THE PENNSYLVANIA STATE UNIVERSITY

SCHREYER HONORS COLLEGE

EBERLY COLLEGE OF SCIENCE FORENSIC SCIENCE PROGRAM

EXAMINING EFFECTS OF PCR CONDITIONS ON MITOCHONDRIAL DNA HETEROPLASMY DETECTION USING NEXT GENERATION SEQUENCING

LAUREN ELIZABETH ROTHWELL

Spring 2012

A thesis
submitted in partial fulfillment
of the requirements
for a baccalaureate degree
in Forensic Science
with honors in Forensic Science

Reviewed and approved* by the following:

Mitchell M. Holland
Director, Forensic Science
Associate Professor, Biochemistry and Molecular Biology
Thesis Supervisor & Honors Adviser

Ronald Porter
Director, Graduate Studies
Associate Professor of Microbiology and Molecular Genetics
Faculty Reader

* Signatures are on file in the Schreyer Honors College

ABSTRACT

Aim To examine the effects of varying PCR conditions on the detection and differentiation of low level mtDNA heteroplasmy and the artificial production of chimeric sequences when performing deep pyrosequencing.

Methods A hypervariable segment (HV1) of the mtDNA control region was analyzed from several individuals using the 454 GS Junior instrument. Similar to previous research experiments, mock mixtures of two individuals were analyzed to evaluate low level heteroplasmy and deconvolute mtDNA mixtures. During previous experiments (Holland et al 2011), the mtDNA mixtures showed chimeric sequences, or sequences that contained the known polymorphisms of both individuals in the mixture. The current study used alterations in first round amplification conditions to reduce or eliminate chimeric sequences in order to more accurately examine low level heteroplasmic variants. Amplicon sequencing was performed on PCR products generated with fusion primers that included multiplex identifiers (MID) and adaptors for pyrosequencing. Data analysis was performed using Next*GENe*® software.

Results Dilution experiments showed a decrease in observed chimeric sequences at 10x dilution. Using input DNA concentrations less than the protocol's optimal range showed higher percentages of chimeric reads than 1:10 dilution. Decreasing input DNA to these concentrations resulted in little correlation of dilution and observed chimeric minor reads. Increasing the first round PCR extension time by one minute resulted in decreased chimeric reads at both 28 cycles (protocol) and 38 cycles (hair shaft protocol). Notable increases in chimeric reads were seen at 38 cycles.

Conclusions Changing the first round PCR settings such as input DNA concentrations, cycle number and extension times do have an effect on downstream mtDNA sequencing results when using NGS pyrosequencing technology. While increasing cycle number to 38 cycles increases the observance of chimeric minor sequences in a mixture, diluting tenfold and increasing extension time by at least 1 minute decreases total chimeric reads. The highest observed chimeric percentage reached 30%, meaning 30% of the total minor component sequences in the mixture sample showed polymorphisms from both contributors of the mixture. This percentage must be reduced in order to properly analyze minor contributor sequences, especially when low level heteroplasmy is observed and could provide additional information for that sample.

TABLE OF CONTENTS

ACKNOWLEDGMENTS

I would like to thank Dr. Mitchell Holland for his help and support throughout the process of this thesis project. Also, I acknowledge and thank Dr. Robert Shaler and Dr. Walther Parson for their input as integral members of my research committee. Our research group also acknowledges the assistance of John McGuigan at Soft Genetics. Finally, I thank my research colleague and friend, Ms. Kerry McGinley for her support, among everything else. Thank you.

Introduction

Though most forensic DNA laboratories test chromosomal DNA for human identification, there is a need for sequencing mitochondrial DNA (mtDNA) for certain applications. In forensic biology, mtDNA has been routinely utilized for older, highly degraded samples and for samples without detectable amounts of chromosomal DNA. It is common knowledge that a single human cell contains hundreds of copies of mitochondrial DNA, whereas the same cell possesses only one nucleus with two copies of genomic DNA. It is because mtDNA is far more abundant in copy number and more robust than genomic DNA that it can be used for more degraded and smaller samples. Biological evidence that is degraded due to heat, long periods of time or exposure to different environmental insults warrant mtDNA analysis [1-3]. This type of DNA is thus inherently important to the forensic DNA community. It provides additional information about maternal relatedness and is a last resort avenue to travel when all other DNA resources are unavailable or possibly inconclusive.

The presence of oxygen radicals and weak repair mechanisms in the mtDNA replication process leads to a high mutation rate in mitochondrial DNA. Because the genomes are highly mutable, there is a chance that there will be differences in their sequences. Bottlenecking of mtDNA genomes occurs during the development of an individual, resulting in the uneven distribution or ratio of different genomes in the body [1,2]. The detection of more than one mtDNA genome type in an individual is called heteroplasmy, and its presence in evidence samples can add to the utility of samples during analysis.

Heteroplasmy is defined as mixtures of two or more subpopulations of mtDNA molecules occurring within an individual [1]. There are two types of this variation: point heteroplasmy,

which involves single base differences (i.e. single transition or transversion events) and length heteroplasmy, where differences are caused by changes in the number of nucleotides in a given sequence. It is believed that mtDNA does not undergo recombination but is highly variable, with high mutation rates due to oxidative damage and poor repair mechanisms during replication. In contrast, *in vitro* PCR polymerase infidelity occurs through the misincorporation of the incorrect nucleotide. Misincorporation of these bases may also occur during the sequencing reaction. The well studied hypervariable segments HV1 and HV2 of the control region have been the targets of heteroplasmy examination [1, 3-8]. After directly sequencing the mtDNA, it can be difficult to gauge whether or not the mixed sites are artifacts of the *in vitro* PCR amplification or are naturally occurring, inherited base mixtures otherwise known as heteroplasmy [1, 3-5]. With Sanger sequencing the observance of artifacts is minimal because the assay 'averages' the different sequences from each mtDNA molecule to one main sequence. This technology does not have the capability to pull out every mtDNA sequence present in the sample [1]. Examining the thousands of sequences per sample allows for the detection of both lower level heteroplasmic variants and artificially introduced PCR artifacts. The ability to discern the artifacts from point heteroplasmy using Next Generation Sequencing is currently a major goal of our laboratory.

The utility of heteroplasmy includes its abilities to discriminate between mtDNA sequences of maternally linked relatives and help in identifying human remains The first instance of the use of heteroplasmy as an aid in human identification in the world of forensic DNA came in 1994 [9]. The authenticity of the alleged remains of Russia's Tsar Nicholas II was and had been questioned for decades. The forensic DNA community aimed to finally answer this question by comparing the mtDNA of the Tsar's relatives to the unknown sample. During analysis, the mitochondrial sequence of a maternal relative of the Tsar proved to match at every

base except for one location. This raised a question by some members of the forensic DNA community: was contamination to blame? With that news, Russia allowed the body of the late Tsar's brother to be exhumed for further DNA testing to determine if the remains truly were authentic. The mtDNA Sanger sequencing on the Grand Duke Georgij Romanov (Nicholas II's brother) yielded very important results for the field of forensic DNA. The brother possessed the same heteroplasmic site as the Tsar, (albeit in different ratio), thus proving a familial match and introducing the world to heteroplasmy as a useful tool in human identification [9].

The presence of multiple mtDNA genomes can possible both help and hinder forensic human identification testing. Heteroplasmy provides more strength to an inclusion between reference and evidence samples. If both samples shared a heteroplasmic site, it would be like sharing a rare STR allele, something very uncommon in both samples provides a stronger indication of a match. It can also discriminate between maternally related individuals.

Not only are heteroplasmic variant ratios different from person to person, but these have been found to be different from tissue to tissue within an individual, thus requiring more in-depth analysis of the mtDNA sequence from sample to sample by the forensic DNA analyst [10,11]. It has been shown in a number of published forensic articles that the occurrence of heteroplasmy is higher in hair shafts, muscle tissue and tissues with higher metabolic rates than in blood and saliva [11-13]. This kind of information is important to understand when examining biological evidence found at crime scenes. In particular, hair is a challenging specimen to analyze because even a single hair shaft can show different ratios of the heteroplasmic variants in the mtDNA sequence, depending on the hair or section of hair being examined.

The Sanger method is the universal gold standard for most sequencing applications, but the recently developed Next Generation platforms could be the future of forensic DNA analysis, especially for sequencing mtDNA heteroplasmy. There was a growing demand for faster, cheaper and higher throughput DNA sequencing in the world of biotechnology/genomics earlier in the previous decade. To answer this demand, a few companies developed a technology that has undoubtedly changed the landscapes of molecular biology, medicine and now the smaller field of forensic DNA [1,14-16]. Since the advent of this technology, it has been coined 'Next Generation Sequencing' (NGS) or 'Second Generation Sequencing' as it refers to the next or second sequencing methodology post the Sanger era [1, 14-16]. It is a method that greatly increases the throughput and depth of DNA sequencing by using a massive parallel reaction system and micro fluidic technology. Also, the NGS technologies aim to lower costs by increasing throughput and decreasing the number of total sequencing runs per sample [14-16]. Biotechnology companies such as Applied Biosystems, Roche and Illumina have developed their own versions of Next Gen platforms using various advanced systems. These platforms offer slightly different advantages and disadvantages, depending on the application of their use. In a forensic context, the applications vary greatly from STR research to human leukocyte antigen (HLA) compatibility testing to mtDNA sequencing for heteroplasmy [1,14-18]. Due to its higher throughput, this technology may have the potential to expand the size of the control region mtDNA sequence reference database for use in forensic casework [6]. Expanding the size of the current mtDNA database would not only result in a better sampling of populations, but also increase the statistical weight of inclusions during comparison analysis. High throughput, deep coverage and long read length characteristics makes NGS on the pyrosequencing platform suitable for such a forensic application [1, 14-17].

The Roche/454 Life Science platform has been utilized by our lab to detect mtDNA heteroplasmy with excellent results [1]. Next Gen technologies generally have very short read lengths, but the 454 platform has longer reads than the others. Pyrosequencing technologies allow for longer read lengths (400-500 bp), which is advantageous for mtDNA heteroplasmy detection. The ability to use longer amplicons allows more of the hypervariable region 1 or 2 to be examined in a single sequencing reaction, thereby eliminating the need to reconstruct sequences from multiple shorter reads The Next Gen pyrosequencing platform also allows for amplicon sequencing, is a more sensitive assay than Sanger, and subsequently maximizes detection of lower levels of heteroplasmy that are commonly found in forensically relevant biological samples. The previous study [1] showed a sensitivity range down to 0.1% heteroplasmic minor variant, possibly the lowest percentage seen (and is in no way the standard or routine during mtDNA analysis). The Roche/454 GS Junior Titanium pyrosequencing platform boasts the parallel sequencing of 100,000 DNA molecules, which is a one thousand fold increase of output sequences per run between Sanger and NGS [1]. One 8-hour 454 pyrosequencing run is comparable to approximately one thousand 96-well plates using the Sanger sequencing method. With the sequencing depth of a cloning assay and the speed and cost efficiency of only Next Gen technologies, 454 pyrosequencing was the obvious choice for the studies done in this laboratory [1].

Much research has already been done using the Next Generation platforms. This technology has proven to be a large step in the direction of massive parallel cloning assays. Although using Next Gen technologies have already been applied to mtDNA heteroplasmy studies in biological samples [5], much of the research has been focused primarily in the medical field (studying mtDNA disease and cancer [18,19]). The research done in our lab has focused on

the forensic applications. As there has also been research done in examining heteroplasmy in mtDNA in the forensic biology community, much of this research has been limited to using Sanger sequencing only. In the Holland et. al. 2011 paper, our lab expounded upon the use of Second Generation or Next Gen technology for forensic STR analysis and more in depth studies on mtDNA heteroplasmy and its implications for further use in the forensic DNA field.

A current area of interest in our lab also includes the observation of chimeric sequences produced during previous mixture studies while using the Next Gen approach. When two different samples were mixed, an in-depth analysis showed a percentage of mixed species, otherwise known as chimeric sequences. These sequences contained the known polymorphisms of both individuals in the mixture, making analysis more complicated (Refer to Figure 1). In samples where the minor contributor makes up less than 5% of the mixture, it becomes difficult to identify the correct mtDNA sequence if there are too many chimeric sequences. In Sanger sequencing, mixtures of mtDNA sequence would be nearly impossible to analyze. However, the NGS technologies allow for deconvolution of mixtures and for the observation of low level heteroplasmy [1]. Unfortunately, the chimeric sequences caused by the 'jumping PCR' phenomenon may obscure low level heteroplasmic frequencies in mixtures, thus hindering in depth analyses of forensically important samples [1].

Methods & Materials

All work for this study was conducted under The Pennsylvania State University internal review board (IRB) approved project number 32063. Three samples (from the group of 27 total samples used in the previous Holland et al study) were obtained from unrelated individuals in the form of buccal swabs. The samples used in this study were labeled M19, M9 and F13. These samples had very different profiles, with no shared polymorphisms. In order for the deconvolution of these mixtures it was imperative that the samples had no shared polymorphisms, because a shared profile leads to inconclusive results which would not benefit the study.

This study focused only on the hypervariable segment 1 (HV1: 16,024-16,365) of the control region. All primers used for first round PCR and emulsion PCR are listed in Holland et. al. 2011. [1] Multiplex identifiers (MIDs) were used in the same application as the previous experiments, using a total of 5 MIDs for each amplification performed in this study (therefore 5 samples per run).

Genomic DNA was extracted from samples collected in the aforementioned study and quantified using Applied Biosystems Quantifiler assay (Foster City, CA, USA). Those DNA extracts were frozen at 4°C and available to use as new experiments were performed. The starting quantity of input DNA for first round PCR amplification (frPCR) was 20 ng [20], where dilutions were based on the quantification of nuclear DNA. Therefore actual mtDNA quantities were unknown and mixture ratios were determined empirically. Three samples (M19, M9 and F13) were chosen as the components of the mock mixtures. Mixture ratios of approximately 1:5 or 1:100 (i.e. M19:F13) were obtained. The total starting concentration of nuclear DNA was approximately 20 ng for each mixture. Actual mixture ratios were recorded during data analysis.

First round amplifications (frPCR) included 1 x PCR buffer, 10 mM dNTPs, 2.5 units of AmpliTaq Gold polymerase (all Applied Biosystems), and 10 pmols of each primer (Integrated DNA Technologies, Coralville, IA, USA) [20]. PCR cycling conditions varied along with the different experiments performed. In frPCR, both the cycle number and extension times were varied and compared to a control frPCR: 95°C soak for 10 minutes followed by 28 cycles of 94°C for 1 minute, 60°C for 1 minute and 72°C for 1 minute [20]. Following frPCR, the concentration of mtDNA amplicons was estimated for each sample by running 20% of the product on a 2% agarose gel and comparing band intensities to known standards.

Further dilutions for emulsion PCR (emPCR) involved the dilution of samples down to 1 x 10^9 molecules/uL using the following equations: $\frac{molecules}{ul} = \frac{sample\ conc\left[\frac{ng}{ul}\right] x\ 6.022\ x10^{23}}{656.6\ x\ 10^9 x\ 504\ bp\ (amplicon\ length)}$

Then diluting each amplicon separately to 1 x 10^9 molecules in diH_20 by adding 1 µl of each sample in the volume of diH_20 as calculated: $\left(\frac{\frac{molecules}{ul}}{10^9} - 1\right) ul$ [20]. The 5 sample amplicons were pooled equally (either 10 µL each or 1 µL each), and 2 µL of the pooled DNA added into 198 µL deionized water. An additional dilution of 3 µL DNA to 27 µL deionized water was also performed for a final total of 30 uL sample DNA. These dilutions were performed to obtain the correct concentration of DNA for downstream sequencing [1,20-22]. 20 µL of this final dilution mixture was used in the emPCR reaction and subsequently the pyrosequencing assay.

Pyrosequencing Assay

The first step of the Roche/Life Science 454 pyrosequencing process involves targeting the desired amplicons from frPCR with 454 fusion primers that allow the PCR product to be added directly into the emulsion PCR amplification step and sequencing without any other library preparation necessary [20,21].

The amplicons are then attached to beads which are emulsified as PCR reagents are captured in the water-in-oil microreactors. Inside the microreactors clonal amplification occurs. [1,17,21] The microreactor emulsions are broken and the DNA strands are then denatured and enriched for DNA-positive beads during the enrichment process. In the final step, a single amplified and enriched DNA bead is deposited in each well of the picotiter plate and is coated in sulfurylase and luciferase as well as other reagents. Each well is its own sequencing factory [1].

The first deoxyribonucleotide triphosphate (dNTP) is added to the sequencing reaction. The DNA polymerase catalyzes the incorporation of a dNTP into the DNA strand. After each incorporation of a dNTP, an eqimolar quantity of pyrophosphate (PP_i) is released. Sulfurylase converts the PP_i to ATP in the presence of adenosine 5' phosphosulfate (APS). ATP drives the reaction of luciferin to oxyluciferin, resulting in the production of visible light proportional to the amount of ATP [17]. Each sequencing reaction flows one nucleotide across the plate so when light is generated, the system can determine which base was incorporated. The amount of light given off by each well correlates to the number of bases incorporated. [1,17] The bases are flowed over the plate in a specific order (via laminar flow) so the order of the bases can be recorded to determine the sequence. A lens array is used to focus the generated luminescence from each individual well of the microtiter plate onto the chip of a CCD (charge coupled device)

camera. The CCD camera sees the light and records it as a peak in the raw data output, resulting in a Pyrogram. Peak heights in the pyrogram are proportional to the number of nucleotides incorporated during the sequencing event [1,17].

Data Analysis

The data analysis for the 454 sequencing assay was performed using NextGENe™ software (Soft Genetics, State College PA). An SFF file from the Roche 454 GS Jr instrument computer was imported to the software for sequencing run analysis. This analysis method was also used in the previous study (Holland et al 2011a). Samples were sorted by their multiplex identifier (MID) through the software so that samples could be analyzed separately. The number of total reads per sample was recorded, actual mixture ratios were recorded, and the number of possible chimeric minor reads were counted and divided by total number of minor reads. These numbers were counted manually.

Experimental Methods

Run 1

The first run performed on the GS Junior instrument was a practice run to ensure similar results could be obtained by a new user. The samples used for this run are unrelated to this study and were therefore not listed in the results of experiment.

Run 2

Sensitivity experiments were performed for three reasons: 1. To test out the lower range (nanograms to picograms) of the 454 GS Junior instrument, beyond manufacturer's protocol of 20 ng nucDNA; 2. To simulate real casework sample concentrations and 3. To determine if decreased input DNA reduces presence of chimeric mtDNA sequences. Estimated total input nucDNA ranged from 20 ng to 20 pg, with sample mixture ratios of approximately 1:5 and 1:100. (See Table 1). These samples were amplified at 28 cycles with the protocol PCR conditions except for the input DNA concentration.

Table 1: Run 2 Parameters

MID	Mixture Ratio	Samples	Dilution	Input nucDNA (ng)	Extension Time (min)
1	1:5	M19:F13	none	20	1
9	1:100	M19:F13	none	20	1
10	1:100	M19:F13	1:10	2	1
11	1:100	M19:F13	1:100	0.2	1
12	1:100	M19:F13	1:1000	0.02	1

Runs 3 and 4

Runs 3 and 4 were duplicate experiments that examined varying PCR conditions, performed in order to observe how changing the initial PCR step affects the presence of chimeric sequences. The first condition altered from the control frPCR protocol [20] was increasing the cycle number to 38 cycles [23]. This was done to simulate casework sample PCR when the sample is either a hair shaft or is degraded. Additionally, certain samples were subjected to an increased extension time of 2 minutes. Increasing the extension was hypothesized to increase the efficiency of the polymerase during frPCR by allowing the polymerase more time to complete the synthesis of the amplicon. A more efficient polymerase could result in decreased artifacts and chimeric sequences.

Table 2: Run 3 Parameters

MID	Mixture Ratio	Samples	Dilution	frPCR Cycles	Input nucDNA (ng)	Extension Time (min)
1	1:100	M19:F13	1:10	28	2	1
9	1:100	M19:F13	1:10	38	2	1
10	1:100	M19:F13	1:10	38	2	2
11	1:100	M19:F13	1:100	38	0.2	1
12	1:100	M19:F13	1:100	38	0.2	2

Table 3: Run 4 Parameters

MID	Mixture Ratio	Samples	Dilution	frPCR Cycles	Input nucDNA (ng)	Extension Time (min)
1	1:100	M19:F13	1:10	28	2	1
9	1:100	M19:F13	1:10	38	2	1
10	1:100	M19:F13	1:10	38	2	2
11	1:100	M19:F13	1:100	38	0.2	1
12	1:100	M19:F13	1:100	38	0.2	2

Run 5

Run 5 was another sensitivity experiment performed to test out the limitations of the 454 GS Junior pyrosequencing assay. This experiment used dilutions ranging from 1:10 to 1:2500 (of the protocol's 20 ng input nucDNA). Therefore the range of input nucDNA was approximately 2 ng to 8 pg. The samples were each given a MID primer and the following descriptions:

Table 4: Run 5 Parameters

MID	Mixture Ratio	Samples	Dilution	frPCR Cycles	Input nucDNA (ng)	Extension Time (min)
1	None	M9	None	28	20	1
9	1:100	M9:F13	1:100	28	0.2	1
10	1:100	M9:F13	1:250	28	0.08	1
11	1:100	M9:F13	1:1000	28	0.02	1
12	1:100	M9:F13	1:2500	28	0.008	1

Runs 6 and 7

Following the initial five runs, two optimization runs were performed. This took into consideration both the diluted input DNA concentration as well as the frPCR conditions. This run was done to optimize the input DNA concentration at the normal frPCR at 28 cycles. (Table 3). Finally, the last experiment was run to observe the optimal extension time utilized during frPCR. Using the optimized input nucDNA (2 ng), extension times of 1 minute, 2 minutes 3 minutes and 4 minutes were used in frPCR.

Table 5: Run 6 Parameters

MID	Mixture Ratio	Samples	Dilution	frPCR Cycles	Input nucDNA (ng)	Extension Time (min)
1	None	F13	None	28	20	1
9	1:100	M19:F13	None	28	20	1
10	1:100	M19:F13	1:10	28	2	2
11	1:100	M19:F13	None	28	20	1
12	1:100	M19:F13	1:10	28	2	2

Table 6: Run 7 Parameters

MID	Mixture Ratio	Samples	Dilution	frPCR Cycles	Input nucDNA (ng)	Extension Time (min)
1	None	F13	None	28	20	1
9	1:100	M19:F13	1:10	28	2	1
10	1:100	M19:F13	1:10	28	2	2
11	1:100	M19:F13	1:10	28	2	3
12	1:100	M19:F13	1:10	28	2	4

Results

Experiments performed using the GS Junior instrument were carried out in several runs with five samples in each run. Every sample has its own multiplex identifiers (MIDs). All samples are labeled as one of the following: MID 1, MID 9, MID 10, MID 11 and MID 12 (Roche). A total of 7 runs were performed. Run 1 (not listed) was used as a practice run. Dilution experiments were performed in Runs 2 and 5, and PCR condition experiments were carried out during Runs 3 and 4. Finally, Run 6 was included as an optimization study.

Each run was analyzed using Next GENe software (Soft Genetics LLC, State College PA). The number of reads per sample differed depending on the type of sample (i.e. mixture, dilution, etc) and from run-to-run. Tables 7-12 summarize the quantitative information obtained from each run. The tables include a description of the sample, total number of reads in the run, percentage of minor component, number of reads per sample and the percentage of minor reads that appear to be chimeric. The empirically determined mixture ratios were analyzed by recording the actual percentage of the minor component. Ideally this percentage should be 1.0% for a 1:100 mixture and 20% for a 1:5 mixture. These percentages ranged from 0.45% to 4.13% for the 1:100 mixtures. Total number of reads per run varied from approximately 16,500 to 51,000. Total number of reads per sample ranged from 879 to 19,299. This high variability could possibly be due to dilution errors or stochastic sampling.

The major reason for quantifying the chimeric minor reads was to ensure the correct minor sequence variants were identified. When the percentage of chimeric minor sequences is elevated (around 30% of the total minor reads), analysis must be done with caution so that the heteroplasmic variants and their ratios are correct. The known heteroplasmic variant in sample

F13 was also examined in each mixture to ascertain the variability from sample to sample. Throughout the study, this position (16192C) showed to be heteroplasmic in all instances, but the ratios varied slightly. The percentage of low level heteroplasmy in F13 ranged from 1.43% to 3.85%. In the previous study, this site ranged from 2.64- 4.50%.

The mixture samples diluted to different concentrations were analyzed to record quantitative and qualitative information. For these experiments, input nucDNA concentrations ranged from approximately 20 nanograms to 8 picograms. Possible chimeric sequences were determined by recording the known mtDNA profiles of the samples and comparing those to the thousands of reads (or sequences). When a sequence showed a mixture of the known polymorphisms for each sample (i.e. samples M19 and F13), that read was recorded as a chimeric read (Refer to Figure 1). The number of possible chimeric sequences was divided by the number of minor component sequences and those percentages were recorded.

In Run 2 (Table 7), input nucDNA ranged from 20 nanograms to 0.02 nanograms, where the dilutions ranged from no dilution to 1000 fold. In this run, the more diluted samples actually showed higher percentages of observed chimeric minor reads. In fact, as the samples decreased in input DNA, the percentage of chimeric sequences increased. MID 1 of Run 2 was a 1:5 mixture and showed the lowest percentage of chimeric minor reads. The previous study performed by this lab [26] also found samples with 1:5 mixture ratios had fewer observed chimeric sequences. All further runs performed for this study used samples at 1:100 mixture ratios.

Run 5 (refer to Table 10) input nucDNA ranged from 20 nanograms to 8 picograms, therefore dilutions ranged from no dilution to 2500 fold. In this run, the dilutions did not

correlate with the percentage of observed chimeric minor sequences. The 1000x dilution once again showed the highest percentage, however the 2500x dilution showed the lowest percentage of chimeric reads. Overall, the percentages of chimeric reads were similar to Run 2 percentages. Both Runs 2 and 5 began with a 28 cycle frPCR.

Two instrument runs were dedicated to testing different PCR parameters. Runs 3 and 4 incorporated change in cycle number from 28 to 38 and increasing extension time from 1 minute to 2 minutes. Additionally, three different input concentrations were chosen: 20 ng, 2 ng and 0.2 ng. As shown in tables 8 and 9, these runs showed similar chimeric sequence results. It is clear that increasing cycle number to 38 drastically increases chimeric reads overall. These tables also illustrate that lengthening the extension time to 2 minutes at 38 cycles decreases number of chimeric reads. Also, both input concentrations showed decrease in chimeric reads when the extension time was increased. In both runs however, the 2 ng (10x dilution) mixture samples had the fewest chimeric reads at 38 cycles.

Table 7: Run 2 Results

RUN 2	MID 1	MID 9	MID 10	MID 11	MID 12
Description	1:5 mixture no dilution	1:100 no dilution	1:100 mixture; 1:10 dilution	1:100 mixture 1:100 dilution	1:100 mixture 1:1000 dilution
Reads	3551	4126	3520	4477	879
Minor Contributor %	NA	3.47%- 4.13%	2.39%- 2.44%	1.79%- 1.85%	0.82-1.32%
Chimeric Sequences/Total Minor Contributor Reads	5.56%	7.84%	9.76%	13.47%	17.46%
Run	**Total Reads: 16,553**				

Table 8: Run 3 Results

RUN 3	MID 1	MID 9	MID 10	MID 11	MID 12
Description	1:100 mixture 1:10 dilution 28 cycles	1:100 mixture 1:10 dilution 38 cycles 1 min ext.	1:100 mixture 1:10 dilution 38 cycles 2 min. extension	1:100 mixture 1:100 dilution 38 cycles 1 min. extension	1:100 mixture 1:100 dilution 38 cycles 2 min extension
Reads	19299	2106	1527	1748	2068
Minor Contributor %	1.91-1.94%	1.61-1.85%	1.70-2.36%	1.49-2.52%	1.93-2.22%
Chimeric Sequences/Total Minor Contributor Reads	3.04%	23.30%	16.50%	29.30%	21.30%
Run	**Total Reads: 26,748**				

Table 9: Run 4 Results

RUN 4	MID 1	MID 9	MID 10	MID 11	MID 12
Description	1:100-10 mixture; 28 cycles 'control	1:100-10 mixture; 38 cycles	1:100-10 mixture; 38 cycles 2 min. extension	1:100-100 mixture; 38 cycles	1:100-100 mixture, 38 cycles with 2 min extension
Reads	10627	13842	9095	10504	6987
Heteroplasmy	2.63%	2.98%	3.85%	2.72%	2.21%
Minor Contributor %	2.41-2.61%	1.54-2.55%	2.40%-2.70%	1.05%-1.78%	1.12%-1.57%
Chimeric Sequences/Total Minor Contributor Reads	only 1 chimeric read observed	27.20%	12.43%	33.14%	15.84%
Run	**Total Reads: 51,055**				

Table 10: Run 5 Results

RUN 5	MID 1	MID 9	MID 10	MID 11	MID 12
Description	M9 'control'	1:100 mixture 1:100 dilution	1:100 mixture 1:250 dilution	1:100 mixture 1:1000 dilution	1:100 mixture 1:2500 dilution
Reads	12418	12180	14084	4888	3218
Heteroplasmy	NA	2.73%	2.82%	1.43%	2.67%
Minor Contributor %	NA	Range: 1.04%-2.73%	1.48%-2.82%	0.45%-1.82%	1.14%-2.67%
Chimeric Sequences/Total Minor Contributor Reads	NA	7.52%	4.41%	14.89%	1.83%
Run	**Total Reads: 46,788**				

Table 11: Run 6 Results

RUN 6	MID 1	MID 9	MID 10	MID 11	MID 12
Description	no mixture no dilution 28 cycles	1:100 mixture no dilution 28 cycles 1 min ext.	1:100 mixture 1:10 dilution 28 cycles 1 min. extension	1:100 mixture no dilution 28 cycles 2 min. extension	1:100 mixture 1:10 dilution 28 cycles 2 min extension
Reads	4283	6953	2774	4422	2811
Minor Contributor %	NA	2.44-2.55%	2.63-2.67%	2.01-2.13%	2.56-2.67%
Chimeric Sequences/Total Minor Contributor Reads	NA	29.71%	21.62%	20.88%	9.46%
Run	**Total Reads: 21,243**				

Discussion

The initial results of the dilution studies seemed counter intuitive. The original hypothesis regarding the input nucDNA concentration stated that decreasing the amount of DNA in the PCR reaction would allow the polymerase to synthesize to completion. Removing the excess molecules in the reaction would thus help increase the efficiency of the polymerase and possibly reduce the unwanted artifacts associated with the polymerase. However, the samples diluted 1000 fold showed the highest percentage of chimeric reads. Going outside the Roche protocol's recommended input DNA concentration may have had some impact on the results. At these lower concentrations, the number of observed chimeric reads showed no trends or patterns consistent with the original hypothesis. While the original hypothesis suggested that 100 fold dilutions may show fewer chimeric sequences than the samples diluted 10 fold, the results disproved that theory. In Run 2, decreasing the DNA concentration below recommended values increased mixed sequences, however Run 5 showed more stochastic results. A 1:2500 dilution performed in Run 5 (or 8 picograms of input DNA) appeared to decrease the unwanted sequences; however, this dilution was carried out once during this study. This should be replicated to confirm such a result.

Although the 2 nanogram sample mixtures varied from run to run, overall they had the lowest average chimeric minor read percentage with 8.78% (28 cycles) and 19.86% (38 cycles). In runs 3 and 4 these samples showed the lowest chimeric read percentages 3.04% and 0.009% (Runs 3 and 4 respectively). Also, the 2 nanogram samples had fewer chimeric reads than the 0.2 nanogram samples at 38 frPCR cycles in Runs 3 and 4. The same trend was seen at the normal

extension time and the lengthened extension time in both runs. This concentration was therefore used in the optimization studies at 28 cycles. In Run 6, input nucDNA amounts of 2 ng and 20 ng were compared at 28 cycles during frPCR. Once again, an additional 1:10 dilution from the protocol 20 ng decreased the overall chimeric reads. This held true for samples with extension times at 1 minute and 2 minutes. When diluting 10x, the observed chimeric reads dropped from 29.71% to 21.62% at the 1 minute extension and dropped from 20.88% to 9.46% at the 2 minute extension. These results indicate that the additional extension time has a more significant impact in the overall reduction of chimeric minor reads.

Coming across hair shafts as evidence in a forensic investigation is quite common. The protocol for sequencing the mitochondrial DNA found in hair shafts calls for increasing the cycle number because the amount of DNA most likely will be less than optimal. In this study, increasing the frPCR cycle number to 38 dramatically increased the observed chimeric sequences. Run 4 showed jumps in chimeric read percentages from less than 1% at 28 cycles to 33% at 38 cycles. If it is absolutely necessary to use a high cycle number during analysis, it is better to lengthen the extension time to at least 2 minutes. The results of the experiment showed increasing extension times at 38 cycles decreases the unwanted mixed sequences.

The final frPCR parameter examined was the length of the extension time. It was hypothesized that increasing the extension time at 38 cycles would allow the polymerase to complete the synthesis of the targeted amplicons before the next round at denaturing temperatures. The results do show that an increased extension time leads to reduction in chimeric minor sequences. This was true for both sets of diluted samples in Runs 3 and 4, as seen in tables 8 and 9 respectively. Because these results were duplicated, the hypothesis was then applied to the optimization studies. If increasing the extension times at 38 cycles was successful, then

perhaps the same would hold true for 28 cycles in frPCR. Another hypothesis applied to the extension optimization study was to increase the times even further to 3 minutes and 4 minutes.

In Run 6 (Table 11), increasing the extension time to 2 minutes during frPCR did decrease the number of chimeric minor reads. At 20 ng input DNA (protocol amount), lengthening the extension time by one minute decreased the percentage of chimeric reads by approximately 9%. At 2 ng, the total decreased by roughly 12%. An increased extension time coupled with a 1:10 dilution showed the largest difference in percent of chimeric reads: 20 ng at 1 minute extension showed 29.71% chimeric minor reads while 2 ng at a 2 minute extension showed approximately 9.5% chimeric minor reads, a difference of more than 20%. These results should be replicated in future studies in order to provide more data points from which to draw conclusions.

Because increasing the extension time seemed to have more of an effect on the chimeric reads than the 1:10 dilution, an additional extension optimization experiment was performed. If doubling the extension time decreased the chimeric reads by roughly 10%, the hypothesis was that tripling and quadrupling the extension time would decrease these artifacts even more. The results of this experiment are still to be determined.

Next Gen Sequencing technology, specifically on the Roche pyrosequencing platform allowed for the generation of copious amounts of data for this study. Many tens of thousands of sequences were analyzed by performing only several instrument run, while cloning experiments followed by sequencing via the Sanger method would have taken an enormous amount of time and tens of thousands of Sanger reactions. That being said, the newer technology does have room for improvement. As discussed in Holland et. al. 2011, the pyrosequencing platform experiences

difficulty in successfully sequencing through homopolymeric stretches of five or more bases, typically cytosines in mtDNA. This phenomenon was observed again during this study, making the analysis of polymorphisms in these regions more difficult, as it commonly resulted in single base deletions.

Additional studies using the Roche/454 GS Junior instrument are needed to enhance the findings of this study as well as the previous studies done in our laboratory. It is imperative that the error rates of this technology are quantified. Any experiments to help pinpoint where artifacts are being introduced into the mtDNA sequences would also be important. Additional replication studies could be performed to solidify these results. Comparison studies with alternative NGS platforms (e.g. Illumina's HiSeq 2000 and MiSeq and SOLiD 4 from Applied Biosytems) could also be utilized to determine whether this technology is definitely the best fit for these experiments. Because platforms such as Illumina's HiSeq 2000 uses different sequencing chemistry than the 454 GS Jr, it may not encounter the same difficulties in homopolymeric stretches in the mtDNA sequence. As technology advances, newer methods could be used and compared. Using Next Gen technologies is most certainly the future of mtDNA sequencing. Improving these platforms and reducing unwanted artifacts is paramount to the successful use of NGS in forensic biology.

References

[1] Holland MM, McQuillan MR, O'Hanlon KA. Second generation sequencing allows for mtDNA mixture deconvolution and high resolution detection of heteroplasmy. Croat Med J. 2011 Jun 15;52(3):299-313.

[2] Holland, M. M., Parsons, T. J.. Mitochondrial DNA Sequence Analysis-Validation and Use for Forensic Casework. Forensic Science Review, 1999, 11, 22-50

[3] White HE, Durston VJ, Seller A, Fratter C, Harvey JF, Cross NCP. Accurate Detection and Quantitation of Heteroplasmic Mitochondrial Point Mutations by Pyrosequencing. *Genet. Test* 2003, .(3):190-199;

[4] Li, M., Schonberg, A., Stoneking, M., Shaefer, M., et. al. Detecting Heteroplasmy from High-Throughput Sequencing of Complete Mitochondrial DNA Genomes. The American Journal of Human Genetics, 2010, 87, 237-249

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

[6] Irwin, J.A., Parson, W., b, Coble, M.D, Just, R.S, mtGenome reference population databases and the future of forensic mtDNA analysis, *Forensic Sci. Int. Genet.* 2010

[7] W. Parson, H.-J. Bandelt , Extended guidelines for mtDNA typing of population data in forensic science. *Forensic Science International: Genetics,* 2007 13–19

[8] Brandstatter, A., Niederstatter H., Parson, W. Monitoring the inheritance of heteroplasmy by computer-assisted detection of mixed base call in the entire human mitochondrial DNA control region. *Int J Legal Med* (2004) 118 : 47–54

[9] Ivanov, P., Wadhams.M., Roby, R., Holland, M., Weedn, V., and Parsons, T. Mitochondrial DNA sequence heteroplasmy in the Grand Duke of Russia Georgij

Romanov establishes authenticity of the remains of Tsar Nicholas II. *Nature Genetics.* April. Vol. 12. 1996. 417-420.

[10] Tully LA, Parsons TJ, Steighner RJ, Holland MM, Marino MA, Prenger VL. A Sensitive Denaturing Gradient-Gel Electrophoresis Assay Reveals a High Frequency of Heteroplasmy in Hypervariable Region 1 of the Human mtDNA Control Region; *Am J Hum Gen* 67:432-443; 2000.

[11] Roberts, K. A., Calloway, C. Characterization of Mitochondrial DNA Sequence Heteroplasmy in Blood Tissue and Hair as a Function of Hair Morphology. *J Forensic Sci*, 2011 Jan. 56, 11

[12] Paneto GG, Martins JA, Longo LV, Pereira GA, Freschi A, Alvarenga VL, Chen B, Oliveira RN, Hirata MH, Cicarelli RM. Heteroplasmy in hair: differences among hair and blood from the same individuals are still a matter of debate**.** Forensic Sci Int. 2007 Dec 20;173(2-3):117-21

[13] Calloway, C.D, et al The Frequency of Heteroplasmy in the HVII Region of mtDNA Differs across Tissue Types and Increases with Age. *Am. J. Hum. Genet.*, 2010, 66:1384–1397

[14] Mikkelsen M, Rockenbauer E, Wächter A, Fendt L, Zimmermann B, Parson W, Nielsen SA,Gilbert T, Willerslev E, Morling N. Application of full mitochondrial genome sequencing using 454 GS FLX pyrosequencing; *For Sci Inter:Genetics*, 2009 2:518-519;

[15] Hert, D.G., Fredlake, C. P, Barron, A.E. Advantages and limitations of next-generation sequencing technologies: A comparison of electrophoresis and non-electrophoresis methods. *Electrophoresis* 2008, 29, 4618-4626

[16] Rothberg, J. M., Leamon, J. H., The development and impact of 454 sequencing. *Nature Biotechnology* 2008, 26, 1117-1124

[17] Mostafa, R. Pyrosequencing Sheds Light on DNA Sequencing. *CSH Press Genome Research*, 2001, 11, 3-11

[18] Morozova, O., Marra, M.A., Applications of next-generation sequencing technologies in functional genomics. *Genomics*, 2008, 92, 255-264

[19] Voelkerding, K. V., Dames, S. A., Durtschi, J. D., Next-Generation Sequencing: From Basic Research to Diagnostics *Clin Chem*, 2009, 55, 641-658

[20] Amplicon Library Preparation Method Manual; GS Junior Titanium Series, Roche. 454 Sequencing May 2010 1-9

[21] emPCR Amplification Method Manual – Lib A; GS Junior Titanium Series, Roche. 454 Sequencing May 2010 1-14

[22]Sequencing Method Manual; GS Junior Titanium Series, Roche. 454 Sequencing. May 2010 1-26

[23] Mitochondrial DNA Amplification and Sequencing Protocol. PSU Protocol #35 Version #1

ACADEMIC VITA

Education

The Pennsylvania State University
Eberly College of Science
The Schreyer Honors College

Integrated Undergraduate Graduate (IUG) Program

Masters of Professional Study Graduating May 2012
Forensic Science
Bachelor of Science Degree Graduating May 2012
Forensic Science Biology Option

Work Experience

Course Assistant FRNSC 201W August 2009 – Present
Penn State Forensic Science Department
107 Whitmore Laboratory University Park, PA 16802

Course Assistant FRNSC 200 August 2010 - Present
Penn State World Campus

Forensic Technician Intern July 2010
Montgomery County Coroner's Office
425 Swede St. Norristown, PA 19401

IUG Research Project

Title: Examining Chimeric Sequences and Low Level Heteroplasmy in the mtDNA Genome using Next Generation Sequencing Technology

Details: Using Roche/454's NGS pyrosequencing platform via the GS Junior Titanium sequencing instrument, work is being done to minimize the presence of PCR induced chimeric amplicons. Through the manipulation of PCR conditions, the hope is to minimize chimeric sequences and examine true heteroplasmic variants in forensically relevant samples.

Related Skills and Knowledge

Criminalistics

- trace evidence skills such as: fingerprinting, evidence collection, microscopy, fiber analysis, crime scene investigation techniques

Forensic Biology

- serology laboratory techniques such as: evidence screening , quality assurance/quality control techniques, presumptive and confirmatory testing

Forensic DNA

- DNA laboratory skills such as: DNA extraction, DNA quantification, STR analysis, statistical analysis, and work with Sanger and NGS mtDNA sequencing

General Laboratory skills

- chemistry, microbiology, forensic trace evidence, biochemistry and molecular biology, forensic biology, and forensic DNA

Independent graduate/undergraduate research experience

Course assistant experience

Internship experience

Leadership experience

- Penn State Forensic Science Club Vice President

Instrumentation & Software Knowledge

- ABI 3130XL
- GS Junior Titanium Sequencer
- Applied Biosystems® 7500 Real-Time PCR System
- GeneAmp™ PCR System 9700
- GeneMarker™ Version 1.7
- GeneMapper IDX ™
- NextGENe ™
- Sequencher™
- Mutation Surveyor™

Awards and Scholarships

- Dean's List
    - Spring 2010, Fall 2010, Spring 2011, Fall 2011
- Eberly College of Science Scholarship
- The Wieland Scholarship
- Unger A & A Memorial Scholarship
- Schreyer Honors College Academic Scholarship