixtures and mtDNA lacks discrimination power as only the HVI/HVII regions are sequenced. While STR and mtDNA markers allow for detection of mixtures, they do not allow for separation of components in a mixture. Mitochondrial DNA markers are ideal targets for detecting mixtures, since with few exceptions a single sequence is the expected result due to its haploid nature. However, unlike STRs, peak areas or heights in sequence electropherograms are not necessarily indicative of the amount of DNA contributed to a mixture. As a result, peak height ratios for two bases cannot be used to determine the relative proportions of components of a mixture. For this reason, Sanger sequencing does not allow for determining the mtDNA haplotype of mixed samples. Furthermore, minor components present at less than 10-20% in a DNA mixture are not detectable by Sanger sequencing.

The 454 DNA sequencing technology is a scalable, highly parallel pyrosequencing system that can be used for *de novo* sequencing of small whole genomes or *direct* sequencing of DNA products generated by PCR. The technology uses emulsion PCR (emPCR) to amplify a single DNA sequence to 10 million identical copies. The "clonal sequencing" aspect of the technology enables separation of individual components of a mixture as well as analysis of highly degraded DNA. T he clonal sequencing approach used with the 454 technology and other next-generation sequencing technologies provides a digital readout of the number of reads or individual sequences allowing for a quantitative determination of the components in a mixture. The 454

DNA sequencing technology has been successfully used to analyze mixtures in clinical samples and for analysis of highly degraded DNA from "ancient DNA" samples, including an ~40,000 year old bone fragment. These studies demonstrate the potential utility of using the 454 sequencing technology for forensic applications.

We proposed to use the 454 DNA sequencing technology to analyze mixed, limited, and/or degraded DNA samples using mtDNA and miniplex STR markers. Specifically, we proposed to develop a 'front-end' PCR based approach which uses multiplex identification (MID) tags to amplify mtDNA and STR targets prior to emPCR. This 'front-end' approach is especially useful for sequencing specific target regions in the genome and allows for massive parallel sequencing of pooled samples. We initially proposed to develop a mtDNA 454 multiplex assay for targeting and sequencing ~20-30% of the mitochondrial genome. However, we modified this specific aim and developed a 454 duplex assay targeting the HVI/HVII regions and a solution phase probe sequence capture assay for enrichement and sequencing of the entire mitochondrial genome. This aim was modified as new probe capture enrichment methods became available and had significant advantages to developing a multiplex targeting only 20-30% of the mitochondrial genome. Using the probe capture method for enrichment, the entire mitochondrial genome could be sequenced greatly improving the discrimination power and using probe capture potentially will allow for analysis of highly degraded DNA as intact priming sites are not necessary.

Our primary goals were to 1) to develop and optimize a "front-end" system targeting nuclear STR and mtDNA markers for use with the 454 sequencing technology and 2) validate and apply the system to forensically relevant samples (i.e. mixed and limited DNA). Specifically, over the granting period the following objects were met. 1) We developed, optimized and validated a 454 HVI/HVII mtDNA PCR assay using fusion primers for sequencing up to 64 samples in parallel and used this system to analyze mixtures, heteroplasmic samples and limited DNA samples. 2) We designed and tested a solution phase probe capture assay for enrichment and next generation sequencing of the entire mitochondrial genome and used this system to show proof of concept for sequencing the mitochondrial genome of limited, mixed and degraded samples. 3) We developed a mini-STR multiplex 454 PCR assay using fusion primers and a two step multiplex 454 PCR using universal primers to greatly reduce the number of required primers and showed proof of concept for sequencing limited DNA samples. 4) In collaboration with SoftGenetics we modified commercially available next generation sequencing software for improved analysis of mtDNA and STR markers and with Dr. Richard Green we modified existing MIA software for analysis of mtDNA mixtures and developed a windows interface for ease of use. The research design and findings for each of these primary objectives are described in more detail below.

i. 454 NGS HVI/HVII mtDNA Assay

We successfully developed a "front-end" duplex PCR assay targeting the non-coding HVI and HVII regions of the mitochondrial genome for use with the 454 sequencing technology. A total of eight optimized sets of duplex multiplex identifier (MID) tagged fusion primers which can be used to generate 64 different combinations of forward and reverse tagged primer sets targeting the HVI/HVII regions of the mitochondrial genome were developed for the assay. This combinatory approach greatly reduces the number of primers required to achieve the maximum number of samples that can be pooled in one sequencing run. Unique 10 bp MID tags are used as sample identifiers in order to pool and sequence multiple samples in a single 454 sequencing run. Fusion primers used in the initial amplification consist of the following parts starting from the 5' end: 1) an adapter sequence that serves as a universal primer for emPCR and pyrosequencing, 2) a 4 base library key for signal normalization, 3) a unique 10 base sample-specific internal sequence tag (MID tag), and 4) a target specific forward or reverse PCR primer sequence. PCR parameters and conditions were optimized as well as an AMPure purification step to remove short fragments such as primer dimer for improved 454 sequencing results.

Mixtures

A mixture study was conducted to determine the sensitivity of the mtDNA HVI/HVII duplex 454 sequencing assay for detecting minor components in a mixture. Two DNA samples were mixed together at various ratios for a simple mixture study while three, four, and five DNA samples were mixed together at different ratios for the complex mixture study. Each mixture ratio was based on mtDNA copy number determined by qPCR. This 454 HVI/HVII assay was shown to be highly sensitive for detecting mixtures with low level variants (~1%) as well as complex mixtures (≥3 contributors). The minor base was reliably detected at each of the mixed base positions and the observed frequencies were similar to the expected frequencies using 454 sequencing with ~600-1000 reads per amplicon but not by Sanger sequencing. Further, we have characterized and observed the jumping PCR effect that arises in amplification of mixed samples that are often encountered in the forensic field and concluded that the frequency of the phenomenon can be decreased by lowering the amplification cycle number. Heteroplasmic samples of various tissues from monozygotic twins were studied using the developed assay and the 454 sequencing results were compared to the Sanger sequencing results. Heteroplasmy at position 16093 was detected in the buccal samples by both methods but was only detectable in blood using the more sensitive 454 NGS method. We show that not only is NGS more sensitive for detecting minor components in a mixture, but it is also more quantitative because it provides a digital read-out counting the number of sequence reads corresponding to the individual components.

Sensitivity

The HVI/HVII 454 sequencing assay was also shown to be highly sensitive. A sensitivity study was conducted by amplifying for 34 cycles various dilutions of a DNA sample which was quantified to determine the mtDNA copy number. Amplification and sequencing was successful for samples with DNA amounts as low as ~1 pg or 100-500 mtDNA copies. This study was successful in showing that NGS results can be obtained from significantly lower amounts of DNA than previously reported or recommended by the manufacturer (pg amounts compared to ng or µg amounts). This study also demonstrates feasibility of using NGS for analysis of forensically relevant samples which are often limited in amount of DNA.

Population Study

Additionally, a small population database was generated using the optimized 454 HVI/HVII NGS assay and compared to Sanger sequencing results which showed concordance between the two sequencing methods for all base substitutions. However, as expected pyrosequencing errors were detected in the homopolymer C-stretch regions, but all base substitution mutations were identified. Sequence alignment issues were minimized in these regions using SoftGenetics NextGENe software which has improved alignment software and allows for filtering sequencing errors in homopolymer regions. We also conducted concordance and reproducibility studies in collaboration with California Department of Justice (CA DOJ) who began validating the HVI/HVII mtDNA assay on the 454 GS Jr sequencing instrument for sequencing backlogged reference samples from their missing persons section. These studies showed concordance between the 454 sequencing data generated from both labs as well as compared to Sanger sequencing.

ii. Whole Mitochondrial Genome Probe Capture

We explored a probe capture method for enrichment and sequencing of the entire mitochondrial genome as an alternative to amplifying select regions of the mtDNA genome described previously as part of our aim to developing a mtDNA 'front end' assay. Probe capture for enriching for mtDNA would allow for full mtDNA sequencing for increased discrimination power as well as the potential for capture of degraded DNA because intact priming sites are not needed. We successfully designed and showed proof-of-principle for whole mitochondrial genome sequencing using a liquid phase probe capture and a 454 NGS assay. After considering several capture technologies we chose the Nimblegen SeqCap EZ platform due to their extensive tiling design and ability to efficiently synthesize hundreds of thousands of probes. A majority base consensus sequence of the mitochondrial genome was used as our target sequence due to the high density and distribution of polymorphisms within the mitochondrial genome. The frequencies used for our consensus sequence were based on published data from the mtDB-Human Mitochondrial Genome Database website.[1] Probes designed to target minor base alleles at polymorphic sites was also considered, but were not included since the SeqCap hybridization conditions tolerate up to five base mismatches. To increase the specificity of our probes, we considered the circular nature of mtDNA, the high density and distribution of polymorphisms, and nuclear pseudogenes in our probe design strategy. Also, we addressed the circular nature of the mitochondrial genome by adding the first 100 bases of mtDNA sequence to the end of the target sequence. During the design process, we took into consideration mtDNA nuclear pseudogenes by restricting the homology of our probes to the nuclear genome to ten sites or less. The probe length was allowed to vary between 50 bp and 100 bp based on optimal melting temperature per Nimblegen's specifications. The design used in our preliminary study directly targeted 99.9% of the mitochondrial genome with unique probes with only two small gaps at positions 2,506-2,513 (7bp) and 2,962-2,972 (10bp).

Pre and post capture PCR parameters for the method was optimized for the forensically relevant limited samples. The amplification cycle number was increased from the manufacturer's recommended 11 cycles to 28 cycles to accommodate limited sample input. Additionally, primer concentration optimized to be 1 uM compared to the manufacturer's recommended concentration of 4 uM. The primer concentration was decreased in order to minimize the amount of primer dimer forming during the amplification. As our studies involved limited DNA amounts (100 pg to1 ng in comparison to the manufacturer's recommended 1 ug), we observed high amount of excess primer dimers forming during library preparation. We also investigated several methods for DNA fragmentation including nebulization, enzymatic digestion using Fragmentase, and mechanical shearing using Covaris. Nebulization was not an ideal fragmentation method for forensic DNA samples due to potential risk of contamination with aerosolizing DNA. Two fragmentation systems using the enzyme Fragmentase from New England Biolabs and the mechanical shearing technology from Covaris was developed and optimized. However, it was soon determined that the enzymatic digestion is dependent on the DNA quantity and quality, thus not practical for forensic applications while mechanical shearing, using Covaris uses Adaptive Focused Acoustic (AFA) technology, was sample quality and quantity independent. A series of experiments were conducted using a Covaris ultrasonicator to show proof-of-principle that this mechanical shearing technique is independent of starting DNA fragment size. DNA was naturally or artificially degraded to different levels (20kb, 5kb, 3kb, 1kb, 700bp, and 500bp). The degraded samples were then all sheared using the manufacturer's recommended protocol targeting 500 bp. Results showed that regardless of the initial level of degradation, the fragment size for each sample was tightened to 500 bp without losing the smaller fragments. These preliminary results show proof-of-principle that a single set of parameters can be used with the Covaris AFA technology for shearing DNA with various levels of degradation without leading to loss of smaller fragments.

We tested the sensitivity, specificity, mixture detection, and reduction of hybridization time of the whole mitochondrial genome capture assay. The systems resulted in 100% capture of the mitochondrial genome with an ~80% on target rate and an average sequence coverage of ~750 reads per base. The distribution of reads was similar across the mitochondrial genome. To improve the efficiency of the probe capture method, the hybridization time was reduced from the recommended three days to one day. No significant differences in probe capture efficiency were observed between the three day or one day hybridization times. Based on the preliminary results, we expect to be able to reduce the hybridization time further.

We have tested sensitivity by reducing the starting amount of DNA from the manufacturer's recommended 1 ug of DNA to 100 pg (tested 100 ng, 1 ng, 100 pg), which shows the proof-of-principle for forensic applications. Based on preliminary results, we expect to be able to reduce the starting DNA amount further. For proof-of-principle, a 10%, 5%, and 1% DNA mixture were analyzed. Preliminary data show detection of the 10% and 5% minor component in the mixture which is below the limits of Sanger sequencing (10-20% dependent on base position and background noise). All SNPs previously detected by Sanger sequencing were detected by 454 sequencing for the 10% and 5% mixture. Both commercially available SoftGenetics NextGENe software and a customized prototype windows based version of a MIA software were used for data analysis and an average of 11.4% and 11.9% minor base frequency was observed at each of the mixed base positions, respectively for the 10% mixture sample. However, the read depth (average 100 reads/base) was not sufficient for detection of the minor component at all mixed based sites in the 1% mixture. A 1% minor component in a mixture is expected to be detected by increasing the read depth to greater than 1000 fold sequence coverage. Further experiments are needed to determine the limitations of the assay, including sensitivity and mixture detection, as well as analysis of degraded DNA samples.

iii. STR

Although the 454 NGS HVI/HVII duplex PCR assay has shown great sensitivity for limited DNA samples as well as for mixture detection, the maternal inheritance pattern of mtDNA and sequence information from only the HV regions provides limited discrimination power, particularly within the Caucasian population. This aspect of the HVI/HVII assay highlights the great need for an assay with increased discrimination power while continuing to utilize the mixture detection power of mtDNA analysis. Thus we have developed 454 NGS assays targeting the CODIS STR loci using 454 mini STR fusion primers in a multiplex PCR as well as an approach using a universal 454 primer set for amplification of a mini-STR multiplex. A STR assay with M13 universal primer design using 454 Next Generation Sequencing technology was developed as an alternative approach to the fusion primer approach to minimize the number of primer sets required for pooling multiple samples per sequencing run. In this unique design the samples are first amplified by inner primers consisting of STR locus target specific sequence and a universal M13 linker sequence. The inner products are then amplified with outer primers consisting of complementary M13 linker sequence, a 10 bp MID tag, and a 25 bp 454 specific sequence. We have showed proof-of-concept for generating a full STR profile with 50 pg input DNA through a small sensitivity study, but the lower sensitivity limit is not established yet. Different purifications methods for small fragment removal (primer dimer) were examined for both STR assays. Although the current method is optimized with the AMPure purification, the potential product loss with AMPure XP small fragment removal need to be addressed and explored further. Further analysis is needed.

iv. Software

We have worked with 2 external collaborators to modify next generation sequencing software for mtDNA and STR next generation sequence alignment and analyses. A commercially available software NextGENe was customized for the analysis of data generated for HVI/HVII assay and the STR assays. Two analysis options, mitochondrial amplicon analysis and STR analysis, were added to the software, and modifications to the sorting tool were made to accommodate the combinatorial use of MID barcode tagging. The STR sequence alignment algorithm was optimized for alignment of STR repeats, assigning lower gap penalty for IN/DEL sizes corresponding to the repeat size (4 bp). The modified algorithm results in improvement for alignment of the STR repeats. However, due to the nature of the repeats in the STR markers, further optimization and modification of the software is needed for the correct and optimal alignment of the STR sequence. A high pyrosequencing error rate was observed in the homopolymer C stretch region in HVI and HVII amplicons and resulted in sequence mis-alignment in these regions using the 454 AVA software. However, base substitutions were correctly identified using the NextGENe software in these homopolymer regions and the sequence alignment was somewhat improved when filters for removing sequence errors were applied using this software. Additional modifications addressing these issues are still needed to effectively analyze the data and are on-going.

We have also developed a flexible front-end for running and analyzing the results from Mapping Iterative Assembler (MIA) for mtDNA mixture analysis, which operates by aligning each of the input 454 (or other platform) reads against a defined reference mtDNA assembly in collaboration with Dr. Richard Green at UC Santa Cruz. MIA was able to detect a minor component (10%) of an alternative haplotype mixed into a background of mtDNA from another haplotype. Recently a modification to the algorithm was made and is currently being tested for analysis of mixtures for the HVI/HVII Amplicon assay. Incorporating an algorithm to identify sequence haplotypes of detected sequences is currently being explored for mixture analysis of the whole mitochondrial genome. A feature to take into account the circular nature of the entire mitochondrial genome is in the process of being added. Further modifications to the software are on-going.

Implications For Criminal Justice Policy and Practice

The massively parallel and clonal nature of next generation sequencing (NGS) technologies has the potential to revolutionize the forensics DNA field. The implementation of NGS in forensics labs will have a significant impact for many different reasons. NGS is capable of analyzing sequence polymorphisms (e.g. mtDNA, SNPs) as well as length polymorphisms (e.g. STRs) on the same platform, and, in principle, in the same run. Moreover, sequencing STR markers will detect any sequence polymorphisms in the STR loci, thus increasing the discrimination potential of these widely used markers. The high throughput nature of NGS makes it technically feasible and cost-effective to sequence the entire mtDNA genome rather than just the HVI and HVII regions, thereby greatly increasing the discrimination potential of mtDNA analysis. In addition, the sensitivity of NGS systems facilitates the analysis of forensics specimens with limiting or degraded DNA.

The clonal nature of NGS provides a powerful and quantitative means of deconvoluting mixtures, a particularly challenging category of forensics specimens. Different contributors to a mixture can be detected by analyzing the different clonal sequence reads identified in the sequence analysis and their contribution quantified by simply counting the number of sequence reads. This digital analysis is much more precise than estimating peak height or area for STR markers or analyzing Sanger electropherograms for mtDNA sequences. MtDNA is a particularly useful marker for deconvoluting mixtures because, potential heteroplasmy aside, each distinct mtDNA sequence corresponds to (at least) one contributor. Statistical analyses can capture and incorporate the probability that one mtDNA sequence could correspond to more than one contributor. Thus, NGS analysis of mtDNA is the most robust way of estimating the number of contributors to a mixture.

One limitation of this study is that it was performed on the 454 platform (GS Junior instrument), a technology that will be discontinued in 2016. However, all of the approaches using the 454 platform developed and reported here can be adapted to Illumina or Ion Torrent with minor modifications. Furthermore, the cost of sequencing, the ease of use, and the length of sequence reads, have all been improving on all platforms, making it likely that NGS on relatively inexpensive desktop sequencers will be broadly implemented in forensics labs over the next 5 years. Most of the NGS work we have done thus far has been based on the 454 GS junior instrument system but, we anticipate that, over the next 2 years, much of our work will be carried out on a MiSeq instrument. with some minor protocol modifications.

In general, for forensics as well as other genetic analyses, the two approaches to target enrichment for library preparation are PCR or probe capture. In our experience, PCR has proved effective for NGS analysis of STRs and the HVI and HVII regions of mtDNA. We have found, however, that probe capture is a robust, sensitive, and efficient way to enrich for the whole mtDNA genome. Given the high concentration of mtDNA, the probe capture is efficient with less starting material and allow for shorter hybridization times than for nuclear markers. The optimized probe capture methods described have high promise to overcome many of the major challenges routinely encountered with analysis of limited and degraded DNA by greatly increasing the discrimination power of current mtDNA assays and through its capability of analysis of degraded samples since it is not dependent on specific priming sites.

The customized forensic NGS software developed for this project will be made publicly available and has broad applicability for NGS applications. Advances in oligo synthesis have markedly reduced the cost of probes, making probe capture an affordable method for routine forensic applications. The use of multiplex tags for sample pooling greatly reduces per sample costs and implementation feasible. NGS technologies are capable of analyzing both mitochondrial and nuclear DNA targets as well as SNP and STR markers simultaneously. Overall, the proposed capture and NGS assays offer increased resolution and discrimination power for mtDNA as well as improved success for the analysis of degraded and limited DNA analysis.

Next generation sequencing (NGS) promises to have a major impact on the practice of forensics labs over the next 5-10 years and, for those labs experienced in incorporating new technologies, within the next 2 years. Over the next few years, the throughput, cost, and read length obtained with existing platforms should improve and the availability of relatively inexpensive desktop sequencers will make access to NGS affordable and cost-efficient for many forensics labs.

IV. Main Body

A. Introduction

1. *Statement of the Problem*

Degraded and mixed DNA samples are often encountered in forensic cases and pose interpretation challenges[2]. Alternative markers (non STR) such as nuclear bi-allelic SNPs and mtDNA are often used to analyze limited and/or degraded DNA[3]. However, there are some limitations to these approaches. Nuclear bi-allelic SNP markers do not allow for efficient detection and analysis of mixtures and mtDNA lacks discrimination power[4]. While STR and mtDNA markers allow for detection of mixtures, they do not allow for separation of components in a mixture. Other approaches allow for physical separation of various cellular components or DNA molecules (e.g. laser microdissection and DHPLC)[5-8]. However, these approaches may not completely separate components or be compatible with standard lab work flows. In addition, current forensic assays do not allow for simultaneous analysis of nuclear and mtDNA markers.

The 454 DNA sequencing technology is a scalable, highly parallel pyrosequencing system that can be used for *de novo* sequencing of small whole genomes or *direct* sequencing of amplified DNA (PCR products)[3]. This "next generation" or massively parallel" sequencing technology uses emulsion PCR (emPCR) to amplify a single DNA fragment (sheared genomic DNA or amplified DNA) to 10 million identical copies (www.454.com). The "clonal sequencing" aspect of the technology enables separation and characterization of individual components of a mixture as well as analysis of highly degraded and limited quantities of DNA. The 454 DNA sequencing technology has been successfully used to analyze mixtures for clinical applications[9] and highly degraded DNA in "ancient DNA" cases[3]. Recently next generation sequencing technology has been used to sequence the complete mitochondrial genome of an ~40,000 year old bone fragment (DNA extracted from 30mg of bone powder)[3]. The 454 sequencing technology has also been used to sequence single STR marker[10]. These studies demonstrate the potential utility of using the 454 sequencing technology or other NGS platforms for forensic applications.

2. Review of Relevant Literature

i. Mixtures

Interpretation of mixtures from known and unknown sources, such as in sexual assault and "contact DNA" cases, can be challenging for forensic DNA analysts when targeting nuclear DNA[11,12]. A four year retrospective study showed that 6.7% of casework samples typed in one laboratory had a mixed STR profile[13]. Mixtures are also encountered in mtDNA analysis (e.g. heteroplasmy and multiple contributors)[3]. In a five year retrospective study of mtDNA analysis of 691 casework hair samples, a mixture of mtDNA sequences attributed to a secondary source was observed in 8.7% of the hairs and sequence heteroplasmy was observed in 11.7% of the cases[14]. In such cases, the ability to resolve mixtures can benefit the overall DNA analysis.

Mixed sample stains (blood, semen, saliva) are present in many forensic investigations and may arise when two or more individuals contribute to the sample at a crime scene or an accidental transfer of DNA (contamination) to the sample occurs. There are also a number of situations in routine forensic casework where mixed DNA profiles may be *expected*, including finger nail clippings and swabs taken from the skin or body. For example, in sexual assault cases where a mixture of suspect and victim DNA is present, it would be valuable to separate the various components of a DNA mixture. Likewise, mixtures are commonly encountered in contact DNA cases, resulting from a transfer of cellular material onto a surface handled or touched by one or more persons. For example, distinguishing DNA mixtures from triggers and slide areas from swabs collected from firearms can be a useful forensic and investigative tool.

Mixed DNA profiles or sequences can also result from somatic or germ-line mutations. Chromosomal abnormalities can occur and will appear as an extra allele peak (three peaks) at a single STR locus[2]. However, in a mixed DNA sample with two or more contributors, extra allele peaks (3, 4… peaks) will likely be observed at more than one STR locus[15]. A mixture of mtDNA sequences resulting from mutation (heteroplasmy) may also be observed during analysis of mtDNA in forensic casework[3]. Point mutations, insertions, and deletions can occur and accumulate in the mitochondrial genome resulting in heteroplasmy[16-18]. Heteroplasmy can differ between tissues within an individual or between maternally related individuals[16,19-22]. Both a mixture of DNA sequences resulting from mutation (heteroplasmy) or a secondary source appear as multiple sequences in a sequence electropherogram[23]. The current method for directly sequencing mtDNA does not easily allow for resolution of components of a mixture or the determination of mtDNA haplotype of mixed samples.

Standard approaches for DNA analysis targeting multi-allelic STR markers where the alleles are distinguished by mobility by gel electrophoresis allow for *detection* of mixtures[24]. A DNA profile of a mixture will likely show three or more prominent peaks at one or more loci[15]. A severe peak height imbalance may also be observed between two alleles at the same locus, suggestive of a mixture[25,26]. Peak areas or heights in an electropherogram have been shown to be directly related to the amount of DNA template present[27]. Therefore, STR locus peak areas or heights can be used to estimate the relative proportions of DNA from the contributors of a mixture, although preferential amplification and stutter bands can complicate the quantification. However, minor components present at less than 5% in a DNA mixture are not detectable with the standard platform used for STRs[28]. Although mixtures are detectable using standard STR approaches and software, mixture interpretation is still difficult because the individual genotypes cannot be determined. STRs allow for establishing that a mixture is present but not for determining the "paired" or linked alleles that constitute a genotype. Therefore, all possible genotype combinations must be considered in the mixture interpretation. Furthermore, in low copy number cases, allelic dropout is commonly observed resulting in greater difficulty interpreting mixtures[25,29].

Several statistical approaches are used for interpretation of complex forensic STR mixtures[29-32]. Deconvolution tools are now available to aid in the interpretation of STR mixtures and include FSS-i3® v4.1.3 i-STReam[33], Least-Square Deconvolution (LSD)[32] and USACIL's DNA_DataAnalysis v2.1.3. The FSS mixture interpretation software is based on the PENDULUM technology, a guideline based approach for mixture interpretation which takes into account peak height imbalance among other heuristic rules, and allows for deconvolution of most two person mixtures[33]. Interpreting mixtures is even more challenging in cases where DNA from more than two individuals is present, and in some cases, three and four person mixtures can be misinterpreted as two or three person mixtures[34]. Sequencing STR loci using the 454 system gives a more quantitative estimate of the relative proportions of STR sequences from multiple contributors. The frequency estimates for STR alleles obtained from the 454 system (counting sequence reads) can be used in place of peak height estimations for statistical mixture analysis and interpretation using existing software.

Mixtures can also be detected by direct sequencing of mtDNA or using multi-allelic SNP probe-based systems[35-38]. Mitochondrial DNA markers are ideal targets for detecting mixtures, since with few exceptions a single sequence is the expected result due to its haploid nature. For example, a mixture of two or more mtDNA sequences will likely have two overlapping peaks at one or more sites in a mtDNA sequence electropherogram (Sanger). However, unlike STRs, peak areas or heights in sequence electropherograms are not necessarily indicative of the amount of DNA contributed to a mixture[3]. As a result, peak height ratios for two bases cannot be used to determine the relative proportions of components of a mixture. For this reason, Sanger sequencing does not allow for determining the mtDNA haplotype of mixed samples. Furthermore, minor components present at less than 10-20% in a DNA mixture are not detectable by Sanger sequencing[37,39]. In addition, a mixture of sequences of varying lengths (length heteroplasmy), although detectable, will result in poor quality sequence downstream of the IN/DEL making the sequence unreadable, such is often the case in the poly C-stretches in the HVI/HVII regions[40-43]. However, the clonal sequencing approach used with the 454 technology will allow for analysis of individual sequences and therefore, length heteroplasmy will not result in unreadable sequence downstream.

While these approaches for analysis of STR and mtDNA markers allow for detection of mixtures, they do not allow for *separation* of components in a mixture. Other approaches allow for physical separation of various cellular components of a mixture. Differential extraction procedures are routinely used for separation of sperm and epithelial cells in sexual assault cases[44]. However, this extraction method does not always allow for complete separation of the female and male fractions. More recently, laser microdissection systems have been used to isolate sperm and non-sperm cellular mixtures[6-8]. However, this method does not always allow for use with standard STR approaches because of the limited amount of starting material. Another method of resolving mixtures focuses on separating DNA molecules rather than cells and uses Denaturing High-Performance Liquid Chromatography (DHPLC)[5,45-47]. This method allows for physical separation and fraction collection of components in a mixture as demonstrated for mtDNA analysis, but does not always allow for complete physical separation[46]. Therefore, in order to reliably resolve mtDNA mixtures using DHPLC, relative changes in electrophoretic peak heights must also be used to determine linkage phase[3]. In addition to these complexities, DHPLC may not be easily integrated into the standard lab work flow or be applied to nuclear STR markers. The 454 system uses a "clonal" sequencing approach in which each component is sequenced individually; therefore it would not be necessary to first separate components of a mixture.

ii. Degraded DNA

Samples with degraded or limited DNA are often encountered in forensic cases[48,49].

Several approaches are currently taken to overcome this problem. These include miniplex STR, nuclear bi-allelic SNP analyses, and mtDNA[50-57]. Miniplex STR and nuclear bi-allelic SNP systems have increased success of amplifying degraded DNA because the targets are shorter compared to standard STR markers. However, the number of mini STR loci that can be multiplexed in a single PCR is limited because of PCR product size constraints and the limited number of dyes available for the capillary electrophoresis platform[52]; the number of loci per multiplex would not be limited when using the 454 sequencing technology. Although nuclear bi-allelic SNP systems are useful for the analysis of degraded DNA, mixtures are difficult to detect with these systems since there are typically only two alleles for each locus. For this reason, it is often difficult to distinguish heterozygosity from a mixture when using nuclear bi-allelic SNP systems[58]. Therefore, multi-allelic systems targeting length polymorphism (STR) or haploid markers (mtDNA or Y chromosome) are more appropriate for detecting mixtures using standard approaches currently used by forensic laboratories.

Mitochondrial DNA is often used to analyze degraded samples because of the high copy number per cell. The HV I /HVII regions are most often targeted and in some cases mini primer sets are used to increase sensitivity. However, mini primer sets are not always successful particularly when the DNA is highly degraded. In addition, the discrimination power of analysis of the mtDNA HVI/HVII regions is somewhat limited[4,59]. To improve discrimination power select coding regions are also targeted using various PCR based methods[4,59-64]. These methods are limited in the number of mtDNA regions that can be analyzed in a single run and cannot analyze mtDNA and STR markers simultaneously. Next generation sequencing of mtDNA coupled with mini STRs for increased discrimination would be an ideal system for analysis of mixed and/or degraded forensic samples.

iii. Next Generation Sequencing Technologies and Applications

Next Generation Sequencing (NGS) is a massively parallel, clonal sequencing method which produces vast amounts of sequencing data in a more cost effective and timely manner than has ever been possible.[65,66] There are several different NGS platforms which utilize different sequencing chemistries resulting in unique characteristics and limitations associated with each platform. The longer read platforms include Roche 454 GS and PacBio RS system which can reliably sequence products of ~500 bp and over 1 kb respectively while the Illumina and Ion Torrent systems have average read lengths of ~200 bp.[67] The differences in producible read lengths result directly from the chemistries employed by each technology. The Roche 454 system is based on pyrosequencing, which measures the amount of light produced by a luciferase reaction created from the pyrophosphate released during nucleotide incorporation. This pyrosequencing chemistry allows for longer read lengths, however, it produces homopolymer-associated in/del errors. Illumina chemistry utilizes reversible dye-terminating fluorescently labeled nucleotides which are incorporated one at a time and detected by a camera. The Roche 454 system produces longer but a fewer number of reads while Illumina produces shorter but a much greater number of reads per run. Longer reads are ideal for determining linkage phase or haplotyping and reading through long repeat regions while systems with a greater number of reads would be more advantageous for larger genomes or a higher number of targets for higher throughput.

Next Generation Sequencing is useful for many different applications including de novo genome assembly, whole exome sequencing, and rare SNP detection.[68] Genomes of many species which were too large or complex to make conventional sequencing possible are now able to be sequenced and assembled with this technology.[69,70] NGS for SNP detection is helping to shape current cancer research methods by individual tumor characterization which allows individualized treatment options to be explored.[71] Also, NGS has been used to sequence a single-cell for cancer applications.[72] This sequencing technology has also greatly impacted microbial research.[73] The dynamics of whole microbial populations can be studied through rRNA sequencing with NGS.[74,75] The human microbiome was recently characterized and was shown to be relatively stable over time and distinct within individuals.[76]

Next Generation Sequencing has also proven useful for the ancient DNA community. Ancient specimens typically yield limited amounts of DNA that consist of short fragments with substantial chemical damage.[3] What is more, ancient DNA preparations generally consist mostly of DNA from microbes that have colonized the bone or other sample in the thousands of years the bone has been in the ground.[77] Nevertheless, through direct high-throughput sequencing, it has been possible to successfully complete whole ancient genomes or mitochondrial genomes.[3] A complete Neandertal mitochondrial genome was assembled using 454.[3] More recently, deep sequencing has revealed the entire nuclear genome of a Denisovan girl,[78] an hominoid who lived during the same time period as the Neanderthals. Probe capture enrichment methods coupled with NGS technologies have been applied to highly degraded ancient DNA samples with great success despite the very short fragment size.[79] Many of the lessons learned in analysis of large datasets of ancient DNA are directly applicable to forensic analysis. For example, the biases incurred by mapping short, damaged DNA reads against a known reference[80,81] are important to understand and mitigate in both applications.

These NGS systems have the potential to address several challenging issues in forensics analysis.[82] The analysis of forensics specimens that are mixtures (greater than one contributor) remains one of the most problematic issues in forensics from both a technical and statistical perspective. The clonal sequencing aspect of NGS provides the opportunity to recover different sequence reads for every genetic variant (allele) present in the mixture. Thus, the number of sequence reads recovered for a particular variant provides a digital read-out and allows a quantitative analysis of the contributors. Massively parallel NGS technologies also offer unparalleled capacity and allow for simultaneous analysis of STRs and SNPs as well as nuclear and mtDNA. The ability to analyze both mtDNA and nuclear markers in a single run would have broad forensic applications. Recently, there has been research exploring the use of NGS in forensic applications.[82,83] These initial studies have been primarily limited to a few STR or mtDNA markers with high initial amounts of DNA.[84,85]

iv. 454 DNA Sequencing Technology

The 454 DNA sequencing technology is a scalable highly parallel sequencing system which uses 1) emulsion PCR (emPCR) to amplify a single DNA fragment immobilized on a bead to 10 million identical copies and 2) pyrosequencing of DNA templates generated by the emulsion PCR[86]. Using the 454 GS FLX Titanium chemistry, >1 million high quality 400-500 bp reads with ~99% accuracy can be obtained in 10 hours (www.454.com). With the 454 GS Junior system, on average one hundred thousand 400-500 bp reads can be generated with 99% accuracy in a single 10 hour run. The clonal sequencing aspect of this technology coupled with the longer read lengths provides the ability to determine the haplotype and the relative ratio of mixed DNA samples[9]. The very large number of sequence reads allows the detection of sequences present in mixtures at ~1%[9]. The longer read lengths of the GS Titanium chemistry make it an ideal system for targeting STR markers using miniplex primer sets or mtDNA since the generated PCR products are typically 400 bps or less. The 400-500 bp or 800-1000 bp read lengths obtained using the 454 GS Titanium chemistry or 454 Flx Plus respectively far exceed the average read lengths generated using competing next generation sequencing technologies (Illumina Solexa and Ion Torrent) which average 150-300 bp per read[87]. The ability to determine mtDNA haplotypes and to read through repetitive STR sequence is critical to the success of this project and would be difficult using other next generation sequencing technologies because of the shorter read length, although the read length achieved by other platforms has been increasing. The longer read length allows for sequencing of the entire PCR product in a single read and therefore sequence fragment assembly is not required using the 454 sequencing system. Also, both strands can be sequenced (forward and reverse reads). Using competing technologies it would currently not be possible to sequence PCR products generated with the longer miniplex

STR primer sets in a single read. Also, the shorter read length of the competing technologies would make assembling STR sequences difficult or impossible because the repeat length sometimes exceeds the read length. The data file size is also much less per run for the 454 GS FLX system (~15 gigabytes) and even less for the 454 GS Junior making storing data more manageable compared to the Illumina Solexa (~1 terabyte) and ABI SOLiD (15 terabytes) systems

One limitation of the 454 pyrosequencing approach is its difficulty in identifying the number of repeats in long *homopolymeric* tracts (n>4)[9] similar to the Ion Torrent. For example, it is difficult to determine the number of C's in the poly C-stretches within the HVI/HVII regions of the mitochondrial genome. However, these poly C-stretches are often not considered when reporting differences because of the high frequency of length heteroplasmy in these regions[88-90]. This limitation of pyrosequencing would not apply to STR markers, since the repeats are not homopolymeric in nature but are di-, tri-, or tetra- nucleotide repeats. Also, the total number of base pairs of sequence generated using competing technologies is greater than the sequence yield of the 454 GS systems. However, unique sequence multiplex identification (MID) tags encoded in the PCR primers can be used to uniquely tag a large number of samples (i.e. 48, 64, 96…). Using MID tags to increase the number of samples processed in parallel significantly reduces the per sample cost, while still achieving the high level of sensitivity required to detect minor sequences in a mixed DNA sample. For these reasons, we chose to develop a mtDNA and STR next-generation sequencing assay using the 454 sequencing platform.

3. Project Research Design and Goals

Mitochondrial DNA sequencing using the next-generation sequencing technology is ideal for resolving mixed DNA samples. Mitochondrial DNA analysis has also proven useful for analysis of degraded samples, but has lower discrimination power. STR markers are ideal for detecting mixtures, but do not allow for resolution of mixtures. Using the 454 sequencing technology, additional sequence information is gained, and in some cases, may allow for further differentiation of homozygous STR alleles, that is, alleles that are indistinguishable in length but differ in sequence. However, this sequence information may not be enough to fully resolve mixed DNA samples using STR markers alone. While the 454 sequencing technology provides a more accurate measurement of relative ratios in a mixture (based on the number of sequence reads), STR's alone are not sufficient for resolving mixed samples (i.e. 50/50 mixture). Therefore, we proposed to target both mtDNA and nuclear STR markers using the 454 sequencing technology. Together, mtDNA and STR markers will allow for resolving mixtures and analysis of degraded DNA while providing the higher power of discrimination required in forensic cases. Our primary goals were to 1) develop and optimize "front-end" enrichment assays for sequencing mtDNA using the 454 NGS platform and test and apply the systems to forensically relevant samples, 2) develop, optimize, and test "front-end" multiplex mini-STRs PCR assays for sequencing the 13 core CODIS loci using the 454 NGS platform, and 3) modify next generation sequencing softwares for mtDNA and STR analyses with external collaborators. The research design, methods and results for each of the three major goals of this project are detailed in the sections below.

B. Methods and Results: 454 mtDNA HVI/HVII and Whole Mitochondrial Genome NGS Assay

Mitochondrial DNA is an ideal marker for 1) detection of degraded or limited DNA because of its high copy number and 2) resolving mixtures because it is haploid in nature. The 454 clonal sequencing technology is well suited for forensic applications because it has the capability to sequence individual sequence targets to determine an individual's mtDNA haplotype. We initially proposed to develop 1) a "front-end" fusion primer PCR for sequencing the HVI/HVII regions of the mitochondrial genome using the 454 sequencing technology and 2) a multiplex PCR assay targeting fifteen to twenty ~250 bp regions of the mitochondrial genome using three to four 5-plex PCRs using fusion primers sets to cover ~20-30% of the mitochondrial genome. While the multiplex PCR method for enriching mtDNA would likely be successful, recently probe based enrichment methods for capture and sequencing have drastically decreased in price making this a feasible alternative. As the probe capture enrichment method allows for capture and sequencing of the entire mtDNA genome and potentially degraded DNA samples, we explored this as an alternative to amplifying select regions of the mtDNA genome. We therefore modified our goal and developed 1) an HVI/HVII 454 fusion primer assay and 2) a solution phase probe capture assay for enrichment and sequencing of the entire mitochondrial genome using the 454 sequencing technology. An overview of the assay design and methods as well as the results from the optimization and application experiments are described in detail below for both the 454 HVI/HVII fusion primer PCR assay and the probe capture whole mitochondrial genome 454 sequencing assay.

1.*Design, Methods, and Results: HVI/HVII 454 Fusion Primer assay*

i. Design : 454 HVI/HVII Fusion Primer PCR

A total of eight sets of multiplex identifier (MID) tagged fusion primers targeting the HVI/HVII regions of the mitochondrial genome were designed and tested. Fusion primers used in the initial amplification consist of 1) a common forward or reverse adapter sequence that serves as a universal primer for emPCR and pyrosequencing, 2) a 4 base library key for pyrosequencing signal normalization, 3) a unique 10 base sample-specific internal sequence tag (MID tag) used for sample 'barcoding', and 4) the forward or reverse target specific PCR primer sequence. Fusion primer sequences for amplification of the HVI and HVII regions of the mitochondrial genome are provided below in Figures 1a and 1b respectively.

Figure 1a. HVI Forward and Reverse Fusion Primers

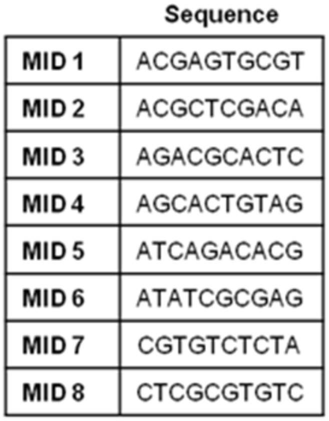
mtDNA HVI	ADAPTOR SEQUENCE	LIBRARY TAG	MID TAG	LOCUS SPECIFIC PRIMER
F15975-93_1	CGTATCGCCTCCCTCGCGCCA	TCAG	ACGAGTGCGT	CTCCACCATTAGCACCCAA
R16481-01_2	CTATGCGCCTTGCCAGCCCGC	TCAG	ACGCTCGACA	ATTTCACGGAGGATGGTG

Figure 1b. HVII Forward and Reverse Fusion Primers

mtDNA HVII	ADAPTOR SEQUENCE	LIBRARY TAG	MID TAG	LOCUS SPECIFIC PRIMER
F15-34	CGTATCGCCTCCCTCGCGCCA	TCAG	ACGAGTGCGT	CACCCTATTAACCACTCACG
R429-410	CTATGCGCCTTGCCAGCCCGC	TCAG	ACGAGTGCGT	CTGTTAAAAGTGCATACCGC

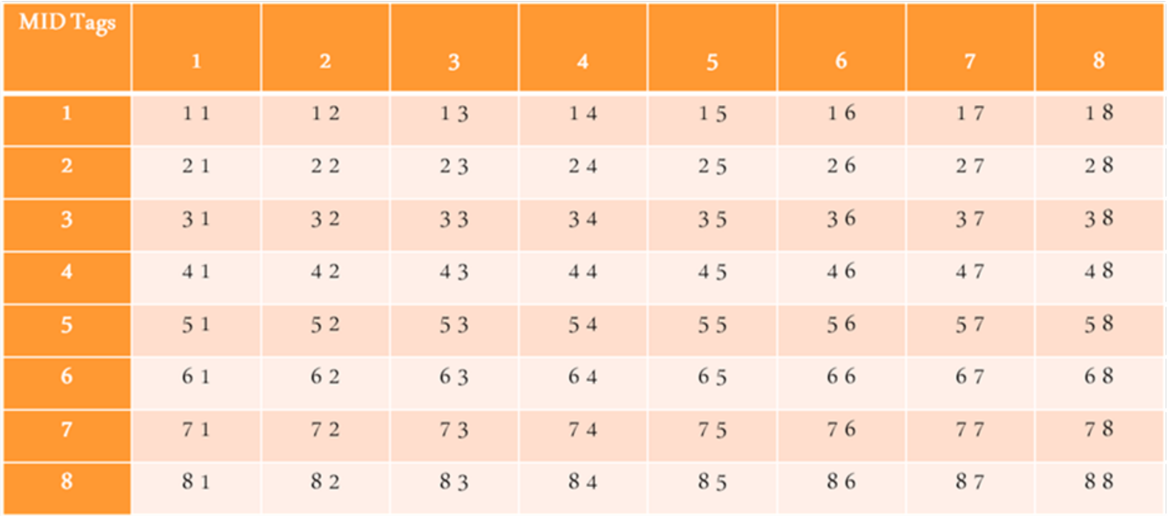
Unique 10 base MID tags are used as sample identifiers in order to pool and sequence multiple samples in a single 454 sequencing run with each sample being 'tagged' or 'barcoded' with a different MID tag. The eight sets of MID tags used for the fusion primer sets are provided in Table 1 below. The eight sets of MID tagged fusion primers were designed to be used in a combinatory approach to generate 64 different combinations of forward and reverse MID tagged HVI/HVII PCR products (see Table 2 for the 64 different MID combinations).

Table 1. 10 Base Extended Multiplex Identifier (MID) Set Sequences (MID Tags 1-8)



	Sequence
MID 1	ACGAGTGCGT
MID 2	ACGCTCGACA
MID 3	AGACGCACTC
MID 4	AGCACTGTAG
MID 5	ATCAGACACG
MID 6	ATATCGCGAG
MID 7	CGTGTCTCTA
MID 8	CTCGCGTGTC

Table 2. 64 Combinatory MID Sequence Tags for Sample Pooling



MID Tags	1	2	3	4	5	6	7	8
1	1 1	1 2	1 3	1 4	1 5	1 6	1 7	1 8
2	2 1	2 2	2 3	2 4	2 5	2 6	2 7	2 8
3	3 1	3 2	3 3	3 4	3 5	3 6	3 7	3 8
4	4 1	4 2	4 3	4 4	4 5	4 6	4 7	4 8
5	5 1	5 2	5 3	5 4	5 5	5 6	5 7	5 8
6	6 1	6 2	6 3	6 4	6 5	6 6	6 7	6 8
7	7 1	7 2	7 3	7 4	7 5	7 6	7 7	7 8
8	8 1	8 2	8 3	8 4	8 5	8 6	8 7	8 8

This combinatory approach for MID sample tagging greatly reduces the number of primers required to achieve the maximum number of samples that can be pooled in one sequencing run as illustrated in the schematic in Figure 2 below. Using a combinatorial approach for amplification of the HVI/HVII regions using the eight sets of optimized duplex 454 MID tagged fusion primers, a total of 64 different samples can be uniquely tagged and pooled in a single 454 sequencing run. Using this approach, only eight sets of fusion primers are needed for each PCR target compared to 64 sets of fusion primers, greatly reducing the overall cost and quality control effort.

Figure 2.

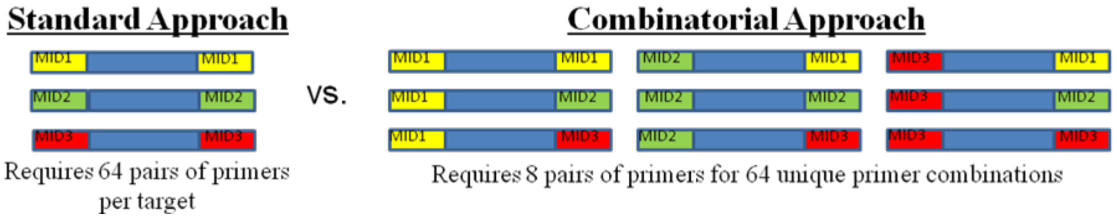


Figure 2. Schematic of Standard MID Tagging approach VS Combinatorial Approach. The standard approach for sample MID tagging uses the same MID barcode for both forward and reverse fusion primers while the combinatory approach for sample MID tagging barcodes each sample with a combination of two different MID sequences per sample. This figure illustrates that

with 3 sets of MID tagged fusion primers, 3 unique samples are barcoded using the standard approach while 9 samples can be uniquely tagged using the combinatory approach.

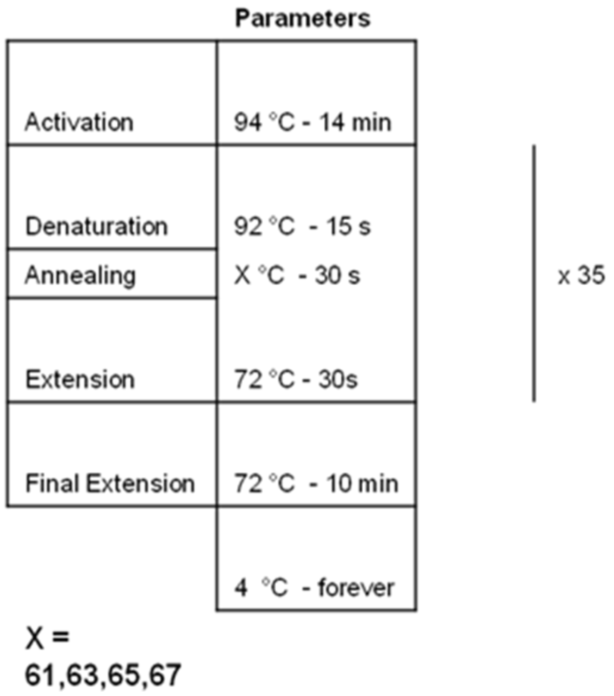
ii. Methods

a. Methods Optimization

1) HVI/HVII 454 Fusion primer PCR optimization

The HVI/HVII duplex PCR cycling parameters and master mix formulation were used as a starting point for the 454 HVI/HVII fusion primer PCR as the target specific primer sequences are the same between the assays. Due to the significant increase in the 454 fusion primer length and resulting melting temperature (TM), the PCR annealing temperature required optimization to minimize primer dimer and improve amplification yield and specificity for the 454 HVI/HVII fusion PCR. Two different strategies were tested with four sets of MID tagged 454 HVI/HVII fusion primers: 1) a standard approach using a single higher annealing temperature and 2) a two stage annealing temperature approach with a low annealing temperature for the initial 10 cycles then shifting to a higher annealing temperature for the later 25 cycles. Four sets of 454 MID tagged fusion primers were tested over a range of annealing temperatures (61° C, 63° C, 65° C, 67° C) increased from the 59° C duplex HVI/HVII PCR (Table 3). PCR products were visualized using gel electrophoresis to assess the level of primer dimer, non-specific amplification, and overall product yield. Results showed that 61° C and 63° C annealing temperatures resulted in overall higher amounts and incidences of primer dimer with lower product yields compared to 65° C and 67° C. Higher product yields were observed at annealing temperatures of 65° C and 67° C with minimal primer dimer (sporadic occurrences).

Table 3. Range of Temperatures Tested for Single Step Annealing

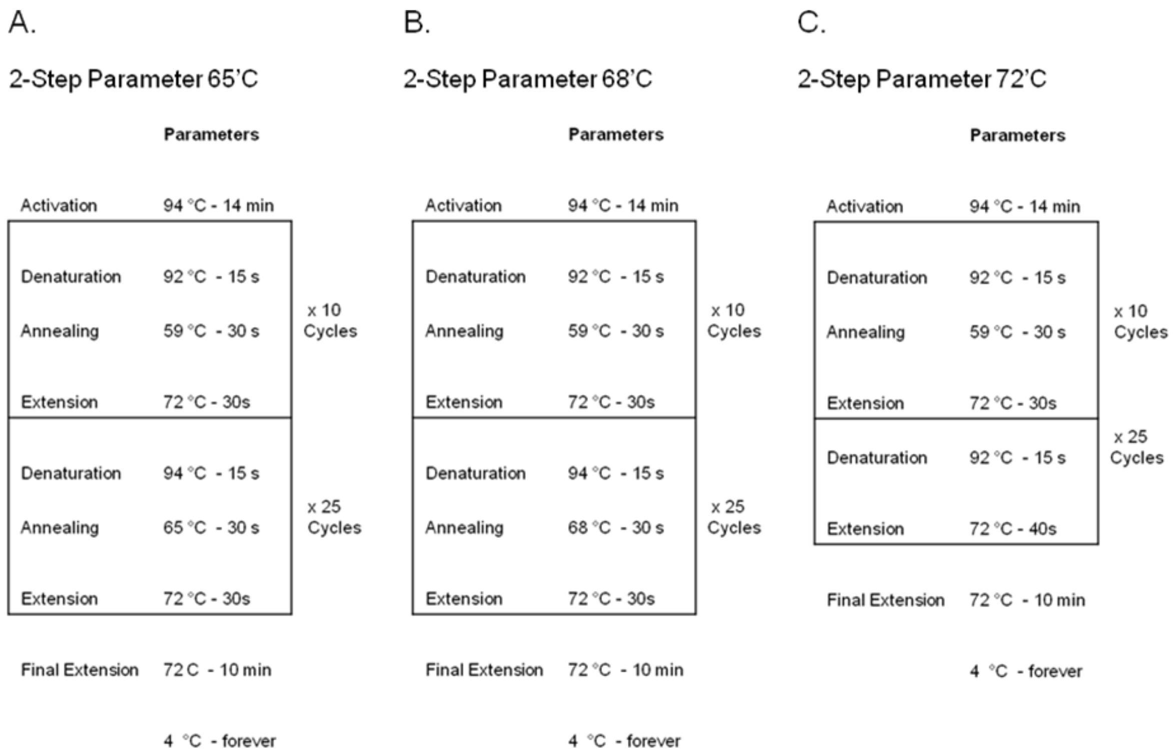


	Parameters	
Activation	94 °C - 14 min	
Denaturation	92 °C - 15 s	
Annealing	X °C - 30 s	x 35
Extension	72 °C - 30s	
Final Extension	72 °C - 10 min	
	4 °C - forever	

X =
61,63,65,67

To try to further minimize primer dimer, a two stage annealing temperature approach was investigated. A lower annealing temperature during early cycles may result in higher product yields since during the initial cycles of the PCR, only the target specific sequence of the primer anneals to the DNA template which has a lower TM; a higher annealing temperature in later cycles may result in increased specificity and yield as once the 454 and MID sequence specific regions of the primer are incorporated into the product, the entire fusion primer anneals to the product and has a much higher TM. Three different sets of parameters for the 2 stage annealing temperature approach were tested (Table 4). Overall, results for the 2 stage annealing temperature resulted in lower yield or increased primer dimer or were similar to yields observed with a single 65° C or 67° C annealing temperature. Based on these results, a single 65° C annealing temperature was selected for the finalized PCR parameters.

Table 4. Parameters for 2 Step Annealing



A.

2-Step Parameter 65'C

	Parameters	
Activation	94 °C - 14 min	
Denaturation	92 °C - 15 s	x 10 Cycles
Annealing	59 °C - 30 s	
Extension	72 °C - 30s	
Denaturation	94 °C - 15 s	x 25 Cycles
Annealing	65 °C - 30 s	
Extension	72 °C - 30s	
Final Extension	72 C - 10 min	
	4 °C - forever	

B.

2-Step Parameter 68'C

	Parameters	
Activation	94 °C - 14 min	
Denaturation	92 °C - 15 s	x 10 Cycles
Annealing	59 °C - 30 s	
Extension	72 °C - 30s	
Denaturation	94 °C - 15 s	x 25 Cycles
Annealing	68 °C - 30 s	
Extension	72 °C - 30s	
Final Extension	72 °C - 10 min	
	4 °C - forever	

C.

2-Step Parameter 72'C

	Parameters	
Activation	94 °C - 14 min	
Denaturation	92 °C - 15 s	x 10 Cycles
Annealing	59 °C - 30 s	
Extension	72 °C - 30s	
Denaturation	92 °C - 15 s	x 25 Cycles
Extension	72 °C - 40s	
Final Extension	72 °C - 10 min	
	4 °C - forever	

In addition to annealing temperature, $MgCl_2$ concentrations and the PCR volume were varied to determine the optimal conditions. A range of $MgCl_2$ concentrations (2.4 and 1.6 mM) were tested to determine the optimal concentration for increased specificity and amplification yield. A control sample HL-60 was used for testing. Four different MID-tagged primer pairs were used to amplify the samples in duplicates. Results showed significantly higher product yields with 2.4 mM $MgCl_2$ compared to 1.6 mM $MgCl_2$. Therefore, 2.4 mM $MgCl_2$ was kept as the final concentration in the master mix. Additionally, two PCR volumes were tested, 50 µL and a reduced 25 uL volume. Results showed an overall greater yield with 50 µL compared to the reduced 25 µL volume, and hence a 50 uL PCR volume was maintained. Based on these results, no changes were made to the original HVI/HVII duplex PCR master mix formulation for the 454 HVI/HVII fusion PCR.

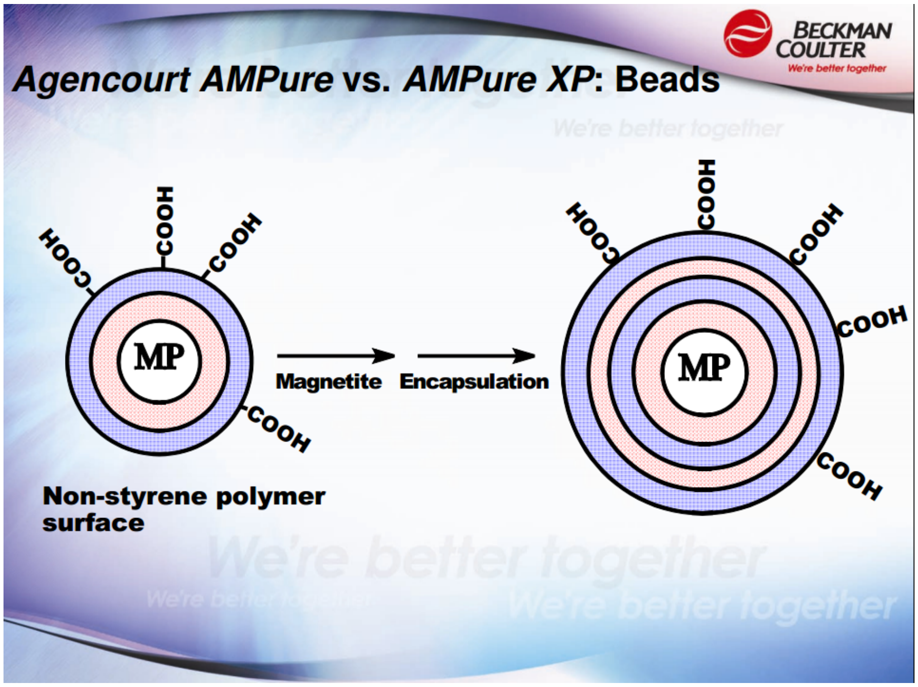
2) AMPure Purification to remove small products (primer dimer)

While PCR conditions were optimized to minimize primer-dimer, it was observed sporadically and at a low frequency in specific MID tagged primer sets. Primer-dimer can ultimately reduce the overall target sequence reads in a 454 sequencing run if not removed. To remove excess primer as well as primer-dimer, a purification step can be added after amplification. Purification using filtration columns was tested and we found that while excess primer was removed, the primer-dimer products from the long fusion primers was too large (125-150 bp) to be removed efficiently by filtration. Filtration columns typically remove <70-100 bp DNA fragments depending on the pore size of the membrane of the filtration system.

We then explored PCR product purification using Agencourt AMPure XP beads to selectively purify primer dimer products from the PCR products. Purification using AMPure beads is ideal due to its DNA concentration independence and the ability of removing varying sizes of short DNA fragments by altering the AMPure:DNA solution volumetric ratio. Agencourt AMPure XP beads reversibly bind DNA in the presence of the "crowding agent" polyethylene glycol (PEG) and salt (20% PEG, 2.5 NaCl). PEG causes the negatively-charged DNA to bind with the carboxyl groups on the bead surface (See Figure 3 below). The volumetric ratio of AMPure XP beads to DNA is critical as the immobilization is dependent on the concentration of PEG and salt in the reaction. Larger fragments bind to the SPRI beads and displace smaller fragments with lower concentration of PEG. Smaller fragments are retained with higher PEG concentrations as they get crowded and bind to the SPRI beads. The size of the DNA fragment that binds to SPRI beads is dependant on the PEG concentration. As PEG and salt concentration is constant in AMPure solution, the volumetric ratio of the AMPure:DNA solution determines the PEG concentration. The higher the volumetric ratio of AMPure solution to DNA sample volume, the higher the PEG, and the smaller the size of the fragment that will be retained. Size selection of DNA fragments using

AMPure beads is DNA concentration independent because DNA copy number will not affect the concentration of PEG in AMPure solution.

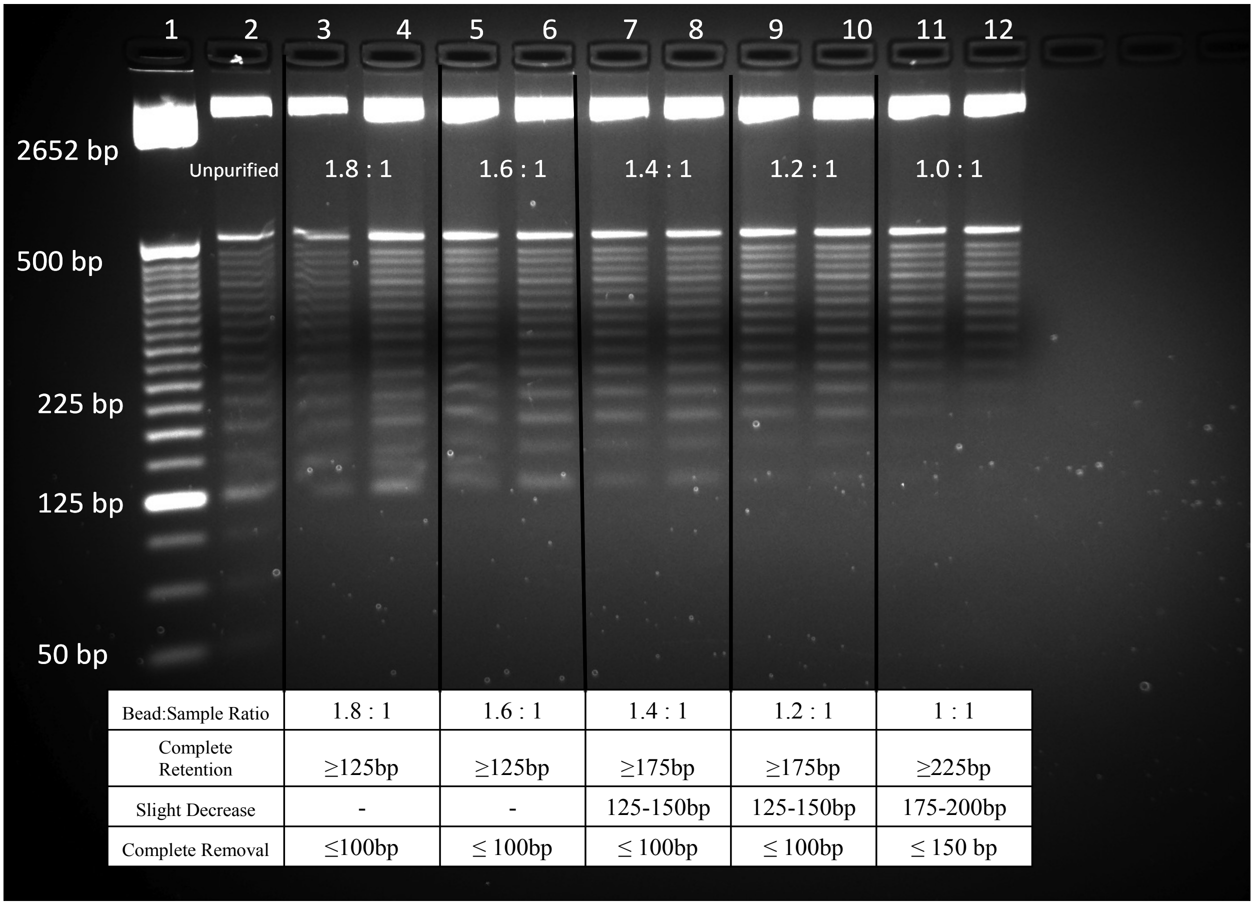
Figure 3. Agencourt AMPure XP magnetite Bead with Non-styrene polymer surface with carboxyl coating, suspended in PEG and salt solution



Agencourt AMPure vs. AMPure XP: Beads

The critical aspect for effective removal of small fragments as discussed above is applying an optimal AMPure bead solution to DNA sample volume, therefore we varied the volumetric ratios to determine the optimal ratio for removal of the long primer dimer (~125 bp) while retaining the HVI/HVII mtDNA products (>450 bp). We tested the effectiveness of small fragment removal using a range of AMPure Bead:DNA volume ratios (1.8:1, 1.6:1, 1.4:1, 1.2:1, 1.0:1) by purifying a 25 bp DNA ladder. A gel image is shown below in Figure 4 with a table summarizing the results. We found that 1.8:1 – 1:1 successfully removed products 100 bp or less. However, complete removal of 150 bp or less required a 1:1 ratio. Partial removal of 125-150 bp products was achieved with 1.4 :1 - 1.2 :1 ratio. We then tested a range of AMPure bead:DNA sample volume ratios with HVI/HVII amplified products with apparent primer dimer to determine the optimal bead:sample ratio and found that a 1:1 – 1.2:1 ratio removed the primer-dimer without removing the HVI/HVII products. Based on these results, it was determined that a 1:1 AMPure bead solution to sample ratio was optimal to successfully remove the large primer dimer products and still retain the HVI/HVII products.

Figure 4. Testing the Effectiveness of Small DNA Fragment Removal by Varying the AMPure:Bead Volume Ratio



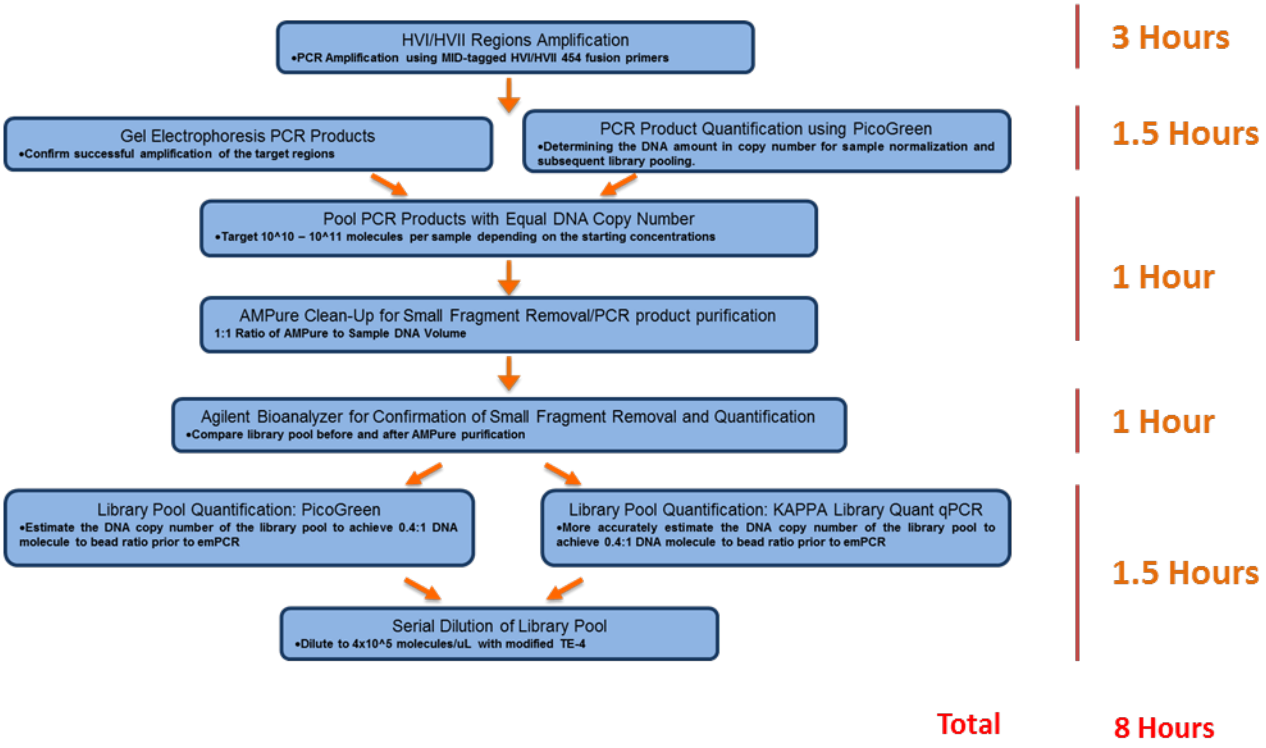
Bead:Sample Ratio	1.8 : 1	1.6 : 1	1.4 : 1	1.2 : 1	1 : 1
Complete Retention	≥125bp	≥125bp	≥175bp	≥175bp	≥225bp
Slight Decrease	-	-	125-150bp	125-150bp	175-200bp
Complete Removal	≤100bp	≤ 100bp	≤ 100bp	≤ 100bp	≤ 150 bp

b. Final Optimized “Front-end” 454 HVI/HVII Library Preparation Methods

A general workflow of the final optimized “Front-end” 454 HVI/HVII library method used for the validation and testing is provided in the Figure 5 below and the methods used are described in detail below.

Figure 5.

454 HVI/HVII Library Preparation Workflow



HVI/HVII Regions Amplification
- PCR Amplification using MID-tagged HVI/HVII 454 fusion primers

Gel Electrophoresis PCR Products
- Confirm successful amplification of the target regions

PCR Product Quantification using PicoGreen
- Determining the DNA amount in copy number for sample normalization and subsequent library pooling.

Pool PCR Products with Equal DNA Copy Number
- Target 10^10 – 10^11 molecules per sample depending on the starting concentrations

AMPure Clean-Up for Small Fragment Removal/PCR product purification
1:1 Ratio of AMPure to Sample DNA Volume

Agilent Bioanalyzer for Confirmation of Small Fragment Removal and Quantification
- Compare library pool before and after AMPure purification

Library Pool Quantification: PicoGreen
- Estimate the DNA copy number of the library pool to achieve 0.4:1 DNA molecule to bead ratio prior to emPCR

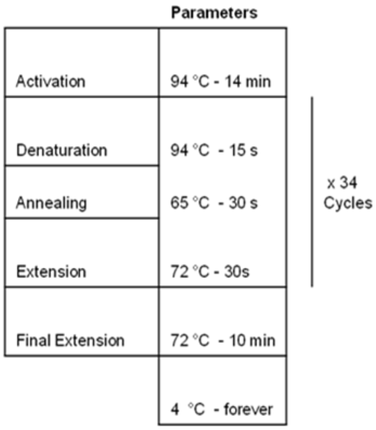
Library Pool Quantification: KAPPA Library Quant qPCR
- More accurately estimate the DNA copy number of the library pool to achieve 0.4:1 DNA molecule to bead ratio prior to emPCR

Serial Dilution of Library Pool
- Dilute to 4x10^5 molecules/uL with modified TE-4

1) HVI/HVII Regions Amplification

Each genomic DNA sample is amplified using a unique MID-tagged 454 HVI/HVII fusion primer set in order to enrich for the HVI and HVII target regions of the mitochondrial genome and incorporate the 454 sequencing primer, library key, and unique MID barcode sequence necessary for emPCR, pyrosequencing, and sample identification. The optimized PCR parameters for the 454 HVI/HVII fusion primer PCR are presented in Table 5 below.

Table 5. Final PCR Parameters for the 454 HVI/HVII Fusion Primer PCR



	Parameters	
Activation	94 °C - 14 min	
Denaturation	94 °C - 15 s	x 34 Cycles
Annealing	65 °C - 30 s	
Extension	72 °C - 30s	
Final Extension	72 °C - 10 min	
	4 °C - forever	

2) Gel Electrophoresis PCR Products

The amplified HVI/HVII products were then visualized via gel electrophoresis using 3% SeaKem agarose gel prepared with 1X TAE to confirm successful amplification of the target regions and assess primer dimer and PCR specificity.

3) PCR Product Quantification using PicoGreen

Each PCR product was quantified using Quant-iT™ PicoGreen® dsDNA kit (Invitrogen, Carlsbad, CA) following the manufacturer's protocol to determine the DNA amount in copy number for sample normalization and subsequent library pooling. Each amplified sample was combined into a single library pool, varying the volume of the PCR product to target 10^10 – 10^11 molecules per sample depending on the starting concentrations.

Sample normalization is important to achieve similar coverage or number of sequence reads per sample. Pooling the individual sample PCR products to a single DNA library prior to purification greatly reduces the number of individual samples needed to be processed. We observed that pooling prior to purification does not result in a significant difference in the number of sequence reads between individual samples when compared to normalizing and pooling individual samples after the purification process.

4) AMPure Clean-Up for Small Fragment Removal/PCR Product Purification

Small fragment DNA or primer dimers were removed from the pooled library using Agencourt AMPure XP (Beckman Coulter, Pasadena, CA) with 1:1 AMPure to DNA sample volumetric ratio. Agencourt AMPure XP beads were used as the purification method due to its DNA concentration independence and its ability of removing varying sizes of short DNA fragments by altering the DNA:AMPure volumetric ratio.

A schematic showing an overview of the method for small fragment removal using AMPure beads is shown in Figure 6 below. The DNA and AMPure bead solution are mixed together at a determined ratio to effectively remove small fragments while retaining the larger PCR products of interest. Then the captured fragments are separated using a magnet, washed with ethanol and eluted from the beads resulting in purified PCR products and removal of small fragments.

Figure 6.

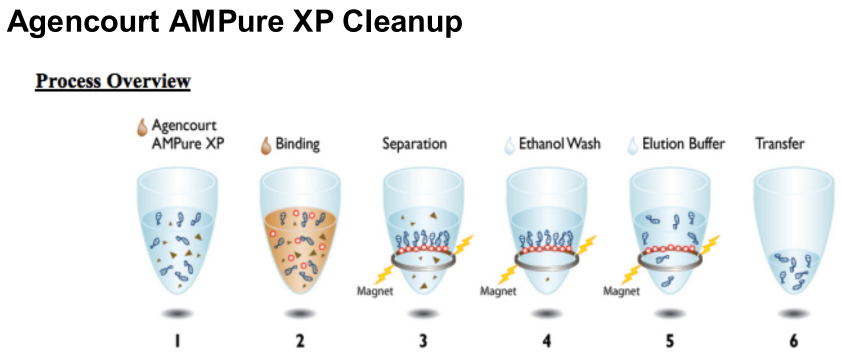


Figure 6. Overview of Small Fragment DNA Removal Using AMPure XP. Sample DNA or PCR product is bound to the Agencourt AMPure XP beads with appropriate volumetric ratio. Unbound small DNA fragments are discarded as supernatant while the captured DNA fragments are retained using a magnet. Trace PEG and salt solution are washed with 70% ethanol, and size selected PCR product is re-eluted with elution buffer (TE, 0.1 mM EDTA, pH 8.0) or molecular

grade water. The purified PCR product is then transferred to a new container while AMPure beads are separated from the product using the magnet.

5) Agilent Bioanalyzer for Confirmation of Small Fragment Removal and Quantification

Agilent DNA 1000 kit (Agilent Technologies, Santa Clara, CA) was used following manufacturer's protocol in order to compare the library pool before and after AMPure purification to confirm the removal of primer dimers.

6) Library Pool Quantification: PicoGreen

The concentration of the purified library pool was estimated by using Quant-iT™ PicoGreen® dsDNA kit (Invitrogen, Carlsbad, CA) following manufacturer's protocol to estimate the DNA copy number of the library pool to achieve 0.4:1 DNA molecule to bead ratio prior to emPCR.

7) Library Pool Quantification: KAPA Library Quant qPCR

Alternatively, the concentration of the purified library pool in DNA copy numbers was determined by using Library Quantification Kit - 454 FLX (Kapa Biosystems, Wilmington, MA) to achieve 0.4:1 DNA molecule to bead ratio prior to emPCR. The Kapa Library Quantification qPCR provides a more accurate estimate of the DNA copy number of the library pool as the qPCR assay measures the copy of each molecule with 454 primer sequence, rather than intercalating double stranded DNA molecules non-specifically and using estimated product size. The manufacturer's recommended protocol was used with the following modifications. The standards were run in duplicate instead of the manufacturer's recommended triplicate of each standard in order to decrease the cost of each qPCR run after showing experimentally that duplicates were sufficient based on the replicate data. The dilution series of the pooled library was modified from the manufacturer's recommendation based on the consistently higher starting concentration of our pooled libraries.

8) Serial Dilution of Library Pool

The quantified and cleaned library pool was then diluted with modified TE-4 (10mM TRIS, 0.1mM EDTA, pH 8.0) to the concentration of 4x10^5 molecules/uL in order to add a total of 10 uL to achieve a 0.4:1 DNA to bead ratio in emPCR. The optimal DNA to bead ratio was determined to be 0.4:1 based on the number of high quality pass filter reads. Targeting the optimal DNA:bead ratio is critical to avoid the incorporation of more than one molecule of DNA per bead or insufficient bead recovery.

9) EmPcR and 454 Sequencing

a) Emulsion PCR

The 454 sequencing method begins with an emulsion PCR for clonal amplification of sequencing targets. The emulsion PCR consists of oil, water, PCR reagents and sequencing beads mixed to create microreactors. Within each microreactor is a sequencing bead which is affixed with the sequencing primer, allowing for clonal amplification of a library target on each bead. Inside the droplet, the individual DNA molecule undergoes amplification resulting in millions of clonal copies attached to each bead. This process occurs for each individual DNA molecule in the enriched sample library. Clonal amplification allows for better detection of rare mutations because each molecule is copied separately, helping in resolving mixtures and heteroplasmy. An overview of the clonal amplification steps are shown in figure II-2, and the amplification conditions are as follows: 4 minutes at 94C, 50 cycles of 94C for 30 seconds, 58C for 4.5 minutes, 68C for 30 seconds, and then hold at 10C forever.

Emulsion PCR Workflow

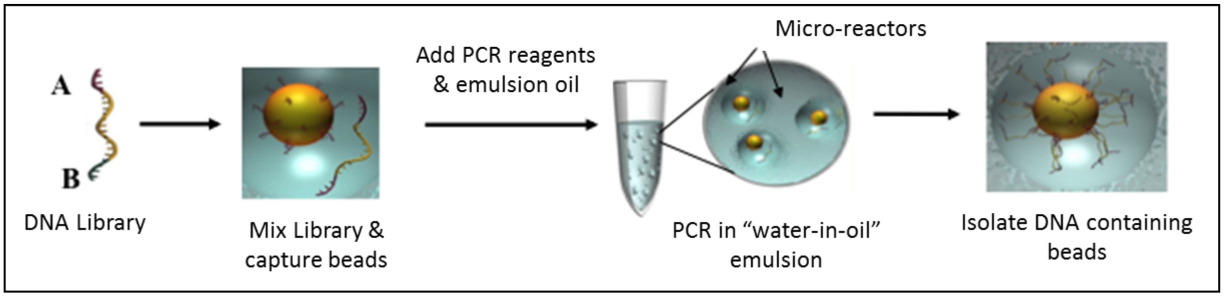

A
B
DNA Library
Mix Library & capture beads
Add PCR reagents & emulsion oil
Micro-reactors
PCR in "water-in-oil" emulsion
Isolate DNA containing beads

Figure 7. Workflow of 454 Emulsion PCR. The emulsion is a water in oil mixture which creates micro-reactors where the amplification occurs. The mico-reactors contain one bead bound with A/B primers, one library template, and PCR components allowing for clonal amplification over the surface of the bead. After amplification, enrichment of beads containing DNA occurs before they are loaded onto the 454 instrument.

After the emulsion PCR, the emulsions needs to be broken and the beads are subjected to a series of washes to remove all excess PCR reagents and oil. Then the beads are enriched for only ones containing DNA, so space on the instrument is not wasted by loading empty beads. Once the beads are enriched they are ready for sequencing on the 454 sequencer.

b) 454 Sequencing Chemistry and Instrumentation

The 454 sequencing chemistry is a sequencing-by-synthesis process which utilizes pyrosequencing chemistry to determine the nucleotide sequence. First, the enriched beads are loaded onto the PicoTiter plate (see Figure 8) which only allows one sequencing bead per well. Then they are packed in with enzyme beads which aid in the pyrosequencing process.

The Roche 454 technology uses a method that obtains the target sequence as the DNA fragments are synthesized, called pyrosequencing. Serving as a template strand, the sample DNA together with six other components—DNA polymerase, deoxribonucleotide triphosphates (dNTP), ATP sulfurylase, apyrase, luciferase, and adenosine 5' phosphosulfate (APS)—causes the sequencing reaction. Cycles of each dNTP flow through the reaction; if that base is complementary to the template strand, DNA polymerase will incorporate the dNTP into the growing strand, releasing a pyrophosphate in the reaction. The pyrophosphate in the presence of APS reacts with ATP sulfurylase to create ATP. The ATP created in the reaction then interacts with luciferase in a reaction that releases visible light, which is read by a camera and interpreted by a computer. Apyrase is used at the end of each flow cycle to degrade unused dNTPs. This process continues for a determined number of cycles (200), resulting in sequence reads of about 400-500 bp.

Figure 8. Loading of DNA Bound Beads and Enzyme Beads on the PicoTiter Plate.

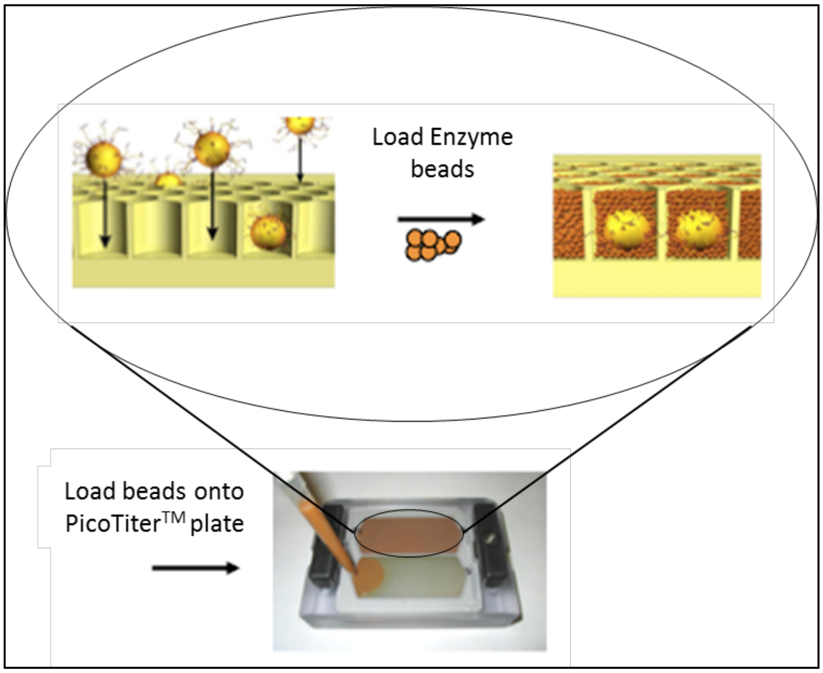
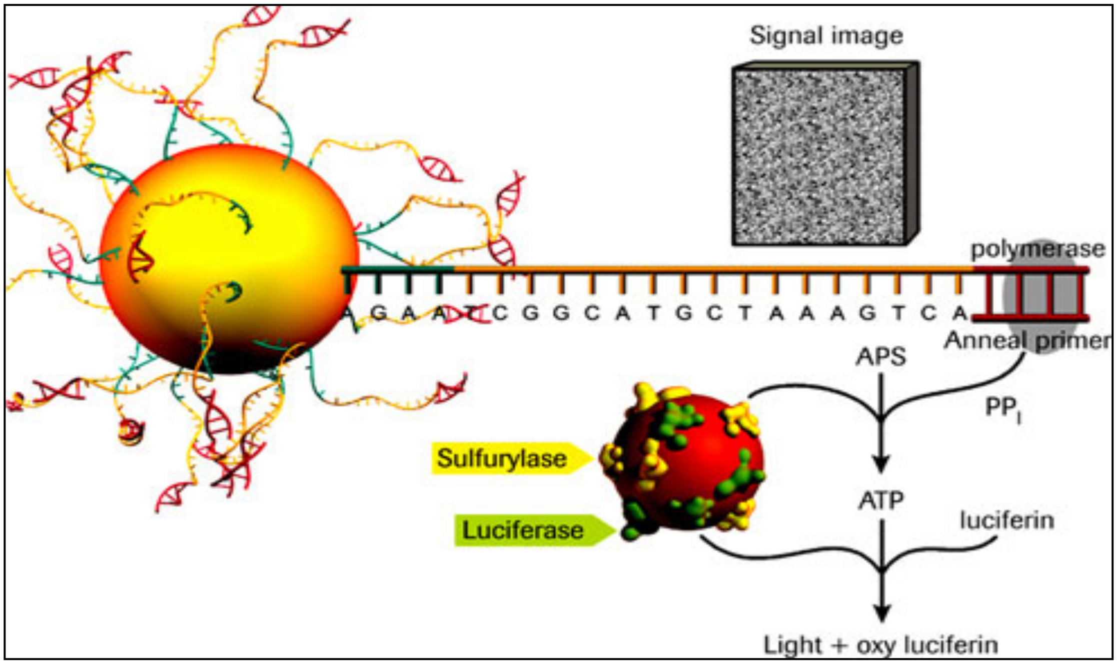


Figure 9. 454 Pyrosequencing Chemistry.



iii. Results

a) Overview of 454 mtDNA HVI/HVII Sequencing Results

A total of eight 454 NGS runs were conducted for the assay targeting the HVI/HVII regions. The sequencing runs were dedicated to validation studies such as mixture studies investigating mixture samples with 2-5 contributing sequences with varying mixture ratios, sensitivity studies and various experiments to characterize the jumping PCR effects and stochastic effects of duplex PCR assay. Further, several runs were dedicated to generating 454 NGS population database and for characterization of heteroplasmy in different tissue types (blood, buccal, and 5 hair samples) of 6 monozygotic twin pairs. All runs were successful and met all 454 benchmarks for amplicon sequencing. The run statistics for the sequencing run results generated using HVI/HVII assays are summarized in the Table 6 below.

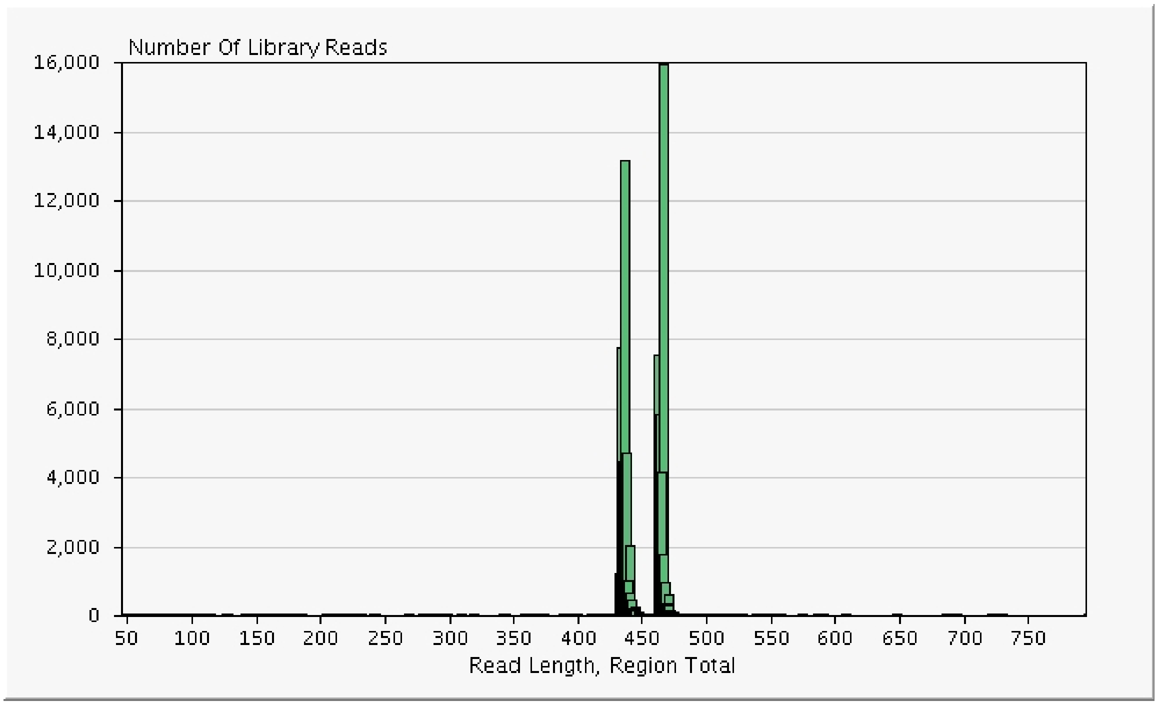
Table 6. 454 GS Jr. Run Statistics from HVI/HVII Assay Runs

Raw wells		100,478 - 220,201
Key Pass Wells		92,403 - 207,537
Failed	*Dot*	904 - 5,053
	Mixed	2,246 - 17,797
	Short quality	27,965 - 120,433
	Short Primer	17 - 3,887
Passed Filter Wells		44,050 - 82,284
	% Dot+ Mixed	3.01 - 13.28
	% Short	28.49 - 53.19
	% Passed Filter	26.18 - 68.49

All runs showed >90% total key pass wells, with the projected ~50,000-75,000 high quality filtered reads. The percent dot and mixed filters that process reads based on signal quality fell in the range of .301-13.28%, well below the recommended <20%. The low dot+mixed % indicate the optimized DNA:bead ratio for emPCR for high quality reads. The percent trimmed short quality filter was observed in the range of 28.49 – 53.19%, indicating the optimized AMPure method was successful for short fragment removal. Prior to the sequencing run, all libraries are purified using the Ampure beads to eliminate all short products (unused primers and primer dimers) and short product removal is confirmed using a Bioanalyzer. Further, short fragment removal was validated during optimization by analyzing the product DNA libraries by a QC PCR (Roche technical Bulletin) to ensure complete removal of all short products. The optimization of small fragment removal using the Agencourt AMPure XP purification method was further confirmed to be effective, demonstrated by the distribution of readlengths collected in the 454 HVI/HVII Next Generation

Sequencing runs. The read length distribution of a 454 HVI/HVII sequencing run presented in Figure 10 below demonstrates two clean peaks around 450 bp each corresponding to amplified HVII and HVI regions. The absence of any peak below ~430 bp indicates there was not any DNA fragments smaller the expected HVI and HVII amplicons present, and hence the primer dimers were removed sufficiently.

Figure 10. Distribution of Reads for 454 HVI/HVII NGS Assay



The 454 sequencing results also confirmed that the PCR parameters and library preparation steps including primer dimer removal were optimized. The Table 7 below shows the number of reads for HVI/HVII forward and reverse for the three mtDNA samples sequenced in the first run, which is similar for all runs. The average total number of reads obtained was similar for both the HVI and HVII regions. We found that there was less variation in the number of reads between regions than between samples and sequencing direction. We show here that using an optimized PCR system similar read depth can be achieved across regions by co-amplifying the regions in a multiplex PCR. Co-amplification is important for conserving sample material in cases where DNA may be limited or degraded and will be essential if the number of targets is increased.

Table 7.

		Number Reads		
		Forward	Reverse	Total
HVII	Sample 1	558	881	1439
	Sample 2	330	470	800
	Sample 3	253	370	623
	Average	380	574	954
HVI	Sample 1	730	652	1382
	Sample 2	554	444	998
	Sample 3	334	307	641
	Average	539	468	1007

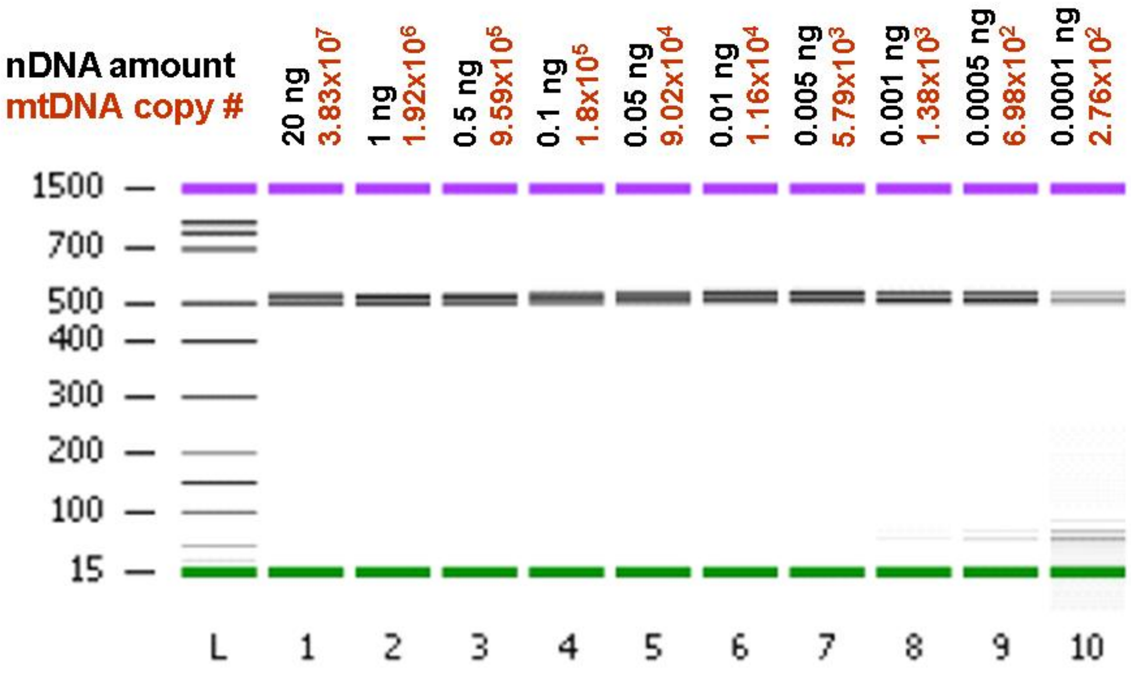
b. Results for Validation Studies and Testing of Forensically Relevant Samples

1) Sensitivity Study

A sensitivity study using MID tagged fusion primers for co-amplification of the HVI/HVII regions was conducted at 35 cycles. The nuclear DNA concentration as well as the mtDNA copy number was determined using a duplex TaqMan qPCR assay (Timken et al) for two DNA samples. Two samples were amplified in duplicate at 10 different DNA concentrations ranging from 20 ng – 0.1 pg. Mitochondrial copy number differed between the two samples and ranged from 3.83 x 10^7 – ~275 copies for one sample (K562) and 4.02 x 10^6 - ~90 copies for the second sample (G147A). The nDNA:mtDNA ratio varied between samples which is not unexpected. Based on the detectable yields using a BioAnalyzer, amplification was successful down to ~90 mtDNA copies or

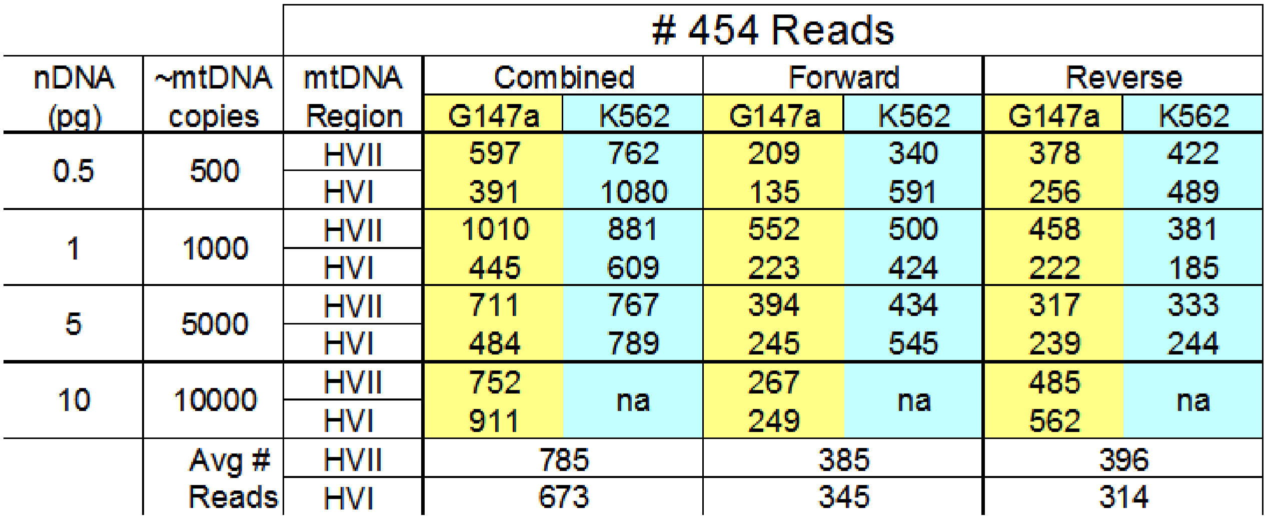
0.1- 0.5 pg of DNA at 35 cycles. Bioanalyzer data for sample K562 is shown below in Figure 11 for the HVI/HVI duplex at 35 cycles.

Figure 11. Bioanalyzer Results from a Sensitivity Study



A subset of the samples prepared for the sensitivity study was sequenced using the 454 to determine the sensitivity (10 pg, 5pg, 1pg and 0.5 pg) ranging from 500-10,000 mtDNA copies. The total combined number of 454 reads as well as the number of forward and reverse reads for each of the HVI and HVII regions are reported in Table 8 below for each sample amount tested. A similar number of reads were observed for combined, forward and reverse for each region for all sample amounts tested ranging from 0.5 pg – 10 pg. These results show successful amplication and 454 sequencing of 0.5 pg or DNA or ~500 mtDNA copies at 35 cycles using the 454 fusion Primer PCR.

Table 8. Sensitivity Study 0.5-10 pg DNA at 35 Cycles



		# 454 Reads						
nDNA (pg)	~mtDNA copies	mtDNA Region	Combined		Forward		Reverse	
			G147a	K562	G147a	K562	G147a	K562
0.5	500	HVII	597	762	209	340	378	422
		HVI	391	1080	135	591	256	489
1	1000	HVII	1010	881	552	500	458	381
		HVI	445	609	223	424	222	185
5	5000	HVII	711	767	394	434	317	333
		HVI	484	789	245	545	239	244
10	10000	HVII	752	na	267	na	485	na
		HVI	911		249		562	
	Avg # Reads	HVII	785		385		396	
		HVI	673		345		314	

We also conducted a more extensive sensitivity study to determine the detection limit of the starting DNA for the mtDNA HVI/HVII duplex 454 sequencing assay at the optimized 34 cycles as well as to explore the stochastic effect in the PCR amplification step of the assay. A control DNA sample (K562) was quantified using a mt:nu qPCR assay (Timken et al.2005) and was amplified at varying concentrations prepared via serial dilutions as follows: 1 ng, 500 pg, 100 pg, 50 pg, 10 pg, 5 pg, 1 pg, 0.5 pg, 0.1 pg, 0.05 pg, 0.01 pg, 0.005 pg, and 0.001 pg. The range of starting DNA concentrations was broadened from the initial sensitivity study conducted to determine the lower sensitivity and the stochastic threshold of the PCR assay. The PCR products from a starting amount of 0.5 pg of DNA and higher have resulted in clear two bands when visualized under 3% agarose gel stained with ethidium bromide, indicating successful amplification of HVI/HVII regions (Figure 12). The starting amounts of 0.1 pg and 0.05 pg showed faint bands corresponding to HVI/HVII regions, while samples below 0.05 pg showed no visible bands on the gel. Stochastic amplification was not observed on the gel electrophoresis (i.e. amplification of a single hypervariable region).

Figure 12.

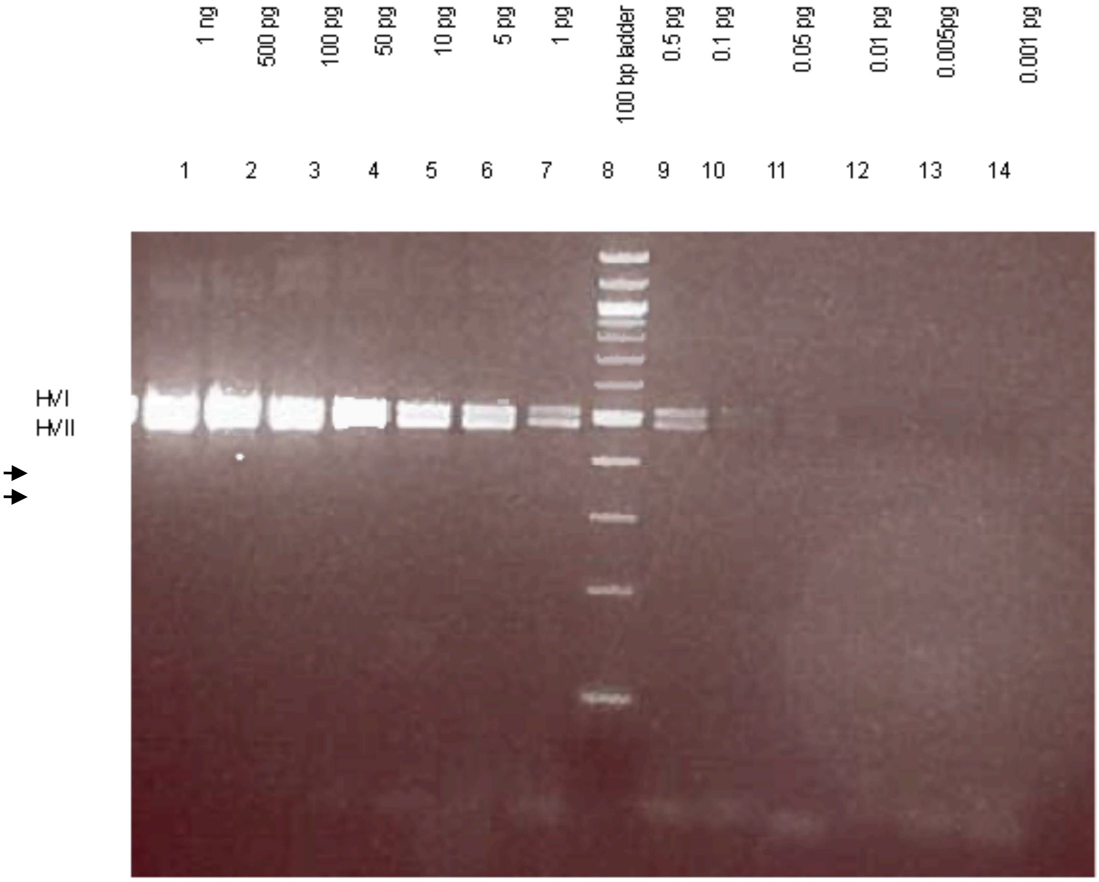
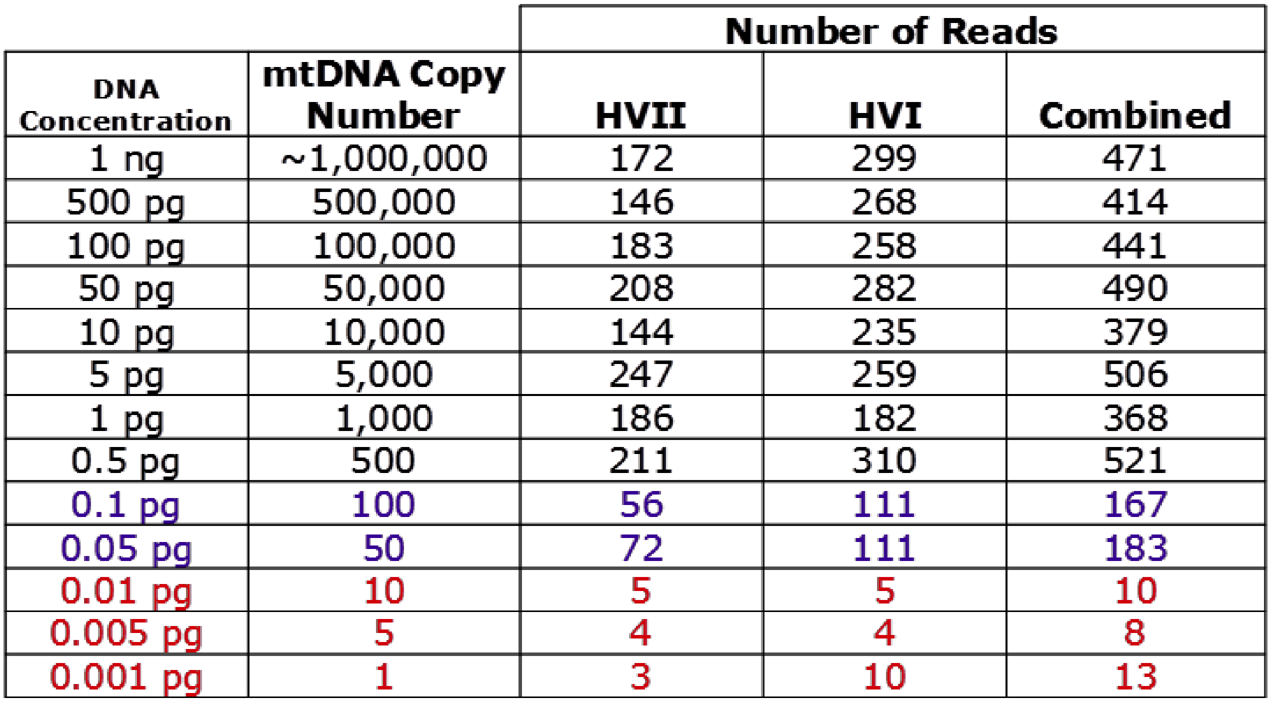


Figure 12. 3% Agarose Gel Electrophoresis of PCR Products with 120V, 60 minutes. Figure 1.1-9) Sample K562 amplified successfully at HVI/HVII regions with starting amount 1ng, 500 pg, 100 pg, 50 pg, 10 pg, 5 pg, 1 pg, (100 bp ladder), and 0.5 pg. Figure 1.10-11) Starting amount of 0.1 pg and 0.05pg show faint bands.

To further determine the sensitivity and stochastic effects of the PCR and sequencing, all thirteen libraries prepared from varying starting DNA sample amounts were included in one 454 sequencing run. During library preparation, 100 pg of a control DNA HL-60 was included as the positive control, and TE-4 was used for negative control to ensure that the reagents were not contaminated. To normalize the copy number added to the library pool, each of the 13 amplicon libraries were diluted from 1:6 to 1:500 based on QuBit DNA quantification values. The library pool was then further diluted ~2 fold to achieve the 0.3 molecule/bead target ratio for 454 sequencing. DNA amounts ranging from 0.5 pg- 1 ng resulted in an average of 261 combined reads ranging from 182-310 reads (Table 9). It was observed that the number of reads decreased two-fold with samples amplified with 0.1 pg and 0.05pg of starting material relative to higher concentration starting materials. It was also noted that stochastic effects of PCR were observed with DNA starting amounts below 0.05 pg or ~50 mtDNA copies and the sequencing results were not reliable due to the extremely low number reads observed, which were often unidirectional and for only a single HV region (Table 9). Based on these results, the current assay was determined to be sensitive down to 0.5 pg or ~500 copies. However, the limit of starting DNA amount of the assay could be possibly lowered to 0.1 pg or 0.05 pg with the elimination of the dilution process after the amplification of the low copy number samples along with increasing the amount of PCR product added to emPCR for sequencing for those specific samples.

Table 9. Number of Reads Obtained for HVI/HVII Regions for Sensitivity Study



		Number of Reads		
DNA Concentration	mtDNA Copy Number	HVII	HVI	Combined
1 ng	~1,000,000	172	299	471
500 pg	500,000	146	268	414
100 pg	100,000	183	258	441
50 pg	50,000	208	282	490
10 pg	10,000	144	235	379
5 pg	5,000	247	259	506
1 pg	1,000	186	182	368
0.5 pg	500	211	310	521
0.1 pg	100	56	111	167
0.05 pg	50	72	111	183
0.01 pg	10	5	5	10
0.005 pg	5	4	4	8
0.001 pg	1	3	10	13

2) Single Plex vs Multiplex

A comparison study between single-plex PCR and duplex PCR was completed to ensure that duplex amplification of samples does not affect the efficacy of each primer set. Amplification of a target region may drop out when multiplex PCR design is used due to competition amongst primers. A control DNA sample was amplified targeting HVI and HVII regions separately and together to compare the results. 3% agarose gel stained with ethidium bromide confirmed that each single plex primer only amplified target region and the duplex primer blend has amplified both regions successfully. The 454 sequencing data shows that separate amplification and the simultaneous amplification of HVI/HVII were comparable since a similar numbers of reads were observed in single-plex PCR and duplex PCR. The number of reads for each hypervariable region in single-plex PCR was approximately two fold higher relative to the number of HVI/HVII reads amplified in the duplex PCR; the result was as expected since two times the number of HVI or HVII copies were sequenced as equal copy numbers of the single-plex and duplex products were added to the library pool. It was also noted that HVI region consistently had higher number of reads compared to HVII region for this run, regardless of single-plex amplification or duplex amplification (Table 10).

Table 10. Number of Reads Obtained for HVI and HVII Regions in Single Plex PCR vs Duplex PCR.

	K562								
	HVI Single Amplification			**HVII Single Amplication**			**HVI/HVII Duplex Amplification**		
	Forward	**Reverse**	**Combined**	**Forward**	**Reverse**	**Combined**	**Forward**	**Reverse**	**Combined**
HVI	175	339	514	0	0	0	198	93	291
HVII	0	0	0	100	216	316	42	101	143

3) Population Studies

A 454 HVI/HVII sequence population database was generated by amplifying and sequencing 173 blood derived DNA samples previously collected from four population groups (45 of African

American, 43 of Caucasian, 40 Hispanic, and 45 Japanese). A control DNA sample of K562 was amplified with 100pg starting amount simultaneously with the population samples as positive control, and TE-4 buffer was used as the negative control. Sanger sequencing results for the HVI and HVII regions were available from a subset of these samples and were compared to the 454 sequencing results for concordance. A total of 119 HVI 454 sequences from 26 African American, 24 Caucasian, 27 Hispanic and 42 Japanese individuals and a total of 96 HVII 454 sequences from 21 African American, 21 Caucasian, 22 Hispanic, and 32 Japanese individuals were compared to Sanger sequencing results for concordance.

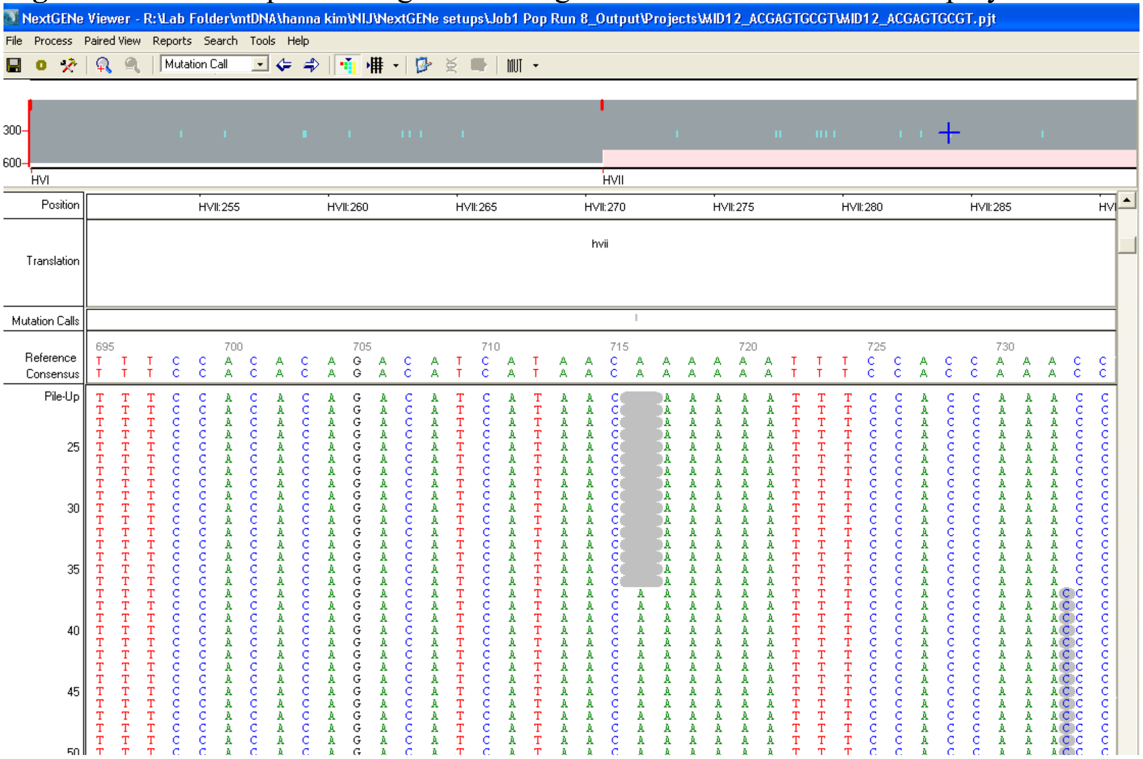
454 sequences generated by the 454 HVI/HVII assay showed concordance with the Sanger sequencing results at all base substitution positions (point mutations) when a 10% threshold was applied to the data analyzed using the 454 Amplicon Variant Analyzer (AVA) software. A 10% threshold was applied to account for the lower sensitivity of Sanger sequencing for minor base detection (10-15%) compared to 454 sequencing (1-3%). Minor base differences were observed at very low frequency below this 10% threshold with 454 sequencing. Data analysis is ongoing to fully characterize the frequency and extent of these putative minor base mutations.

However, as anticipated, 454 pyro-sequencing errors were observed as insertion/deletions (in/del) above the 10% threshold within the sequence surrounding the two homopolymer C stretches in the HVI/HVII regions which resulted in sequence alignment errors using the AVA software. Typically, these in/del sequencing errors were observed in only a single direction and/or were imbalanced. Specifically, pyro-sequencing errors and/or length heteroplasmy in the HVI region spanning base positions 16180 – 16195 (5’-AAAACCCCCTCCCCAT-3’) resulted in alignment issues using the AVA software. Specifically, an A/C mutation at position 16183 and any C/T transition observed between base positions 16184 to 16193 of this region was reported as a combination of in/del of A/C or C/T, with imbalanced frequencies between read directions. The SNPs in the homopolymer region were detected, but were not reported correctly due to alignment issues from homopolymer pyro-sequencing errors or potentially length heteroplasmy using the 454 AVA software.

Additionally, pyro-sequencing errors and/or length heteroplasmy within the homopolymer region surrounding the C-stretch in the HVII region spanning positions 286 – 315 (5’-AAAAAATTTCCACCAAACCCCCCCTCCCCC-3’) were observed. Specifically within this region, an A deletion was reported at base position 291 and position 295 was reported as a mixture of C/T followed by C deletion both with directionally imbalanced frequencies using 454 AVA software. These observations can be partially explained as a sequence alignment issue for the specific regions with homopolymers within the HVI/HVII sequences using the 454 sequencing platform as it is established that pyro-sequencing has a high error rate in homopolymer regions (Luo et al, 2012).

These sequence misalignment issues observed in the homopolymer stretches may potentially be resolved with better sequence alignment tools. For example, directionally imbalanced base differences can be filtered out as sequencing errors when certain filter thresholds are applied and are not reported as mutations using the NextGENe software by Soft Genetics, LLC. (Figure 13). This figure 13 shows an example of an A deletion at position 291 resulting from a pyro-sequencing error observed in a subset of the reads with a higher frequency observed in the forward direction as well as a C insertion in the poly-C stretch which were filtered out as sequencing error and not called a base mutations. Similarly, the C insertions resulting from pyrosequencing error or length heteroplasmy in the HVI homopolymer C stretch were aligned properly and were not called as mutations as they were filtered out as errors by the software; the C to T transition at position 16189 within the HVI homopolymer C stretch was properly aligned and the mutation was called using the NextGENe software (See Figure 14).

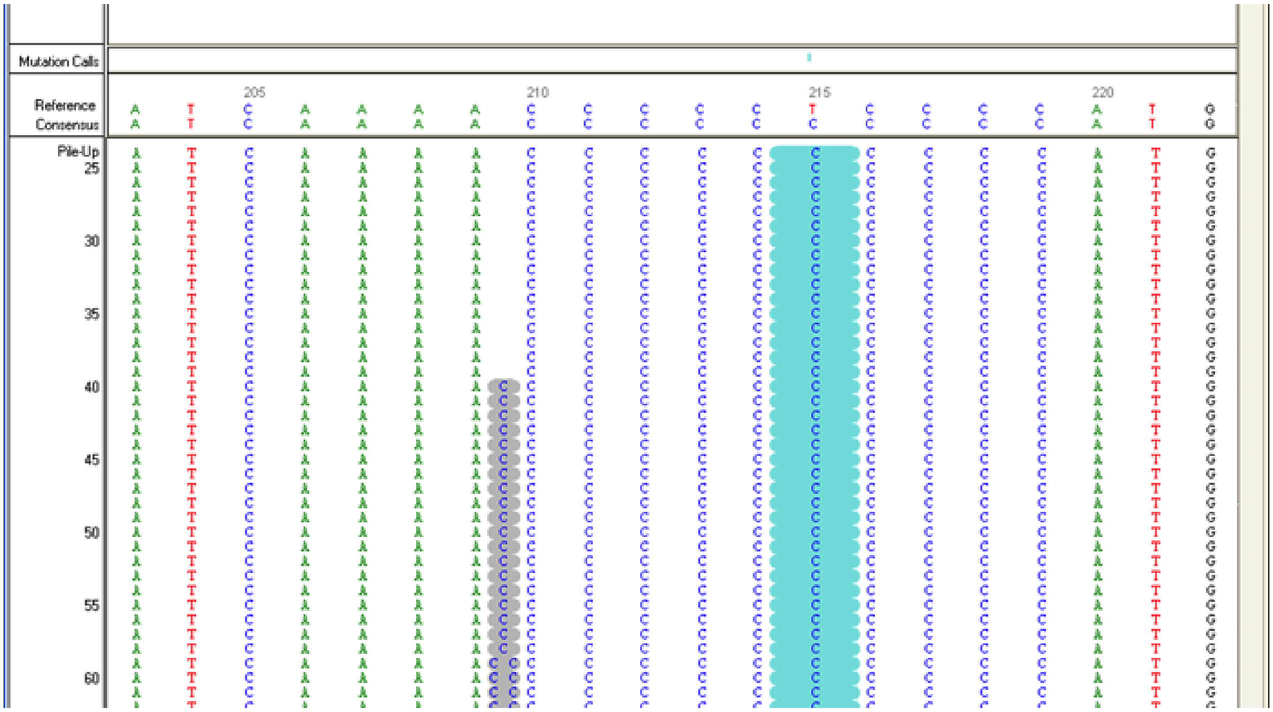
Figure 13. HVII Sequence Alignment using NextGENe within the Homopolymer C Stretch


NextGENe Viewer - R:\Lab Folder\mtDNA\hanna kim\NIJ\NextGENe setups\Job1 Pop Run 8_Output\Projects\MID12_ACGAGTGCGT\MID12_ACGAGTGCGT.pjt

Reference: T T T C C A C A C A G A C A T C A T A A C A A A A A A T T T C C A C C A A A C C

Consensus: T T T C C A C A C A G A C A T C A T A A C A A A A A A T T T C C A C C A A A C C

Figure 14. HVI Sequence Alignment using NextGENe within the Homopolymer C Stretch



Reference: A T C A A A A C C C C C T C C C C A T G

Consensus: A T C A A A A C C C C C C C C C C A T G

4) Interlaboratory Validation and Concordance Study at California Department of Justice

Validation at the CA DOJ

The California Department of Justice Bashinski DNA Laboratory in Richmond is in the process of validating the CHORI duplex HVI-HVII mitochondrial NGS assay on the Roche 454 GS Jr instrument platform. Implementation of the assay was planned for the Missing Persons DNA Program (MPDP), where it would first be used for the sequencing of reference samples, and subsequently for both MPDP reference and evidence samples. It was also planned for the assay to be used in the Casework DNA Analysis Program, where it would supplement current procedures for obtaining DNA information from low-yield samples, particularly from hair-shaft extracts. However, with the recently announced discontinuation of the 454 sequencing platform in 2016, the planned implementation is currently under discussion. Currently, HVI-HVII sequencing at CA DOJ laboratory is performed using the "standard" Sanger method. The move to NGS methods, particularly the ability to pool samples for processing, is expected to provide significant gains in sequencing efficiency and throughput.

To date, scientists at the CA DOJ have received NGS training both from Roche and from CHORI staff, and are in the initial stages of the validation studies. The Roche training focused on the basics of the 454 sequencing process, with hands-on training for emulsion PCR, bead enrichment, sequencing set-up, and basic analysis procedures. The CHORI training included several days of hands-on work that focused on library preparation procedures: (i) to quantify the DNA; (ii) to use the CHORI duplex PCR to label the HVI-HVII sequences with 454 adaptors and also to uniquely "barcode" each sample with index sequences; (iii) to quantify the library samples for pooling; and (iv) to purify, quantify, and quality check the pooled library samples for subsequent emulsion PCR and sequencing. In addition, the CHORI staff provided assistance with data analysis. Also, CHORI staff provided training to 10 CA DOJ criminalists over the course of eight days through lecture and hands on laboratory training covering the entire 454 HVI/HVII amplicon library preparation through 454 sequencing and data analysis.

5) Concordance and Reproducibility between Facilities

A subset of population samples (15African American, 16 Caucasian, 16 Hispanic, and 15 Japanese) from the above 454 HVI/HVII population database were used for the concordance and reproducibility study as part of the validation studies. The pooled, purified, and quantified library pool prepared and sequenced in CHORI laboratory was provided to California Department of Justice Jan Bashinski DNA Laboratory (Cal DOJ) to be sequenced in their facility. Additionally, the genomic DNA samples of the same library were also provided to the Cal DOJ to be prepared as libraries and sequenced, following the same protocols for the optimized assay, in their facility by the CalDOJ staff to confirm the reproducibility of the assay. The 454 sequence data was concordant between the two laboratories and both 454 datasets were concordant with Sanger sequencing results for the HVI and HVII regions outside of homopolymer stretch regions as identified and discussed above.

Internal validation studies are on-going at the CA DOJ and will be guided by the FBI Quality Assurance Guidelines to investigate contamination, sensitivity, mixtures, and reference and case-type samples. Through these studies, the CA DOJ will continue to rely on expertise and training from CHORI staff as they develop protocols for data analysis and reporting, and as they develop new methods for sequencing challenging (e.g., degraded) samples.

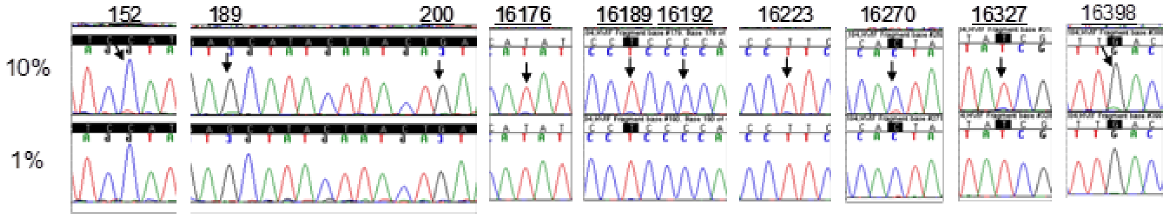
6) Mixture Studies

a) Mixture Sensitivity Study with 2 Contributors

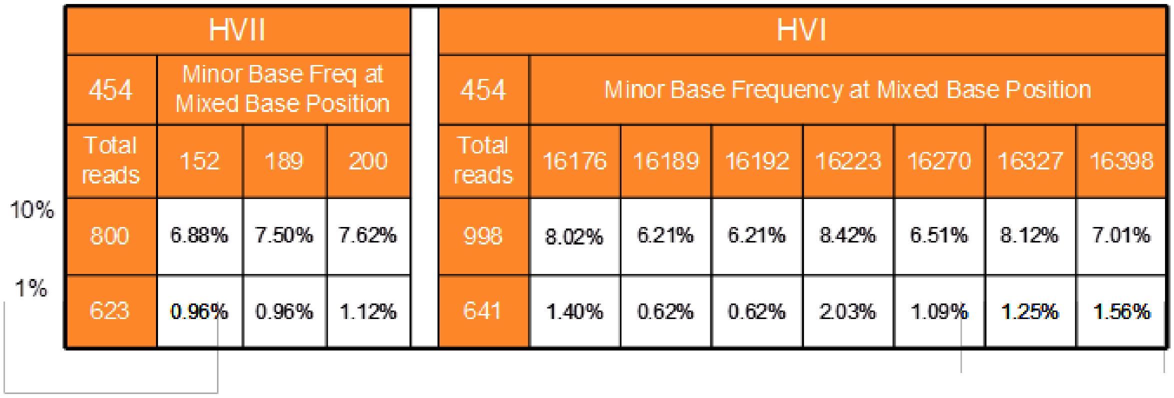
We also conducted a mixture study to determine the sensitivity of the mtDNA HVI/HVII duplex 454 sequencing assay for detecting minor components in a mixture. Two DNA samples were mixed together to give the following ratios based on mtDNA copy number determined by qPCR: 100:0, 99.5:0.5, 99:1, 95:5, 90:10, 75:25, 50:50, 25:75, 10:90, 5:95, 1:99, 0.5:99.5, 0:100. All mixed base positions were identified and minor components down to the tested 0.5% were detected. Sanger and 454 sequencing results for two of the mixed samples (10% and 1%) are shown in the Figure 15 and table 11 below. As expected mixed base positions of the minor sequences at 10% and 1% were not detectable by Sanger sequencing but were detected by 454 sequencing. The minor sequence for the mixed base positions ranged from 0.96% - 2.03% for the 1% sample mixture with an average of 1.01% for HVII and 1.47% for HVI. The minor sequence for the 10% sample mixture ranged from 6.88%-8.42% at the mixed base positions with an average of 7.33% and 7.62% for the HVII and HVI regions respectively. The 10% minor sequence was detected at a ~2.5% lower frequency than expected, but this slight difference may be contributed to error in quantifying the mtDNA copy number or the initial mixing of the sample. These results also show that the 454 fusion primers do not interfere with Sanger sequencing quality and that the products can be sequenced by both methods.

Figure 15. Mixture Study 10% and 1%: Sanger vs 454 Sequencing

Sanger Sequencing: Detection of minor sequence >15-20%



454 Next-Generation Sequencing: Detection of minor sequence 1% with 600-1000 reads



	HVII				HVI							
	454	Minor Base Freq at Mixed Base Position			454	Minor Base Frequency at Mixed Base Position						
	Total reads	152	189	200	Total reads	16176	16189	16192	16223	16270	16327	16398
10%	800	6.88%	7.50%	7.62%	998	8.02%	6.21%	6.21%	8.42%	6.51%	8.12%	7.01%
1%	623	0.96%	0.96%	1.12%	641	1.40%	0.62%	0.62%	2.03%	1.09%	1.25%	1.56%

In addition, a more extensive two person mixture study was conducted to determine the limit of sensitivity for detection of minor sequences over a larger range of mixture ratios. Nuclear and mtDNA copy number of two DNA samples (Corielle 110 and Corielle 073) were quantified using a mt:nu qPCR assay (Timken et al.2005). Based on the mtDNA copy number estimation resulted from the qPCR assay, the two samples were mixed at the following 19 different mixture ratios (Table 11). The mixture samples were then amplified using the optimized 454 HVI/HVII PCR assay and sequencing using the 454 GS Jr.

Table 11. Mixture Ratio of Artificially Mixed Samples with 2 Contributors.

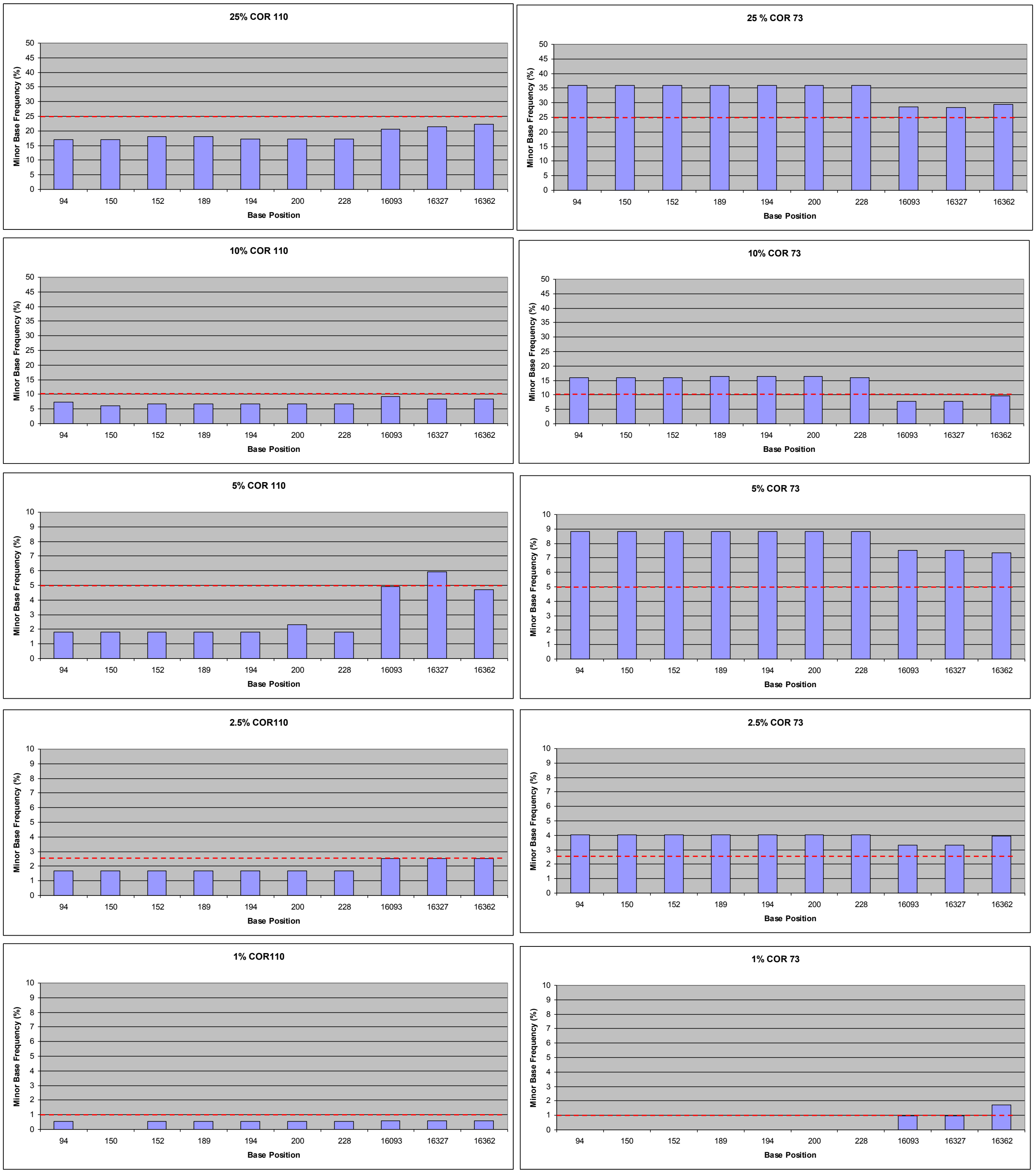
Corielle 110	Corielle 073
Mixture %	**Mixture %**
100	0
99.9	0.1
99.75	0.25
99.5	0.5
99	1
97.5	2.5
95	5
90	10
75	25
50	50
25	75
10	90
5	95
2.5	97.5
1	99
0.5	99.5
0.25	99.75
0.1	99.9
0	100

The minor and major sequences were detected with confidence using 454 NGS sequencing at all mixed based positions for samples mixed at a 2.5% ratio and above. The observed frequency of the minor base for each of the mixed based positions as well as the number of 454 reads are presented for a subset of the mixture samples in Table 12 and illustrated in Figure 16 below. For the 1% mixture ratio with COR 110 as the minor sequence, the minor base was detected for all mixed based positions except position 150 in the HVII region with a 370 read depth. For the 1% mixture ratio with COR 073 as the minor sequence, it was observed that the minor bases at all mixed base positions were not detected in HVII but were observed in HVI. However, 274 reads were obtained for the HVII region compared to 516 reads for HVI region for this sample, and therefore a mixture with lower than 1% contributor would not likely be resolved with such low read coverage. It is expected that even lower than 1% minor base can be resolved by increasing the number of reads. To reliably detect the minor sequence in a 1% mixture, greater than 1000 read coverage would be ideal. An approximately 20-40% difference from the expected mixture ratio was observed in each experimental mixture sample, with the number of reads for one sample in the mixture (COR 110) consistently being under represented (Figure 16). Based on these results, the mt:nu qPCR most likely over estimated the mtDNA copy number of COR 110 and thus the actual mixture ratio could have been skewed from the expected mixture ratio.

Table 12. Frequency of Minor Base in 25%-1% Mixture Samples with a) Corielle 110 as the Minor Contributor and b) Corielle 073 as the Minor Contributor.

	HVII								**HVI**			
	Observed Frequency of Minor Base (%)								**Observed Minor Base Freq (%)**			
Base Position	94	150	152	189	194	200	228		16093	16327	16362	
COR 73 Sequence	G	T	C	G	C	G	G	**# of Reads**	T	T	T	**# of Reads**
25% COR073	35.97	35.97	35.97	35.97	35.97	35.97	35.97	303	28.47	28.28	29.48	583
10% COR073	15.87	15.87	15.87	16.35	16.35	16.35	15.87	416	7.69	7.81	9.7	845
5% COR073	8.82	8.82	8.82	8.82	8.82	8.82	8.82	238	7.54	7.54	7.35	517
2.5% COR073	4.04	4.04	4.04	4.04	4.04	4.04	4.04	223	3.31	3.31	3.93	484
1 % COR073	0	0	0	0	0	0	0	274	0.97	0.97	1.74	516

Figure 16. Frequency of Minor Contributor Detected per Base Position.



25% COR 110

25 % COR 73

10% COR 110

10% COR 73

5% COR 110

5% COR 73

2.5% COR110

2.5% COR 73

1% COR110

1% COR 73

Figure 16. Frequency of Minor Contributor Detected per Base Position. Red dashed line indicates expected frequency.

b) Complex Mixture Study

We conducted more extensive mixture studies to test the 454 HVI/HVII assay's capability to detect multiple sequences in a sample during this reporting period by studying mixture samples composed of 3, 4, and 5 contributors with varying compositions. Individuals from different population groups (Caucasian, African American, and Hispanic) were selected to prepare the mixture samples. Four samples of each mixture (3 individuals, 4 individuals, and 5 individuals) were formulated at different ratios prior to HVI/HVII amplification. The HVI/HVII amplicon libraries of the complex mixture samples were prepared using the final validated assay and sequenced using 454 GS Jr. Preliminary results show that multiple sequence haplotypes were observed for each of the complex mixture samples. NGS sequence alignment and sequence analysis is currently underway to fully characterize all the observed sequence haplotypes. Preliminary evidence of jumping PCR was observed in each of the complex mixtures as chimeric sequences were present at low levels. A table summarizing the preliminary sequence results including the expected and observed frequency of each contributor sequence as well as jumping PCR artifacts for a three person mixture is presented below (see Table 13).

Table 13. Complex Mixture Composed of Three Contributors and Frequencies of Individual Sequences Detected Using 454 Sequencing.

			Sequence					
Contributor	Expected%	Observed%	73	93	150	189	195	263
Contributor 1	67.8	49.9	A	G	C	A	T	G
Contributor 2	21.7	36.3	G	A	C	C	C	G
Contributor 3	10.5	7.5	G	A	T	A	C	G
Contributor 2/1	-	2	G	A	C	A	T	G
Contributor 1/2	-	1.7	A	G	C	C	C	G
Contributor 1/3	-	0.9	A	G	T	A	C	G
Contributor 1/2	-	0.9	A	A	C	C	C	G
Contributor1/3	-	0.4	A	G	T	A	C	G
Contributor 2/1	-	0.4	G	G	C	A	T	G

7) Jumping PCR Studies

Jumping PCR, a phenomenon where partially amplified single strand of DNA act as a primer to generate an amplicon with chimeric sequences (Paabo et al. 1989), can occur during PCR typically during late PCR cycles when the number of DNA copies is very high. This event may be problematic for the forensic application of the assay if not properly recognized and considered for the analysis, especially with mixed samples, as this may lead to ambiguous results if observed at a high frequency. In order to explore the effect of jumping PCR of mtDNA HVI/HVII duplex 454 sequencing assay, 100 pg of a 50:50 mixture of two DNA samples, Corielle 037 and Corielle 110, were amplified at varying cycle numbers: 34 cycles, 32 cycles, 30 cycles, 26 cycles, and 24 cycles. Each amplification included a negative control with TE-4 buffer. Higher number of amplification cycles resulted in an increase in number and frequency of chimeric sequences resulting from jumping PCR. The 50:50 mixed sample amplified with 34 cycles resulted in 10 different chimeric sequences with the total frequency of 15.41%. At 24 cycles, one chimeric sequence was detected with frequency of 0.62% (Table 13). A secondary sensitivity study was conducted using 24 amplification cycles. As expected, results based on the agarose gel electrophoresis show that a much higher starting DNA amount (~500 pg) would be required with 24 amplification cycles, which may be too high for forensic applications. However, lowering the cycle number to 30 or 28 cycles from 34 also significantly lowered the frequency and number of jumping PCR events and could be considered if DNA amounts allow for repeat amplification when mixtures are encountered. Additionally, further improvements to NGS software are being explored to aid with mixture analysis, including adding filters to flag or remove jumping PCR chimeras.

Table 13.

A. 24 Cycles

		Base Position														
Frequency		73	89	93	94	146	150	152	182	194	195	198	228	263	266	325
54.18	A	G	C	G	G	C	T	C	T	C	C	T	G	G	C	T
45.2	B	G	T	A	A	T	C	T	C	T	T	C	A	G	T	C
0.62	A/B	G	T	A	A	T	C	T	C	T	T	C	A	G	T	T

B. 34 Cycles

		Base Position														
Frequency		73	89	93	94	146	150	152	182	194	195	198	228	263	266	325
47.21	A	G	C	G	G	C	T	C	T	C	C	T	G	G	C	T
37.41	B	G	T	A	A	T	C	T	C	T	T	C	A	G	T	C
1.56	A/B	G	C	G	G	T	C	T	C	T	T	C	A	G	T	C
1.34	A/B	G	C	G	G	C	T	C	T	C	C	T	G	G	C	C
0.9	A/B	G	C	G	G	C	T	C	T	C	C	T	A	G	T	C
0.45	A/B	G	C	G	G	C	T	C	T	C	C	T	A	G	T	T
1.56	A/B	G	T	A	A	C	T	C	T	C	C	T	G	G	C	T
0.45	A/B	G	T	A	A	C	T	C	T	C	C	T	G	G	T	C
0.9	A/B	G	T	A	A	T	C	T	T	C	C	T	G	G	C	T
0.45	A/B	G	T	A	A	T	C	T	C	T	T	T	G	G	C	T
5.57	A/B	G	T	A	A	T	C	T	C	T	T	C	A	G	T	T
2.23	A/B	G	T	A	A	T	C	T	C	T	T	C	A	G	C	T

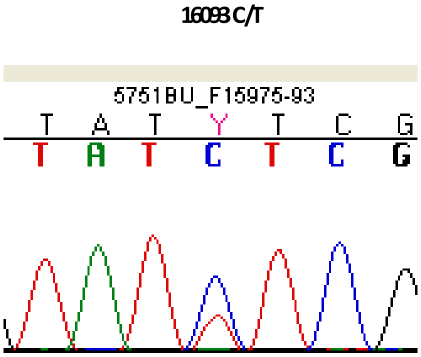
Table 13. Jumping PCR Effect in A.) 24 Cycle Amplification and B.) 34 Cycle Amplification.

8) Heteroplasmy Studies

To determine the detection limits of the 454 HVI/HVII assay for detection of heteroplasmy, we sequenced a subset of buccal, blood and hair samples (23 samples) from twins previously identified as heteroplasmic in at least one tissue (REF). Sanger sequencing and 454 sequencing results from a buccal sample which was heteroplasmic at position 16093 are shown in the Figure

17 below. For this sample, ~65% of the sequences showed a C at position 16093 in the forward direction and ~70% in the reverse direction or a combined average of ~67.8% corresponding to the C >T at this position in the Sanger sequencing electropherogram. The 454 NGS sequencing method provides a digital readout and is more quantitative than Sanger sequencing.

Figure 17.

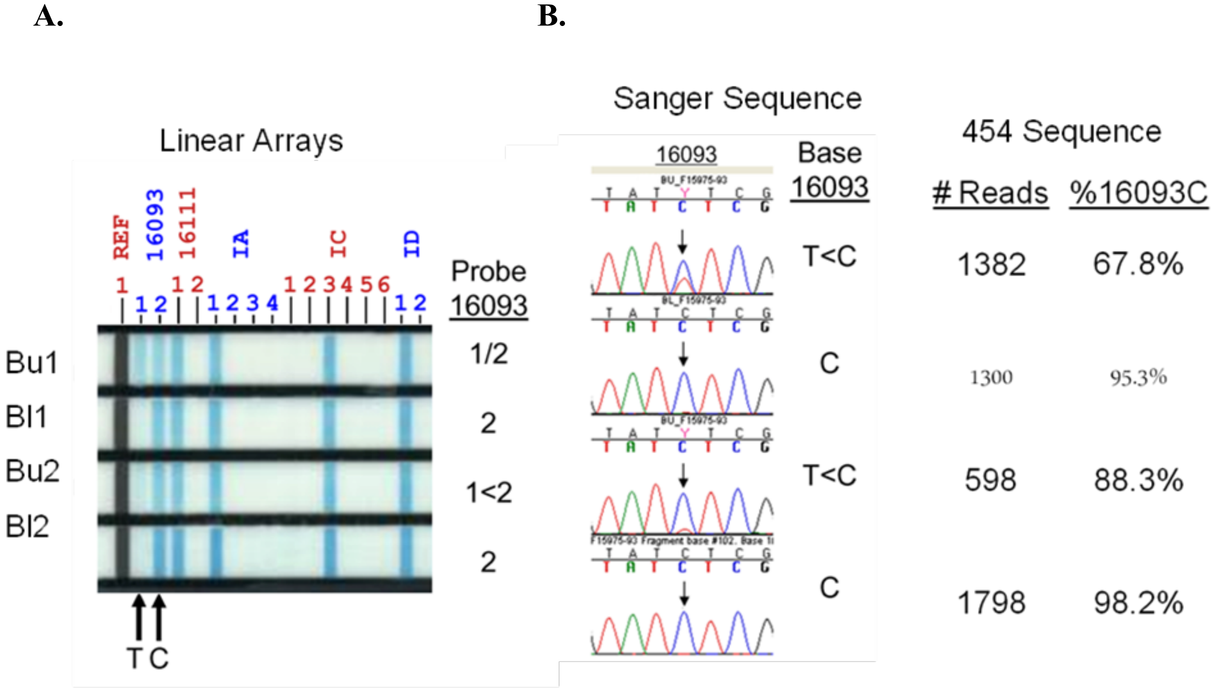


	% Variant Sequence by 454 Sequencing		
	Forward	Reverse	Total
16093C	65.62%	70.25%	67.80%

We also sequenced using 454 a subset of samples from the heteroplasmic twins where heteroplasmy was not detected by probe analysis or Sanger sequencing. Since the 454 sequencing method uses clonal amplification and sequencing of each individual molecule, this method of sequencing is expected to be more sensitive for detecting minor components in a mixture or rare sequence mutations. The sensitivity is dependent on the read depth; the greater the read depth, the greater chance at detecting minor components. In Figure 18 (A-C) below, heteroplasmy was detected by 454 sequencing but not by Sanger sequencing or by probe analysis in the blood sample and in two hairs above 1%. In panel A and B, heteroplasmy was not detected in the blood sample of dizygotic (DZ) twin two but was detected at ~2% level by 454 sequencing.

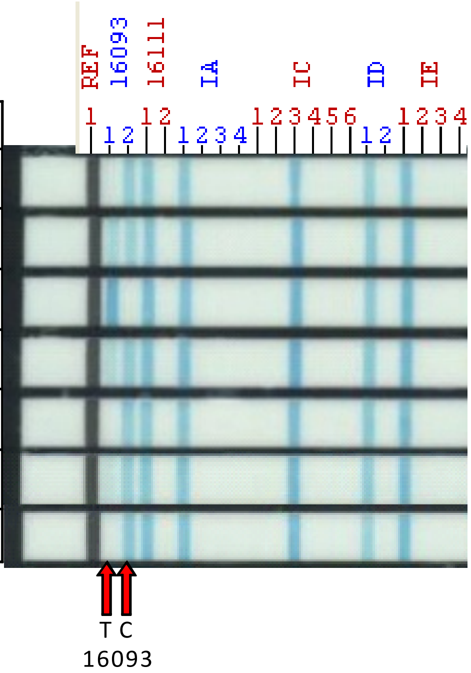
In addition, heteroplasmy was not detected in three of the five hairs by probe analysis but was detected at very low levels in two of the three hairs (3.5% and 1.5%) in both directions and in one hair in the forward direction (0.7%). Heteroplasmy in hair three could not be confirmed as sequences corresponding to a 16093C were only obtained in the forward direction.

Figure 18.



www.454.com

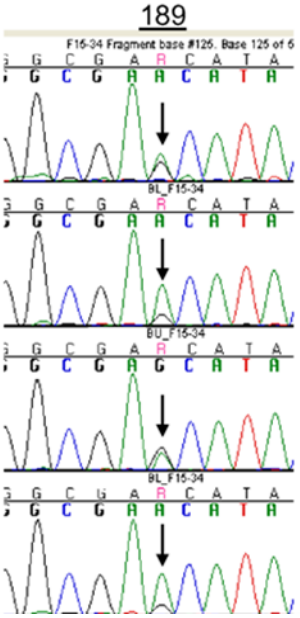
C.



		% C at 16093 Detected using 454		
DZ Twin 2	16093	Forward	Reverse	Combined
Hair (#1)	C	96.76	96.9	96.83
Hair (#2)	T < C	-	-	-
Hair (#3)	T	0.76	0	0.38
Hair (#4)	T < C	-	-	-
Hair (#5)	C	99.11	97.98	98.545
Buccal	T < C	86.51	90.14	88.325
Blood	C	98.04	98.43	98.235

As previously observed with Linear Array and Sanger sequencing analysis, the frequency of heteroplasmy differed across tissues. Figure 19 and 20 below shows varying levels of heteroplasmy detected at position 189 differing between twin sibs as well as tissues. In MZ twin 1 Hair 3 only an A was detected at position 189 while in MZ twin 2 Hair 4 primarily G was detected (99.7%). The 454 NGS assay provides a digital readout of the number of reads and the percent heteroplasmy at each position can be determined whereas Sanger sequencing is not quantitative. The 454 NGS method is also more sensitive for detecting heteroplasmy or mixtures because of the clonal nature of the amplification.

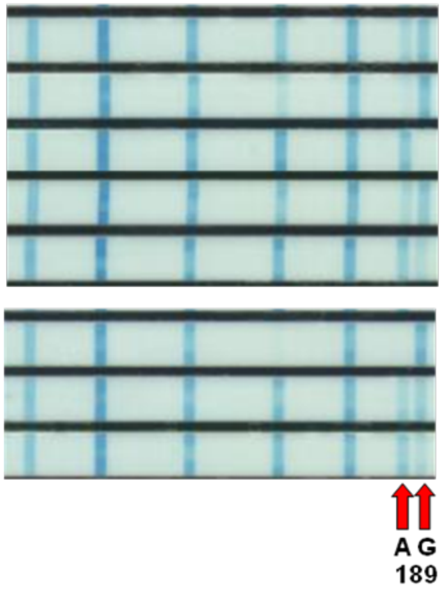
Figure 19.



	Sanger Sequencing Base 189	454 Sequencing # Reads	189 %G
MZ Twin 1 Buccal	A/G	915	42.8
MZ Twin 1 Blood	A>G	1144	19.9
MZ Twin 2 Buccal	A/G	961	61.5
MZ Twin 2 Blood	A>G	918	14.8

Figure 20.

	Tissue	189
MZ Twin 1	Hair (1)	A ≈ G
	Hair (2)	A < G
	Hair (3)	A
	Buccal	A < G
	Blood	A > G
MZ Twin 2	Hair (4)	G
	Buccal	A ≈ G
	Blood	A > G



454 Sequence	
% 189G	Total # Reads
53.27	511
84.87	1159
0	921
61.48	915
14.84	1144
99.77	849
42.81	961
19.90	918
Avg Reads	922

To further determine the sensitivity of the HVI/HVII 454 assay for detection of low level heteroplasmy, the study was expanded to include a blood, buccal, and 5 hair samples from

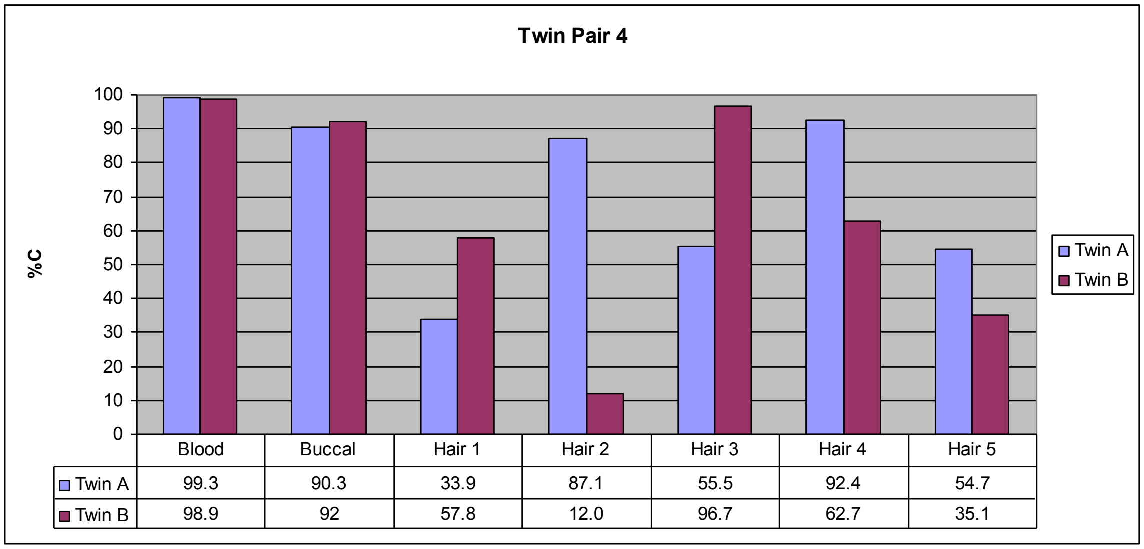
5 twin pairs previously shown to exhibit heteroplasmy at the 16093 position. The total number of reads as well as the percentage of the C base at position 16093 for the samples completed to date is presented in Table 4 below. As previously observed with Linear Array and Sanger sequencing analysis, the frequency of heteroplasmy differed across tissues. However, using the 454 NGS method which is more quantitative and sensitive, heteroplasmy was detected in samples not previously observed using the less sensitive methods. Heteroplasmy was generally not detected below 10% for the other methods while it was detected as low as ~1% using the 454 method. Also, since the 454 NGS technology allows for 'counting' of reads and a digital read out of the proportion of sequence bases at a position, differences between the level of the minor base sequences at a position are now detectable. The level of heteroplasmy was significantly lower in blood compared to buccal samples and was highly variable in hair samples. Figure 10 is shown below to further illustrate the differences in the level of heteroplasmy across tissues within two twin pairs.

Table 14. Frequency of Major Base at Position 16093 Across Different Tissues in Twins

		Tissue Type	Total # reads	%C
Twin Pair 1	Twin A	Hair 1	668	99.1
		Hair 2	710	5.4
		Hair 3	564	98.4
		Hair 4	328	58.5
		Hair 5	591	83.8
		Buccal	770	96.5
	TwinB	Hair 1	341	52.2
		Hair 2	620	84.4
		Hair 3	401	64.8
		Hair 4	483	14.9
		Hair 5	550	99.6
		Buccal	587	82.3
Twin Pair 2	Twin A	Hair 1	284	46.8
		Hair 2	862	98.7
		Hair 3	637	54.6
		Hair 4	2711	37.8
		Hair 5	484	4.1
		Blood	442	98.2
		Buccal	566	62.7
	TwinB	Hair 1	573	76.6
		Hair 2	6161	30.4
		Hair 3	1296	5.9
		Hair 4	644	86.6
		Hair 5	594	93.6
Twin Pair 4	Twin A	Hair 1	496	33.9
		Hair 2	830	87.1
		Hair 3	532	55.5
		Hair 4	734	92.4
		Hair 5	691	54.7
		Blood	668	99.3
		Buccal	680	90.3
	TwinB	Hair 1	622	57.8
		Hair 2	841	12.0
		Hair 3	1084	96.7
		Hair 4	916	62.7
		Hair 5	485	35.1
		Blood	710	98.9
		Buccal	759	92
Twin Pair 5	Twin A	Hair 1	590	79.5
		Hair 2	440	100.0
		Hair 3	726	89.7
		Hair 4	317	1.3
		Hair 5	138	100.0
		Blood	991	97.3
		Buccal	952	94
	TwinB	Hair 1	831	81.9
		Hair 2	831	95.1
		Hair 3	647	69.2
		Hair 4	967	33.9
		Hair 5	767	42.8
		Blood	932	96.4
		Buccal	1105	93

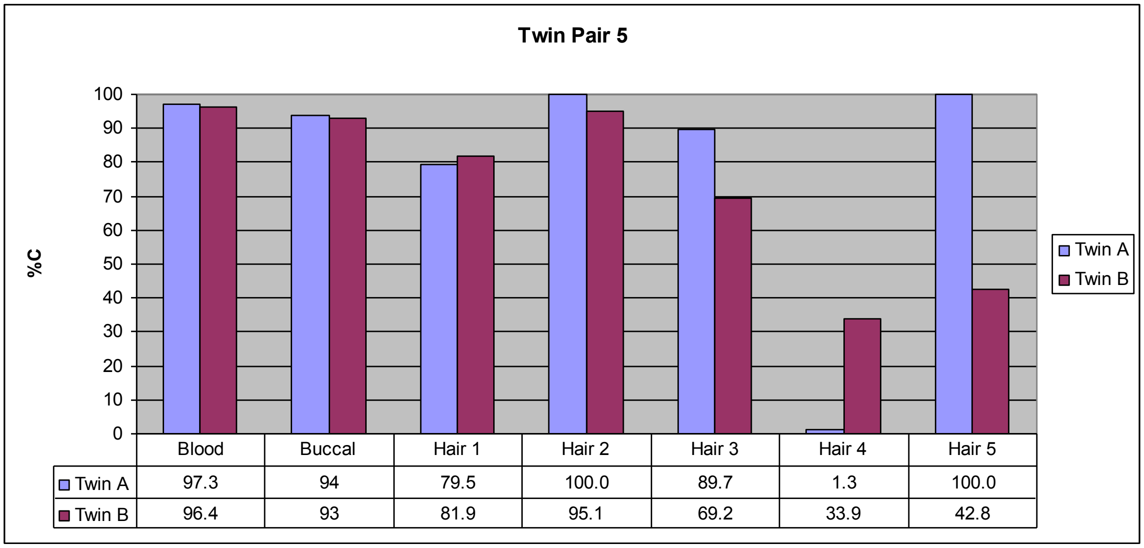
Figure 21. Frequency of Major Base on Position 16093 Across Different Tissues in Two Monozygotic Twin Pairs.

a.



	Blood	Buccal	Hair 1	Hair 2	Hair 3	Hair 4	Hair 5
■ Twin A	99.3	90.3	33.9	87.1	55.5	92.4	54.7
■ Twin B	98.9	92	57.8	12.0	96.7	62.7	35.1

b.



	Blood	Buccal	Hair 1	Hair 2	Hair 3	Hair 4	Hair 5
■ Twin A	97.3	94	79.5	100.0	89.7	1.3	100.0
■ Twin B	96.4	93	81.9	95.1	69.2	33.9	42.8

2. Design, Methods, and Results of Probe Capture and 454 Sequencing Assay for Whole Mitochondrial Genome Sequencing

We initially proposed to extend this PCR based assay to include the rest of the non-coding control region (2 PCR targets) and 10-15 100- 350 bp coding regions representing ~20% of the mtDNA genome. While this multiplex PCR method for enriching mtDNA would likely be successful, recently probe based enrichment methods for capture and sequencing the entire mtDNA genome have drastically decreased in price making this a feasible alternative. Probe capture for enriching for mtDNA would allow for full mtDNA sequencing as well as the potential for capture of degraded DNA. As the probe capture method could allow for capture and sequencing of the entire mtDNA genome, we explored this as an alternative to amplifying select regions of the mtDNA genome described previously as part of our aim to developing a mtDNA 'front end' assay. Several probe capture methods were explored and the solution-phase Nimblegen SeqCap EZ platform was chosen due to their extensive tiling design (Figure 22) and ability to efficiently synthesize hundreds of thousands of unique probes for a specific target sequence of interest. The probe design strategy and optimization of the probe capture system are described below.

i) Design: Sequence Probe Capture Design

The Nimblegen SeqCap EZ Library capture probes are typically designed by NimbleGen using proprietary software targeting a given reference DNA sequence provided by the user. Nimblegen employs a tiling approach to probe design which results in a high redundancy of unique probes for enrichment of the target sequence of interest which increases capture efficiency compared to alternate methods which use a simple probe design strategy with non-overlapping probes (Figure 22). In general, each probe is allowed to vary between 50-100bp in size, while keeping the GC content and melting temperature similar for all probes. To prevent capture of off target DNA, the probes are constrained to the amount of homology to the h19 Human nuclear reference sequence using the standard parameters of their proprietary algorithm. The probe design process is proprietary to Roche Nimblegen, however, we were able to work closely with Nimblegen to help guide the design of our probes. The parameters specific to our designs are described below.

Figure 22. Nimblegen Tiling Design

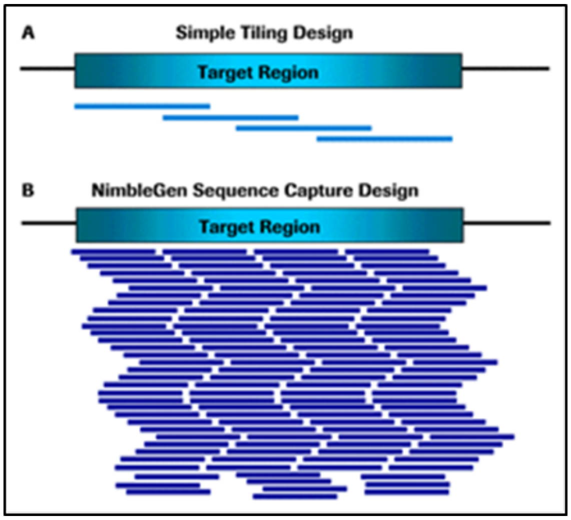


Figure 22. Nimblegen Tiling Design of Capture Probes. A) A simple tiling design of probes with single overlaps. B) The Nimblegen design for extensive tiling of probes over the target region. Figure courtesy of Roche Nimblegen.

For each of our probe designs, a majority mtDNA consensus was used as our reference sequence. A majority mtDNA consensus sequence was used rather than simply the revised Cambridge Reference Sequence (rCRS)[91], since the rCRS has several minority frequency mutations at several base positions. The majority base consensus reference sequence was constructed using the majority base frequency from the over 2000 sequences from a global mtDB-Human Mitochondrial Genome population Database[1]. In addition to a single majority base consensus reference sequence, we considered using multiple mtDNA reference sequences targeting the most common sequence haplotypes to ensure capture of samples with multiple sequence differences compared to the reference sequence. The Nimblegen hybridization conditions allow for capture of target sequences differing at up to five base positions from the ~75 base probe. To determine the density and distribution of polymorphisms within the mitochondrial genome, the global mtDB-Human Mitochondrial Genome Database was surveyed to determine if multiple references sequences would be necessary (Figure 23). While a high density of polymorphism is found in the mitochondrial genome, particularly in the non-coding region, it was determined that sequence haplotypes typically did not differ from the majority reference sequence by more than five bases within a 75 base region. Further, due to the high redundancy tiling approach used by Nimblegen for probe design, it was predicted that most if not all sequence haplotypes would be captured using the single majority mtDNA consensus sequence as the reference.

Figure 23. Frequency of Most Abundant Polymorphism at Each Position in the Mitochondrial Genome

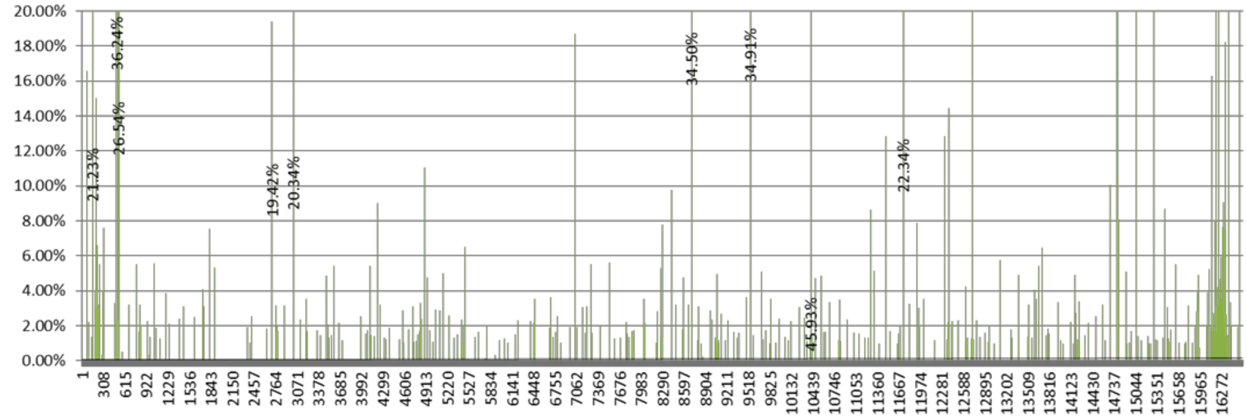


Figure 23. Frequency of Most Abundant Polymorphism at Each Position in the Mitochondrial Genome. Frequencies are from mtDB database.

The first probe design was created using the standard settings determined by Nimblegen, and is shown in Panel A of Figure 24 below. Nimblegen uses a proprietary algorithm to design probes for your target region including removing probes with homology to other regions of the genome using the h19 human reference genome. Design one covers 97% of the mitochondrial genome (green) with 14 different regions did not have unique probes producing gaps (noted as white gaps Figure 24 panel A). Due to the length of the probes Nimblegen assumes regions within 100bp from the ends of unique probes will still be captured, giving us a theoretical 100% probe coverage. However, we were concerned that it did not take into consideration any of the unique features of the mitochondrial genome, in particular its circular structure as well as nuclear pseudogenes. Nuclear pseudogenes are portions of the mitochondrial genome which have been copied into the nuclear genome. Subsequent designs took into account these unique features of the mitochondrial genome.

For the second design, we reduced the tiling distance between probes in an attempt increase the coverage of the mitochondrial genome with unique probes. Also, 50bps from the beginning of the consensus sequence were added to the end of the sequence in order to replicate the circular nature of the mitochondrial genome. The distribution of unique probes for design 2 is shown in Figure 24 Panel B below and has unique probes covering ~98.9% (green) of the mitochondrial genome. There were 183bps (1.1%) over seven regions with no unique probes (white gaps). Design 3 used a similar selection algorithm as for design 2 probe but used an algorithm to rebalance the probes in order to gain more coverage over the gaps. The rebalanced design is shown in Figure 24 Panel C. This design covered 99.8% of the mitochondrial genome and only 34bps (0.2%) were uncovered by unique probes. However, the overall number of unique probes across the mitochondrial genome was greatly decreased using the rebalancing algorithm and thus this design was rejected. The fourth and last design included adding a 100bp from the beginning of the consensus sequence to the end of the reference sequence to ensure full length probes to span the ends of our consensus sequence. We also slightly reduced the stringency of the allowable homology of the probes to the nuclear genome to increase coverage. To prevent capture of off target DNA, the probes are typically constrained to a set percent shared homology to the h19 Human nuclear reference sequence. For design 4, we took into consideration mtDNA nuclear pseudogenes by restricting the homology of our probes to the nuclear genome to ten sites or less (changed from the standard shared homology constraints). This final design directly targeted 99.999% of the mitochondrial genome with only two small gaps at positions 2,506-2,513 (7bp) and 2,962-2,972 (10bp) (Figure 24 Panel D). This design 4 was selected as the final probe design and was used to produce the Nimblegen SeqCap Ez Library capture probes used for testing.

Figure 24. Mitochondrial Capture Probe Design Iterations

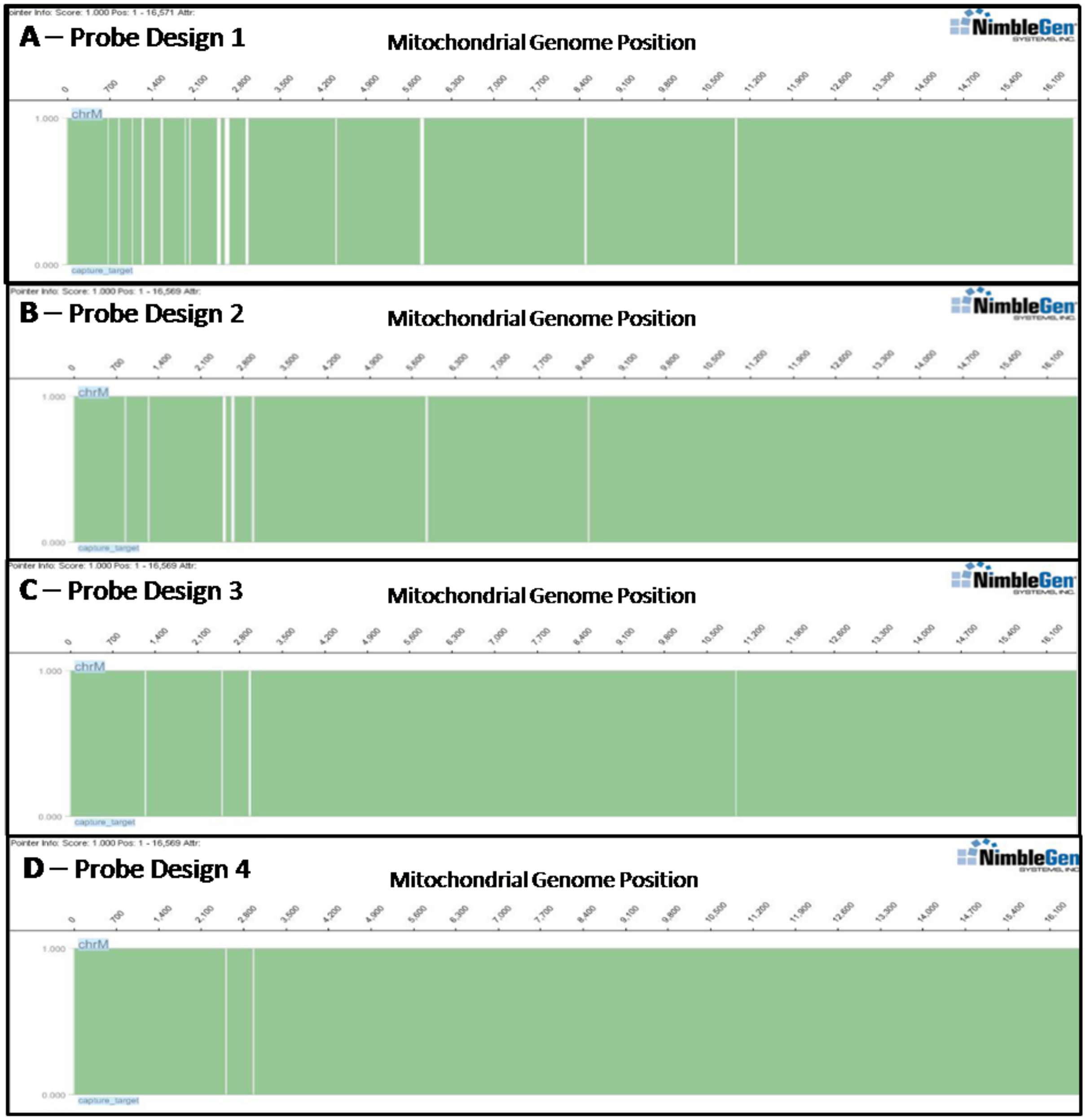


Figure 24. Unique Probe Maps Across a Consensus Mitochondrial Sequence. The probe graph shows the portions of the mitochondrial genome which are directly covered by unique probes in green and the gaps represent regions without any probes. Graphs provided by Roche Nimblegen.

ii. Methods

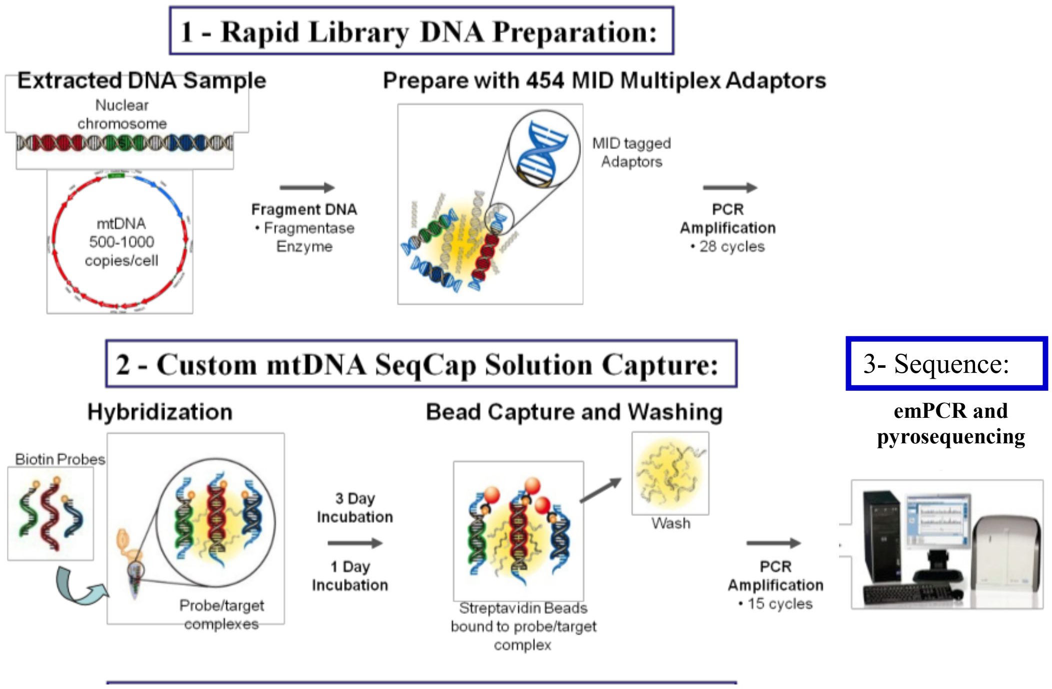
a. Methods Optimization

1) Rapid library Prep Optimization

Optimization of several of the steps in both the 454 sequencing and Nimblegen capture methods was necessary for the application to mitochondrial DNA. An overview of the methods is provided in the figure 25 below and a description of the optimization follows.

Figure 25.

Mitochondrial DNA Capture Method



2) Fragmentation Optimization

First, fragmentation of total genomic DNA to make 454 libraries was investigated by four different methods; mechanical shearing using sonication via a waterbath sonicator, nebulization, enzymatic digestion, and mechanical shearing using Covaris. The first fragmentation method we tested was sonication using a water bath sonicator to physically shear the DNA at varied time points to determine the optimum conditions to produce 454 libraries. This method was the least consistent method. Although we controlled the temperature to less than 25°C, and reduced the number of tubes within the sonicator to one at a time, we were not able to consistently produce fragmented DNA in the size range needed. We also tested nebulization for fragmentation which is the protocol recommended by the manufacturer. Nebulization also physically shears DNA by using nitrogen to aerosolize the DNA and force it through a small hole in a nebulizer to produce fragmented DNA. Following the 454 manufactured protocol, we were also able to produce fragmented DNA in the correct range for 454 libraries. However, since nebulization results in aerosolized DNA, this method was not a favorable option for forensic DNA samples.

3) Optimization of Conditions for DNA Fragmentation using Fragmentase

We also tested enzymatic digestion using the enzyme blend Fragmentase from New England Biolabs for DNA fragmentation. Fragmentase is a blend of two enzymes, one which makes random nicks in double stranded DNA, and the other which recognizes the breaks and cuts the second strand of DNA. The enzyme concentration and length of digestion was optimized to produce fragmented DNA in the size range of 500-1000bp which is recommended for 454 libraries. The protocol recommended by the manufacturer suggests a 15 minute incubation with a final fragmentase enzyme concentration of 0.025ug/ul for 1ug of DNA for DNA fragmentation between the recommended 454 fragment target size range of 600-800bp. As the end fragmentation size is dependent on the initial DNA concentration as well as enzyme concentration and incubation time, conditions using a final fragmentase enzyme concentration of 0.025ug/ul and 0.05ug/ul over a range of incubation times (10, 15, 30 and 60 minutes) for 500 ng of DNA (high DNA amount required for visualization on gel) were tested to determine the optimal enzyme concentration and incubation time for lower starting DNA amounts. Both final enzyme concentrations gave similar results across all incubation times, therefore the recommended 0.025ug/ul was selected for the final enzyme concentration. Incubation times below 30 minutes resulted in DNA fragment sizes above

1.5 kb and the 60 minute incubation resulted in fragments between 100-400 bp, while 30 minutes resulted in fragments between 100-800 bp which was within the 454 target size range. Based on these results, a final fragmentase enzyme concentration of 0.025ug/ul was tested over a range of incubation times (25, 30 and 35 minutes) for lower DNA inputs (100 ng, 1 ng, 100 pg) which were then amplified for 28 cycles and visualized by gel electrophoresis. Incubation for 30 minutes resulted in DNA fragments ranging from 100-800 bp with an ~500 bp average fragment size. The final fragmentase protocol is provided in Table X below and was used for DNA fragmentation for the initial proof of concept experiments.

Table 15. Final Fragmentase Protocol

Reaction Components:	Volume (ul)
DNA (500ng-100pg)	X
TE	(14-X)
10X Fragmentase Reaction Buffer	2
10X BSA	2
dsDNA Fragmentase	2
Final Volume	20

Table 15. Final Fragmentase Protocol for DNA Inputs Down to 100pg. The final fragmentase reaction requires an incubation time of 30 minutes at 37C to shear the DNA to a fragment size of 500-800bp, recommended for 454 libraries. The reaction is stopped by addition of 5ul of 0.5M EDTA. The final fragmentase protocol is integrated in the final capture protocol found in Appendix A.

4) Optimization of Conditions for DNA Fragmentation using Covaris

While enzymatic fragmentation following the optimized protocol was successful in shearing DNA from pristine samples to the desired 454 recommended size range, the optimal enzyme concentration and length of digestion required for fragmentation to the appropriate size range was found to be dependent on the DNA quality as well as quantity, thus not practical for forensic applications. For this reason, we also explored alternate mechanical shearing methods for fragmentation. For a library preparation protocol to be effective it needs to result in a standard degree of fragmentation regardless of initial sample quality and concentration. Forensic samples are often degraded and the ideal technology would cut larger fragments while leaving smaller fragments intact. This would decrease sample loss from over fragmenting. Mechanical shearing consists of breaking DNA by means of cavitation bubbles. These bubbles are created by wavelengths traveling through water. DNA wraps itself around the expanding bubble until the bubble implodes breaking the DNA. Adjusting the frequency of the wavelength changes the size of the cavitation bubble and in turn the length of DNA that can wrap around it. Covaris uses Adaptive Focused Acoustic (AFA) technology to focus high-energy wavelengths into a very small area inside a sample tube resulting in a very tight and more reproducible distribution of DNA size. Other shearing mechanisms work by creating cavitation bubbles throughout a water bath and are DNA concentration dependent.

The Covaris technology was chosen because it claimed to be concentration and quality independent because of its mechanism of DNA shearing. There is a much higher energy requirement to break smaller fragments when using cavitation bubbles and thus significantly more time is needed to shear short DNA fragments. Therefore, we developed and optimized a protocol for the DNA fragmentation step of the library preparation using mechanical shearing using the Covaris ultra sonicator. When DNA reaches ~200 bp, the treatment time needed to cut that fragment increases by more than 10 orders of magnitude. Because of the properties of mechanical shearing, this method was predicted to work well for degraded samples by cutting the large fragments while leaving the small ones intact. Covaris® is unique because it uses Adaptive Focusing Acoustics (AFA), which can target the ultrasonic waves into a focal point that is no bigger than a grain of rice and is centered inside the sample tube. The highly focused energy creates heat, and for that reason, Covaris® makes glass tubes that dissipate heat. Covaris® use of AFA technology is responsible for its accuracy and reproducibility.

Figure 26

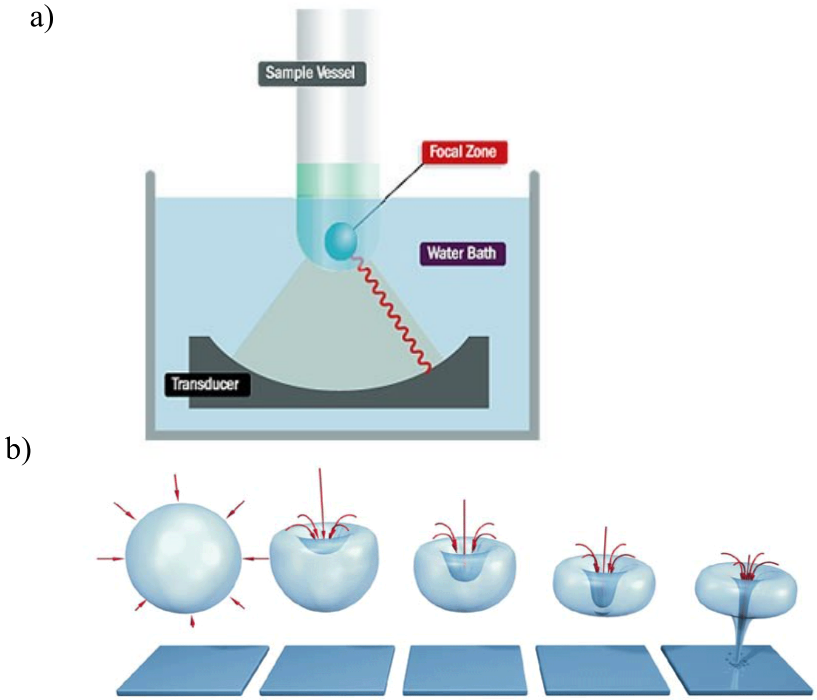


Figure 26. Covaris® Adaptive Focusing Acoustics Technology: a) AFA technology uses ultrasonic acoustic energy that travels through water to form cavitation bubble b) The cavitation bubbles form inside the sample tube shear DNA using the inward energy (red arrows) as they implode. Images from Covaris®

There are four parameters that can be changed to manipulate the fragment size of DNA: Cycles per Burst, Duty Factor, Peak Incident Power, and treatment time. Cycles per burst are the amount of wave cycles in each burst. Duty factor is a percentage of the total treatment time that is actively releasing energy. Peak incident power is the energy in watts being released by the transducer. Treatment time is the process total time in seconds. Covaris® protocols for different sample volumes and fragment size depend on changes in these four parameters (Covaris® Protocol).

Covaris® has two protocols for the same DNA size fragments depending on the sample volume. Table 16 compares the two set protocols for 50 and 130 µl. The protocols differ in peak incident power and treatment time. To fragment a sample in a 50 µl volume, a higher energy is used for a shorter period of time. In contrast, to fragment a DNA sample in a 130 µl volume employs a lower energy and a longer treatment time. The difference in protocols is due to the sample tube. MicroTUBES manufactured by Covaris® hold 130 µl but are used for both the 50 and 130 µl treatments. When the 50 µl protocol is used, droplets will separate from the sample and may not receive treatment. If droplets form, part of the sample will not receive treatment, resulting in reduced reproducibility. To reduce reproducibility inconsistencies, the protocol used on 50 µl reactions is shorter and stronger. For increased reproducibility, the manufacturer recommended to use the 130 µl protocol if possible. However, because forensic samples are often limited and the subsequent steps in library preparation require small volumes, a 50 µl sample volume is preferred. To test reproducibility of the two sample volumes, Four 50 µl and four 130 µl samples were prepared to contain 200 ng and then sheared in the S220 Covaris® machine using the 500 bp protocol for each volume size (see Table 16).

Table 16 – Covaris® parameters for shearing DNA to 500bp in 50 µl sample volume and in 130 µl sample volume.

	500 bp Target 130 µl Sample	500 bp Target 50 µl Sample
Peak Incident Power (W)	105	175
Duty Factor	5%	5%
Cycles per Burst	200	200
Treatment time (S)	80	35

After treatment, the samples were analyzed on a 1% agarose gel alongside the intact DNA and a 100 bp ladder for size reference (see Figure 27). The size range of DNA was 100-1,200 bp with similar distribution between samples for both sample volumes. There was also no observable difference in size distribution within replicates at each sample volume. The results of this experiment suggest that the 50 µl sample volume is as reproducible as the 130 µl samples. A larger study would be needed to establish the true variability; however, based on these results, we adopted 50 µl as the preferred sample processing volume.

Figure 27.

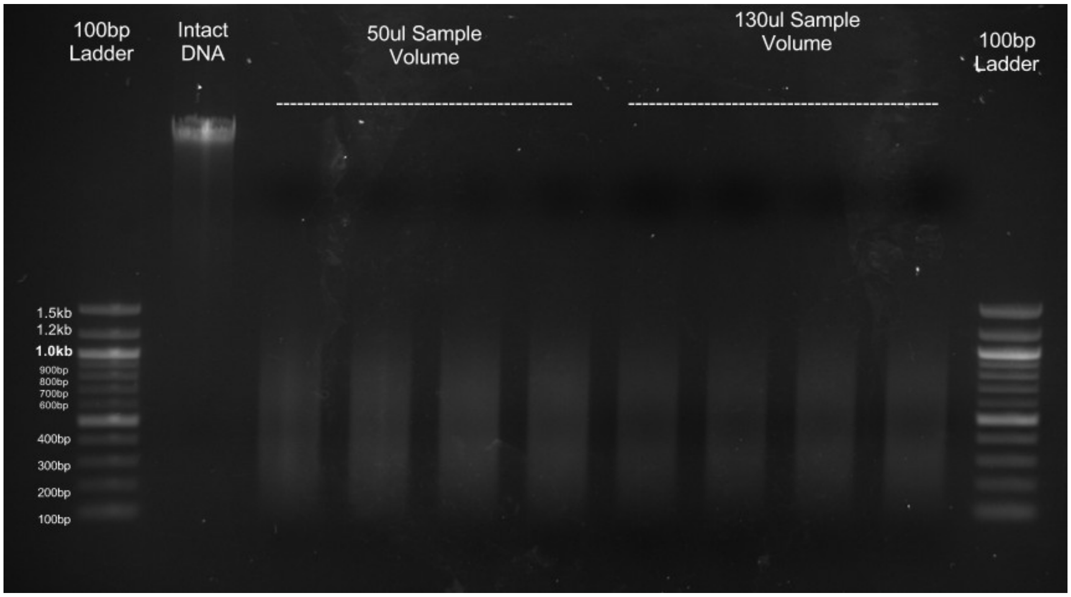
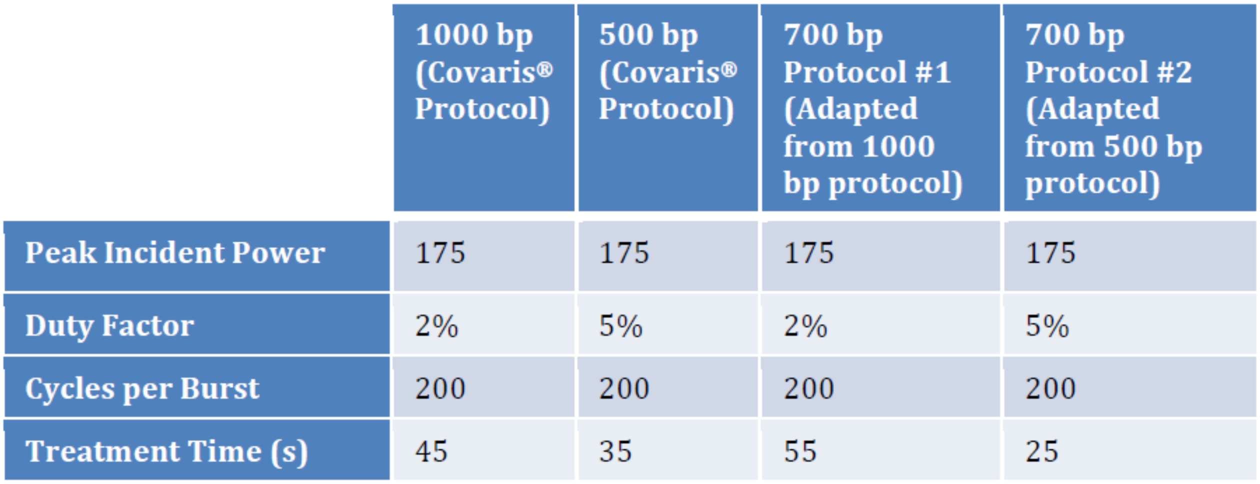


Figure 27 Comparing Covaris® sample volumes. DNA to a total of 200ng in 50 and 130 µl volumes, is sheared to 500 bp using Covaris. The gel loading dye runs at 500 bp, casting a shadow or artifact in the section of interest.

Additionally, a protocol was optimized for a targeted 700 bp fragment size which is optimal for the 454 GS to maximize read length. Using both 500 and 1000 bp as guidelines, two 700 bp protocols were designed. Table 17 shows the Covaris® protocols for 500 and 1000 bp compared to the designed 700 bp protocol parameters. Both 700 bp protocols were tested by shearing 200 ng of intact control DNA. The samples were sheared in duplicate and then analyzed on a gel to visualize the results. The protocol that was adapted from 1000 bp (700 bp Protocol #1) had a longer exposure time of 55 seconds to shear the DNA more, while keeping all other parameters constant. The protocol adapted from 500 bp (700 bp Protocol #2) had a shorter treatment time cutting DNA less than 500 bp, while also keeping all other parameters equal. When the Duty Factor is higher there is more active treatment time during the protocol and therefore the total amount of time needed to cut DNA is shorter. In general, a short treatment time leads to inconsistent results. Figure 28 shows both protocols resulted in similar size distributions. Protocol #1 was adopted due to its longer treatment time.

Table 17. Covaris® shearing parameters to target 1000 bp and 500 bp fragments, as well as the two adapted protocols to target 700 bp fragments.



	1000 bp (Covaris® Protocol)	500 bp (Covaris® Protocol)	700 bp Protocol #1 (Adapted from 1000 bp protocol)	700 bp Protocol #2 (Adapted from 500 bp protocol)
Peak Incident Power	175	175	175	175
Duty Factor	2%	5%	2%	5%
Cycles per Burst	200	200	200	200
Treatment Time (s)	45	35	55	25

Figure 28

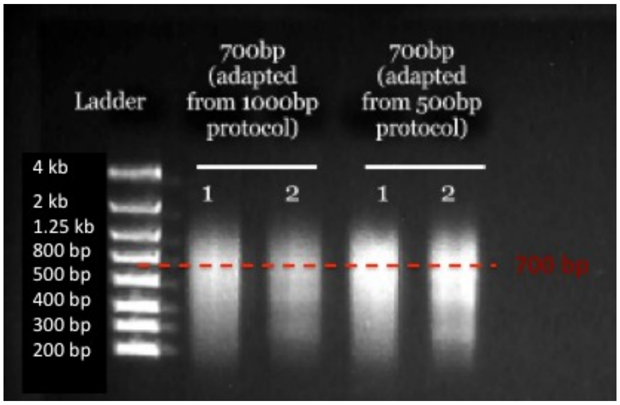

Ladder
700bp (adapted from 1000bp protocol)
700bp (adapted from 500bp protocol)
1
2
1
2
4 kb
2 kb
1.25 kb
800 bp
500 bp
400 bp
300 bp
200 bp
700 bp

Figure 28– Covaris® fragmentation targeting 700 bp fragments using protocols adapted from the manufacturer's 1000 bp and 500 bp parameters. See Table 1 for parameters. Run on a 1.2% Lonza Flash Gel.

a) Covaris Degraded DNA

A series of experiments were conducted on a Covaris S220 to assess if this mechanical shearing technique was independent of starting DNA fragment size. To test DNA quality independence, DNA was naturally or artificially degraded to different levels ranging from 20kb to 500 bp. DNA was naturally degraded to ~20 kb by incubating blood at room temperature for one month. To obtain other levels of degradation, we artificially degraded DNA using enzymatic fragmentation and the Covaris® technology. To obtain the large size range of samples, DNA was sheared with the Covaris® technology to 20 kb using a special centrifuge tube shaped like an hourglass containing a very small ruby between the two compartments and centripetal force and to 5kb and 3kb using the Covaris cavitation bubble technology. In order to test the Covaris® technology on highly fragmented DNA samples, control DNA was artificially degraded to 1.25 kb to 500 bp with Fragmentase® for 10, 20, 30, 40 minutes. Following natural or artificial degradation, the samples were then all sheared using the manufacturer's recommended protocol targeting 500 bp. Results were compared on both a 1% Agarose gel and a 1.2% Lonza Flash Gel. Results show that regardless of the initial level of degradation, the fragment size distribution was similar after subsequent shearing using the Covaris without loss of smaller fragments (See Figure 29 and 30). These preliminary results show proof-of-principle that a single set of parameters can be used with the Covaris AFA technology for shearing DNA with various levels of degradation without leading to loss of smaller fragments.

Figure 29.

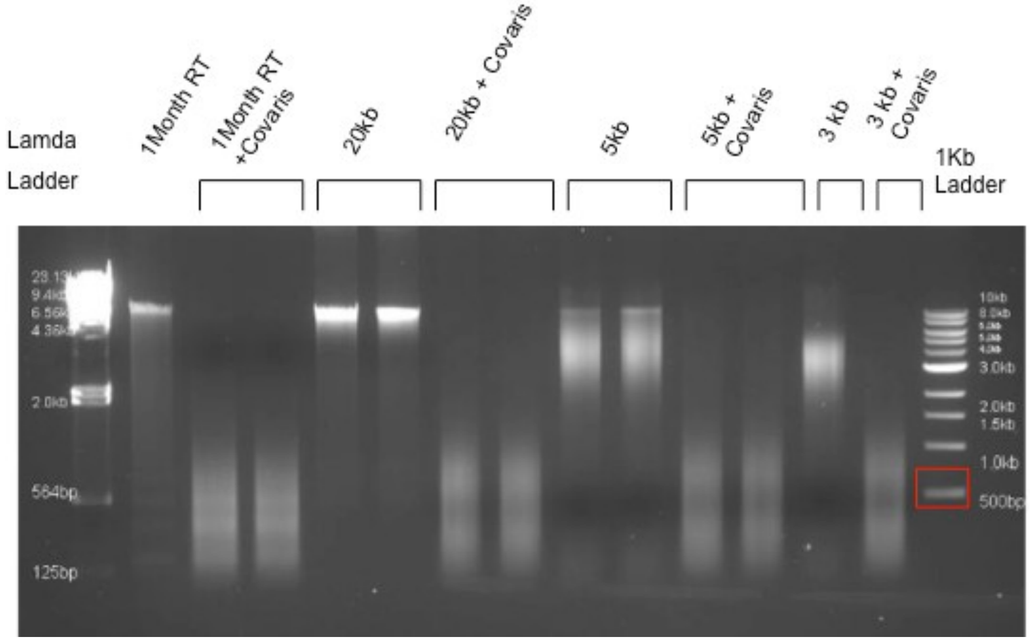


Figure 29 Various levels of DNA degradation before and after Covaris® fragmentation targeting 500 bp fragments, demonstrating uniform fragmentation independent of initial DNA quality. To analyze the DNA, it was run on a 1% Agarose gel.

Table 18 Fragment size distribution range for samples treated with 500 bp Covaris® shearing.

	1 Month RT + Covaris®	20 kb + Covaris®	5 kb + Covaris®	3 kb + Covaris®
Size Distribution	100-2,000 bp	100-2,000 bp	100-2,000 bp	100-2,000 bp

Figure 30.

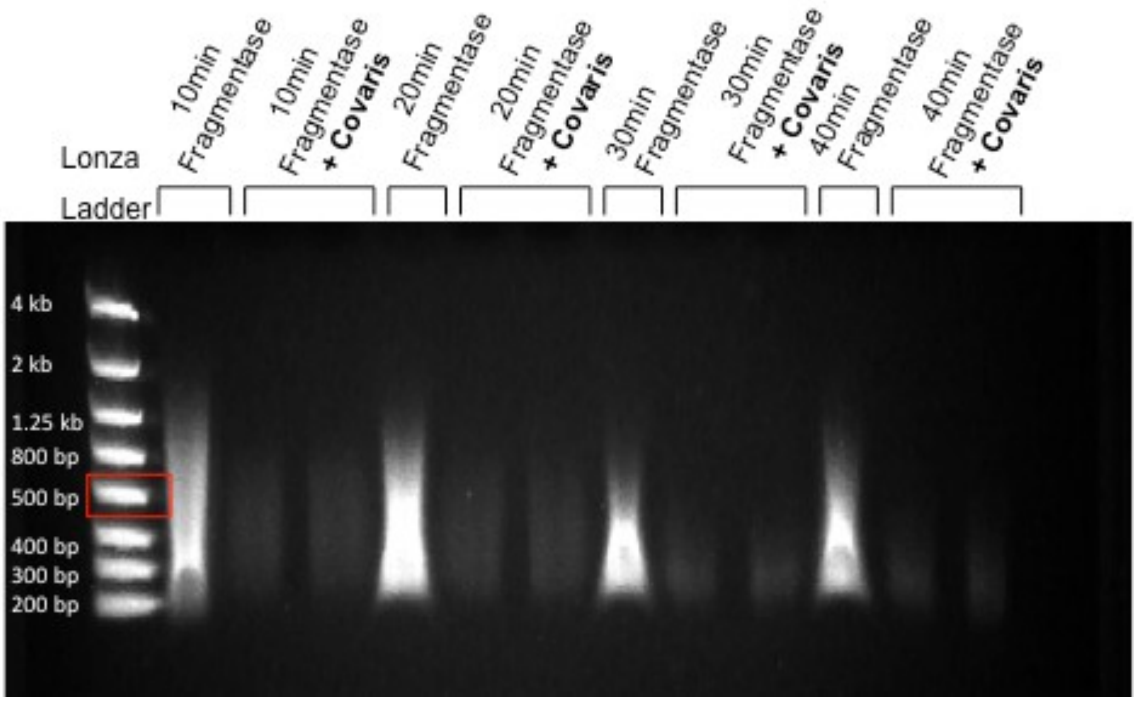


Figure 30. Highly Degraded DNA before and after Covaris® fragmentation targeting 500 bp fragments, demonstrating no DNA loss secondary to over---fragmentation. Run on a 1.2% Flash Gel.

Table 19 Fragment size distribution range for samples treated with 500 bp Covaris® shearing.

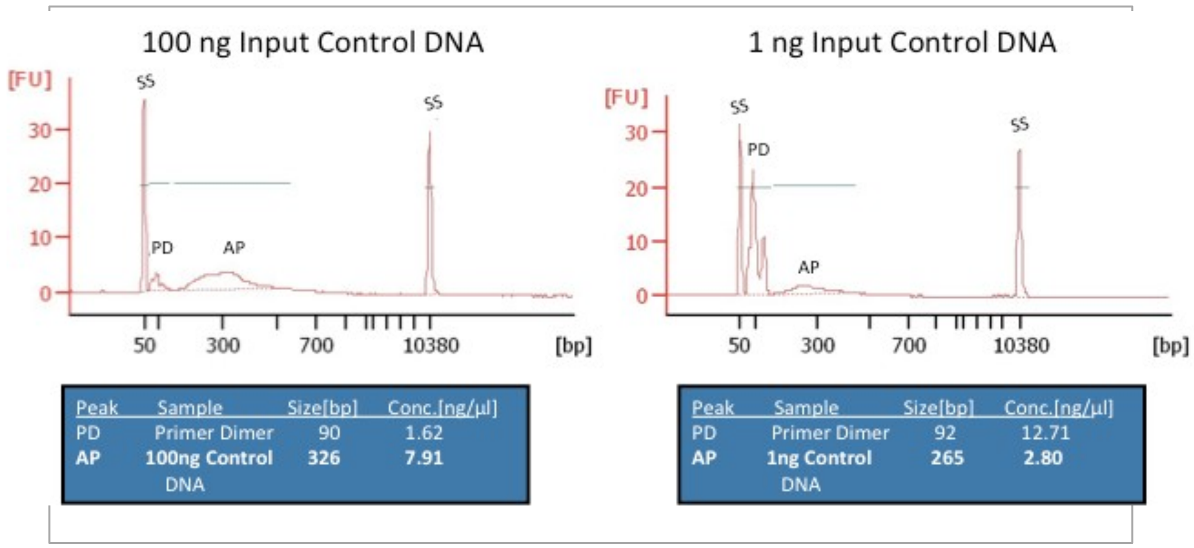
	10 min Fragmentase + Covaris®	20 min Fragmentase + Covaris®	30 min Fragmentase + Covaris®	40 min Fragmentase + Covaris®
Size Distribution	250-800bp	250-800bp	250-500bp	200-500 bp

5) Pre and Post Capture PCR Optimization

We also optimized the pre-capture PCR step of the rapid library preparation. Due to our much lower starting amounts of DNA (1ng-100ng), we increased the number of PCR cycles to 28 from 11 recommended by the manufacturer to obtain a minimum of 1ug of PCR product prior to the capture, per manufacturer specifications.

In addition, the Post-capture PCR was optimized by decreasing the PCR primer concentration. The manufacturer recommended 4 uM of primer for 1 ug of DNA. However, a much lower DNA sample amount is typically encountered in forensic samples (<1 ng). Using the recommended PCR protocol with a primer concentration of 4 μM with lower amounts of input DNA resulted in increased primer dimer with decreasing sample DNA amounts due to the excess primer amounts. As shown in Figure 31, the amount of primer dimer increased approximate eight fold when amplifying the 1 ng sample compared to 100 ng of sample DNA.

Figure 31.



Peak	Sample	Size[bp]	Conc.[ng/µl]
PD	Primer Dimer	90	1.62
AP	**100ng Control** DNA	**326**	**7.91**

Peak	Sample	Size[bp]	Conc.[ng/µl]
PD	Primer Dimer	92	12.71
AP	**1ng Control** DNA	**265**	**2.80**

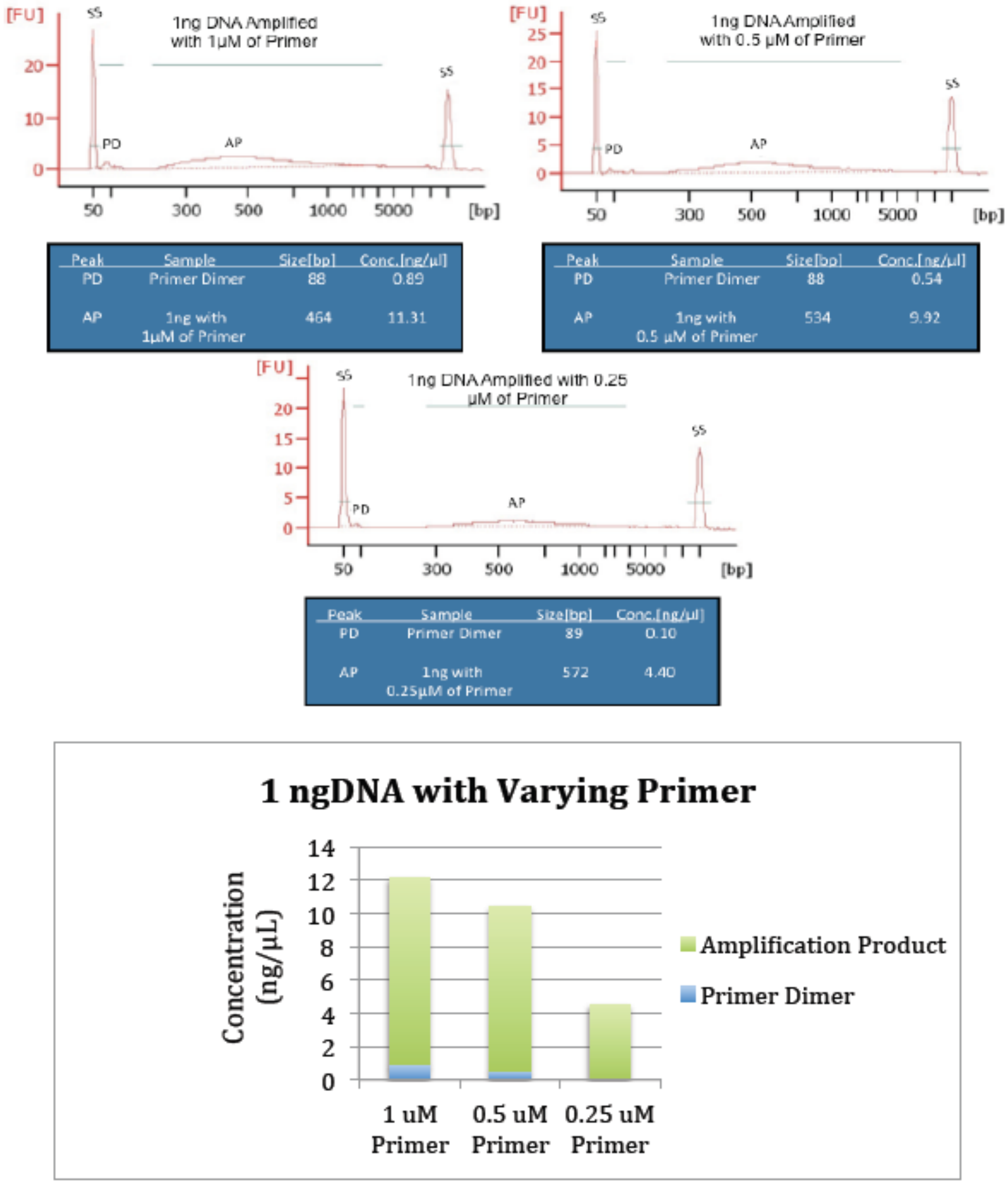
Figure 31 Primer dimer and amplification product peaks for 100 and 1 ng amplification under manufacturer recommended primer concentration are observed and compared in both flowgram and bar graph form. The first and last peaks are size standards (SS), PD peak is the primer dimer and AP is the amplification primer peak. Sample libraries were analyzed on an Agilent Bioanalyzer 12000 Chip.

To decrease generation of primer dimer and improve product yields, the PCR primer concentration was reduced. A range of primer concentrations (1, 0.5 and 0.25 µM) was tested to determine the optimal primer concentration with DNA amounts within the range of forensic samples (1 ng). A primer concentration of 1 uM resulted in decreased primer dimer without compromising PCR product yields (Figure 32 below).

When 4 µM of primer was used, a 12.71 ng/µl primer dimer concentration was observed (Figure 32). Reducing the primer to 1 µM significantly decreased primer dimer to 0.89 ng/µl. Primer dimer is further reduced as primer concentration is decreased to 0.5 and 0.25 µM. However, amplification efficiency is reduced as the primer concentration decreases to 0.5 and 0.25 µM; the lower primer concentration appears to be rate limiting resulting in lower product yields with an observed DNA concentration decline from 11.31 ng/µl with 1µM, to 9.92 ng/µl with 0.5 µM and

4.40 ng/µl with 0.25 µM primer. Based on these results, the optimal primer concentration for lower starting input DNA amplification was determined to be 1 µM.

Figure 32.



Peak	Sample	Size[bp]	Conc.[ng/µl]
PD	Primer Dimer	88	0.89
AP	1ng with 1µM of Primer	464	11.31

Peak	Sample	Size[bp]	Conc.[ng/µl]
PD	Primer Dimer	88	0.54
AP	1ng with 0.5 µM of Primer	534	9.92

Peak	Sample	Size[bp]	Conc.[ng/µl]
PD	Primer Dimer	89	0.10
AP	1ng with 0.25µM of Primer	572	4.40

Figure 32 Primer Concentration experiment using 1 µM, 0.5 µM, 0.25 µM of primer on 1 ng of initial sample DNA represented in both flow gram and bar graph form. The flow gram's first and last peaks are size standards (SS), PD peak is the primer dimer and AP is the amplification primer peak. Samples were visualized with an Agilent Bioanalyzer 12,000 chip.

b. Final Optimized Methods for Whole Mitochondrial Genome Library Preparation, Sequence Capture and 454 NGS Sequencing

The whole mitochondrial capture and NGS assay is composed of three main steps: library preparation, capture, and sequencing. The final optimized methods used to show proof of concept of the whole mitochondrial genome sequence capture and 454 NGS sequencing are briefly described below.

1) Rapid Library Preparation

DNA samples were prepared for capture and sequencing following a modified rapid DNA library preparation method following the 454 manufacturer's protocol. An overview of the library preparation method is provided in the schematic in Figure 33 below. The library preparation begins with fragmentation of the sample DNA to a size that can be optimally sequenced. For the current Roche 454 GS Titanium platform, the maximum sequencing length is 400□500 bp, and a recommended average target fragment size of 700 bp. DNA was sheared to an average size of 500 bp using enzymatic fragmentation following the optimized fragmentase protocol described above in Table 15 or using mechanical shearing using the Covaris Focused Ultrasonicator (Covaris, MA) following the optimized protocol described in Table 17 above. Irrespective of the fragmentation method used, the ends of the DNA fragments have irregular overhangs which need to be transformed to Adenosine (A) overhangs for proper adaptor ligation. End Repair and A-tailing was performed following the manufacturer's protocol to repair the ends of the fragments and add an additional Adenosine to the end of the DNA fragments using Taq Polymerase which naturally adds an Adenosine to fragment ends. After end repair and A-tailing, the DNA fragments are now ready to be ligated to the 454 MID adaptors.

Figure 33

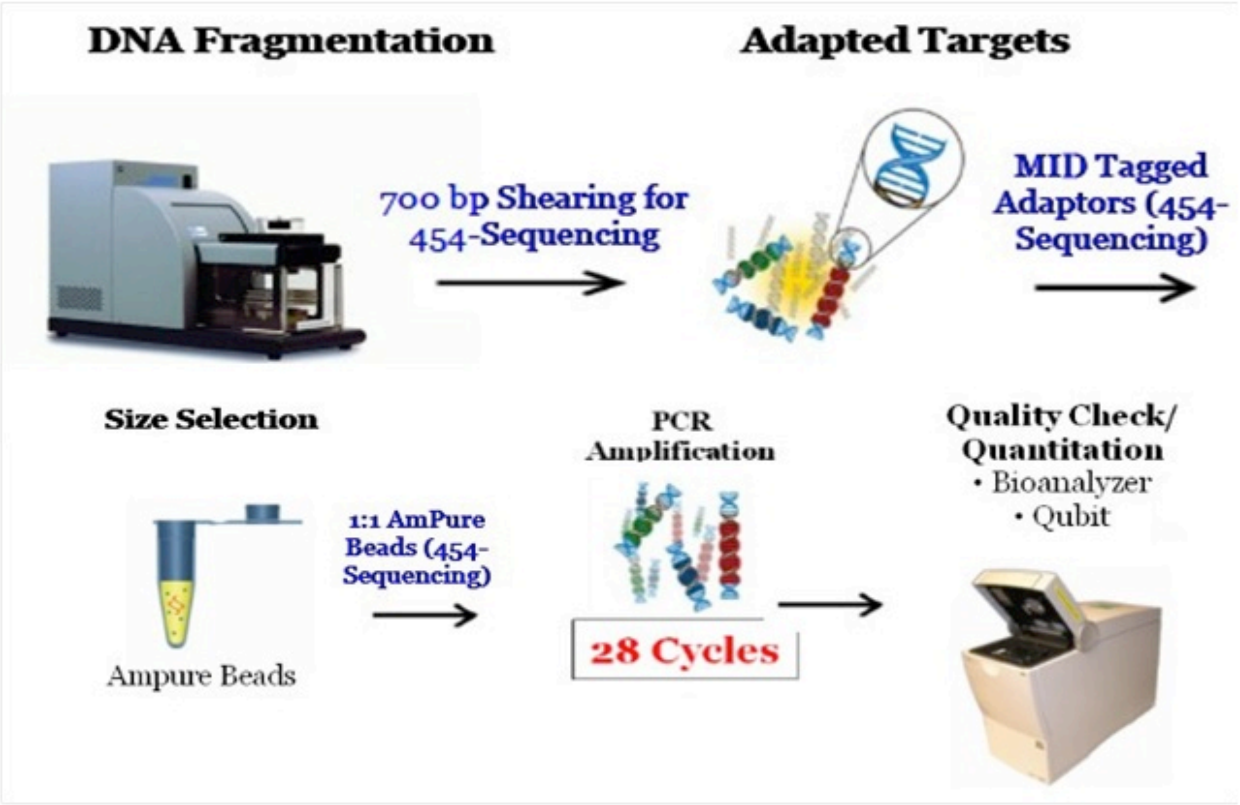

DNA Fragmentation
Adapted Targets
700 bp Shearing for 454-Sequencing
MID Tagged Adaptors (454-Sequencing)
Size Selection
PCR Amplification
Quality Check/ Quantitation
• Bioanalyzer
• Qubit
1:1 AmPure Beads (454-Sequencing)
Ampure Beads
28 Cycles

Figure 33. **Library Preparation of Sample DNA**: Sample preparation starts by fragmentation of input DNA. Once the DNA is fragmented to a size that can be sequenced, adaptors (containing a primer and a barcode) are ligated to the DNA fragments. Once the sample DNA has been tagged, a purification step removes small fragments before the DNA is amplified for 28 cycles. Checking the quality of the final libraries is the last step. Images adapted from Covaris®, Roche 454, and Beckman Coulter.

The 454 MID adaptors have a complex structure which is shown in Figure II-1. Each adaptor consists of a primer sequence followed by a generic sequence and ending in a multiplex identifier sequence (MID). The primer sequence is used during the amplification steps pre and post capture as well as during the emulsion PCR and pyrosequencing steps of 454 sequencing. The MID sequence is a unique sequence used to identify a particular sample, which allows multiple samples to be captured and sequenced at one time. MID barcoding and sequencing multiple samples in one sequencing run reduce the per sample cost while still maintaining adequate sequence coverage of each sample. The 454 MID adaptors are also double stranded and forked due to the incorporation of primer sequences within the adaptors. The specific MID adaptor sequences are given in Table 34 and were purchased through Integrated DNA Technologies, Inc., San Diego, CA.

Figure 34. 454 MID Adaptor Components and Orientation with Target DNA Fragment

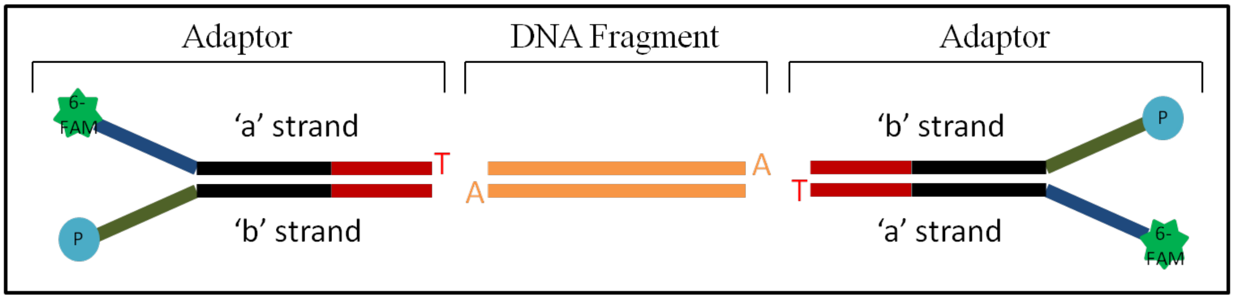


Figure 34. Components of 454 MID Adaptors and Structure of Completed Library Fragments. Each 454 MID adaptor is double stranded with an 'a' strand and a 'b' strand. The 'a' strand contains a 5' 6-FAM, primer oligo A (blue), an intermediate sequence (black) and a unique MID tag with a 3' tyrosine overhang (red). The 'b' strand contains a 5' phosphate, primer oligo B (green), an intermediate sequence complimentary to the 'a' strand (black) and the complimentary sequence of the unique MID tag in the 'a' strand without the last adenosine base to preserve the T overhang on the 'a' strand (red). Fragments of the genomic DNA of interest are prepared to have 3' adenosine overhangs (orange) allowing ligation of adaptors to each end. A completed library consists of fragments of DNA as shown, with adaptors ligated to both ends.

After adaptor ligation, the 454 libraries need to be cleaned up and a size selection performed to remove excess adaptors and targets below 300bp. Using Agencourt AMPure XP (Beckman Coulter, Brea, CA) with 1:1 DNA to AMPure bead volumetric ratio as a purification step, the unattached adaptors and small fragments are separated from the sample DNA. The sample DNA/MID fragments are then amplified using the MID adaptor sequences as primers. By using the 454 adaptor sequence as the PCR primer, the sample DNA is amplified without an intact target sequence. The PCR cycle number was increased to 28 cycles from the recommended 15 cycles to ensure that the minimum PCR product amount required for capture is obtained.

2) SeqCap EZ Probe Capture System

The fragmented, adaptor ligated and purified libraries then undergo probe capture procedure to enrich for the mtDNA sequences. The Nimblegen SeqCap EZ Probe Capture System is a solution based capture which consists of hybridization followed by a series of wash steps to remove unbound and non-specific targets and finally a low cycle amplification and quality check. Probe capture was performed following the recommended manufacturer's procedure using a three day or one day hybridization. The overview of the probe-capture process is summarized in the Figure 35 below.

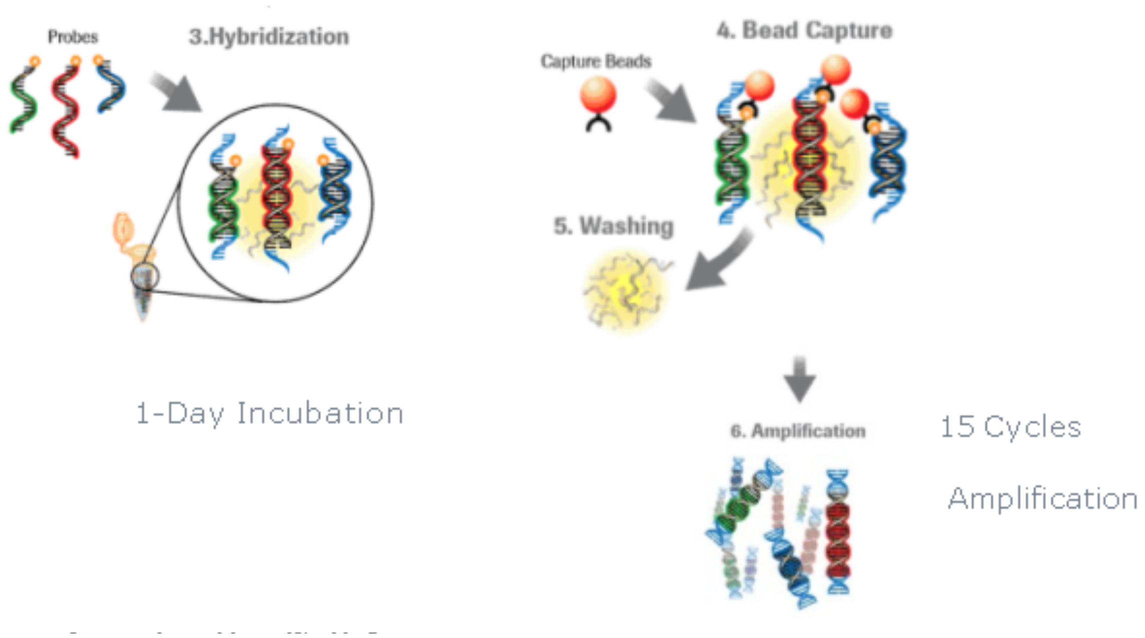


Figure 35 An overview of the probe capture step of enriching for the mitochondrial genome. Here we start by adding biotinylated mtDNA probes to our sample libraries and hybridizing them for one day. Using streptavidin beads, we capture the probe-target complex and wash away non-specific DNA. The enriched DNA is then amplified for 15 cycles. Image adapted from NimbleGen.

The capture process starts with hybridizing sample libraries to the mtDNA probes. The Nimblegen capture protocol recommends that a total of 1ug of the amplified DNA library is added to hybridization reaction along with the custom capture probes. Using MID tags label and pool samples prior to capture reduces the total PCR product amount required per sample and increasing the number of PCR cycles to 28 ensured adequate PCR product amounts were obtained. In order to increase the on target capture rate of the probes, COT DNA and hybridization enhancing oligos are also added to the hybridization reaction. COT DNA is a pool of short DNA fragments consisting of all of the repetitive sequences throughout the human genome. It is included in the reaction to bind to all repetitive elements which may be in the target library. However, all the samples have an MID adaptor that could hybridize to each other, forming concatemers that can lower the capture's specificity. For this reason, there are blocking oligos that match the MID-454 adaptor sequence, leaving only the target sequences free to interact with the probes. The hybridization enhancing oligos are designed to block the primer and adapter sequences to keep the probes from potentially binding to them as well as to prevent secondary structures from occurring between targets within the library. After a three day or one day hybridization, magnetic streptavidin beads are added, which bind to the biotinylated probe-target complex. With the use of a magnet, the streptavidin probe target complex is pelleted and the non-specific DNA fragments are washed away. Last, the enriched DNA goes through a low-cycle amplification.

After the amplification a final clean-up using AMPpure beads is required to remove excess primer or adapter dimer. The captured library is now ready for a quality check using the bioanalyzer followed by a qPCR to analyze the capture effectiveness. To determine the size distribution and quality of the capture libraries, we use Quant-iT™ dsDNA High-Sensitivity Assay (Invitrogen, Grand Island, NY) and Agilent Bioanalyzer chip (Agilent Technologies, Santa Clara, CA). The libraries are expected to exceed a 1 µg yield and have a fragment size distribution between 500-1,500 bp, with the distribution peak between 600--1000 bp. Nimblegen includes probes from four control loci within each custom capture design, which allows for analysis of the capture enrichment. This analysis is completed using a targeted SYBER green qPCR assay comparing pre-capture libraries to post-capture libraries at all four loci as well as a positive genomic DNA control and a negative control each in triplicate. The enrichment qPCR can only determine the capture efficiency of the control loci, however, it is used as an indication the capture was successful. The capture is considered successful if enrichment of the control loci is achieved, as seen by lower Ct values of the post capture sample compared to the Ct values of the pre capture sample. After determining successful capture and a quality post capture library, it is ready to sequence.

3) 454 Next Generation Sequencing

The third and final step of the whole mitochondrial genome sequencing assay is the 454 pyrosequencing of the captured mitochondrial DNA fragments. An overview of the emulsion PCR and 454 pyrosequencing is provided in the Figure 36 below and is described in more detail in section 9) of the Method for HVI/HVII assay above. EmPCR and 454 pyrosequencing are performed following the recommended manufacturer's protocol.

Figure 36.

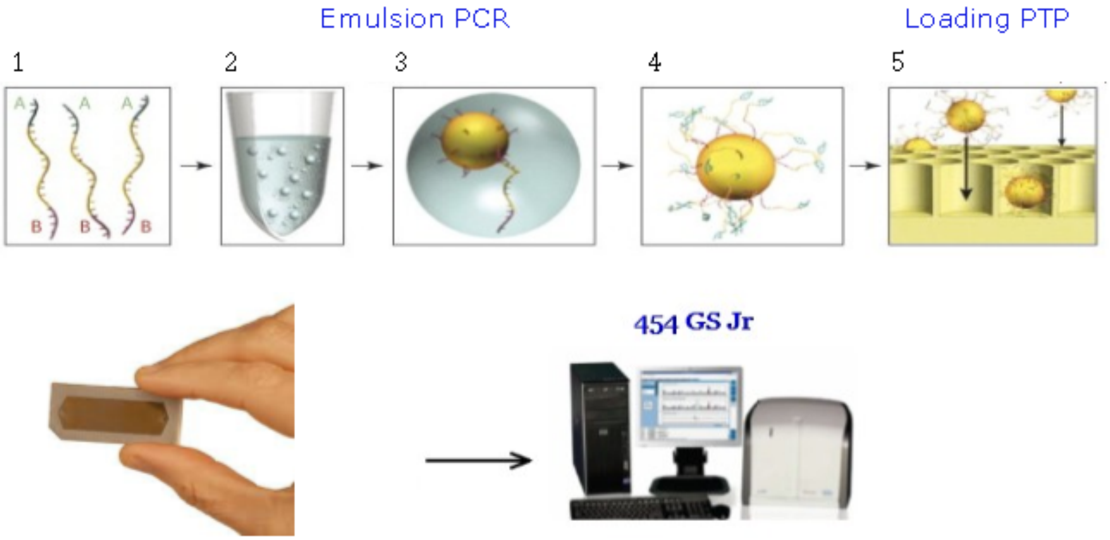


Figure 36 Emulsion PCR and Pyrosequencing: The third and final portion of the assay is the clonal amplification and pyrosequencing using the 454 GS Jr platform. Image adapted from Roche 454.

iii. Results

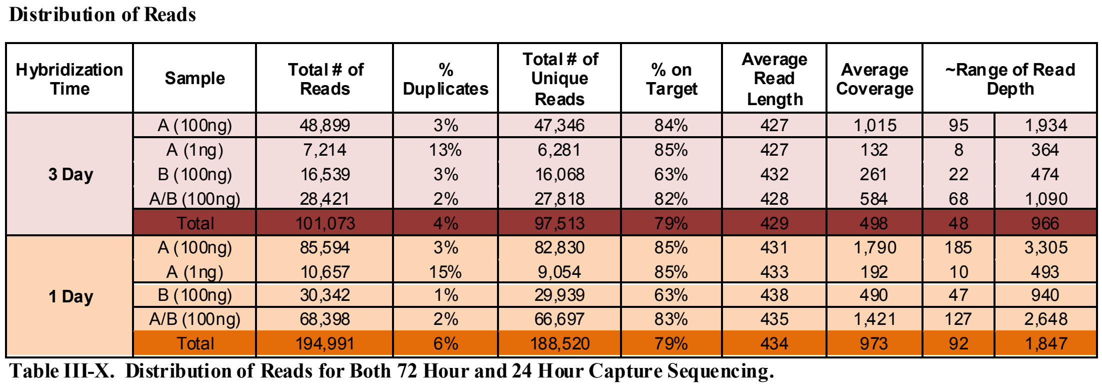
The full mtDNA sequence capture assay was initially tested for proof-of-concept by completing two captures and sequencing runs of four samples at two different hybridization times. The four samples consisted of Sample A at 100ng, Sample A at 1ng, Sample B at 100ng, and a 90:10 mixture of Samples A:B at 100ng. The DNA library was generated using the optimized fragmentase protocol for shearing the DNA enzymatically and the four DNA samples were uniquely tagged with different MIDs to allow for sample pooling. Two different DNA concentrations of a single sample were studied to determine if the starting DNA amount could be reduced from the manufacturer's recommendation of 1 ug to more forensically relevant amounts of DNA, two different samples were sequenced to determine the probe capture specificity, a mixed sample was included to determine if a 10% minor sequence could be detected which is below the limits of Sanger sequencing, and the DNA library was hybridized at two different hybridization times to determine if the hybridization time could be reduced. Samples A and B were selected as the entire mitochondrial genome of these two samples had been previously sequenced using Sanger sequencing. For the first experiment, we followed the Nimblegen manufacturer's protocol and performed a three day hybridization. After successful capture and sequencing using the 72 hour hybridization incubation time, the capture was then repeated with the same four samples DNA library with a 24 hour hybridization to directly compare hybridization efficiency at the two time points and determine if the hybridization time could be reduced to increase throughput of the assay.

Table 20. Experimental Design

Sample Type	DNA Input	Samples into Capture	Hybridization Time
Individual A	100ng	Multiplex all 4 samples into capture	72 Hours (recommended)
Individual A	1ng		
Individual B	100ng		
A/B 90%:10% Mixture	100ng		
Individual A	100ng	Multiplex all 4 samples into capture	24 Hours
Individual A	1ng		
Individual B	100ng		
A/B 90%:10% Mixture	100ng		

These sets of experiments demonstrated proof-of-concept for using sequence probe capture for enrichment and next-generation sequencing of the entire mitochondrial genome. The optimized system resulted in 100% capture of the mitochondrial genome with an ~80% on target rate with an average 500-1000 fold sequence coverage per base. The average read length was ~430 bp. Results for the three day and one day hybridization are summarized in Table 21 below.

Table 21. Distribution of Reads for 72 and 24 hour Hybridization



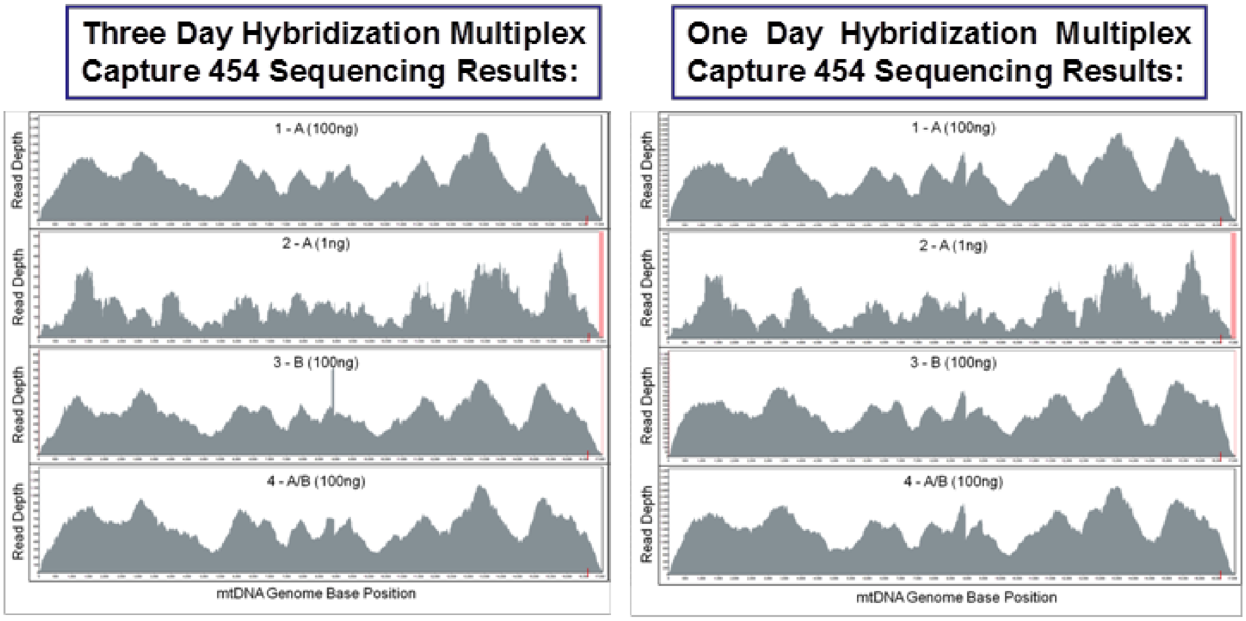
Distribution of Reads

Hybridization Time	Sample	Total # of Reads	% Duplicates	Total # of Unique Reads	% on Target	Average Read Length	Average Coverage	~Range of Read Depth	
3 Day	A (100ng)	48,899	3%	47,346	84%	427	1,015	95	1,934
	A (1ng)	7,214	13%	6,281	85%	427	132	8	364
	B (100ng)	16,539	3%	16,068	63%	432	261	22	474
	A/B (100ng)	28,421	2%	27,818	82%	428	584	68	1,090
	Total	101,073	4%	97,513	79%	429	498	48	966
1 Day	A (100ng)	85,594	3%	82,830	85%	431	1,790	185	3,305
	A (1ng)	10,657	15%	9,054	85%	433	192	10	493
	B (100ng)	30,342	1%	29,939	63%	438	490	47	940
	A/B (100ng)	68,398	2%	66,697	83%	435	1,421	127	2,648
	Total	194,991	6%	188,520	79%	434	973	92	1,847

Table III-X. Distribution of Reads for Both 72 Hour and 24 Hour Capture Sequencing.

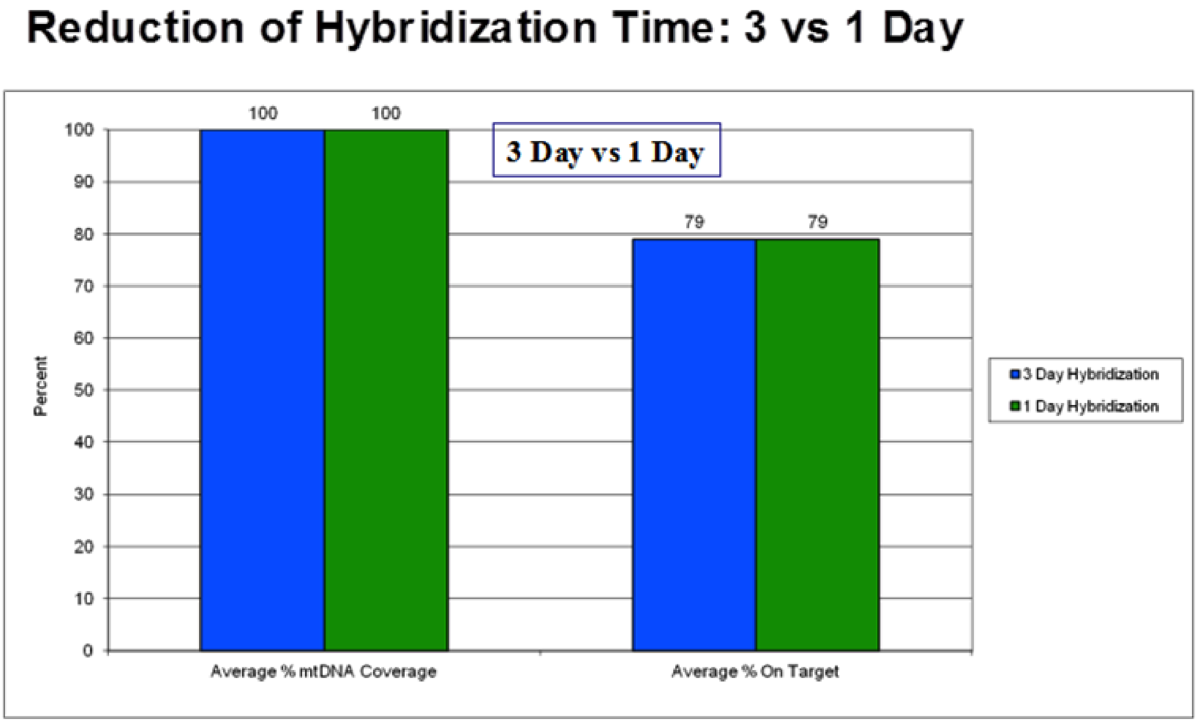
The distribution of reads differed across the mitochondrial genome but was similar between samples at the same concentration (100 ng) (Figure 37). The differences in the distribution are likely due to differences in probe capture efficiency.

Figure 37.



To increase the efficiency of the entire process, we investigated reducing the hybridization time from the recommended three days down to one day. No significant differences were observed between the 3 day or 1 day hybridization times likely due to the vast number of probes and high copy number of mitochondrial DNA favorable for probe capture (Figure 38).

Figure 38

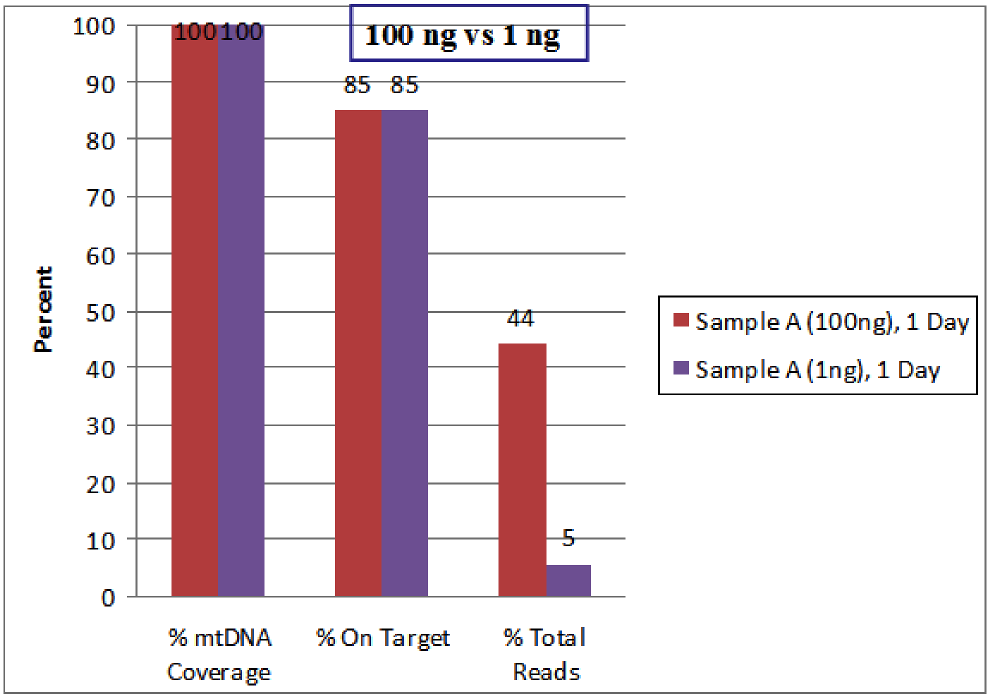


- Decreasing the hybridization time to one day did not result in a loss of probe specificity or increase in off target capture rate

We initially tested sensitivity by reducing the starting amount of DNA to 100ng and 1ng. The manufacturer protocol recommends a starting DNA amount of 1 ug of DNA, significantly higher than DNA amounts encountered in forensic cases. Our results show we were able to successfully capture 100% of the mitochondrial genome with both 100 ng and 1ng of DNA input with an average of 230-1000 fold coverage and an average on target capture rate of 79% for both 3 and 1 day hybridization times. No difference in on target capture rate was observed between the sample tested at 100 ng and 1 ng (85% on target rate) (Figure 39). An ~7x difference in the percent total reads for the run was observed between the 100 ng and 1 ng sample but this is likely due to quantification error and not differences due to DNA amounts.

Figure 39

Sensitivity of mtDNA Capture: 100ng vs 1ng

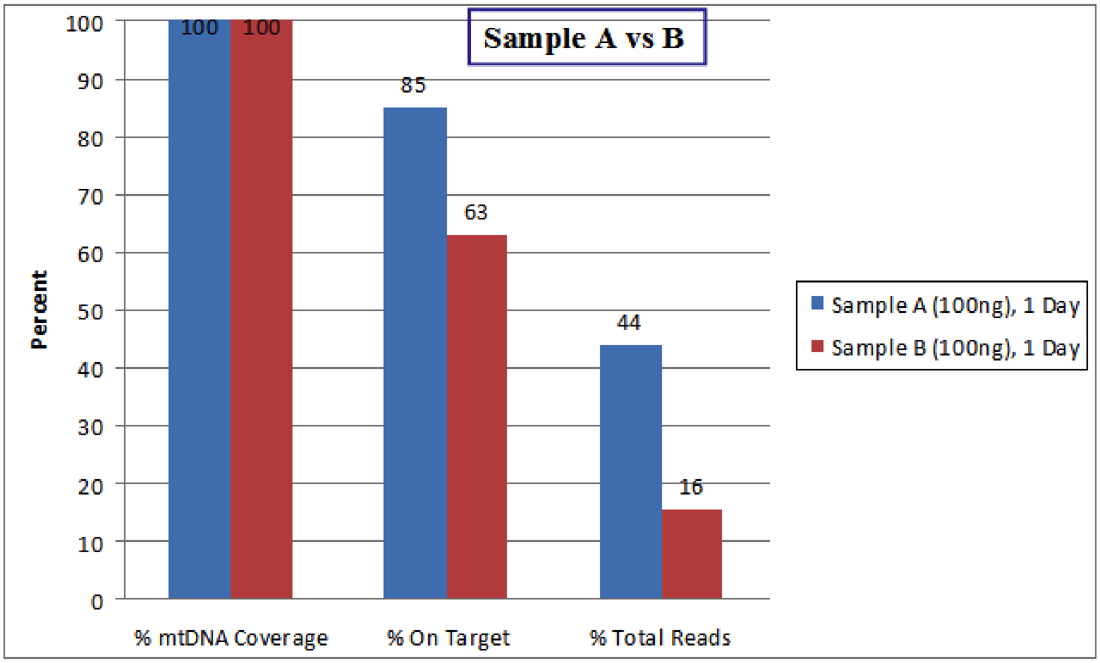


- Difference of % total reads likely due to quantitation error
- 500 fold reduction of the input DNA did not effect the capture or on target rate

We also tested probe specificity by testing two different DNA samples at both hybridization times by comparing 3 days to 1 day hybridization. No differences between hybridization times were observed. Mitochondrial DNA sequences differences did not lead to a loss of coverage (100% coverage observed). Differences in the percent on target rate were observed between the samples (85% vs 63%) (Figure 40). The inconsistency could be due to differences in nuclear DNA sequence or capture efficiency. Additional samples need to be tested prior to drawing any conclusions. Differences in the percent total reads were observed between the samples, but as discussed above are likely due to quantification error.

Figure 40.

Capture Specificity: Comparison of Samples

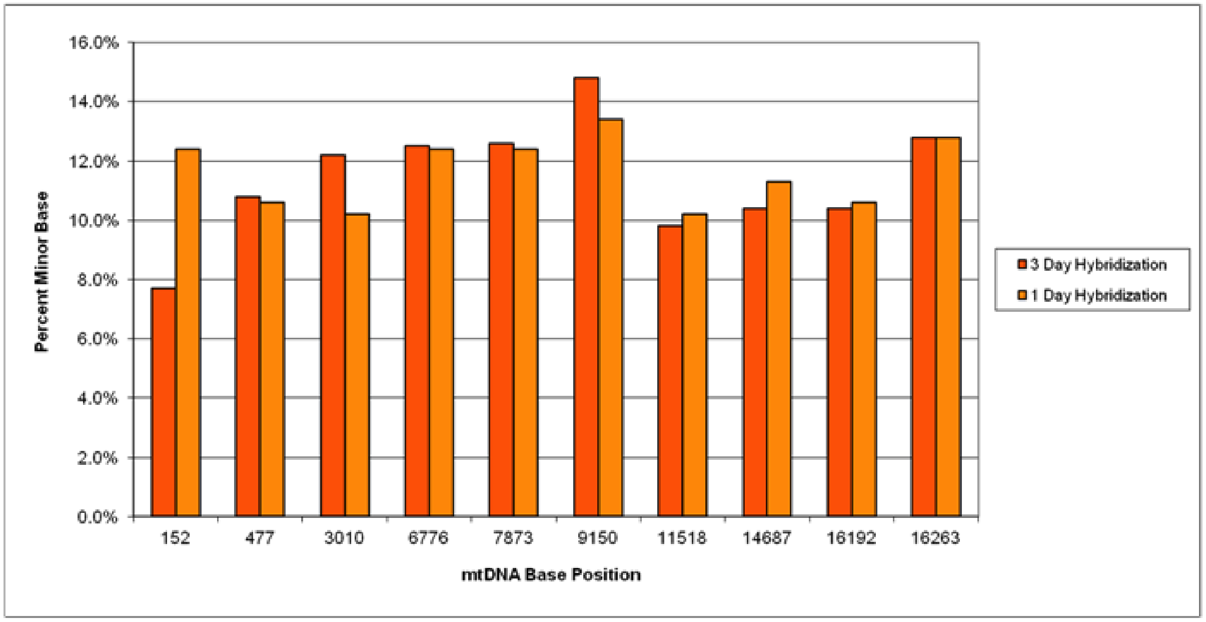


- On target capture rate varied between samples, likely due to DNA sequence differences
- Additional samples will need to be tested to better determine the on target capture rate

To determine the ability to detect a mixed DNA sample, the mtDNA copy number was determined using a qPCR assay for two samples and mixed at 90:10 ratio. All SNPs previously detected by Sanger sequencing were detected by 454 sequencing and an ~10% mixture was detected at each of the 10 mixed base positions. Results are summarized below in Table 22 and Figure 41.

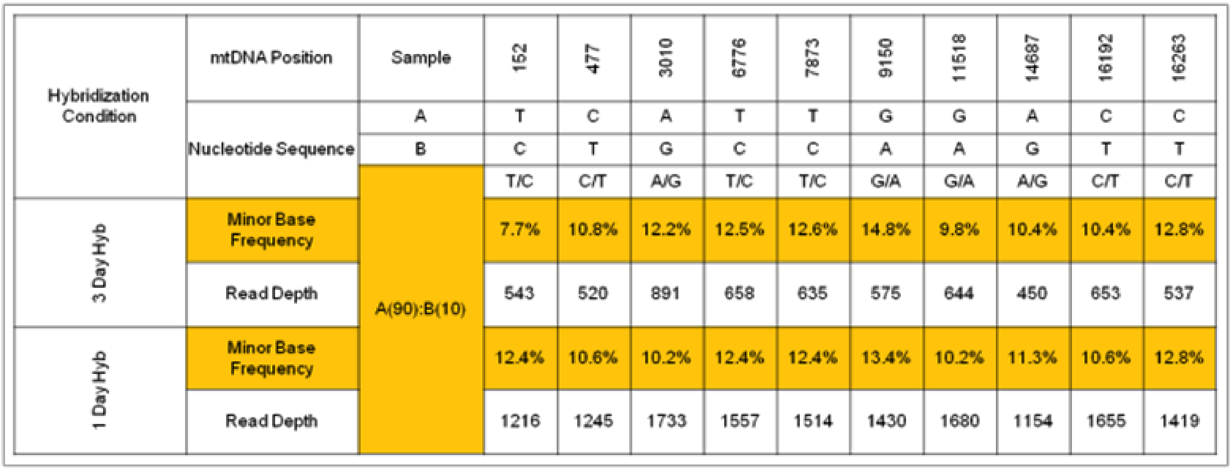
Figure 41.

Mixture Detection of 10% Minor Sequence



- **Successful detection of the minor sequence (10%) at all mixed base positions**
- **Analysis of additional mixed DNA samples needed**

Table 22. 10% Mixture Results



Hybridization Condition	mtDNA Position	Sample	152	477	3010	6776	7873	9150	11518	14687	16192	16263
	Nucleotide Sequence	A	T	C	A	T	T	G	G	A	C	C
		B	C	T	G	C	C	A	A	G	T	T
			T/C	C/T	A/G	T/C	T/C	G/A	G/A	A/G	C/T	C/T
3 Day Hyb	Minor Base Frequency	A(90):B(10)	7.7%	10.8%	12.2%	12.5%	12.6%	14.8%	9.8%	10.4%	10.4%	12.8%
	Read Depth		543	520	891	658	635	575	644	450	653	537
1 Day Hyb	Minor Base Frequency		12.4%	10.6%	10.2%	12.4%	12.4%	13.4%	10.2%	11.3%	10.6%	12.8%
	Read Depth		1216	1245	1733	1557	1514	1430	1680	1154	1655	1419

a. Proof-of-Concept Experiments for Covaris Fragmented Library Preparation

We also tested proof-of-concept for capture and 454 NGS sequencing of mechanically sheared DNA prepared using the optimized Covaris protocol for DNA library fragmentation. This SeqCap experiment included 100 ng and 1 ng Sample A to allow for a direct comparison of the mechanical and enzymatic fragmentation methods, 100 pg of Sample A to expand the sensitivity study, and a 5% and 1% mixture of Samples A/B to expand the mixture detection study.

This set of experiments showed proof-of-concept for using Covaris mechanical shearing for DNA fragmentation for DNA library preparation and subsequent capture and NGS sequencing of the entire mitochondrial genome. The optimized system using Covaris mechanical shearing resulted in 100% coverage of the mitochondrial genome with similar percent on target capture rate (~79%) compared to the enzymatic fragmentation method. However, the average read length was lower ~280 bp for the Covaris fragmentation method compared to ~440 bp for the enzymatic shearing method using Fragmentase (See Table 23 below). Differences in the average read length are likely directly attributed to the difference in the fragmentation method. To increase the average read length using the mechanical shearing method, the Covaris protocol can be further optimized to increase the average target fragment length.

Table 23. Sequencing statistics comparing the Enzymatic fragmentation vs. Covaris® mechanical shearing.

	Percent on Target	Average Read Length	Percent Coverage
Fragmentase (avg. 100ng & 1ng)	84.12%	437.5	100%
Covaris (avg 100ng & 1ng)	78.56%	279.5	100%

Further sensitivity of the assay was also demonstrated by successful capture and sequencing of Covaris prepared libraries with as low as 100 pg of starting DNA. The distribution of the number of reads across the mitochondrial genome is shown in Figure 42 below and the run statistics are presented in the Table 24 below. For each of the 3 concentrations tested (100ng, 1 ng and 100 pg), 100% sequence coverage was obtained with ~79% on target rate. Further experiments with lower DNA starting amounts would need to be conducted to determine the lower limit of sensitivity.

Figure 42. Distribution of Sequence Reads Across the Mitochondrial Genome

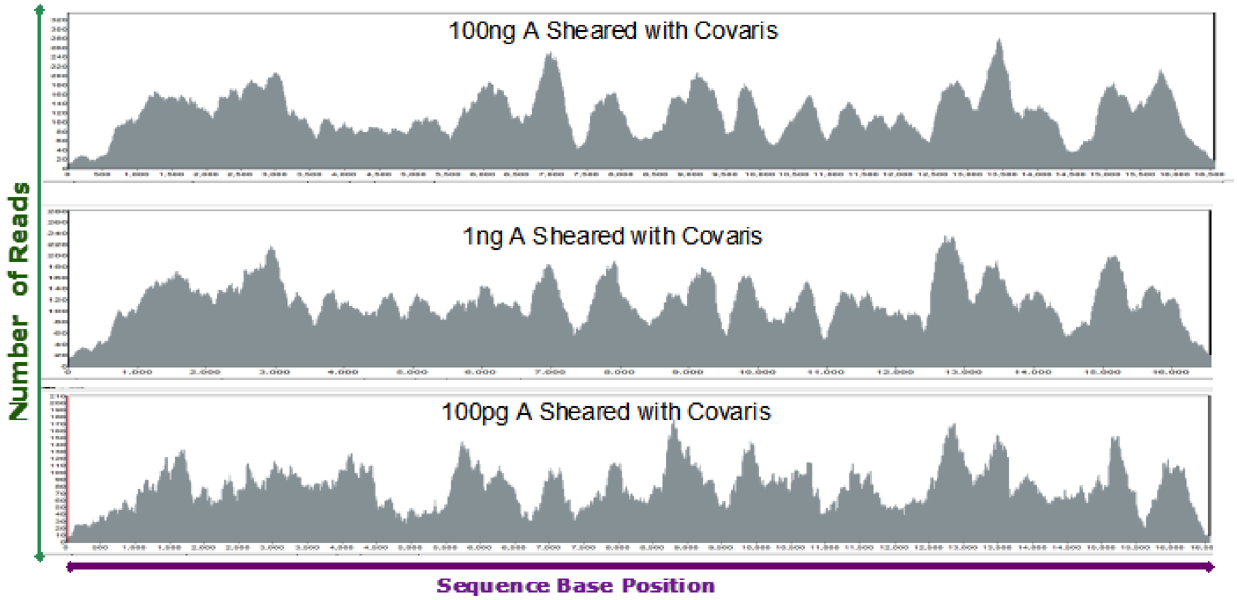
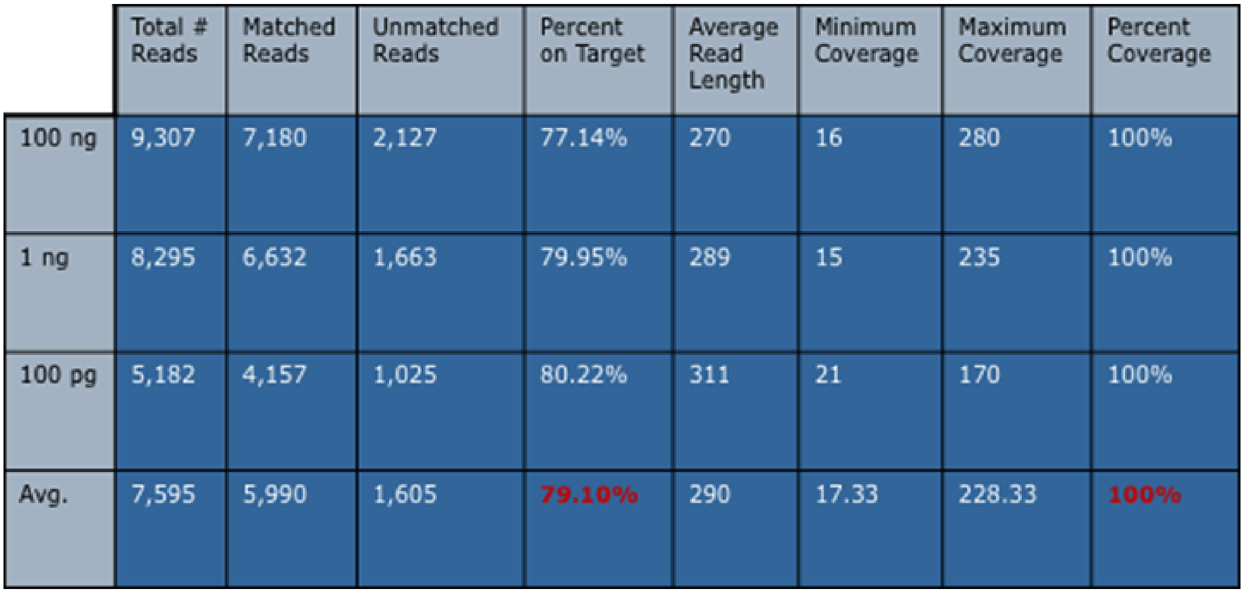


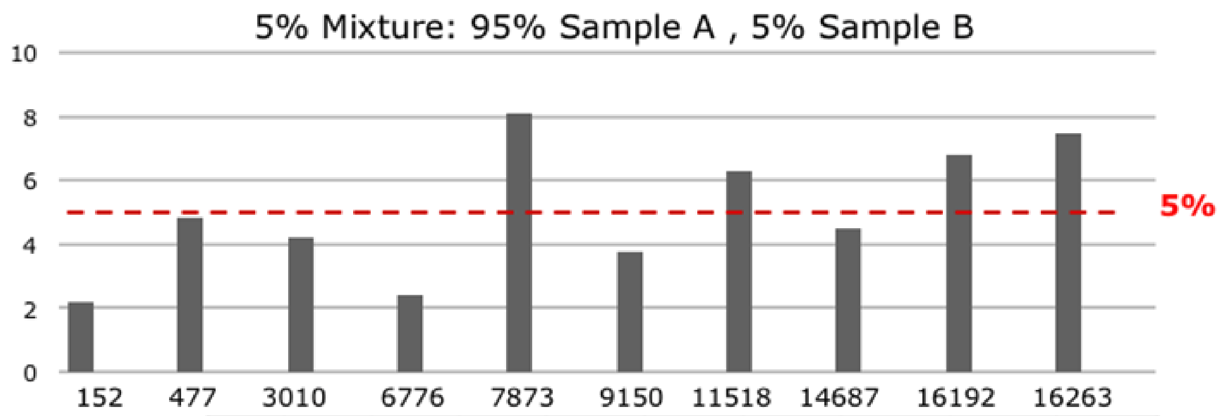
Table 24. Sequence Read Values



	Total # Reads	Matched Reads	Unmatched Reads	Percent on Target	Average Read Length	Minimum Coverage	Maximum Coverage	Percent Coverage
100 ng	9,307	7,180	2,127	77.14%	270	16	280	100%
1 ng	8,295	6,632	1,663	79.95%	289	15	235	100%
100 pg	5,182	4,157	1,025	80.22%	311	21	170	100%
Avg.	7,595	5,990	1,605	79.10%	290	17.33	228.33	100%

Additionally, further sensitivity for detecting minor components in a mixture was demonstrated by successful detection of each of the minor bases at each of the mixed base positions of a 5% mixture. Results are graphed in the Figure 43 below. A low coverage was obtained for this sample for this run. The minor base frequency would be expected to be less variable with a greater read depth, but even with the low read depth, all minor bases were detected and were within 3% of the expected 5% frequency. A 1% mixture was also, tested but the average read depth for this sample was ~100 base coverage, which was too low for detecting a 1% mixture. Increased read depth would be required (minimum of 1000 fold coverage per base) for detection of such a low level mixture. Further experiments with higher fold coverage would need to be conducted to determine the lower limit of mixture detection sensitivity.

Figure 43 Probe Capture 5% Mixture Sensitivity Study



	152	477	3010	6776	7873	9150	11518	14687	16192	16263
Sample A	T	C	A	T	T	G	G	A	C	C
Sample B	C	T	G	C	C	A	A	G	T	T
Sample B %	2.17	4.8	4.23	2.41	8.07	3.78	6.30	4.5	6.78	7.49
Coverage	46	41	189	135	161	185	132	95	75	54

C. Development and Testing of 454 STR mini-Plex fusion Primer and Universal Primer Multiplex PCR Assay

1. Research Design and Methods

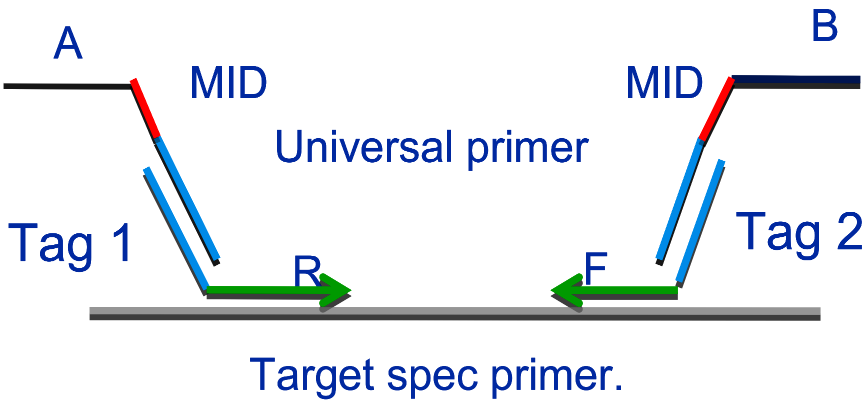
i. Develop and Optimize "Front-End" 454 Assay Targeting STR Markers

Two assays targeting CODIS CORE STR markers with amelogenin gender marker using 454 Next Generation Sequencing technology were developed using mini-STR primer sets[2]. Two designs were made in an effort to incorporate 454 sequencing primer and 10-bp barcode MID tags to the targeted STR loci while minimizing the number of primers needed.

a. Amplification of CODIS STRs using 2 Step PCR using M13 linker and MID Tags

A STR assay design using 454 Next Generation Sequencing technology was developed using M13 universal primer sets targeting CODIS STR loci as an approach to minimize the number of primer sets required for pooling multiple samples per sequencing run. In this design the samples are first amplified by inner primers consisting of STR locus target specific sequence (green) and a universal M13 linker sequence (blue) (see Figure 44). The inner products are then amplified with outer primers consisting of complementary M13 linker sequence (blue), a 10 bp MID tag (red), and a 25 bp 454 specific sequence (black). We have developed a universal set of CODIS STR loci specific inner primers with M13 linkers as well as eight sets of 454 MID tagged outer primers with universal M13 linkers.

Figure 44. 454 STR Assay Approach 2: Universal Primer (M13).



Each primer set was first tested in single-plex and then optimal conditions were determined for a multiplex. The following amplification parameters were optimized: reaction mix composition; annealing temperature; denaturation, annealing, and extension times; and primer concentrations. Two different compositions of reactions mix were tested; one was prepared following protocols provided by Butler et al. (2003) and the other using a composition developed in our lab for the mtDNA assay. It was concluded that the composition developed for the mtDNA assay in our lab produced more robust results as higher yields were observed for all STR loci. To determine the optimal annealing temperature, annealing temperatures ranging from 53-57°C were tested for inner amplification and annealing temperatures ranging from 54 - 67°C were tested for outer amplification. Further, two step annealing temperatures of 54°C increasing to 63°C for inner amplification was also tried to account for the earlier stages of amplification and the increased length of primers with M13 sequences which are not complementary to the target sequence. Both 15 seconds and 30 seconds were tested for each of the cycled steps (denaturation, annealing, and extension) and 30 seconds showed higher yields. Primer concentrations ranging from 0.05µM to 0.2µM and from 0.2µM to 1.8µM for inner and outer primers, respectively, were tested. The optimized PCR parameters for the inner and outer amplification are summarized below (see Table 25).

Table 25. Final Parameters for Inner and Outer Amplification of STR Products Using Universal Primer Approach.

Inner Amplification			Outer Amplification		
No. of Cycles	Temperature	Length	No. of Cycles	Temperature	Length
1	95°C	10 min	1	95°C	10 min
10	95°C	30 sec	15	95°C	30 sec
	54°C	30 sec		54°C	30 sec
	72°C	30 sec		72°C	30 sec
20	95°C	30 sec	1	72°C	7 min
	63°C	30 sec		4°C	Forever
	72°C	30 sec			
1	72°C	7 min			
	4°C	Forever			

During the optimization, all conditions tested resulted in primer dimer with the 13 plex. A series of experiments were conducted to identify the STR primers that produced dimers in the 13 plex and these sets were removed from the multiplex, resulting in a 9 plex and 4 plex. Using a control DNA sample, inner amplicons and outer amplicons were prepared following the parameters provided above for the 4 plex, 9 plex, and 13 plex. For the inner amplification, 4 plex primers produced minimal amount of primer dimer while 9 plex and 13 plex primers produced dimers in amounts comparable to that of the STR products (see Figure 45a). After the outer amplification, however, minimal amounts of primer dimer were observed for the 4 plex and 9 plex primer sets while 13 plex primers produced high amount of excess dimers (see Figure 45b).

Figure 45a. Inner Products of Universal Primer Approach with 4 Plex, 9 Plex, and 13 Plex

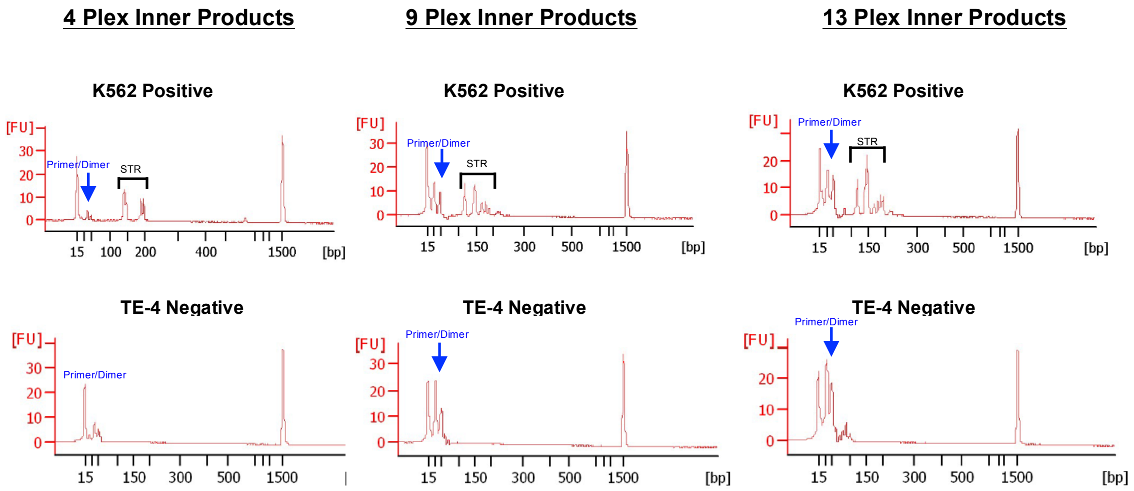
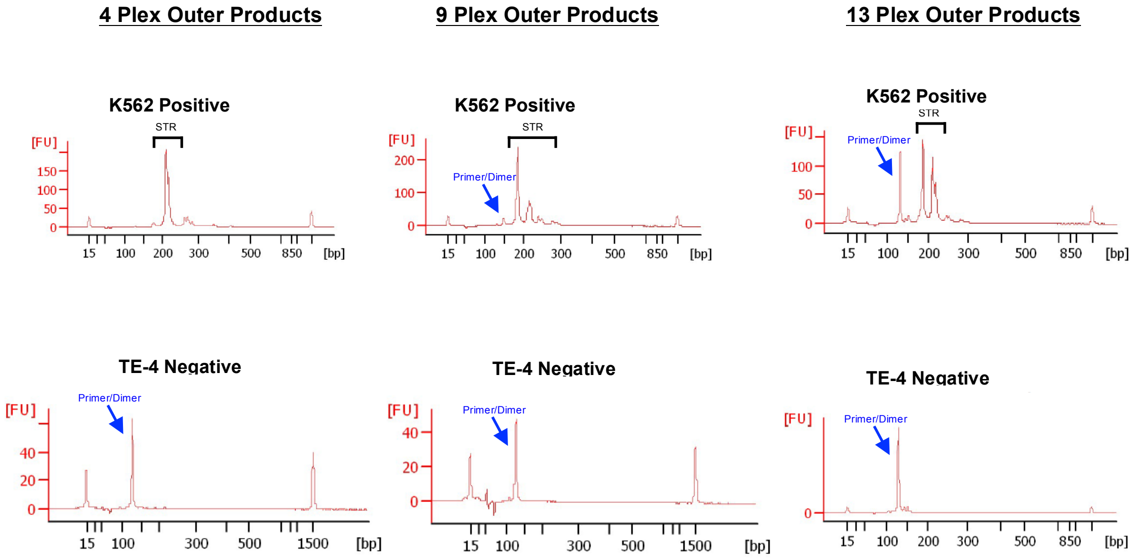
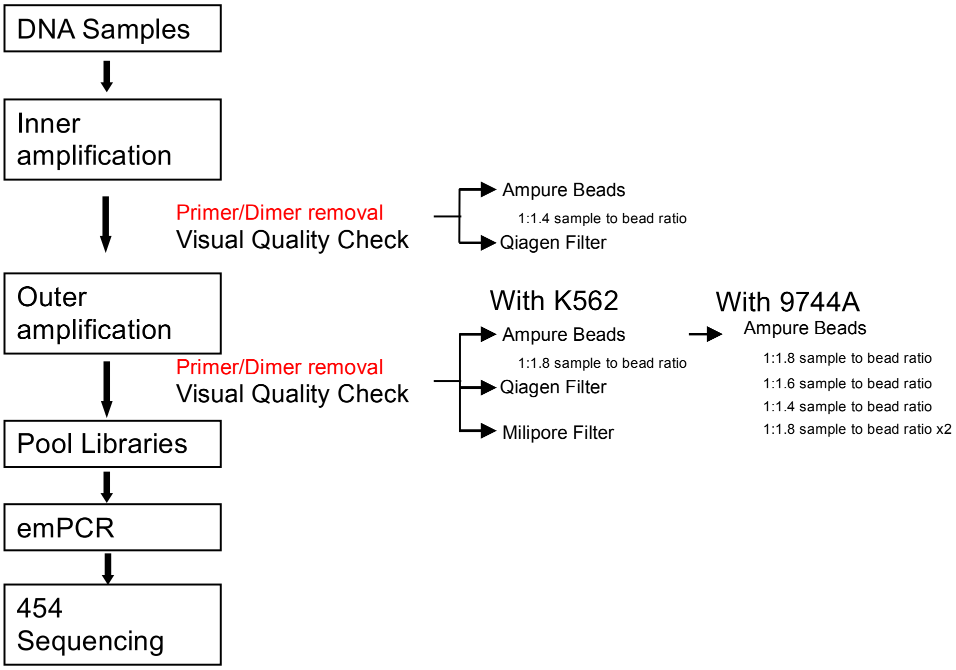


Figure 45b. Outer Products of Universal Primer Approach with 4 Plex, 9 Plex, and 13 Plex.



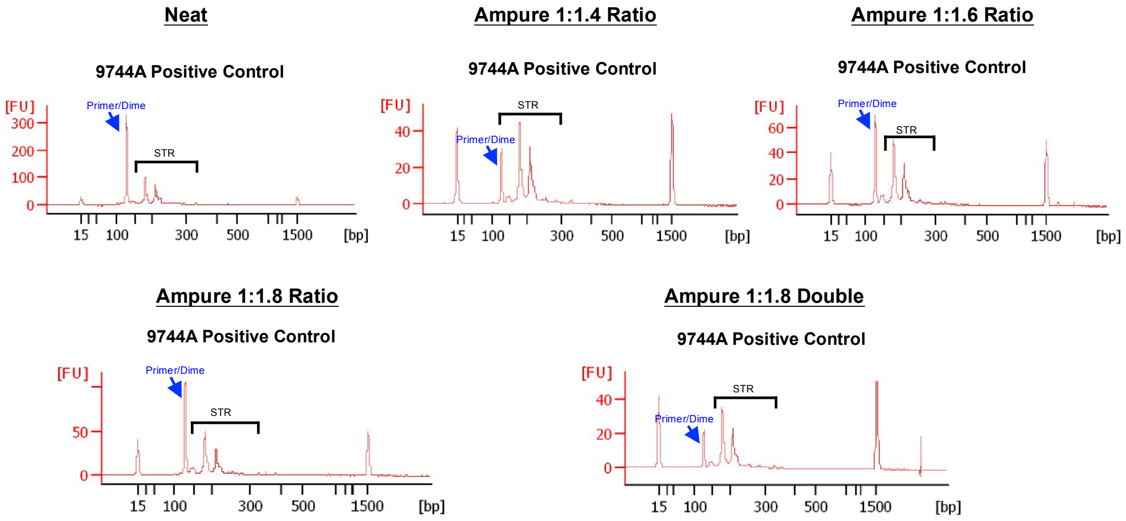
In order to remove the primer dimers produced, several different purification methods were explored including Agencourt AMPure XP beads (Life Technologies, Carlsbad), QIAmp Mini Elute columns (Qiagen, Valencia), and Millipore filters. To determine the optimal sample to bead ratio to effectively remove the dimer artifacts while minimizing STR product loss, the following different ratios were tested to purify the outer amplicon of a control DNA sample: 1:1.4, 1:1.6, 1:1.8, and 1:1.8 ratio 2X purification. A flowchart of the library preparation steps including the different purification methods explored is presented below (see Figure 46).

Figure 46. Flowchart of Workflow for STR 454 Assay with Universal Primer Approach and the Different Purification Methods Explored.



The results showed that Agencourt AMPure XP beads removed the dimers from the product the most effectively compared to filtration columns. Furthermore, the bioanalyzer results showed that the 1:1.4 ratio and 1:1.8 ratio double purification resulted in the maximal removal of dimers as the primer dimer peak heights are reduced under 1:1.4 ratio and 1:1.8 2X conditions. However, the STR product peak heights were sustained (>40 FU) with 1:1.4 ratio while they were lowered in 1:1.8 2X condition (see Figure 47). Based on these results, the 1:1.4 ratio was selected as the optimal ratio for purification and was carried out for the inner primer dimer removal experiments.

Figure 47. Varying Ampure Ratio Experiment on Outer Products.



Neat

Ampure 1:1.4 Ratio

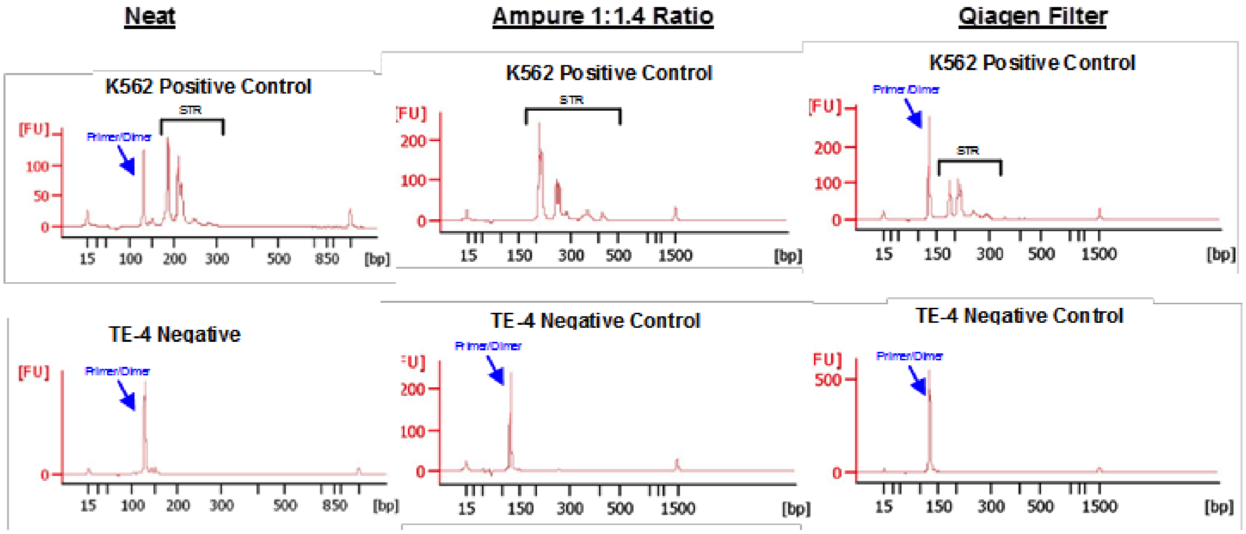
Ampure 1:1.6 Ratio

Ampure 1:1.8 Ratio

Ampure 1:1.8 Double

As another approach to minimize the production of primer dimers, we tested 1:1.4 DNA sample to AMPure bead ratio and size selection filtration units for removal of excess primers and primer dimers after inner amplification. The bioanalyzer results showed that the primer dimers were effectively removed with the optimized sample to AMPure bead ratio of 1:1.4 after the inner amplification but not with the filtration units (see Figure 48). The peak corresponding to primer dimer is not present for the AMPure purified inner products after outer primer amplification. Based on these results, the AMPure clean-up step (1:1.4 sample to AMPure bead ratio) after inner product amplification appeared to optimally remove the dimers without STR product loss for the 13 plex. However, to determine successful amplification and retention of all 13 STR loci, next generation sequencing data was necessary. A subset of the STR libraries prepared with different purification methods were sequenced using the 454 GS technology.

Figure 48. 13 Plex Outer Products Purified with Different Methods After Inner Amplification.



1. Purification Method Comparison – M13 Universal Primer Approach

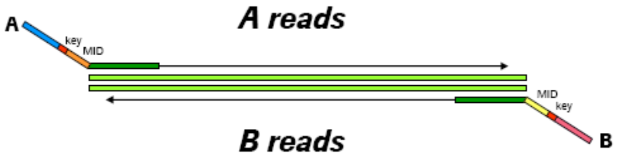
A control DNA sample K562 was amplified for CODIS STR loci using 4-plex, 9-plex, and 13-plex inner STR primers. Each multiplex PCR products were treated with five different purification methods as following: neat (<70 bp filter removal), AMPure purification with DNA to AMPure volumetric ratio1:1.4 after the inner amplification, AMPure purification with DNA to AMPure volumetric ratio 1:1.8 after the outer amplification, filter purification (removal of <100 bp) after inner amplification, and finally filter purification (removal of <100 bp) after outer amplification.

Several loci drop outs were observed using the NextGENe software (SoftGenetics, LLC, PA). On average 12 loci were observed for 13 plex, 8 loci were detected for 9 plex, and 4 loci were detected for 4 plex. In order to confirm the source of the loci drop-out, a sequencing experiment of running the multiplex products with single plex products was conducted.

b. Amplification of CODIS STRs using 454 MID tagged Fusion Primers

A second STR assay design using Multiplex Identifiler (MID) tagged fusion primers, targeting 13 CODIS STRs and amelogenin marker were developed using 454 Next Generation Sequencing technology. The primers consist of a 454 library key, 10 bp MID sequence, and STR locus specific sequence, similar to the mitochondrial HVI/HVII fusion primer design (Figure 49). The mini STR primer sequences published by Butler et al. (2003) were used.

Figure 49. 454 STR Assay Approach 1: Fusion Primer.



Forward primer (Primer A-Key):

5' -CGTATCGCCTCCCTCGCGCCATCAG-{MID}-{template-specific-sequence}-3'

Reverse primer (Primer B-Key):

5' -CTATGCGCCTTGCCAGCCCGCTCAG-{MID}-{template-specific-sequence}-3'

A range of annealing temperatures (55°C, 57°C, 59°C, and 61°C) were re-tested as part of the optimization of the PCR parameter for the PCR amplification. Annealing temperature optimization was conducted by single-plex amplification of each locus separately before optimizing the 14 plex PCR. While there was no significant change in level of primer dimer in each annealing temperature, 57°C was observed to produce the highest product yield based on bioanalyzer results. Additionally, two step annealing temperatures of 54°C to 61°C and 54°C and 63°C were tested and compared to one-step 57°C PCR amplification in order to account for the delayed complete amplification to the 454 library sequence and MID tags. No significant difference was observed in product yield or the level of primer dimmer between the two step amplification to one step 57 °C annealing step. Different PCR parameters tested is summarized in the Figure 50 below.

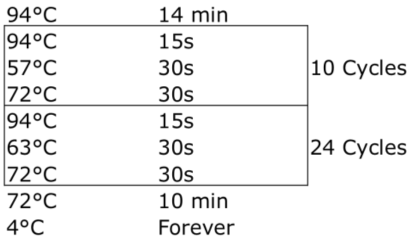
Figure 50. Different PCR Parameters Tested Using 454 Fusion Primers Targeting CODIS STRs with a) Single Annealing Temperature and b) Two Step Amplification

a).

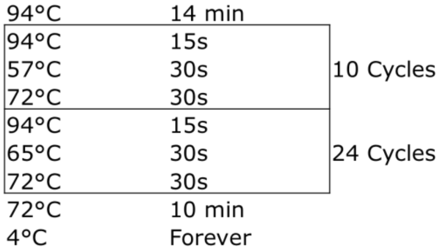
94°C	14 min	
94°C	15s	
X°C	30s	34 Cycles
72°C	30s	
72°C	10 min	
4°C	Forever	

X = 55, 57, 59, 61

b)



94°C	14 min	
94°C	15s	
57°C	30s	10 Cycles
72°C	30s	
94°C	15s	
63°C	30s	24 Cycles
72°C	30s	
72°C	10 min	
4°C	Forever	



94°C	14 min	
94°C	15s	
57°C	30s	10 Cycles
72°C	30s	
94°C	15s	
65°C	30s	24 Cycles
72°C	30s	
72°C	10 min	
4°C	Forever	

Once the PCR parameter was optimized for each locus, the 14 plex amplification parameters were optimized by testing the annealing temperatures 55°C ,57°C , and 59°C . Similar to the single plex amplification, 57°C annealing temperature resulted in the highest yield of product, and no significant difference was observed in level of primer dimmer. The two step amplification conditions were once again compared to the 57°C single step amplification for the 14 plex amplification. No significant difference in the level of primer dimmer and product yield was observed between the two step amplification condition and the single amplification condition. The final, optimized PCR parameter is summarized below in Figure 51 below.

Figure 51. Final PCR Parameter for 14 Plex Amplification of CORE STR Loci using Fusion Primers

94°C	14 min	
94°C	15s	
		34
57°C	30s	Cycles
72°C	30s	
72°C	10 min	
4°C	∞	

Primer concentration for amplification was optimized in an effort to minimize the level of primer dimer. Varying primer concentrations of 0.3µM, 0.6 µM and 0.9µM were tested. 0.3µM primer concentration produced the least amount of primer dimer without compromising the product yield based on the bioanalyzer results. The PCR reaction mix composition was kept constant as the M13 linker model.

1) Purification Method Comparison – 454 Fusion Primer Approach

To compare the efficiency of different purification method for the 454 STR fusion primer assay, the 14 loci amplified from the control DNA K562 using the 14-plex primer blend was treated with three different purification method as following: neat (filter purification <70bp removal for excess primers), AMPure purification with DNA:AMPure volumetric ratio of 1:1.4, and filter purification removing <100 bp fragments. Additionally, four loci amplified in single-plex were sequenced along with the 14 plex products to confirm the successful amplification of the loci. Loci vWA, FGA, and D21S11 resulted in the least amount of product yield from the PCR amplification compared to the other targeted STR loci based on the bioanalyzer result. Locus TPOX was sequenced for the direct comparison for the three mentioned loci as it was observed that the highest PCR product yield was observed using the primers targeting the TPOX.

2) Single Plex vs 14 Plex Amplification Comparison – M13 Universal Primer Approach and 454 Fusion Primer Approach

In order to confirm the locus drop-out observed in the purification method experiments in both fusion primer and M13 linker approach, a control DNA K562 was amplified targeting each 13 STR loci and amelogenin separately in single-plex as well as in 14-plex using both fusion primer and universal M13 primer designs. An additional locus, vWA was added to form the 14-plex primer blend for the M13 universal primer design to mirror the 14-plex design of the fusion primer approach.

For both single-plex PCR products and 14 plex products, 1:1.4 DNA:Ampure clean-up step was used after the amplification as it was finalized to be the optimal primer dimer removal method for the fusion primer assay. For the universal primer design, 1:1.4 DNA:Ampure clean-up was performed after inner amplification of each PCR product. Both primer model libraries were pooled and ran through a filter for removing < 70 bp fragments before the emPCR step. All 14 loci of interest were detected in both samples amplified multiplexed or single plexed using fusion primer or M13 linker primer with varying frequencies when analyzed using NextGENe software. The detected loci from each amplification method are presented with corresponding frequencies in the Figure 52 below.

Figure 52. Locus Frequency Distribution for 14 Loci Amplified.

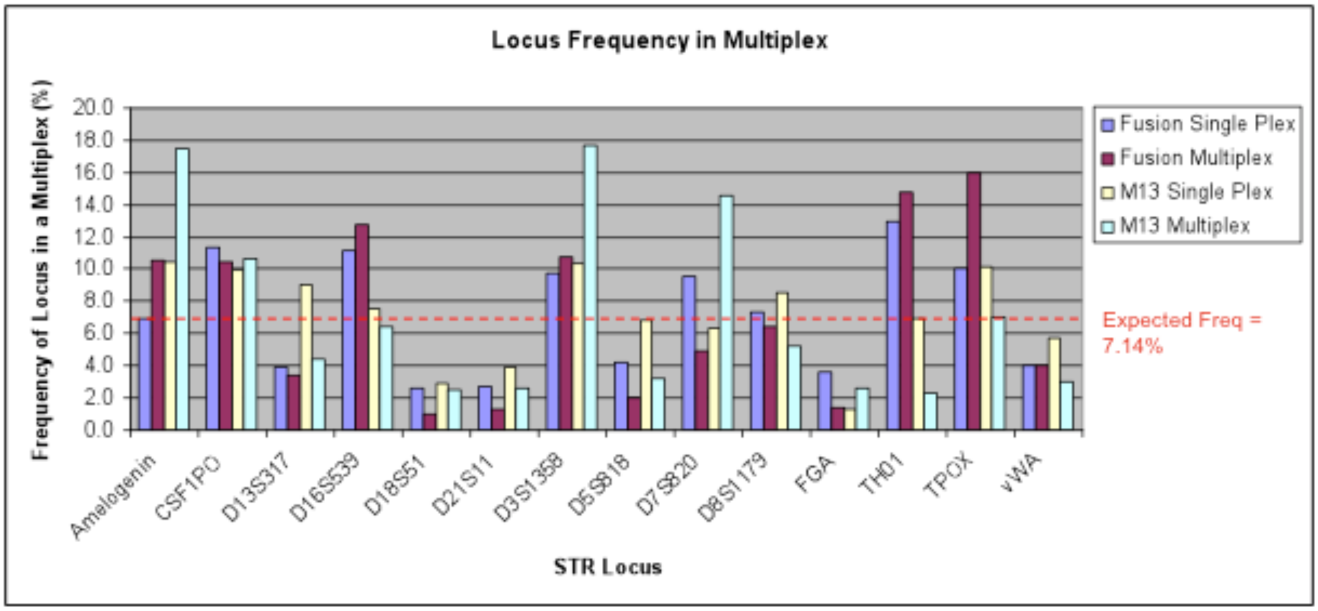


Figure 52. All 13 CODIS STR loci with amelogenin marker were detected in samples amplified multiplexed and in single plex. All loci were detected regardless of primer design, (i.e. fusion primer approach or M13 universal linker approach) but the frequency of each locus present in the run was not balanced.

2. Results

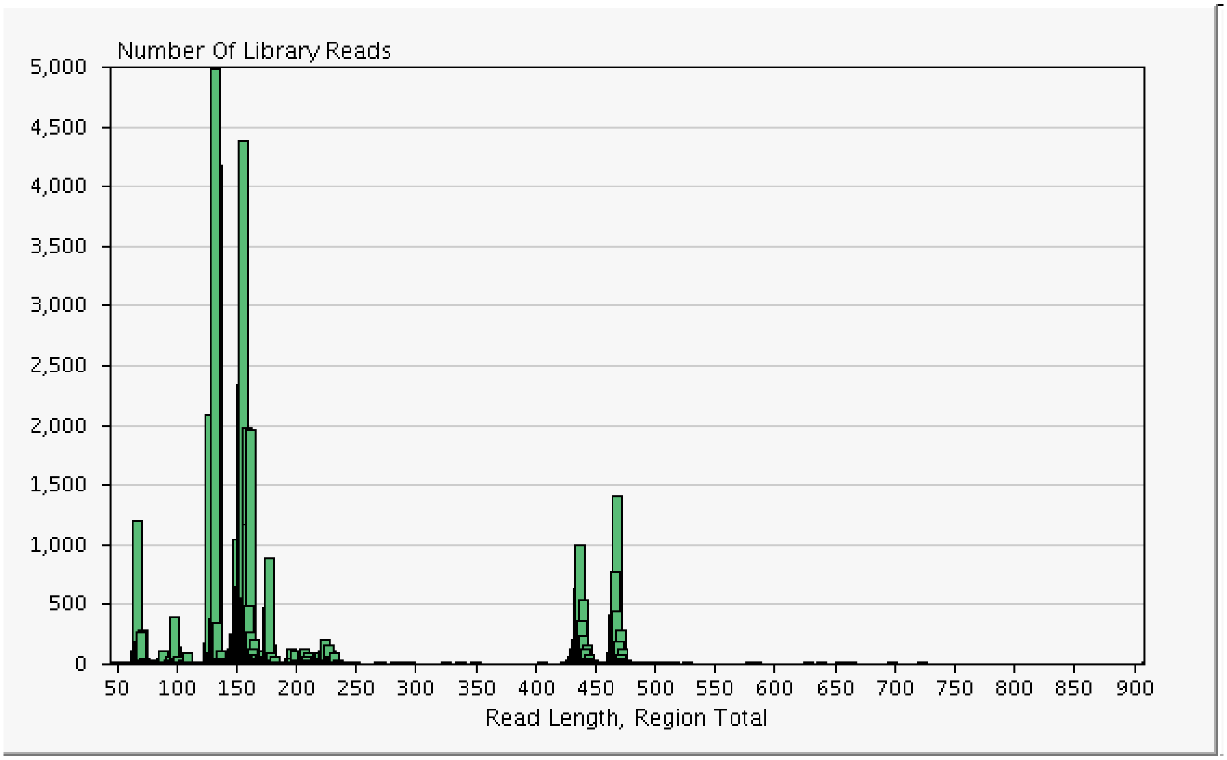
i. Sensitivity Study

To determine the sensitivity of the 454 STR M13 linker assay, control DNA samples (k562) varying in initial input amounts from 0.1pg – 20 ng were amplified using the 454 9 plex mini-STR assay. Samples below 50 pg clearly showed only partial amplification of the 9 STRs based on gel electrophoresis while no drop out was apparent for samples above 500 pg. Based on these results, samples amplified with the initial input of 50 pg and 500 pg were selected, purified using the AMPure beads and sequenced using the 454 GS Jr. In this run, ~40,000 of the 50,000 reads corresponded to STR sequence products. Only sequence reads for three of the nine markers were successfully assigned using the 454 AVA software and was determined to not be an ideal sequence alignment software program for analysis of STRs. However, all 9 loci (CSF1PO, D16S59, D5S818, THO1, TPOX, D21S11, D18S51, D8S1179, and FGA) that were expected to be present were detected by the alignment software NextGENe version 2.4.0 for the limited samples (500 pg and 50 pg). The NextGENe software using an algorithm setting specifically optimized for short repeats, ideal for analyzing STRs. We have worked closely with SoftGenetics software to modify NextGENe to make further improvements. Modifications to the Software developed under our collaboration are discussed in more detail in the software section below. Data analysis for the STR sensitivity study is ongoing, and the software modification for the STR analysis is communicated on regular basis with Soft Genetics, LLC.

ii. Combined System: STR and HVI/HVII Mitochondrial Markers

One advantage of the NGS technologies is that both mtDNA and STRs can be run on the same platform and we initially proposed to sequence both mtDNA and mini-STRs in a single run using the 454 DNA sequencing technology. To test if both mtDNA and STRs could be successfully sequenced in a single 454 run, a sequencing experiment was conducted where STR products amplified using the M13 universal linker system and HVI/HVII products amplified using the 454 HVI/HVII assay were prepared together in a single emPCR and 454 sequencing run. The distribution of reads lengths obtained for the STRs and mtDNA HVI/HVII regions from the combined run are as shown in the Figure 53 below.

Figure 53. Read Length Distribution for the STR and HVI/HVII Combined Sequencing Run



Results show that both mtDNA and STR were successfully sequenced in a single run as reads for both mtDNA and STR targets were obtained. Reads in the 400-500 bp range correspond to the mtDNA HVI/HVII products. Across the run, on average ~1000 reads for each region of the mtDNA sequenced was obtained. Reads in the 75-250 bp range correspond to STR products while reads less than 75 are likely primer dimer products or truncated products. However, a high percentage of short reads was observed in the run compared to the other 454 sequencing runs conducted in the project, indicating conditions were less optimal. It is advised by 454 Roche that in order to maximize the sequencing yield and sample preparation efficiency, pooled amplicons do not span a significant length range that is wider than 150 bp whether or not MIDs are used to tag them. 454 identifies that pooling mixed-length Amplicons through emPCR amplification and sequencing can have negative effects. The DNA library beads derived from amplicons of different length can result in a different amount of amplified target and, consequently, in a wider signal distribution, potentially lowering the sequencing run yield. Additionally, if two templates of different sizes are present in any given emulsion droplet, the shorter fragment will amplify preferentially over the longer one. This will result in either an increase in mixed reads that will be discarded by the data processing filters, or an over representation of the shorter fragments in the sequencing results. Therefore, sequencing amplicons with great size difference would introduce a potential sequencing bias and a reduction in run yield.

In order for the amplicons to be processed together in a single 454 sequencing run, emPCR conditions would need to be modified or STR or mtDNA primers redesigned to target similar target sizes to optimally run the mini-STRs and HVI/HVII products in a single emPCR. With the current 454 HVI/HVII assay and STR PCR systems, it was concluded that the mtDNA and STR amplicons should be prepared in separate emPCRs and then pooled in a single 454 Run or run separately in order to yield optimal sequencing results.

D. Next Generation Sequence Alignment and Analysis Software Development

1. SoftGenetics NextGENe Software

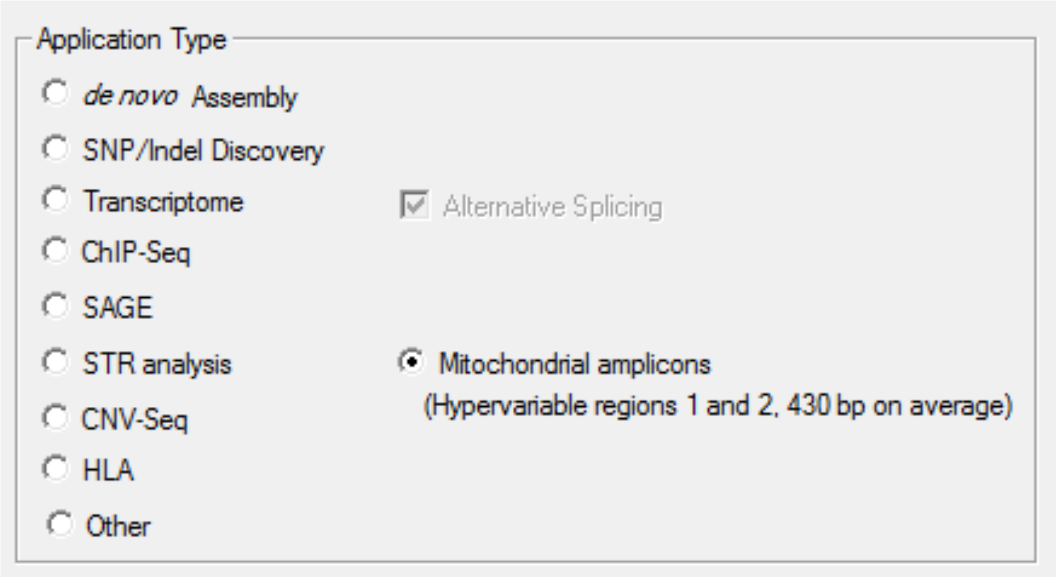
i. Overview

Collaboration was established with Soft Genetics, LLC, to customize and add modifications to improve their commercially available NextGENe next generation sequencing data analysis software for analysis of mtDNA and STR data. Several commercial softwares were explored including DNAStaR, GeneCodes, 454 AVA and NextGENe from Soft Genetics for analysis of mtDNA and STR NGS data. The NextGENe software was chosen for further consideration based on the STR specific algorithm developed for analysis of forensic STRs. Other softwares that were investigated had difficulty aligning repeat regions due to the higher gap penalties assigned to IN/DEL mutations compared to SNPs. The NextGENe forensic STR sequence alignment algorithm was optimized for alignment of STR repeats, assigning lower gap penalty for IN/DEL sizes corresponding to the repeat size (4 bp). The modified algorithm results in optimal alignment of the STR repeats. We are currently collaborating with Soft Genetics to further modify the software to include additional filters for homopolymer sequencing errors, STR mixture analysis and reporting as well as mtDNA haplotype identification, mixture analysis and reporting. The .sff files of 454 sequencing data generated using HVI/HVII assay and both universal and fusion primer STR assays were provided to Soft Genetics, LLC to develop modified software for beta testing of different versions of NextGENe software modified throughout the granting period. The modified software was tested extensively internally as well as demonstrated via a webinar and hands on training as part of the CCI course at the California Department of Justice. NextGENe software is a user-friendly software that does not require scripting and is a versatile program that can handle analysis of data generated by different platforms including 454, Illumina, Ion Torrent, and SOLiD. Further, the program can be used for multiple applications including de novo assembly, targeted sequencing, and forensic science applications.

ii. mtDNA: HVI/HVII

A specific application option for mitochondrial amplicon analysis, customized for the HVI/HVII assay, was added to the software from this collaboration, and this modification of the software demonstrated in the screen shot of the set up screen for the software in Figure 54 below.

Figure 54.



Application Type

○ *de novo* Assembly

○ SNP/Indel Discovery

○ Transcriptome ☑ Alternative Splicing

○ ChIP-Seq

○ SAGE

○ STR analysis ● Mitochondrial amplicons (Hypervariable regions 1 and 2, 430 bp on average)

○ CNV-Seq

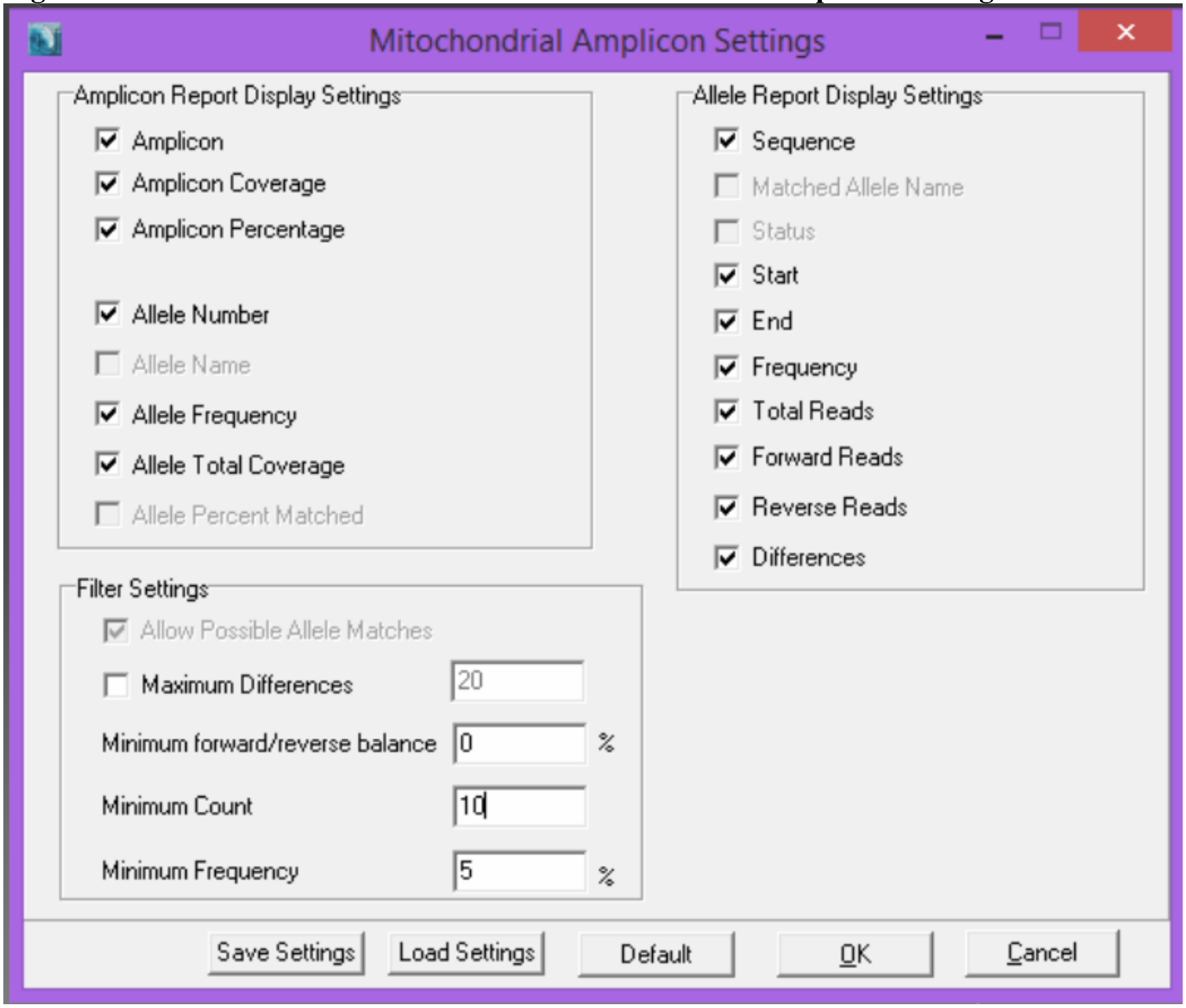
○ HLA

○ Other

Figure 54. Screen shot of NextGENe v.2.4.0 Set Up Screen. An option for mitochondrial amplicon analysis was added as one of the Application Type to the software. Identification of the application type of the alignment project is needed to determine the optimal alignment algorithm.

This mitochondrial amplicon analysis option allows for different customization for the analysis such as designating regions of interest within mitochondrial genome or adjusting the report and filter setting using the mitochondrial amplicon setting option shown in the Figure 55 below.

Figure 55. Screen shot of NextGENe v.2.4.0 Mitochondrial Amplicon Settings Screen.



Amplicon Report Display Settings

☑ Amplicon
☑ Amplicon Coverage
☑ Amplicon Percentage

☑ Allele Number
□ Allele Name
☑ Allele Frequency
☑ Allele Total Coverage
□ Allele Percent Matched

Filter Settings

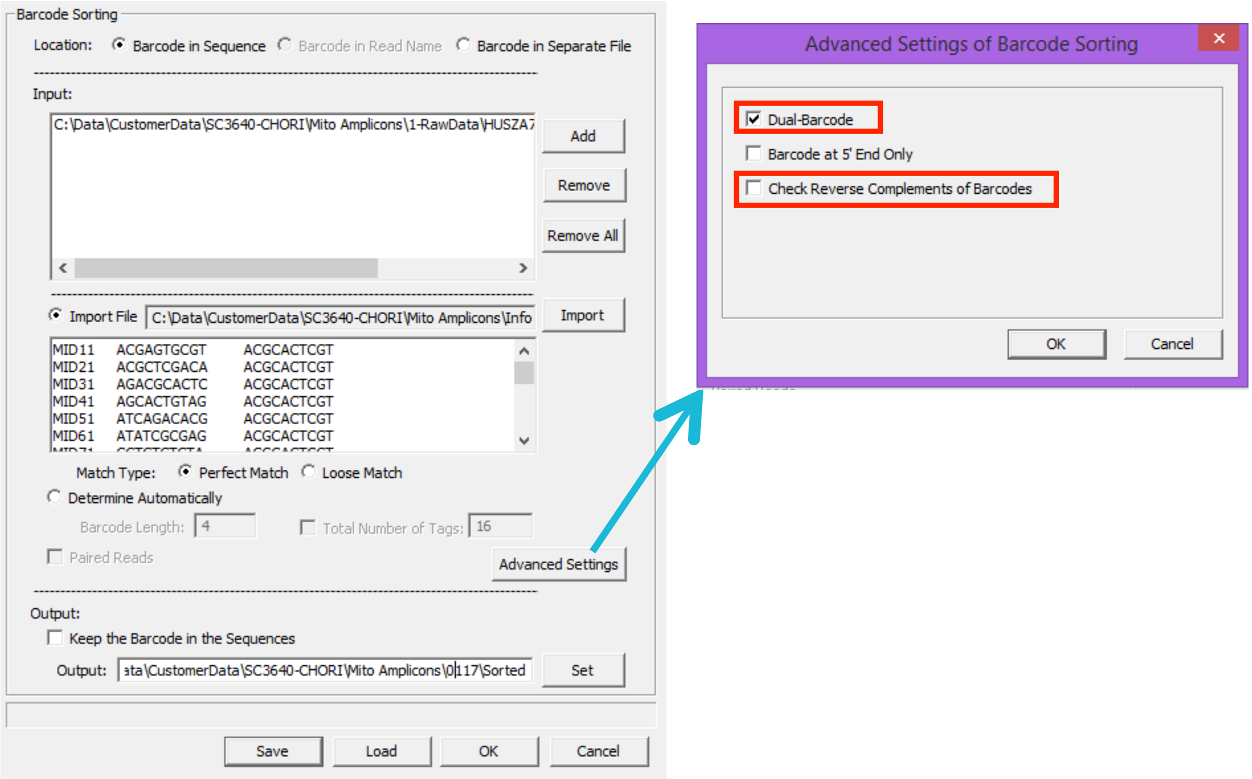
☑ Allow Possible Allele Matches
□ Maximum Differences 20
Minimum forward/reverse balance 0 %
Minimum Count 10
Minimum Frequency 5 %

Allele Report Display Settings

☑ Sequence
□ Matched Allele Name
□ Status
☑ Start
☑ End
☑ Frequency
☑ Total Reads
☑ Forward Reads
☑ Reverse Reads
☑ Differences

Prior to this collaboration, NextGENe software did not support parsing and sorting the next generation sequencing data based on both directions of reads, unable to correctly sort the data barcoded with MID using the combinatorial approach. Initially, the sequence data was sorted with barcode sequence in 5'-end of the read. It was suggested during the early stage of the collaboration to input the 3'-end MID barcode sequence reverse complemented into the barcode sorting reference file, but it was quickly recognized that certain reverse complemented MID sequence was identical to the forward direction of another MID tag, resulting in incorrect sorting of the data. Hence it was necessary for the software to recognize the directionality of the read as well as the MID tag sequence. This resulted in the options for dual-barcode and check Reverse Compliments of Barcodes was implemented into the NextGENe software to accommodate the combinatorial approach of barcoding samples.

Figure 56. Screen shot of NextGENe v.2.4.0 Barcode Sorting Tool Screen.



Barcode Sorting

Location: ● Barcode in Sequence ○ Barcode in Read Name ○ Barcode in Separate File

Input:

C:\Data\CustomerData\SC3640-CHORI\Mito Amplicons\1-RawData\HUSZA7

● Import File C:\Data\CustomerData\SC3640-CHORI\Mito Amplicons\Info

```
MID11   ACGAGTGCGT   ACGCACTCGT
MID21   ACGCTCGACA   ACGCACTCGT
MID31   AGACGCACTC   ACGCACTCGT
MID41   AGCACTGTAG   ACGCACTCGT
MID51   ATCAGACACG   ACGCACTCGT
MID61   ATATCGCGAG   ACGCACTCGT
```

Match Type: ● Perfect Match ○ Loose Match

○ Determine Automatically

Barcode Length: 4 □ Total Number of Tags: 16

□ Paired Reads

Output:

□ Keep the Barcode in the Sequences

Output: ata\CustomerData\SC3640-CHORI\Mito Amplicons\0117\Sorted

Advanced Settings of Barcode Sorting

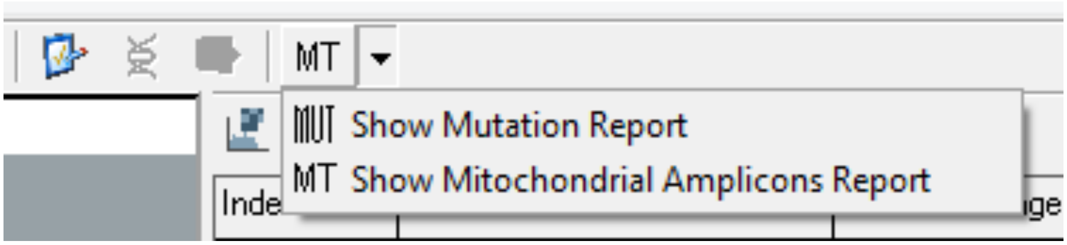
☑ Dual-Barcode

□ Barcode at 5' End Only

□ Check Reverse Complements of Barcodes

Further, a large effort was put into customizing the output and report format of the sequencing data. As of version 2.4.0, NextGENe is capable of generating Mitochondrial Amplicons report for each sample (Figure 57.) in an addition to a mutation report as shown in the screen shot Figure 57.

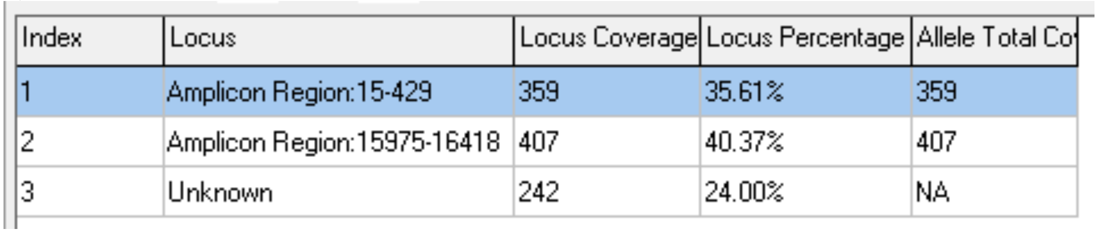
Figure 57. Screen Shot of NextGENe v.2.4.0 For Generation of Mitochondrial Amplicons Report



The mitochondrial Amplicon Report generates two separate reports per sample, a locus report and an allele report. A locus report shows the coverage of each HVI and HVII region amplicon. The allele report records different alleles, or unique sequences, within each amplicon, and the number of reads per observed sequence is recorded. An example of each report for the mitochondrial amplicon analysis is shown in Figure 58.

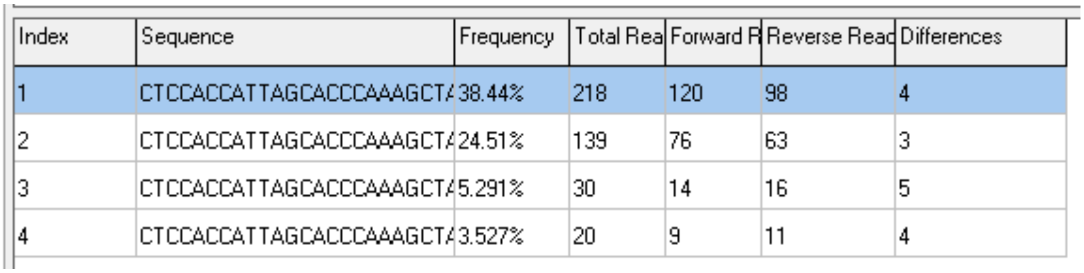
Figure 58. Screen Shot of a) Locus Report and b) Allele Report Generated Using NextGENe v.2.4.0 for Mitochondrial Amplicon Project

a)



Index	Locus	Locus Coverage	Locus Percentage	Allele Total Co
1	Amplicon Region:15-429	359	35.61%	359
2	Amplicon Region:15975-16418	407	40.37%	407
3	Unknown	242	24.00%	NA

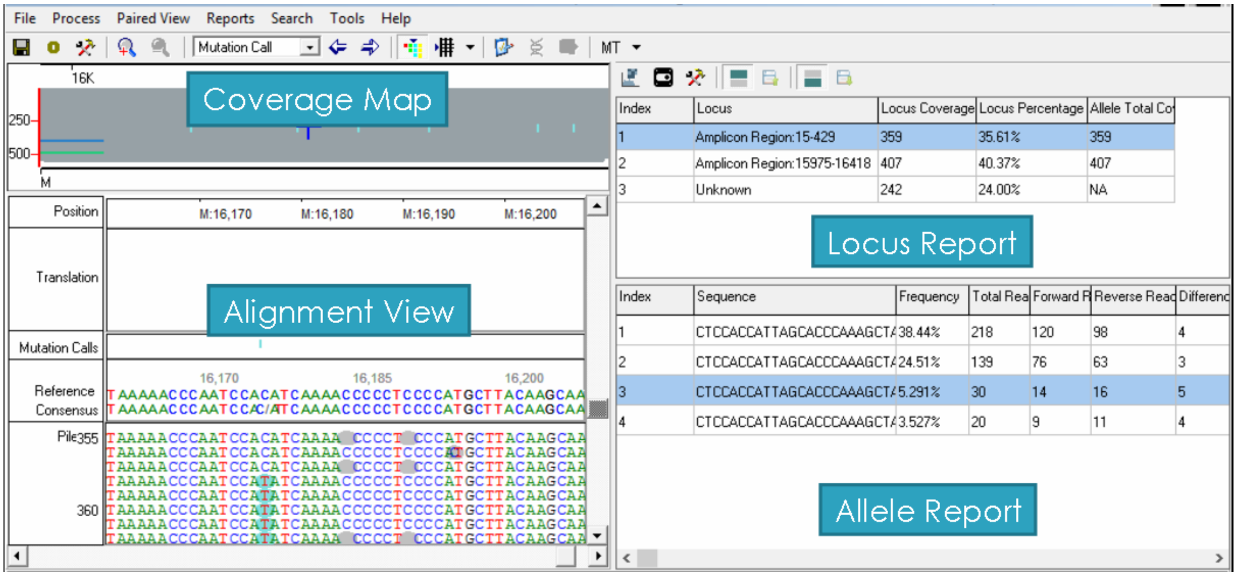
b)



Index	Sequence	Frequency	Total Rea	Forward R	Reverse Reac	Differences
1	CTCCACCATTAGCACCCAAAGCT/	38.44%	218	120	98	4
2	CTCCACCATTAGCACCCAAAGCT/	24.51%	139	76	63	3
3	CTCCACCATTAGCACCCAAAGCT/	5.291%	30	14	16	5
4	CTCCACCATTAGCACCCAAAGCT/	3.527%	20	9	11	4

The interface of the current version 2.4.0 of NextGENe software allows the user to view coverage map, alignment of sequences to the reference, locus report and allele report all in one window as demonstrated in the Figure 59 below.

Figure 59. Screen Shot of Interface of NextGENe v.2.4.0 for a Mitochondrial Amplicon Project. NextGENe versions 2.4.0 presents coverage map, alignment view, locus report, and allele report of a sample in one screen.



```
Reference  TAAAAACCCAATCCACATCAAAACCCCCTCCCCATGCTTACAAGCAA
Consensus  TAAAAACCCAATCCAC/ATCAAAACCCCCTCCCCATGCTTACAAGCAA
Pile355    TAAAAACCCAATCCACATCAAAA CCCCT CCCATGCTTACAAGCAA
           TAAAAACCCAATCCACATCAAAACCCCCTCCCCACGCTTACAAGCAA
           TAAAAACCCAATCCACATCAAAA CCCCT CCCATGCTTACAAGCAA
           TAAAAACCCAATCCATATCAAAACCCCCTCCCCATGCTTACAAGCAA
           TAAAAACCCAATCCATATCAAAACCCCCTCCCCATGCTTACAAGCAA
360        TAAAAACCCAATCCATATCAAAACCCCCTCCCCATGCTTACAAGCAA
           TAAAAACCCAATCCATATCAAAACCCCCTCCCCATGCTTACAAGCAA
           TAAAAACCCAATCCATATCAAAA CCCCT CCCATGCTTACAAGCAA
```

Index	Locus	Locus Coverage	Locus Percentage	Allele Total Co
1	Amplicon Region:15-429	359	35.61%	359
2	Amplicon Region:15975-16418	407	40.37%	407
3	Unknown	242	24.00%	NA

Index	Sequence	Frequency	Total Rea	Forward R	Reverse Read	Differenc
1	CTCCACCATTAGCACCCAAAGCTA	38.44%	218	120	98	4
2	CTCCACCATTAGCACCCAAAGCTA	24.51%	139	76	63	3
3	CTCCACCATTAGCACCCAAAGCTA	5.291%	30	14	16	5
4	CTCCACCATTAGCACCCAAAGCTA	3.527%	20	9	11	4

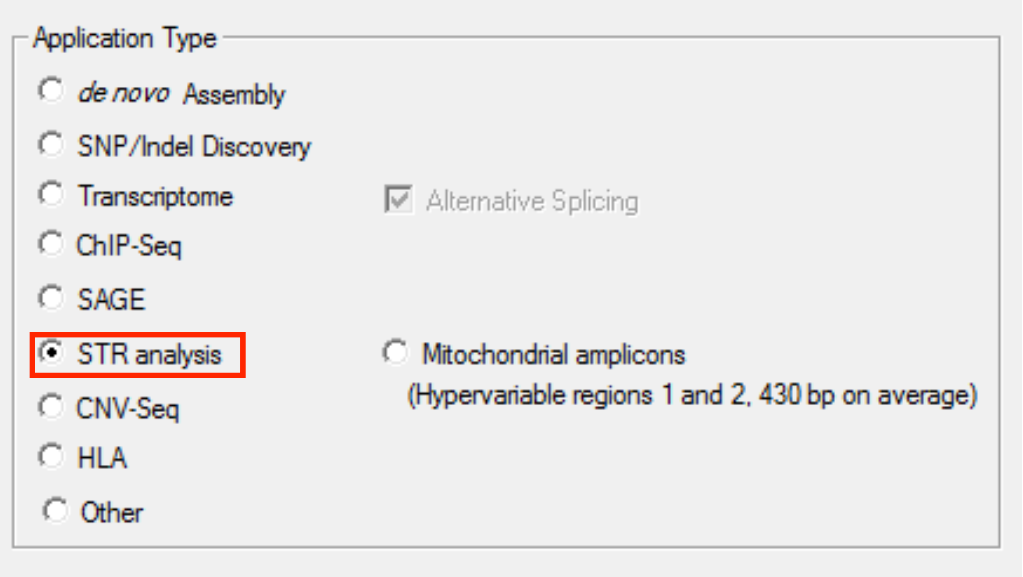
iii. STR: Alignment Algorithm

As discussed above, SoftGenetics has created a modified sequence alignment algorithm specifically for analyzing STRs. The specifics of the modified STR sequence alignment program are detailed here. First, every possible perfect match between 12-mers in the reads and the reference are found. These positions are used to locate clusters of 12-mers where the position in the reference correlates to position in the original read. Each cluster position is scored based on uniqueness of the 12-mers (number of occurrences in the reference) and number of bases of the read matched to the reference. The reads are locally aligned to the chosen cluster by selecting the top scoring alignment out of all possible alignments.

In order to score the alignment, mismatches have a penalty of 1, while in/dels have a penalty of 3 times the length. The forensic alignment has a modification that calculates different penalties for homopolymer in/dels in order to account for the high error rate with homopolymers using 454 pyrosequencing. The penalty is the length divided by three. For example, an insertion of one base in a homopolymer would have a penalty of 1/3. A deletion of two bases in a hompolymer repeat would have a penalty of 2/3. This allows us to penalize missing or extra STR repeats more harshly than the expected homopolymer in/del errors. Forensic alignment also has an extra filter requiring that the read length is at least some fraction of the total reference contig (STR allele) length.

A specific application option for STR analysis was added to the software as demonstrated in the screen shot of the set up screen for the software in Figure 60 below.

Figure 60.



Application Type

○ *de novo* Assembly

○ SNP/Indel Discovery

○ Transcriptome ☑ Alternative Splicing

○ ChIP-Seq

○ SAGE

● STR analysis ○ Mitochondrial amplicons (Hypervariable regions 1 and 2, 430 bp on average)

○ CNV-Seq

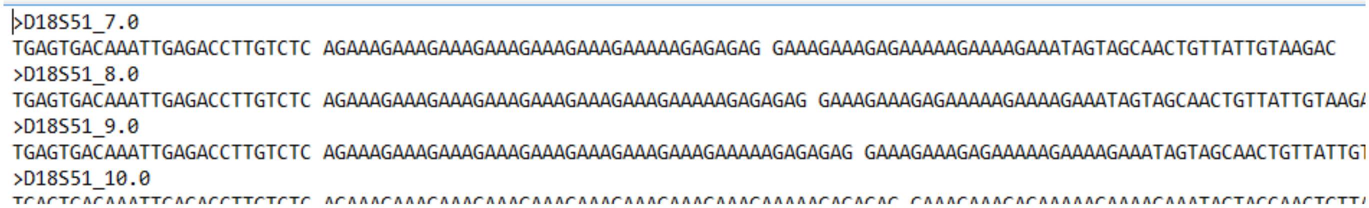
○ HLA

○ Other

Figure 60. Screen shot of NextGENe v.2.4.0 Set Up Screen. An option for STR analysis was added as one of the Application Type to the software. Identification of the application type of the alignment project is needed to determine the optimal alignment algorithm.

For the STR analysis application, custom reference sequences for STR analysis were generated based on the mini-STR primers used for sequencing and alleles listed on the STRBase website. Generating reference sequences of common alleles per locus is necessary because it is ideal to have reference sequences that identically match hypothetical sequence results. An example of the custom STR reference generated for the NextGENe software is presented in the Figure 61 below.

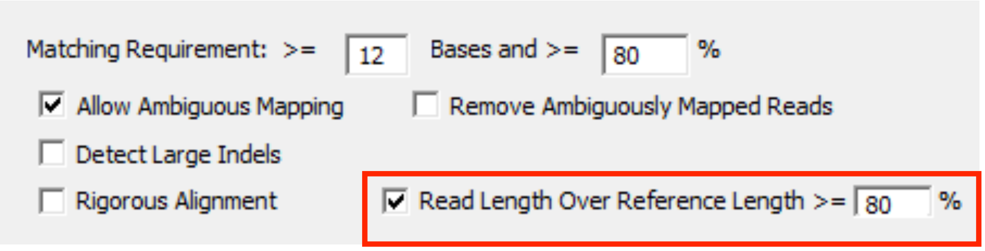
Figure 61.



```
>D18S51_7.0
TGAGTGACAAATTGAGACCTTGTCTC AGAAAGAAAGAAAGAAAGAAAGAAAGAAAAAGAGAGAG GAAAGAAAGAGAAAAAGAAAAGAAATAGTAGCAACTGTTATTGTAAGAC
>D18S51_8.0
TGAGTGACAAATTGAGACCTTGTCTC AGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAAAGAGAGAG GAAAGAAAGAGAAAAAGAAAAGAAATAGTAGCAACTGTTATTGTAAGA
>D18S51_9.0
TGAGTGACAAATTGAGACCTTGTCTC AGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAAAGAGAGAG GAAAGAAAGAGAAAAAGAAAAGAAATAGTAGCAACTGTTATTGT
>D18S51_10.0
TGAGTGACAAATTGAGACCTTGTCTC AGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAAAGAGAGAG GAAAGAAAGAGAAAAAGAAAAGAAATAGTAGCAACTGTTA
```

Figure 61. Screen Shot of the Custom STR References for Alleles per Locus Partial custom STR reference sequences for locus D18S51 alleles 7 – 10 shown in the screen shot.

The user is enabled to customize the filter stringency on the match of a read to a reference for an allele calling as demonstrated in the screen shot of the setting windows for the STR analysis in NextGENe version 2.4.0 below.



Matching Requirement: >= 12 Bases and >= 80 %

☑ Allow Ambiguous Mapping ☐ Remove Ambiguously Mapped Reads

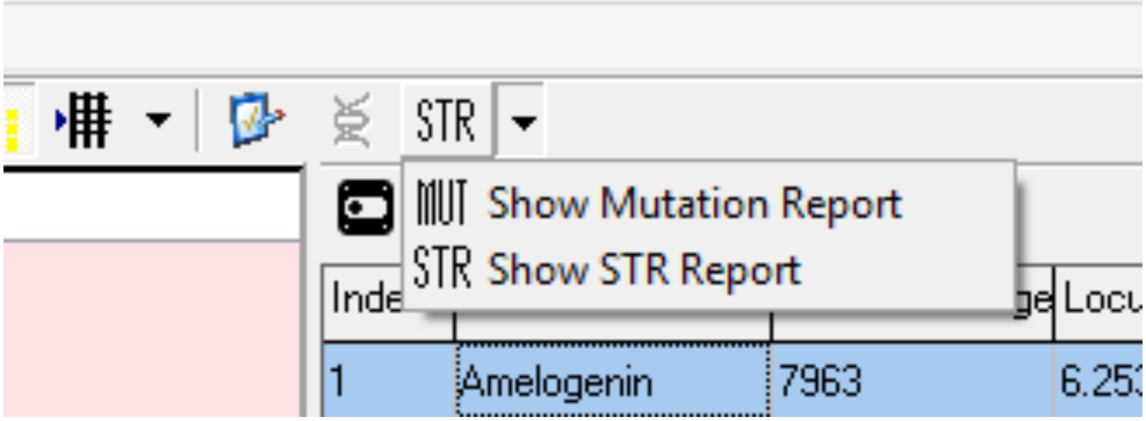
☐ Detect Large Indels

☐ Rigorous Alignment ☑ Read Length Over Reference Length >= 80 %

Figure 62. Screen Shot of the Settings Window for STR Analysis. The setting shown indicates that the observed sequence in the data must match >80% with the reference in length in order to be assigned the corresponding allele in the analysis.

Further, a large effort was put into customizing the output and report format of the sequencing data. As of version 2.4.0, NextGENe is capable of generating STR report for each sample (Figure 63.) in an addition to a mutation report as shown in the screen shot Figure 63.

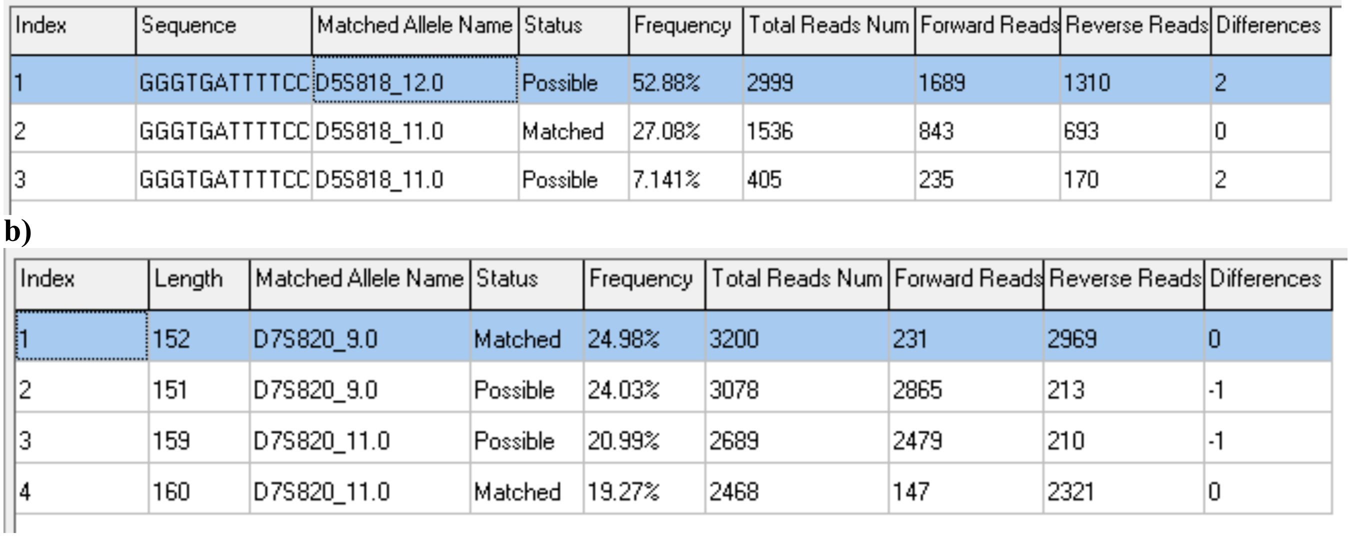
Figure 63. Screen Shot of NextGENe v.2.4.0 For Generation of STR Report



Two different methods of classifying reads to specific STR alleles were established for the software. One of the main alternatives explored is to identify the locus of a read by searching for the primer sequence, and matching the length of the read to a list of known alleles. This approach was suggested in order to mirror the current STR sequencing method using the capillary electrophoresis, which assigns alleles per locus by identifying the length of the amplified DNA fragment. This method will work very well for perfect sequencing reads (the output is essentially equivalent to traditional STR analysis, where peaks are found for different alleles). However, it is necessary to correct for in/del errors in the sequences for this approach. The data we have analyzed to date indicate that this in an acceptable approach but further analysis is needed. Comparing reads from the forward and reverse direction can help.

The second approach was to identify and assign alleles per locus by perfect match sequence alignment of the sequence reads against the custom STR reference sequences. The sequencing results are reported into an allele report based on sequence-based allele assignment or an allele report with length-based allele assignment based on the user's preference. An example of each allele report is presented in Figure 64 below.

Figure 64. Allele Report Based on a). Sequence-based Allele Calling and b) Length-based Allele Calling



a)

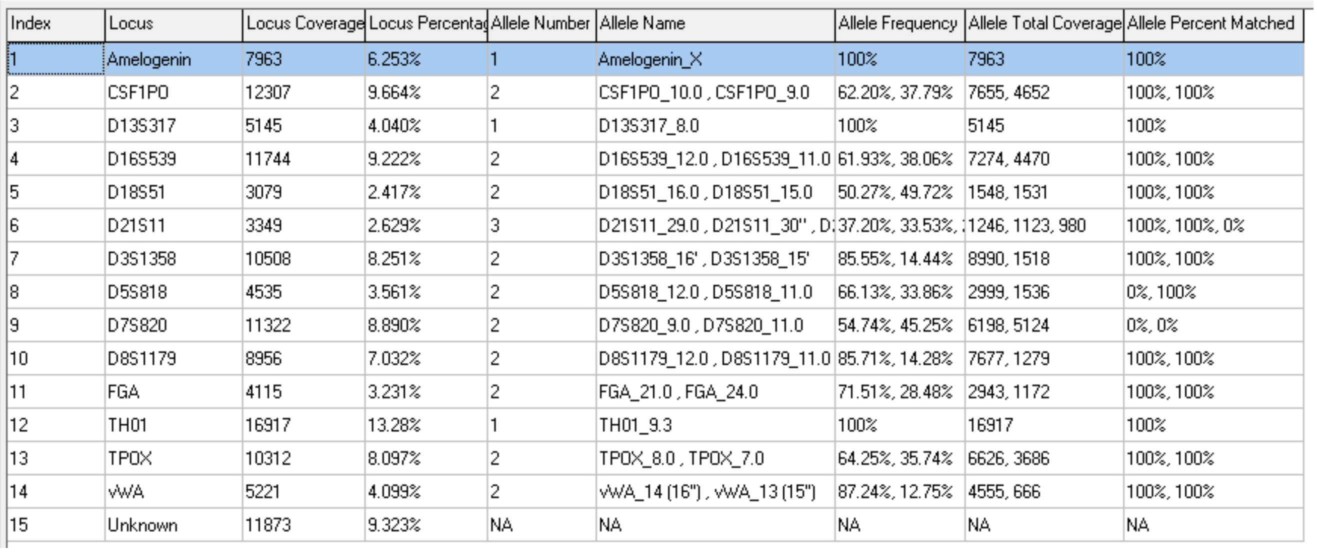
Index	Sequence	Matched Allele Name	Status	Frequency	Total Reads Num	Forward Reads	Reverse Reads	Differences
1	GGGTGATTTTCC	D5S818_12.0	Possible	52.88%	2999	1689	1310	2
2	GGGTGATTTTCC	D5S818_11.0	Matched	27.08%	1536	843	693	0
3	GGGTGATTTTCC	D5S818_11.0	Possible	7.141%	405	235	170	2

b)

Index	Length	Matched Allele Name	Status	Frequency	Total Reads Num	Forward Reads	Reverse Reads	Differences
1	152	D7S820_9.0	Matched	24.98%	3200	231	2969	0
2	151	D7S820_9.0	Possible	24.03%	3078	2865	213	-1
3	159	D7S820_11.0	Possible	20.99%	2689	2479	210	-1
4	160	D7S820_11.0	Matched	19.27%	2468	147	2321	0

In an addition to allele reports, a locus report is also generated per sample to allow the user to check the number of loci detected with their coverage. An example of a locus report is presented in the Figure 65 below.

Figure 65. Locus Report of a 14-plex Sample Generated by NextGENe version 2.4.0

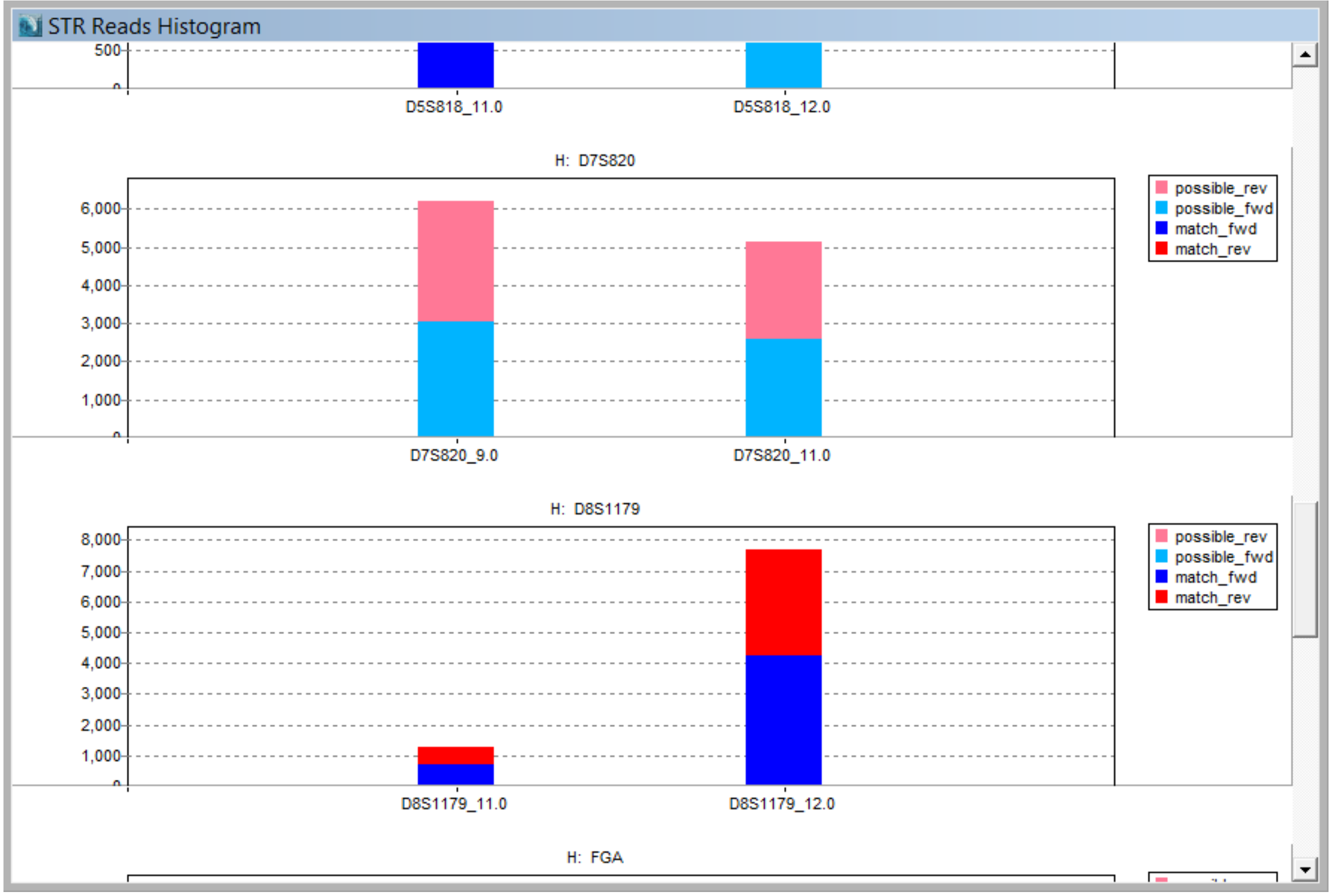


Index	Locus	Locus Coverage	Locus Percentag	Allele Number	Allele Name	Allele Frequency	Allele Total Coverage	Allele Percent Matched
1	Amelogenin	7963	6.253%	1	Amelogenin_X	100%	7963	100%
2	CSF1PO	12307	9.664%	2	CSF1PO_10.0 , CSF1PO_9.0	62.20%, 37.79%	7655, 4652	100%, 100%
3	D13S317	5145	4.040%	1	D13S317_8.0	100%	5145	100%
4	D16S539	11744	9.222%	2	D16S539_12.0 , D16S539_11.0	61.93%, 38.06%	7274, 4470	100%, 100%
5	D18S51	3079	2.417%	2	D18S51_16.0 , D18S51_15.0	50.27%, 49.72%	1548, 1531	100%, 100%
6	D21S11	3349	2.629%	3	D21S11_29.0 , D21S11_30'' , D	37.20%, 33.53%,	1246, 1123, 980	100%, 100%, 0%
7	D3S1358	10508	8.251%	2	D3S1358_16' , D3S1358_15'	85.55%, 14.44%	8990, 1518	100%, 100%
8	D5S818	4535	3.561%	2	D5S818_12.0 , D5S818_11.0	66.13%, 33.86%	2999, 1536	0%, 100%
9	D7S820	11322	8.890%	2	D7S820_9.0 , D7S820_11.0	54.74%, 45.25%	6198, 5124	0%, 0%
10	D8S1179	8956	7.032%	2	D8S1179_12.0 , D8S1179_11.0	85.71%, 14.28%	7677, 1279	100%, 100%
11	FGA	4115	3.231%	2	FGA_21.0 , FGA_24.0	71.51%, 28.48%	2943, 1172	100%, 100%
12	TH01	16917	13.28%	1	TH01_9.3	100%	16917	100%
13	TPOX	10312	8.097%	2	TPOX_8.0 , TPOX_7.0	64.25%, 35.74%	6626, 3686	100%, 100%
14	vWA	5221	4.099%	2	vWA_14 (16") , vWA_13 (15")	87.24%, 12.75%	4555, 666	100%, 100%
15	Unknown	11873	9.323%	NA	NA	NA	NA	NA

In order to screen for the balanced directionality in sequence reads per allele, a function of generating read histogram is added to the software. The Histogram of reads reports the number of read for identified alleles for each locus. Each bar in the histogram represents read depth per identified allele and is color coded to indicate the ratio of sequence reads in forward direction compared to the sequence reads in reverse direction. An example of a read histogram is presented in Figure 66 below.

The effort in optimizing the NextGENe software with Soft Genetics, LLC., is ongoing. Addition of an option for alignment against a circular genome is anticipated in order to accommodate with the whole mitochondrial genome sequencing using the nimblegen probe-capture method.

Figure 66. Reads Histogram Generated by NextGENe v.2.4.0. Histogram of reads for identified alleles for each locus is generated using NextGENe software version 2.4.0.



2. Modification of MIA Software for Analysis of mtDNA

One goal of our research effort was to develop a software analysis procedure that determines (1) if a sample has multiple mtDNA components and (2) assign the fractional composition of these multiple mtDNA haplotypes if present. While this goal was not fully attained, we have made the progress outlined below and fully expect to soon have a working implementation.

Our originally outline approach, based on modifying the Conexio Genomics software was abandoned. Instead, we took on a new collaborator, Dr. Richard Green, who has extensive expertise in the field of ancient DNA analysis. Through this collaboration, Green has begun to

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

modify an existing code base called MIA (mapping iterative assembler) that was developed for assembly and analysis of mtDNA data from ancient samples (Figure 67). This code base has many useful features for forensics work including the ability to learn and use the DNA error profile specific to each sample, a maximally sensitive alignment scheme, rich features for reporting non-consensus sequences in the analysis, and the ability to natively handle circular genomes.

As shown in the schematic Figure 67 for MIA, it is an iterative assembler that considers each read anew at each round of assembly. In this way, it converges on a correct assembly and can sensitively detect reads that may have had significant divergence from the original reference. The program uses the MIA iterative mapping algorithm to align the sequences to a reference sequence illustrated in the figure below. Each read is mapped to the starting reference sequence using the Needleman-Wunsch algorithm and a position-specific scoring scheme that can be trained for the chemical damage typical of degraded samples (Figure 67a). This alignment methodology is maximally sensitive in detecting and aligning sequence reads. Once the reads are aligned, a new consensus sequence is called (Figure 67b). This new consensus replaces the starting reference sequence and all reads are re-mapped to it. This process continues until the consensus no longer changes. The alignment scores for each read during each round are analyzed to dynamically determine an appropriate cutoff for inclusion in the assembly (Figure 67c). In this way, extraneous sequences can be excluded making this approach appropriate for samples with large amounts of extraneous sequence, like mixed forensic samples. In performance evaluations, we have been able to correctly assemble genomes with more than 10% divergence to the starting reference using this iterative approach. This sensitivity will be useful for forensics applications.

Figure 67. MIA Algorithm

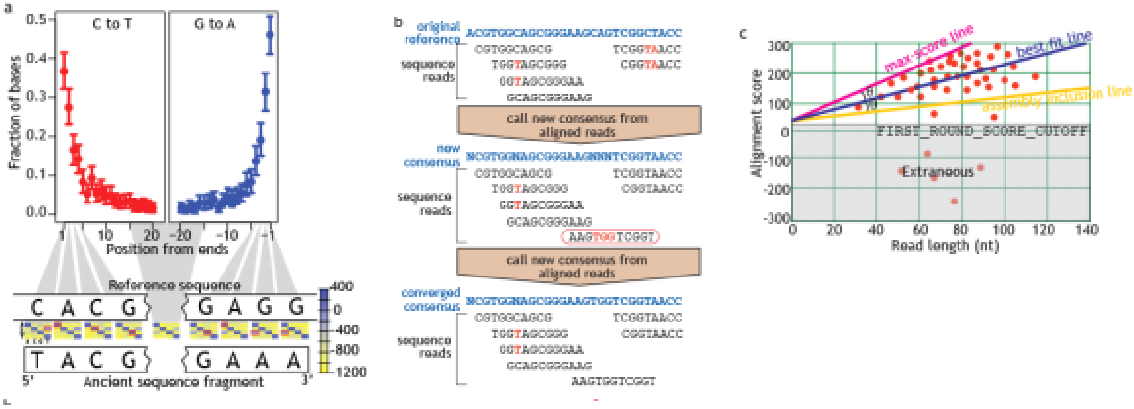


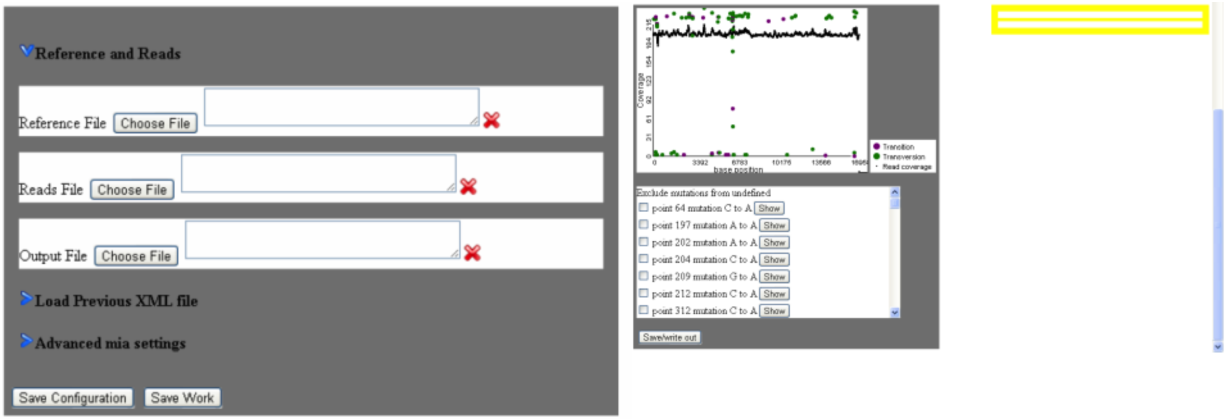
Figure 67. MIA algorithm. (a) Each input sequence read is aligned with a custom position-specific alignment matrix that is aware of the mutational tendencies of degraded DNA. (b) Each read is aligned to the consensus. These alignments are used to call a new consensus. This continues until no further changes to the consensus are found. (c) Inclusion of a read into the consensus is based on a dynamic score cutoff such that reads that may originally look dissimilar to the consensus can be included in later rounds.

Our efforts thus far have included (1) writing a web-based front-end for MIA for 454 assembly and analysis and (2) performing a proof-of concept analysis using data from a "spike-in" experiment to test detection efficiency of minor haplotypes. Having satisfactorily completed these tasks, our final goal is to implement the extension to MIA wherein the variants detected in a complex mixture are further analyzed to give a haplotype based summary. Each of these three tasks is described in detail below.

i. Developed a web-based front-end for MIA

We have developed a flexible front-end for running and analyzing the results from MIA. In brief, MIA operates by aligning each of the input 454 (or other platform) reads against a defined reference mtDNA assembly. Each read with a statistically significant alignment score is recorded in the output file. The web-based front-end handles both invoking MIA on input data (Figure 68a) and interpreting the output data (Figure 68b).

Figure 68. a) **b)**



Exclude mutations from undefined

☐ point 64 mutation C to A Show
☐ point 197 mutation A to A Show
☐ point 202 mutation A to A Show
☐ point 204 mutation C to A Show
☐ point 209 mutation G to A Show
☐ point 212 mutation C to A Show
☐ point 312 mutation C to A Show

Figure 68. MIA web-based front end. A)In the left panel is shown the interface for invoking MIA. One simply specifies the input files (mtDNA reference, reads) and output file. Optionally, one can specify custom alignment parameters for the assembly. B)In the right panel is shown the graphical output summary for the assembly. The percent or reads with a consensus base is shown along with various filtering criteria for examining positions where some reads differ from the consensus.

ii. Evaluate sensitivity of MIA in "spike-in" mixture experiment

We tested the ability of MIA to detect a minor component (10%) of an alternative haplotype mixed into a background of mtDNA from another haplotype. In this experiment, a sample was prepared that contained 90% of a known haplotype and 10% of an alternative haplotype. These haplotypes differ at several known positions. A library of this mixture was prepared and sequenced. After alignment and filtering of the data, we analyzed the fraction of reads that contained the variants associated with the major and minor haplotypes. As shown in Figure 3, on average MIA detected the minor haplotype variant in 11.9% of the reads. While this result is encouraging, our future goals are to determine the lower limits of detection efficiency and to test more complex mixtures, i.e., with more than two haplotypes.

Figure 69

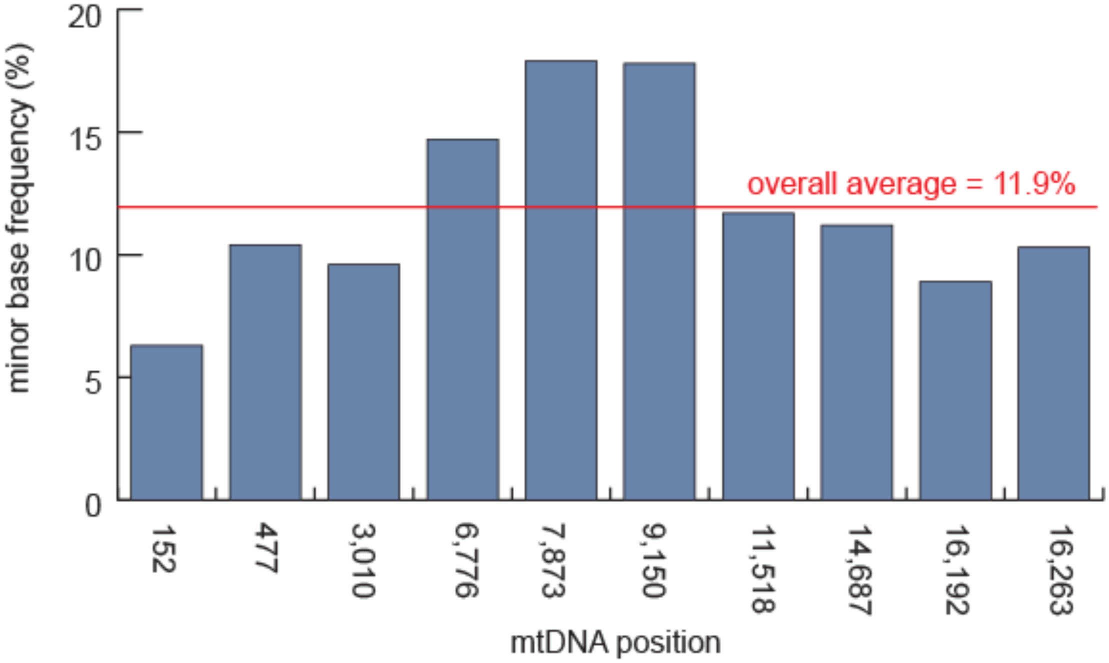


Figure 69. mtDNA mixture experiment. The major mtDNA haplotype comprised 90% of the input. Shown are the ten mtDNA positions where the major and minor haplotype have a known sequence difference. The y-axis shows the fraction of reads containing the minor haplotype base. The overall average of reads containing the minor haplotype base (red line) was 11.9%.

iii. Haplotype-level summary of complex mtDNA mixtures

While the number of minor variants at each mtDNA position where one is observed can be manually analyzed to infer the mixture of haplotypes present, this task lends itself to a more automated approach. We will use the large datasets available on the phylogeny of human mtDNA variation to do this. The approach is shown schematically in Figure 70.

Figure 70.

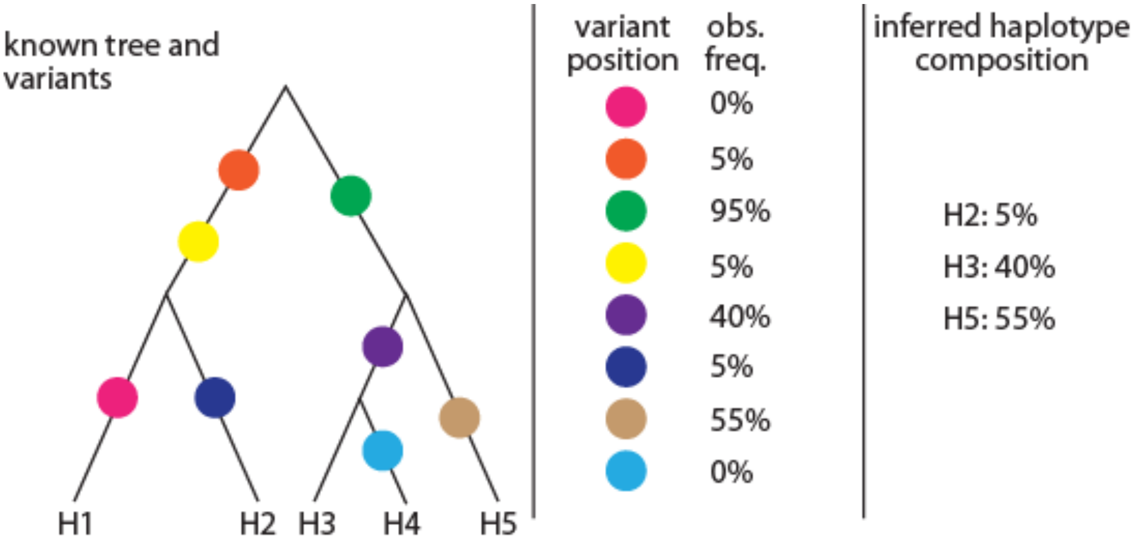


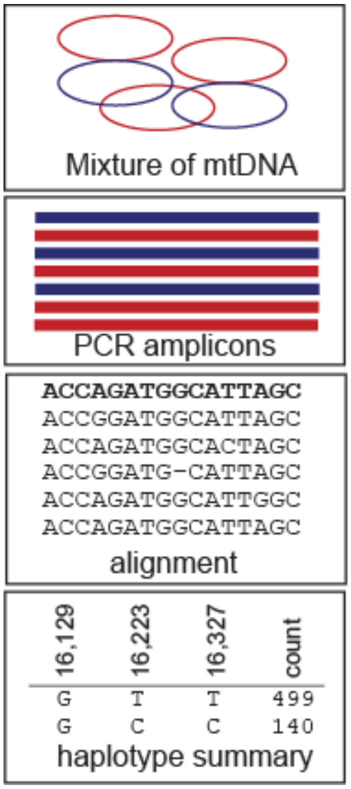
Figure 70. Inference of haplotype composition of complex mixtures. An additional input is required (left panel): the known haplotype phylogeny of human mtDNAs and the variants associated with each branch in the tree. After calling variants and their proportions (middle panel), we will make a maximum likelihood determination of the number and relative abundance of each haplotype present.

This approach requires that use of an additional, static input: the known phylogeny of human mtDNA and its associated variation. We will combine the counts for observed variants at each position to find a maximum likelihood solution for detecting which haplotypes are present and their relative abundances.

iv. Hap-summary.pl

We have developed a program called hap-summary.pl to provide read counts of mtDNA sequence haplotypes from a sample which can consist of a mixture of mtDNA sequences due to mutation (heteroplasmy) or a mixture of mtDNA sequences from multiple individuals. The procedure is shown schematically in Figure 71.

Figure 71. Schematic of Haplotype Sequence Analysis of mtDNA mixtures

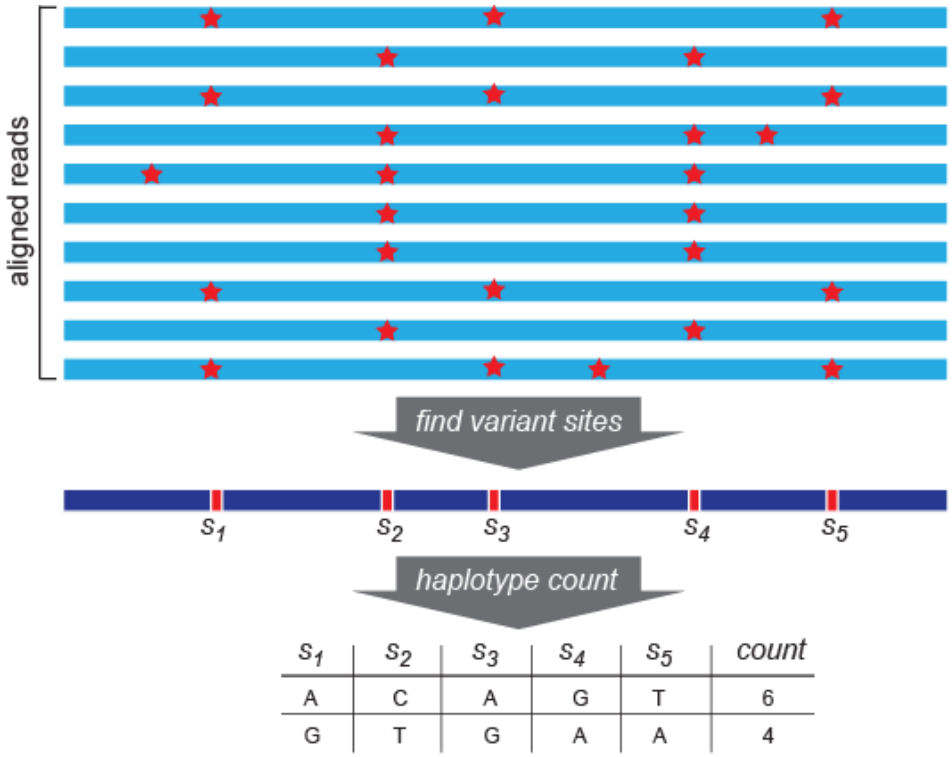


```
ACCAGATGGCATTAGC
ACCGGATGGCATTAGC
ACCAGATGGCACTAGC
ACCGGATG-CATTAGC
ACCAGATGGCATTGGC
ACCAGATGGCATTAGC
```

16,129	16,223	16,327	count
G	T	T	499
G	C	C	140

A pool of mtDNA is used as template for PCR. This pool may be from a single or multiple individuals. The PCR uses primers that will amplify the target region from conserved segments of the mtDNA. Thus, the components of the mixture should be present in the PCR product is concentrations that are proportional to their template concentration. This amplicon product is then converted into a high-throughput sequencing library and sequenced using settings appropriate for amplicon sequencing. The sequence data are then aligned to the segment of the revised Cambridge reference sequence that corresponds to the HVRI or HVRII region that has been amplified. For this step, we have developed a custom alignment program, *mia*, that performs a full Needleman-Wunsch alignment for maximal alignment sensitivity and accuracy. Finally, the aligned amplicon sequences are analyzed to determine which sites carry variable positions (SNPs) and which of these variants co-occur. That is, the presence and proportion of each discrete haplotype is measured. For this analysis, we have developed and tested a program called, *hap-summary.pl*. A schematic of the hap-summary is shown in Figure 72 and a description below.

Figure 72. hap-summary Schematic



s_1	s_2	s_3	s_4	s_5	count
A	C	A	G	T	6
G	T	G	A	A	4

The input consists of a set of reads aligned to the rCRS. Each position is examined to determine the fraction of all reads that carries a variant base at that position (red stars). Sites where more than a user-defined percentage of reads are observed to be variant are flagged as **variant sites**. Then, each read is assigned to a haplotype according to the bases it carries at the variant sites. In this way, sporadic sequencing error can generally be ignored for determining the haplotype configuration. The output is a list of the variant sites, the observed haplotype configurations, and their counts from most to least common. This algorithm was used to analyze the 454 single source and mixture data and a summary is presented below. Data analyzed using hap-summary are consistent with data previously analyzed with alternate software. However, the haplotype sequence summary is much easier for mixture analysis and essential for mixtures with greater than 2 contributors as it provides sequence haplotype counts and not just mutation variant counts.

Analysis of single sample data

We analyzed using hap-summary data collected from two "Population Studies" wherein a battery of mtDNAs was individually amplified and barcoded. These products were then pooled and sequenced together. The resulting data, after barcode identification, should be the product of a single mtDNA haplotype. These data were used processed with the goal of determining which, if any, sites are called variant in the presumed absence of any *bona fide* variation. This is an important experiment for several reasons. First, genuine heteroplasmy within an individual can generate data with variable sites. These should be identified and ignored in haplotype assignment. Second, sequencing error and especially homopolymer-related sequencing error using the 454 sequencing system, can generate what appear to be variant sites. It is a well-established observation that sequence context can generate reproducible sequencing error using any sequencing technology and especially pyro-sequencing. Third, there are some regions of low complexity where alignment can be ambiguous. These regions must be identified by visual inspection of the alignment data.

The two Population Studies contained pooled data from 62 and 63 individuals. We aligned these data to the HVRI and HVRII using *mia* and summarized the haplotypes using *hap-summary.pl*. In each sample, we identified the sites where at least 2% of the amplicon data contain a base that differs from the consensus base at that position. Each such position is called a variant position for that sample. Then, we counted the number of samples in which that position was found to be variant. The results of this accounting are shown in Figure 73 for the HVRI region. Ideally, these datasets, each of which derives from a single mtDNA haplotype would have no variant positions in any sample. However, there are several positions that consistently have variant reads. As shown in the figure, variant positions are typically associated with the base miscalls and alignment ambiguities in the vicinity of homopolymer stretches.

To account for this issue, we implemented in *hap-summary.pl* a position-specific filtering regime. The program can take as argument an arbitrary list of positions to ignore for variant detection and haplotype summary. Further, it can be told the length of the PCR primers that flank the amplicon sequence and ignores any variant positions that are within the primer regions.

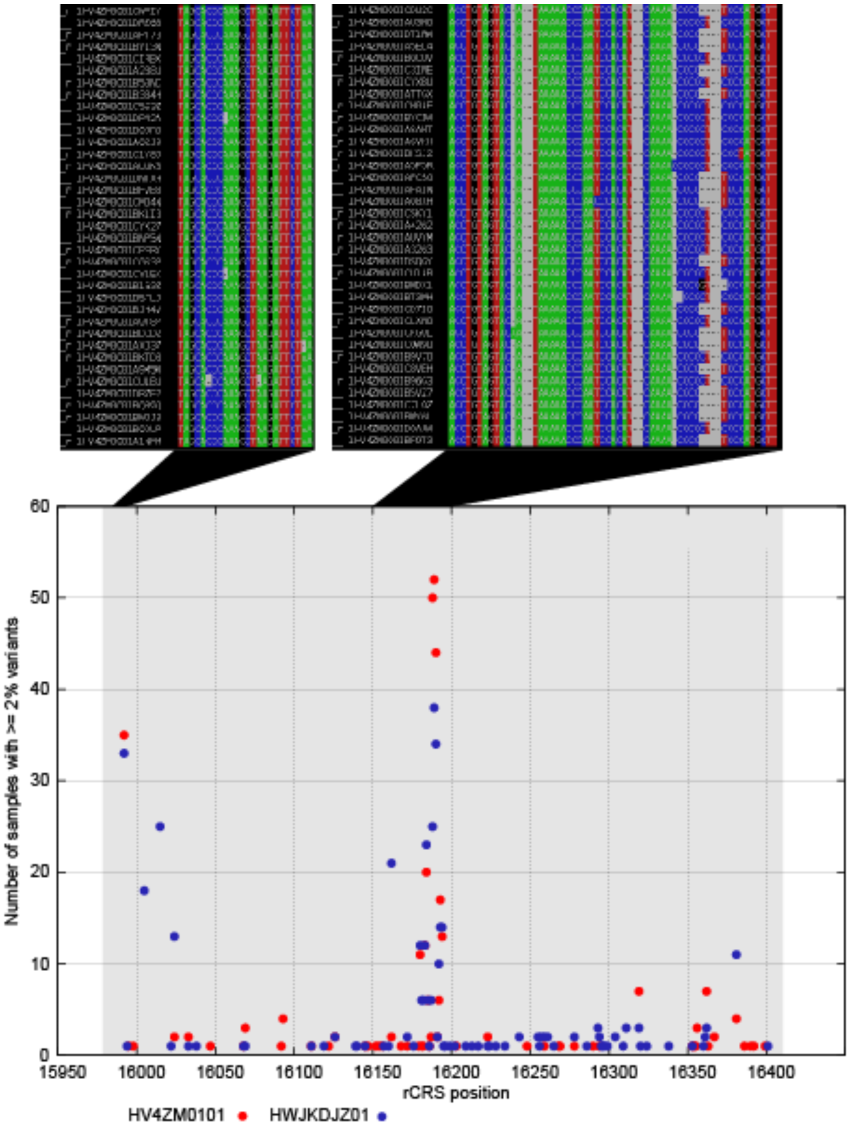
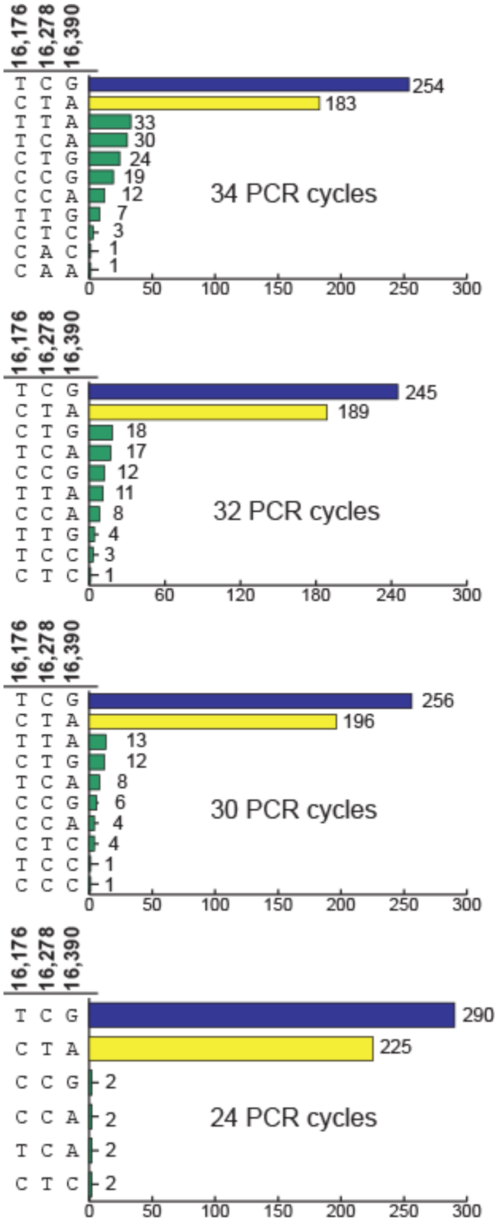


Figure 73. – Sequence variants in HVRI single-sample amplicon data. 62 samples of single source were individually amplified, bar-coded, and analyzed. The x-axis shows the positions where variants were detected. The y-axis shows the number of samples with at least 2% of the reads being variant at the given position. The panels above show sample alignment data at the indicated regions of high-number of variants. As shown, these are often associated with homopolymer regions.

Examination of the Population Study data led to exclusion of the following sites in HVRI: 16,005, 16,015, 16,024, 16,162, 16,179-16,199, and 16,381. For further analyses, these sites are excluded from variant detection and haplotype definition due to their propensity for being falsely indicative of variation.

Jumping PCR

It is known that PCR can sometimes switch templates during amplification. This phenomenon, sometimes called "jumping PCR" can result in amplicon products that are chimeric between molecules. For mtDNA mixtures, this could result in haplotypes that are a combination of the input templates.



We tested the relationship between the number of PCR cycles and haplotype chimerism with the following experiment. We amplified from a 1:1 mixture of COR037 and COR110 using an increasing number of cycles. Then, we performed the alignment, variant detection, and haplotype calling procedure as described above. The results of this experiment are summarized in Figure 74. As expected, fewer PCR cycles results in few template-switched products. After 34 PCR cycles, 437 of 567 sequenced products (77.1%) were not template-switched. In contrast, after 24 PCR cycles, 515 of 529 sequenced products (97.4%) were not template-switched. Thus, for accurate haplotype identification and quantification from a complex mixture, it is important to keep the number of PCR cycles to a reasonable minimum.

2 Person and Complex Mixtures

To test our ability to reliably determine the ratio of a simple mixture of templates, we engineered mixtures of two templates (COR073 and COR110) with ratios of 100% down to 0.1% of each. We then performed the PCR, sequencing, and analysis process as described above. These samples were found to differ in the HVRI region at position 16,327. COR073 carries a T whereas COR110 carries a C. To estimate the observed proportion of each sample in the sequence data we consider only the correct haplotype, i.e., disregard all other differences, and report the proportion of correctly sequenced and aligned amplicon reads differing at position 16,327. These results are shown in Table 21. Overall, there is a expected percentage of reads that match the COR110 haplotype and the observed percentage. However, we note that the limit of detection is about 1% of the minor component. Notably, regardless of whether the minor component was COR110 or COR073, if it was mixed at 1% or less, we failed to recover any sequence reads corresponding to the minor component. This could be due to the complete absence of the minor product in the PCR amplicon pool, limits of detection from a few hundred reads, or a combination of these factors. We are in the process of using Hap- to analyze complex mixtures (3 to 5 person mixtures). Preliminary data show that the algorithm successfully provides read counts for each observed

COR110 % Expected	COR110% Observed
0	0
50	49.2
25	29.7
10	8.2
5	7.4
2.5	2.9
1.0	0
0.5	0
0.25	0
0.1	0
100	100
50	50
75	78.3
90	89.7
95	94
97.5	97.2
99	100
99.5	100
99.75	100
99.9	100

sequence haplotype (table 22 below). However, additionally analysis is needed to identify ‘true’ sequence haplotypes from jumping PCR chimeras, sequence error, and heteroplasmy mutations. Additional modifications to account for these issues are in progress.

```
The sites that are variant using cutoff of 2 percent
16129 16223 16327 16391 count
G     T     T     G     504
G     C     C     G     142
A     C     C     A     40
G     T     C     G     15
A     C     C     G     13
G     C     T     G     12
A     T     T     G     9
G     C     C     A     7
A     C     T     G     6
```

E. Conclusions

1. Discussion of Findings

We have successfully developed a duplex PCR assay targeting the mtDNA hypervariable regions I and II (HVI/HVII) using eight sets of 454 MID tagged fusion primers in a combinatorial approach for deep sequencing 64 samples in parallel on a 454 GS Jr. The library preparation methods including the PCR parameter and the AMPure purification method were proven to be optimized through quality check steps and the sequencing results. This assay was shown to be highly sensitive for sequencing limited DNA amounts (~100 mtDNA copies) and detecting mixtures with low level variants (~1%) as well as “complex” mixtures (≥3 contributors). The minor base was reliably detected at each of the mixed base positions and the observed frequencies were similar to the expected frequencies using 454 sequencing with ~600-1000 reads per amplicon but not by Sanger sequencing. Further, we have characterized and observed the jumping PCR effect that arises in amplification of mixed samples that are often encountered in the forensic field and concluded that the frequency of the phenomenon can be decreased by lowering the amplification cycle number. Heteroplasmic samples of various tissues from monozygotic twins were studied using the developed assay and the 454 sequencing results were compared to the Sanger sequencing results. Heteroplasmy at position 16093 was detected in the buccal samples by both methods but was only detectable in blood using the more sensitive 454 NGS method. We show that not only is NGS more sensitive for detecting minor components in a mixture, but it is also more quantitative because it provides a digital read-out counting the number of sequence reads corresponding to the individual components. In addition, a population study was conducted in order to confirm the concordance with Sanger Sequencing data and to generate a NGS population data base. All confirmed SNPs from Sanger Sequencing were detected, and further low level mutation that was not detected through Sanger Seuqencing method was observed in 454 population database. Such low level heteroplasmy needs further investigation. A concordance and reproducibility study was conducted between CHORI and the Jan Bashinski DNA Laboratory of The California Department of Justice (CA DOJ), and the 454 sequencing results were concordant between labs and compared to Sanger Sequencing.

We have also successfully designed and showed proof of concept for a solution phase sequence capture and NGS assay for targeted enrichment and deep sequencing of the entire mitochondrial genome for increased discrimination power. Using this SeqCap NGS assay, 100% sequence coverage of the mitochondrial genome with an ~80% on target rate was achieved. Additionally, a DNA fragmentation method using mechanical shearing was optimized for rapid library preparation. This method was shown to be DNA quantity and quality independent, essential for preparation of highly degraded or limited samples often encountered in forensic cases. This optimized fragmentation method coupled with the SeqCap NGS assay was successfully used for sequencing the entire mitochondrial genome of limited DNA samples as well as detection of minor sequences in a mixture. The input DNA amount was successfully lowered to the forensically relevant level of 100 pg, but the definitive sensitivity limit is to be determined. 10% and 5% mixture resolution was established, and even the lower mixture detection limit is expected to be established with higher number of reads. Proof of concept for the application of the assay for degraded samples was established with artificially degraded samples, and further exploration for such application is currently undergoing.

Also, 454 NGS assays targeting the CODIS STR loci were developed using 454 mini STR fusion primers in a multiplex PCR as well as an approach using a universal 454 primer set for amplification of a mini-STR multiplex. A STR assay with M13 universal primer design using 454 Next Generation Sequencing technology was developed as an alternative approach to the fusion primer approach to minimize the number of primer sets required for pooling multiple samples per sequencing run. In this unique design the samples are first amplified by inner primers consisting of STR locus target specific sequence and a universal M13 linker sequence. The inner products are then amplified with outer primers consisting of complementary M13 linker sequence, a 10 bp MID tag, and a 25 bp 454 specific sequence. We have showed proof of concept for generating a full STR profile with 50 pg input DNA through a small sensitivity study, but the lower sensitivity limit is not established yet. Different purifications methods for small fragment removal (primer dimer) were examined for both STR assays. Although the current method is optimized with the AMPure purification, the potential product loss with AMPure XP small fragment removal need to be addressed and explored further.

Further analysis is needed.

We have worked with 2 external collaborators to modify next generation sequencing software for mtDNA and STR next generation sequence alignment and analyses. A commercially available software NextGENe was customized for the analysis of data generated for HVI/HVII assay and the STR assays. Two analysis options, mitochondrial amplicon analysis and STR analysis, were added to the software, and modifications to the sorting tool were made to accommodate the combinatorial use of MID barcode tagging. The STR sequence alignment algorithm was optimized for alignment of STR repeats, assigning lower gap penalty for IN/DEL sizes corresponding to the repeat size (4 bp). The modified algorithm results in improvement for alignment of the STR repeats. However, due to the nature of the repeats in the STR markers, further optimization and modification of the software is needed for the correct and optimal alignment of the STR sequence. A high pyrosequencing error rate was observed in the homopolymer C stretch region in HVI and HVII amplicons and resulted in sequence mis-alignment in these regions using the 454 AVA software. However, base substitutions were correctly identified using the NextGENe software in these homopolymer regions and the sequence alignment was somewhat improved when filters for removing sequence errors were applied using this software. Additional modifications addressing these issues are still needed to effectively analyze the data and are on-going.

We have also developed a flexible front-end for running and analyzing the results from Mapping Iterative Assembler (MIA) for mtDNA mixture analysis, which operates by aligning each of the input 454 (or other platform) reads against a defined reference mtDNA assembly in collaboration with Dr. Richard Green at UC Santa Cruz. MIA was able to detect a minor component (10%) of an alternative haplotype mixed into a background of mtDNA from another haplotype. Recently a modification to the algorithm was made and is currently being tested for analysis of mixtures for the HVI/HVII Amplicon assay. Incorporating an algorithm to identify sequence haplotypes of detected sequences is currently being explored for mixture analysis of the whole mitochondrial genome. A feature to take into account the circular nature of the entire mitochondrial genome is in the process of being added. Further modifications to the software are on-going.

2. Implications for Criminal Justice Policy and Practice

The massively parallel and clonal nature of next generation sequencing (NGS) technologies has the potential to revolutionize the forensics DNA field. The implementation of NGS in forensics labs will have a significant impact for many different reasons. NGS is capable of analyzing sequence polymorphisms (e.g. mtDNA, SNPs) as well as length polymorphisms (e.g. STRs) on the same platform, and, in principle, in the same run. Moreover, sequencing STR markers will detect any sequence polymorphisms in the STR loci, thus increasing the discrimination potential of these widely used markers. The high throughput nature of NGS makes it technically feasible and cost-effective to sequence the entire mtDNA genome rather than just the HVI and HVII regions, thereby greatly increasing the discrimination potential of mtDNA analysis. In addition, the sensitivity of NGS systems facilitates the analysis of forensics specimens with limiting or degraded DNA.

The clonal nature of NGS provides a powerful and quantitative means of deconvoluting mixtures, a particularly challenging category of forensics specimens. Different contributors to a mixture can be detected by analyzing the different clonal sequence reads identified in the sequence analysis and their contribution quantified by simply counting the number of sequence reads.

This digital analysis is much more precise than estimating peak height or area for STR markers or analyzing Sanger electropherograms for mtDNA sequences. For chromosomal markers, one must make assumptions about which alleles in a mixture "go together" as a genotype and many statistical approaches to mixture analysis consider all possible genotype combinations and, in some cases, different numbers of contributors. MtDNA, on the other hand, is a particularly useful marker for deconvoluting mixtures because, potential heteroplasmy aside, each distinct mtDNA sequence corresponds to (at least) one contributor. Statistical analyses can capture and incorporate the probability that one mtDNA sequence could correspond to more than one contributor. Thus, NGS analysis of mtDNA is the most robust way of estimating the number of contributors to a mixture.

In general, for forensics as well as other genetic analyses, the two approaches to target enrichment for library preparation are PCR or probe capture. In our experience, PCR has proved effective for NGS analysis of STRs and the HVI and HVII regions of mtDNA. We have found, however, that probe capture is a robust, sensitive, and efficient way to enrich for the whole mtDNA genome. Given the high concentration of mtDNA, the probe capture is efficient with less starting material and allow for shorter hybridization times than for nuclear markers. The optimized probe capture methods described have great promise to overcome many of the major challenges routinely encountered with analysis of limited and degraded DNA. As noted above, the probe capture technique allows the entire mitochondrial genome to be analyzed, greatly increasing the discrimination power of current mtDNA assays which use Sanger sequencing for analysis of just the HV regions and can be applied to analysis of degraded samples since it is not dependent on specific priming sites. The NimbleGen SeqEZ probe capture method in conjunction with NGS technology also provides high sensitivity and requires lower starting DNA amounts compared to other NGS enrichment methods.

The customized forensic NGS software developed for this project will be made publicly available and has broad applicability for NGS applications. Advances in oligo synthesis have markedly reduced the cost of probes, making probe capture an affordable method for routine forensic applications. The use of multiplex tags for sample pooling greatly reduces per sample costs and implementation feasible. NGS technologies are capable of analyzing both mitochondrial and nuclear DNA targets as well as SNP and STR markers simultaneously. Overall, the proposed capture and NGS assays offer increased resolution and discrimination power for mtDNA as well as improved success for the analysis of degraded and limited DNA analysis.

One limitation of this study is that it was performed on the 454 platform (GS Junior instrument), a technology that will be discontinued in 2016. At the outset of this study, the 454 platform was chosen because it was the first NGS technology available and because it could achieve far longer sequence reads then the subsequently developed NGS platforms (Illumina, Ion Torrent). However, all of the approaches using the 454 platform developed and reported here can be adapted to Illumina or Ion Torrent with minor modifications. Furthermore, the cost of sequencing, the ease of use, and the length of sequence reads, have all been improving on all platforms, making it likely that NGS on relatively inexpensive desktop sequencers will be broadly implemented in forensics labs over the next 5 years. However, some challenges facing implementation will include start-up cost of instrumentation, training, and validation.

3. Implications for Further Research

Next generation sequencing (NGS) promises to have a major impact on the practice of forensics labs over the next 5-10 years. Over the next few years, the throughput, cost, and read length obtained with existing platforms should improve and the availability of relatively inexpensive desktop sequencers will make access to NGS affordable and cost-efficient for many forensics labs. One of the reasons for implementing NGS in forensics labs is that sequence markers (e.g. mt DNA, SNPs) and length polymorphisms (e.g. STRs) can be run on the same platform. Moreover, detecting sequence polymorphism within the repeat region or the flanking region for STR markers will increase their discrimination potential. Also, the clonal nature of NGS means that, in the future, NGS will be the method of choice for deconvoluting mixtures. In particular, mtDNA is the marker that is most informative for determining the number of contributors to a forensic mixture and, given the ease with which whole mtDNA genome sequences can be obtained with NGS from both reference and forensics samples, we anticipate that mtDNA analyses will move from HVI and HVII alone to whole mtDNA genome.

NGS analysis will need to be "backward" compatible with currently used markers so that the NGS panels of the future (potentially including phenotypic and ancestry markers) will need to include the current CODIS STR markers. Similarly, software capable of using whole genome mt DNA sequences to search missing person data bases populated with only HVI and HVII data will be needed; this software will need to take into account "hot spots" for heteroplasmy to maximize the discrimination potential of mtDNA sequences. To increase the discrimination potential of "lineage" markers such as Y-chromosome STRS and SNPs as well as mtDNA sequences, the databases for these lineage markers will need to be significantly increased and software designed to estimate the probability of a random match from such lineage databases.

With respect to our study, we will continue to optimize the probe capture system for mtDNA. Increasing the balance of coverage over the genome and reducing the time of probe capture hybridization will be a major focus. We will be determining the sensitivity of input DNA as well as the proportion of a minority sequence that can be detected as well as characterizing heteroplasmy patterns from different tissues. We will also focus on analysis of degraded samples, such as DNA from bones shown to be degraded using a qPCR assay. We will use the NextGENe software for this analysis. We will also use custom software developed in collaboration with Richard green for mtDNA mixture analysis.

We will test our currently optimized NimbleGen capture system on various kinds of challenging forensics specimens as well as samples from many different populations and we will be submitting our whole mt DNA genome sequences to EMPOP. We also plan on adapting the whole genome sequencing for use with Illumina platforms.

For STR analysis, we will be comparing different software packages for STR allele calling. Most of the NGS work we have done thus far has been based on the 454 GS junior instrument system but, with some minor protocol modifications, the systems we have developed can be adapted to other NGS platforms (e.g.Ion Torrent and Illumina). We anticipate that, over the next 2 years, much of our work will be carried out on a MiSeq instrument.

V. References

References

1 Ingman, M. & Gyllensten, U. mtDB: Human Mitochondrial Genome Database, a resource for population genetics and medical sciences (2006) *Nucleic Acids Res* **34**, D749-751.

2 Butler, J. M. Short tandem repeat analysis for human identity testing (2004) *Curr Protoc Hum Genet* **Chapter 14**, Unit 14 18.

3.

4 Just, R. S., Irwin, J. A., O'Callaghan, J. E., Saunier, J. L., Coble, M. D., Vallone, P. M., Butler, J. M., Barritt, S. M. & Parsons, T. J. Toward increased utility of mtDNA in forensic identifications (2004) *Forensic Sci Int* **146 Suppl**, S147-149.

5 LaBerge, G. S., Shelton, R. J. & Danielson, P. B. Forensic utility of mitochondrial DNA analysis based on denaturing high-performance liquid chromatography (2003) *Croat Med J* **44**, 281-288.

6 Di Martino, D., Giuffre, G., Staiti, N., Simone, A., Le Donne, M. & Saravo, L. Single sperm cell isolation by laser microdissection (2004) *Forensic Sci Int* **146 Suppl**, S151-153.

7 Elliott, K., Hill, D. S., Lambert, C., Burroughes, T. R. & Gill, P. Use of laser microdissection greatly improves the recovery of DNA from sperm on microscope slides (2003) *Forensic Sci Int* **137**, 28-36.

8 Sanders, C. T., Sanchez, N., Ballantyne, J. & Peterson, D. A. Laser microdissection separation of pure spermatozoa from epithelial cells for short tandem repeat analysis (2006) *J Forensic Sci* **51**, 748-757.

9 Bentley, G., Higuchi, R., Hoglund, B., Goodridge, D., Sayer, D., Trachtenberg, E. A. & Erlich, H. A. High-resolution, high-throughput HLA genotyping by next-generation sequencing (2009) *Tissue Antigens* **74**, 393-403.

10 Holland M, M. M., et al. Next Generation Sequencing of Forensic DNA Loci Using 454 Life Science Technology. 20th International Symposium on Human Identification. Las Vegas, NV. (2009).

11 Irwin, J. A., Leney, M. D., Loreille, O., Barritt, S. M., Christensen, A. F., Holland, T. D., Smith, B. C. & Parsons, T. J. Application of low copy number STR typing to the identification of aged, degraded skeletal remains (2007) *J Forensic Sci* **52**, 1322-1327.

12 von Wurmb-Schwark, N., Heinrich, A., Freudenberg, M., Gebuhr, M. & Schwark, T. The impact of DNA contamination of bone samples in forensic case analysis and anthropological research (2008) *Leg Med (Tokyo)* **10**, 125-130.

13 Torres, Y., Flores, I., Prieto, V., Lopez-Soto, M., Farfan, M. J., Carracedo, A. & Sanz, P. DNA mixtures in forensic casework: a 4-year retrospective study (2003) *Forensic Sci Int* **134**, 180-186.

14 Melton, T., Dimick, G., Higgins, B., Lindstrom, L. & Nelson, K. Forensic mitochondrial DNA analysis of 691 casework hairs (2005) *J Forensic Sci* **50**, 73-80.

15 Butler, J. M. *Forensic DNA typing : biology, technology, and genetics of STR markers*. 2nd edn, (Elsevier Academic Press, 2005).

16 Calloway, C. D., Reynolds, R. L., Herrin Jr, G. L. & Anderson, W. W. The frequency of heteroplasmy in the HVII region of mtDNA differs across tissue types and increases with age (2000) *Am J Hum Genet* **66**, 1384-1397.

17 Melton, T. Mitochondrial DNA heteroplasmy (2004) *Forensic Sci Rev* **16**.

18 Tully, L. A., Parsons, T. J., Steighner, R. J., Holland, M. M., Marino, M. A. & Prenger, V. L. A sensitive denaturing gradient-gel electrophoresis assay reveals a high frequency of heteroplasmy in hypervariable region 1 of the human mtDNA control region (2000) *Am J Hum Genet* **67**, 432-443.

19 Huhne, J., Pfeiffer, H., Waterkamp, K. & Brinkmann, K. Mitochondrial DNA in human hair shafts--existence of intra-individual differences? (1999) *Int J Legal Med* **112**, 172-175.

20 Lee, H. Y., Chung, U., Park, M. J., Yoo, J. E., Han, G. R. & Shin, K. J. Differential distribution of human mitochondrial DNA in somatic tissues and hairs (2006) *Ann Hum Genet* **70**, 59-65.

21 Lutz, S., Weisser, H. J., Heizmann, J. & Pollak, S. Mitochondrial heteroplasmy among maternally related individuals (1999) *Int J Legal Med* **113**, 155-161.

22 Paneto, G. G., Martins, J. A., Longo, L. V., Pereira, G. A., Freschi, A., Alvarenga, V. L., Chen, B., Oliveira, R. N., Hirata, M. H. & Cicarelli, R. M. Heteroplasmy in hair: differences among hair and blood from the same individuals are still a matter of debate (2007) *Forensic Sci Int* **173**, 117-121.

23 Prieto, L., Alonso, A., Alves, C., Crespillo, M., Montesino, M., Picornell, A., Brehm, A., Ramirez, J. L., Whittle, M. R., Anjos, M. J., Boschi, I., Buj, J., Cerezo, M., Cardoso, S., Cicarelli, R., Comas, D., Corach, D., Doutremepuich, C., Espinheira, R. M., Fernandez-Fernandez, I., Filippini, S., Garcia-Hirschfeld, J., Gonzalez, A., Heinrichs, B., Hernandez, A., Leite, F. P., Lizarazo, R. P., Lopez-Parra, A. M., Lopez-Soto, M., Lorente, J. A., Mechoso, B., Navarro, I., Pagano, S., Pestano, J. J., Puente, J., Raimondi, E., Rodriguez-Quesada, A., Terra-Pinheiro, M. F., Vidal-Rioja, L., Vullo, C. & Salas, A. 2006 GEP-ISFG collaborative exercise on mtDNA: reflections about interpretation, artefacts, and DNA mixtures (2008) *Forensic Sci Int Genet* **2**, 126-133.

24 Clayton, T. M., Whitaker, J. P., Sparkes, R. & Gill, P. Analysis and interpretation of mixed forensic stains using DNA STR profiling (1998) *Forensic Sci Int* **91**, 55-70.

25 Barbaro, A. & Cormaci, P. DNA analysis from mixed biological materials (2004) *Forensic Sci Int* **146 Suppl**, S123-125.

26 Tomsey, C. S., Kurtz, M., Flowers, B., Fumea, J., Giles, B. & Kucherer, S. Case work guidelines and interpretation of short tandem repeat complex mixture analysis (2001) *Croat Med J* **42**, 276-280.

27 Gill, P., Sparkes, R., Pinchin, R., Clayton, T., Whitaker, J. & Buckleton, J. Interpreting simple STR mixtures using allele peak areas (1998) *Forensic Sci Int* **91**, 41-53.

28 A, B. AmpFISTR Profiler Plus PCR Amplification Kit User's Manual. (1998).

29 Gill, P., Brenner, C. H., Buckleton, J. S., Carracedo, A., Krawczak, M., Mayr, W. R., Morling, N., Prinz, M., Schneider, P. M. & Weir, B. S. DNA commission of the International Society of Forensic Genetics: Recommendations on the interpretation of mixtures (2006) *Forensic Sci Int* **160**, 90-101.

30 Gill, P., Curran, J., Neumann, C., Kirkham, A., Clayton, T., Whitaker, J. & Lambert, J. Interpretation of complex DNA profiles using empirical models and a method to measure their robustness (2008) *Forensic Sci Int Genet* **2**, 91-103.

31 Ladd, C., Lee, H. C., Yang, N. & Bieber, F. R. Interpretation of complex forensic DNA mixtures (2001) *Croat Med J* **42**, 244-246.

32 Wang, T., Xue, N. & Birdwell, J. D. Least-square deconvolution: a framework for interpreting short tandem repeat mixtures (2006) *J Forensic Sci* **51**, 1284-1297.

33 Bill, M., Gill, P., Curran, J., Clayton, T., Pinchin, R., Healy, M. & Buckleton, J. PENDULUM--a guideline-based approach to the interpretation of STR mixtures (2005) *Forensic Sci Int* **148**, 181-189.

34 Buckleton, J. S., Curran, J. M. & Gill, P. Towards understanding the effect of uncertainty in the number of contributors to DNA stains (2007) *Forensic Sci Int Genet* **1**, 20-28.

35 Andreasson, H., Nilsson, M., Budowle, B., Frisk, S. & Allen, M. Quantification of mtDNA mixtures in forensic evidence material using pyrosequencing (2006) *Int J Legal Med* **120**, 383-390.

36 Hancock, D. K., Tully, L. A. & Levin, B. C. A Standard Reference Material to determine the sensitivity of techniques for detecting low-frequency mutations, SNPs, and heteroplasmies in mitochondrial DNA (2005) *Genomics* **86**, 446-461.

37 Reynolds, R., Walker, K., Varlaro, J., Allen, M., Clark, E., Alavaren, M. & Erlich, H. Detection of sequence variation in the HVII region of the human mitochondrial genome in 689 individuals using immobilized sequence-specific oligonucleotide probes (2000) *J Forensic Sci* **45**, 1210-1231.

38 Stewart, J. E., Aagaard, P. J., Pokorak, E. G., Polanskey, D. & Budowle, B. Evaluation of a multicapillary electrophoresis instrument for mitochondrial DNA typing (2003) *J Forensic Sci* **48**, 571-580.

39 Chong, M. D., Calloway, C. D., Klein, S. B., Orrego, C. & Buoncristiani, M. R. Optimization of a duplex amplification and sequencing strategy for the HVI/HVII regions of human mitochondrial DNA for forensic casework (2005) *Forensic Sci Int* **154**, 137-148.

40 Asari, M., Azumi, J., Shimizu, K. & Shiono, H. Differences in tissue distribution of HV2 length heteroplasmy in mitochondrial DNA between mothers and children (2008) *Forensic Sci Int* **175**, 155-159.

41 Nelson, K. & Melton, T. Forensic mitochondrial DNA analysis of 116 casework skeletal samples (2007) *J Forensic Sci* **52**, 557-561.

42 Rasmussen, E. M., Sorensen, E., Eriksen, B., Larsen, H. J. & Morling, N. Sequencing strategy of mitochondrial HV1 and HV2 DNA with length heteroplasmy (2002) *Forensic Sci Int* **129**, 209-213.

43 Stewart, J. E., Fisher, C. L., Aagaard, P. J., Wilson, M. R., Isenberg, A. R., Polanskey, D., Pokorak, E., DiZinno, J. A. & Budowle, B. Length variation in HV2 of the human mitochondrial DNA control region (2001) *J Forensic Sci* **46**, 862-870.

44 Yoshida, K., Sekiguchi, K., Mizuno, N., Kasai, K., Sakai, I., Sato, H. & Seta, S. The modified method of two-step differential extraction of sperm and vaginal epithelial cell DNA from vaginal fluid mixed with semen (1995) *Forensic Sci Int* **72**, 25-33.

45 Danielson, P. B., Shelton, R. J. & LaBerge, G. S. Clinical applications of denaturing high-performance liquid chromatography-based genotyping (2003) *Croat Med J* **44**, 447-454.

46 Danielson, P. B., Sun, H. Y., Melton, T. & Kristinsson, R. Resolving mtDNA mixtures by denaturing high-performance liquid chromatography and linkage phase determination (2007) *Forensic Sci Int Genet* **1**, 148-153.

47 Kristinsson, R., Lewis, S. E. & Danielson, P. B. Comparative analysis of the HV1 and HV2 regions of human mitochondrial DNA by denaturing high-performance liquid chromatography (2009) *J Forensic Sci* **54**, 28-36.

48 Budowle, B., Bieber, F. R. & Eisenberg, A. J. Forensic aspects of mass disasters: strategic considerations for DNA-based human identification (2005) *Leg Med (Tokyo)* **7**, 230-243.

49 Fondevila, M., Phillips, C., Naveran, N., Fernandez, L., Cerezo, M., Salas, A., Carracedo, A. & Lareu, M. V. Case report: identification of skeletal remains using short-amplicon marker analysis of severely degraded DNA extracted from a decomposed and charred femur (2008) *Forensic science international. Genetics* **2**, 212-218.

50 Budowle, B. SNP typing strategies (2004) *Forensic Sci Int* **146 Suppl**, S139-142.

51 Budowle, B. & van Daal, A. Forensically relevant SNP classes (2008) *Biotechniques* **44**, 603-608, 610.

52 Butler, J. M., Shen, Y. & McCord, B. R. The development of reduced size STR amplicons as tools for analysis of degraded DNA (2003) *J Forensic Sci* **48**, 1054-1064.

53 Coble, M. D. & Butler, J. M. Characterization of new miniSTR loci to aid analysis of degraded DNA (2005) *J Forensic Sci* **50**, 43-53.

54 Divne, A. M. & Allen, M. A DNA microarray system for forensic SNP analysis (2005) *Forensic Sci Int* **154**, 111-121.

55 Hill, C. R., Kline, M. C., Coble, M. D. & Butler, J. M. Characterization of 26 miniSTR loci for improved analysis of degraded DNA samples (2008) *J Forensic Sci* **53**, 73-80.

56 Sobrino, B., Brion, M. & Carracedo, A. SNPs in forensic genetics: a review on SNP typing methodologies (2005) *Forensic Sci Int* **154**, 181-194.

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

58 Egeland, T., Dalen, I. & Mostad, P. F. Estimating the number of contributors to a DNA profile (2003) *International journal of legal medicine* **117**, 271-275.

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

60 Coble, M. D., Vallone, P. M., Just, R. S., Diegoli, T. M., Smith, B. C. & Parsons, T. J. Effective strategies for forensic analysis in the mitochondrial DNA coding region (2006) *Int J Legal Med* **120**, 27-32.

61 Irwin, J. A., Saunier, J. L., Strouss, K. M., Sturk, K. A., Diegoli, T. M., Just, R. S., Coble, M. D., Parson, W. & Parsons, T. J. Development and expansion of high-quality control region databases to improve forensic mtDNA evidence interpretation (2007) *Forensic Sci Int Genet* **1**, 154-157.

62 Niederstatter, H., Coble, M. D., Grubwieser, P., Parsons, T. J. & Parson, W. Characterization of mtDNA SNP typing and mixture ratio assessment with simultaneous real-time PCR quantification of both allelic states (2006) *Int J Legal Med* **120**, 18-23.

63 Parsons, T. J. & Coble, M. D. Increasing the forensic discrimination of mitochondrial DNA testing through analysis of the entire mitochondrial DNA genome (2001) *Croat Med J* **42**, 304-309.

64 Vallone, P. M., Just, R. S., Coble, M. D., Butler, J. M. & Parsons, T. J. A multiplex allele-specific primer extension assay for forensically informative SNPs distributed throughout the mitochondrial genome (2004) *Int J Legal Med* **118**, 147-157.

65 Mardis, E. R. The impact of next-generation sequencing technology on genetics (2008) *Trends Genet* **24**, 133-141.

66 Metzker, M. L. Sequencing technologies - the next generation (2010) *Nat Rev Genet* **11**, 31-46.

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

68 Morozova, O. & Marra, M. A. Applications of next-generation sequencing technologies in functional genomics (2008) *Genomics* **92**, 255-264.

69 Thudi, M., Li, Y., Jackson, S. A., May, G. D. & Varshney, R. K. Current state-of-art of sequencing technologies for plant genomics research (2012) *Briefings in functional genomics* **11**, 3-11.

70 Qin, N., Li, D. & Yang, R. [Next-generation sequencing technologies and the application in microbiology--a review] (2011) *Wei Sheng Wu Xue Bao* **51**, 445-457.

71 Aburatani, H. [Cancer genome analysis through next-generation sequencing] (2011) *Gan To Kagaku Ryoho* **38**, 1-6.

72 Navin, N. & Hicks, J. Future medical applications of single-cell sequencing in cancer (2011) *Genome medicine* **3**, 31.

73 Goldberg, S. M., Johnson, J., Busam, D., Feldblyum, T., Ferriera, S., Friedman, R., Halpern, A., Khouri, H., Kravitz, S. A., Lauro, F. M., Li, K., Rogers, Y. H., Strausberg, R., Sutton, G., Tallon, L., Thomas, T., Venter, E., Frazier, M. & Venter, J. C. A Sanger/pyrosequencing hybrid approach for the generation of high-quality draft assemblies of marine microbial genomes (2006) *Proc Natl Acad Sci U S A* **103**, 11240-11245.

74 Yergeau, E., Lawrence, J. R., Sanschagrin, S., Waiser, M. J., Korber, D. R. & Greer, C. W. Next-generation sequencing of microbial communities in the Athabasca River and its tributaries in relation to oil sands mining activities (2012) *Appl Environ Microbiol* **78**, 7626-7637.

75 Fierer, N., Breitbart, M., Nulton, J., Salamon, P., Lozupone, C., Jones, R., Robeson, M., Edwards, R. A., Felts, B., Rayhawk, S., Knight, R., Rohwer, F. & Jackson, R. B. Metagenomic and small-subunit rRNA analyses reveal the genetic diversity of bacteria, archaea, fungi, and viruses in soil (2007) *Applied and environmental microbiology* **73**, 7059-7066.

76 Costello, E. K., Lauber, C. L., Hamady, M., Fierer, N., Gordon, J. I. & Knight, R. Bacterial community variation in human body habitats across space and time (2009) *Science* **326**, 1694-1697.

77 Green, R. E., Krause, J., Ptak, S. E., Briggs, A. W., Ronan, M. T., Simons, J. F., Du, L., Egholm, M., Rothberg, J. M., Paunovic, M. & Paabo, S. Analysis of one million base pairs of Neanderthal DNA (2006) *Nature* **444**, 330-336.

78 Reich, D., Green, R. E., Kircher, M., Krause, J., Patterson, N., Durand, E. Y., Viola, B., Briggs, A. W., Stenzel, U., Johnson, P. L., Maricic, T., Good, J. M., Marques-Bonet, T., Alkan, C., Fu, Q., Mallick, S., Li, H., Meyer, M., Eichler, E. E., Stoneking, M., Richards, M., Talamo, S., Shunkov, M. V., Derevianko, A. P., Hublin, J. J., Kelso, J., Slatkin, M. &

Paabo, S. Genetic history of an archaic hominin group from Denisova Cave in Siberia (2010) *Nature* **468**, 1053-1060.

79 Briggs, A. W., Good, J. M., Green, R. E., Krause, J., Maricic, T., Stenzel, U., Lalueza-Fox, C., Rudan, P., Brajkovic, D., Kucan, Z., Gusic, I., Schmitz, R., Doronichev, V. B., Golovanova, L. V., de la Rasilla, M., Fortea, J., Rosas, A. & Paabo, S. Targeted retrieval and analysis of five Neandertal mtDNA genomes (2009) *Science* **325**, 318-321.

80 Green, R. E., Briggs, A. W., Krause, J., Prufer, K., Burbano, H. A., Siebauer, M., Lachmann, M. & Paabo, S. The Neandertal genome and ancient DNA authenticity (2009) *Embo J* **28**, 2494-2502.

81 Prufer, K., Stenzel, U., Hofreiter, M., Paabo, S., Kelso, J. & Green, R. E. Computational challenges in the analysis of ancient DNA (2010) *Genome Biol* **11**, R47.

82 Berglund, E. C., Kiialainen, A. & Syvanen, A. C. Next-generation sequencing technologies and applications for human genetic history and forensics (2011) *Investig Genet* **2**, 23.

83 Bandelt, H. J. & Salas, A. Current next generation sequencing technology may not meet forensic standards (2012) *Forensic Sci Int Genet* **6**, 143-145.

84 Holland, M. M., McQuillan, M. R. & O'Hanlon, K. A. Second generation sequencing allows for mtDNA mixture deconvolution and high resolution detection of heteroplasmy *Croat Med J* **52**, 299-313.

85 Bornman, D. M., Hester, M. E., Schuetter, J. M., Kasoji, M. D., Minard-Smith, A., Barden, C. A., Nelson, S. C., Godbold, G. D., Baker, C. H., Yang, B., Walther, J. E., Tornes, I. E., Yan, P. S., Rodriguez, B., Bundschuh, R., Dickens, M. L., Young, B. A. & Faith, S. A. Short-read, high-throughput sequencing technology for STR genotyping (2012) *BioTechniques* **0**, 1-6.

86 Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y. J., Chen, Z., Dewell, S. B., Du, L., Fierro, J. M., Gomes, X. V., Godwin, B. C., He, W., Helgesen, S., Ho, C. H., Irzyk, G. P., Jando, S. C., Alenquer, M. L., Jarvie, T. P., Jirage, K. B., Kim, J. B., Knight, J. R., Lanza, J. R., Leamon, J. H., Lefkowitz, S. M., Lei, M., Li, J., Lohman, K. L., Lu, H., Makhijani, V. B., McDade, K. E., McKenna, M. P., Myers, E. W., Nickerson, E., Nobile, J. R., Plant, R., Puc, B. P., Ronan, M. T., Roth, G. T., Sarkis, G. J., Simons, J. F., Simpson, J. W., Srinivasan, M., Tartaro, K. R., Tomasz, A., Vogt, K. A., Volkmer, G. A., Wang, S. H., Wang, Y., Weiner, M. P., Yu, P., Begley, R. F. & Rothberg, J. M. Genome sequencing in microfabricated high-density picolitre reactors (2005) *Nature* **437**, 376-380.

87 ten Bosch, J. R. & Grody, W. W. Keeping up with the next generation: massively parallel sequencing in clinical diagnostics (2008) *J Mol Diagn* **10**, 484-492.

88 Bar, W., Brinkmann, B., Budowle, B., Carracedo, A., Gill, P., Holland, M., Lincoln, P. J., Mayr, W., Morling, N., Olaisen, B., Schneider, P. M., Tully, G. & Wilson, M. DNA Commission of the International Society for Forensic Genetics: guidelines for mitochondrial DNA typing (2000) *International journal of legal medicine* **113**, 193-196.

89 Carracedo, A., Bar, W., Lincoln, P., Mayr, W., Morling, N., Olaisen, B., Schneider, P., Budowle, B., Brinkmann, B., Gill, P., Holland, M., Tully, G. & Wilson, M. DNA Commission of the International Society for Forensic Genetics: guidelines for mitochondrial DNA typing (2000) *Forensic Sci Int* **110**, 79-85.

90 Wilson, M. R., Allard, M. W., Monson, K., Miller, K. W. & Budowle, B. Recommendations for consistent treatment of length variants in the human mitochondrial DNA control region (2002) *Forensic Sci Int* **129**, 35-42.

91 Andreasson, H., Nilsson, M., Styrman, H., Pettersson, U. & Allen, M. Forensic mitochondrial coding region analysis for increased discrimination using pyrosequencing technology (2007) *Forensic Sci Int Genet* **1**, 35-43.

VI. Dissemination of Research Findings

A. Publications:

1. Book Chapter

Cassandra D. Calloway, PhD and Henry Erlich, PhD. “Evolving Technologies in Forensic DNA Analysis”. *Forensic DNA Applications: An Interdisciplinary Perspective,* Editors Dragan Primorac and Moses Schanfield. Taylor & Francis Group (2014)

2. Master's Thesis

Daniela Cuenca. “Optimization and Validation of a Probe Capture/NGS Assay for Sequencing the Whole Mitochondrial Genome on Forensically Relevant Samples.” Master's Thesis (2013). University of California, Davis, Forensic Science Graduate Group

B. California Criminalists Institute 2 Week Intensive Laboratory and Lecture Course

CCI Course R602 Mitochondrial DNA Analysis: Linear Array, Sanger Sequencing, and Next-generation Sequencing. California Criminalists Institute continuing education 2 week lecture and lab course covering mtDNA biology, applications of mtDNA analysis to forensics, Linear Array methods and interpretation issues, Sanger sequencing methods, Next-generation sequencing (NGS) chemistry and methods, 8 day hands on lab and training on mtDNA HVI/HVII 454 NGS PCR assay, 454 sequencing, software training and data interpretation. Developed course materials and binder including laboratories protocols, training exercises, laboratory training materials, software demonstrations and webinars.
Instructor: Cassandra Calloway, PhD; Assistant Instructors: Hanna Kim and Daniela Cuenca
California DOJ Jan Bashinski DNA Laboratory, Richmond, CA
December 3-5, 2013; January 22-24; January 27-31, 2014 (84 instruction hours)

C. Software
MIA available upon request from Dr. Richard Green UC, Santa Cruz.
NextGENe Version 2.4.0_02172014 beta version available for demo upon request from SoftGenetics, LLC .

D. Presentations

1. **Resolution of DNA Mixtures and Analysis of Degraded DNA Using the 454 DNA Sequencing Technology.**
   Cassandra D. Calloway, PhD, Saloni Pasta, PhD, Christina Parodi, Damian Goodridge, PhD, and Henry A. Erlich, PhD.
   Poster presentation as DNA grantee at the annual NIJ conference. Arlington, VA, June 20-22nd, 2011

2. **Resolution of DNA and Degraded DNA by Next-Generation Sequencing.**
   Henry Erlich and Cassandra Calloway.
   Invited Speaker at the 7th ISABS conference. Bol, Croatia, June 20-24th, 2011

3. **The development of a sensitive assay for mitochondrial and nuclear DNA analysis using next-generation sequencing for forensic and genetic applications.**
   Saloni Pasta, Christina Parodi, Henry Erlich and Cassandra Calloway.
   Poster presentation at the 39th annual ASCLD Symposium.
   Denver, CO September 18-22nd, 2011

4. **Decoding Next-Generation Sequencing for Forensic Applications.**

Cassandra D. Calloway
Invited Speaker at the California Department of Justice Lunch and Learn Series
Richmond, CA October 6th, 2011

5. **mtDNA Analysis Using the 454 Next-Generation Sequencing Technology.**

Cassandra D. Calloway
Invited Speaker at the California Association of Criminalists DNA Workshop
Sacramento, CA October 25th, 2011

6. **A Sensitive Next-Generation Sequencing Assay for Resolution of Mixtures and Analysis of Forensic Samples.**

Cassandra D. Calloway, PhD, Saloni Pasta, PhD, Christina, Parodi, BS, and Henry Erlich, PhD
oral presentation at the annual AAFS Conference
Atlanta, GA February 20-25th, 2011

7. **Resolution of DNA Mixtures Using 454 Next-Generation Sequencing**

Cassandra D. Calloway, PhD
Invited Speaker at the Green Mountain DNA conference
Burlington, VT, August 1-3, 2012

8. **PCR and Sequence Capture Enrichment Assays for Sequencing mtDNA**

Cassandra D. Calloway
Invited lecture CAL DOJ seminar

Richmond, CA August 9th, 2012

9. **Resolution of DNA Mixtures Using 454 Next-Generation Sequencing**

Cassandra Calloway
Invited Lecture NYOCME seminar
NY, NY September 24th, 2012

10. **Analysis of mtDNA Heteroplasmy in Human Hair from Monozygotic Twins**

Amy Yam, Sarah Stuart, and Cassandra Calloway
Oral presentation at Midwestern Association of Forensic Science Fall meeting
Milwaukee, WI September 27th, 2012

11. **PCR and Probe Based Enrichment Assays for Deep Sequencing mtDNA**

Cassandra D. Calloway, Valarie McClain, George Sensabaugh, Henry Erlich
Oral presentation at the ISHG conference
Nashville, TN October 15-18th, 2012

12. **PCR and Probe Based Enrichment Assays for Deep Sequencing mtDNA**

Valerie McClain, Cassandra Calloway, George Sensabaugh
Oral Presentation at California Association of Criminalists conference
San Jose, CA November 8th, 2012

13. **Forensic Applications of mtDNA Typing and Sequencing Analysis**

Cassandra D. Calloway
Invited lecture for UC Davis Forensic Science Graduate Seminar Series
Davis, CA, November 29th, 2012

14. **Whole Mitochondrial Genome Sequence Analysis Using Probe Capture and 454 Next Generation Sequencing**

Valerie McClain
Master's Thesis Presentation UC, Davis
Davis, CA December 6th, 2012

15. **Whole Mitochondrial Genome Sequencing Using Probe Capture and 454**

Valerie McClain, Cassandra Calloway, Daniela Cuenca, George Sensabaugh
Accepted for oral presentation at the AAFS Annual meeting
Washington, DC, February 18-23rd, 2013

16. **Analysis of mtDNA Mixtures Using 454 Sequencing**

Cassandra D. Calloway, PhD
Invited speaker at the DNA workshop at the121st CAC Spring Seminar, Pasadena, CA, May 21st, 2013

17. **Optimization of a Probe Capture/NGS Assay for Sequencing the Whole Mitochondrial Genome of Limited and Degraded Samples**

Daniela Cuenca, Valerie McClain, George Sensabaugh, Cassandra Calloway

Poster presentation at the 121st CAC Spring Seminar, Pasadena, CA, May 20th–24th 2013

18. **Mitochondrial DNA Analysis of Complex Mixtures and Heteroplasmy Using 454 Next Generation Sequencing**

Hanna Kim and Cassandra Calloway
Poster presentation at the 121st CAC Spring Seminar, Pasadena, CA, May 20th–24th 2013

19. **Optimization and Validation of a Probe Capture/NGS Assay for Sequencing the Whole Mitochondrial Genome on Forensically Relevant Samples**

Daniela Cuenca
Thesis presentation for Master's of Science in the Forensic Science Graduate Group, UC Davis, Davis, CA, June 5th, 2013

20. **Analysis of DNA Mixtures and Degraded DNA by Clonal Sequencing**

Cassandra D. Calloway, Hanna Kim, Valarie McClain, Daniela Cuenca, George Sensabaugh, and Henry Erlich
Invited speaker and distinguished faculty at the 8th ISABS Conference in Forensic, Anthropologic and Medical Genetics and Mayo Clinic Lectures in Translational Medicine, Split, Croatia, June 24th-28th 2013

21. **Next-Generation Sequencing (NGS) and Forensic DNA Analysis**

Mark Timken
CADOJ BFS, May 7, 2013;

22. **The Basics of Next-Generation Sequencing (NGS)**

Mark Timken
Oral presentation at CAC Meeting
Modesto, CA, October 21, 2013).