UNIVERSIDAD DE GRANADA

TESIS DOCTORAL

High Throughput Mitochondrial DNA Analysis

Optimization of Sequence Chemistry, Characterization of Local Dye Terminator Sequencing Frames, and Tools for the Development of an Expert System

RHONDA KAY ROBY

11 de Septiembre 2008

Editor: Editorial de la Universidad de Granada
Autor: Rhonda Kay Roby
D.L.: GR. 2089-2008
ISBN: 978-84-691-6330-6

Universidad de Granada

High Throughput Mitochondrial DNA Analysis

Optimization of Sequence Chemistry, Characterization of Local Dye Terminator Sequencing Frames, and Tools for the Development of an Expert System

Memoria presentada por

Rhonda Kay Roby

para optar al grado de Doctor por la Universidad de Granada

Dirigida por los Doctores

José Antonio Lorente Acosta
Juan Carlos Álvarez Merino

11 de Septiembre 2008

JOSÉ A. LORENTE ACOSTA, Profesor Titular de la Universidad de Granada
y JUAN CARLOS ÁLVAREZ MERINO, Titulado Superior Docencia e Investigación de la Universidad de Granada

CERTIFICAN:

Que el trabajo que presenta para aspirar al Grado de Doctor D ª. RHONDA KAY ROBY, titulado: **High Throughput Mitochondrial DNA Analysis: Optimization of Sequence Chemistry, Characterization of Local Dye Terminator Sequencing Frames, and Tools for the Development of an Expert System**, se ha realizado bajo su dirección, y reúne los requisitos académicos, formales y de calidad necesarios como para que pueda ser defendido públicamente ante la Comisión que se constituya al efecto.
Granada, 18 de julio, 2008

Dr. José Antonio Lorente Dr. Juan Carlos Alvarez

High Throughput Mitochondrial DNA Analysis

Optimization of Sequence Chemistry, Characterization of Local Dye Terminator Sequencing Frames, and Tools for the Development of an Expert System

by

Rhonda Kay Roby

A.B. 1985, Washington University, St. Louis, Missouri, United States

M.P.H. 1989, University of California at Berkeley, Berkeley, California, United States

M.S. 2006, University of Granada, Granada, Spain

A Dissertation Submitted to

The Faculty of

University of Granada

in Partial Satisfaction of the Requirements for the

Degree of Doctor of Philosophy

September 11, 2008

Dissertation Directed By

José Antonio Lorente Acosta
and
Juan Carlos Álvarez Merino

To my parents:

Pauline Garner Roby

Richard Ellis Roby

And my brothers:

David Garner Roby

Meir Rick Roby

TO VICTIMS AND THEIR FAMILIES

DEDICATION

I have dedicated the last 20 years of my career to the field of forensic sciences. I have been entrusted with the scientific analysis and review of cases worldwide in parentage testing, human identification of remains, and forensic casework including robbery, sexual assaults, and murders. Further, I have educated scientists, victims' family members, lawyers, and jury panels on the technology used for forensic DNA analysis. I have been on the development teams for new technology that has been adapted in forensic science, on committees focused on quality assurance, and an evaluator of released products.

On April 22, 2008 at a public presentation, I vowed to dedicate the next 20 years of my career to family members to help identify their loved ones through case processing and review and to the future development and advancement of technology to the field of forensic sciences.

The work embodied here is just a small part towards that dedication and to the field that I so love and to the victims and their family members that I will probably never meet but that I vow to give my up most in scientific integrity and ethics in helping in every way that I can for their cases.

ACKNOWLEDGMENTS

I am indebted to many people. I am grateful to those that have influenced me academically, those who have been my constant supporters professionally, and my loved ones. So many people have persuaded me to pursue this aspiration – be it my immediate and distant family, friends, colleagues, or the transitional business traveler at the airport bar.

First and foremost, I want to thank my family. My profession is unique to my family; however, their undying love and support have filled me with energy and enthusiasm.

Next, Arthur J. Eisenberg and Bruce Budowle. It was a simple conversation amongst the three of us and my major advisor that sent me along this wonderful journey. It is with their support and guidance that I have been able to complete this piece of work.

My friends at Granada University. Juan Carlos Álvarez Merino; Miguel C. Botella López, Carmen Entrala Bernal; Francisco Fernández Rosado; Encarni 'Encarniti' García Ruíz; Margarita Jiménez Alcaide; Blanca Gutiérrez Martínez; Esther 'Esther Uno' Martínez Espín, Luis Javier Martínez González; Alberto López Galindo; Olga López Guarnido; and Esther 'Esther Dos' Molina Riva. And my friends at the Residence de "Amor de Dios"

It would be remiss of me not to mention my colleagues at Cellmark Diagnostics, the Department of Defense Armed Forces Institute of Pathology Armed Forces DNA Identification Laboratory, and Applera Corporation. It is through my work and interaction with the scientists and equally important, the administrators, that have shaped me into the scientist and manager I am today. And most importantly, J. Craig Venter. Craig Venter is a friend who pushes the technology and is quite a visionary. This thesis is in large part a direct result of his influence in my work on our team's efforts in identifying the victims from 9/11.

I also want to acknowledge my professors at the University of California at Berkeley. George Sensabaugh and John Thornton gave me my foundation in forensic sciences. Their guidance to their students has shaped many areas of forensic sciences as they are known today. And lastly, Carol A. Langhauser. My friend and lecturer of biostastics. She gave d me the tools in which to analyze this complex and enormous data set. She has been alongside me to the last calculation and evaluation even after being her student over 20 years ago. Never could I have completed this work without her support and critical evaluation of the numerical data.

There are friends and role models that will never know the encouragement they have given me: Deborah Carbullido, Robin Cotton, Michael Koo, Margaret Kuo, John Paul Jones, Demris Lee, Maria Cristina de Mendonça, Susan Narveson, Melissa Smart, Kimberly Stevens, Mark Stolorow, Bridget Tincher, Lois Tully.

I want to thank the scientists at the University of North Texas and Coimbra University who have used techniques and ideas that I have designed. Their willingness to try these techniques and put them into practice is quite rewarding. It has been a wonderful pleasure to work with these talented scientists: Angela Ambers, Filipa Balsa Sa, Suzanne Gonzalez, Maria João Porto, and Jennifer Thomas. Their work is also reflected in this document of which I give my utmost thanks.

And lastly, I must acknowledge my scientific partner and programmer Pedro Fernández de Mendonça. I met Pedro one evening and explained my research interests and struggles with my massive amount of data. He shared with me his skills in programming and data organization. Pedro and I embarked on a scientific expedition that has led to the organization of my data and advanced degrees for us both. It was an instant friendship and one that will last for the rest of our lives.

Finally, my major advisor, José Antonio Lorente and committee. José Lorente welcomed me into his program, his laboratory, his country, and his home. I am grateful that he made this opportunity possible for me. I also extend my deepest gratitude to my dissertation advisory/defense committee members: Enrique Villanueva Cañadas, President; Esther Viseras, Secretary; Francisco Corte-Real; Arthur Eisenberg; and Francisco Etxeberria Gabilondo. Thank you for your time and useful comments. Also, I must thank Margarita Rivera Sánchez and Luis Javier Martínez González. Both Marga and Luis Javier have guided me through the program’s requirements over the past four years and the paperwork in filing for this degree, have opened up their homes, have taught me most of my Spanish, and have been the best of friends.

This has been a fantastic journey. Not just the scientific work captured in this dissertation, but the career and field that I love and the countries and people that I have come to know.

ABSTRACT

Mitochondrial DNA (mtDNA) sequence analysis is a technique that is well-characterized, validated, and useful in the analysis of forensic evidence and identification of human remains. The technique has become more popular as a result of successes in identifying hair samples and older skeletal remains. Mixed DNA samples are very common in casework. Analysis of mixed specimens for nuclear DNA is well-documented and accepted in forensic casework. Additionally, forensic analysts have recently adopted Y-STRs which amplify only the male DNA and are ideal for sexual assault evidence. However, an additional tool that may prove to be very useful in the forensic community would be the analysis of mixed mitochondrial DNA. As of this date, it is not commonplace to report mixed results obtained from mtDNA sequence analysis.

In this study, I will present a procedure to improve sequence quality which reduces background noise and artifacts. Further, I will show the work that has been conducted in conjunction with a software company to automatically output base sequence relative heights. With these output files, I have been able to develop a bioinformatics tool to automatically characterize the patterns seen with the sequence data.

This study focuses on the improved chemistry and output of data using today's higher throughput multicapillary ABI PRISM 31xx instrumentation. This combination of chemistry and instrumentation is routine in the forensic mtDNA sequencing laboratory. An evaluation of the patterns, trends, and statistical analysis of these patterns will be reported. Further, I will present new bioinformatic tools developed for this project to critically evaluate sequence data. This model is used for the statistical evaluations and to demonstrate the non-random patterns displayed in a sequence frame of three bases due to the different rates of incorporation of the different dye chemistries.

Identifying these patterns in local sequence frames from single-source samples can increase base calling accuracy. A single base change can affect the peak heights, or patterns, in these local sequence frames. Patterns in peak heights are characterized using dRhodamine and Big Dye™ terminator sequencing on an ABI PRISM 3100

DNA Sequencer. Now there is an automated system in evaluating these local sequence frames.

With the characterization of the peak patterns, common sequence data can be confirmed with ease. If the sequence of a sample is modified or is different from the defined local sequence frame, then the change may be due to a single base position mutation, or it may be due to a single position that is heteroplasmic, or it may be due to a true mixed sample with multiple mixed positions. More confidence in calling mixed positions can be assured by understanding peak patterns within the local sequence frame in single source samples.

The research from this work could add another tool in the investigation of important criminal matters and identifications. If the forensic scientist could accurately report the results of the mixed sequences for otherwise reported inconclusive data, then the knowledge gained from knowing the different sequences could potentially include or exclude a suspect. Further, identifying the two contributors from a mixed specimen in a mass fatality case would prove that the DNA from both individuals had been extracted. This information could lead investigators to possible locations, e.g., body bags, for further testing and identification of human remains.

In summary, this study will focus on: 1) process development to reduce the time and costs associated with mtDNA processing of forensic samples; 2) a comparison with previous procedures with the results from the more automated process; and 3) the tools developed to critically evaluate sequence data. A study of other sequence data (i.e., non-forensic specimens and non-mtDNA sequence data) will also be included since this overall procedure has applications in other research than just forensic mtDNA analysis. Further, these tools can assist programmers in the development of software intelligent expert systems for a more automated, more accurate base calling method of sequence analysis.

RESUMEN

El análisis de la secuencia de ADN mitocondrial (mtDNA) es una técnica bien caracterizada, validada, y útil en el análisis de evidencias forenses e identificación de restos humanos. La técnica se ha hecho más popular debido a los éxitos obtenidos en la identificación de muestras de pelo y restos antiguos de hueso. Las muestras donde aparece mezcla de ADN son muy comunes en muchos casos. El análisis de mezclas en casos en los que se estudia el ADN nuclear están bien documentados y aceptados en la comunidad forense. Además, los analistas forenses han comenzado a utilizar recientemente los STRs en cromosoma Y, los cuales amplifican sólo el ADN masculino y son ideales para evidencias obtenidas en agresiones sexuales. Sin embargo, un instrumento adicional que puede resultar ser muy útil en la comunidad forense sería el análisis de las mezclas en ADN mitocondrial, que por el contrario, a diferencia del nuclear, hasta ahora se desestiman los resultados obtenidos de mezclas de distintos mtDNA.

En este estudio, presentaré un procedimiento para mejorar la calidad de las secuencias que reduce los ruidos de fondo y artefactos. En adelante, mostraré el trabajo que ha sido realizado junto con una compañía de software para automatizar la altura de la señal obtenida en estudios de secuencias. Con estos datos brutos (output files), he sido capaz de desarrollar un instrumento bioinformatico para caracterizar automáticamente los modelos observados con los datos de secuencias obtenidos.

Mi tesis se centra en la mejora química y el tratamiento de los datos brutos usando los equipos de secuenciación (electroforesis multicapilar) de más alto rendimiento ABI PRISM 31xx. Basándonos en la combinación de química e instrumentación, rutinaria en los laboratorios forenses que realizan estudios de mtDNA, realizaremos una descripción de la evaluación de los modelos, tendencias, y análisis estadísticos de los resultados obtenidos en la secuenciación. Además de esto, presentaré nuevas herramientas bioinformaticas desarrolladas para una evaluación crítica de los datos de secuencia obtenidos. Este nuevo modelo es usado para las evaluaciones estadísticas y demostrar los modelos no arbitrarios que encontramos al estudiar un *marco de secuencia* de tres bases debido a los diferentes porcentajes de incorporación de los distintos terminadores químicos (marcadores).

La identificación de estos modelos de secuencia en *marcos locales,* estudiando cada una de las muestras va a aumentar la reproductividad de estos acercándose a la exactitud. El cambio en una sola base (nucleótido) puede afectar la altura de la señal recibida (altura de picos en electrofluorograma),por tanto, también un cambio en los modelos de estos *marcos locales.* Los modelos de secuenciación para estudiar la señal recibida son caracterizados utilizando dos tipos de marcadores (dos químicas de secuenciación) dRhodamine y Big Dye terminator sequencing y la secuenciación se realizará en un secuenciador ABI PRISM 3100. Tras la secuenciación se aplica un sistema automatizado para la evaluación de estos marcos de secuencia locales.

Con la caracterización de estos modelos de tres bases (señal obtenida), los datos de secuencia comunes pueden ser confirmados con facilidad. Si la secuencia de una muestra es modificada o es diferente del marco de secuencia local definido, entonces el cambio puede ser debido a una mutación de una sola base, o a una heteroplasmica, o a una mezcla de varios ANDs con lo que tendrá muchas posiciones heteroplásmicas. Cuantas más posiciones con mezcla se hayan estudiados más modelos de marcos locales obtendremos y podremos entender y reproducir estos modelos de picos.

La investigación de este trabajo podría añadir otro instrumento para la investigación de casos criminales importantes e identificaciones. Si el científico forense pudiera asegurar exactamente los resultados de las secuencias con mezcla para datos inconcluyentes, entonces el conocimiento ganado al conocer las diferentes secuencias incluida en la mezcla podría inculpar o exculpar a un sospechoso. En adelante, la identificación en casos de catástrofes masivas donde aparecen muestras con mezclas de ADN esta técnica demostraría que el ADN de varios individuos ha sido extraído.

En resumen, este estudio se divide en: 1) optimizar proceso de secuenciación de mtDNA mediante la reducción de los costes y el tiempo; 2) comparar con otros procedimientos los resultados obtenidos con el proceso más automatizado; y 3) evaluación crítica de los instrumentos desarrollados para la edición de la secuencia. Se realizará un estudio de otros datos de secuencia (es decir, no forenses y no exclusivamente para mtDNA), ya que esta técnica podrá ser aplicada en todos los estudios de secuenciación. Además, este estudio puede ayudar a programadores en el desarrollo de software de expert systems más automatizados y más exactos.

TABLE OF CONTENTS

Chapter 1. Introduction

Standard fluorescent sequencing techniques are of tremendous value in producing base sequence information to researchers. These techniques are used to sequence genomes world-wide, to characterize genes, and to target specific regions. Whereas the majority of the work represented here is focused on a specific target, that is, mitochondrial DNA, the techniques discussed in this thesis can be used for fluorescent Sanger DNA sequencing projects globally. In this thesis, I have sought to describe a sequencing technique that I have found to produce beautiful sequence traces that are less expensive to generate than traditional sequencing protocols.

Mitochondrial DNA (mtDNA) sequence analysis is a technique that is well-characterized, validated, and useful in the analysis of forensic evidence and identification of human remains. The technique has become more popular as a result of successes in identifying hair samples and older skeletal remains. In contrast to fragment analysis, the laboratory demands of sequencing can be laborious, time-consuming, and expensive. To address the increasing demand for base sequence information and to help reduce costs, a number of steps have been put into practice. All steps in this procedure are amenable to higher throughput operations and can be easily implemented into a robotic workflow. Robotic liquid handling techniques can help ensure consistency in pipetting and increase throughput capabilities in a laboratory. I also introduce other time-saving and cost-saving techniques to characterize mtDNA results for quantification and as a screening tool.

This study focuses on the improved chemistry and output of data using today's higher throughput multicapillary ABI PRISM 31xx instrumentation. I have executed several steps for the high throughput sequencing while maintaining the quality of previous procedures. These steps include the adoption of BigDye® Terminator v.1.1 Cycle Sequencing Kits (Applied Biosystems, Foster City, CA, USA) to replace of dRhodamine Terminator Cycle Sequencing Kits (Applied Biosystems); reduction of dye chemistry kit consumption by using a sequence enhancing and dilution buffer; and a simple bead purification method to remove unincorporated BigDye® terminators. In this study, this procedure also improves sequence quality which reduces background noise and artifacts.

I have characterized the sequence data with respect to patterns that are observed within a frame of three bases. Additionally, the analysis of sequence data can be laborious, time-consuming, and thus, expensive with respect to the time that is required for scientific review. Further, I will show the work that has been conducted in conjunction with a software company to automatically output base sequence relative heights. With these output files, I have been able to develop a bioinformatics tool to automatically characterize the patterns seen with the sequence data. I present many software tools that when combined together and with the characterization of patterns, could produce an expert system for sequence analysis.

An evaluation of the patterns, trends, and statistical analysis of these patterns will be reported. Further, I will present new bioinformatic tools developed for this project to critically evaluate sequence data. This model is used for the statistical evaluations and to demonstrate the non-random patterns displayed in a sequence frame of three bases due to the different rates of incorporation of the different dye chemistries, a term I have coined as “sequence biometrics.”

Identifying these patterns in local sequence frames from single-source samples can increase base calling accuracy. A single base change can affect the peak heights, or patterns, in these local sequence frames. Patterns in peak heights are characterized using Big Dye™ and dRhodamine terminator sequencing on a ABI PRISM 31xx DNA Sequencer. I introduce an automated system in evaluating these local sequence frames.

With the characterization of the peak patterns, common sequence data can be confirmed with ease. If the sequence of a sample is modified or is different from the defined local sequence frame, then the change may be due to background noise and interference, a single base position mutation, a single position that is heteroplasmic, or a true mixed sample with multiple mixed positions. More confidence in calling mixed positions can be assured by understanding peak patterns within the local sequence frame in single source samples. The research from this work could add additional tools in the toolbox for the forensic scientist in the investigation of important criminal matters and identifications, and for the research investigator.

In summary, this study will focus on: 1) process development to reduce the time and costs associated with mtDNA processing of forensic samples; 2) a comparison with previous procedures with the results from the more automated process; and 3) the tools developed to critically evaluate sequence data. Further, a study of other sequence data (i.e., non-forensic specimens and non-mtDNA sequence data) will also be included since this overall procedure has applications in other research than just forensic mtDNA analysis. Further, these tools can assist programmers in the development of software intelligent expert systems for a more automated, more accurate base calling method of sequence analysis.

1.1. General Sequencing

Sequencing is quite commonplace today. Officials of Mars, Incorporated recently announced that they will invest in the sequencing of the cocoa plant to better understand and cultivate this lucrative product (Anonymous 2008). Sequencing is the process of determining the order of nucleotide bases, similar to the order of the music notes in a *concerto*. In the vocabulary of DNA sequencing, there are four bases: adenine, guanine, cytosine, and thymine. The order of these four bases is critical in scientific discoveries and identifications. The order of these bases is fundamental to many basic and applied sciences. Sequencing of DNA is used in research associated with archaeology, anthropology, genetics, biotechnology, molecular biology, forensic sciences, and more. Now, there is a focus on developing new sequencing methodologies that will improve read lengths, throughput, and cost (Chan 2005). DNA sequencing technology, including the chemistry, the equipment, and the software, has significantly advanced in recent years. With these new techniques, new discoveries and medical advances will surely accelerate.

Frederick Sanger and his colleagues invented the Sanger method, or chain-termination method. Sanger and Coulson’s (Sanger and Coulson 1975) original paper described a method using *Escherichia coli* DNA polymerase I and DNA polymerase from bacteriophage T4 (Englund 1971;Englund 1971;Englund 1972) with different limiting nucleoside triphosphates. The products generated by the polymerases were resolved by ionophoresis on acrylamide gels (França et al. 2002).

Two years later, Sanger and his co-workers described a new method for sequencing oligonucleotides via enzymatic polymerization (Sanger et al. 1977). This method, known as the chain termination method or the dideoxynucleotide method consists of a catalyzed enzymatic reaction that polymerizes the DNA fragments complementary to the template DNA of interest, or the unknown DNA. A ^{32}P-labelled primer anneals to a specific known region on the template DNA, which provides a starting point for DNA synthesis. Catalytic polymerization of deoxynucleoside triphosphates (dNTPs) of the DNA takes place in the presence of

DNA polymerases. The polymerization extends until the enzyme incorporates a modified nucleoside, known as terminator or dideoxynucleoside triphosphate (ddNTPs), into the growing chain of the synthesized DNA.

The original Sanger method was performed in four different tubes, each containing one of the four terminators. The generated fragments have the same 5'-end whereas the residue at the 3'-end are determined by the dideoxynucleotide triphosphates, or ddNTPs, used in the reaction as chain terminators since they lack the 3'-OH group required for the formation of a phosphodiester bond between two nucleotides during strand elongation. After the reactions with the four terminators, the mixture of different-sized DNA fragments are resolved by electrophoresis on a denaturing polyacrylamide gel, usually in four lanes immediately adjacent to the other. The pattern of bands are visualized and read from of the autoradiography which corresponds to the radiolabelled terminated fragments of different lengths in the synthesized strand of DNA.

The Sanger method has served as the foundation for genome sequence production since 1977. This method has produced an abundance of information over the years leading to today's techniques.   Today, fluorescent DNA sequencing technology, which led to other base-determining technologies, has been credited with many of the large-scale sequencing projects, most importantly, to the completion of the human genome sequence, or Human Genome Project (Venter et al. 2001).

Today, fluorescent dye-labelled nucleotides are used instead of radiolabelled nucleotides. With dye-labelled dideoxy terminators, a single reaction tube can be used(Tracy and Mulcahy 1991;Rosenthal and Charnock-Jones 1992). The energy dyes are excited by a laser in the instrumentation and read by the charged-couple device (CCD) optimal camera. These data are then translated into human readable data since one color is associated with one base and can be analyzed by software. Originally, Smith et al. (Smith et al. 1986) designed four different fluorescent dyes that when combined together could be electrophoresed in a single lane and read separately since each of the dyes had its own spectral properties. This method used labels attached at the 5'-end of the primer. The fluorescent light was then separated by four different filters. The fluorescent labels are attached to the ddNTP terminators called dye-labelled terminator chemistry. The fluorescent dyes chosen for use today have their maximum emission spectra relatively even-spaced for better base calling. These techniques have led to high-throughput automated DNA sequencing.

1.2 Mitochondrial DNA

Mitochondria contain their own DNA and are responsible for the bulk of ATP synthesis through oxidative phosphorylation. Mitochondria are often referred to as the energy powerhouse of the cell. They are the site of cellular respiration and capture energy generated by the breakdown of food during the oxidation of simple organic compounds (Copeland 2002). The small number of polypeptides encoded with the mtDNA genome represents only a small fraction of the total proteins necessary for mitochondrial function. Most of these proteins are encoded in the nuclear DNA genome and are subsequently exported to the mitochondria.

Mitochondria were first visualized as discrete cytoplasmic organelles in 1840. These double-membrane organelles were isolated in 1948 using zonal centrifugation techniques. Mitochondria are rod-shaped organelles that are present in all nucleated eukaryotic cells that use oxygen. Mitochondria are approximately 1 to 10 micrometer (μm) in length and approximately 0.5 to 1.0 μm in diameter. Unlike nuclear DNA (nDNA) where there is only one copy from the mother and one copy from the father, most cells contain hundreds to tens of thousand copies of discrete mitochondria ((Robin and Wong 1988). There are exceptions, however, where some cells contain one mitochondrion to other cells containing as many as 100,000 mitochondria.

In the 1960s it was determined that these organelles contain their own DNA. A team of scientists at the Cambridge Research Institute completely sequenced the reported 16,569 bases of the mitochondrial genome (mtGenome) (Anderson et al. 1981). In fact, historically, this is the first component of the human genome that was completely sequenced over 25 years ago. The DNA inside the mitochondrion is circular in structure and double-stranded. Mitochondrial DNA (mtDNA) codes for 13 polypeptides required for oxidative phosphorylation and 22 transfer RNAs and 2 ribosomal RNA subunits (see Figure 1). The heavy strand is purine-rich and the light strand is pyrimidine-rich. Many scientists, especially evolutionary biologists, attribute the mere genetics of the mitochondrion as a primitive aerobic bacterium that was once engulfed by the ancestor of present-day eukaryotic cells (Gray 1992;Grivell 1997).

This closed, double-stranded, circular genome can be classified according to function: the coding region (about 15.5 kb of the genome) and the non-coding control region (about 1.1 kb of the genome). The control region has a regulatory function for the mitochondria and contains sequences to initiate both transcription and DNA replication of the heavy strand. Many laboratories have focused on sequences within the non-coding control region of the mtGenome, specifically hypervariable regions 1 and 2 (HV1 and HV2), since there are a large number of polymorphisms and there is a high degree of variation between individuals. HV1 covers, in theory, positions 16024-16365 and HV2 covers positions 73-340. These positions are ordered according to the origin of replication and numbered according to the published standard reference sequence (Anderson et al. 1981).

The mtGenome is not subjected to recombination during sexual transmission. The mtGenome is strictly maternally inherited; it is passed to the offspring from the oocyte (Giles et al. 1980). That is to say that progeny of both males and females inherit the mtDNA from their mother (barring mutations), whereas only the daughter passes on the mtDNA to the next generation. Differences between two people indicate not sharing a common maternal line since it has been shown that mtDNA does not recombine in humans (Ingman et al. 2000).

The displacement loop, or D-loop, is a non-coding segment of the mtGenome that maintains elements for initiation of transcription and replication but does not code for any gene products. It is the D-loop region of the mtGenome that the forensic community routinely sequences for forensic casework. The D-loop is 1.1 Kb and is often referred to as the control region. Since the D-loop is a non-coding segment of DNA, variability within this region is observed and is not lethal to the growing fetus because of this reason. It is this nucleotide, or base, variability that is significant to the forensic scientist. Differences are observed between individuals not of the same maternal line.

The evolution of mtDNA has been studied in such detail that evolutionary biologists have determined that the Mother Eve, or “mitochondrial Eve,” of all surviving mtDNA profiles lived in Africa between 140,000 and 290,000 years ago (Cann et al. 1987). The low fidelity of mtDNA polymerase and the apparent lack of mtDNA repair mechanisms have led to a higher rate of mutation in the mtGenome as compared to the nuclear genome making it an excellent marker for human evolution research. Some regions of the mtGenome appear to evolve five to ten times the rate of single copy nuclear genes (Brown et al. 1979;Budowle et al. 2000).

Hypervariable Region 1

```
15971            ttaactccac cattagcacc caaagctaag attctaattt aaactattct
16021 ctgttctttc atggggaagc agatttgggt accacccaag tattgactca cccatcaaca
16081 accgctatgt atttcgtaca ttactgccag ccaccatgaa tattgtacgg taccataaat
16141 acttgaccac ctgtagtaca taaaaaccca atccacatca aaaccccctc cccatgctta
16201 caagcaagta cagcaatcaa ccctcaacta tcacacatca actgcaactc caaagccacc
16261 cctcacccac taggatacca acaaacctac ccacccttaa cagtacatag tacataaagc
16321 catttaccgt acatagcaca ttacagtcaa atcccttctc gtccccatgg atgacccccc
16381 tcagataggg gtcccttgac caccatcctc cgtgaaatca atatcccgca caagagtgct
16441 actctcctcg ctccgggccc ataacacttg ggggtagcta aagtgaactg tatccgacat
16501 ctggttccta cttcagggtc ataaagccta aatagcccac acgttcccct taaataagac
16561 atcacgatg
```

Hypervariable Region 2

```
    1 gatcacaggt ctatcaccct attaaccact cacgggagct ctccatgcat ttggtatttt
   61 cgtctggggg gtatgcacgc gatagcattg cgagacgctg gagccggagc accctatgtc
  121 gcagtatctg tctttgattc ctgcctcatc ctattattta tcgcacctac gttcaatatt
  181 acaggcgaac atacttacta aagtgtgtta attaattaat gcttgtagga cataataata
  241 acaattgaat gtctgcacag ccActttcca cacagacatc ataacaaaaa atttccacca
  301 aacccccccct CCCCCgcttc tggccacagc acttaaacac atctctgcca aaccccaaaa
  361 acaaagaacc ctaacaccag cctaaccaga tttcaaattt tatcttttgg cggtatgcac
  421 ttttaacagt caccccccaa ctaacacatt attttcccct cccactccca tactactaat
  481 ctcatcaata caacccccgc ccatcctacc cagcacacac acaccgctgc taaccccata
  541 ccccgaacca accaaacccc aaagacaccc cccaca
```

http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?val=AC_000021

Table 1. Comparisons of nuclear DNA to mitochondrial DNA.

Features	Nuclear DNA	Mitochondrial DNA
Structure	Linear genome	Closed circular genome
Size	3,200,000 Kb	16.5 Kb
Copy Number	1 from mother; 1 from father	100s to 10,000s
Inherited	50% from mother; 50% from father	100% from mother
Ploidy	Diploid	Haploid
Mutation Rate	Low	Higher than nDNA
Recombination	Yes	No

1.2.1 Forensic Applications

There are specific cases in the forensic community where mtDNA analysis is an appropriate method and has played a significant role. For example, with highly charred remains, oftentimes, it is not possible to obtain a full STR profile. Due to the high copy number of mitochondria in a single cell, it is frequently possible to recover a sufficient quantity of mtDNA for analysis. Degraded specimens, either through environmental insults or exposure to chemical challenges, can produce a mtDNA profile when limited or no nuclear DNA profiles are possible. It is possible to obtain full nuclear DNA profiles from hair with root ends attached, fresh bone, fresh fingernails, and obviously stains containing biological materials. However, hair shafts (as small as 1 cm), dried skeletal remains, older fingernails (Anderson et al. 1999) as well as very, very small samples sizes can produce full mitochondrial DNA sequence information when it is a challenge for nuclear DNA testing techniques.

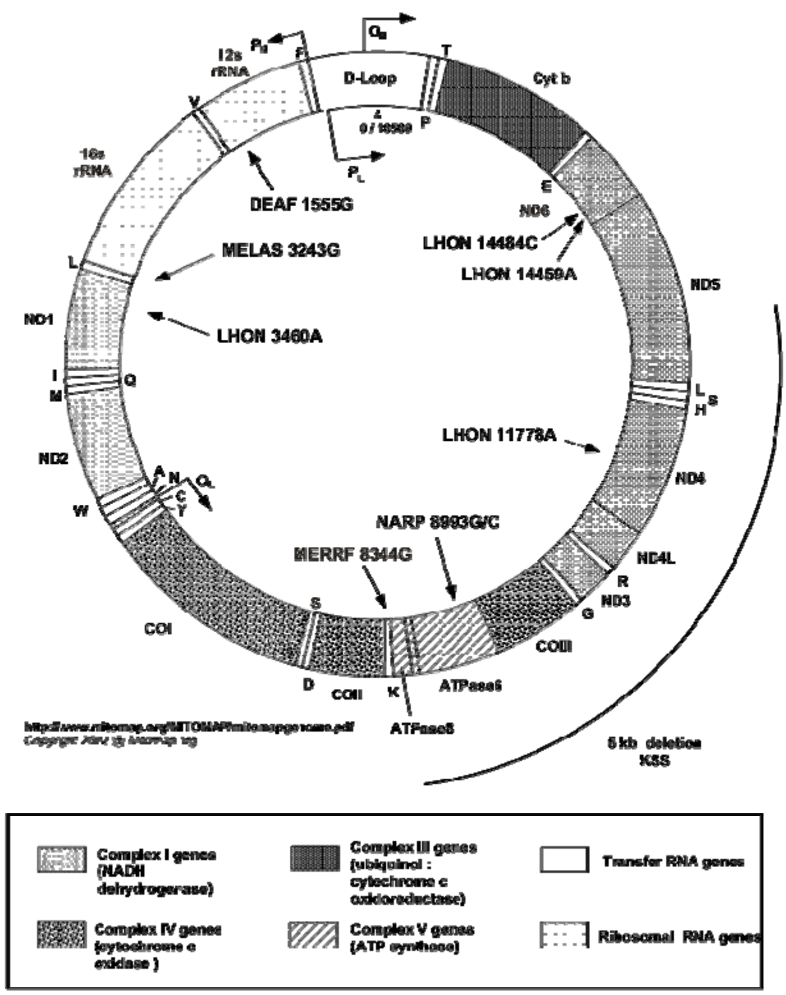
Additionally, mtDNA is quite useful in forensic investigations of remains recovered from a missing person or mass disaster. Biological material from known, maternal relatives, even quite distant, can be used as a reference for direct comparison to the recovered remains.

Variability in the D-loop region is measured for forensic and identification casework. Two regions within this 1.1 Kb fragment demonstrate multiple variations between individuals. The two variable regions are amplified, detected, and analyzed. Other regions within the mtGenome have been successfully analyzed as well (Lutz et al. 2000;Bini et al. 2003).

In Caucasians, the most common mtDNA types with data from HV1 and HV2 occurs in approximately 7% of the population. On average, there are eight nucleotide differences between individuals amongst unrelated Caucasians and 15 nucleotide differences between individuals amongst unrelated Africans (Budowle et al. 1999;Melton et al. 2001).

Unlike STR analysis where discrete alleles according to size are reported, the base sequence information is reported for mtDNA analysis. Insomuch, the standard for the forensic community is to report the sequence information as compared to the revised Cambridge Reference Sequence (rCRS) (Andrews et al. 1999). When the base sequence is the same, then no reference to that particular position is noted. However, if the base sequence at a position is different at one or more positions when compared to the rCRS, then the difference(s) are noted. For a common example of a transition, the rCRS is an A at position 73 (p73) and the sequenced sample is a G at p73, then a report is generated showing an A to G transition at p73.

Mitochondrial DNA (mtDNA) sequencing has been used for over 18 years for sequencing of skeletal remains and hair samples. The majority of forensic laboratories that conduct mtDNA sequencing focus on two regions of the mtGenome: hypervariable regions 1 and 2 (HV1 and HV2, respectively) which consists of approximately 623 bases in the control region. The distribution of mtDNA types is highly skewed toward rare types, making the significance of a match for a mtDNA type previously unseen in a database quite substantial. There are also a number of common types observed in various populations. One limitation of mtDNA testing is the low power of discrimination associated with common HV1 and HV2 types. For example, in the European Caucasian forensic database, there are approximately twenty common HV1 and HV2 types that occur at a population frequency of 0.5% or greater, for an aggregate frequency of about twenty-one percent of the population.



<u>Figure 1. The Human Mitochondrial DNA Genome.</u> The genes encoded by the mitochondrial DNA (mtDNA) genome are noted. Point mutations associated with mitochondrial diseases are noted in the center of the genome. Diagram provided by MitoMap (http://www.mitomap.org/).

1.3 Current Technology and Procedures

The products, regardless of the size, are sequenced using the ABI PRISM® dRhodamine Terminator Cycle Sequencing Kit (Applied Biosystems, Foster City, CA) and purified with Performa® DTR V3 96-Well Short Plates (Edge Biosystems, Gaithersberg, MD) according to the manufacturers' instructions. The dichloro-rhodamine terminators, or dRhodamine terminators, are used for cycle sequencing with a thermostable DNA polymerase and dye-labelled dideoxy chain terminators.

The Performa® DTR plates remove excess dye-labelled terminators (DTR, Dye Terminator Removal Systems), dNTPs, and salts; are optimized for low volumes; and have pre-hydrated matrices for ease of use. The use of the Performa® DTR plates requires the transfer of sequencing products to the plate, centrifugation, and then transfer again to another plate. The mtDNA products are sequenced on the multi-capillary ABI PRISM® 3130*xl* Genetic Analyzer (Applied Biosystems) according to the manufacturer's instructions and the mtDNA sequence data are analyzed using Sequencher™ 4.7 software (Gene Codes, Ann Arbor, MI, USA).

The results in our laboratory using dRhodamine dideoxy terminator chemistry are excellent (see Figure 3, a). Due to the reduced spectra overlap with a narrow dye emission spectra, we produce sequence data with little background noise and high the signal-to-noise ratios. We are very pleased with the results and quality of the sequence data using our current procedures; however, we have been informed by the manufacturer that the ABI PRISM® dRhodamine Terminator Cycle Sequencing Kit will soon be discontinued. Furthermore, we need to implement a more automated and cost-saving procedure for our sequencing protocols in order to meet the expected demand for our respective programs.

Mitochondrial DNA testing is a laborious process with extraction, amplifying minimally two regions, and sequencing using multiple primers in both the forward and reverse directions. The sequence obtained is compared to a standard sequence, the revised Cambridge Reference Sequence (rCRS). Nucleotide differences at base positions are noted and these differences make up an individual's mtDNA profile.

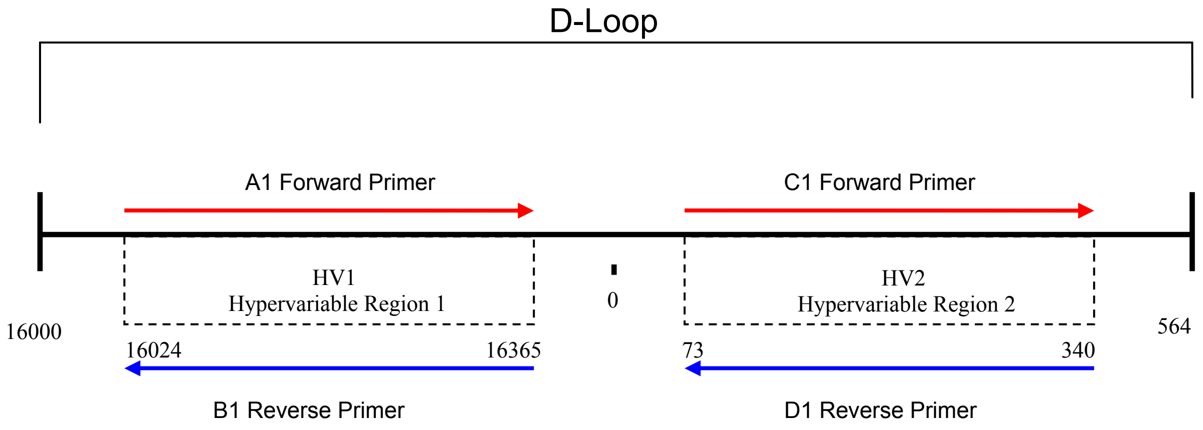


Figure 4. Schematic of mtDNA HV1 and HV2 and their respective amplification primers.

CHAPTER 2. Higher Throughput Laboratory Analysis for Mitochondrial DNA

2.1 Sequence Chemistry Optimization with mtDNA and Other Amplicons

Standard fluorescent sequencing techniques are of tremendous value in producing base sequence information to researchers. Sequencing is laborious, time-consuming, and expensive. In order to address the increasing demand for base sequence information and to help reduce costs, a number of steps have been put into practice. All steps in this procedure are amenable to higher throughput and can be easily implemented into a robotic workflow. Robotic liquid handling techniques can help ensure consistency in pipetting and increase throughput capabilities in a laboratory.

I have executed several steps for the high throughput sequencing while maintaining the quality of previous procedures. These steps include the adoption of BigDye® Terminator v.1.1 Cycle Sequencing Kits (Applied Biosystems, Foster City, CA, USA) to replace of dRhodamine Terminator Cycle Sequencing Kits (Applied Biosystems); reduction of dye chemistry kit consumption by using a sequence enhancing and dilution buffer; and a simple bead purification method to remove unincorporated BigDye® terminators.

The section will focus on: 1) the process developed to reduce the time and costs associated with sequencing; and, 2) a comparison of previous procedures with the results from the more automated process.

MATERIALS AND METHODS

Extraction and Amplification

Mitochondrial DNA

For the mtDNA sequencing results obtained in this study, buccal swabs were pre-processed using the Slicprep™ 96 Device (Promega Corporation, Madison, WI, USA) and the DNA is extracted using the DNA IQ™ System (Promega Corporation) in conjunction with the Tecan Freedom EVO® 100 robot (Tecan Group Ltd., Zurich, Switzerland) using fixed tips (see Section 2.2). The DNA IQ™ System is a system of isolating DNA and producing a consistent DNA concentration from those samples such as reference samples that have an ample amount of starting material. The procedure developed and used in our laboratory maintains an approximate mean of 1.0 ng/µL nuclear DNA (nDNA).

The extracted DNA for buccal swabs is amplified for both nDNA and mtDNA. For mtDNA, both HV1 and HV2 are tested. Following PCR amplification, mtDNA amplicons are quantified with the Agilent 2100 Bioanalyzer (Agilent Technologies, Palo Alto, CA, USA) or 1% agarose mini-gel analysis and visualized with ethidium bromide. Unused primers and dNTPs are removed using ExoSAP-IT® for PCR Product Clean-Up (USB Corporation, Cleveland, OH, USA). By reducing our reaction volume, we were further able to reduce costs by reducing the amount of ExoSAP-IT® for each reaction. Regardless of the cycle sequencing kit or purification reagents used, the above steps are performed with both procedures described in this paper.

New Sequencing Procedure

For the sequencing procedure currently used in our laboratory, see Section 1.3. The goal in evaluating a new sequencing procedure was to implement a more automated procedure for sequencing of mtDNA, reduce costs, implement energy transfer dyes due to the possible discontinuance of the dRhodamineTerminator Cycle Sequencing Kit, all while maintaining or improving the current mtDNA sequence data quality.

For the new sequencing procedure, the mtDNA products are sequenced using the ABI PRISM® BigDye® Terminator v1.1 Cycle Sequencing Kit (Applied Biosystems). The BigDye® Terminators, or energy transfer dichloro-rhodamine dyes, are designed for those products requiring optimal basecalling immediately adjacent to the primer, for short PCR product templates, and for challenging templates. Whereas our samples may not meet all of these design necessities, it is our hope that the same energy transfer dye version can be used for our forensic evidentiary challenged samples. Thus, requiring the design specifications reported by the manufacturer.

The ABI PRISM® BigDye® Terminator v1.1 Cycle Sequencing Kit reaction volume has been reduced to save costs. In order to decrease the manufacturer's terminator sequencing recommended input volume, BetterBuffer (The Gel Company, San Francisco, CA, USA), an enhancement buffer, has been incorporated into the new sequencing procedure. BetterBuffer is a diluting agent that enhances the binding of primers and increases polymerase performance with energy transfer dichloro-rhodamine dyes such as ABI PRISM® BigDye® Terminator v1.1 Cycle Sequencing Kit. By lowering the concentrations of the terminator sequencing reactions, we are able to reduce the terminator sequencing chemistry 5-fold.

As noted in the Extraction and Amplification section, no differences are noted in amplifying the DNA. However, specific quantification using the Agilent 2100 Bioanalyzer of the DNA product has proven to lead to the success of the new sequencing procedure. Since we have dramatically reduced the dye chemistry in a single tube/well, the template used for each reaction must also be remarkably reduced. Quantification of the template is more critical than in the current sequencing procedure and a good reaction cleanup procedure as described below is vital to remove unnecessary salts and unused diluent buffer.

Instead of using the Performa® DTR V3 96-Well Short Plates, BigDye® XTerminator™ Purification Kit (Applied Biosystems) was used to remove the unincorporated BigDye® terminators. It is a simple method of transferring a mixture of XTerminator™ Solution and SAM™ Solution directly to the original 96-well plate (Phenix Research Products, Hayward CA, USA) with strip caps or plate (Axygen, Union City, CA, USA) with sealing film used for the energy transfer dichloro-rhodamine dye sequencing reactions. The plate is sealed, vortexed for 30 minutes, centrifuged at 1000 x *g*, and then placed directly on the sequencer. There is no further transfer of the sample after setting up the sequencing reaction.

A unique run module is used since the BigDye® XTerminator™ Purification Kit reagents have been placed directly in the 96-well plate. The plate is displaced by a few millimeters so the capillaries do not enter directly in the XTerminator™ reagents. The XTerminator™ Solution contains beads to bind to excess terminators and salts. If these beads electrokinetically or mechanically entered into the capillaries, it could cause them to be obstructed.

The XTerminator™ Solution looks for unincorporated dye terminators and free salts from the post-sequencing reaction. The XTerminator™ Solution contains beads that require the use of a large bore pipette tip for transfer. The SAM™

Solution, which contains a detergent, stabilizes the post-purification reactions and enhances the performance of the XTerminator™ Solution.

Cycle sequencing (Chart 1) was performed in the GeneAmp® PCR System 9700 (Applied Biosystems) using ABI PRISM® BigDye® Terminator v1.1 Cycle Sequencing Kit (containing buffer, dNTPs, dye-labelled ddNTPs, and *Taq* polymerase) with BetterBuffer, specific primer, 0.25 ng amplified product, and water for a total of 14.0 uL reagents and 1 uL product using the cycling procedure in Chart 2. After cycling, the plates are centrifuged at 1000 x *g* for 2 min.

BigDye Terminator v.1.1	=	1.0 uL
BetterBuffer	=	5.0 uL
10µM Primer	=	1.5 uL
Sterile water	=	6.5 uL
0.25 ng amp. product	=	1.0 uL
TOTAL	=	15.0 uL

Chart 1. New cycle sequencing procedure.

96˚C, 3:00 min, HOLD

25 cycles of:
96˚C, 0:15 min
50˚C, 0:10 min
60˚C, 3:00 min

4˚C, ∞, HOLD

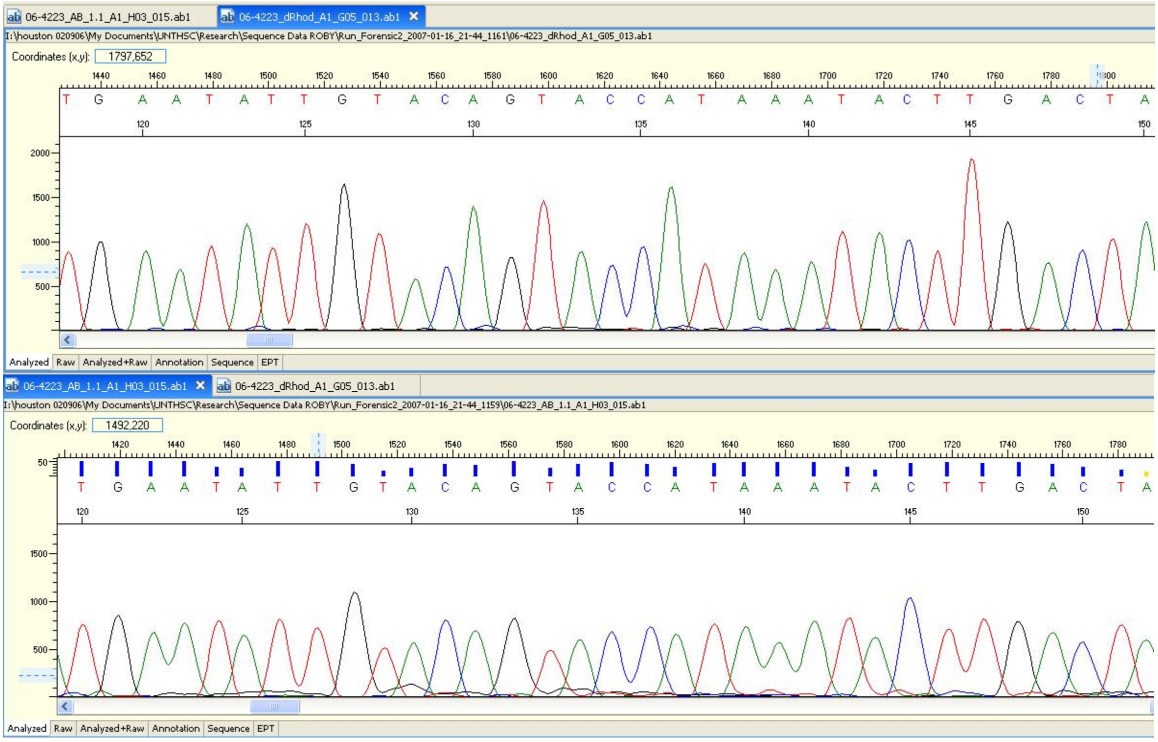
Chart 2. Cycling parameters for new cycle sequencing procedure.

Excess dye-labelled terminators were removed from the extension products by adding 50 uL of the pre-mix XTerminator/SAM solution (Chart 3). Care must be taken when pipetting the XTerminator™ solution, wide bore pipette tips should be used to accurately aspirate the beads that bind to the excess dyes. The pre-mix solution is diluted from the manufacturer's recommended procedure. We are able to dilute the XTerminator/SAM solution since we have drastically diluted the energy transfer dichloro-rhodamine dye. Due to the dilution, the water is added to accommodate the same volume since a specific run BigDye® XTerminator™ module (BDx_RapidSeq36_POP6_1) is used. The plate wells are covered and thoroughly mixed for 30 minutes. Centrifuge the plate at 1000 x *g* for 2 min. Run the plate on the 3130*xl* Genetic Analyzer using the BigDye® XTerminator™ run module. The major feature of this run module is that the plate on the 31xx Genetic Analyzers is not raised as high. The XTerminator™ Solution contains beads that bind to the excess dyes. So the beads are not introduced into the capillary, the tray is not raised to the same level as the other modules.

SAM™ Solution	=	16.4 uL
BigDye® XTerminator™	=	3.7 uL
Sterile water	=	29.9 uL
TOTAL	=	50.0 uL

Chart 3. XTerminator/SAM solution. Use a wide bore pipette tip when pipetting the BigDye® XTerminator™ Solution.

Sequence Data



I:\houston 020906\My Documents\UNTHSC\Research\Sequence Data ROBY\Run_Forensic2_2007-01-16_21-44_1161\06-4223_dRhod_A1_G05_013.ab1

Coordinates (x,y): 1797,652

I:\houston 020906\My Documents\UNTHSC\Research\Sequence Data ROBY\Run_Forensic2_2007-01-16_21-44_1159\06-4223_AB_1.1_A1_H03_015.ab1

Coordinates (x,y): 1492,220

Figure 1. Sequence data visualized with Sequence Scanner v1.0 (Applied Biosystems) produced with a) dRhodamineTerminator Cycle Sequencing Kit and b) ABI PRISM® BigDye® Terminator Cycle Sequencing Kit using the respective procedures described in this paper. Note that the peak signal and patterns are different between the two dye chemistries (Roby *et al.,* 2007).

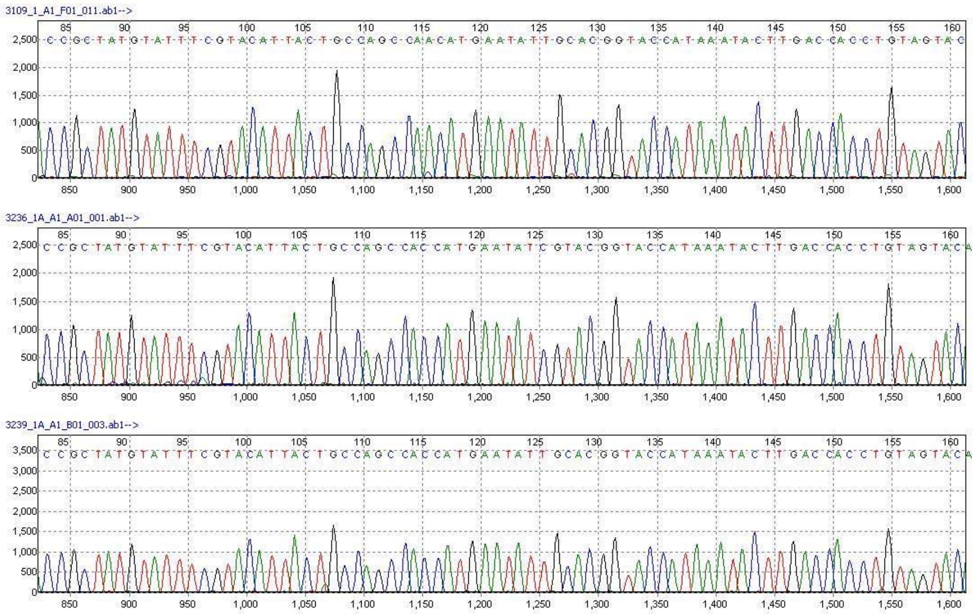


Figure 2. Sequence data visualized with Mutation Surveyor v3.1 (SoftGenetics LLC, State College, PA) from 3 different samples using the new sequencing procedure with ABI PRISM® BigDye® Terminator v1.1 Cycle Sequencing Kit, BetterBuffer, and BigDye XTerminator Solution from HV1 products and the Primer A1. Note the similarity of peak signal and patterns for each of the different samples (Roby *et al.*, 2007).

2.2 High Throughput Robotic Processes for mtDNA

During the analysis of samples for the identification of remains for the World Trade Center incident, Celera Genomics designed a high throughput mtDNA sequencing laboratory like had never been done before. Robotics were put into place for re-arraying extracted DNA from 96-well plates to 384-well plates. These 384-well plates were then used in another robotic process to amplify the extracted DNA. Once the mtDNA was amplified, a large amplicon for each of the reference samples and four smaller amplicons for each of the evidentiary samples and personal effects, these amplified products were then quantitated by a Cytofluor assay. Once quantitated, the samples were cycle sequenced with another robotic assay. In summary, the processing of over 42,000 samples was accomplished in the first high throughput robotic processing for mtDNA sequencing of forensic samples.

The laboratory at the University of North Texas Center for Human Identification has adapted a method using the Slicprep™ 96 Device and DNA IQ™ System from Promega Corporation in conjunction with the Tecan Freedom EVO® 100 robot using fixed tips (Figure 1). The addition of the Maxwell 16 DNA Purification instrument will help in the extraction of DNA from human remains.

The University of North Texas (UNT) System Center for Human Identification is a partner in the National DNA Index System database for missing persons. In order for the missing persons program to be successful, building a family reference database for comparison to recovered remains is critical. Since the DNA from many of the recovered remains is highly degraded, both nuclear DNA and mtDNA results are evaluated. The UNT System Center for Human Identification anticipates its program growing to over 20,000 family reference samples per year. With the incorporation of multi-capillary instruments speeding up the data acquisition, bottlenecks in sample processing exist in DNA extraction and PCR amplification setup. Many robotics systems are used in forensic and vendor laboratories to process convicted offender samples worldwide. However, the use of robotics to extract DNA for both nuclear DNA and mtDNA testing with fixed pipette tips is not commonplace. Our laboratory has adapted a method using the Slicprep™ 96 Device and DNA IQ™ System from Promega Corporation in conjunction with the Tecan Freedom EVO® 100 robot using fixed tips. The Slicprep™ 96 Device is used to pre-process the buccal swab samples submitted in the Family Reference Sample collection kits. Buccal swab heads arranged in batch format are incubated in lysis buffer and then centrifuged using the collar expander of the Slicprep™ 96 Device. The 2.2mL deep-well collection plate from the Slicprep™ 96 Device is conveniently placed on the Tecan Freedom EVO® 100 deck. The Tecan Freedom EVO® 100 is configured with 8 fixed liquid-sensing tips, a magnetic block, and a heat block. The fixed tip format was selected to reduce the cost and storage requirements involved with disposable tips. Using a paramagnetic resin and magnetic block, the DNA IQ™ System eliminates the need for any hands-on manipulation. Scripts provided by Promega Corporation for the DNA IQ™ System using the Tecan were modified to reduce the number of transfers made between plates, to reduce the amount of sample lysate and resin used, and to introduce multiple wash procedures for use with a fixed-tip system. Since the Tecan Freedom EVO® 100 can be programmed to independently control each tip, well positions on the plate can be skipped to allow for addition of controls and/or ladders later in the testing process with no waste of additional reagents. Studies were performed to reduce the amount of DNA IQ™ Resin used in the extraction procedure. Even with this reduced amount of DNA IQ™ Resin, consistent amounts of extracted DNA were obtained. Cross-contamination studies were performed for both STRs and mtDNA using checkerboard and zebra-stripe patterns. Cleaning procedures were developed for the fixed-tip platform, eliminating the impact of any inadvertent well to well crosstalk originally detected through mtDNA sequencing. Following DNA extraction, STR and mtDNA amplification setup is performed by a Tecan MiniPrep 75 Sample Processor. Together, the overall process incorporating Slicprep™ 96 pre-processing, DNA IQ extraction on the Tecan Freedom EVO® 100, and amplification setup on the Tecan MiniPrep 75 Sample Processor has resulted in consistency in sample yield, elimination of cross-contamination effects, and reproducibility for both STR and mtDNA analysis.

Buccal swabs were pre-processed in a Slicprep™ 96 Device (Promega Corp., Fig. 1) and then placed on the fixed-tip Tecan Freedom EVO® 100 for DNA extraction with the DNA IQ™ System (Madison, WI). After extraction, real-time PCR was performed on the ABI 7500® Sequence Detection System (Applied Biosystems, Foster City, CA) using the Quantifiler™ Human DNA Quantification Kit. The Applied Biosystems AmpFℓSTR® Profiler Plus ID and COfiler kits were utilized for STR amplification and the mtDNA regions HV1 and HV2 were also amplified. Following PCR amplification, mtDNA amplicons were quantified with the Agilent

2100 Bioanalyzer (Agilent Technologies, Palo Alto, CA). Unused primers and dNTPs were removed using ExoSAP-IT® (USB Corp., Cleveland, OH) and then the mtDNA products were sequenced using the ABI PRISM® dRhodamine Terminator Cycle Sequencing Kit and purified with an Edge Perfroma Plate (Edge Biosystems, Gaithersberg, MD). Both STRs and mtDNA were capillary electorphoresed on the ABI PRISM® 3100 Genetic Analyzer. STR and mtDNA analyses were then conducted using GeneMapper™ ID (Applied Biosystems) and Sequencher™ (Gene Codes, Ann Arbor, MI) software, respectively.

In this study, 61 buccal swabs and 35 reagent blanks were extracted. The swabs and reagent blanks were arranged in both a checkerboard and a zebra stripe pattern. The quantity of nuclear DNA was determined using Quantifiler™ kit. Nine of the 35 reagent blanks produced results with a concentration in the range of 0.002 – 0.054 ng/μL. All reagent blanks were amplified with the AmpFLSTR® COfiler PCR Amplification Kit. RB11 (0.054 ng/μL) gave a partial STR profile (Fig. 2). Comparisons were made to STR profiles obtained from previous casework and all results were concordant. All reference samples and reagent blanks were amplified at HV1. Interpretable sequence was obtained for all samples including the reagent blanks. Sources of contamination were: 1) adjacent well (6); 2) same tip (11); 3) random well (3); and 4) sequence mixture, multiple source (15). Contamination may have resulted from: 1) carryover on fixed tips; 2) tip touching; and 3) shaking. Washing tips with water between transfers was not enough to eliminate carryover. It was determined that bleach washes should be added to the automated extraction protocol. To determine if adding less resin will result in consistent, adequate concentrations of DNA extract the volume of DNA IQ™ Lysis Buffer was increased as the volume of DNA IQ™ Resin decreased. Hence, the total volume of the Resin/Lysis Buffer mixture remained the same (40μL). As resin volume decreased, mean concentration and standard deviation of DNA extract was decreased (Fig. 4)

Lysis Buffer/Sample Reduction Study 1

- An additional measure to minimize the amount of carryover contamination
- 600ul DNA IQ™ Lysis Buffer was added to each sample to fully cover the ejectable swab heads
- Two different volumes of sample lysate (25μL and 50μL) mixed with DNA IQ™ Resin/

Lysis Buffer mixture were tested

- Two different volumes of DNA IQ™ Resin were tested: 3µL and 5µL
- Bleach washes with 10 second holds were inserted between each transfer step on the Tecan Freedom

EVO® 100

- Standard deviation was lower for samples in which 3µL of DNA IQ™ Resin

was added (Fig. 5)

- Concentrations obtained with 3µL DNA IQ™ Resin were low at times, therefore

3.5µL was tested (this is ½ of volume stated in Promega's DNA IQ™ protocol

- 400µL, 500µL and 600µL of DNA IQ™ Lysis Buffer addition was tested
- Standard deviation was higher for samples incubated with 400µL DNA IQ™ Lysis Buffer
- This discrepancy may be due to variations in the amount of DNA deposited on the buccal swabs

rather than the amount of DNA IQ™ Lysis Buffer added

- 84 reference samples were amplified with AmpFℓSTR® Profiler Plus ID
  - A set volume (1µL, 2µL or 3µL) of DNA sample was added to the amplification for each

extraction group

  - Concentration of each sample was not normalized
- 10 of these samples required reloading on the ABI PRISM® 3100 Genetic Analyzer
  - One overblown (diluted and reloaded)
  - 8 with alleles below threshold (2µL or 3µL loaded)
  - One issue with ROX Internal Lane Standard (reloaded 1µL)
- Improvement seen in all reloaded samples - only 5 samples with partial STR profiles
- Success rate – 94%

- For future extractions, all samples were incubated in 400µL DNA IQ™ Lysis Buffer and

3.5µL of DNA IQ™ Resin was added to 20µL sample lysate

- Mean concentration of 1.002 ng/µL is what is expected from DNA IQ™

- Peak height ratio is not significantly lower than for other extraction groups
- 400μL DNA IQ™ Lysis Buffer volume is specified by Promega's DNA IQ™ protocol
- 47 buccal swabs and 44 reagent blanks were extracted in a checkerboard pattern
- No real-time PCR was performed
- All reference samples and reagent blanks were amplified with the AmpFℓSTR® COfiler Kit
    - No alleles present for reagent blanks
    - 91% of reference samples yielded complete STR profiles
- The mtDNA region HV1 was amplified and sequenced for all reference samples and reagent blanks
    - 9 reagent blanks (20%) had interpretable sequence (Fig. 9)
    - All reference samples had interpretable sequence with no signs of mixture

Extraction of buccal swabs on the Tecan Freedom EVO® 100 using the DNA IQ™ System yields good quality STR profiles with a success rate greater than 90% and clean interpretable mtDNA sequence. STR failures are most likely due to swab quality. A minimum amount of carryover is occurring, however the level of DNA is below the level of detection when in competition with the actual DNA sample. The source of the contamination is not always from the adjacent well on the left (same tip). Following Contamination Study 2, the order of steps on the Tecan Freedom EVO® 100 were altered slightly to further reduce the risk for contamination. Initially the mixture of DNA IQ™ Resin/Lysis Buffer was added to the 2.2mL plate from Slicprep™ 96 Device which contained the sample lysate. This new mixture of resin with sample lysate then had to be transferred again to a 1.2mL plate which fits onto the MagnaBot®. The script was altered so that the DNA IQ™ Resin/Lysis Buffer mixture is transferred directly into a clean 1.2mL plate and then the sample lysate is added to this mixture. This eliminates two key transfer steps which are the most likely place where carryover contamination was occurring. This new deck setup is shown in Figure 8b. Other recommendations are to use a UV crosslinker for all consumables before extraction and mtDNA amplification.

A plate layout has been designed so that samples and a reagent blank will always be placed into the same wells. Nine wells are left empty so that they can be utilized for negative and positive amplification controls and for allelic ladders later in the process. This enables the plate to be placed directly onto the MiniPrep 75 Sample Processor for PCR amplification setup and later onto the ABI 3100® Genetic Analyzer without any plate rearrangement. However, all sample amplifications performed for these studies were set up manually. Testing is currently ongoing for amplification setup with the AmpFℓSTR® Identifiler Kit on the MiniPrep 75 Sample Processor. With a sample pre-processing time of approximately 2-3 hours and an automated extraction script that takes about 4 hours, the total time required to extract a plate of swabs is 6-7 hours. A total of 86 samples can be processed per plate for a total of 17,200 samples per year assuming 200 working days, one technologist and one plate extracted per day. With the amount of hands on time greatly reduced, analysts can devote time to other responsibilities such as analysis.

Additionally, automation scripts have been designed for the amplification and sequencing of these multiple reactions required for each sample

2.3 Decrease in Cycle Number for Amplification

Mitochondrial DNA testing is a laborious process with amplifying minimally two regions and sequencing using multiple primers in both the forward and reverse directions. Automation scripts have been designed for the amplification and sequencing of these multiple reactions required for each sample as well as decrease reaction volume. By using a smaller reaction volume, fewer reagents are used, thereby saving time and money.

RESEARCH DESIGN

Mitochondrial DNA Protocol

Currently at the University of North Texas Center for Human Identification (UNTCHI), the mtDNA protocol is not amenable to high throughput analysis. The first step in mtDNA analysis, following DNA extraction and quantification is amplification of HV1 and HV2. The master mix recipe is in Table 4.

Master Mix for mtDNA	
	Per reaction
Sterile H_2O	5.5µL
10X PCR Buffer II	2.5µL
25mM $MgCl_2$	2.0µL
dNTP mix (10mM)	1µL
BSA (1.6µg/µL)	2.5µL
Primer 1 (10µM)	0.5µL
Primer 2 (10µM)	0.5µL
AmpliTaq Gold (5U/µL)	0.5µL

Table 4. mtDNA Amplification Master Mix for HV1 and HV2

The master mix is made for both HV1 and HV2, with Primer 1 being A1 and C1 for HV1 and HV2, respectively and Primer 2 being B1 and D1 for HV1 and HV2, respectively. A total volume of 15µL of master mix is added to a tube, or a well in a 96-well plate, depending on the total number of amplifications to be performed. A total volume 10µL of each DNA sample is then added to the appropriate tubes, or wells. DNA dilutions are performed when necessary, typical dilutions are to 1.0ng/μL of quantified nDNA. Additionally, 0.34ng/μL nDNA HL60 diluted in molecular biology grade water is added to each positive control tube or well, with 15μL of master mix and 10μL of sterile water is added to each negative control tube, or well. The amplification parameters are as follows in Table 5.

mtDNA Amplification		
HOLD	95°C	11 minutes
32 Cycles		
95°C	10 seconds	
61°C	30 seconds	
72°C	30 seconds	
HOLD	70°C	10 minutes
HOLD	4°C	∞

Table 5. mtDNA Amplification Thermal Cycler Parameters

After amplification, the products must then be visualized using an agarose yield gel. This allows an analyst to determine if sufficient amplified product was obtained or not. For the positive control, approximately 5μL of amplified product shall be cycle-sequenced. Depending on the intensity of the band produced by a sample, when compared to the band produced by the positive control, up to 9μL of product is cycle-sequenced. Next, amplicon purification is performed by removing unused primers and nucleotides, using ExoSAP-IT®. A total volume 5μL of Exo-SAP-IT® is added to the amplified product and then placed on the thermal cycler, using the parameters shown in Table 6.

ExoSAP-IT®	
37°C	15 minutes
80°C	15 minutes
4°C	∞

Table 6. Exo-SAP-IT® Thermal Cycler Parameters

After Exo-SAP-IT® the samples are ready for cycle sequencing using the ABI PRISM™ dRhodamine Terminator Cycle Sequencing Reaction Kit. First, 8µL of *Sequencing Ready Reaction Mix* is added to each amplification tube to be used. Then, 3.2µL of 10µM primer is added to the tubes and the reaction volumes are adjusted to 20µL with sterile water and amplified DNA product, depending on the gel visualization. The tubes are then capped and placed on the thermal cycler using the parameters seen in Table 7.

mtDNA Cycle Sequencing		
HOLD	96°C	1 Minute
25 Cycles		
	96°C	15 seconds
	50°C	1 second
	60°C	1 minute
HOLD	15°C	10 minutes
HOLD	4°C	∞

Table 7. mtDNA Cycle Sequencing Parameters

Following cycle sequencing, Edge Gel Filtration columns are used to purify the amplified product by removing dye terminators, dNTPs, salts and other low molecular weight materials from sequencing reactions, as well as DNA primers and fragments, among others. This is done using Performa® DTR spin columns or 96-well tray. The gel column is placed on a microcentrifuge tube, or a 96-well plate. The entire 25µL reaction is removed from the tubes and loaded into the top of the gel, this is centrifuged and the column is discarded, as the purified fragments remain in the tube. The samples are then ready to be placed on the 3130*xl* Genetic Analyzer for capillary electrophoresis.

Electropherogram Analysis

Automated Mitochondrial DNA Analysis for High Throughput Testing

Script Design

In order to begin evaluating the amenability of high throughput mtDNA analysis, the previously described, current analysis procedure was evaluated. First robotic scripts were designed for the Tecan MiniPrep 75 ® Sample Processor, which allowed for automated amplification setup as opposed to manual setup. A picture of the Tecan MiniPrep 75® Sample Processor used in this project is pictured below in Figure 5.

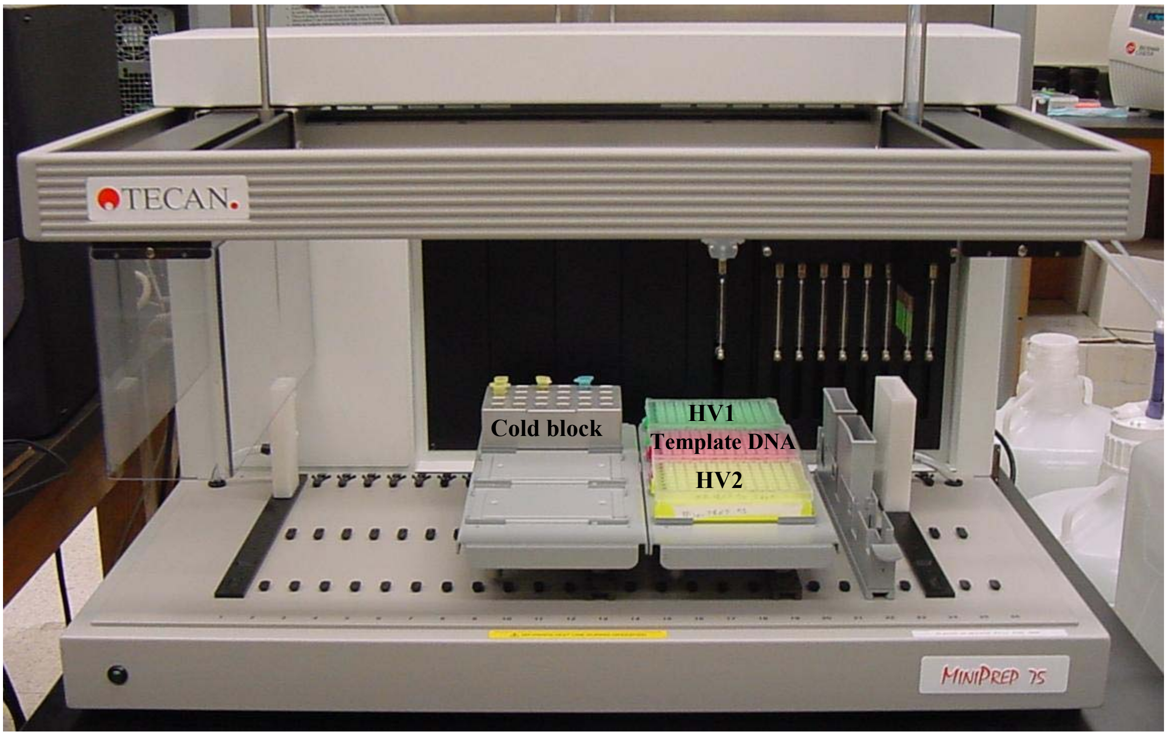


Figure 5. Tecan MiniPrep 75® Sample Processor.

The Tecan MiniPrep 75® Sample Processor contains two robotic arms. One of these arms (on the left) has one fixed-tip pipette, and the other (on the right) has eight fixed-pipette tips. With the script designed, the single tip pipette first goes into the HV1 master mix tube and distributes 15μL of master mix into all sample or control wells in the HV1 plate. The tip then undergoes a thorough wash with bleach, a bleach hold, and then a thorough rinsing with water, and then enters the HV2 master mix and puts it in the sample and control wells in the HV2 plate. Note, because there numerous washes and flushes to avoid contamination and because of the air space required for robotics, an increase in the amount of master needed for setup is 25%. After another thorough wash, the single-tip pipette adds the positive control and negative control into the appropriate wells in the HV1 and HV2 plates. After the two master mix plates are made, the right arm goes into the first column of the template plate and intakes a specified amount DNA template and then distributes it into Column 1 of the HV1 plate. The tips then undergo a bleach wash and again re-enter Column 1 of the template plate and dispensed into Column 1 of HV2 plate. This continues into successive columns until the plates are complete. The scripts have been designed to ask for the number of columns in the plates of template DNA. When amplification setup is complete, the plate is ready for amplification on the thermal cycler.

Input DNA & Cycle Number Evaluation

In order to save analyst time and eliminate potentially unnecessary steps in the current UNTCHI protocol, the overall process was thoroughly examined. The proposed steps to eliminate were: dilutions, post-amplification visualization using an agarose yield gel; reduction in reaction volume and the amount of template DNA input into the amplification reaction; and reduction in cycles for amplification. In the modified BigDye® Terminator v.1.1 Cycle Sequencing with Dilutor Buffer, for reference samples protocol, it states that amplified DNA be diluted to 0.25ng/µL [16]. Therefore, to eliminate the agarose yield gel visualization and dilution steps, this is the amount targeted in the decreased cycle number and amount of input DNA evaluation. The initial step was to decrease the amount of input template in amplification from 10µL to 1µL. In order to evaluate the cycle number, this 16µL total reaction volume was then amplified for 22, 24, 26, 28, 30 and 32 cycles, using the same thermal cycler parameters seen in Table 5. Because the target concentration was 0.25ng/µL, the cycle number which produced this target amount was the one chosen and further optimized. In order to decrease waste of reagents and sample, only the first column was amplified and cycle sequenced. This contained the positive and negative controls, the reagent blank and samples 1-5. Following the amplification of these, they were quantified using the Agilent 2100 Bioanalyzer chip. This was done in order to accurately target which cycle number produced the targeted amount of amplified mtDNA. The results are shown in Table 8.

Moreover, in the first amplification, samples 2 and 3 had a homopolymeric-C (PolyC) stretch, and samples 4 and 5 were not properly quantified; therefore only sample one is shown in Table 8.

Cycle Number	Quantity (ng/µL)
32	5.40
30	4.14
28	3.42
26	2.29
24	1.59
22	0.14

Table 8. Effects of Decreasing Cycle Number

These results showed that the cycle number which produced results closest to 0.25ng/µL was achieved in 22 cycles of amplification. I again reevaluated this finding by rerunning the samples using 21, 22, 23, 24 and 25 cycles. After this, 22 cycles was determined to be the chosen number of amplification cycles. It was also noticed that the positive control was not properly amplifying. We determined this to be because of the decreased reaction volume. Therefore, we needed to increase the concentration of the positive HL60. After determining that the concentration of the original HL60 was 0.17 ng/µL, a new positive control was made that was approximately three times more concentrated. This was done using 1.5µL of stock HL60 [340ng/µL] added to 998.5µL of distilled water to make a new concentration of 0.51ng/µL. Because of this change the robot script was redesigned to add only 1µL of the positive and negative controls into the 15µL of master mix. The samples and positive control were then ready to be cycle sequenced.

Cycle Sequencing

Because the total reaction volume is only 16µL with this proposal, the amount of Exo-SAP-IT® was decreased from 5µL to 3µL. This 19µL was now ready to be cycle-sequenced. To determine whether the modifications made to the amplification parameters would still yield a sequence equivalent or cleaner, samples 1-5 were cycle sequenced using only the A1 primer. This was done to prevent unnecessary waste of amplified product and reagents. Using the modified BigDye® Terminator v.1.1 protocol seen in Table 9, the samples were cycle sequenced.

BigDye® v.1.1 Cycle Sequencing Master Mix	
BigDye® Terminator v.1.1	1µL
BetterBuffer	5µL
Primer	1.5µL
Sterile Water	6.5µL*

Table 9. BigDye® v.1.1 Cycle Sequencing Master Mix (*variable amounts)

14µL of master mix was added to a new 96-well plate and 1µL of purified amplified product was added to this. Cycle Sequencing was performed under the parameters seen in Table 10 below.

BigDye® v.1.1 Cycle Sequencing		
HOLD	96°C	3 minutes
25 cycles		
	96°C	15 seconds
	50°C	10 seconds
	60°C	3 minutes
HOLD	4°C	∞

Table 10. BigDye® v.1.1 Cycle Sequencing Parameters

The samples then undergo BigDye® XTerminator™ Purification according to the master mix recipe in Table 9. Because the BigDye® XTerminator™ is a solution containing numerous beads, the 3130*xl* is programmed to only lift the plate to a certain height so that the capillaries avoid aspirating the beads. Therefore, since the reaction volume has been decreased, 5µL of water was added to the sample wells to ensure the beads not be taken up. So instead of 50µL, which was typically added to the previously cycle-sequenced product, 55µL shall be added.

BigDye® Xterminator™ Purification	
SAM™ Solution	16.4µL
BigDye® Xterminator™	3.7µL
Sterile Water	35µL*

Table 11. BigDye® XTerminator™ Purification Master Mix (*originally 30µL but increased to 35µL to maintain a high reaction volume)

The sample plate is then centrifuged and ready to be placed on the 3130*xl* for sequencing via capillary electrophoresis. The results obtained were fantastic. The baseline was very low and the sample sequences were very easily interpreted. The same procedure was done for samples 1-5 using the B1, C1 and D1 primers. The results here varied. Some of the sequences obtained for HV2 produced a very low signal quality. Therefore it was proposed to increase the amount of template added to the 14µL of master mix. Thus the amount of sterile water was decreased to 5.5µL and 13.0 µL master mix was added to 2µL of purified amplified product, so that the final volume was still 15µL. Samples 1-5 and the positive control were cycle sequenced using these parameters. The results obtained from this showed that adding 2µL of template resulted in consistently high baseline and made the interpretation of mtDNA sequence somewhat problematic. Because results obtained were consistent, it was determined to run the plate of all 44 samples using the determined parameters; 22 cycles of amplification and 1µL and 1.5µL of input template in cycle sequencing for HV1 and HV2 respectively.

Therefore, amplification of HV1 and HV2 was set up by the Tecan MiniPrep 75® Sample Processor. The plates were then amplified and Exo-SAP-IT was performed. The parameters used are in Tables 5 and 6 respectively. After Exo-SAP-IT, four master mixes were made, one for A1, B1, C1 and D1. Master mix was added to the wells according to the previously mentioned determinations. The samples were cycle sequenced and then placed on the 3130*xl* for capillary electrophoresis.

RESULTS

mtDNA using Automated mtDNA Analysis for High Throughout Testing

Using the designed method, of 15µL amplification master mix plus 1µL of template DNA, undergoing 22 cycles of amplification, followed by immediate Exo-SAP-IT®, using 3µL, then BigDye® Terminator™ Cycle Sequencing using 1µL and 1.5µL of input template DNA for HV1 and HV2 primers, respectively, followed by XTerminator® purification and then sequence analysis on the 3130*xl*. Forty-six samples were analyzed with an approximate 65% pass rate. The amount of input template was 1µL for the cycle sequencing reactions, since HV1 previously used 1µL of input. Approximately 29% of the samples produced some data interpretable, and contained useful information. Approximately 6% of the samples failed.

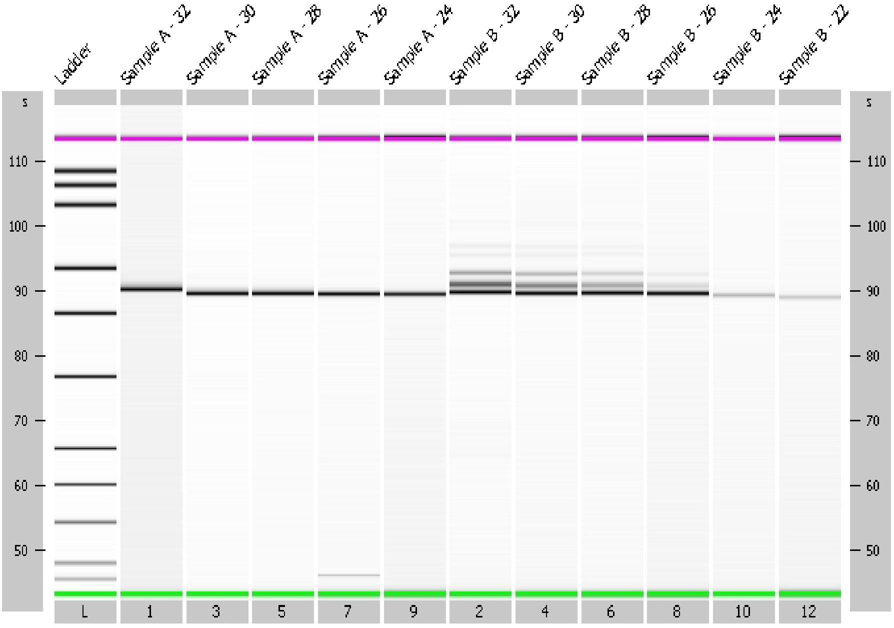
DISCUSSION AND CONCLUSIONS

mtDNA Protocol Design and Optimization

The amplification set-up script designed on the Tecan MiniPrep 75® Sample Processor is the first step for mtDNA automation. Scripts to robotically perform Exo-SAP-IT® and BigDye® Terminator™ Cycle Sequencing are also in the process of being designed to further automation. Also, undergoing only 22 cycles of amplification, as opposed to 32 or more, has proven to be very sufficient. In addition, using only 1µL of DNA extract, instead of 10µL, allows for more DNA extract to remain, which permits additional future testing if necessary.

Cycle No.	Quantity (ng/µL)
32	5.40
30	4.14
28	3.42
26	2.29
24	1.59
22	0.14

Results from the Agilent 2100 Bioanalyzer Chip for Sample A with varying cycle number.



Agilent 2100 Bioanalyzer Chip results showing the effect of decreasing cycle number on Samples A and B. In Sample B, it is obvious with the multiple bands that a homopolymeric stretch is present.

Overview of
Quality for Sample A

32 cycles

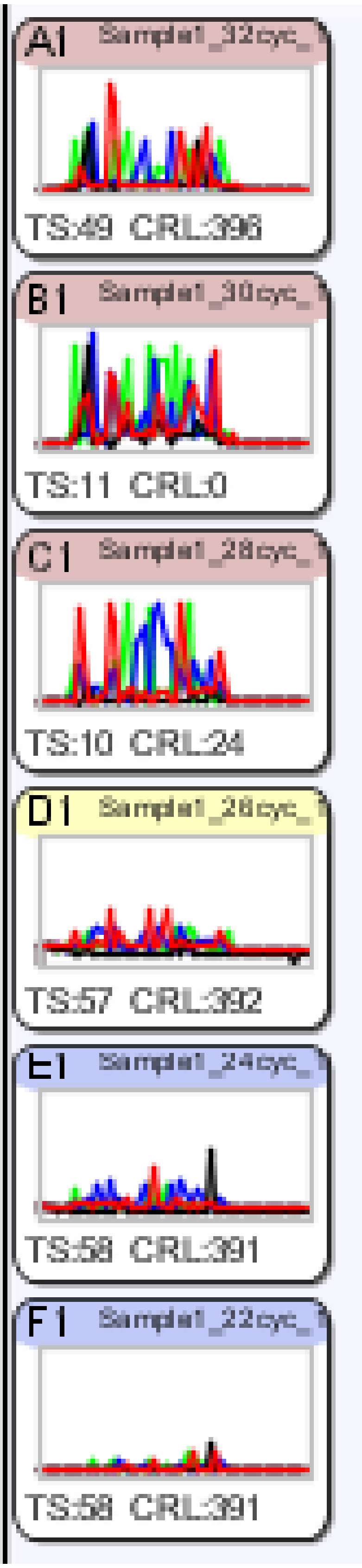
30 cycles

28 cycles

26 cycles

24 cycles

22 cycles



A1 Sample1_32cyc_

TS:49 CRL:396

B1 Sample1_30cyc_

TS:11 CRL:0

C1 Sample1_28cyc_

TS:10 CRL:24

D1 Sample1_26cyc_

TS:57 CRL:392

E1 Sample1_24cyc_

TS:58 CRL:391

F1 Sample1_22cyc_

TS:58 CRL:391

Figure 3. Sequence Scanner v1.0 software (Applied Biosystems, Foster City, CA) showing the quality of sequence obtained from each reaction of Sample A with varying cycle number. The colored bars above the thumbnail images represent quality: pink = low quality; yellow = medium quality

2.4 Quantification of mtDNA

Two real-time quantitative procedures were evaluated. The first procedure was published by Andréasson et al. (Andreasson et al. 2002) Andréasson et al. have designed a duplex assay for co-amplification of the human retinoblastoma susceptibility gene (approximately 108 bp), RB1, and of mitochondrial DNA genes for tRNA lysine and ATP synthase 8 (positions 8294 to 8436, approximately 142 bp). The second procedure was published by Timken et al. (Timken et al. 2005). Timken et al. have designed a duplex assay for co-amplification of the TH01 STR locus (approximately 170 to 190 bp) and of mitochondrial DNA ND1 region (approximately 69 bp). Both assays are being evaluated on the ABI PRISM 7000 Sequence Detection System instrument, Data Collection v.1.1 (Applied Biosystems, Foster City, California, USA). Both assays use TaqMan probes with TAMRA as the non-fluorescent quencher dye.

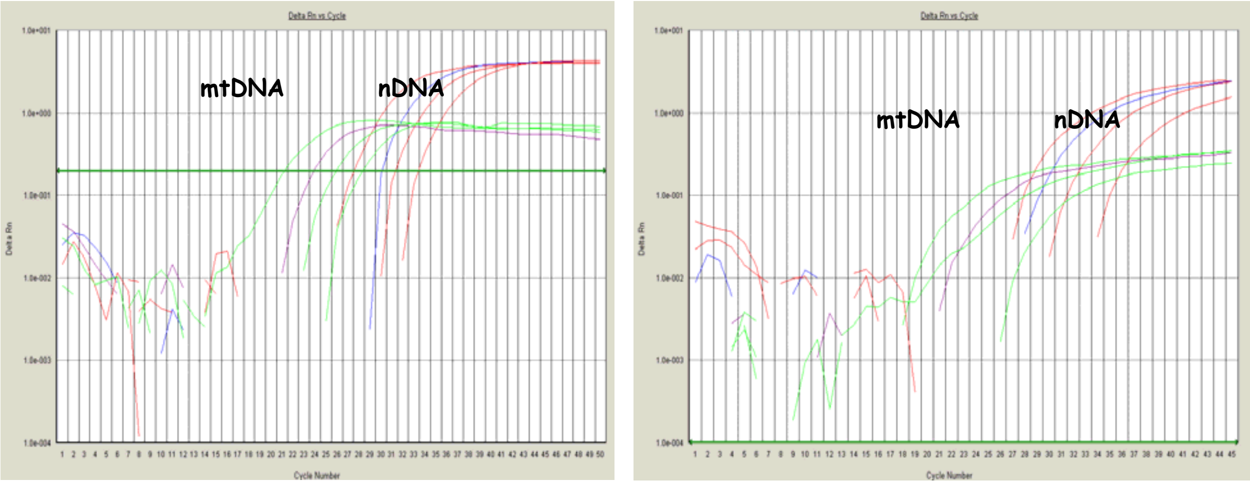


Figure 2. Results obtained in using both assays. The real-time results for (A) are from the Andr'easson procdedure and the results for (B) are from the Timken procedure. A curve using four standards is shown for each the mtDNA and the nDNA results for both assays. Each assay was easy to implement into our laboratory. Decreasing the amount of DNA used in each amplification resulted in an expected higher Ct value for both nDNA and mtDNA.

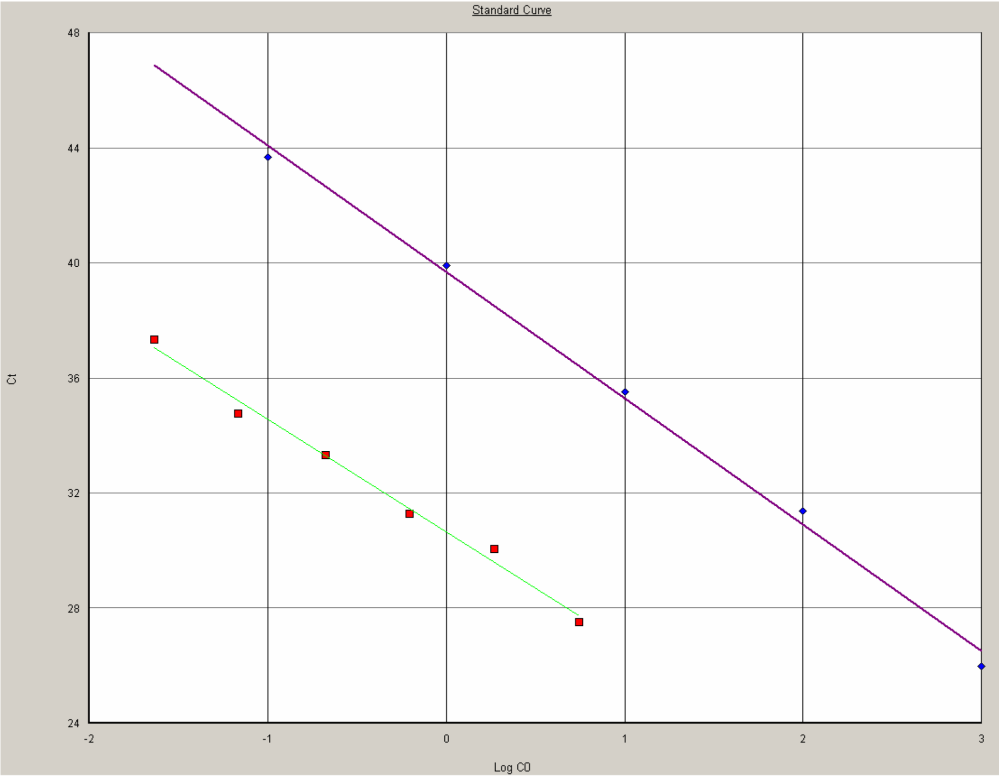


Figure 3. The standard curves for the duplex assay with the nDNA standard and the mtDNA standard for the Andréasson procedure.

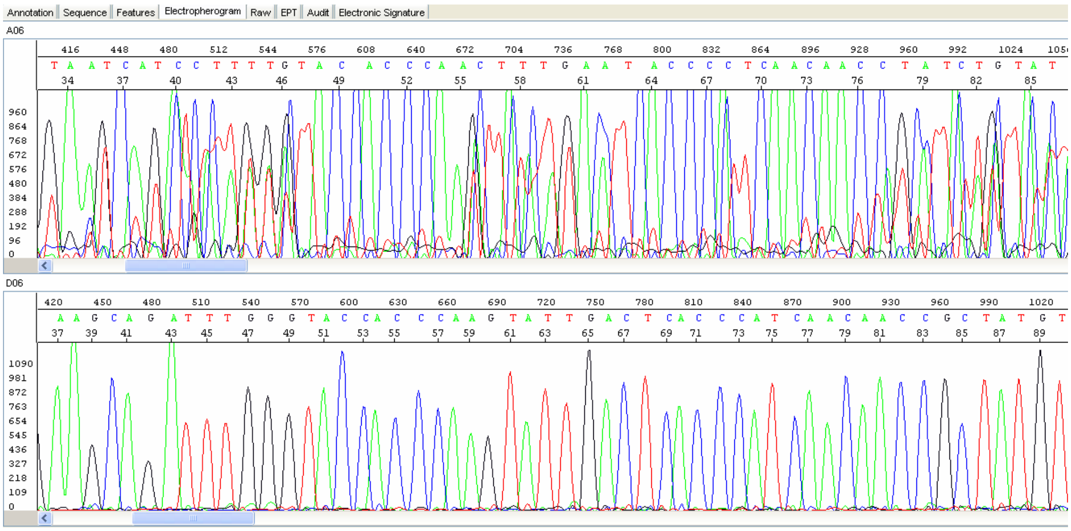

Annotation
Sequence
Features
Electropherogram
Raw
EPT
Audit
Electronic Signature
A06
416 448 480 512 544 576 608 640 672 704 736 768 800 832 864 896 928 960 992 1024 105
T A A T C A T C C T T T T G T A C A C C C A A C T T T G A A T A C C C C T C A A C A A C C T A T C T G T A T
34 37 40 43 46 49 52 55 58 61 64 67 70 73 76 79 82 85
960
864
768
672
576
480
384
288
192
96
0
D06
420 450 480 510 540 570 600 630 660 690 720 750 780 810 840 870 900 930 960 990 1020
A A G C A G A T T T G G G T A C C A C C C A A G T A T T G A C T C A C C C A T C A A C A A C C G C T A T G T
37 39 41 43 45 47 49 51 53 55 57 59 61 63 65 67 69 71 73 75 77 79 81 83 85 87 89
1090
981
872
763
654
545
436
327
218
109
0

Figure 4. Mitochondrial DNA sequence results. These results demonstrate that accurate quantitation of mtDNA for the input in the sequence assay is critical for quality data. In (A), too much DNA was used in the cycle sequencing reaction resulting in noisy data. In (B), an optimal amount of DNA was added to the sequencing assay which resulted in very clean data (Roby et al., 2007).

Newer technologies are more and more sensitive to quantity and quality of DNA to be used in the assay. Today we are routinely using nanogram quantities of DNA in many assays whereas 20 years ago we were striving for microgram levels of DNA. And, in some assays today, we are able to get results with picogram levels of DNA. These new assays are often sensitive to the amount of DNA added. Therefore, quantification of DNA is critical. We are evaluating two assays that will provide us with the quantity of both nDNA and mtDNA. With accurate quantitation, the number of times that we need to repeat our assays is reduced; we can produce quality data on on our first attempt. Accurately estimating the quantity of DNA can conserve valuable DNA sample and can save both time and money. Both TaqMan assays are currently under evaluation to be used in routine forensic casework analysis.

2.5 Quantification of mtDNA PCR Product – Higher Throughput and Cost Savings

Mitochondrial DNA PCR products are often quantified on the Agilent 2100 Bioanalyzer. Each chip holds enough positions for 12 samples. To determine the feasibility of using the chip for duplex assays and thus, duplicated the results obtained from one chip, HV1 and HV2 products were used. This idea not only saves time, but also saves a few minutes. Both HV1 and HV2 amplified products of several samples were added to single wellsof an Agilent DNA 500 labchip. Samples included casework samples, positive control samples, and negative control samples. The volumen added was 1.0uL of each amplicon for a total of 7uL in each well of the chip. The chip was prepared according to the manufacturer's recommendations.

The bands for both HV1 and HV2 were distinctly visible for the combined samples. Concentrations were compared to previous results when run on the chip in separate wells. The concentration differences were small, and would not greatly affect the downstream processing of the samples.

By combining HV1 and HV2 amplicons from a single sample into a single well on the DNA 500 labchip, the Agilent 2100 Bioanalyzer can be used to quantify more amplicons in a shorter amount of time using fewer reagents and supplies. This can be accomplished with no loss of data and can improve throughput in the laboratory.

2.6 Multiplex Screening Tool with mtDNA

Testing of skeletal remains require the expertise of a multi-faceted approach to the identification of unknown human remains for missing person and cold case investigations by integrating both forensic anthropology and odontology with state-of-the-art DNA testing methods. Likewise testing of hair specimens require the expertise of trace evidence hair examiners and DNA examiners together.

Both mitochondrial DNA analysis and nuclear DNA analysis are evaluated since the DNA from many of the recovered remains is highly degraded and/or since family reference samples are often limited. Additionally, a screening tool was requested to quickly determine if more than one subjects' bone samples were found at the bottom of a drain. Similarly, this same technique could be applied for a mass disaster situation, mass graves, or to the screening of multiple hair specimens. In trying to identify victims' remains or stains, many times they are severely fragmented, scattered, and commingled or the number of forensic stains or evidence (e.g., hairs) is quite enormous. DNA testing can help separate commingled specimens or identify the number of contributors to evidence, but it is quite time-consuming and expensive.

To make the most efficient use of time and resources, a DNA-based screening tool could be used early in the investigation in order to separate the fragmented/commingled remains and or forensic samples and help determine the minimum number of victims or donors present. A multiplex screening tool has been proposed that could potentially be used in a massive situation. This screening tool provides Amelogenin, or sexing identification, a nuclear DNA result, and an mtDNA dinucleotide result without the need for sequencing. All of these fragments are sized-based and can be evaluated rapidly, even through an expert system software program. Specifically chosen for the multiplex design are the Amelogenin sex-determining locus, D3S1358, and a 3' CA dinucleotide repeat in the mitochondrial D-loop. A multiplex of these three loci would provide investigators with a quick, convenient, and relatively easy way to initially assess casualty numbers and separate remains for further DNA testing and positive identification.

Statement of Problem:

When skeletal remains are discovered, the first step in the investigation is to determine if the bones are human or animal in origin. Once human origin has been established and when the remains unearthed are skeletal, semi-skeletal, or too badly decomposed for identification by traditional approaches (i.e., fingerprints, dentition, or recognition of facial features), an experienced forensic anthropologist may be called upon to assist in separation and identification of the remains. In some scenarios, the skeletal remains may be fragmented, incomplete, and/or commingled, which complicates identification efforts and makes it difficult to assess the number of victims present. Although forensic geneticists can assist in the identification of victims via nuclear and mitochondrial DNA analysis, obtaining complete genetic profiles from every single skeletal fragment at a mass disaster site or in a mass grave can consume a laboratory's resources in valuable time and reagents. An efficient method to separate commingled and fragmented remains would permit the most cost-effective allocation of resources and speed the identification process.

Research Significance:

Mass death can be precipitated by a variety of scenarios, including explosions, aviation accidents, war, acts of terrorism, fires, natural disasters, homicides, and violations of human rights. Since the remains in these cases are often scattered, commingled, and altered beyond recognition, ascertaining the number of victims and determining the victims' identities is not always a straightforward process. Ultimately, in such situations, the remains require separation to determine how many bodies are present. Since soft tissue is frequently absent or severely degraded in victims from these types of scenarios, the assistance of a forensic anthropologist is often required in the identification process. The remaining bone and teeth are relatively resilient materials from which non-degraded DNA can be obtained. The DNA-based screening tool proposed here is uniquely applicable to the myriad of situations in which fragmented, commingled human remains need to be separated and identified. This multiplex design could prove to be an invaluable investigative tool for the association of separated remains and in the initial assessment of the number of victims present at the scene. By multiplexing several markers into one amplification, time and precious template DNA can be saved (Edwards and Gibbs 1994).

Three loci were selected for incorporation into this multiplex PCR reaction, each for separate reasons. The first and most obvious locus to be included in the proposed multiplex screening tool is Amelogenin. The genes for Amelogenin can be used in sex determination of samples from unknown human origin and would permit separation of commingled/fragmented remains based on gender. Amelogenin is currently the most popular method for sex-typing, and both mass grave and mass disaster investigations greatly benefit from gender identification of the remains. A single primer pair is used for amplification of part of the X-Y homologous genes because it generates different length products from the X and Y chromosomes. These primers flank a 6-bp deletion within intron 1 of the X homologue, resulting in a 106 bp amplification product (amplicon) for the X chromosome and a 112 bp amplicon for the Y chromosome (Sullivan et al. 1993).

The second locus chosen for this project is D3S1358 due to its small amplicon size (97-151 bp) and its variation within the population (Collins et al. 2004). The D3S1358 marker is a compound tetranucleotide repeat found on the short arm of chromosome 3, with a specific repeat structure of TCTA[TCTG]$_{2\text{-}3}$[TCTA]$_n$. It is known that the more degraded the sample, the more difficult it is to amplify larg DNA fragments; hence, D3S1358 was a small amplicon chosen for this multiplex (Whitaker et al. 1995).

The third marker to be incorporated into the multiplex design is a 3' CA dinucleotide repeat in the D-loop of the human mitochondrial genome. This dinucleotide repeat was significant it would give mtDNA results in a quick fashion and possibly lead to the screening of the type of DNA testing to be performed on the specimens to be tested. Oftentimes, mtDNA produces results when nDNA results are not achieved due to its high copy number.

Although analysis of the highly polymorphic control region has become a powerful tool for forensic identity testing, the often severely degraded condition of remains from crime scenes leads to poor PCR amplification of the larger-sized complete mtDNA control region. Therefore, selective targeting and preferential amplification of highly informative variable sites (such as HVIII) is likely to be a more effective and alternative method for forensic mtDNA analysis (Lee *et al.* 2006). Specifically located at positions 514-523 in Hypervariable Region III (HVIII) *(Appendix, Figure 2),* the CA dinucleotide repeats selected for this project display length polymorphism like STRs in nuclear DNA and have been shown to demonstrate relatively high genetic diversity. A total of six different alleles [$(CA)_3$ to $(CA)_8$] have been detected at this locus, with the allele name designating the number of dinucleotide repeats it contains.

Although more population data needs to be collected regarding the CA repeat alleles, several studies have already demonstrated their variation within and among populations. Szibor *et al.* (1997) examined samples from three European populations (Germans, Hungarians, and Russians) and one African Bantu population (Cameroon). Upon analysis, significant differences were demonstrated not only as to the incidence of particular alleles, but also in terms of allele distributions in different populations (*Appendix, Table 3*). Allele 5 was the most frequent allele found in all three Caucasian populations, while allele 7 was rare and allele 3 was not seen. A comparison of the proportion of allele 4/allele 6 was useful when comparing European populations, as differences in the frequencies of these alleles were significant between Hungarians and Germans and between Hungarians and Russians. Furthermore, in the African population, allele 4 was the most common, alleles 6 and 7 were not detected, and allele 3 was found twice in the sample of 105 individuals. Research on the allelic distribution of the CA repeat has also been conducted on a group of unrelated Japanese individuals living in Gifu Prefecture (a central region of Japan) (Nagai *et al.* 2004), a population in Korea (Chung *et al.* 2005), and on two separate populations in Bologna, Italy (Bini *et al.* 2003). The CA repeat allele distributions for these four studies are summarized in *Appendix Table 4.* A study conducted by Tang *et al.* (2003) on two Chinese ethnic groups further provides population data for the CA repeat in the mitochondrial D-loop. *Appendix Table 5* is a compilation of the reported frequencies of the eight CA repeat alleles in the various population groups studied.

The first step in the design of this multiplex screening tool is to amplify each locus individually using the appropriate primer pairs. Initial thermocycling parameters for the singleplex reactions will correspond to previously published protocols. In particular, the recommended PCR cycling conditions for amelogenin and D3S1358 with Applied Biosystems' AmpFℓSTR kits are as follows: enzyme activation at 95°C (11 min.) followed by 28 cycles of denaturation at 94°C for 1 min., annealing at 59°C for 1 min., and extension at 72°C for 1 min.; and a final extension at 60°C for 45 min. (∝ 25°C) (Wallin *et al.* 1998, 2002; Butler 2005). Promega's cycling parameters for the same two loci are 95°C (11 min.); [94°C for 30 sec. (cycles 1-10); 90°C for 30 sec. (cycles 11-30); 60°C (30 sec.); 70°C (45 sec.)] x 30; 60°C (30 min.) and ∝ 4°C (Butler 2005). These optimized and validated multiplex PCR parameters for both Applied Biosystems' AmpFℓSTR kits and Promega's GenePrint® STR kits are summarized in *Appendix Table 6* and will be used as a starting point for the amelogenin and D3S1358 singleplex reactions. The cycling parameters differ because reaction components (particularly the primer concentrations and sequences) vary between the different manufacturers' kits (Butler 2005).

Furthermore, during initial testing, Applied Biosystems' recommendations for final concentrations of PCR reaction mix components will be adhered to, with alterations made if necessary as the development of this multiplex screening tool proceeds. Applied Biosystems recommends using 1-2.5 ng of genomic DNA in 50-μl reaction volumes, with PCR reaction mix final concentrations as follows: 10 mM Tris-HCl (pH 8.3), 50 mM potassium chloride (KCl), 1.25 mM $MgCl_2$, 800 μM blended dNTPs, 0.16 μg/μl bovine serum albumin (BSA), and 4.5 U of AmpliTaq Gold DNA Polymerase (Wallin *et al.* 2002). Several different protocols have been used successfully in terms of the thermocycling parameters for amplification of the

CA repeat in Hypervariable Region III of the mitochondrial D-loop (Szibor *et al.* 1997; Bini *et al.* 2003; Hoong & Lek 2005). The range of experimental PCR cycling conditions for each of these researchers is summarized in *Appendix Table 7*, along with the reaction mix components and concentrations.

After successful singleplex amplifications of each of the three loci, duplex reactions will be attempted and, ultimately, experimentation will proceed in an effort to amplify all three loci simultaneously in a single reaction. This will require adjustments to various PCR parameters (such as relative primer concentrations, buffer concentration, thermocycling temperatures, amount of template DNA, balance between the magnesium chloride and deoxynucleotide (dNTP) concentrations, and amount of Taq DNA polymerase) in order to optimize the multiplex reaction. After electrophoresis of samples on ethidium-bromide stained agarose gels, PCR products will be visualized with a UV transilluminator and photographed with Doc-It®LS Image Acquisition Software (Life Science Software from UVP).

Custom unlabeled oligonucleotide primers for each of the three chosen loci were purchased through Invitrogen™ Life Technologies (Carlsbad, CA). The amelogenin and D3S1358 primer pairs for this experiment correspond to the sequences used in Promega's PowerPlex® 16 kit (Krenke *et al.* 2002). Specifically, the amelogenin primer pair sequences are: (forward) 5'-CCCTGGGCTCTGTAAAGAA-3' and (reverse) 5'-ATCAGAGCTTAAACTGGGAAGCTG-3'; and the D3S1358 primer pair sequences are: (forward) 5'-ACTGCAGTCCAATCTGGGT-3' and (reverse) 5'-ATGAAATCAACAGAGGCTTGC-3'. The primers selected for the mitochondrial D-loop $(CA)_n$ repeat were successfully used in previous research in an effort to establish the value of this dinucleotide repeat polymorphism in forensics (Bodenteich *et al.* 1992; Szibor *et al.* 1997). The primer sequences used for amplification of the CA repeat are: (L00484) (forward) 5'-CTCCCATACTACTAATCTCA-3' and (H00537) (reverse) 5'-TGGTTGGTTCGGGGTATG-3'.

A 200 μM stock solution was prepared for each of the Invitrogen™ primers and the stock solutions were then subsequently diluted into 10 μM working aliquots. PCR reaction components (AmpliTaq Gold™ DNA polymerase, $MgCl_2$, 10X PCR buffer, BSA, dNTPs) were obtained from Applied Biosystems. Gels were prepared with Certified Molecular Biology Grade Agarose (BIO-RAD), *tris*-borate (TBE) or sodium borate (SB) buffer (SIGMA), and ethidium bromide stain. Two different DNA ladders [pBR322/HinfI (JenaBioscience) and DNA Ladder 50631 (Cambrex BioScience)] were used as comparison standards for PCR products, with fragment size ranges of 75-1632 bp and 50-2500 bp, respectively. PCR products on the gels were visualized with Doc-It®LS Image Acquisition Software (Life Science Software from UVP).

Phase I:

Previously extracted DNA was obtained from the UNTHSC DNA Identity Lab in Fort Worth. These reference DNA samples were extracted using the TECAN Schweiz AG robot (Freedom EVO 100 Base), which yields mean DNA quantities of 1.29 ng/μL ± 1.227 (range 0.0298 – 4.63 ng/μL) [per validation study conducted by a UNTHSC lab analyst (results not published)].

Three separate protocols were designed for single locus amplification of the CA repeat to compare band intensities when primer or DNA concentrations were changed, as outlined in Table 1 (below):

Table 1: *Phase I* PCR Master Mix Protocols for the mtDNA CA Repeat (25 μL total reaction volume)	
Standard Protocol:	2 μL $MgCl_2$ (25 mM) 1 μL dNTPs (25 mM) 2.5 μL PCR rxn buffer (10X) 2.5 μL BSA (1.6 mg/ml) 0.5 μL AmpliTaq Gold™ DNA Polymerase (5 U/μL) 0.5 μL Forward Primer (L00484) (10 μM) 0.5 μL Reverse Primer (H00537) (10 μM) 14.5 μL molecular grade H_20 1 μL DNA
Protocol #2: *(addition of more primer)*	2 μL $MgCl_2$ (25 mM) 1 μL dNTPs (25 mM) 2.5 μL PCR rxn buffer (10X) 2.5 μL BSA (1.6 mg/ml) 0.5 μL AmpliTaq Gold™ DNA Polymerase (5 U/μL)

	1.0 μL Forward Primer (L00484) (10 μM) 1.0 μL Reverse Primer (H00537) (10 μM) 13.5 μL molecular grade H_20 1 μL DNA
Protocol #3: *(addition of more DNA)*	2 μL $MgCl_2$ (25 mM) 1 μL dNTPs (25 mM) 2.5 μL PCR rxn buffer (10X) 2.5 μL BSA (1.6 mg/ml) 0.5 μL AmpliTaq Gold™ DNA Polymerase (5 U/μL) 0.5 μL Forward Primer (L00484) (10 μM) 0.5 μL Reverse Primer (H00537) (10 μM) 12.5 μL molecular grade H_20 3.0 μL DNA

Using the Applied Biosystems GeneAmp PCR System 2400, the mtDNA CA repeat was amplified with the following thermocycling parameters: enzyme activation (hot start) at 94°C (11 min.) followed by 32 cycles of denaturation at 94°C for 45 sec., annealing at 56°C for 30 sec., and extension at 72°C for 1 min.; and a final extension at 60°C for 45 min. (∞ at 4°C).

Standard amplification protocols were also designed individually for the amelogenin and D3S1358 singleplex reactions, as well as for a duplex reaction of the two loci (Table 2). Furthermore, the thermocycling parameters for these two nuclear DNA loci were as follows: enzyme activation at 95°C (11 min.) followed by 28 cycles of denaturation at 94°C for 1 min., annealing at 58°C for 1 min., and extension at 72°C for 1 min.; and a final extension at 60°C for 45 min. (∞ at 4°C). The amelogenin and D3S1358 loci were amplified using the Perkin Elmer GeneAmp PCR System 2400.

Table 2: *Phase I* PCR Master Mix Protocols for Amelogenin and D3S1358 (25 μL total reaction volume)	
Standard Protocol: *(Amelogenin only)*	2 μL $MgCl_2$ (25 mM) 1 μL dNTPs (25 mM) 2.5 μL PCR rxn buffer (10X)

	2.5 μL BSA (1.6 mg/ml) 0.5 μL AmpliTaq Gold™ DNA Polymerase (5 U/μL) 1.0 μL *Amelogenin* Forward Primer (10 μM) 1.0 μL *Amelogenin* Reverse Primer (10 μM) 12.5 μL molecular grade H_20 2 μL DNA
<u>Standard Protocol:</u> *(D3S1358 only)*	2 μL $MgCl_2$ (25 mM) 1 μL dNTPs (25 mM) 2.5 μL PCR rxn buffer (10X) 2.5 μL BSA (1.6 mg/ml) 0.5 μL AmpliTaq Gold™ DNA Polymerase (5 U/μL) 1.0 μL *D3S1358* Forward Primer (10 μM) 1.0 μL *D3S1358* Reverse Primer (10 μM) 12.5 μL molecular grade H_20 2 μL DNA
<u>Standard Protocol:</u> *(Amelogenin &* *D3S1358 duplex)*	2 μL $MgCl_2$ (25 mM) 1 μL dNTPs (25 mM) 2.5 μL PCR rxn buffer (10X) 2.5 μL BSA (1.6 mg/ml) 0.5 μL AmpliTaq Gold™ DNA Polymerase (5 U/μL) 1.0 μL *Amelogenin* Forward Primer (10 μM) 1.0 μL *Amelogenin* Reverse Primer (10 μM) 1.0 μL *D3S1358* Forward Primer (10 μM) 1.0 μL *D3S1358* Reverse Primer (10 μM) 10.5 μL molecular grade H_20 2 μL DNA

Post-amplification, 5 μL of each sample (along with 2 μL loading dye OG•XC) was loaded into the successive wells of a 2% agarose (1X TBE) mini-gel (dimensions 9 cm x 6 cm), with the first well containing 10 μL pBR322/HinfI DNA ladder. A negative control lane was included, and the gel was electrophoresed at 137 volts for 1 hour using a Hsi Hoefer Transphor/Electrophoresis DC power supply (no photo available due to broken camera). A few days later, the mtDNA CA Repeat samples were re-run on a mini-gel to duplicate the results that couldn't previously be photographed (Figure 1). Additionally, amplified samples from the amelogenin standard protocol, the D3S1358/amelogenin duplex standard protocol, and CA Repeat protocol #2 were further analyzed with the Agilent 2100 (Agilent Technologies, Inc.), as shown in Figures 2-5 in the results section.

In a subsequent experiment, the thermocycling parameters for amelogenin and D3S1358 were changed. Specifically, the annealing temperature for these two loci was re-adjusted back to 59°C and, in addition, new PCR master mix protocols were designed. Amelogenin and D3S1358 hence were then re-amplified (both individually and as a duplex reaction) using new protocols which included use of the AmpFlSTR PCR Reaction Mix (rather than individual components), as described in Table 3 (below):

Table 3: Alternate *Phase I* PCR Master Mix Protocols for Amelogenin & D3S1358 (50 µL total reaction volume)	
Protocol A1: *(D3S1358 only)*	20 µL AmpFlSTR PCR reaction mix 1 µL AmpliTaq Gold™ DNA Polymerase 2 µL *D3S1358* Forward Primer (10 µM) 2 µL *D3S1358* Reverse Primer (10 µM) 24 µL molecular grade H$_2$0 1 µL DNA
Protocol B1: *(D3S1358 only)*	20 µL AmpFlSTR PCR reaction mix 1 µL AmpliTaq Gold™ DNA Polymerase 5 µL *D3S1358* Forward Primer (10 µM) 5 µL *D3S1358* Reverse Primer (10 µM) 18 µL molecular grade H$_2$0 1 µL DNA
Protocol C1: *(D3S1358 only)*	20 µL AmpFlSTR PCR reaction mix 1 µL AmpliTaq Gold™ DNA Polymerase 2 µL *D3S1358* Forward Primer (10 µM) 2 µL *D3S1358* Reverse Primer (10 µM) 22 µL molecular grade H$_2$0

	3 µL DNA
<u>Protocol D1:</u> *(D3S1358 only)*	20 µL AmpFlSTR PCR reaction mix 1 µL AmpliTaq Gold™ DNA Polymerase 5 µL *D3S1358* Forward Primer (10 µM) 5 µL *D3S1358* Reverse Primer (10 µM) 16 µL molecular grade H_20 3 µL DNA
<u>Protocol E1:</u> *(Amelogenin only)*	20 µL AmpFlSTR PCR reaction mix 1 µL AmpliTaq Gold™ DNA Polymerase 2 µL *Amelogenin* Forward Primer (10 µM) 2 µL *Amelogenin* Reverse Primer (10 µM) 24 µL molecular grade H_20 1 µL DNA
<u>Protocol F1:</u> *(Amelogenin only)*	20 µL AmpFlSTR PCR reaction mix 1 µL AmpliTaq Gold™ DNA Polymerase 5 µL *Amelogenin* Forward Primer (10 µM) 5 µL *Amelogenin* Reverse Primer (10 µM) 18 µL molecular grade H_20 1 µL DNA
<u>Protocol G1:</u> *(Amelogenin only)*	20 µL AmpFlSTR PCR reaction mix 1 µL AmpliTaq Gold™ DNA Polymerase 2 µL *Amelogenin* Forward Primer (10 µM) 2 µL *Amelogenin* Reverse Primer (10 µM) 22 µL molecular grade H_20 3 µL DNA

Protocol H1: *(Amelogenin only)*	20 µL AmpFlSTR PCR reaction mix 1 µL AmpliTaq Gold™ DNA Polymerase 5 µL *Amelogenin* Forward Primer (10 µM) 5 µL *Amelogenin* Reverse Primer (10 µM) 16 µL molecular grade H$_2$0 3 µL DNA
Protocol I1: *(D3S1358 & Amelogenin)*	20 µL AmpFlSTR PCR reaction mix 1 µL AmpliTaq Gold™ DNA Polymerase 2 µL *D3S1358* Forward Primer (10 µM) 2 µL *D3S1358* Reverse Primer (10 µM) 2 µL *Amelogenin* Forward Primer (10 µM) 2 µL *Amelogenin* Reverse Primer (10 µM) 20 µL molecular grade H$_2$0 1 µL DNA
Protocol J1: *(D3S1358 & Amelogenin)*	20 µL AmpFlSTR PCR reaction mix 1 µL AmpliTaq Gold™ DNA Polymerase 5 µL *D3S1358* Forward Primer (10 µM) 5 µL *D3S1358* Reverse Primer (10 µM) 5 µL *Amelogenin* Forward Primer (10 µM) 5 µL *Amelogenin* Reverse Primer (10 µM) 8 µL molecular grade H$_2$0 1 µL DNA
	20 µL AmpFlSTR PCR reaction mix 1 µL AmpliTaq Gold™ DNA Polymerase 2 µL *D3S1358* Forward Primer (10 µM)

Protocol K1: *(D3S1358 & Amelogenin)*	2 µL *D3S1358* Reverse Primer (10 µM) 2 µL *Amelogenin* Forward Primer (10 µM) 2 µL *Amelogenin* Reverse Primer (10 µM) 18 µL molecular grade H_20 3 µL DNA
Protocol L1: *(D3S1358 & Amelogenin)*	20 µL AmpFlSTR PCR reaction mix 1 µL AmpliTaq Gold™ DNA Polymerase 5 µL *D3S1358* Forward Primer (10 µM) 5 µL *D3S1358* Reverse Primer (10 µM) 5 µL *Amelogenin* Forward Primer (10 µM) 5 µL *Amelogenin* Reverse Primer (10 µM) 6 µL molecular grade H_20 3 µL DNA

Post-amplification, 5 µL of each sample (along with 1 µL loading dye OG•XC) was loaded into the successive wells of a 2% agarose (1X TBE) mini-gel (dimensions 9 cm x 6 cm), with the fifth and tenth wells containing 10 µL and 5 µL Cambrex DNA ladder 50631, respectively. The gel was electrophoresed at 134 volts for 1 hour with an HSI Hoefer Transphor/Electrophoresis DC power supply and then photographed, as shown in Figure 6.

Phase II:

As a result of the data generated from *Phase I* of this study, quality control checks on the primers were performed. Primer sequences were rechecked and the primer concentrations were verified with a Beckman DU-50 spectrophotometer. New DNA samples (buccal swabs) were obtained via phenol-chloroform extraction and Microcon 100™ concentration. Resulting DNA concentrations were determined using the gel quantitation method with 1% agarose in 1X TBE. The PCR master mix protocols used for each of the three loci were modeled after the *Phase I* protocols and are described in Table 4. Thermocycling parameters for these singleplex reactions were as follows: enzyme activation at 94°C (3 min.) followed by 30 cycles of denaturation at 94°C for 1 min., annealing at 54°C-66°C (gradient) for 30 sec., and extension at 72°C for 30 sec.; and a final extension at 72°C for 2 min. (∞ at 4°C). The samples were amplified using the MJ Research PTC-200 Peltier Thermal Cycler (Gradient Cycler). Post-amplification, 25 µL of each sample (along with 5 µL loading dye OG•XC) (30µL total) was loaded into the successive wells of 2% agarose gels in 1X TBE (dimensions 14.5 cm x 12 cm). The gels were electrophoresed at 115 volts for 1 hour with an Hsi Hoefer Transphor/Electrophoresis DC power supply, stained with ethidium bromide, and then visualized with Doc-It®LS Image Acquisition Software (Life Science Software from UVP) (Figures 7-8). Subsequently, the same thermocycling parameters used for the singleplex reactions were used for a multiplex attempt, with an increase in agarose percent from 2% to 2.5% (Figure 9).

Table 4: *Phase II* PCR Master Mix Protocols for Singleplex Amplifications of the mtDNA CA Repeat, Amelogenin, & D3S1358 (25 µL total reaction volume)

Locus	Master Mix	Final Conc.
CA Repeat:	2 µL $MgCl_2$ (25 mM) 1 µL dNTPs (25 mM) 2.5 µL PCR reaction buffer (10X) 2.5 µL BSA (1.6 mg/ml) 0.5 µL Taq Gold™ DNA polymerase (5 U/µL) 1.0 µL Forward Primer (L00484) (5 µM) 0.5 µL Reverse Primer (H00537) (10 µM) 14 µL molecular grade H_20 1 µL DNA (1 ng/µL)	2 mM $MgCl_2$ 250 µM dNTPs 1X rxn buffer 0.16 µg/µl BSA 2.5 U Taq Gold 0.2 µM primer 0.2 µM primer

D3S1358:	2 µL $MgCl_2$ (25 mM) 1 µL dNTPs (25 mM) 2.5 µL PCR reaction buffer (10X) 2.5 µL BSA (1.6 mg/ml) 0.5 µL Taq Gold™ DNA polymerase (5 U/µL) 1.0 µL *D3S1358* Forward Primer (10 µM) 1.0 µL *D3S1358* Reverse Primer (10 µM) 13.5 µL molecular grade H_20 1 µL DNA (1 ng/µL)	2 mM $MgCl_2$ 250 µM dNTPs 1X rxn buffer 0.16 µg/µl BSA 2.5 U Taq Gold 0.2 µM primer 0.2 µM primer
Amelogenin:	2 µL $MgCl_2$ (25 mM) 1 µL dNTPs (25 mM) 2.5 µL PCR reaction buffer (10X) 2.5 µL BSA (1.6 mg/ml) 0.5 µL Taq Gold™ DNA polymerase (5 U/µL) 1.0 µL *Amelogenin* Forward Primer (10 µM) 1.0 µL *Amelogenin* Reverse Primer (10 µM) 13.5 µL molecular grade H_20 1.0 µL 1 µL DNA (1 ng/µL)	2 mM $MgCl_2$ 250 µM dNTPs 1X rxn buffer 0.16 µg/µl BSA 2.5 U Taq Gold 0.2 µM primer 0.2 µM primer

After completing amplifications of the singleplex reactions, PCR reaction mix protocols were carried out for duplex reactions as described in Table 5 (below):

Table 5: *Phase II* PCR Master Mix Protocols for Duplex Amplifications of the mtDNA CA Repeat, Amelogenin, & D3S1358 (25 µL total reaction volume)

Amelogenin & D3S1358	2 µL $MgCl_2$ (25 mM) 1 µL dNTPs (25 mM) 2.5 µL PCR reaction buffer (10X) 2.5 µL BSA (1.6 mg/ml) 0.5 µL Taq Gold™ DNA polymerase (5 U/µL) 1.0 µL *Amelogenin* Forward Primer (10 µM)

	1.0 µL *Amelogenin* Reverse Primer (10 µM) 1.0 µL *D3S1358* Forward Primer (10 µM) 1.0 µL *D3S1358* Reverse Primer (10 µM) 11.5 µL molecular grade H_20 1 µL DNA (1 ng/µL)

Amelogenin & CA Repeat	2 µL $MgCl_2$ (25 mM) 1 µL dNTPs (25 mM) 2.5 µL PCR reaction buffer (10X) 2.5 µL BSA (1.6 mg/ml) 0.5 µL Taq Gold™ DNA polymerase (5 U/µL) 1.0 µL *Amelogenin* Forward Primer (10 µM) 1.0 µL *Amelogenin* Reverse Primer (10 µM) 1.0 µL *CA Repeat* Forward Primer (L00484) (5 µM) 0.5 µL *CA Repeat* Reverse Primer (H00537) (10 µM) 12.0 µL molecular grade H_20 1 µL DNA (1 ng/µL)
D3S1358 & CA Repeat	2 µL $MgCl_2$ (25 mM) 1 µL dNTPs (25 mM) 2.5 µL PCR reaction buffer (10X) 2.5 µL BSA (1.6 mg/ml) 0.5 µL Taq Gold™ DNA polymerase (5 U/µL) 1.0 µL *D3S1358* Forward Primer (10 µM) 1.0 µL *D3S1358* Reverse Primer (10 µM) 1.0 µL *CA Repeat* Forward Primer (L00484) (5 µM) 0.5 µL *CA Repeat* Reverse Primer (H00537) (10 µM) 12.0 µL molecular grade H_20

	1 μL DNA (1 ng/μL)

Duplex samples were amplified using the same thermocycling parameters as the singleplex reactions [MJ Research PTC-200 Peltier Thermal Cycler (Gradient Cycler)] but with an annealing temperature range of 54°C-62°C . Post-amplification, 7 μL of each sample (along with 3 μL loading dye OG•XC) (10μL total) was loaded into the successive wells of 2.5% agarose gels in 1X sodium borate (SB) (dimensions 14.5 cm x 12 cm). Ten μL of pBR322/HinfI ladder was added to the first lane of the gel for PCR product comparison. The gels were electrophoresed at 226 volts with an Hsi Hoefer Transphor/Electrophoresis DC power supply, stained with ethidium bromide, and then visualized with Doc-It®LS Image Acquisition Software (Life Science Software from UVP). Results of the duplex reactions are illustrated in Figures 10-12.

After completion of the duplex attempts, an amplification master mix was made for the multiplex reaction, as described in Table 6 (below):

Table 6: *Phase II* PCR Master Mix Protocol for Multiplex Amplification of the mtDNA CA Repeat, Amelogenin, & D3S1358 (25 μL total reaction volume)

2 µL $MgCl_2$ (25 mM)

1 µL dNTPs (25 mM)

2.5 µL PCR reaction buffer (10X)

2.5 µL BSA (1.6 mg/ml)

0.5 µL Taq Gold™ DNA polymerase (5 U/µL)

1.0 µL *Amelogenin* Forward Primer (10 µM)

1.0 µL *Amelogenin* Reverse Primer (10 µM)

1.0 µL *D3S1358* Forward Primer (10 µM)

1.0 µL *D3S1358* Reverse Primer (10 µM)

1.0 µL *CA Repeat* Forward Primer (L00484) (5 µM)

0.5 µL *CA Repeat* Reverse Primer (H00537) (10 µM)

10 µL molecular grade H_20

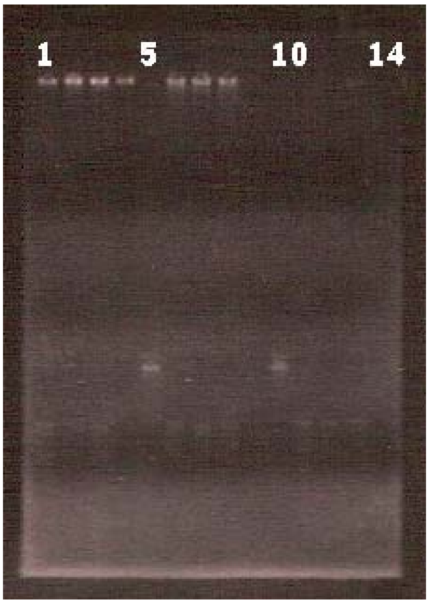
1 µL DNA (1 ng/µL)

The multiplex samples were amplified using the same thermocycling parameters as the duplex reactions [MJ Research PTC-200 Peltier Thermal Cycler (Gradient Cycler)] with an annealing temperature range of 54°C-62°C . Post-amplification, 7 µL of each sample (along with 3 µL loading dye OG•XC) (10µL total) was loaded into the successive wells of a 2.5% agarose gel in 1X sodium borate (SB) (dimensions 14.5 cm x 12 cm). Ten µL of pBR322/HinfI ladder was added to the first lane of the gel for PCR product comparison. The gel was electrophoresed at 226 volts with an Hsi Hoefer Transphor/Electrophoresis DC power supply, stained with ethidium bromide, and then visualized with Doc-It®LS Image Acquisition Software (Life Science Software from UVP). Results of this multiplex amplification are shown in Figure 13.

Phase I:

In the *Phase I* thermocycling parameters for the amplification of the mtDNA CA repeat, the total number of cycles was decreased from the standard 35 cycles used to amplify mtDNA to 32 cycles. Since amelogenin and D3S1358 have both previously been successfully amplified using 28-30 cycles, the decrease in cycle number for the CA repeat was a forethought to the upcoming multiplex attempts. In addition, the annealing temperature for the amplification of the two nuclear DNA loci was decreased from 59°C to 58°C in order to see if these loci would amplify at a temperature closer to the standard annealing temperature used for mtDNA samples. The first *Phase I* gel ran with the various single locus amplification protocols and the amelogenin/D3S1358 duplex could not be photographed due to broken equipment. All negative controls performed as expected (clear lanes), but the ladder only faintly appeared on the gel. Smears were seen in lanes 1-3 (D3S1358 standard protocol) and lanes 5-7 (amelogenin standard protocol), but bright bands were seen in lanes 10-12 (amelogenin/D3S1358 duplex standard protocol) and lanes 14-16 (CA repeat standard protocol). Lanes 19-21 (CA repeat protocol #2) and lanes 23-25 (CA repeat protocol #3) also produced bright dense bands in the ladder region that corresponds with the expected amplicon sizes of interest. Figure 1 is a 2% agarose (1X TBE) gel that represents the results of the mtDNA CA repeat amplifications.

Figure 1: *Phase I* PCR Amplification of the D-loop 3' $(CA)_n$ Dinucleotide Repeat



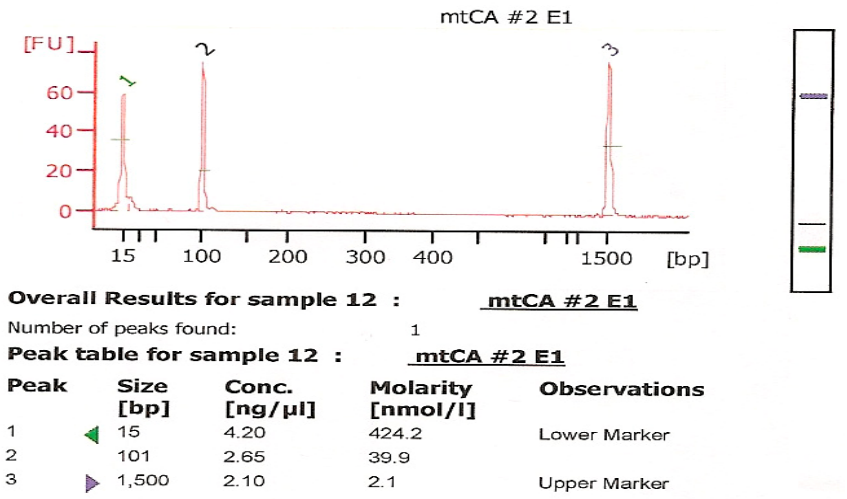
Lane 1	Standard Protocol – DNA Sample A2
Lane 2	Standard Protocol – DNA Sample D1
Lane 3	Standard Protocol – DNA Sample E1
Lane 4	Negative Control
Lane 5	Cambrex DNA Ladder #50631
Lane 6	Protocol #2 – DNA Sample A2
Lane 7	Protocol #2 – DNA Sample D1
Lane 8	Protocol #2 – DNA Sample E1
Lane 9	Negative Control
Lane 10	Cambrex DNA Ladder #50631
Lane 11	Protocol #3 – DNA Sample A2
Lane 12	Protocol #3 – DNA Sample D1

Lane 13	Protocol #3 – DNA Sample E1
Lane 14	Negative Control

Lanes 5 and 10 both contained Cambrex 50631 DNA ladder, which only faintly appeared after ethidium bromide staining. Lanes 4, 9, and 14 were negative controls and produced no bands. Faint bands were distinguishable in lanes 1-3 (CA repeat standard protocol), lanes 6-8 (CA repeat protocol #2), and lanes 11-13 (CA repeat protocol #3).

To further investigate the nature of these bands, the mtDNA CA repeat samples were analyzed with the Agilent 2100, as shown in Figure 2.

Figure 2: Agilent 2100 Analysis of CA repeat (Protocol #2)



Overall Results for sample 12 : **mtCA #2 E1**

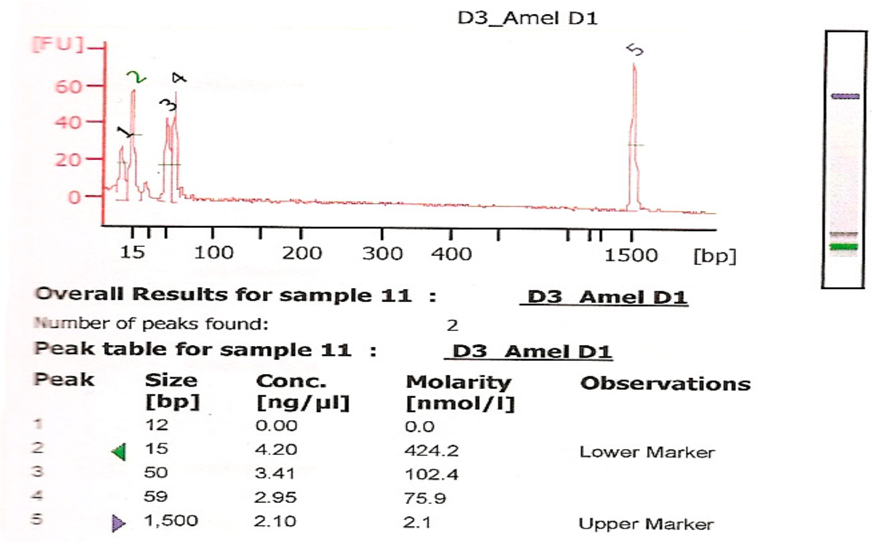
Number of peaks found: 1

Peak table for sample 12 : **mtCA #2 E1**

Peak	Size [bp]	Conc. [ng/µl]	Molarity [nmol/l]	Observations
1	◀ 15	4.20	424.2	Lower Marker
2	101	2.65	39.9	
3	▶ 1,500	2.10	2.1	Upper Marker

As shown in Figure 2, peak 2 in the Agilent electropherogram corresponds to a 101 bp PCR product, which conflicts with the expected 86-94 bp amplicon size range for the mtDNA CA repeat. Agilent analysis was also performed on the D3S1358/amelogenin duplex (Figure 3) and the amelogenin singleplex standard protocol samples (Figure 4).

Figure 3: Agilent 2100 Analysis of D3S1358/Amelogenin Duplex



Overall Results for sample 11 : **D3 Amel D1**

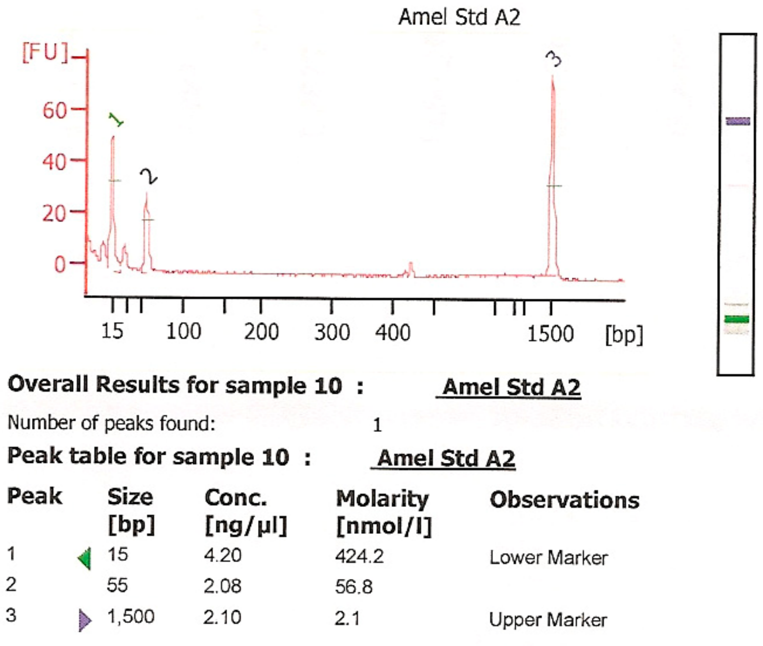
Number of peaks found: 2

Peak table for sample 11 : **D3 Amel D1**

Peak		Size [bp]	Conc. [ng/µl]	Molarity [nmol/l]	Observations
1		12	0.00	0.0	
2	◀	15	4.20	424.2	Lower Marker
3		50	3.41	102.4	
4		59	2.95	75.9	
5	▶	1,500	2.10	2.1	Upper Marker

None of the peaks in Figure 3 correspond to the expected amplicon sizes of 97-151 bp and 106/112 bp for D3S1358 and amelogenin, respectively. Considering the length of the primers chosen for this study, it is likely that peaks 3 and 4 represent primer-dimers generated during the PCR amplification reaction.

Figure 4: Agilent 2100 Analysis of Amelogenin (Standard Protocol)



Overall Results for sample 10 : **Amel Std A2**

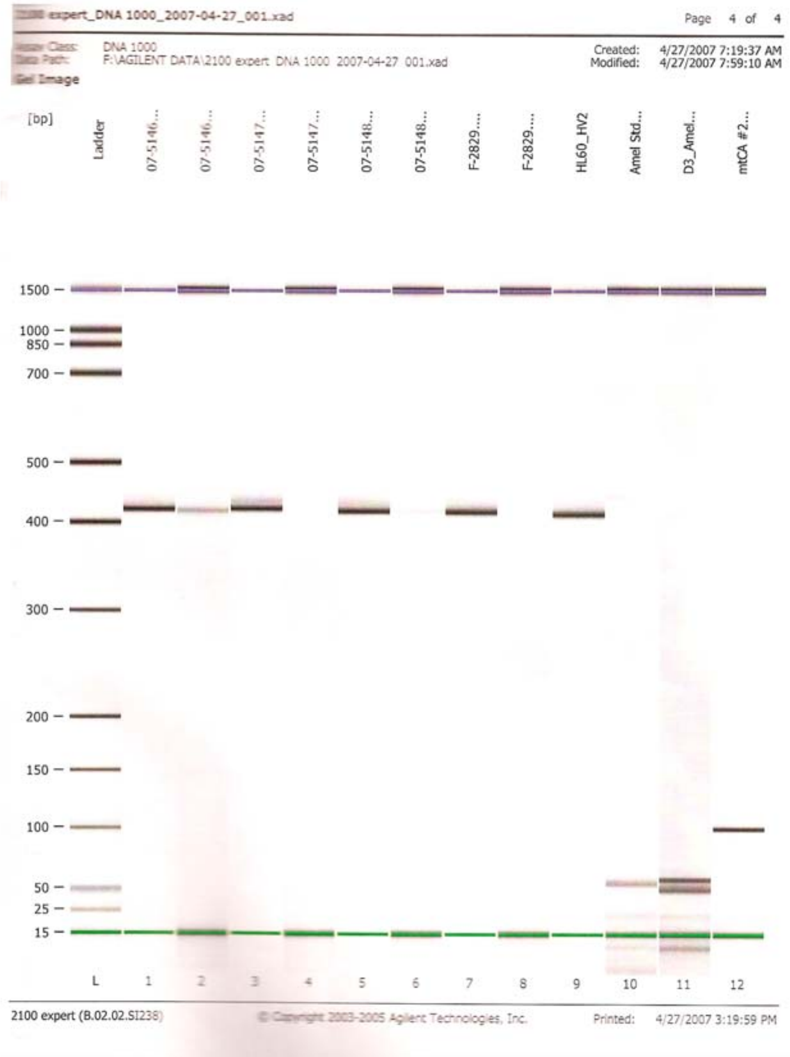
Number of peaks found: 1

Peak table for sample 10 : **Amel Std A2**

Peak		Size [bp]	Conc. [ng/µl]	Molarity [nmol/l]	Observations
1	◀	15	4.20	424.2	Lower Marker
2		55	2.08	56.8	
3	▶	1,500	2.10	2.1	Upper Marker

Similarly, none of the peaks in Figure 4 correspond to the 106 bp and 112 bp expected PCR product sizes for amelogenin. Peak 2 in the electropherogram is likely a primer-dimer formed during the amplification of this sample. Figure 5 represents a gel image that summarizes the Agilent 2100 analysis of these samples. Lane 11 in the figure represents the results from the amelogenin standard protocol, while lanes 12 and 13 correspond to the D3S1358/amelogenin duplex reaction and protocol #2 for the mtDNA CA repeat, respectively.

Figure 5: Agilent 2100 Analysis of the mtDNA CA Repeat (Protocol #2), D3S1358/Amelogenin Duplex, and Amelogenin (Standard Protocol)



2100 expert_DNA 1000_2007-04-27_001.xad Page 4 of 4

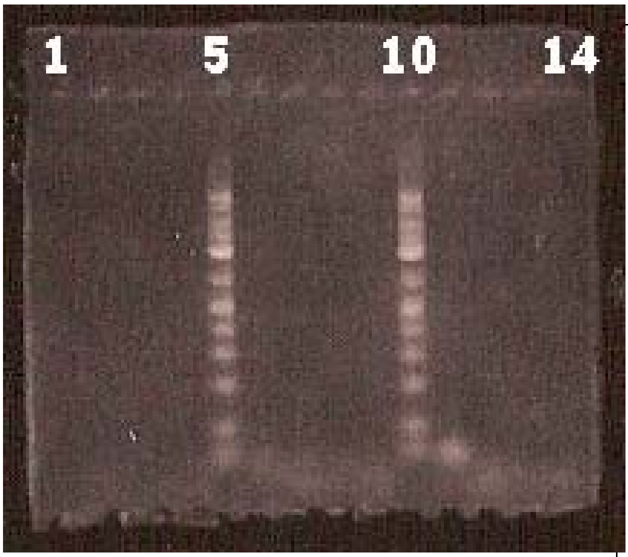
Assay Class: DNA 1000
Data Path: F:\AGILENT DATA\2100 expert DNA 1000 2007-04-27 001.xad
Created: 4/27/2007 7:19:37 AM
Modified: 4/27/2007 7:59:10 AM

Gel Image

2100 expert (B.02.02.SI238) © Copyright 2003-2005 Agilent Technologies, Inc. Printed: 4/27/2007 3:19:59 PM

Figure 6 represents the results of PCR after changing the thermocycling parameters for amelogenin and D3S1358 (specifically, re-adjusting the annealing temperature back to 59°C) and utilizing the AmpFlSTR PCR Reaction Mix (instead of individual components). As can be seen in the gel, the ladder (Cambrex DNA ladder 50631) performed as expected, but most of the lanes contained no visible PCR product.

Figure 6: *Phase I* Amplification of Amelogenin and D3S1358 with AmpFlSTR PCR Reaction Mix (2% Agarose in 1X TBE)



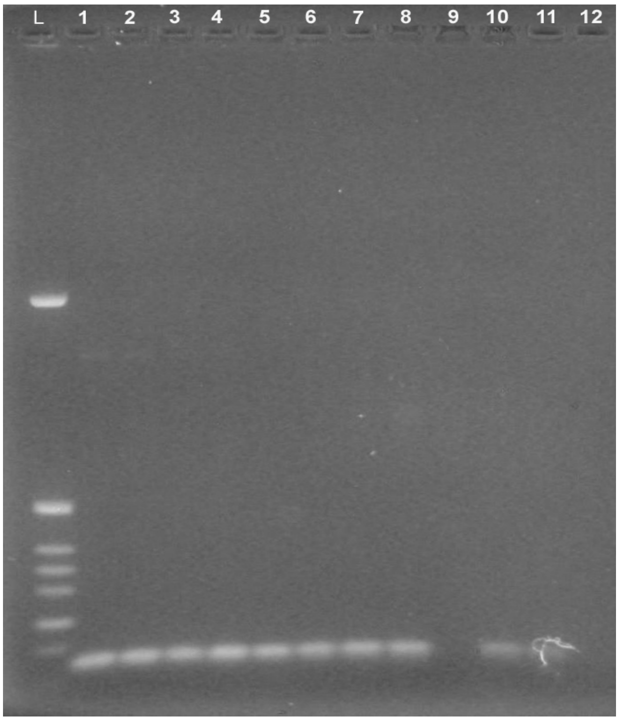
Lane 1	Protocol A1 (D3S1358)
Lane 2	Protocol B1 (D3S1358)
Lane 3	Protocol C1 (D3S1358)
Lane 4	Protocol D1 (D3S1358)
Lane 5	Cambrex Ladder 50631
Lane 6	Protocol E1 (Amelogenin)
Lane 7	Protocol F1 (Amelogenin)
Lane 8	Protocol G1
Lane 9	(Amelogenin)
Lane 10	Protocol H1
Lane 11	(Amelogenin)
Lane 12	Cambrex Ladder 50631
Lane 13	Protocol I1 (Both)
Lane 14	Protocol J1 (Both)
	Protocol K1 (Both)
	Protocol L1 (Both)

A faint band, however, is seen in Lane 6 (amelogenin standard protocol E1) and a brighter band appeared in Lane 11 (D3S1358/amelogenin duplex protocol I1). Primer-dimer bands appear to be present in lanes 6-9 and lanes 11-13.

Phase II:

Due to indeterminate results from the *Phase I* protocols, new DNA samples were extracted from buccal swabs, primer sequences were double-checked, and primer concentrations were verified via spectrophotometry. During these quality control checks, primer sequences were determined to be correct, but verification of primer concentrations revealed that the mtDNA CA repeat forward primer (L00484) was approximately 100 µM (only half of the previously indicated concentration of 200 µM). Appropriate adjustments were made to the *Phase II* amplification protocols. As depicted in Figures 7-8, singleplex reactions for each of the three loci were carried out on an annealing temperature gradient of 54°-66°C. This range was programmed into the MJ Research PTC-200 Peltier Thermal Cycler, and the instrument automatically (and seemingly arbitrarily) determined the temperature increments for each well. The range of 54°-66°C was chosen because it spans the successful annealing temperatures used for these three loci in previous research (as indicated in the literature).

Figure 7: *Phase II* Amelogenin Singleplex in 2% Agarose (1X TBE) With Annealing Temperature Gradient 54°-66°C



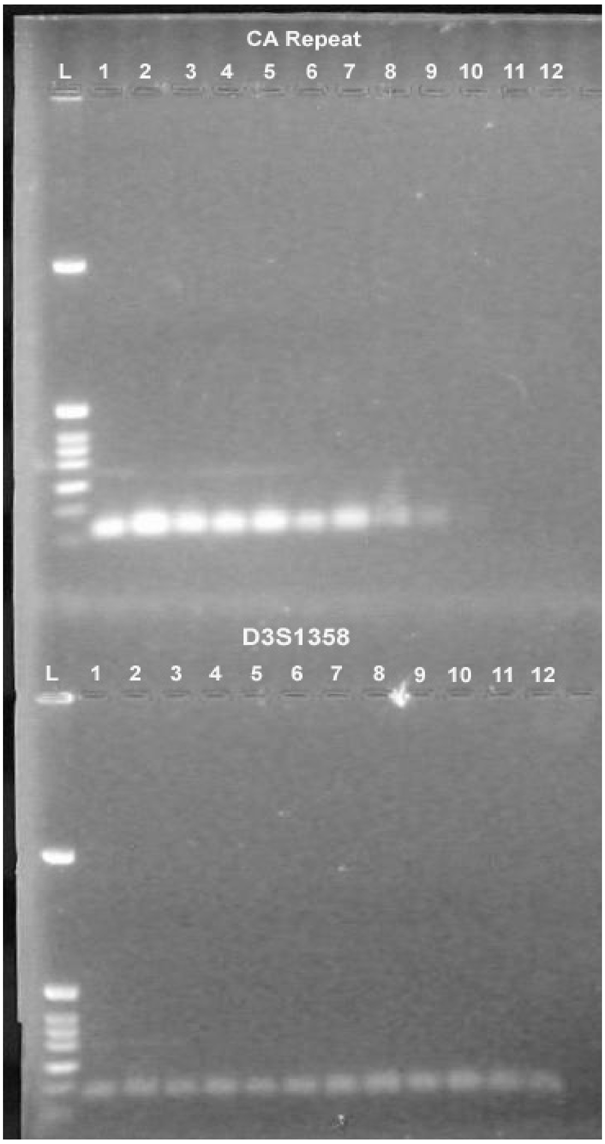
Amelogenin

Lane "L"	pBR322/HinfI DNA Ladder
Lane 1	Annealing Temp. 54°C
Lane 2	Annealing Temp. 54.3°C
Lane 3	Annealing Temp. 55.1°C
Lane 4	Annealing Temp. 56°C
Lane 5	Annealing Temp. 57.4°C
Lane 6	Annealing Temp. 59.2°C
Lane 7	Annealing Temp. 61.2°C
Lane 8	Annealing Temp. 62.9°C
Lane 9	Annealing Temp. 64.2°C
Lane 10	Annealing Temp. 65.1°C
Lane 11	Annealing Temp. 65.8°C
Lane 12	Annealing Temp. 66°C

PCR product was observed in all lanes throughout the range of annealing temperatures, with the exception of lane 9 and lane 12. The bands in lane 10 and lane

11 are very faint, indicating that the annealing temperatures in these lanes (65.8°C and 66°C, respectively) are not optimum for this locus.

Figure 8: *Phase II* CA Repeat and D3S1358 Singleplexes in 2% Agarose (1X TBE) w/Annealing Temperature Gradient 54°-66°C



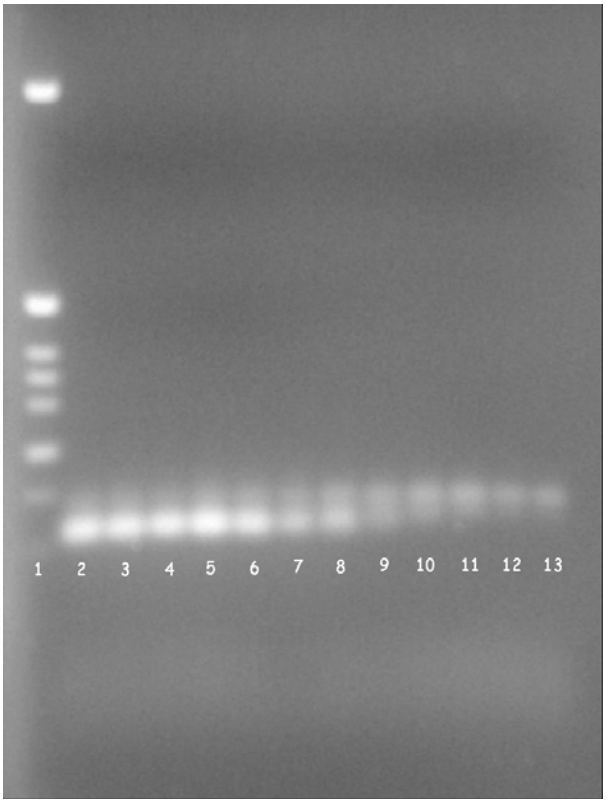
Lane "L"	pBR322/HinfI DNA Ladder
Lane 1	Annealing Temp. 54°C
Lane 2	Annealing Temp. 54.3°C
Lane 3	Annealing Temp. 55.1°C
Lane 4	Annealing Temp. 56°C
Lane 5	Annealing Temp. 57.4°C
Lane 6	Annealing Temp. 59.2°C
Lane 7	Annealing Temp. 61.2°C
Lane 8	Annealing Temp. 62.9°C
Lane 9	Annealing Temp. 64.2°C
Lane 10	Annealing Temp. 65.1°C
Lane 11	Annealing Temp. 65.8°C

Lane 12	Annealing Temp. 66°C

PCR product was observed in lanes 1-9 for the mtDNA CA repeat, although the bands in lanes 8 and 9 are faint. No PCR product was observed in lanes 10-12. Since lane 7 corresponds to an annealing temperature of 61.2°C, the faintness or lack of band presence in the lanes past this point indicate that annealing temperatures above 62°C are not optimum for this mtDNA CA repeat.

Furthermore, PCR product was observed in all lanes for the D3S1358 locus, spanning the entire 54°-66°C annealing temperature range. Due the fact that each individual loci amplified successfully in the singleplex reactions, a multiplex was attempted. Since the amplicon sizes for each of these three loci are close in range, the agarose percentage was increased from 2% to 2.5% in an effort to better resolve and separate the bands. Figure 9 (next page) shows the results of this multiplex attempt. The gel depicted in Figure 9 was left in destaining solution for too long and, as a result, the bands had dissipated/diffused before it could be photographed. Nonetheless, more than one PCR product band appears to be present in all of the lanes (most notably in lanes 2-8, which correspond to annealing temperatures of 54°-61.2°C).

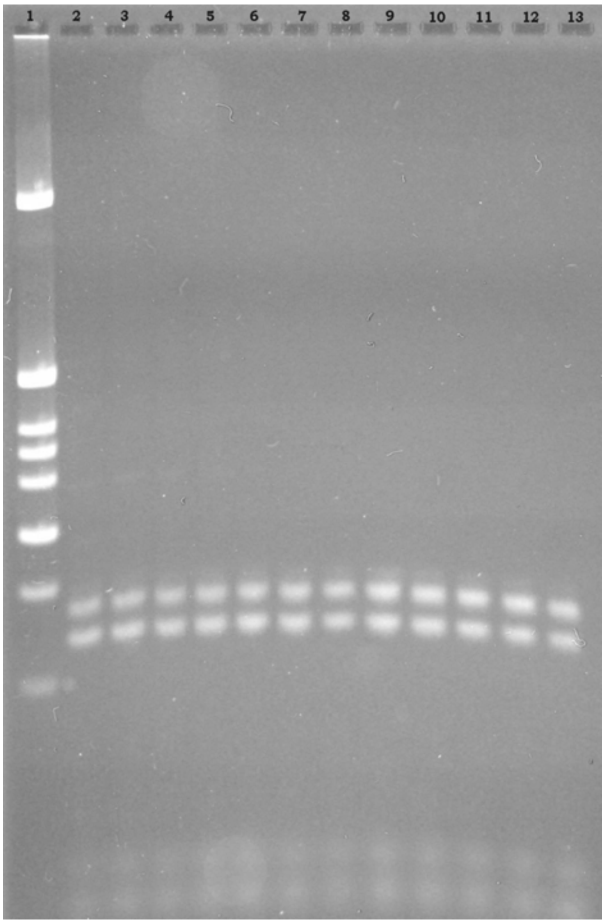
Figure 9: *Phase II* Multiplex Attempt in 2.5% Agarose (1X TBE) w/Annealing Temperature Gradient 54°-66°C



Lane 1	pBR322/HinfI DNA Ladder
Lane 2	Annealing Temp. 54°C
Lane 3	Annealing Temp. 54.3°C
Lane 4	Annealing Temp. 55.1°C
Lane 5	Annealing Temp. 56°C
Lane 6	Annealing Temp. 57.4°C
Lane 7	Annealing Temp. 59.2°C
Lane 8	Annealing Temp. 61.2°C
Lane 9	Annealing Temp. 62.9°C
Lane 10	Annealing Temp. 64.2°C
Lane 11	Annealing Temp. 65.1°C
Lane 12	Annealing Temp. 65.8°C
Lane 13	Annealing Temp. 66°C

Based on the results of the singleplex reactions and the latter mentioned multiplex attempt, the annealing temperature range for the duplex reactions was adjusted to 54°-62°C to accommodate a temperature range that is optimum for all three loci. The 2.5% agarose gels for the duplex reactions were run in 1X sodium borate (SB) buffer instead of 1X TBE because sodium borate has a lower conductivity, produces sharper bands, and can be run at higher speeds than gels made with TBE buffer. Results from the duplex amplification reactions are shown in Figures 10-12.

Figure 10: *Phase II* Amelogenin & D3S1358 Duplex in 2.5% Agarose (1X SB) w/Annealing Temperature Gradient 54°-62°C

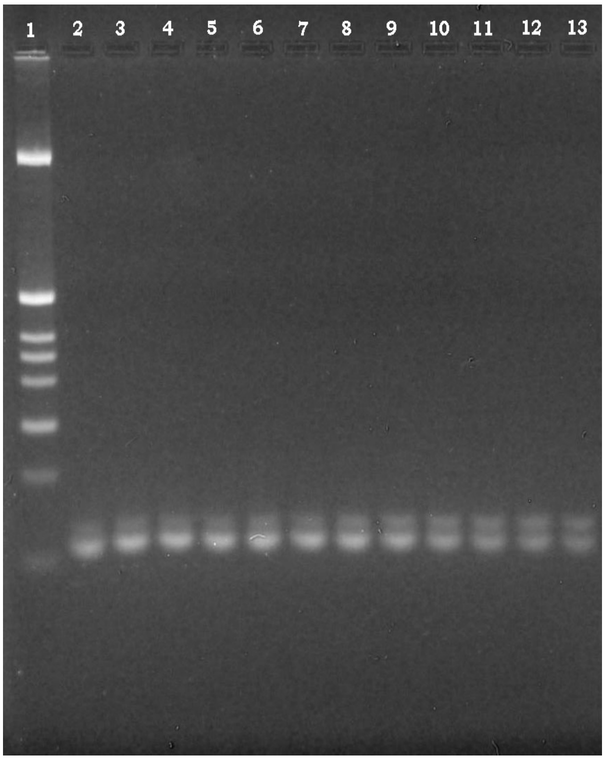


Amelogenin/D3S1358 Duplex

Lane 1	pBR322/HinfI DNA Ladder
Lane 2	Annealing Temp. 54°C
Lane 3	Annealing Temp. 54.2°C
Lane 4	Annealing Temp. 54.7°C
Lane 5	Annealing Temp. 55.3°C
Lane 6	Annealing Temp. 56.3°C
Lane 7	Annealing Temp. 57.5°C
Lane 8	Annealing Temp. 58.8°C
Lane 9	Annealing Temp. 60°C
Lane 10	Annealing Temp. 60.8°C
Lane 11	Annealing Temp. 61.4°C
Lane 12	Annealing Temp. 61.9°C
Lane 13	Annealing Temp. 62°C

Both amelogenin and D3S1358 amplified successfully throughout the entire selected annealing temperature range. However, the darkest bands (and hence the most DNA) were found in lanes 9 and 10 (annealing temperatures 60°C and 60.8°C, respectively). The two bands seen in lanes 2-13 on this gel fall between the 154 bp and 75 bp bands in the ladder, an appropriate location for loci with expected amplicon sizes of 97-151 bp (D3S1358) and 106/112 bp (amelogenin). The top band in the row represents the D3S1358 locus, while the bottom band represents PCR product from amplification of the amelogenin sex-determining locus.

Figure 11: *Phase II* Amelogenin & CA Repeat Duplex in 2.5% Agarose (1X SB) w/Annealing Temperature Gradient 54°-62°C



Amelogenin/CA Repeat Duplex

Lane 1	pBR322/HinfI DNA Ladder
Lane 2	Annealing Temp. 54°C
Lane 3	Annealing Temp. 54.2°C
Lane 4	Annealing Temp. 54.7°C
Lane 5	Annealing Temp. 55.3°C
Lane 6	Annealing Temp. 56.3°C
Lane 7	Annealing Temp. 57.5°C
Lane 8	Annealing Temp. 58.8°C
Lane 9	Annealing Temp. 60°C

Lane 10	Annealing Temp. 60.8°C
Lane 11	Annealing Temp. 61.4°C
Lane 12	Annealing Temp. 61.9°C
Lane 13	Annealing Temp. 62°C

Again, both amelogenin and the CA repeat amplified successfully throughout the entire selected annealing temperature range. However, preferential amplification of the CA repeat (bottom band) occurred through lane 9 (annealing temperature 60°C). The two bands seen in lanes 2-13 on this gel again fall between the 154 bp and 75 bp bands on the ladder, which is consistent with the expected amplicon sizes for amelogenin (106/112 bp) and the mtDNA CA repeat (86-94 bp). The top band in the row represents the amelogenin locus and the bottom band represents PCR product from the amplification of the mtDNA CA repeat.

In Figure 12, preferential amplification of the mtDNA CA repeat is again seen in lanes 2-9 (which corresponds to annealing temperatures of 54°-60°C), but both loci amplified across the entire annealing temperature range. The best results here are seen in lanes 10-13, where the two bands are relatively equal in intensity (and hence one locus is not outcompeting the other for components in the master mix).

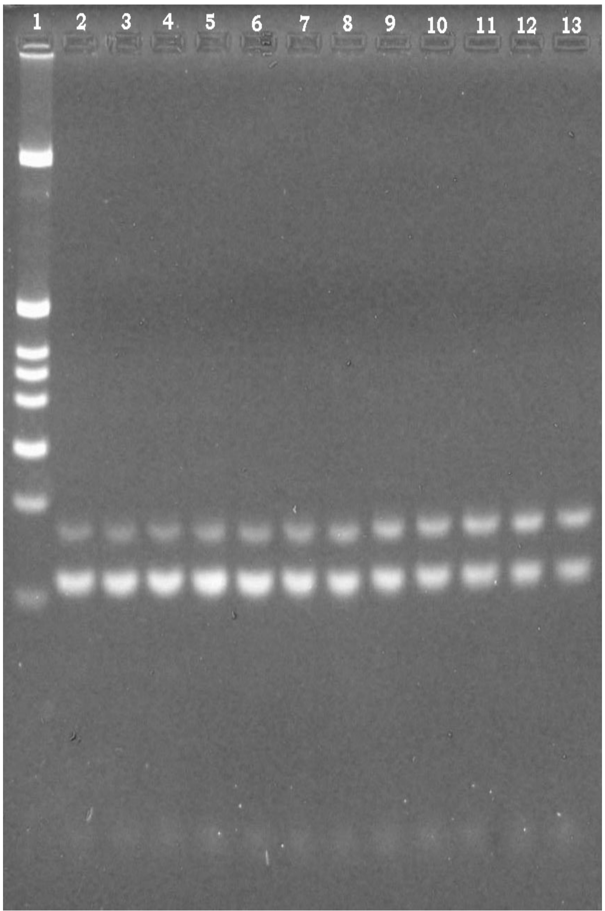


Figure 12: *Phase II* D3S1358 & CA Repeat Duplex in 2.5% Agarose (1X SB) w/Annealing Temperature Gradient 54°-62°C

D3S1358/CA Repeat Duplex

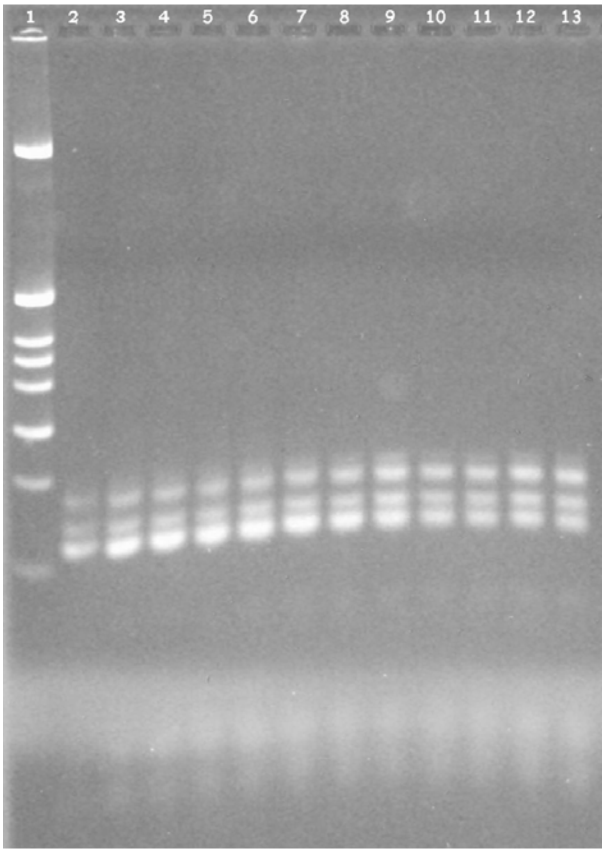
Lane 1	pBR322/HinfI DNA Ladder

Lane 2	Annealing Temp. 54°C
Lane 3	Annealing Temp. 54.2°C
Lane 4	Annealing Temp. 54.7°C
Lane 5	Annealing Temp. 55.3°C
Lane 6	Annealing Temp. 56.3°C
Lane 7	Annealing Temp. 57.5°C
Lane 8	Annealing Temp. 58.8°C
Lane 9	Annealing Temp. 60°C
Lane 10	Annealing Temp. 60.8°C
Lane 11	Annealing Temp. 61.4°C
Lane 12	Annealing Temp. 61.9°C
Lane 13	Annealing Temp. 62°C

The top band in each row in Figure 12 represents the D3S1358 locus, while the bottom band is PCR product from amplification of the mtDNA CA repeat. These two bands again fall between the 154 bp and 75 bp bands on the ladder, an appropriate location for loci with expected amplicon sizes of 97-151 bp (D3S1358) and 86-94 bp (CA repeat).

After successful amplification of the three different duplex combinations, a multiplex was attempted with the same gel conditions and same annealing temperature gradient (54°-62°C). Preferential amplification of the mtDNA CA repeat was again seen in several lanes (lanes 2-8), with the competition for components of the master mix balancing out at 60°C (lane 9). Figure 13 depicts the multiplex results.

Figure 13: *Phase II* Multiplex in 2.5% Agarose (1X SB)
w/Annealing Temperature Gradient 54°-62°C

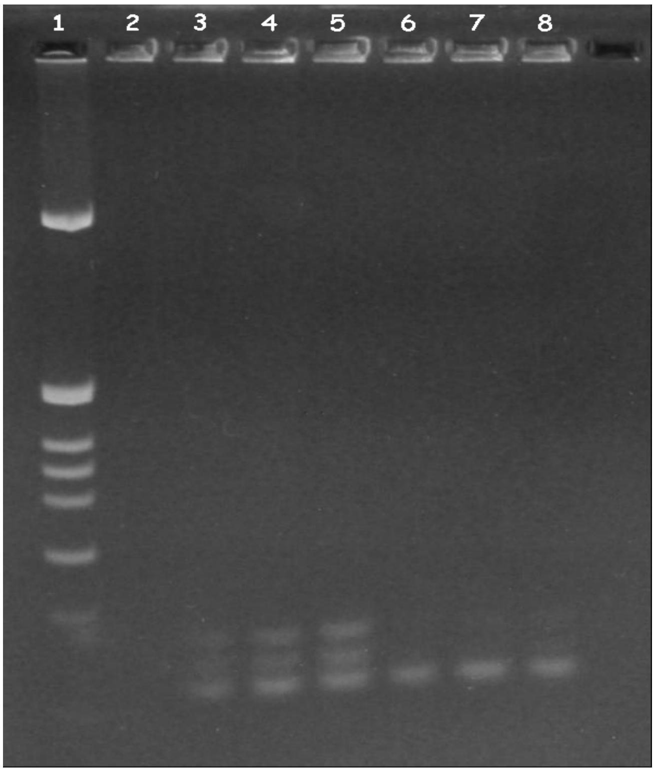


	Multiplex
Lane 1	pBR322/HinfI DNA Ladder
Lane 2	Annealing Temp. 54°C
Lane 3	Annealing Temp. 54.2°C
Lane 4	Annealing Temp. 54.7°C
Lane 5	Annealing Temp. 55.3°C
Lane 6	Annealing Temp. 56.3°C
Lane 7	Annealing Temp. 57.5°C
Lane 8	Annealing Temp. 58.8°C
Lane 9	Annealing Temp. 60°C
Lane 10	Annealing Temp. 60.8°C
Lane 11	Annealing Temp. 61.4°C
Lane 12	Annealing Temp. 61.9°C
Lane 13	Annealing Temp. 62°C

All three loci amplified successfully throughout the entire annealing temperature range. The darkest bands (and hence the most DNA) were found in lane 9 (annealing temperature 60°C). However, the bands in lanes 10-13 are also close in intensity and hence indicate a more optimum protocol. After completion of this multiplex the dNTP supplies were exhausted, so new dNTPs (New England BioLabs® Inc.) were ordered so that the protocol and results could be replicated. Figure 14 represents the first attempt to replicate the multiplex results with other

DNA samples and with the new dNTPs. The DNA used for the previous multiplex attempts was the investigator's female buccal swab DNA. The DNA used for the gel in Figure 14 includes male buccal swab DNA and female DNA from known cell line K562 (GenePrint™, Madison, WI).

Figure 14: *Phase II* Multiplex Replication Attempt with Annealing Temperatures 60°C, 60.8°C, and 61.4°C



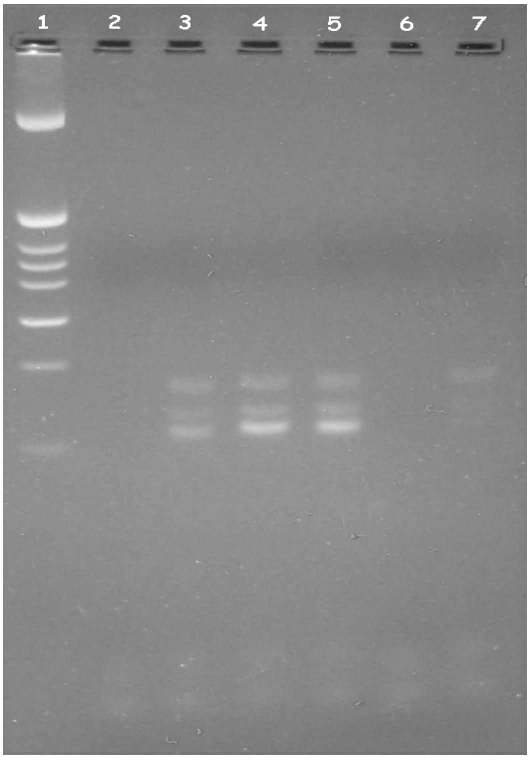
Lane 1	pBR322/HinfI DNA Ladder
Lane 2	Negative Control
Lane 3	Annealing Temp. 60°C (Male DNA)
Lane 4	Annealing Temp. 60.8°C (Male DNA)
Lane5	Annealing Temp. 61.4°C (Male DNA)
Lane 6	Annealing Temp. 60°C (K562 Female DNA)
Lane 7	Annealing Temp. 60.8°C (K562 Female DNA)
Lane 8	Annealing Temp. 61.4°C (K562 Female DNA)

The ladder in lane 1 and negative control in lane 2 performed as expected. The bands in lanes 3-5 represent male buccal swab DNA amplified at annealing temperatures of 60°C, 60.8°C, and 61.4°C, respectively. Lanes 6-8 represent PCR product from amplification of K562 (female) DNA at annealing temperatures of 60°C, 60.8°C, and 61.4°C, respectively. The mtDNA CA repeat amplified successfully across all lanes and preferentially to the other two loci. The three bands present in lanes 3-5 are weaker and not as bright as the bands in Figure 13. The bands for

D3S1358 and amelogenin in lanes 6-8 are only faintly visible, possibly due to inadequate amounts of BSA in the amplification master mix. For the K562 DNA master mix, only half the required amount of BSA was left to add to the PCR tube (1.25 μL per reaction as opposed to 2.5 μL per reaction). Also, since new dNTPs (from a different manufacturer) were being used for this replication, concentrations needed to be checked for accuracy. More BSA was ordered (New England BioLabs™ Inc.) and a test gel was run as a quality control test for both the new BSA and new dNTPs (as shown in Figure 15 on next page).

In Figure 15, the DNA in Lanes 2-3 was amplified using the newly ordered New England BioLabs dNTPs and BSA. Specifically, 1 μL dNTPs (25 mM each)/0.25 μL BSA (10 mg/ml) and 0.25 μL dNTPs (25mM each)/0.25 μL BSA (10 mg/ml) were used, respectively. Lanes 4-5 tested the previously used dNTPs, using 1 μL dNTPs (25 mM)/0.25 μL BSA (1.6 mg/ml) and 1 μL dNTPs (25 mM)/2.5 μL BSA (1.6 mg/ml), respectively. Lanes 6-7 tested use of the newly ordered New England BioLabs™ dNTPs with the previously used BSA concentration [1 μL new dNTPs (25 mM each)/2.5 μL
BSA (1.6 mg/ml) and 0.25 μL new dNTPs (25 mM each)/2.5 μL BSA (1.6 mg/ml)].

Figure 15: Quality Control Permutations of dNTPs and BSA in the *Phase II* Multiplex at Annealing Temperature 60°C



Lane 1: pBR322/HinfI DNA Ladder
Lane 2: 1 μL dNTPs (25 mM each)
(NEB)
0.25 μL BSA (10 mg/ml)
Lane 3: 0.25 μL dNTPs (25mM each)
(NEB)
0.25 μL BSA (10 mg/ml)
Lane 4: 1 μL dNTPs (25 mM) (ABI)
0.25 μL BSA (1.6 mg/ml)

Lane 5: 1 μL dNTPs (25 mM) (ABI)
2.5 μL BSA (1.6 mg/ml)
Lane 6: 1 μL dNTPs (25 mM each) (NEB)
2.5 μL BSA (1.6 mg/ml)
Lane 7: 0.25 μL dNTPs (25 mM each) (NEB)
2.5 μL BSA (1.6 mg/ml)

It was determined from these permutations that the reactions run on the gel in Figure 14 used a four-fold excess of dNTPs due to a mistake in preparing the working dNTP dilution from the newly ordered tube from New England BioLabs™. The four-fold excess of dNTPs was exceeding the capacity of the PCR buffer and dropping the pH of the reaction mixture to unacceptable levels. The 10X PCR buffer is made with the expectation that 200 μM will be the final concentration of the dNTPs. Extra BSA used in some of these permutations helped overcome the excess of dNTPs (hence the appearance of three bands in some of the lanes). All of the reactions with 250 μM final concentration of dNTPs worked, although the New England BioLabs™ dNTPs performed better with more BSA.

New reaction mixtures were prepared using 0.2 μL dNTPs (25 mM each)(final concentration 200 μM) and 2.5 μL of 1.6 mg/ml BSA per 25 μL reaction volume. The results of the first successful replication of the multiplex are shown in Figure 16 (below).

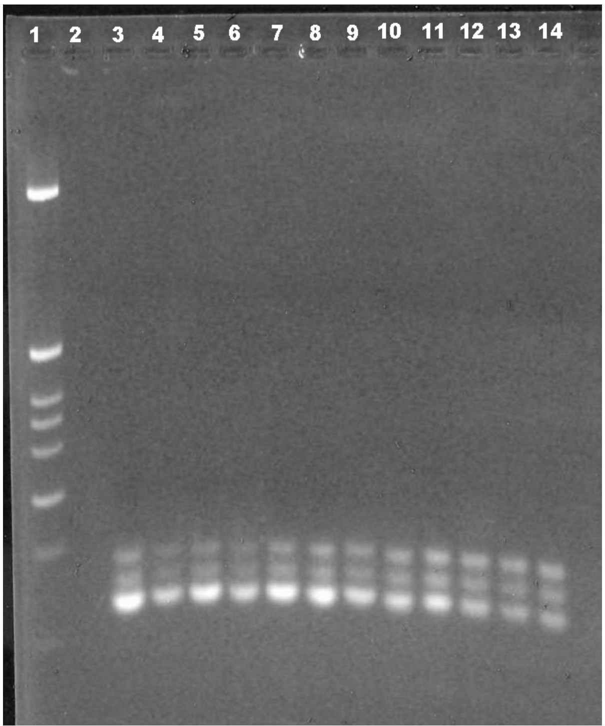
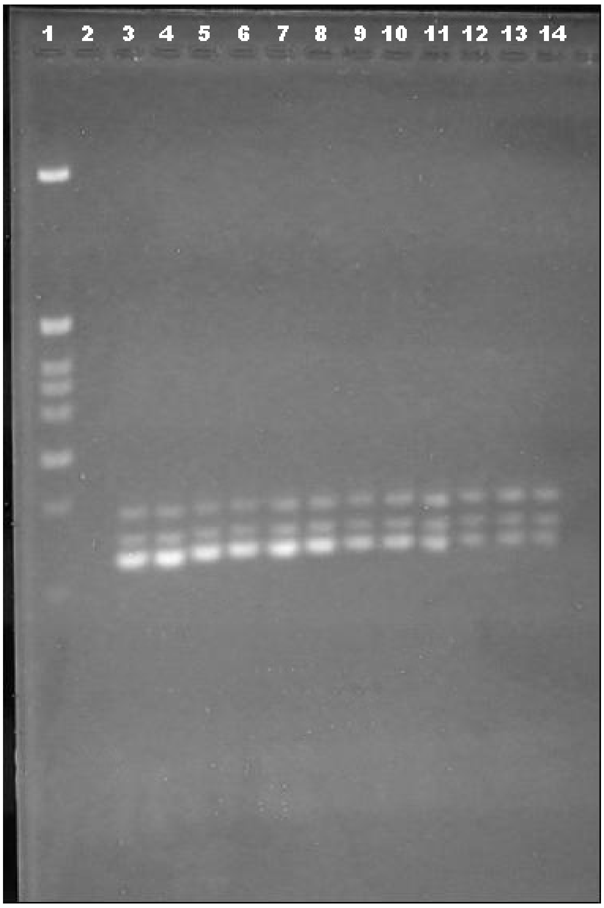


Figure 16: First Successful *Phase II* Replication of Multiplex in 2.5% Agarose (1X SB) w/Annealing Temperature Gradient 54°-62°C

Lane 1	pBR322/HinfI DNA Ladder
Lane 2	Negative Control
Lane 3	Annealing Temp. 54°C
Lane 4	Annealing Temp. 54.2°C
Lane 5	Annealing Temp. 54.7°C
Lane 6	Annealing Temp. 55.3°C
Lane 7	Annealing Temp. 56.3°C
Lane 8	Annealing Temp. 57.5°C
Lane 9	Annealing Temp. 58.8°C
Lane 10	Annealing Temp. 60°C
Lane 11	Annealing Temp. 60.8°C
Lane 12	Annealing Temp. 61.4°C
Lane 13	Annealing Temp. 61.9°C
Lane 14	Annealing Temp. 62°C

The ladder in lane 1 and negative control in lane 2 performed as expected. The bands in lanes 3-14 represent PCR products from female buccal swab DNA. All three loci amplified successfully throughout the entire annealing temperature range. The darkest bands (and hence the most DNA) were found in lane 11 (annealing temperature 60.8°C). However, the bands in lanes 12-14 are also close in intensity and therefore represent an optimized protocol. Figure 17 (below) depicts the second successful replication of the *Phase II* multiplex and is identical in protocol to that depicted in Figure 16, except that the source of the DNA in this replication is a male buccal swab.

Figure 17: Second Successful *Phase II* Replication of Multiplex in 2.5% Agarose (1X SB) w/Annealing Temperature Gradient 54°-62°C



Lane 1	pBR322/HinfI DNA Ladder
Lane 2	Negative Control
Lane 3	Annealing Temp. 54°C
Lane 4	Annealing Temp. 54.2°C
Lane 5	Annealing Temp. 54.7°C
Lane 6	Annealing Temp. 55.3°C
Lane 7	Annealing Temp. 56.3°C
Lane 8	Annealing Temp. 57.5°C
Lane 9	Annealing Temp. 58.8°C
Lane 10	Annealing Temp. 60°C
Lane 11	Annealing Temp. 60.8°C
Lane 12	Annealing Temp. 61.4°C
Lane 13	Annealing Temp. 61.9°C
Lane 14	Annealing Temp. 62°C

Although the multiplex results in this study were replicated with various buccal swab DNA samples (male and female), the fact that the original design of this screening tool was intended for use on fragmented and commingled skeletal remains dictates that the same PCR reaction protocols be performed on DNA that has been extracted from human bone. Attempts to replicate this multiplex on skeletal remains will pose additional challenges since DNA obtained from bone is likely to be much more degraded than the fresh buccal swab DNA used in this research.

Additional analyses need to be conducted on the PCR amplification products generated during this study to verify that the bands visualized on the agarose gels actually represent alleles from the three target loci (amelogenin, D3S1358, and CA repeat). Although the agarose gel electrophoresis used in this research was useful for initial post-amplification confirmation of PCR product, further development of this screening tool would greatly benefit from the high-resolution capabilities of polyacrylamide gel electrophoresis (PAGE). Due to their smaller pore size, polyacrylamide gels resolve smaller DNA fragments better than agarose gels and can even separate fragments differing by a single base pair. The high-resolution capability of PAGE is important in regards to the design of this multiplex, since the amplicon sizes of the three loci chosen are very similar and may even overlap in some samples. In addition to transitioning to a different separation technique (i.e. a higher-resolution gel), detection methods could be modified as well. Aside from the hazards of using radioactive detection methods, research has shown that sensitivity is approximately 100 times higher with silver staining than with ethidium bromide staining (Merril *et al.* 1998).

Capillary electrophoresis (CE) is yet another separation method that would prove useful in the further development of this multiplex. Since the remains to be analyzed with this screening tool will frequently be limited in both quantity and quality, capillary electrophoresis offers an important advantage in that it consumes only minute quantities of sample during injection and samples can be retested if necessary. Aside from being rapid and automated, fluorescent labeling offers an advantage over other detection methods in its ability to record two or more fluorophores separately using optical filters and a matrix. This multi-color capability enables the components of complex mixtures to be labeled individually and identified separately in the same sample. Hence, due to the similar amplicon sizes of the three loci in this multiplex, primers for each locus should be labeled with a different fluorescent dye so that the alleles for each locus can be correctly distinguished and genotyped. Although the Agilent 2100 Bioanalyzer provides another option for analysis, capillary electrophoresis is a more appropriate choice since it is the more widely used technology in forensic laboratories today.

Another important factor to consider in the development of this screening tool is stutter, which impacts interpretation of DNA profiles (especially in cases where two or more individuals have contributed to a DNA sample). Considering the fact that this screening tool is designed to be used with remains that are fragmented and commingled, mixtures will be prevalent and accurate interpretation of results is critical to ensure proper separation of remains. The mtDNA locus incorporated into this multiplex design is a dinucleotide repeat. Stutter percentage is high (30% or more) for dinucleotide repeats, which will complicate the interpretation of mixtures (Butler 2005).

In addition to the challenges introduced by stutter, interpretation could be further complicated by length heteroplasmy in the D-loop 3’ CA dinucleotide repeat. Heteroplasmy is the presence of more than one mtDNA type in an individual. In 2005, Chung et al. reported on the presence of length heteroplasmy in the Hypervariable Region III CA repeats in Koreans. Interpretation guidelines associated with this screening tool will need to take these two factors into account. Although heteroplasmy can sometimes certainly complicate the interpretation of mtDNA results, its presence at identical sites can improve the probability of a match, as occurred in the positive identification of the Romanov family (Ivanov *et al*. 1996).

Although D3S1358 was originally selected as the STR locus for this multiplex screening tool, it may be prudent to consider incorporating the D5S818 locus into the design (either in combination with the D3S135 marker or in place of it). D5S818 is a smaller marker (amplicon size range 119-178) with a relatively high diversity in the population (15 observed alleles) and, like D3S1358, is more likely to amplify in degraded samples than larger STR loci. Incorporating both of these loci into the multiplex would certainly increase the tool’s discriminating power. The power of discrimination (a measure of how powerful a locus is at individualizing) for both of these loci has been calculated. The reported power of discrimination (PD) of the D3S1358 locus is 0.903 for African-Americans, 0.920 for Caucasians, and 0.880 for Hispanics. Using the D5S818 marker, the powers of discrimination for African-

Americans, Caucasians, and Hispanics are 0.879, 0.834, and 0.880, respectively (Budowle et al. 1999).

Ultimately, there are limitations to the application and use of this screening tool. Due to the environmental conditions to which mass grave and mass disaster victims are often exposed, many samples may be limited in quantity or quality. Since positive identification of the remains is the ultimate goal --- and given that the success of DNA typing relies on isolation of DNA of sufficient quality, quantity, and purity --- the condition of the remains upon discovery must be considered and it may not be prudent to utilize this screening tool with all samples.

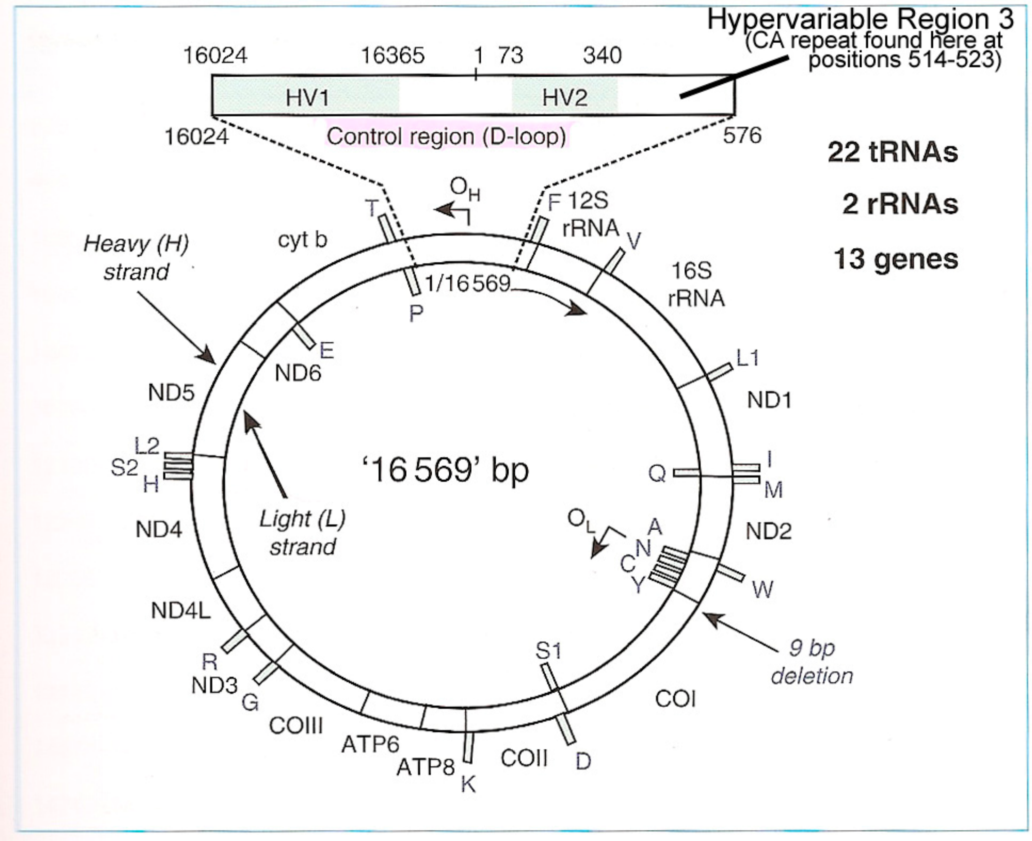
Additionally, if many of the victims found in mass graves or mass disasters are relatives, the usefulness of this screening tool will be limited and such considerations should be taken into account when assessing number of victims. Most notably, maternal inheritance of the mitochondrial genome is important here, as children will have the same mitochondrial DNA sequence as their mother, grandmother, and so on (and hence would carry the same 3’ CA dinucleotide repeat in the D-loop). Similarly, screened samples could potentially contain the same allele at the D3S1358 locus if the remains of parents and their biological children are commingled at the site.

Lastly, if sufficient published population data for the D3S1358 locus and the 3’ CA repeat in the mitochondrial D-loop is not available for the region in which the remains are located, the discriminatory power of the screening tool could not be determined.

FIGURE 1:

FIGURE 2:

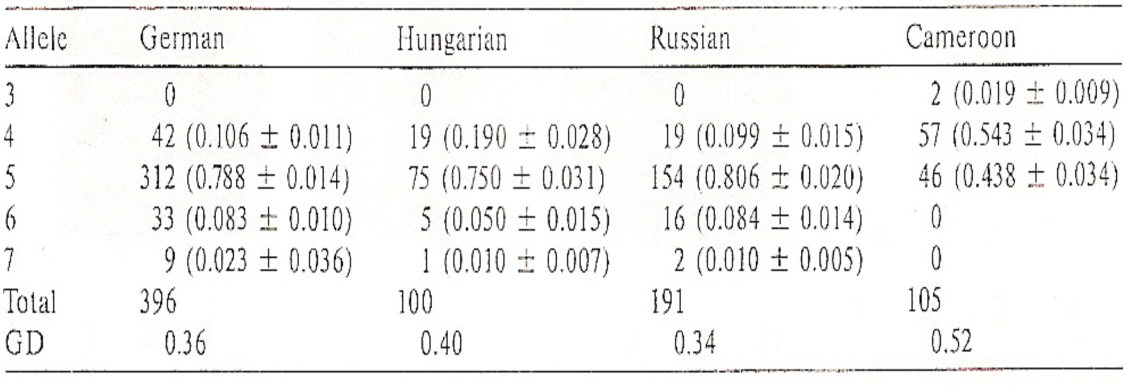
Schematic showing the location of the 3' CA dinucleotide repeat in Hypervariable Region III of the D-loop in the mitochondrial DNA genome



SOURCE: Butler, J.M. (2005). *Forensic DNA Typing: Biology, Technology, and Genetics of STR Markers (2nd Edition)* (p. 243). Burlington, MA: Elsevier Academic Press.

TABLE 3:

Frequency ± standard error of the various alleles of the D-loop 3' $(CA)_n$ dinucleotide repeat polymorphism in three European and one African population



Allele	German	Hungarian	Russian	Cameroon
3	0	0	0	2 (0.019 ± 0.009)
4	42 (0.106 ± 0.011)	19 (0.190 ± 0.028)	19 (0.099 ± 0.015)	57 (0.543 ± 0.034)
5	312 (0.788 ± 0.014)	75 (0.750 ± 0.031)	154 (0.806 ± 0.020)	46 (0.438 ± 0.034)
6	33 (0.083 ± 0.010)	5 (0.050 ± 0.015)	16 (0.084 ± 0.014)	0
7	9 (0.023 ± 0.036)	1 (0.010 ± 0.007)	2 (0.010 ± 0.005)	0
Total	396	100	191	105
GD	0.36	0.40	0.34	0.52

SOURCE: Szibor *et al.* (1997) Mitochondrial D-loop 3' $(CA)_n$ repeat polymorphism: Optimization of analysis and population data. *Electrophoresis* 18:2857-2860

TABLE 4:

Allelic distribution of the D-loop 3' $(CA)_n$ dinucleotide repeat in unrelated individuals from populations in the Gifu Prefecture (Japan), Korea, and Bologna (Italy)

Allele	Gifu Prefecture (Japan)[1]	Korea[2]	Bologna (Italy)[3]	Bologna (Italy)[4]
3	0	0	0	0
4	53	186	8	14
5	92	303	87	126
6	3	7	4	8
7	2	4	1	2
Total	150	500	100	150

SOURCES:

[1] Nagai *et al.* (April 2004) Sequence analysis of mitochondrial DNA HVIII region in a Japanese population. *Progress in Forensic Genetics 10,* Vol. 1261:410-412

[2] Chung *et al.* (2005) Mitochondrial DNA CA dinucleotide repeats in Koreans: the presence of length heteroplasmy. *International Journal of Legal Medicine* 119:50-53.

[3] Bini *et al.* (2003) Different informativeness of the three hypervariable mitochondrial DNA regions

in the population of Bologna (Italy). *Forensic Science International* 135(1):48-52

[4] Bini *et al.* (2003) Population data of mitochondrial DNA region HVIII in 150 individuals from Bologna (Italy). *International Congress Series* 1239:525-528

TABLE 5:

Allele frequencies for the D-loop 3' (CA)$_n$ dinucleotide repeat
in Hypervariable Region III of the human mitochondrial genome

				(N=105)		(N=100)	(N=100)	(N=100)	(N=1
	(N=396)	(N=100)	(N=191)	Cameroon	(N=500)	Uygur ethnic group	Han ethnic group	Bologna	Bolo
ALLELE	**Germany**	**Hungary**	**Russia**	**Africa**	**Korea**	**China**	**China**	**Italy**	**Ita**
3	*	*	*	0.019	*	0.010	*	*	*
4	0.106	0.190	0.099	0.543	0.372	0.200	0.330	0.080	0.0
5	0.788	0.750	0.806	0.438	0.606	0.690	0.670	0.870	0.8
6	0.083	0.050	0.084	*	0.014	0.010	*	0.040	0.0
7	0.023	0.010	0.010	*	0.008	*	*	0.010	0.0
8	*	*	*	*	*	0.090	*	*	*

** = allele not seen in sample population*

SOURCES:

Bini *et al.* (2003) Population data of mitochondrial DNA region HVIII in 150 individuals from Bologna (Italy). *International Congress Series* 1239:525-528.

Bini *et al.* (2003) Different informativeness of the three hypervariable mitochondrial DNA regions in the population of Bologna (Italy). *Forensic Science International* 135(1):48-52.

Chung *et al.* (2005) Mitochondrial DNA CA dinucleotide repeats in Koreans: the presence of length heteroplasmy. *International Journal of Legal Medicine* 119:50-53.

Szibor *et al.* (1997) Mitochondrial D-loop 3' (CA)$_n$ repeat polymorphism: Optimization of analysis and population data. *Electrophoresis* 18:2857-2860.

Tang *et al.* (2003) Allele frequencies of mitochondrial DNA STR locus in two Chinese ethnic groups. *Journal of Forensic Science* 48(2):445-446.

TABLE 7:

Comparison of successfully used thermocycling parameters for amplification of the 3’ $(CA)_n$ dinucleotide repeat in Hypervariable Region III of the mitochondrial D-loop

Cycling condition	Protocol A[1]	Protocol B[2]	Protocol C[3,4]
Initial denaturation	94°C (5 min.)	94°C (3 min.)	94°C (1 min.)
	35 cycles	**35 cycles**	**30 cycles**
Denaturation	94°C (45 sec.)	94°C (45 sec.)	94°C (1 min.)
Annealing	66°C (1 min.)	56°C (30 sec.)	54°C (30 sec.)
Extension	72°C (1 min.)	72°C (1 min.)	72°C (1 min.)
Final extension	72°C (5 min.)	72°C (2 min.)	72°C (7 min.)
PCR Reaction Mix	(Not specified)	**(25 µl rxn volume)** 1X rxn buffer 1.5 mM $MgCl_2$ 200 µM each dNTP 0.4 µM each primer 0.5 U Goldstar polymerase	**(25 µl rxn volume)** 50 mM KCl 10 mM Tris/HCl 1.5 µM $MgCl_2$ 50 µM each dNTP 0.25 µM each primer 1.25 U Taq

SOURCES:

[1] Hoong & Lek (2005) Genetic polymorphisms in mitochondrial DNA hypervariable regions I, II, and III of the Malaysian population. *Asia Pacific Journal of Molecular Biology and Biotechnology* 13(2):79-85.

[2] Szibor *et al.* (1997) Mitochondrial D-loop 3’ $(CA)_n$ repeat polymorphism: Optimization of analysis and population data. *Electrophoresis* 18:2857-2860.

[3] Bini *et al.* (2003) Population data of mitochondrial DNA region HVIII in 150 individuals from Bologna (Italy). *International Congress Series* 1239:525-528.

[4] Bini *et al.* (2003) Different informativeness of the three hypervariable mitochondrial DNA regions in the population of Bologna (Italy). *Forensic Science International* 135(1):48-52.

CHAPTER 3. Higher Throughput Software Analysis

3.1 Introduction to Expert Systems

> ***"Expert systems are computer programs that store human problem solving knowledge, or expertise, and use it to solve difficult problems."*** Dabrowski and Fong in Guide to Expert System Building Tools for Microcomputers (Dabrowski and Fong 1991).

> ***"An expert system is a computer program that reasons, using knowledge, to solve complex problems."*** Feigenbaum in Encyclopedia of Computer Science Third Edition (Ralston and Reilly, Edwin D. 1993).

In the field of forensic DNA analysis, computer automation for data storage and retrieval of convicted offender and forensic DNA data has proven quite useful. Since the FBI's Combined DNA Index System's (CODIS) inception in 1990 to October 2007, over 5 million DNA profiles of individuals convicted of sex offenses (and other violent crimes) with many states now expanding legislation to include other felonies(Anonymous 2007b) have been entered into CODIS including convicted offenders and forensic casework profiles that have aided in over 46,300 investigations and resulted in 45,400 hits (Anonymous 2007a). As advancing computer technology has proven its utility, expert systems are now on the horizon to support the analysis and evaluation of nuclear DNA profiles for convicted offender and forensic DNA data destined for upload into CODIS.

Expert systems will give forensic analysts the ability to rapidly analyze, review, and upload large numbers of convicted offender and forensic nuclear DNA profiles into CODIS. Furthermore, expert systems for mitochondrial DNA analysis could give forensic analysts the ability to rapidly analyze, review, and upload large numbers of missing persons profiles and those of their family members into the

CODIS+Mito database. The CODIS+Mito electronic database is a database of DNA profiles of relatives of missing persons and the second includes profiles of unidentified human remains (Anonymous 2007c). The databases allow local, state, and federal agencies the capability of exchanging and comparing DNA profiles for potential matches. The introduction of higher throughput DNA systems as well as more sophisticated software systems can and has lead to the reduction of samples in backlog for processing and analysis. The number of matches and blind hits clearly points to the importance of databasing and of uploading these profiles into the national databases. Additionally, higher throughput analysis of DNA profiles through expert systems can support the importance of important basic research and discovery work.

With the advances in higher throughput workflow via the implementation of sample card punch instrumentation, robotic pipetting workstations, multicapillary instrumentation, and multiplex single amplification chemistry along with the outsourcing of convicted offender samples, the bottleneck today has shifted from sample processing to the data review required for eventual upload into NDIS for nuclear DNA profilers. Likewise for mtDNA, with the advances in higher throughput workflow via the implementation of robotic extraction for both mtDNA and nDNA as described previously, multicapillary instrumentation, combining the amplification reactions as described previously, and with the production of sequence data with very little background noise, the bottleneck is certainly going to shift from sample processing to data review. Currently, for the submission of convicted offender profiles into NDIS, each sample profile is analyzed by a scientist and its corresponding analytical data is reviewed independently by a second qualified scientist (DNA Advisory Board 2000). The practice of two scientists reviewing each profile prior to CODIS import is tedious and time-consuming. Further, these are the de facto standards that most forensic laboratories use even not analyzing CODIS database samples.

The analysis of sequence DNA data requires considerable training and experience. An expert system, one that interprets data with limited or no human intervention, could prove to be a noteworthy advancement for the forensic DNA community. The incorporation of expert systems into laboratory processes will reduce the time required for data review and expedite the submission of the profiles intoCODIS+Mito or other databases. Currently, expert systems are available for nuclear DNA analysis; they provide quality scoring (or ranking) of analyzed data and are designed such that advanced data analysis and interpretation parameter modifications can be easily customized to laboratory specific guidelines by a technical system administrator. Modifications to the expert system parameters produce an electronic trail of changes. If a similar system could be designed for mtDNA analysis, or sequence analysis in general, it would be quite an advantage for sequencing facilities.

Research in and development of expert systems began in the 1960s and it was in the 1980s that they were first used (Hunt 1986). In the past 30 years, the number and type of expert systems being developed has sustained rapid growth. This growth may be attributed partly to expert system shells, or expert systems building tools, that are used for the purpose of developing expert systems. They are primarily "specialized software systems for automating expert problem-solving for specific types of" scenarios (Dabrowski and Fong 1991).

Expert systems are a subset of artificial intelligence (Association for the Advancement of Artificial Intelligence 2008). Artificial intelligence is the study of theories and methods for automating intelligent behavior. Areas such as business, science, medicine, and engineering have benefited from these artificial intelligence systems (Engelmore and Feigenbaum 1993).

An expert system stores knowledge on how to respond to a particular situation. When an example of the problem is presented, the expert system uses the stored knowledge in its program to find a solution. In summary, the artificial intelligence applications are knowledge-based, meaning that a design engineer has incorporated as much information and knowledge as possible from experts in a particular field. Expert systems are integral at applying the human expert decision-making process; and, they are excellent at automating the human expert decision-making process consistently. By combining computer power with expert system technology, productivity and quality can be improved. An expert system cannot completely replace an analyst, the human expert. Rather, the expert system's knowledge base contains knowledge to solve most commonly encountered problems (Association for the Advancement of Artificial Intelligence 2008). It cannot solve any problem outside its scope of knowledge. The knowledge base is the information stored in the expert system. What is most important is that the expert system must recognize when there is not enough information in its knowledge base to accurately analyze a problem and as a result, a warning, or flag, is triggered bringing the particular problem to the user's attention.

Expert systems differ considerably from non-heuristic every-day computer programs. Algorithms, like those in conventional computer programs, are used along with other intelligent design tools. Expert systems use both factual and heuristic approaches to problem solving. Expert systems don't just calculate, they problem solve via reasoning. As mentioned earlier, to perform as an expert system, the software driving the program must be loaded with a large knowledge base. The knowledge base of expert systems contains the information and reasoning used by human experts. The larger the knowledge base of pertinent information, the more intelligent the expert system. Many rule-based expert systems are programmed with a multitude of *"If…., then…"* statements. These statements follow a chain of reasoning.

The knowledge base can never be complete; and therefore, results in expert systems are weighted, or prioritized. "The set of methods for using uncertain knowledge in combination with uncertain data in the reasoning process is called reasoning with uncertainty (Dabrowski and Fong 1991)." Some form of weighting designed into the expert system would be ideal.

The use of expert systems technology is increasing and the application in industrial and commercial operations is varied. This technology is used in the field of engineering, can assist a physician in making a medical diagnosis, support NASA's space program, manage the inventory for a large factory, and can be used to provide intelligent assistance to a forensic analyst or basic researcher when analyzing genetic profiles and sequence information,. Medical diagnosis was one of the first fields where expert system technology was applied. The most famous example of an early expert system is MYCIN. MYCIN (Shortliffe 1976) is an expert system for diagnosing and recommending treatment of bacterial infections of the blood. It was developed by Shortliffe and Associates at Stanford University. MYCIN is one of several well-known programs and it is one of the earliest documented expert systems; it incorporates artificial intelligence and provides data on the extent to which intelligent behavior can be programmed. There have been considerable advances in medical expert systems for patient diagnosis and care. As a tool to troubleshoot and diagnose, hospitals have integrated expert systems into their hospital information systems (HIS)(Patel and Groen 1991;Barber 1992).

Another use of expert systems that is now available to the everyday consumer is making airline reservations on-line. In order to use a search engine to reserve a flight from city to another, the search engine finds all possible airline companies who have subscribed to the program, determines the routes based on the particular request, and attempts to find the timeframe requested. Once the particular flight plan is chosen, these expert system can be used to determine the class of travel, price, and seat assignments. In this example, the expert system is being used as a scheduling and planning tool.

There are also expert systems that make risk assessments. Automobile insurance companies use expert system software to process information with respect to age, number of speeding tickets, number of accidents, and other germane information that can be used to assess the risk level of insuring a customer. Once the risk is assessed, and the limits of coverage and deductibles are determined, the expert system calculates a price for the insurance policy. Other expert systems that are readily available for the consumer's use on a regular basis are the spelling and grammar warning flags seen when writing a word document or flags and questions when processing personal taxes.

Expert systems have been introduced as computer programs and/or systems that perform at the level of a human expert, and perform it consistently. In the context of expert systems for genetic profiles or sequence data, expert systems should introduce a flag, or an explanation of the reasoning, used to support a decision or conclusion when data fail to meet predefined analysis criteria.

An expert system cannot fully replace the human expert, or DNA analyst, as it relates to data interpretation. The expert systems can be used as a tool to evaluate data and alert the DNA analyst with a flag when results do not meet defined thresholds and rules. However, by defining the rules the expert system uses to initiate the firing of a flag, the human expert is freed from reviewing data that fires no flags, i.e., data that meet the set thresholds. The analyst can focus on those samples that fire a flag, i.e., those samples that did not meet the set thresholds and require human expertise. The final decision-making process is made by the trained DNA analyst who cannot be replaced.

When a particular situation presents itself, for example, a mixed base result, a rule can fire to bring this region of bases (minus ten to plus ten bases) to the analysts' attention. The software will *recognize* that this is called an ambiguous result, or whatever it is defined to be, and fires a rule. Additionally, a frame within a sequence may be meet defined and programmed results. Once again, the software will fire a rule and bring this region of sequence to the analysts' attention. Whereas this example may be over simplified, it illustrates the many data points of information that the expert system must process and the large knowledge base it must have. Add to this example a low Phred score which would fire another rule. With all of this information, expectation would be that the expert system would fire multiple rules. If a sequence fired ten rules, then it may be possible that the analyst would have to review the data for those ten regions of sequence, but would be confident that the other sequence data was good.

The use of expert systems has the potential of expediting accurate analysis and consistent quality review of single source sequence data. Implementing an expert system into a forensic DNA databasing laboratory or even a research facility could significantly increase the speed of data review while ensuring consistency and quality. Most users of the expert systems measure the cost-savings in their implementation; however, I would suggest that equally important is the quality improvement and speed in data analysis. Theoretically, a laboratory using ten analysts to process samples and review convicted offender data could potentially implement an expert system and reorganize the way in which they get the work processed. For example, it may be possible to move nine (9) analysts back into the laboratory to be devoted to processing the samples and one (1) person could be devoted to reviewing the data generated by the other nine (9). The end result is an overall increase in the number of samples sequenced and analyzed. This may not be the ideal reorganization; it is just one example.

3.2 Global Evaluation of Sequence Data Quality

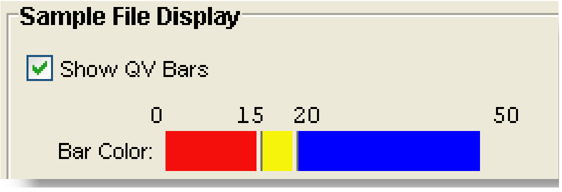
There are several steps in the analysis of sequence data. First, the data are baselined. Then, they run through a basecaller program which calls each of the peaks. After that, a score is defined for each base. Some software programs evaluate the overall quality score of the data. Whenever a quality metric can be assigned to data, a metric with a quantitative result, is can be applied to a software program.

I have evaluated several software programs, both purchased and freeware, and developed some software tools that I believe could be placed in the processing stream for the analysis of sequence data. In this evaluation, I use quality metrics defined for different aspects of the data and combine those metrics to develop my own sequence biometrics. If middleware and glue code were designed to put all of these analysis tools together with the sequence biometric values, an automated sequence analysis system, i.e., a sequence expert system, could be implemented into sequence facilities.

In this section, I will discuss many of the commercially available products and freeware products that have been critically evaluated and their file formats which could be easily integrated into downstream bioinformatic trace processing.

3.2.1 Basecaller Programs

Accurate basecalling is essential in data analysis. KB™ basecaller v1.2, part of DNA Sequencing Analysis Software v5.2 package (Applied Biosystems), is a basecaller program which provides basecalls for short PCR products. It is possible to obtain longer read lengths than previous basecaller programs, more high-quality bases, and increased accuracy at the 5' end with this software. And as the manufacturer claims, increased accuracy is obtained in regions with low signal-to-noise ratios or with anomalous signal artifacts such as spikes or dye blobs. Another feature is the designation of a quality value (QV) associated with each peak. The QV equation is as follows: $QV=-10 \times \log_{10} (P_E)$. P_E denotes the probability that a basecall is erroneous. A basis for data evaluation and defining a threshold for an expert system would be the assignment of a quality value. A QV > 20 is considered acceptable; a QV value can be user-defined. The QV helps determine which bases within a sample file are acceptable. It also helps to determine which bases should be trimmed and reviewed. A QV > 20 is defined that the probability that the base was miscalled is no greater than 1%. Each laboratory should define its own QV. The data generated from a laboratory is affected by its dye chemistry, polymer, run module, protocols, and instrumentation.



Sample File Display

☑ Show QV Bars

Figure 1. Graphic display when defining the QV for your laboratory.

The data in this research were called using KB™ basecaller v1.2. Other basecaller programs may be efficient at basecalling but were not evaluated. This software enables the scientist to basecall and trim data for data analysis and quality control as well as display, edit, and print the data. However, to include these calls in an expert system workflow, only the basecall feature in KB™ basecaller v1.2 is amenable for automation. It was determined that the trim feature in this software would not work in an automated high throughput fashion. The reason that the trim feature could not be used in an automated high throughput fashion is because once the data are trimmed, which is a necessary component in sequence data, the data are still visible in a highlighted format. This software shows that the data have been trimmed, but it is still present. Whereas many scientists may find that it is an extremely important feature to be able to visualize the trimmed data in a highlighted format, it does not make it amenable to automation. It is also extremely important to recognize that the raw data are the raw data; they are not modified in any fashion. Hence, the trimmed data points can always be recovered from the raw .ab1 data files. When it is exported into another program, the highlighted trim data are still exported. This has been a major drawback in developing an automated system since low quality sequence is not part of the overall sequence calls and could adversely affect the final sequence results.

On a positive note, the sequence trimming feature allows for automatic filtering of low quality data. Knowing that these scripts are written and available, it is important to evaluate other trimming programs where quality values can be used to automatically trim the data.

3.2.2 Phred

Phred is a another base calling program for DNA sequence traces. Phred was developed for easy integration into automated data processing by Ewiing et al. (Ewing et al. 1998)at the University of Washington. In fact, Phred was the initial base calling program with quality metrics, and these quality mestrics lead to automation. Phred reads DNA sequence chromatogram files and analyzes the peaks to call bases. Phred assigns Phred scores to each base call. Initially, Phred software had the highest accuracy in base calling. Phred scoring was designed to increase automation. The Phred scores provided information on the consensus sequences which in-turn provided information regarding their accuracy. Additionally, Phred scores could identify those areas that needed additional sequencing to fill in the gaps and provided immediate quality control information immediately following sequencing. The quality metrics were a cornerstone in high throughput sequencing efforts; Phred was initially developed for the Human Genome Project where the quality metrics were just one of its features. Phred also had a graphical user interface (GUI), contig editing features, alignments to a reference sequence, and detection of mutations.

To learn more about how Phred works or about Phred quality values, visit our PHRED page.

3.2.3. Phrap

Phrap is a program also designed for the Human Genome Project; it is a program which enables DNA sequence assembly. For the Human Genome Project, Phred was used to assemble hundreds to thousands sequencing traces very quickly. Using Phred’s quality metrics in Phrap assemblies, more accurate consensus sequences can be built. Phrap uses the quality information of individual sequences to estimate the quality of the consensus sequence. Phrap can also use information such as the sequence chemistry.

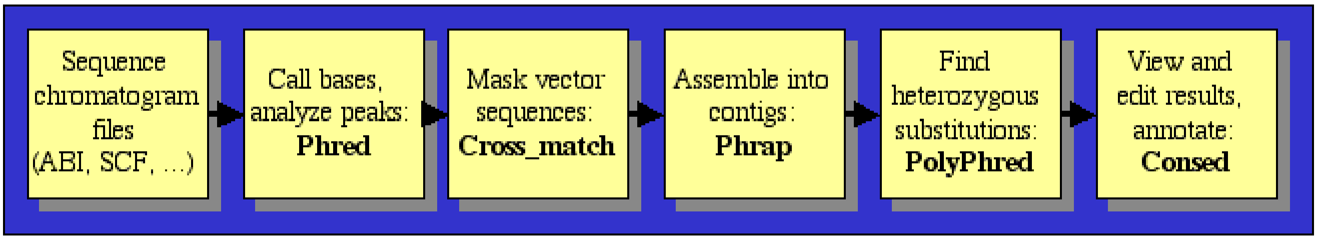
Phrap has been used routinely to assembly bacterial genomes sequenced by the "shotgun" approach, where each project contained tens of thousands of reads. Smaller bacterial genomes (2 million bases or less) could often be assembled in less than three hours.

Phrap uses quality scores to estimate whether discrepancies between two overlapping sequences are more likely to arise from random errors, or from different copies of a repeated sequence. For repeats with 95 to 98% identity (like human Alu sequences) and high quality sequence data, this typically yields correct assemblies (Anonymous).

The combination of Phred and Phrap software and consensus quality metrics can recognize that a sample has two base signals at a specific locus, thus noting a mutation (Rieder et al. 1998).

3.2.4 PolyPhred

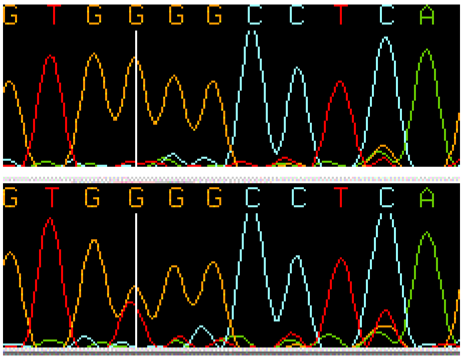
PolyPhred is a program identifying heterozygous single base substitutions in assemblies of DNA sequence traces (Nickerson et al. 1997). PolyPhred identifies putative heterozygous single base substitutions in assembled collections of DNA sequence traces. It is used together with the other programs of the Phred (basecalling)-Phrap (sequence assembly)-Consed (sequence assembly editing) package, as shown in the following diagram:


Sequence chromatogram files (ABI, SCF, ...)
Call bases, analyze peaks: Phred
Mask vector sequences: Cross_match
Assemble into contigs: Phrap
Find heterozygous substitutions: PolyPhred
View and edit results, annotate: Consed

PolyPhred identifies potential heterozygous single-base substitutions by going the all bases in the Phrap-generated contigs, and examining the information about the sequence quality and peaks in each trace. For increased accuracy, PolyPhred ignores the lower quality sequences at the beginning and end of sequence traces. In the regions of sufficiently high quality, PolyPhred looks for the following characteristics that indicate a heterozygous substitution:

- A reduction in relative peak height compare to the other traces
- A secondary peak.

The following example shows a homozygous wild type sequence on top, and trace with a heterozygous point mutation below it:



Note that the second G peak in the sequence at the bottom is only half as high as the peak in the wild type sequence, and that a second red peak indicates that this sequence is a G-T heterozygote.

PolyPhred ranks all putative point mutations it identifies on a scale of 1 to 6, with a 1 indicating highest confidence. It assigns "tags" to all the bases at this point in each sequence, indicating whether a sequence is classified a homozygous or heterozygous sequence at this position, as shown below.

PolyPhred is available for Linux, Unix, and Mac OS X, but not for Windows. For mutation detection on Windows (and OS X), CodonCode Corporation offers CodonCode Aligner. In addition to detecting heterozygous point mutations (SNPs), CodonCode Aligner also can detect and analyze heterozygous insertions and delections.

Screen Shots

The following screen shots shows heterozygous mutations identified by PolyPhred as seen in Consed:

“Often, heterozygous bases are less obvious, because the secondary peak is small or missing; PolyPhred often can correctly identify mutations, too, since it looks for the relative drop in intensity that can typically be seen even if the secondary peak is not obvious.”

Limitations

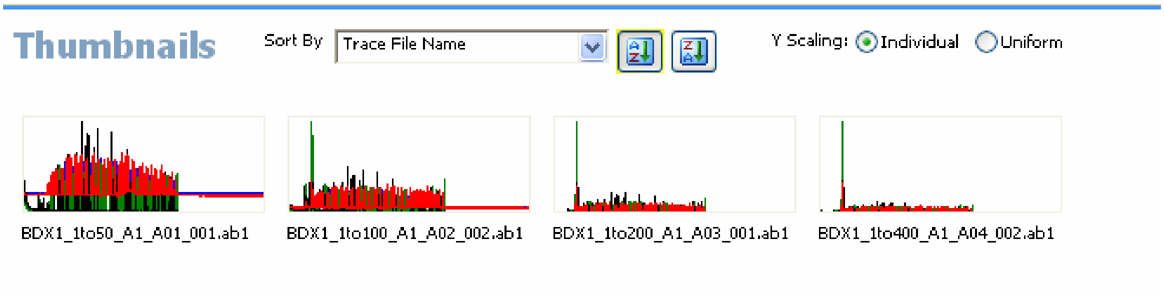
PolyPhred is a research tool, and not perfect. PolyPhred identifications contain both false positive and false negative errors - PolyPhred misses some existing mutations, and falsely flags non-mutated bases as putative substitutions. PolyPhred only identifies heterozygous single-base substitutions, not homozygous substitutions or heterozygous insertions or deletions. PolyPhred requires Phred, Phrap, and Consed,

Another limitation results from PolyPhred's reliance on Phrap-generated assemblies. Phrap sometimes creates more than one contig for a given set of sequences that all cover the same region. This often happens if the sequences contain homozygous mutations - Phrap (which was originally developed for shotgun assembly, not for mutation analysis) thinks that the traces belong to two different copies of a repeat, and therfore puts them into separate contigs.

3.2.5 Sequence Scanner

For the filtering step in the process of evaluating sequence data, Sequence Scanner™ v1.0 Software is used. Sequence Scanner software is a free software program designed to display, edit, trim, export, and generate quality reports for sequencing sample files, specifically those generated from Applied Biosystems instruments. This software reads .ab1 files generated from ABI PRISM 310, 377, 3700 DNA Analyzer and Applied Biosystems 3100-Avant, 3100, 3130/3130xl and 3730/3730xl systems. Sequence Scanner software reads .ab1 files and has multiple export options, e.g., .txt, .seq (only post-trim sequence information), .jpg, .pdf, .fsta (only post-trim sequence information), .phd1. Data processed by the basecaller software can be visualized by Sequence Scanner software. The basecaller takes the raw data obtained from the data collection and converts it to a processed trace, assigning bases for every peak. This processing takes into account smoothing, baselining, and mobility shifts. The KB basecaller allows for quality value bars; if other base calling programs are used, it is possible that no quality values will be assigned. Quality values can be defined by the user. Additionally, the quality values assigned to each peak, i.e., a qualitative value, can be visualized and exported as well. Coordinates of each peak can be obtained by moving the cursor over the processed trace data; however, these data cannot be automatically exported. Exporting of these data has been accomplished in the Translator process. Base calling and auto-trimming features are not supported in Sequence Scanner software.

In Sequence Scanner software, first you import the sample files. Traces can be imported via drag and drop, double-clicking on the file, and the "Import Traces" dialog. After you import the trace files into Sequence Scanner, you can review the data. The first view of your data is in the Details view. You can immediately switch to a Thumbnails view. Thumbnails display raw data and can be scaled uniformly or individually. Traces can be sorted by different categories within both the tabular or Thumbnails formats. With this software, you can quickly scan the data through a Thumbnail view and determine if the data quality is good or bad. Traces can be viewed individually or as a group. One useful tool in Sequence Scanner software is the ability to view each raw image of the sequence traces from an entire plate in a "Thumbnails" view. The thumbnails view allows you to quickly review the run and make informed speedy determinations of data quality.


Thumbnails
Sort By
Trace File Name
Y Scaling:
● Individual
○ Uniform
BDX1_1to50_A1_A01_001.ab1
BDX1_1to100_A1_A02_002.ab1
BDX1_1to200_A1_A03_001.ab1
BDX1_1to400_A1_A04_002.ab1

The thumbnails view seen in this image is a dilution series experiment of a single sample using the BDX procedure presented in this thesis.

Multiple traces can be displayed by tiling individual trace viewers in 2x2, 3x1, 2x1, and 1x2 modes. A Trace Score is assigned to each file. A Trace Score is the average basecall quality value of bases in the post-trim sequence. The Contiguous Read Length (CRL) is the longest uninterrupted stretch of bases with quality higher than a specified limit. In the evaluation of the quality of each base, not only the quality value of that base is used, but also those of adjacent bases within a specified window size. By combining the result of the Trace Score and the Contiguous Read Length, an expert system could use these data to filter the sample files. An additional feature in Sequence Scanner software is the ability of the analyzed and raw data to be displayed simultaneously and the ability to align the raw data with the analyzed data by moving it in either direction by specific key commands. With Sequence Scanner software, you can edit, delete, and insert bases. Additionally, you can undo and/or redo the edits that you have made. One drawback in our application in Sequence

Scanner software is the fact that bases that have been removed after trimming are still available for analysis, i.e., the bases are grayed out.

Sequence Scanner software produces seven reports from the imported sequence trace files: Quality Control, Plate, Trace Score, CRL, CRL Distribution, QV20+, and Signal Strength reports. These reports contain hyperlinks to the trace files and can be viewed in trace viewer. The thresholds for low, medium, and high sequence metrics for Trace Scores and CRL can be changed by the user. Additionally, the user has color control for different sequence metrics.

The metrics used to measure data quality included Q20 clear-range read lengths, Q20 scores, veracity clear-range read lengths, and QV accuracy. The results of this comparative analysis have been published.1

The overall sample score and read length for each file are also determined from the quality value.

Automated Sequence-Trimming The Sequencing Analysis Trimming option, available in the Analysis Protocol of Sequence Analysis Software, automatically removes the low-quality bases that occur at the beginning and end of a sequence and are inherent in sequencing data (Figure 3).

Less Time Required for Screening Data Failures

Sequence Analysis Software saves you and your laboratory staff from having to spend endless hours reviewing data. You no longer have to scan a vast number of sample files to detect failures. Instead, the software’s Analysis QC Report automatically flags failed files (Figure 4). In addition, the per-base QV indicates low-quality regions, thus eliminating the need to scan every base in each sample file.

Longer Read Lengths with High-Quality Base Pairs

KB™ Basecaller increases read lengths on existing instruments and chemistries, which improves assembly success and therefore reduces project costs. Figure 1 results show that regardless of the run module, KB Basecaller provides significantly more accurate basecalling than other options.

Accurate Basecalling of PCR Products

Accurate basecalling of short PCR products is essential for resequencing or mutation detection. Even with PCR products of only 100 base pairs, the KB™ Basecaller is able to call a greater number of bases more accurately at the beginning of a sequence (Figure 5).

Detecting Heterozygous Bases and Assigning Calibrated QV

Laboratories spend many hours manually verifying that heterozygous bases are not missed or miscalled. The complexity of signal characteristics makes correct identification of mixed bases more difficult than that of pure bases, which generally have higher QV. To address the complexity of mixed base signals and to ensure accurate mixed basecalling and correct QV assignment, the KB Basecaller was tested against a large data set (Figure 6). The KB Basecaller simplifies mixed basecalling and is the only solution that provides QV for mixed basecalls calibrated according to the standard relation: Q = -10 x log10(PE).

Designed to Improve Service Laboratory Productivity

Service laboratories face the challenges of providing high-quality sequences to hundreds of researchers who require a short response time. Data submitted by researchers vary significantly in quality, which necessitates hours of troubleshooting. Sequencing Analysis Software provides multiple metrics (sample score, read length, signal-to-noise values) that help pinpoint the cause of data failure and thus improve the productivity of today's high-throughput service laboratory.

Protects Your Data

Sequencing Analysis Software was designed with a set of security features that gives multiple users controlled access to the software and data. These

features include three levels of user access, password protection, and an audit trail that tracks data changes by recording which user edited the data and the corresponding reason for the edits.

Provides Flexible Data Export

Sequencing Analysis Software is designed with flexible data export formats to ensure compatibility with downstream applications and databases. These formats include: .ab1, .seq, .fsta, .scf, and phd1.

Determine the quality of your data using superior metrics from basecalling quality values.

Accelerate quality control using analysis reports with analysis statistics.

Filter out low-quality sequence ends automatically with sequence trimming.

Longer Read Lengths with High-Quality Base Pairs

Our significantly improved basecalling algorithm, the KB™ basecaller now gives you up to 100 more high-quality bases than other basecalling algorithms. You also get longer read lengths with high-quality base pairs, mixed basecalling with quality value, and accurate basecalling of usually difficult-to-sequence short PCR fragments.

Easily Review Sequencing Results with Quality Values

This software enables you to customize and color code the range of the quality values to represent low-, medium-, and high-quality bases. This way, when the basecaller identifies each base and assigns it a quality value, all you have to do is look at the color coding to easily review, discard, or accept it. In addition, the software trims the ends of low-quality bases, grays them out on the user interface for easy identification, and calculates a sample score, which is the average quality value for all the bases in the untrimmed region.

Reduce Data Screening Time

Eliminate manual review of sequencing data batches. With the software's Quality Control (QC) reports, you get read length and sample score (average QV of bases in the clear range) for each sample file, enabling you to sort data by quality. And to make reviewing data even easier, each QC report is hyperlinked back to its source data

Sequencher™ is an alignment program used by the forensic community; this software package allows the user to assemble and analyze sequence data and also has specialized tools for forensic mtDNA profiling.

3.3 Original Analysis of Patterns – Manual

Patterns in sequence data have been a phenomenon that I have wanted to characterize and analyze for many years. I have noted patterns in sequence data in reviewing thousands of sequences using both dichloro-Rhodamine and BigDye™ chemistries. In discussing these data with scientists from the largest manufacturer of sequence chemistry, Applied Biosystems, I was informed that this phenomenon was not reproducible and that the chemistry could not be relied upon to continuously produce such reproducible results. However, Zakeri et al. (Zakeri et al. 1998)established patterns in peak heights with both dichloro-Rhodamine (dRhodamine) and energy transfer dye terminator sequencing using the 377 DNA Sequencer (Applied Biosystems) by measuring the heights and normalizing the data. By characterizing sequence frames, improved accuracy can be obtained in base calling.

To characterize the sequence frames, measurements of peak heights were manually taken. Sequence data was produced from hundreds of samples using dRhodamine, BigDye™ v1.1., and BigDye™ v3.1, all produced on the 3100 Genetic Analyzer (Applied Biosystems). The sequence traces could be evaluated with either KB™ basecaller v1.2 software or ChromasPro. For this study, ChromasPro software was used. Whereas the heights from each software package are relative, they are different. The signal from each software package is different. Both software packages allowed for hovering over the peak to manually document the height associated with that peak. The electropherogram was printed, and the height associated with that peak was documented directly above the corresponding peak. Each frame was tallied with the corresponding height and translated to a particular pattern. Sixty-four (64) sequence frame possibilities (which corresponds to 4 base possibilities, i.e., A, G, T, and C, and three (3) bases per frame, or 4^3) were measured and tallied for 10 traces in HV1 with primer A1.

64 Possible Sequence Frames			
AAA	CAA	GAA	TAA
AAC	CAC	GAC	TAC
AAG	CAG	GAG	TAG

AAT	CAT	GAT	TAT
ACA	CCA	GCA	TCA
ACC	CCC	GCC	TCC
ACG	CCG	GCG	TCG
ACT	CCT	GCT	TCT
AGA	CGA	GGA	TGA
AGC	CGC	GGC	TGC
AGG	CGG	GGG	TGG
AGT	CGT	GGT	TGT
ATA	CTA	GTA	TTA
ATC	CTC	GTC	TTC
ATG	CTG	GTG	TTG
ATT	CTT	GTT	TTT

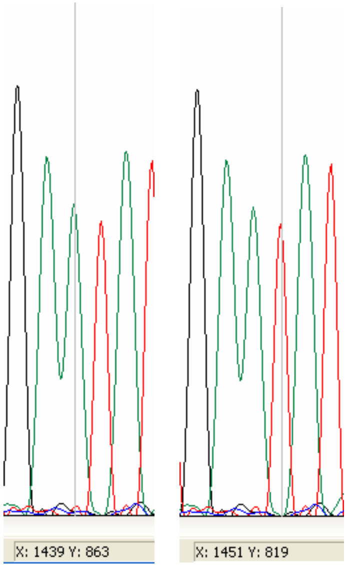


Figure. By placing the cursor at the apex of the peak, the ChromsPro (Technelysium Pty Ltd., Tewantin, Queensland, Australia) software identifies the X and Y coordinates. The X coordinate is the scan data point and the Y coordinate is the relative height of the peak. The Y coordinate was collected with the corresponding base information to analyze each sequence frame.

Initially, it was determined that certain frames were characterized by particular patterns for each of the dye chemistries and that other frames were random. If the frame information were random for each of the patterns, then the expectation would be that each frame would be displayed 1/6 of the time, or 17%. Table XX shows a few examples of the data for 10 sequence traces characterized for several sequence frames and their corresponding patterns. For a sequence frame of AAA with dRhodamine dye chemistry, it is clear that the patterns are not random; in fact, there were no observances of a pattern A, B, C, or D. Interestingly, the BigDye™ Chemistry produced different patterns to dRhodamine chemistry but similar patterns to each other.

Table 1. Percent of patterns for the corresponding frame.

	AAA		
Pattern	dRhodamine	BDT v1.1	BDT v.3.1
A	0%	31%	33%
B	0%	7%	5%
C	0%	31%	29%
D	0%	31%	33%
E	81%	0%	0%
F	19%	0%	0%

	AAC		
Pattern	dRhodamine	BDT v1.1	BDT v.3.1
A	22%	13%	17%
B	0%	17%	20%
C	17%	50%	52%
D	0%	13%	11%
E	52%	7%	0%
F	9%	0%	0%

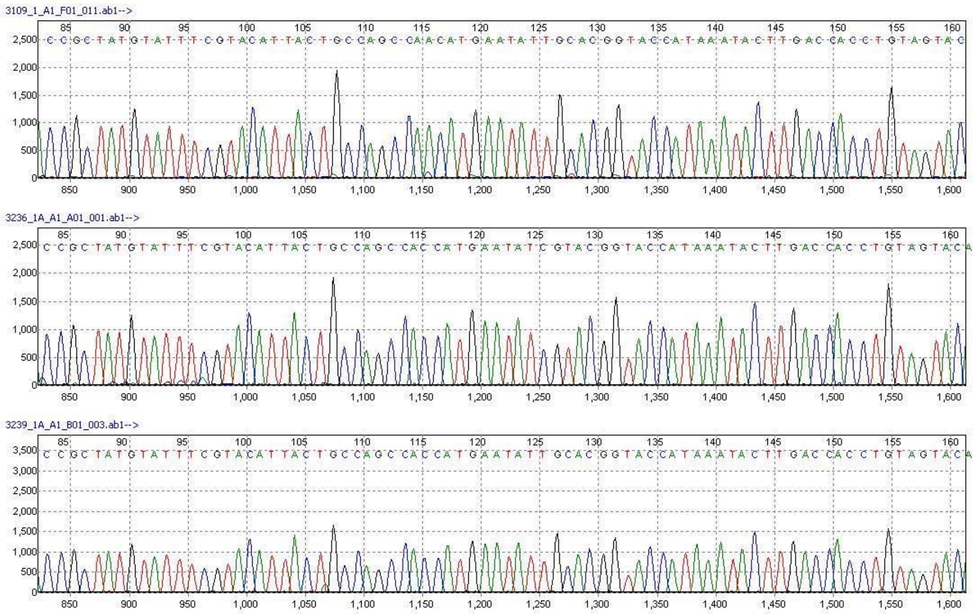
	ACA		
Pattern	dRhodamine	BDT v1.1	BDT v.3.1
A	19%	15%	13%
B	0%	26%	22%
C	58%	26%	26%
D	0%	15%	19%
E	23%	15%	20%
F	0%	3%	0%

	TAC		
Pattern	dRhodamine	BDT v1.1	BDT v.3.1
A	0%	50%	58%
B	0%	6%	8%
C	0%	19%	25%
D	19%	19%	9%
E	11%	0%	0%
F	70%	9%	0%

	TTC		
Pattern	dRhodamine	BDT v1.1	BDT v.3.1
A	40%	11%	9%
B	33%	22%	27%
C	13%	0%	5%
D	13%	67%	59%
E	0%	0%	0%
F	0%	0%	0%

From these preliminary data, it appeared as if some frames demonstrated some significance. Patterns were observed. Patterns were different between the different chemistries and similar between BDT v1.1 and BDT v3.1. It is clear that many of the patterns are not randomly distributed. However, this method of measuring the peaks was laborious, time-consuming, inefficient, and somewhat objective since the method of measurement was hovering over a peak and eye-balling the center of the apex. A method of pattern characterization can support the design of an expert system program. It was determined to be extremely important to develop a more automated process for the output of height data.

I have systematically examined frames of data to determine how base changes influence the peak heights of neighboring bases in sequence traces generated by two commercial dye-terminator chemistries, energy transfer (BigDye) terminators and dichloro-Rhodamine (dRhodamine) terminators.



Sequence data visualized with Mutation Surveyor v3.1 (SoftGenetics LLC, State College, PA) from 3 different samples using the BDX procedure with ABI PRISM® BigDye® Terminator v1.1 Cycle Sequencing Kit, BetterBuffer, and BigDye XTerminator Solution from HV1 products and forward primer A1. Comparable results are also obtained with the reverse primers. Note the similarity of peak signal and patterns for each of the different samples (Roby *et al.*, 2007).

3.4 Patterns in Sequence Data and Output of Data

Reading and exporting trace sequence data

Many software tools are available from the companies that produce sequence chemistry and instrumentation as well as from software companies. There are also many freeware products available over the internet. Obviously, one of the challenges was collecting the data from the output files in order to analyze statistically. The major sequencing companies produce their own file type with many having proprietary source code. With the proprietary source code, research scientists are obliged to purchase the company's software.

Applied Biosystems is, perhaps, the largest major company in the sequencing business. Data produced from sequencers manufactured by Applied Biosystems is stored in a proprietary .ab1 file. These files include the actual raw data produced from the run and append metadata. These metadata include run procedures, dates, settings, basecaller information, etc. Since this kind of files have become so popular, a great effort has been made, throughout the years, to reverse engineer the files and obtain the actual data taken from the instrument. Nowadays, many open source tools have been developed that are capable of reading and exporting data out of AB1 files. Three of those tools are Mutation Surveyor® by Softgenetics, 4Peaks by Mek&Tosj and FinchTV by Geopiza.

All these tools can read and write AB1 files but Mutation Surveyor packs an exclusive feature: it can output all the data taken from the sequencer into a simple, tab-delimited text file. This file produced can then be processed using the methods we will describe later.

This unique feature is of major importance because it allows the manipulation of trace sequence data in ways that were not possible before. It is, of course, of crucial importance for this work.

2.2 – Defining patterns

As it was described in the introduction, the existence of patterns in the heights of DNA sequence trace data is not something unknown, but it was, until today, not well described.

Yet, in order to describe patterns one must first define what patterns are.

As it was described before, for each point were a base call was done, a single dye will be predominant within the four dyes. And the intensity signal at that point, for that dye, will be the height of the peak for that base.

For simplicity sake, frames of triplets were chosen, as they are the ideal candidates for this kind of study. Including three height values allows a descriptive, but not over descriptive analysis of the data. These three peak heights can be are described in Table I:

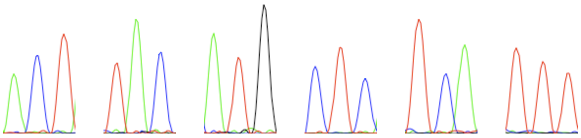
Table I – Height definition for Peaks

Minimum	Medium	Maximum
1	2	3

So, it can be assumed the existence of three different heights for each base call that can be combined in six different ways, named A, B, C, D, E and F as it is explained in Table II:

Table II – Pattern definitions from heights

A	**B**	**C**	**D**	**E**	**F**
1 2 3	1 3 2	2 1 3	2 3 1	3 1 2	3 2 1



Yet sometimes, because of the resolution of the sequencer, peak heights on more than one base call can be identical. In these cases new patterns, called groups, must be defined (Table III).

Table III – Group definitions from heights

Alfa (α)	**Beta** (β)	**Gamma** (γ)	**Phi** (φ)	**Chi** (χ)	**Psi** (ψ)	**Null**
3 1.5 1.5	1.5 3 1.5	1.5 1.5 3	1 2.5 2.5	2.5 1 2.5	2.5 2.5 1	2 2 2

The existence of these groups makes it possible that, what was one type of pattern can become one type of group if the sequencer sensitivity is not enough. These combinations are explicit on Table IV:

Table IV – Possible group combinations by Pattern

	A	**B**	**C**	**D**	**E**	**F**
2 Min	γ	β	γ	β	α	α
2 Max	φ / χ	φ / ψ	φ / χ	φ / ψ	χ / ψ	χ / ψ

But there is still another problem with the sensitivity of the sequencer: Flip-Flopping. If two of the peak heights are too similar, the same sequence, ran twice in the sequencer, can show different values for each of those heights, according to the internal error of the device. If this happens, then it is possible for the two peak heights to exchange places. If one was higher than the other by just a little, it can become lower and hence, produce a different pattern. These are called Minor changes, or flip-flops, and are quite frequent in trace sequence data. Major differences occur when more than two heights in the triplet change. Major differences are much more significant, as they are much less possibly caused by resolution problems as Minor ones.

Table V describes these changes according to each pattern:

Table V - Major/minor changes

	A (123)	**B** (132)	**C** (213)	**D** (231)	**E** (312)	**F** (321)
A (123)	-	Minor	Minor	**Major**	**Major**	**Major**
B (132)	Minor	-	**Major**	Minor	**Major**	**Major**
C (213)	Minor	**Major**	-	**Major**	Minor	**Major**
D (231)	**Major**	Minor	**Major**	-	**Major**	Minor
E (312)	**Major**	**Major**	Minor	**Major**	-	Minor
F (321)	**Major**	**Major**	**Major**	Minor	Minor	-

2.3 – Python and the scripts to evaluate patterns in sequence trace data

Python is a very popular dynamic object-oriented programming language, based in a simple syntax and therefore very simple to learn. Python is also open and platform independent, which makes this particular language widely available, from desktops to mobile devices and web servers. Moreover, given its outstanding capabilities in handling text strings, Python has become a very popular language for bioinformatics. All in all, Python is the best programming language for building applications capable of processing the data used in this work.

Three simple scripts were written to process data out of Mutation Surveyor®:

"Sanger" – The pattern finder script. It does its own base calling, builds triplets and finds the patterns in each of those triplets. This is the core tool. Since its programmed in Python, its easily modifiable according to specific data changes.

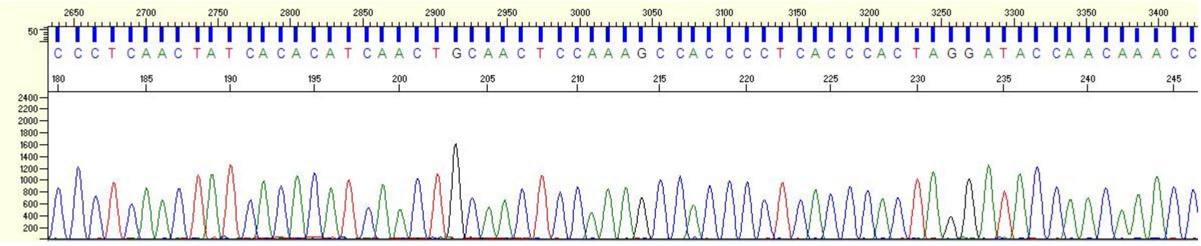
"Cleaner" – An optional script that goes trough the "Sanger" output file and finds repeating mononucleotide triplets, choosing only the first. This cleaning process is important because it allows the elimination of non-independent observations.

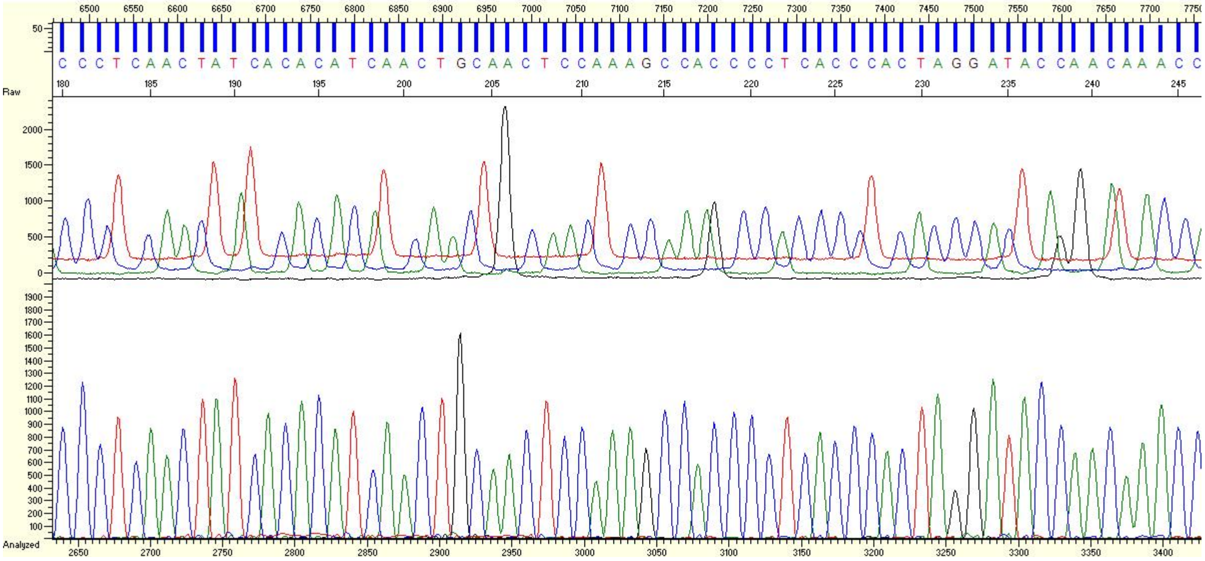
"Reporter" – This is a counting and organizing script. Reporter counts each occurrence of triplets and patterns by scanning and counting the files produced by either "Cleaner" or "Sanger". Also, "Reporter" writes its counts to a text file that can be easily imported into excel for further processing.

Special nomenclature – Some scripts have several versions and therefore they may be named differently. "Cleaner_NI" finds only non-independent mononucleotide repeats.

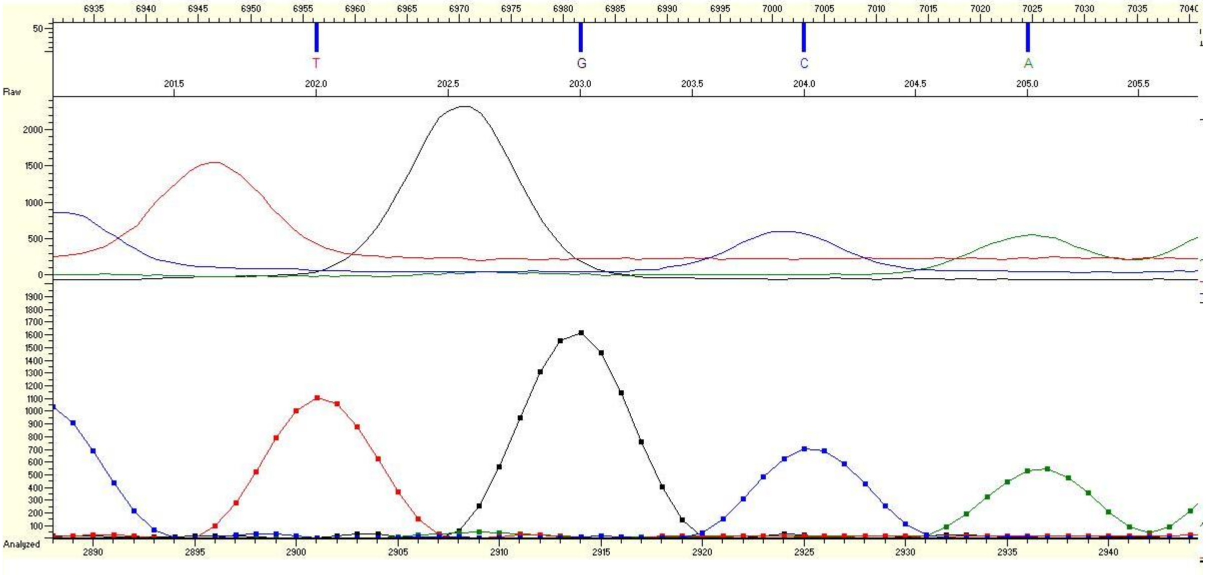
“Reporter_CD” is a version of “Reporter” capable of counting the number of cleaned triplets after the “Cleaner” processing.

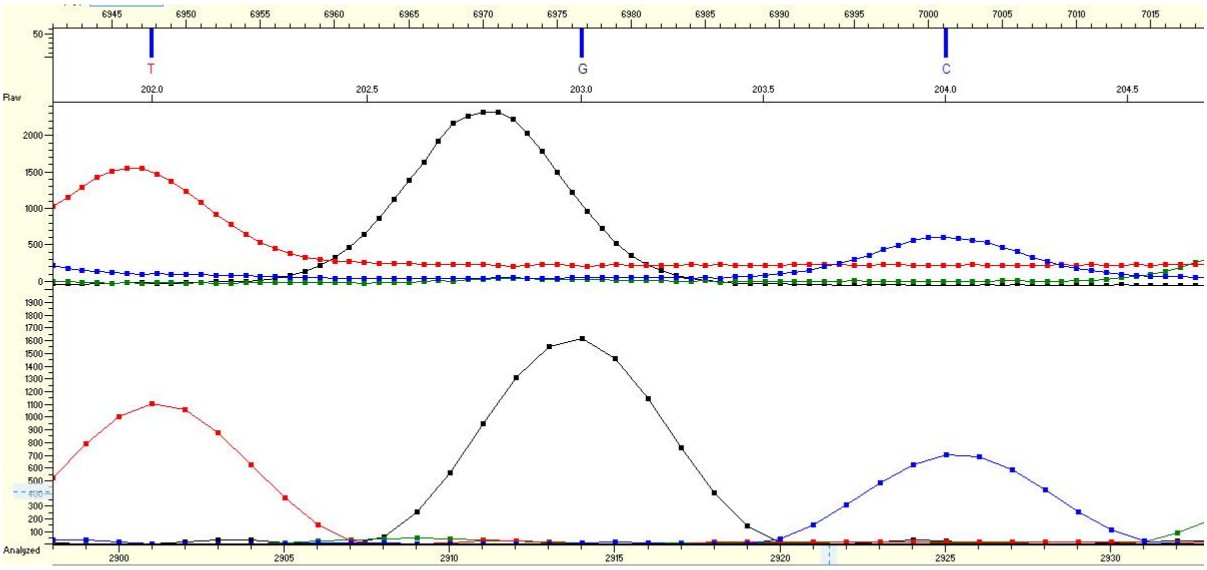
All these scripts are run in a Terminal window by calling the Python interpreter. Moreover, they allow the user to specify the file to process or ask the script to process all the text files in a folder and also allow the user to choose the location and name of the output file.





Raw data peaks are aligned to processed data peaks in the analyzed trace. The Analyzed + Raw view shows both processed and raw trace data side by side. The company discusses how the Analyzed + Raw view is useful for troubleshooting purposes.





Insert MS export features

Exported individual text files

2.4 – Workflow - From the AB1 Trace Data file to Excel®

The workflow followed for each batch of data is as follows (Fig #):

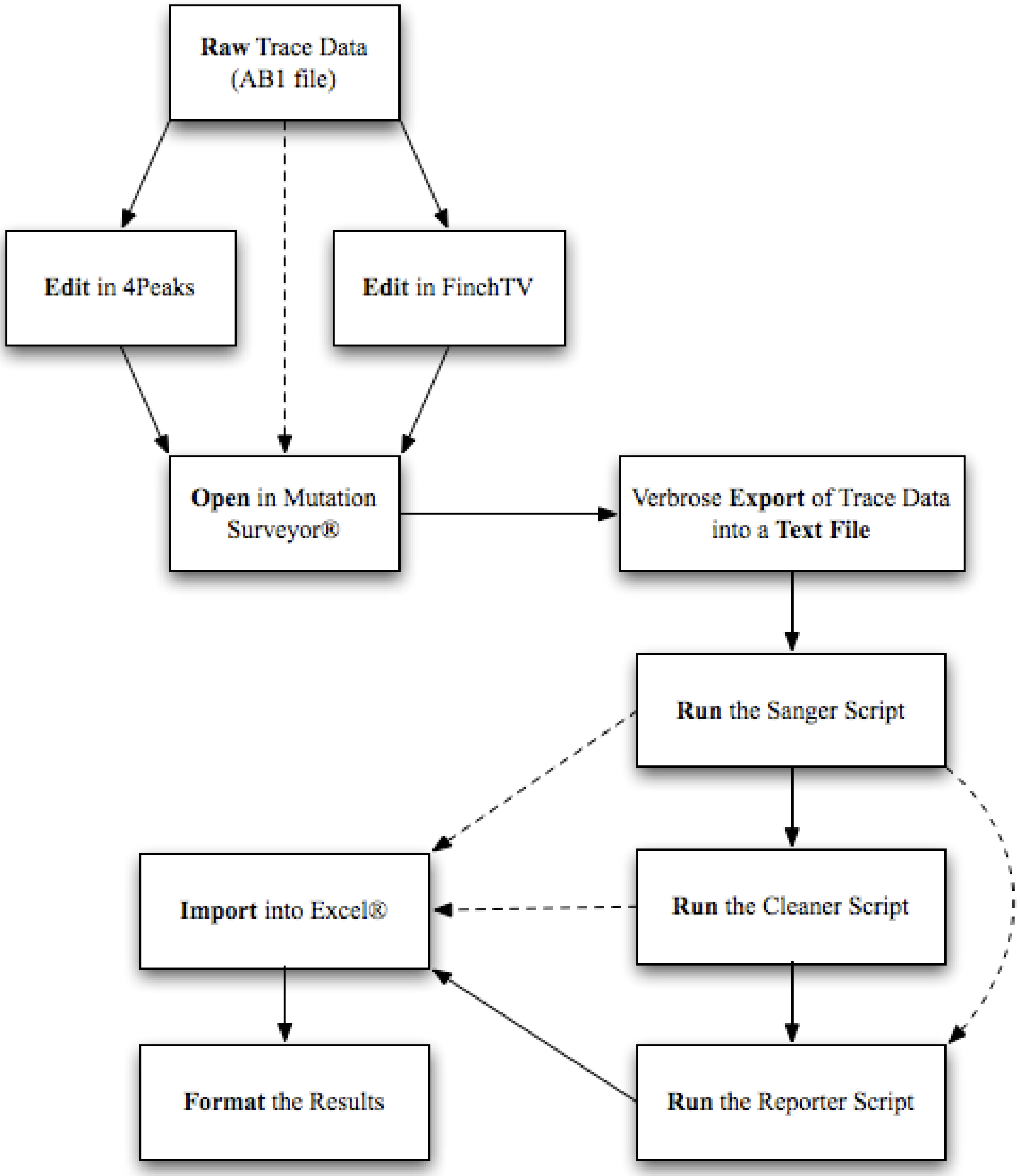
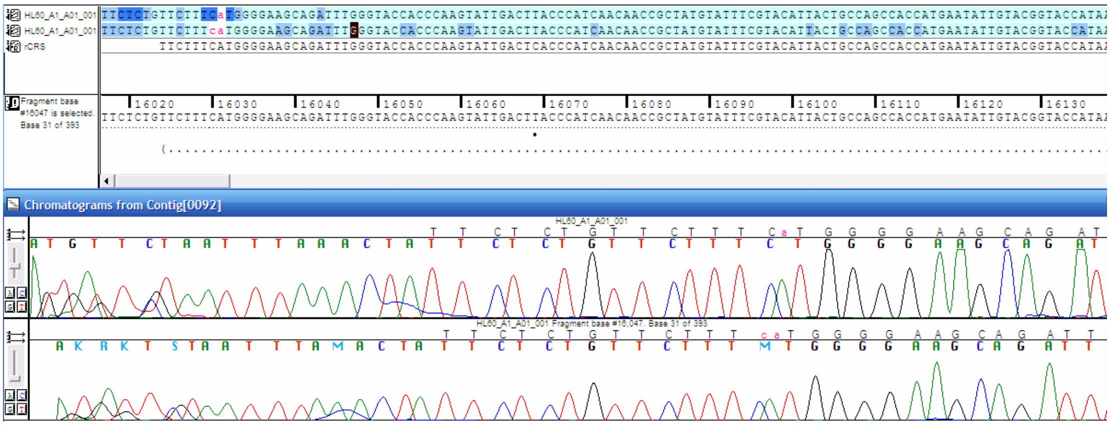


Figure 4 — Data workflow: from the AB1 Trace File to Excel®. Files are opened in either 4Peaks or FinchTV if they need editing, if not they are opened directly in Mutation Surveyor®. Then data is exported into a text file to be processed in the Sanger Script either individually or in a batch. Data is then processed all the way into the Reporter Script and then imported into Excel®. Dashed arrows indicate possible, yet not very common variations of the workflow.

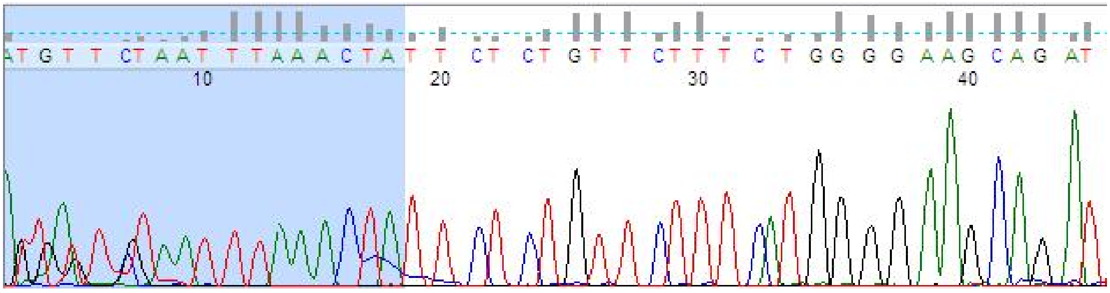
3.1 – The same sample run on the same instrument with all chemistry parameters held constant produce similar sequence patterns. (Patterns are reproducible)



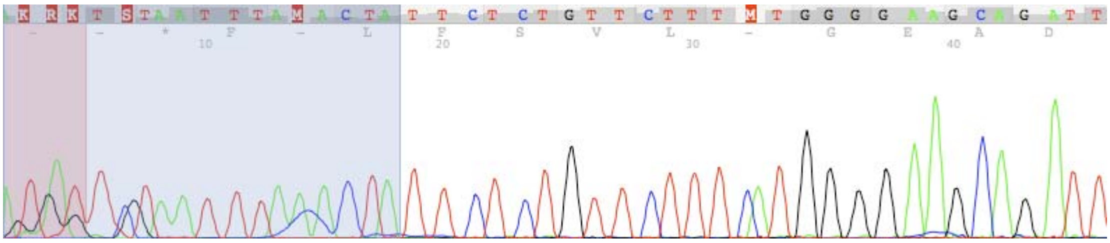
```
HL60_A1_A01_001  TTCTCTGTTCTTTCaTGGGGAAGCAGATTTGGGTACCACCCAAGTATTGACTTACCCATCAACAACCGCTATGTATTTCGTACATTACTGCCAGCCACCATGAATATTGTACGGTACCATAA
HL60_A1_A01_001  TTCTCTGTTCTTTcaTGGGGAAGCAGATTTGGGTACCACCCAAGTATTGACTTACCCATCAACAACCGCTATGTATTTCGTACATTACTGCCAGCCACCATGAATATTGTACGGTACCATAA
rCRS                    TTCTTTCATGGGGAAGCAGATTTGGGTACCACCCAAGTATTGACTCACCCATCAACAACCGCTATGTATTTCGTACATTACTGCCAGCCACCATGAATATTGTACGGTACCATAA

Fragment base     16020     16030     16040     16050     16060     16070     16080     16090     16100     16110     16120     16130
#16047 is selected.
Base 31 of 393   TTCTCTGTTCTTTCATGGGGAAGCAGATTTGGGTACCACCCAAGTATTGACTTACCCATCAACAACCGCTATGTATTTCGTACATTACTGCCAGCCACCATGAATATTGTACGGTACCATAA
```

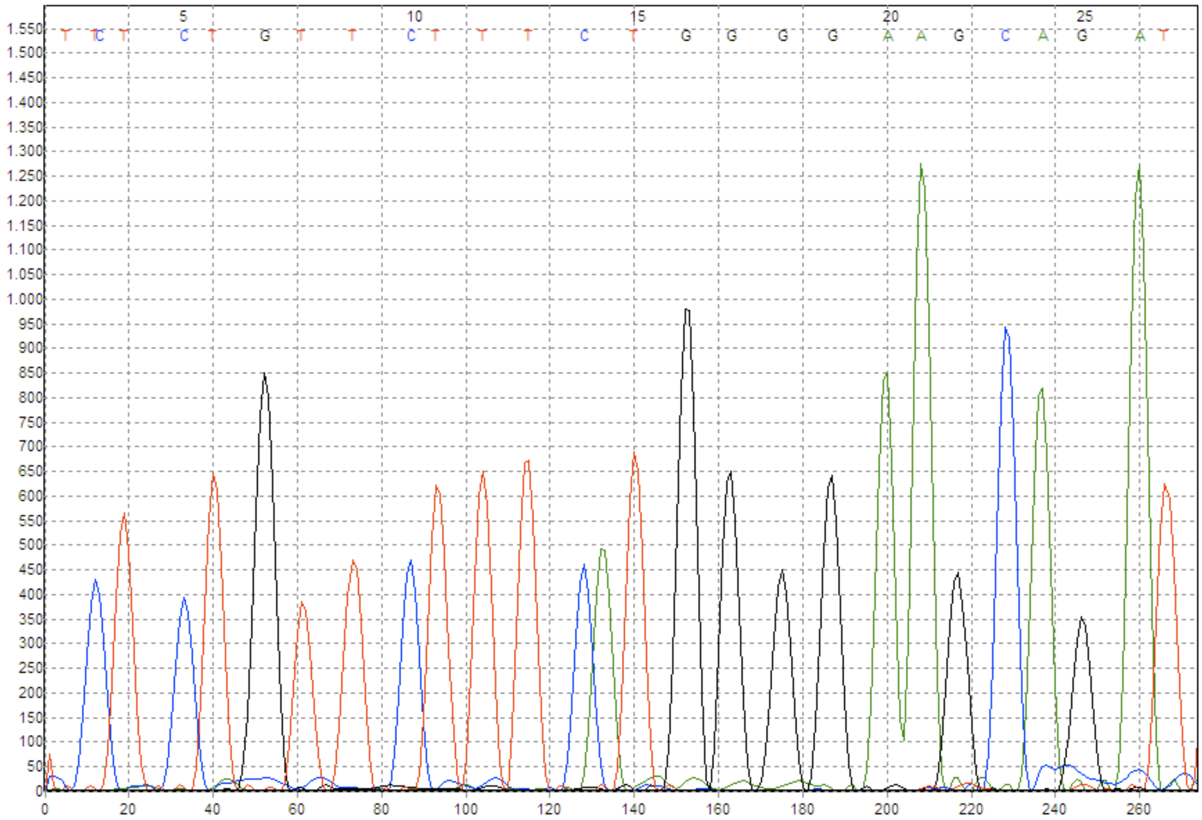
Chromatograms from Contig[0092]

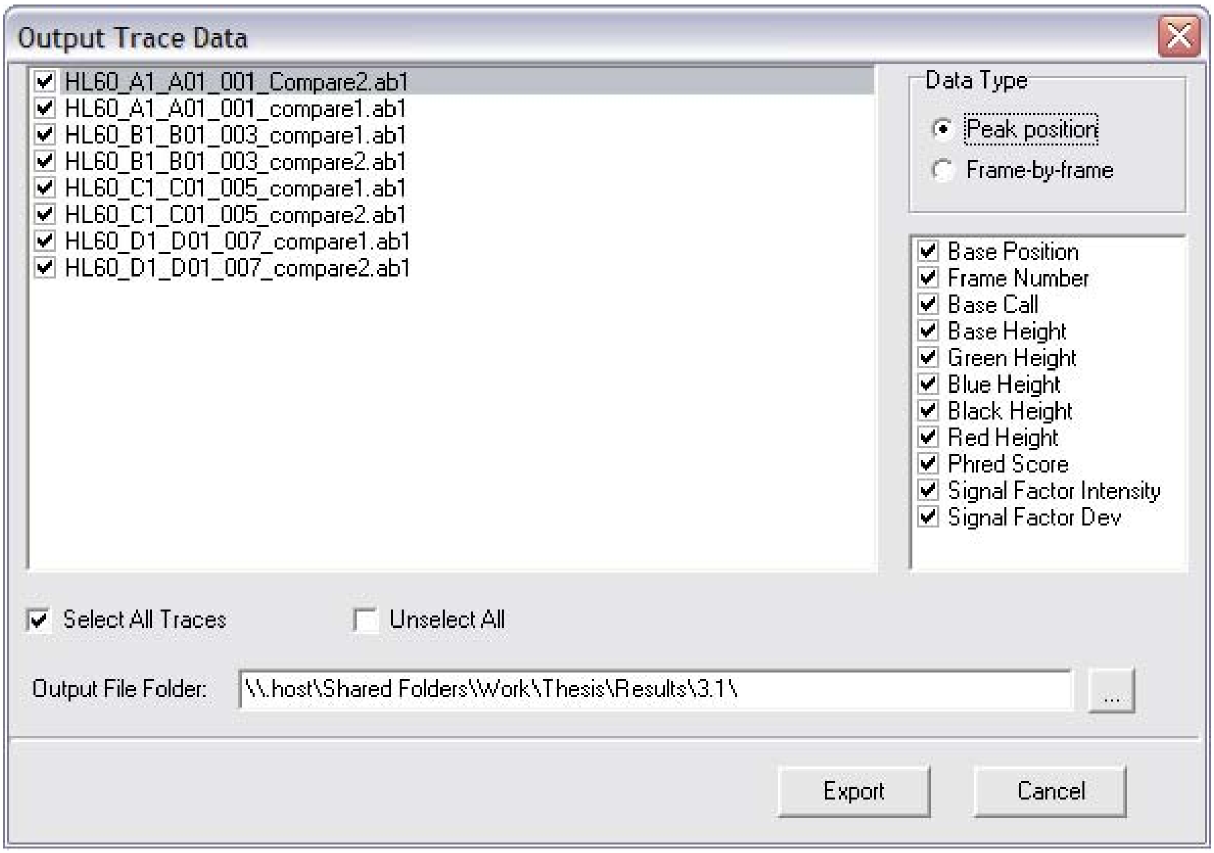


FinchTV



4Peaks





Output Trace Data

☑ HL60_A1_A01_001_Compare2.ab1
☑ HL60_A1_A01_001_compare1.ab1
☑ HL60_B1_B01_003_compare1.ab1
☑ HL60_B1_B01_003_compare2.ab1
☑ HL60_C1_C01_005_compare1.ab1
☑ HL60_C1_C01_005_compare2.ab1
☑ HL60_D1_D01_007_compare1.ab1
☑ HL60_D1_D01_007_compare2.ab1

Data Type

● Peak position
○ Frame-by-frame

☑ Base Position
☑ Frame Number
☑ Base Call
☑ Base Height
☑ Green Height
☑ Blue Height
☑ Black Height
☑ Red Height
☑ Phred Score
☑ Signal Factor Intensity
☑ Signal Factor Dev

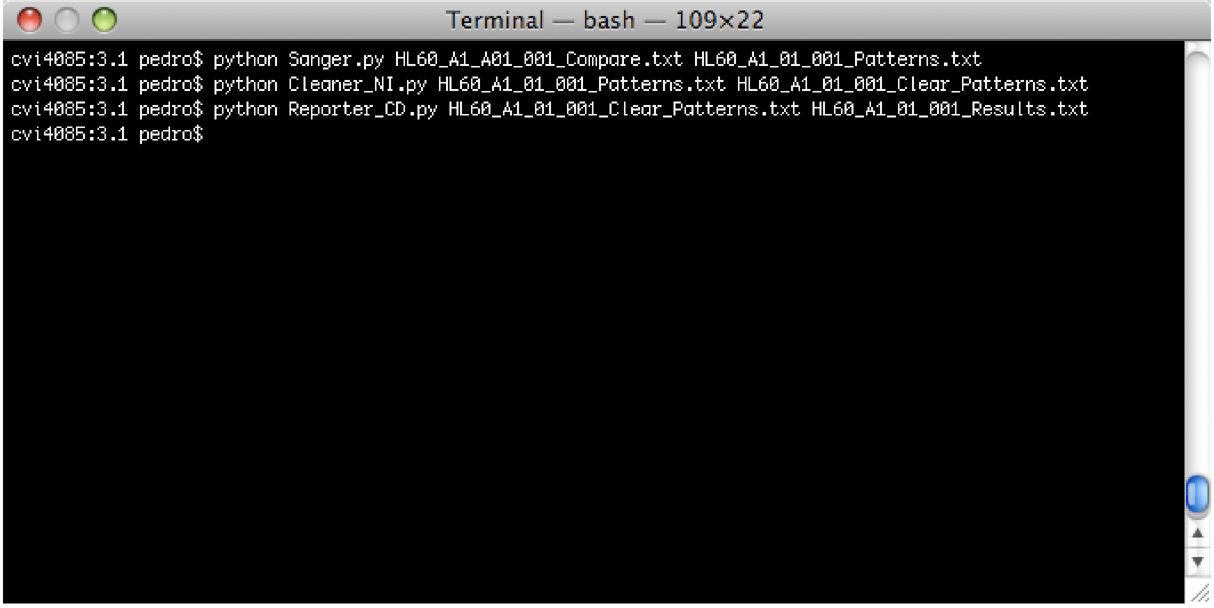
☑ Select All Traces □ Unselect All

Output File Folder: \\.host\Shared Folders\Work\Thesis\Results\3.1\

Mutation suveyor

HL60_A1_A01_001_Compare.ab1

Base Psn.	Frame Number	Base Call	Base Ht.	Green Height	Blue Ht.	Black Height	Red Ht.	Phred Score	Signal Factor Intensity	Signal Factor Dev
0	5	T	11	0	8	0	11	14	0.48	0.00
1	12	T	11	4	429	0	4	18	0.03	0.00
2	13	C	429	4	399	0	0	18	0.96	0.99
3	19	T	566	7	0	0	566	18	0.97	1.00
4	33	C	394	5	394	0	9	18	0.95	0.98
5	40	T	648	7	2	0	648	21	0.93	0.95
6	52	G	849	1	28	849	1	21	0.95	0.99
7	61	T	385	0	7	9	385	21	0.95	0.99
8	73	T	471	2	8	7	471	21	0.96	0.99
9	87	C	470	5	470	8	2	25	0.95	0.98
10	93	T	621	0	0	5	621	23	0.96	0.96
11	104	T	651	0	13	8	651	16	0.92	0.93
12	115	T	673	3	5	2	673	16	0.97	0.99
13	128	C	462	33	462	7	0	13	0.47	0.49
14	140	T	689	7	1	3	689	13	0.95	0.98
15	152	G	979	19	1	979	0	33	0.95	0.97
16	163	G	650	13	1	650	0	33	0.97	0.98
17	175	G	451	11	1	451	0	45	0.95	0.97
18	187	G	640	4	0	640	0	30	0.97	0.98
19	200	A	849	849	0	7	0	30	0.98	0.98
20	208	A	1276	1276	6	5	4	30	0.98	0.99
21	217	G	445	27	0	445	11	30	0.88	0.96
22	228	C	943	14	943	0	3	30	0.97	1.00
23	237	A	820	820	51	2	0	20	0.93	0.96
24	246	G	353	18	35	353	14	20	0.79	0.91
25	260	A	1272	1272	44	8	2	20	0.95	0.97
26	266	T	625	1	5	1	625	20	0.91	0.95



```
cvi4085:3.1 pedro$ python Sanger.py HL60_A1_A01_001_Compare.txt HL60_A1_01_001_Patterns.txt
cvi4085:3.1 pedro$ python Cleaner_NI.py HL60_A1_01_001_Patterns.txt HL60_A1_01_001_Clear_Patterns.txt
cvi4085:3.1 pedro$ python Reporter_CD.py HL60_A1_01_001_Clear_Patterns.txt HL60_A1_01_001_Results.txt
cvi4085:3.1 pedro$
```

Instert output of patterns

Pattern Analysis

Python Sanger script outputs the pattern for each frame.

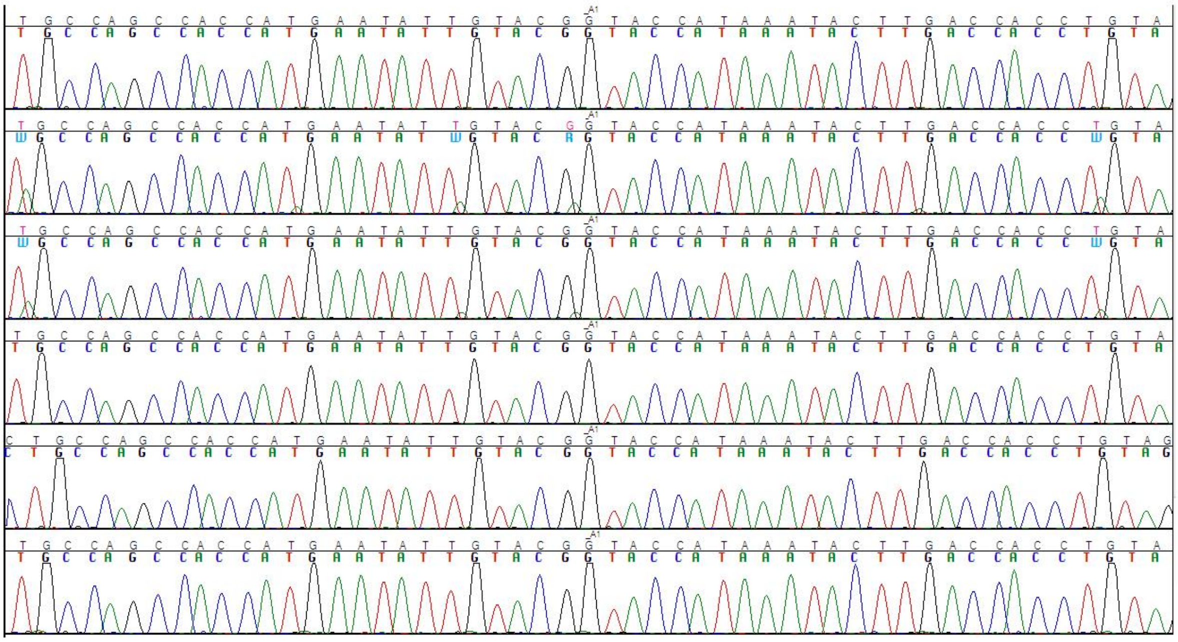
Cleaner scripts cleans them up – after a stretch of 3 bases

Reporter counts the number of patterns for each and the number cleaned.

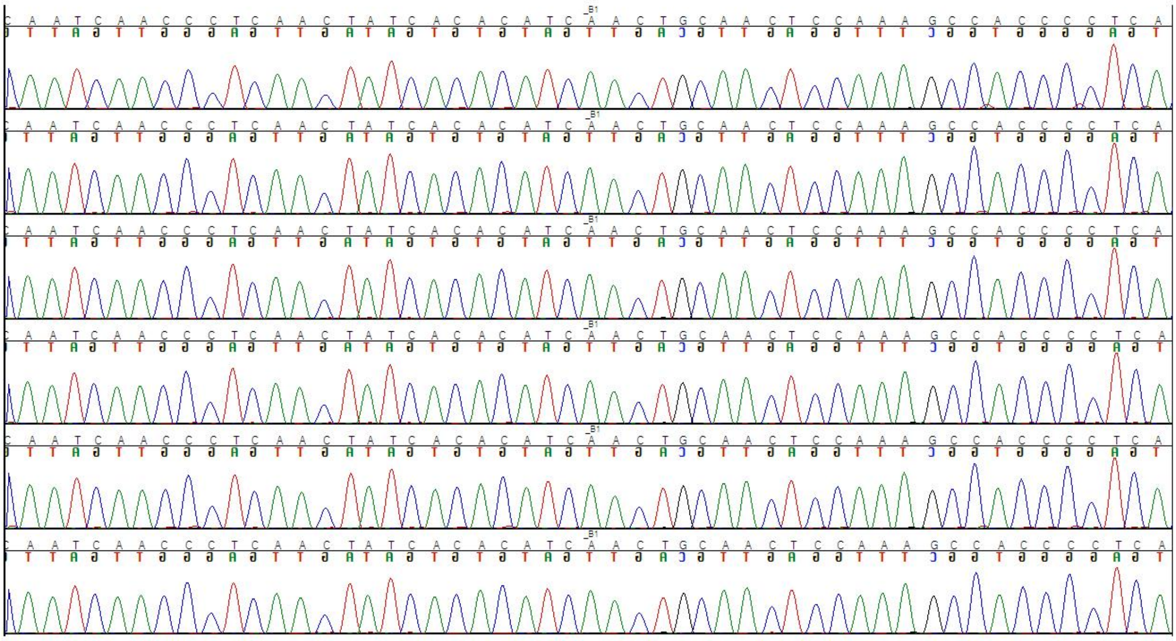
With the same sample sequenced twice on the same instrument, give table.

Picture of major difference – talk about thresholds – this really is not a major difference.

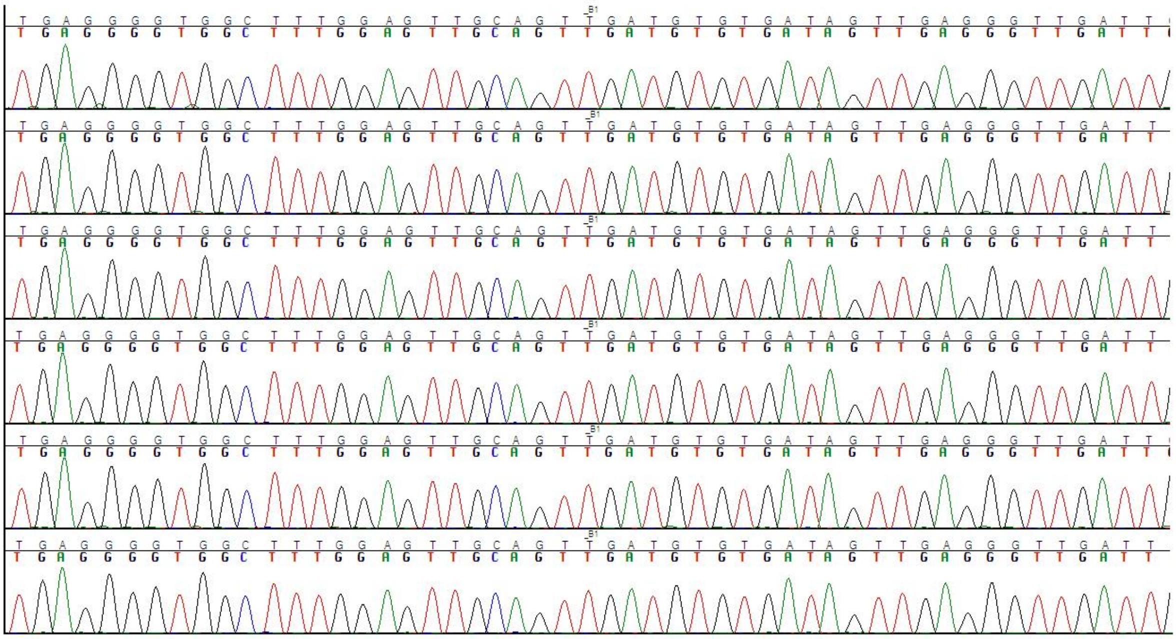
Different samples; same instrument.



This figure displays p16105 to p16155 for the same sample run multiple time on the 3130xl Genetic Analyzer with the A1 primer. Trends and patterns can be seen



p16215 to p16265



p16265 to p16215

Mutation surveyor

The Sanger script evaluates the relative height information.

The first question we need to answer is whether patterns are kept constant if the same sequence is ran several times in the same instrument.

To do so, we took a sample of one individual and distributed it in two wells on the same plate and ran the sequencing procedure. We took these two sequence traces from the same sample, aligned them, ran them trough the processing scripts and compared them. Please refer to Annex I for the raw results. Table I summarizes the results of two different samples that we compared.

Table I – Difference comparison between the same samples ran twice on the same instrument.

Comparison	Total Triplets	Dif. Triplets (minus cleaned)	Dif. Patterns	Same Triplet, dif. Pattern	Major differences
B1	250	0 (0*)	14	14	0 (0**)
C1	357	0 (2*)	12	10	0 (1**)
D1	362	0 (2*)	15	13	0 (0**)

* Different Triplets found at the beginning of the sequence (within the first 10 Triplets).

** Number of major differences preceded by another minor or major difference.

Notice the small number of pattern differences (<5%) and the almost absent number of different triplets. Though one of the comparisons show at least one major difference in their patterns, that difference is preceded by other difference and is not an independent event.

3.2 – Different samples run on the same instrument with all chemistry parameters held constant produce similar sequence patterns. (Similar samples produce similar patterns and can be pooled together)

In the last chapter we saw that patterns are kept constant when we run the same sample, with the same parameters, twice on the same instrument.

Now, the second question we need to answer is whether patterns are kept constant if the samples are different but the parameters are kept the same. That is, we need to know if two different samples behave in a similar way and show similar patterns if the parameters are kept constant and, if that is true, know if we can merge those pattern results.

To do so, first we took two different DNA samples form two different individuals and, keeping all the primer and chemistry parameters constant, we ran the sequencing procedure on the same instrument. We then took these two sequence traces, aligned them, ran them trough the processing scripts and compared them much like we did on chapter 3.1. Please refer to Annex II for the raw results. Table II summarizes the results of our comparisons.

Table II – Difference comparison between two different samples ran on the same instrument.

Comparison	Total Triplets	Dif. Triplets (minus cleaned)	Dif. Patterns	Same Triplet, dif. Pattern	Major differences
A1	148	8 (0*)	35	27	1 (1**)
A2	81	8 (4*)	26	14	2 (0**)
B1	220	18 (1*)	66	47	1 (5**)
C1	241 (194§)	9 (4*)	47	34	1 (3**)
C2	241 (122§)	3 (4*)	36	29	1 (2**)

* Different Triplets found at the beginning of the sequence (within the first 10 Triplets).

** Number of major differences preceded by another minor or major difference.

§ Usable number of triplets (before or after total sequence mismatch). All other values are calculated over this number.

It is clear that, though differences do occur, most of the patterns are kept constant from one sample to the other. Moreover, very few major differences occur (no more than two significant differences in any comparison). This means that most differences are caused by pattern flip-flop. Hence, it is safe to assume that we can pool these sequences together, if we want to study the behavior of patterns using the same chemistry and primers.

3.3 – Samples run on different instruments within the same laboratory with all chemistry parameters held constant produce similar patterns. (The same sample but different instrument produces similar patterns)

Another preliminary question we need to answer is if pattern behavior is dependant on the chemistry and parameters used or the instrument itself.
In order to answer that, we took the same sample and, keeping all the chemistry parameters constant, ran the sequencing procedure on two different instruments. We then took these two sequence traces, aligned them, ran them trough the processing scripts and compared them. Please refer to Annex III for the raw results. Table III summarizes the results of our comparisons.

Table III – Difference comparison between equal samples ran on different instruments.

Comparison	Total Triplets	Dif. Triplets (minus cleaned)	Dif. Patterns	Same Triplet, dif. Pattern	Major differences
A1	394(369§)	0 (0*)	25	25	0 (2**)
B1	350(332§)	0 (0*)	33	33	0 (3**)
C1	90(89§)	0 (0*)	20	20	0 (3**)
C2	120(94§)	0 (0*)	8	8	0 (0**)
D1	334(327§)	0 (2*)	26	24	1 (2**)
D2	172(145§)	0 (0*)	14	14	0 (0**)

* Different Triplets found at the beginning of the sequence (within the first 10 Triplets).
** Number of major differences preceded by another minor or major difference.
§ Usable number of triplets (before or after total sequence mismatch). All other values are calculated over this number.

As it is clear, these results are very similar to those in chapter 3.1. Hence, the same sample ran in two different instruments produce patterns in a similar way as if that sample were to be run in just one instrument, with just one case of a significant major difference in one of the comparisons. Therefore, we cannot only pool together batches of sequences with constant parameters but also batches of sequences from multiple instruments on the same lab.

3.4 – Different samples run on different instruments between different laboratories with all chemistry parameters held constant do not produce similar sequence patterns. (different labs, batches of similar sequences produce similar patterns)

We already know, from chapter 3.1 and 3.3 that the same sample ran in different instruments will produce similar patterns. We also know, form chapter 3.2 that we can pool together different samples if, and only if, they are produced in the same lab using the same the chemistry and keeping all the parameters constant.

Now, do similar sequences produce similar results in two different labs?

To answer this question we pooled together a batch of A1, B1, C1 and D1 sequences from two different labs and compared them.

Since we are pooling together dozens of sequences, we had to compare our results in a different way using a simple chi-square test. Please refer to Annex IV for the raw data. Table IV summarizes the results.

As it is clear, most of the patterns in each triplet cannot be compared from one lab to the other, even using the same chemistry in the two labs. Of notice, A1 results were the most consistent results.

3.5 – Two different samples with different primers (hence different DNA sequences analyzed) but same instrument and chemistry do not exhibit the same patterns. (different primers different patterns)

Again, we can try to answer this question by pooling together similar sequences into batches and then compare different batches.

To do so, we could look at individual samples and compare them, but we will rather look at batches of A1 sequences and compare them to B1, C1 and D1 sequences. Please refer to Annex V for the raw results. Table V summarizes the results of our comparisons.

Since A1 sequences produce different types of patterns, for each frame than B1, C1 or D1 sequences, it becomes clear that different primers will produce different types of patterns for each frame.

3.6 – Same primers, same samples but different chemistry produce different patterns. (different chemistry, different patterns)

At this point we know that we can only reproduce patterns within frames if the lab, chemistry and primer are the same. Now, do different chemistries produce different patterns?

To answer this question we took a batch of sequences from either A1, B1, C1… primers but processed, in the same lab, using different dye chemistries (dRhodamine or BigBye®).

Please refer to Annex VI for the raw results. Table VI summarizes the results of these comparisons.

Again, it is clear that the chemistry used is also a major player in the formation of patterns for each frame.

3.7 – Can sequence patterns be aligned to demonstrate a match or non-match between two samples?

3.8 – Can sequence data produced on the same instrument with different primers be pooled together? (crossing, dependence on primers)

3.9 – Can sequence data produced on different instruments with different primers in different laboratories be pooled together?

2.3 – Excel® as a tool to analyze the statistical significance of patterns

In order to analyze the statistical significance of patterns, special Excel® spreadsheets were developed.

Microsoft™ Excel® is a simple, widespread spreadsheet environment and, even though it is not an application build to deal with heavy statistics, it manages well a simple chi-square test. Moreover spreadsheets in Excel® are easy to program, format and deliver or export to other applications.

Two types of spreadsheets were built to study the significance of peak height patterns. The “6x1” spreadsheet, that performs a chi-square test to a single sample for each frame and analyses the distributions of patterns, deeming them as random or not, and the “6x2” spreadsheet, that also performs a chi-square test for a pair of samples for each frame, calculating the significance of pattern distribution of each frame, comparing the two distributions for each frame and deeming them as comparable or not.

2.4 – Workflow - From the AB1 Trace Data file to Excel®

The workflow followed for each batch of data is as follows (Fig #):

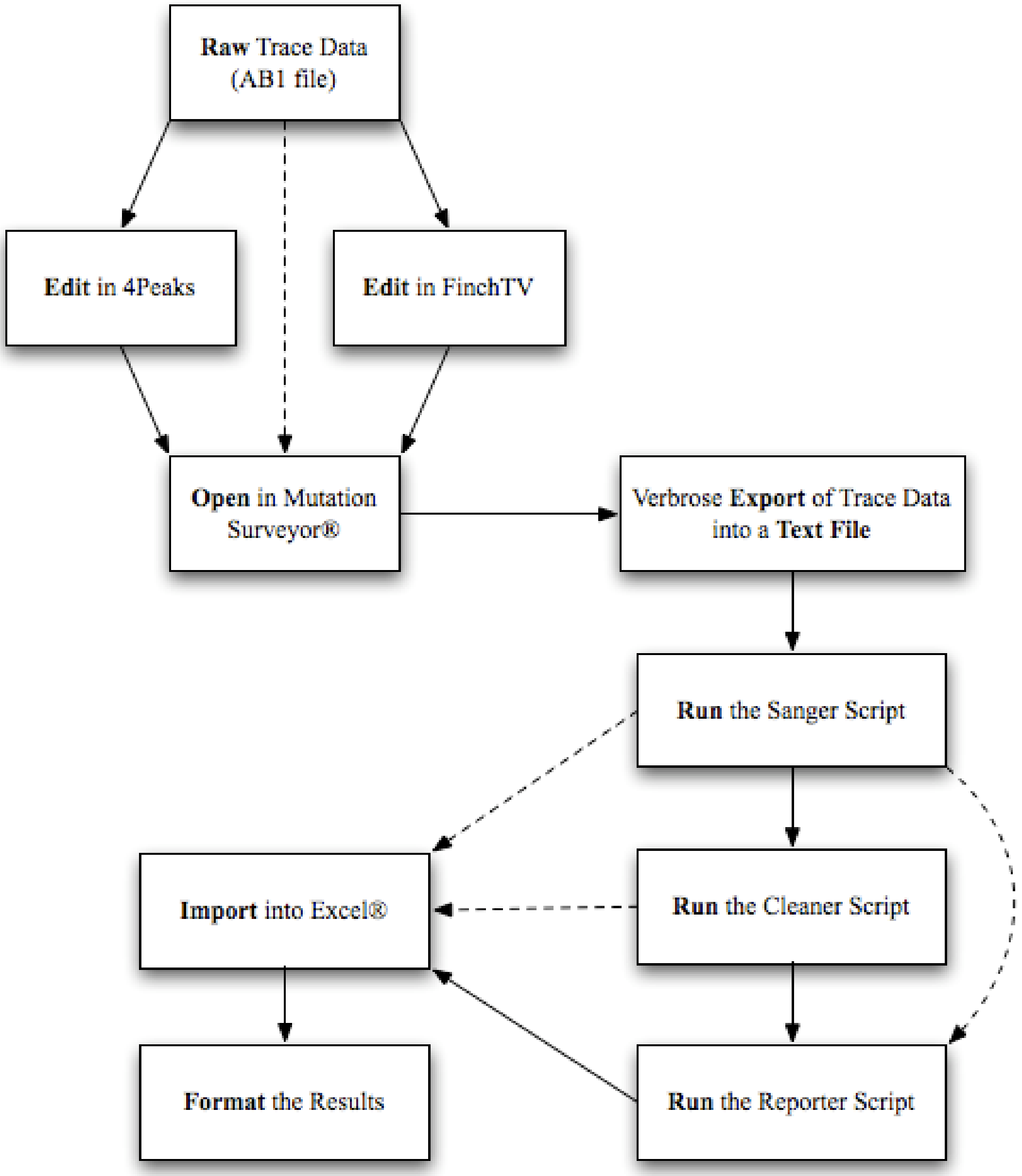


Figure # — Data workflow: from the AB1 Trace File to Excel®. Files are opened in either 4Peaks or FinchTV if they need editing, if not they are opened directly in Mutation Surveyor®. Then data is exported into a text file to be processed in the Sanger Script either individually or in a batch. Data is then processed all the way into the Reporter Script and then imported into Excel®. Dashed arrows indicate possible, yet not very common variations of the workflow.

2.5 – Workflow – Sequence analysis

Most researchers follow a simple procedure to analyze DNA sequence traces, this procedure is depicted in (Fig #):

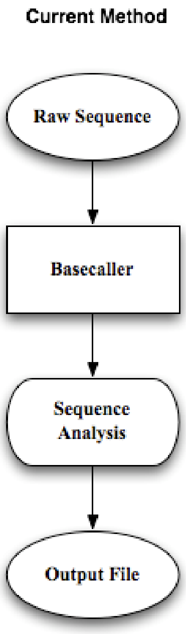


Figure # — Typical workflow: the raw sequence is processes within a software package and sequence analysis is made on the output file. Users don't intervene during the process.

The methodology followed, trough out this thesis is quite different. We came up with a better solution for the workflow. Our workflow is based on the best features of some software tools combined together to empower us to analyze trace sequence data in a more powerful, novel way (Fig #).

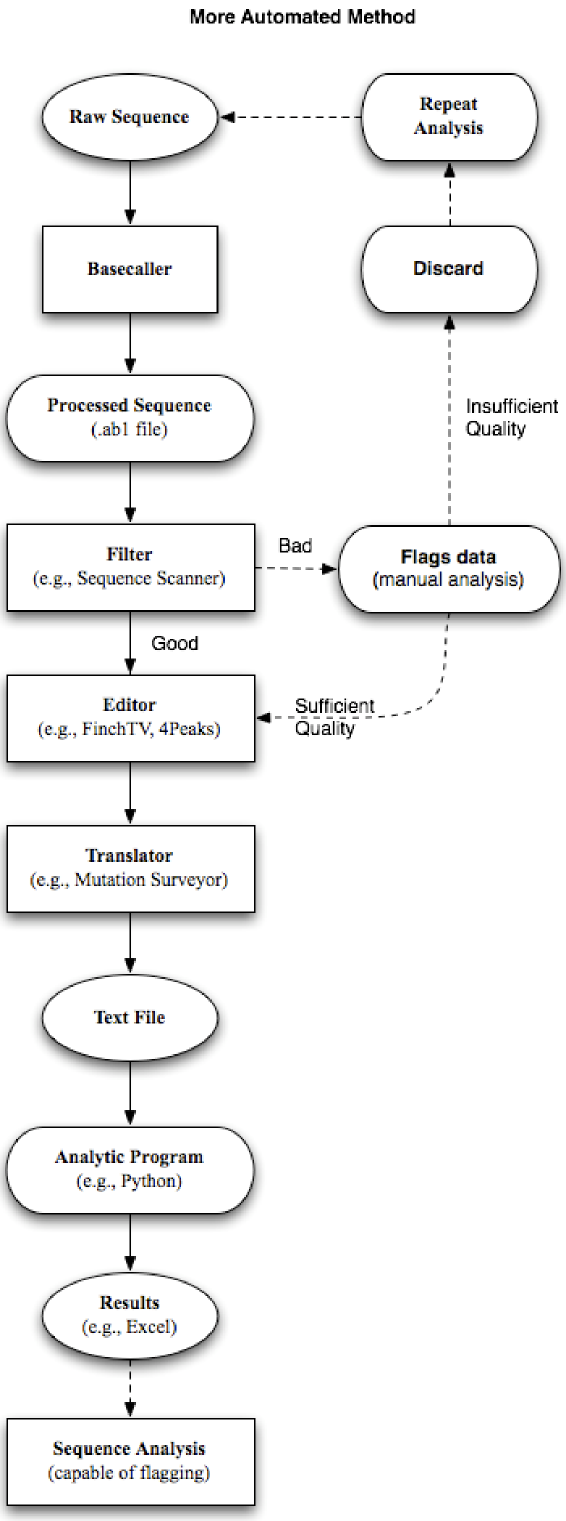

More Automated Method
Raw Sequence
Repeat Analysis
Basecaller
Discard
Processed Sequence (.ab1 file)
Insufficient Quality
Filter (e.g., Sequence Scanner)
Bad
Flags data (manual analysis)
Good
Editor (e.g., FinchTV, 4Peaks)
Sufficient Quality
Translator (e.g., Mutation Surveyor)
Text File
Analytic Program (e.g., Python)
Results (e.g., Excel)
Sequence Analysis (capable of flagging)

Figure 4 — Our workflow is based on a more complex analysis of sequence trace data. Data is processed, filtered, translated and analyzed using different types of software tools.

Looking at our procedure and the way how we were able to took existing tools and use bits and parts of them in order to provide a better sequence analisys, we can now propose a new way to look and analyse sequence data: an expert system. The backbone of such system is shown in figure #.

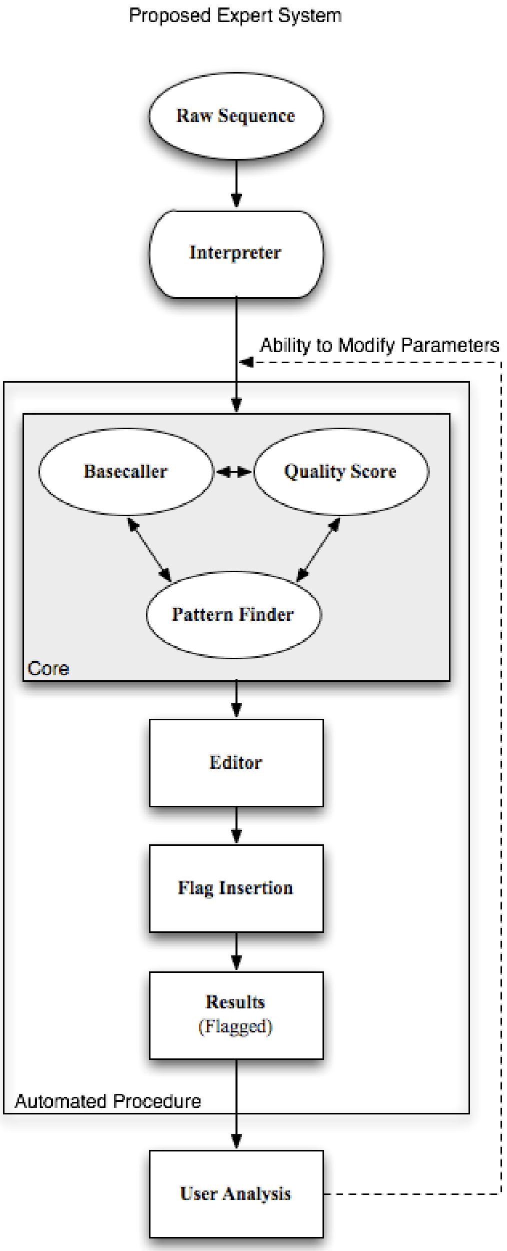


Figure 5 —Proposed expert system. An expert system is an automated software tool that can look at sequence data and, without user intervention flag problems in the sequence.

This expert system should be able to read sequence data, call bases, provide quality scores for those bases and find patterns within frames. Then, gathering these three bits of information, patterns, base call and quality, it would be able to edit and trim bad areas and/or flag bad bases, error points or mismatches. All in all, it would offer the user a digested view of the sequence rather than the simple, unprocessed results nowadays scientists are used to analyse. Such an expert system would therefore allow a much quicker analysis of big amounts of sequences and also a more reliable and consistent way of gathering quality information from the data.

Preliminary note about the results

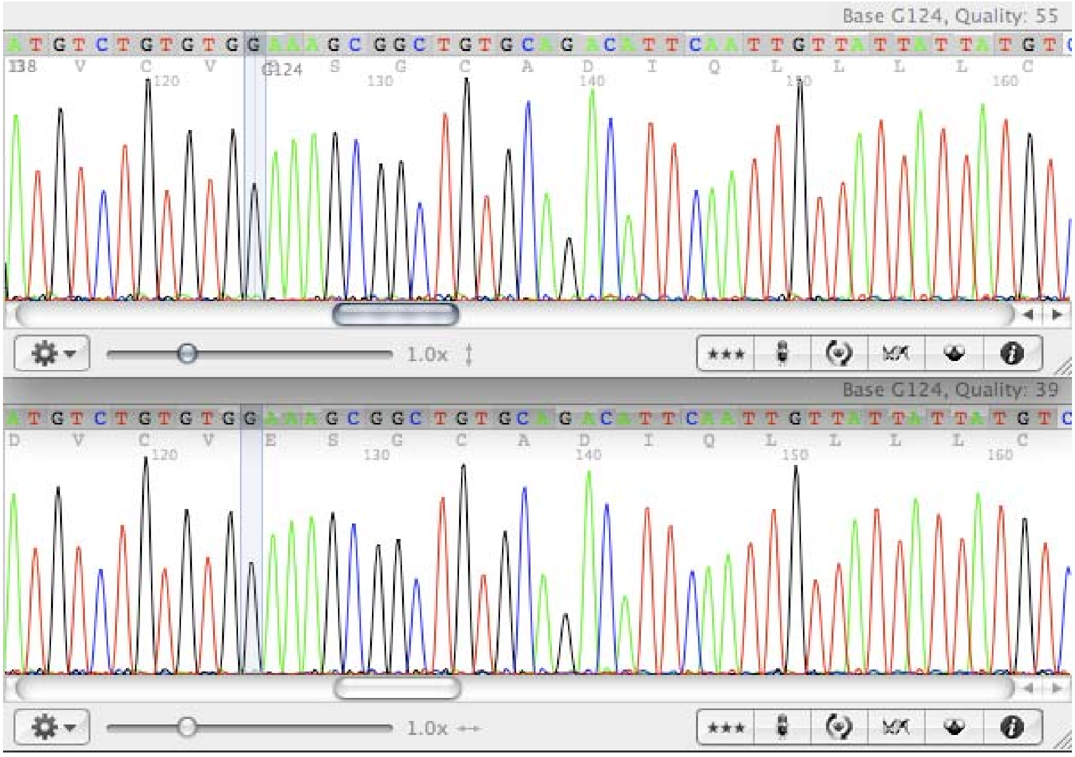
Results are always shown as a summary of the processed data. For each chapter, examples of raw can be found on the indicated annex. This has been done due to the unusual length of the raw data itself, which would force a trimming of most of its contents if it were to be shown or else a very long thesis. Yet, every example, though lengthy, has been chosen carefully to provide you, the reader, with a clear view of every step taken in the elaboration of the results. Remember that all the raw and processed data used throughout the thesis is available electronically on the CD that you can find on the back of this book, organized in folders by chapters. Information about system requirements and copyright notice is also printed on the back of this book.

3.1 – The same sample run on the same instrument with all chemistry parameters held constant produce similar sequence patterns.

In order to prove the existence and meaning of patterns, the first question we need to answer is whether peak height patterns within each frame are kept constant if the same sequence is ran several times in the same instrument. If so, this will mean that the patterns are kept constant and are not a totally random event.

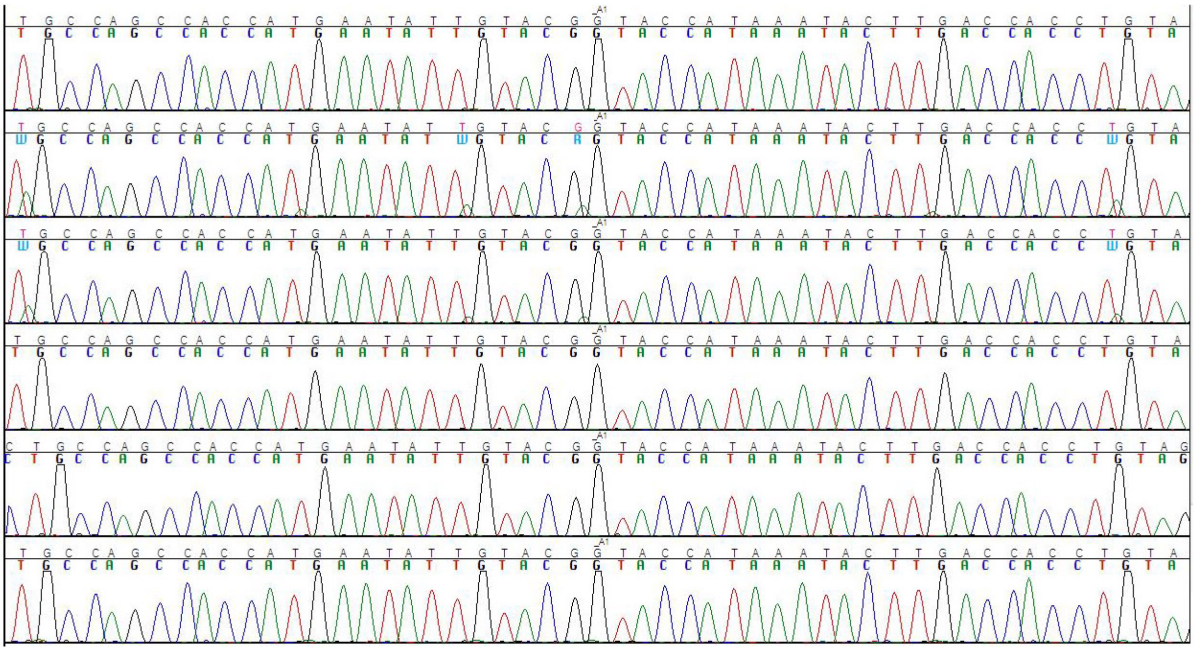
To do so, a sample of one individual was taken and distributed into two wells on the same plate and then the sequencing procedure was run as it is usual on the lab routine. Picture 1 shows the two chromatograms of the same sample ran twice on the same instrument.

Picture 1 – Two chromatograms depicting the same sample ran twice on the same intrument using BigDye 1.1 chemistry and POP-6. The sample is of mitochondrial origin and was sequenced using a D1 primer.



Picture 2 shows six chromatograms of the same sample ran over and over on the same instrument.

Pict2



As seen in both Picture 1 and 2 is clear that peak height patterns are kept constant on multiple runs for the same sample.

To mathematically analyze this observation, two sequence traces were taken from the same sample, aligned, ran trough processing scripts and compared. Then the same was done for other primer combinations. Please refer to Annex I for an example of the raw results. Table I summarizes the results of all these comparisons.

Table I – Difference comparison between the same samples ran twice on the same instrument.

Comparison	Total Frames	Dif. Frames	Dif. Patterns	Same Frame, dif. Pattern	Major differences
A1	380	3	34	31	0
B1	250	0	14	14	0
C1	357	0	12	10	0
D1	362	0	15	13	0

Notice that the first ten frames of each sequence were discarded because of noise and trimming artifacts and that major differences are counted if, and only if, a minor one does not precede them because otherwise it would not be an independent event. Of relevance, notice the small number of pattern differences (<10%) and the almost absent number of different frames.

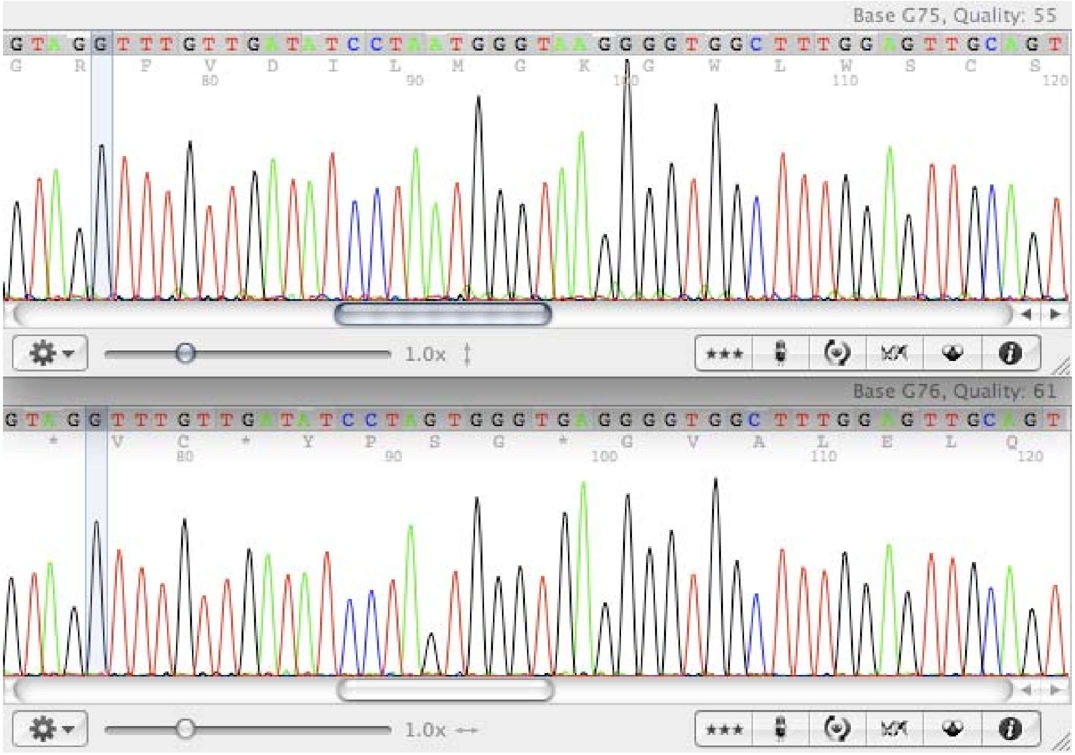
3.2 – Different samples run on the same instrument with all chemistry parameters held constant produce similar sequence patterns.

The last chapter showed that patterns are kept constant when the same sample is run, using the same parameters, more than once on the same instrument.

Now, the second question to answer is whether peak height patterns are kept constant if the samples are different but the parameters are kept the same. Hence, to know if two different samples behave in a similar way and show similar patterns if the parameters are kept constant and, if that is true, know if those results can be pooled together.

To do so, two different mtDNA samples form two different individuals were taken and, keeping all the primer and chemistry parameters constant, the sequencing procedure was run on the same instrument. Picture 3 shows the chromatograms of these two samples run in the same instrument.

PICT 3



B1 Data

As seen in Picture 3 is clear that peak height patterns are kept mostly constant from sample to sample. Bases may change, peak height may also change as so do patterns, but the same frame will show very similar pattern behavior.

To mathematically analyze this, sequence traces were taken from each of two different samples, aligned, ran trough processing scripts and compared. Then the same was done for other primer combinations. Please refer to Annex II for an example of the raw results. Table II summarizes the results of these comparisons.

Table II – Difference comparison between two different samples ran on the same instrument.

Comparison	Total Frames	Dif. Frames	Dif. Patterns	Same Frame, dif. Pattern	Major differences
A1	148	8	35	27	1
A2	81	8	26	14	2
B1	220	18	66	47	1
C1	194*	9	47	34	1
C2	122*	3	36	29	1

* Usable number of frames (before or after sequence mismatch). All values are calculated over this number.

Again, notice that the first ten frames of each sequence were discarded because of noise and trimming artifacts and that major differences are counted if, and only if, a minor one does not precede them because otherwise it would not be an independent event.

It is clear that, though differences do occur, most of the patterns are kept constant from one sample to the other. Moreover, very few major differences occur (no more than two significant differences in any comparison).

This indicates that, though the samples are different, similar sequences will produce similar patterns.

3.3 – The same sample run on different instruments within the same laboratory with all chemistry parameters held constant produces similar patterns.

Another preliminary question to answer is whether pattern behavior within frames is dependant on the chemistry and parameters used or the instrument itself.

In order to answer that, the same sample was taken and, keeping all the chemistry parameters constant, ran trough the sequencing procedure on two different instruments, but within the same lab.

These two sequence traces were taken, aligned, ran trough the processing scripts and compared. Please refer to Annex III for an example of the raw results. Table III summarizes the results of our comparisons.

Table III – Difference comparison between the same sample ran on two different instruments.

Comparison	Total Frames	Dif. Frames (minus cleaned)	Dif. Patterns	Same Frames, dif. Pattern	Major differences
A1	369*	0	25	25	0
B1	332*	0	33	33	0
C1	89*	0	20	20	0
C2	94*	0	8	8	0
D1	327*	0	26	24	1
D2	145*	0	14	14	0

* Usable number of frames (before or after sequence mismatch). All values are calculated over this number.

As usual, notice that the first ten frames of each sequence were discarded because of noise and trimming artifacts and that major differences are counted if, and only if, a minor one does not precede them because otherwise it would not be an independent event.

These results are very similar to those in chapter 3.1. Hence, the same sample ran in two different instruments produces patterns in a similar way as if that sample were to be run twice in just one instrument. All-in-all, just one case of a significant major difference was foundin one of the comparisons.

3.4 – Similar samples from the same lab can be pooled together.

The results form the last chapters suggest that patterns are kept constant within each frame, in the same lab, independently of the sample used if the chemistry and primer used are kept constant.

In order to prove that, batches of A1, B1, C1 and D1 sequences, using BigDye 1.1 chemistry were pooled together and analyzed statistically for two different labs.

Since the process involves pooling together dozens of sequences, result comparison had to be done in a different way using a simple chi-square test. Please refer to Annex IV for an example of the raw data. Table IV summarizes the results.

Table IV – Analysis of the predictability of patterns within frames.

Lab/Primer	Predictable Frames	Too little information	More than 4 blank patterns	Low chi-square	Random patterns
Lab 1 A1	55	2	5	2	0
Lab 2 A1	52	5	6	1	0
Lab 1 B1	54	3	5	2	0
Lab 2 B1	54	4	5	1	0
Lab 1 C1	63	0	1	0	0
Lab 2 C1	56	0	7	1	0
Lab 1 D1	54	0	9	1	0
Lab 2 D1	60	0	3	0	1

Predictable frames are all the frames for witch the p-value of the distribution is higher than 0,00078125 and the chi-square value is higher than 15,085, for a confidence of 95%. Values are only discarded for frames with less than 12 observations or low chi-square value.

Notice that only one case, with 23 observations produced a result that can be considered as a random distribution of patterns. There are 448 totally predictable frames (87.5%) and 41 (8%) of all frame pattern observations were grouped in less than two types of patterns. Only about 5% of the frames had too little information, a low chi-square value or showed a random distribution of patterns. It is then safe to assume that, given these results, peak height patterns within the same lab, for the same chemistry and primer are kept constant.

This is an important milestone. Not only peak height patterns are not random but also they can be grouped according to the primer and chemistry used for the same lab.

3.5 – Samples form different laboratories with all chemistry and primer parameters held constant produce similar, but not equal, sequence patterns.

In chapter 3.1 and 3.3 it was already addressed that the same sample ran in different instruments will produce similar patterns. It is also known, form chapter 3.2, that sequences from different samples can be pooled together into batches if, and only if, they are produced in the same lab using the same the chemistry and keeping all the parameters constant. Also, chapter 3.4 showed that patters are kept constant within a lab, if the chemistry and primer are the same.

Now, the question in hand is whether similar sequences will produce similar results in two different labs or not. It is important to know if sequence peak height patterns can be exporter from one lab to the other if the parameters are kept constant or if this patterns are a characteristic oa the lab that produced them.

To answer this question a batch of A1, B1, C1 and D1 sequences were pooled together from two different labs and compared. Please refer to annex for an example of the raw data. Table V summarizes the comparisons.

Table V – Comparison of frame patterns for the same chemistry in two different labs.

Primer	Comparable	Not comparable	Too little information	More than 6 blank cells
A1	30	19	2	13
B1	38	13	4	9
C1	32	28	0	4
D1	35	16	0	13

Opposed to what was previously shown, only about half of the patterns in each triplet can be compared from one lab to the other, when using the same chemistry in the two labs. Of notice, incomparable frames don’t exhibit, some times, the same types of patterns. Sometimes, though pattern distribution for the frame is different, the one of the pattern types is always missing.

3.6 – Two different samples with different primers (hence different DNA sequences analyzed) but same instrument and chemistry do not exhibit the same patterns.

To this point it has been made clear that, using the same chemistry and primers, the same region of the DNA, when sequenced, will produce similar patterns independently of the instrument used, but patterns may change quite a bit form lab to lab. The next step is to understand if different regions of DNA will produce similar patterns within the same lab, this is, if patterns are determined only by the frame chosen or, if not, dependant also on the entire DNA sequence that is being sequenced. To do so, one could take individual samples and compare them, but it is more clever to
look at batches of A1 sequences and compare them to B1, C1 and D1 sequences.
Please refer to Annex V for the raw results. Table V summarizes the results of our comparisons.

Table VI – Comparison of frame patterns for the same chemistry but two different primers.

Primers	Comparable	Not comparable	Too little information	More than 6 blank cells
A1 – C1	2	52	0	10
B1 – D1	4	51	0	9
A1 - B1	3	52	0	9

Since A1 sequences produce different types of patterns, for each frame than B1, C1 or D1 sequences, it becomes clear that different primers will produce different types of patterns for each frame, even within the same lab. Hence, peak height patterns in sequence data, though predictable, are also dependent on the DNA sequence.

3.7 – Same primers, similar samples but different chemistry produce different patterns.

At this point it is know that patterns can only be reproduced within frames if the lab, chemistry and primer are the same. Now, the remaining question is whether different chemistries produce different patterns or not.

To answer this question a batch of similar samples, processed in the same lab using different dye chemistries (dRhodamine or BigBye®) was taken for either A1, B1, C1 and D1 primers.

These results were then compared for significance.

Please refer to Annex VI for the raw results. Table VI summarizes the results of these comparisons.

Table VII – Comparison of frame patterns for the same primer but different chemistries.

Primer	Comparable	Not comparable	Too little information	More than 6 blank cells
A1	5	54	0	5
B1	3	57	0	4
C1	1	63	0	0
D1	5	59	0	0

* Usable number of frames (before or after sequence mismatch). All values are calculated over this number.

The results show that the chemistry used is also a major player in the formation of patterns for each frame. Different chemistries alter radically the distribution of patterns within each frame. Hence, besides the lab were samples were processed, their DNA sequence will also influence the patterns obtained for each frame, though, within the same lab, frames are kept constant, for a given chemistry within the same DNA sequence.

3.8 – Patterns can be predicted

All the results so far suggest that, if the chemical parameters and primers used are kept constant, patterns can be predicted for a given lab.

This means that one can build a pattern reference table for each frame given the primer and chemistry used in the sequencing.

3.6 Analysis of Patterns: BigDye v1.1

Patterns are not random

Our results show, first and foremost, that peak height patterns in sequence trace data are not random events. As opposed to what was initially though, they can be reproduced and are dependant on a series of technical and chemical factors that we have been able to deduce.

In fact, they are a characteristic of the sequencing procedure because they are dependant on the chemistry and DNA sequence.

We have demonstrated thatthe same sample or similar samples (same chemistry and primer) will produce the same type of patterns for each frame, independently of were the sequence was produced.

Therefore, it is possible to predict, knowing primers used what will be the patterns observed using dRhodamine chemistry or BigDye on any of the DNA zones (HV1 or HV2) sequenced and this prediction is independent of the instrument used or lab.

We can even compare sequences, side-by-side, using the patterns of each frame and tell the difference between different and equal samples.

Patterns change if DNA sequence changes

We have also shown that different labs will not produce different patterns even if the sample, instruments, chemistry, and primers are similar. Yet, the patterns will not be equal, they will be similar but not exactly the same. This phenomenon is associated with small differences in the primer design and in the processing of the sample or even small differences in the DNA zones sequenced and will produce, sometimes, different types of patterns characteristic of that zone.

Different chemistries or primers will produce different patterns in the frames, but if all parameters are kept constant, different samples will produce similar patterns for each zone of DNA analyzed. We can then talk about “sequence biometrics” or predictable patterns for each DNA area when we sequence them. This characteristic is comparable from one lab to the other unless any of the sequencing parameters change but not constant. It is only kept constant in the same lab, independently of the instrument used. Hence, it and can be calibrated. Small differences in DNA, such as point mutations, heteroplasms or mixes can, in theory, be predicted if the sequence biometrics for the lab and sequence has been calibrated. These new observations can lead to the introduction of new methods and procedures of sequence analysis in forensic, and non-forensic sequencing labs.

Factors that affect patterns

All in all, we have been able to identify several factors that do affect the way that patterns are formed. First, the primer used. If the primer changes the start of the sequencing even by one base, the whole spectrum of patterns for each frame in that sequence changes. It is not a radical change but yet it is significant. Second, the DNA zone sequenced. Much such as the primer used affects pattern formation so does the DNA zone sequenced. HV1 and HV2 mtDNA zones do not produce equal patterns for each frame. They do produce similar pattens, but they cannot be compared. Third, the dye chemistry used. The major differences found in pattern formation are caused by different dye chemistry. BigDye and dRhodamine dye chemistry produce radically different patterns for each frame.

In order to understand what might be happening that changes pattern formation so much we could compare peak height patterns to music notes, with tempo, pitch and partitures. We can also imagine the music being played by inexperienced musicians (error prone) using 4 different instruments. For two similar partitures, if one starts after or before the other, they will sound similar, but not the same because the tempo changes and the musicians make some mistakes. Moreover, if the partiture is different, the same notes might not sound the same because they are dependant on the harmonics of the other instruments. Even equal partitures, played with other different instruments, will sound different. So, the same sequence never “sounds” exactly the same and if we change the DNA zone or the chemistry the “music” will inevitably change also.

Therefore, we can assume that the way how patterns are formed is a characteristic of each sequence in that specific sequencing conditions.

Then, the reasons for pattern formation are dependant on the DNA sequence itself and the sequencing procedure used. They are not dependant on the instrument used (going back to the music example that would be the concert hall), but on the way they are produced.

There are molecular and physical reasons for these changes. At this point we can only speculate about those reasons. One thing is sure, pattern formation differences are affected by several factors, and not just one.

Among those factors we could point, as one of the most probable ones, the dye incorporation rate for each chemistry and the DNA sequence annealing kinetics for each PCR cycle. This would influence the amount of dye incorporated to the DNA on each cycle and therefore the signal obtained. The dye concentration can also be considered and different enzymes, with different rate incorporations must be accounted.

The DNA sequence, itself, its also important. Even the smallest time difference in denaturalizing DNA in each PCR cycle could mean noticeable differences in the amount of dye incorporation and therefore signal formation and consequently pattern formation.

The discussion of those factors, though interesting, goes beyond the scope of this thesis.

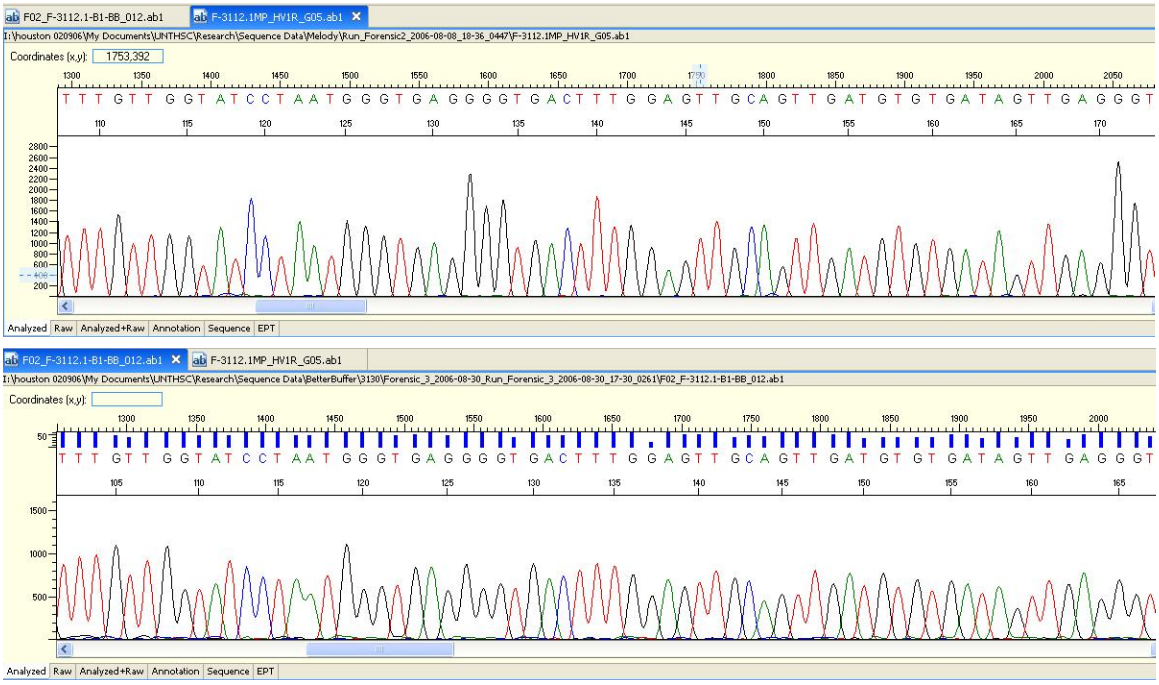
The most important fact to account is that, if all the parameters are kept constant (DNA zone, primer, chemistry and lab procedures) observed patterns will be independent of the instrumet used.

Predicting patterns

We have also been able to show that, for a kind of DNA sequence, with constant parameters, we can predict the patterns for each frame, and, of more interest, filter out anomalies in that frame.

Pattern characterization is, therefore, possible and a good tool for cleaning up the sequences. In this work we’ve been able to provide the starting point for the development of novel tools that, in the future, could be introduced into expert systems for DNA sequence analysis.

3.7 Analysis of patterns: dRhodamine, BigDye v3.1, Roche 454


F02_F-3112.1-B1-BB_012.ab1
F-3112.1MP_HV1R_G05.ab1
I:\houston 020906\My Documents\UNTHSC\Research\Sequence Data\Melody\Run_Forensic2_2006-08-08_18-36_0447\F-3112.1MP_HV1R_G05.ab1
Coordinates (x,y): 1753,392
T T T G T T G G T A T C C T A A T G G G T G A G G G G T G A C T T T G G A G T T G C A G T T G A T G T G T G A T A G T T G A G G G T
Analyzed
Raw
Analyzed+Raw
Annotation
Sequence
EPT
F02_F-3112.1-B1-BB_012.ab1
F-3112.1MP_HV1R_G05.ab1
I:\houston 020906\My Documents\UNTHSC\Research\Sequence Data\BetterBuffer\3130\Forensic_3_2006-08-30_Run_Forensic_3_2006-08-30_17-30_0261\F02_F-3112.1-B1-BB_012.ab1
Coordinates (x,y):
T T T G T T G G T A T C C T A A T G G G T G A G G G G T G A C T T T G G A G T T G C A G T T G A T G T G T G A T A G T T G A G G G T
Analyzed
Raw
Analyzed+Raw
Annotation
Sequence
EPT

Sequence data of the same HV1 product using the reverse primer B1 visualized with Sequence Scanner v1.0 (Applied Biosystems) produced with:

A) dRhodamineTerminator Cycle Sequencing Kit; and

B) ABI PRISM® BigDye® Terminator Cycle Sequencing Kit using the dR/E and BDX procedures, respectively. Note that the peak signal and patterns are different between the two dye chemistries (Roby *et al.*, 2007).

The Impact of Heteroplasmy and Mixtures on Patterns

Heteroplasmy and Mutations

A phenomenon known as heteroplasmy is when an individual exhibits more than one mtDNA type; this is usually observed as a single difference at one position. Heteroplasmy is mitochondrial genetic diversity at a single base position exhibiting either two or more bases at one position or exhibiting a length difference often seen in a polycytosine rich region. Single base heteroplasmy has been documented in forensic casework, most notably, in the identification of Tsar Nicholas II (Ivanov et al., 1996). Length heteroplasmy in the D-loop region has been described in both HV1 and HV2 (Bendall et al., 1995; Marchington et al., 1997; Wilson et al., 2002).

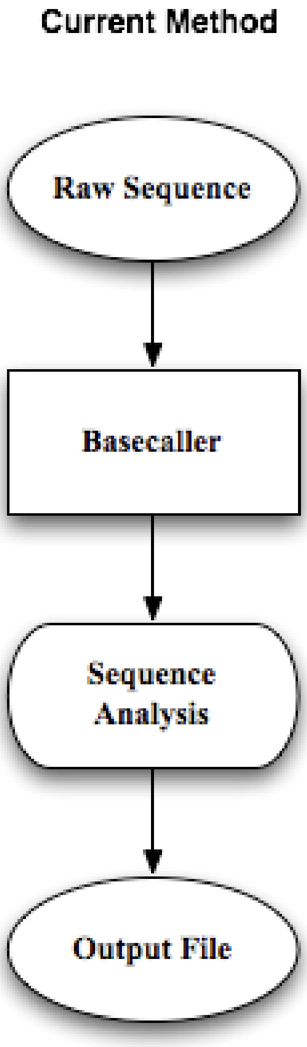
Point mutations or deletions in mtDNA usually increase exponentially with age. Defects in mitochondrial function produce a wide range of human diseases and can be caused by mutations within the mtDNA. The first mutation discovered in mtDNA to be the cause of a mitochondrial disease, Leber's hereditary optic neuropathy (LHON), was first identified in 1988 (Wallace et al.). LHON is a maternally inherited form of adult-onset blindness due to death of the optic nerve.

An even peak pattern improves the accuracy of base assignment by the base-calling software and miakes it easier to spot heterozygous bases in a sequence.

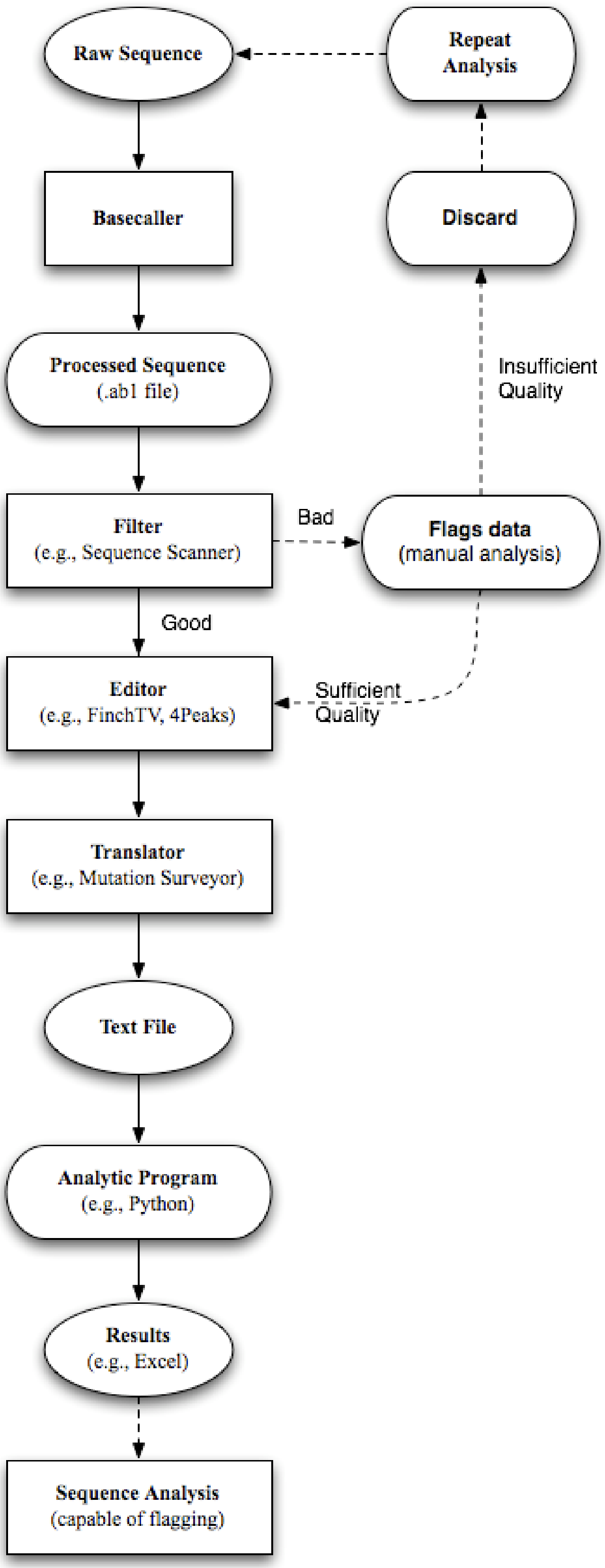
Individual Base Calls

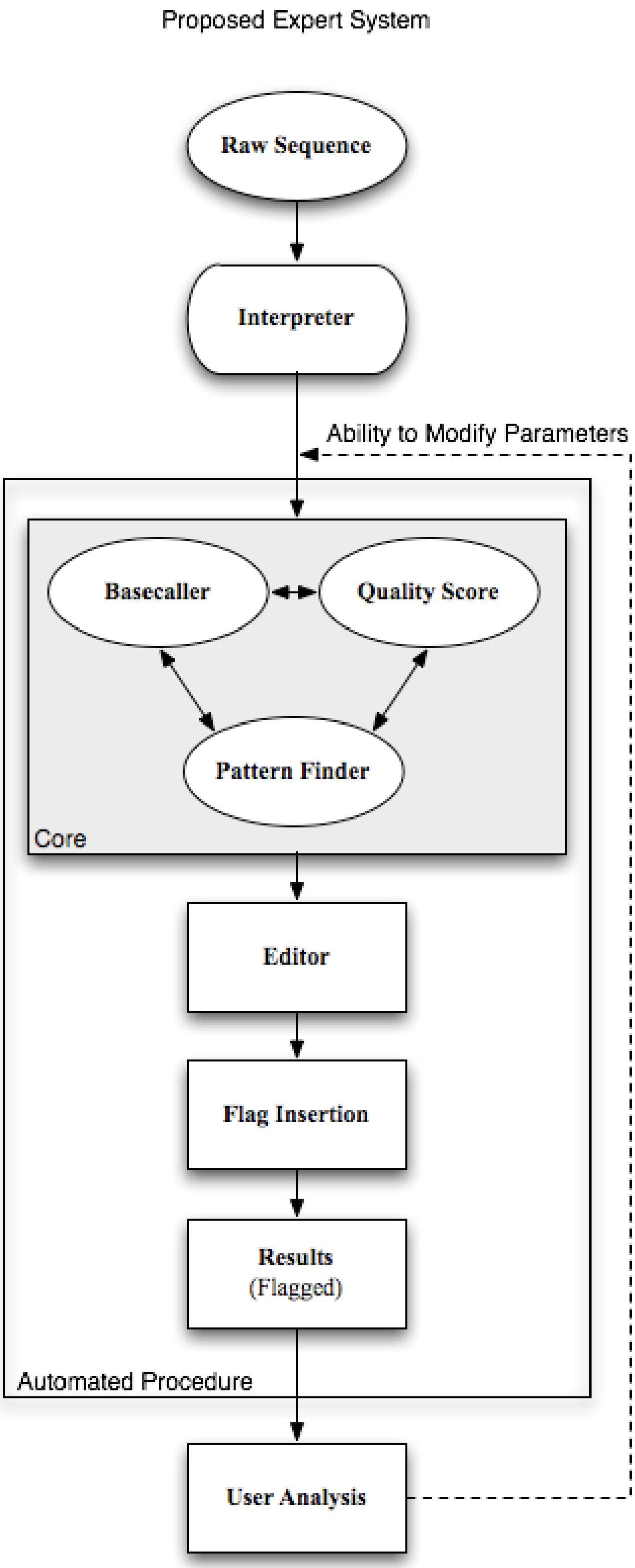
Another fact to consider when programming is that there is a 32-fold bias favoring transitions over transversions and there is a significantly nonrandom distribution of nucleotide substitutions and sequence length variation (Aquadro & Greenberg, 1983).

3.5 Development of an Expert System for Mitochondial DNA Analysis



More Automated Method




Proposed Expert System
Raw Sequence
Interpreter
Ability to Modify Parameters
Basecaller
Quality Score
Pattern Finder
Core
Editor
Flag Insertion
Results
(Flagged)
Automated Procedure
User Analysis

Chapter 4. Conclusion

Today there are eight laboratories in the United States with the ability to upload mtDNA results into the FBI's CODIS+mito databases. With this ability to upload mtDNA results into the CODIS+mito databases, the need to populate such databases and make it useful is ever-so-important. By developing a more streamlined robotic approach to generating quality sequence and an automated analysis method for mtDNA analysis, it will be possible to add more sequences to the database and in such, make more matches. The steps described in this thesis could be linked together as a seamless integrated full systematic method with adoption of robotic methods and middleware designed to put many of these steps into place.

Currently, it takes approximately one month for the staff at the UNTCHI to analyze 50 sets of human remains and 120 reference samples using both nuclear DNA and mtDNA. Using this teamwork batching approach, the average number of human remains that an analyst can process per month is 10 and the average number of reference samples that an analyst can process per month is 30. The length of time to process a single bone sample can vary considerably. On average, to obtain a complete mtDNA profile (hypervariable regions HV1 and HV2) for comparison or upload into CODIS+mito, it takes approximately four weeks from the time the bone is initially examined until completion. Some human remains make require considerably more time based upon sample quality. The length of time to process a single family reference sample is approximately two weeks from start to completion which includes the upload into CODIS+mito.

In order to address the increasing numbers of reference samples in laboratories throughout the world and in order to help reduce costs in this massive undertaking, an evaluation of existing methods and more efficient cutting edge technologies are continuously be assessed. Several steps in the analysis of mtDNA processing have been identified that can be modified to significantly reduce the labor in both the laboratory and in data review, reduce the costs and quantity of reagents, and reduce the overall processing time of a sample. A reduction in labor, reagents, and processing time will ensure savings in money and an increase in overall capacity of mtDNA testing by the laboratory. With the passage of legislation in many states in the United States addressing the missing persons issue, implementation of the proposed methods will increase efficiency of the laboratory operations throughout its scope resulting in more identifications aided and increased throughput. There is a potential of over 20,000 family reference samples to be submitted for missing persons cases, just within the United States. This number does not account for the number of missing persons cases worldwide. With the incorporation of multi-capillary instruments speeding up the data acquisition, bottlenecks in sample processing exist in DNA extraction and PCR amplification setup. Many robotics systems are used in forensic and vendor laboratories to process convicted offender samples worldwide. However, the use of robotics to extract DNA for both nuclear DNA and mtDNA testing with fixed pipette tips is not commonplace. This automated system is more cost effective and efficient in extracting DNA from human remains. We have addressed the DNA extraction step and will build upon this success to address other bottlenecks. These target areas are outlined below.

Amplification and Quantification. Increasing throughput for family reference samples by adopting a single amplification reaction that includes both hypervariable regions, HV1 and HV2, is feasible (Figure 2). A thorough evaluation must be performed to ensure an increase in throughput with minimal need for re-amplification and maximum capacity with robotic capabilities (Figure 3). Another area to be evaluated is the amplification process, where better optimization of the reactions, such as decreasing amplification volumes and reducing the cycle number in the amplification process lead to a decrease in the generated product (Table 1). With a decrease in the amount of product generated, it may be possible to directly perform cycle sequencing reactions with no need for quantification and dilutions of the PCR amplicon. This step would dramatically save time and money if shown to be successful. By decreasing the amount of amplified product, we will also be able to minimize the total reaction volume and the amount of ExoSAP-IT® required for template purification, providing another decrease in costs associated with each sequence generated.

Newer technologies are more sensitive to quantity and quality of DNA to be used in the assay. Today we are routinely using nanogram quantities of DNA in many assays and in some assays being evaluated today, we are able to get results with picogram levels of DNA. Some of these new assays, however, are more restrictive in the range of the amount of DNA being tested. Therefore, quantification of DNA is critical. With accurate quantification, the number of times that an assay fails and needs to be repeated is reduced resulting in quality data produced on the first attempt.

Accurately estimating the quantity of DNA can conserve a valuable DNA sample and can save both time and money. TaqMan® assays that provide the quantity of both nuclear DNA and mtDNA need to be evaluated thoroughly and optimized for use in routine forensic casework analysis (Figure 4).

Chemistry for Sequencing. Our laboratory has initiated a critical evaluation of the chemistry used for conventional mtDNA sequencing. We anticipate adopting a new version of sequencing chemistry that incorporates a dilution buffer resulting in a reduction of the manufacturer's standard input of fluorescent dyes. Extracted DNA from reference and challenged samples has been amplified for both regions of mtDNA, HV1 and HV2. In our current protocol, sequencing is achieved with the ABI PRISM® dRhodamine Cycle Sequencing Kit (Applied Biosystems, Foster City, CA) followed by purification with Performa™ columns and plates (Edge Biosystems Inc., Gaithersburg, MD). Our laboratory, as well as others, have shown that the cycle sequencing reaction can successfully be performed using the ABI PRISM® BigDye® Terminator v1.1 Cycle Sequencing Kit (Applied Biosystems). To enhance our efficiency and data quality, we have found that the reaction volume can be reduced 5-fold and an enhancement buffer, BetterBuffer (The Gel Company, San Francisco, CA), can be incorporated into the new sequencing procedure. These steps are then followed by a simple bead purification method, BigDye® XTerminator™ Purification (Applied Biosystems) to remove unincorporated BigDye® terminators, unnecessary salts, and unused diluent buffer. By adopting this new bead method of cleanup we will reduce the transfer of sequencing products from plate to plate one time, as well as additional centrifugation steps required by the currently used Performa™ plates. Advantages to these procedural modifications make the process amenable to complete automation in addition to the reduction in the overall costs.

Robotics. Many of the process improvements outlined thus far incorporated the use of automated platforms which reduce sample handling by forensic analysts. Not only do these changes promote better sample control, eliminating possible sample switches and fluid transfer accuracy, but also redirect the caseworkers' time to the more technically demanding role for which they are trained, namely, evidence screening and processing and data summary and reporting. For improvements in robotic sample processing, we anticipate developing additional scripts for setting up amplifications, clean-up procedures, cycle sequencing reactions, and dye removal.

Additionally, we would like to evaluate other robotic systems than those that we currently own for more programming flexibility. We currently utilize a Tecan Freedom EVO® 100 for extraction. By incorporating a Tecan Freedom EVO® 150 into post-PCR applications, we will have more flexibility in liquid handling for removal of unincorporated primers, cycle sequencing reactions, and cleanup. The Tecan Mini-Prep® 75s that are currently used are quite limiting due to the design of their liquid handling arms. We also have the need for adding a software package to our existing Tecan Freedom EVO® 100 robot for normalization of quantified mtDNA. With an increase in processing capacity, we need to evaluate other robotics, such as re-arraying robots by Tomtec® which increase sample density from 96-well plates to 384-well plates. Once a full amplicon procedure is developed and validated for the hypervariable regions of reference samples, re-arraying the products from the 96-well plate formats onto a 384-well plate will streamline the processing of the sequencing reactions that have been reduced in volume.

Data Analysis. Current forensic DNA sequencing methods incorporate a level of redundancy in the testing process to ensure quality data; however, as a result, a substantial amount of data is generated that requires analysis and review. Following forensic testing convention, two independent analysts must evaluate the sequence data and arrive at a concordant haplotype definition for each section of sequence and the sample as a whole. The bottleneck in data analysis can be addressed through software systems and program tools currently available to the scientific community; however, there is no package that links these tools together or automatically imports the data. Unlike the availability of expert systems for fragment analysis, there is not an expert system available for mtDNA sequence analysis. Automated data analysis with expert systems for nuclear DNA analysis is being validated and implemented for use with national databases by many laboratories. With automated data analysis, the analyst can be assured that the data meet minimum quality thresholds based on their validated interpretation guidelines. If the data do not meet the minimum quality thresholds, they are automatically marked for manual review by the analyst. Time savings in data analysis can be achieved by not having to review quality data. Automated data sequence analysis would be very beneficial to the casework analyst, as the analysis of mtDNA sequence data on average takes two and half times longer per sample than that of an autosomal STR profile.

In our laboratory, we have evaluated many software programs that can provide tools to quickly evaluate data. We have also developed steps that could be incorporated into a software program to help streamline the data analysis process. We have worked on the massive amounts of data produced from the World Trade Center tragedy and have previously collaborated with bioinformatics experts. It is our intent to work with software development teams to link these steps together to simplify this arduous task of sequence data analysis. Middleware will need to be developed and sophisticated programming for complex algorithms written. We have mathematically evaluated data output that could be programmed to lead to rule firings similar to the expert systems available for STR fragment analysis of reference samples. A laboratory workflow management system for sequence analysis can simplify laboratory processing and management, increase throughput, reduce errors and support real-time data analysis to the scientists. At this time, analysts must manually import data, evaluate each base, and have these data go through a second review by another scientist. This process is then followed by manual import into databasing systems and/or hard copy documents.

Initially, the .ab1 files generated from sequencing on 31xx series capillary electrophoresis instrumentation (Applied Biosystems) could automatically be imported into Sequence Scanner v1.0 software, a free software package available from Applied Biosystems. Sequence Scanner is a software package designed to display data, edit and trim it, export results, and generate reports. The Sequence Scanner software produces quality control (QC) reports, plate reports, trace score reports, and more. One useful tool in Sequence Scanner is the ability to view each raw image of the sequence traces from an entire plate in a “Thumbnails” view. The thumbnails view allows the scientist to quickly review the run and make informed, speedy determinations of data quality. We have determined quality metrics in this software that, if programmed into a pass/fail filter, would allow all samples with a trace score above a pre-defined metric to move directly to the next step in the analysis process. Additionally, those samples below the pre-defined metric could be flagged for human intervention, or the decision making process. “Can this sample move forward to analysis, or does it need to be re-sequenced with less PCR product or more PCR product?” If the sample passes and is not flagged, it could automatically be imported into a program allowing for edit and removal of any low quality data, such as produced at the beginning or end of sequence traces, or be trimmed to predefined start and stop points.

Two programs have been identified that allow for complete removal of identified sequence information: FinchTV (Geospiza™, Inc., Seattle, WA) and 4Peaks® (Mek & Tosj, The Netherlands Cancer Institute, Amsterdam). Other sequence edit programs do not remove these low quality bases, rather, they make them transparent. However, when the data are exported or imported to another program, these bases are still present in the final analysis. By incorporating FinchTV and/or 4Peaks® into the data analysis stream, minimum quality values can be defined and eventually programmed to remove the low quality base data. Once the low quality bases are trimmed, they can then be imported into a program such as Mutation Surveyor® (SoftGenetics LLC, State College, PA). Mutation Surveyor® allows for the sequence quantitative data to be exported into a simple, tab-delimited text file with the fluorescent color information and the relative peak height. Our laboratory has been collaborating with these groups. SoftGenetics has modified its software to fit the needs of our studies so we can evaluate the data as described. Currently, our laboratory utilizes Microsoft® Excel® to process the huge amount of generated data and combine it using macros specifically built for this study. These macros and the formulae allow us to closely evaluate a frame, or triplet, of bases along the DNA sequence. Each frame has a defined pattern due to the incorporation of the sequencing dyes to that particular sequence. The macros calculate the type of pattern for each frame. Note that there are 4^3 (64) possible combinations of triplets and each can show 6 different types of patterns.

Here, we present a first step in using simple bioinformatics tools to critically evaluate sequence data. Non-random patterns are displayed in a sequence window of three bases, or a frame. We have developed an automated system for characterizing each frame. Many of the patterns are reproducible. If the pattern that is displayed by the sequence data is not the pre-determined pattern for that frame, then a flag could fire and bring these sequence frames/trace data to the attention of the analyst for manual review. If, however, no flags are fired, then those data could be accepted and sent directly to a consensus layout program. By studying and understanding the different peak patterns in the data, we can make more accurate base calls and report fewer ambiguous positions.

In summary, the analysis of peak patterns can lead to the development of an automated expert system of base calling for the sequence community. Patterns in peak heights in local mtDNA sequencing frames from the D-loop have been defined and are automatically processed from the program we have written. These patterns are quite easy to establish and allow for predictive modeling in local sequence frames. Identifying these patterns in local sequence frames from single-source samples will increase base calling accuracy. A single base change can affect the peak heights, or patterns, in these local sequence frames. Patterns in peak heights have been characterized using dRhodamine and Big Dye™ terminator sequencing on an ABI PRISM 31xx DNA Sequencer. We now have an automated system to evaluate these local sequence frames. This model has been used in statistical studies to show the non-random nature of these patterns in each sequence frame due to different dye incorporation rates in each chemistry and the order of the string of bases.

The melding of a series of these analytical tools and processes can replicate the cognitive procedure that a trained mtDNA analyst performs in the evaluation of sequence data. Consultation with computer programmers to develop a process by which some of the software tools described, and possibly others, can be daisy-chained into a data analysis stream would create a prototype expert system for the analysis of mtDNA sequence data. The volume of data processed by our laboratory would serve as the perfect testbed to determine the feasibility of implementing a system such as this, as well as a venue to collect accurate time/cost data on its application.

LIMS Overview. Efficient processes and methods within the laboratory are typically thought to have a major impact on increasing throughput and success for forensic identifications. A major component of the whole laboratory package is overlooked: data management. Optimized methods, automation, and expert systems allow for the generation of large amounts of genetic data in a continuous and rapid stream. The management of the electronic data, beginning at sample accessioning and ending with statistical analyses and preparation of batch files for import into CODIS, is crucial in the efficient operation of a mid- to large-sized crime laboratory. The UNTCHI has partnered with the California Department of Justice Crime laboratory in the implementation of the LISA laboratory information system developed by Future Technologies Inc. (Fairfax, VA) for the Armed Forces DNA Identification Laboratory (AFDIL). Both the California Department of Justice and UNTCHI are actively involved in missing persons identification at the state and national levels, and therefore have similar data handling and analysis requirements. To date, several modifications to the sample accessioning and databasing/statistical analysis portions of the package have been made to accommodate the individual and joint requirements of these laboratories. Standard laboratory processes are being addressed within the LISA platform at this time at the UNTCHI. Effective implementation of the high throughput processing and data analysis components outlined in this proposal will require specific programming processes to be included in the LIMS system to accommodate the need for information by the process and the capture of data generated by the process. Ultimately, the goal of the laboratory is to perform its analyses in a completely automated and paperless system. Exemplar models for the practicality of this can be seen in the operations of laboratories, such as the Georgia Bureau of Investigation Crime Laboratory, who have essentially implemented paperless casework files. These applications not only capture critical case file information, but also facilitate faster review and reporting of the results of forensic analyses resulting in better service to the law enforcement community.

The implementation of the high efficiency/high throughput models described thus far in this proposal will require programming loop and feedback systems to be coded for specific lab processing steps captured in the LISA system. Many of the scripts for these processes already exist in some form in the software which is based on AFDIL's laboratory operations and will only require modifications specific to our processes. Major points which require the development of the feedback loop include 1) quantification (both nuclear DNA and mtDNA); 2) amplification setup and amplicon routing; 3) data capture and raw data storage from instrumentation; and 4) integration of expert systems data flow both into and out of the LIMS. An added benefit of interactive LIMS management of these systems is the ability to easily capture accurate metrics on reagent usage, time, and cost for processing of samples from accessioning to reporting. This allows for the accurate assessment of efficiency gains and throughput, while providing the laboratory with an accurate means to capture incremental increases in throughput and data to project realistic budgetary requirements.

Future Technology. The automated electrospray ionization mass spectrometry (ESI-MS) method developed by Ibis™ Biosciences Inc. (Carlsbad, CA) shows great promise for future mtDNA base composition determination and can reduce typing time and labor considerably. The ESI-MS method can produce base composition information for the analysis of the hypervariable regions routinely sequenced today. The sequence composition data can be directly compared to actual DNA sequence data that currently resides at NDIS allowing for a mostly seamless transition between the two technologies. One major advantage of ESI-MS is the huge reduction in processing time at a rate of one sample per minute. Additionally, as previously mentioned regarding the bottleneck in data analysis, the software accompanying the ESI-MS method offers a great advantage over traditional base calling analysis. We anticipate that the base composition data generated would be readily amenable to the expert system software approach that we are proposing. This method is highly amenable to automation, produces results in a higher throughput fashion with a reduction in overall labor of laboratory personnel. We currently have an active collaboration with Ibis™ Biosciences Inc. and we have been informed that the FBI is considering placing an instrument at the UNTCHI for future development and validation.

Summary. The objective of this project is to develop and integrate a workflow from laboratory processing to report generation of mtDNA haplotype data. This will be accomplished by addressing several bottlenecks in the processing and analysis of mtDNA as it is currently performed in our laboratory. The development of a new laboratory process with efficient amplification, sequencing, and analysis of mtDNA will greatly enhance throughput capabilities, decrease unit cost of sequencing, and significantly impact the amount of time for data review by the analyst. All of this will be accomplished by: 1) integration of our laboratory's automation platform; 2) expansion of our current LIMS capabilities; and, 3) through the development of expert system tools for assembly and analysis of mtDNA sequence data.

It is anticipated that funding and implementation of the full systematic method presented in this proposal could reduce the overall processing of skeletal remains and reference samples for mtDNA analysis by more than 35% once the LIMS, the robotics, and the software projects are completed and validated for use in casework. Prior to the implementation of this project measurements will be performed on time and pricing of the current procedures. Each step will be broken down and the overall process will be tallied. Upon successful implementation of each step, a complementary measurement will be conducted for comparison purposes and reported to NIJ. With no change in staffing, we anticipate increasing our capacity by approximately 35%.

FIGURES

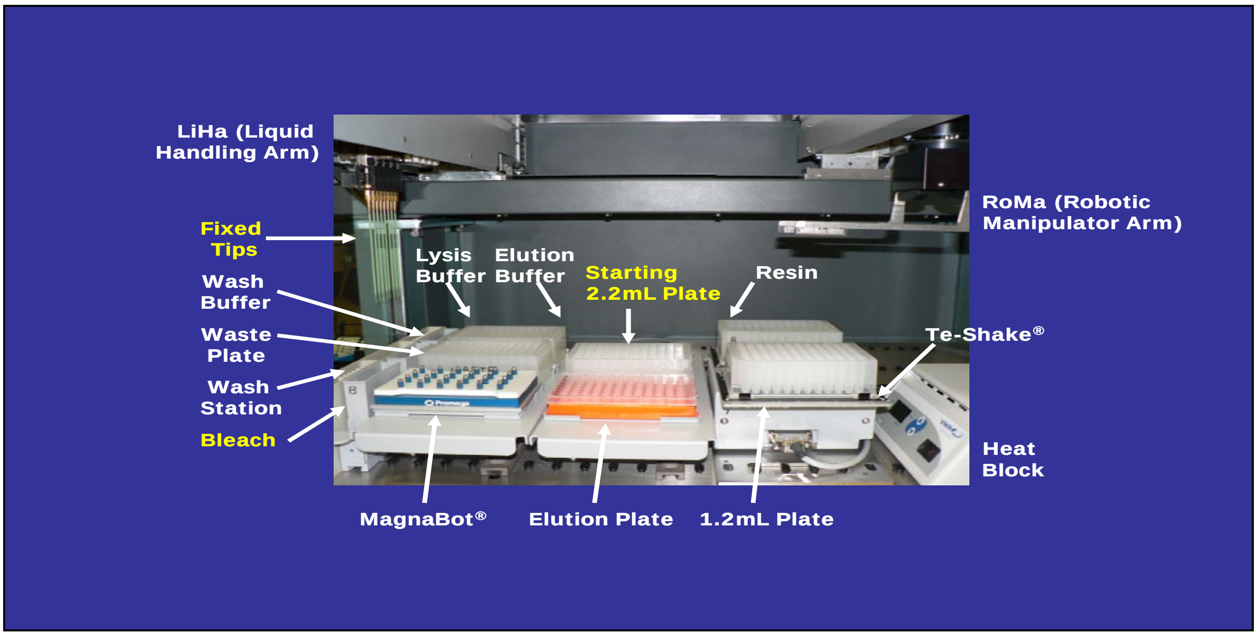


Figure 1. Tecan Freedom EVO® 100 Deck Setup for extraction of DNA from reference buccal samples for both nuclear and mtDNA.

TABLE

Cycle No.	Quantity (ng/µL)
32	5.40
30	4.14
28	3.42
26	2.29
24	1.59
22	0.14

Table 1. Results from the Agilent 2100 Bioanalyzer Chip for a sample with varying cycle number. Approved procedures utilize 32 cycles for sample amplification. By reducing the cycle number it may be possible to achieve optimal quantity of PCR product with no need for further dilutions prior to sequencing.

Discussion Points

1. Mitochondrial DNA analysis is laborious, time-consuming, and expensive. Any steps in the laboratory processing, screening of evidence, or analysis of data will greatly enhance a laboratory's throughput.

2. Preliminary data suggest that decreasing the number of cycles in amplification may alleviate the need for quantification of the amplified products.
3. Several real-time PCR assays have been developed for mtDNA quantification. A laboratory should closely evaluate the assays that have been developed and make a careful consideration in its choice for implementation into the laboratory's workflow.
4. It has been shown that multiplexing both HV1 and HV2 PCR products in the Agilent 2100 Bioanalyzer assay can reduce costs and time in quantification.
5. A multiplex screening tool with Amelogenin, a nuclear STR marker, and a mtDNA dinucleotide repeat has been designed for screening of remains and a multitude of stains or specimens in mass disaster or commingled specimen scenarios.

6. Expert systems are currently being evaluated for the forensic community with nuclear DNA databasing results. Individual quality values and scores are available for sequence data and should be considered in the development of an expert system for mtDNA results.

7. Frames of three (3) bases were subcategorized into patterns. These patterns were statistically evaluated for each of the 64 frames using Chi-square analysis and shown that frames are not distributed uniformly random across the six (6) patterns.

8. When comparing data between two (2) laboratories, it was shown that some frames are not distributed uniformly random and that such data may serve to provide a laboratory with a start optimization point in defining the thresholds for an expert system.

9. Different dye chemistries and different technologies produce different patterns and cannot be compared.

10. With the knowledge of mtDNA sequence data and the tools available in the sequencing community, a design plan for the development and integration of an expert system for mtDNA sequence analysis has been proposed.

Conclusiones

1. El análisis de ADN mitocondrial es laborioso, requiere mucho tiempo y es caro. Algunos pasos del procesamiento en el laboratorio, el rastreo de la evidencia o el análisis de la información mejoraran enormemente el rendimiento de un laboratorio.

2. Los datos preliminares sugieren que la disminución del número de ciclos en la amplificación puede paliar la necesidad de la cuantificación de los productos de amplificación.

3. Se han desarrollado varios ensayos de PCR a tiempo real para la cuantificación de ADNmt. Un laboratorio debería evaluar los ensayos que se han llevado a cabo y hacer una consideración cuidadosa en su elección para la implementación en el volumen de trabajo del laboratorio.

4. Se ha mostrado que el análisis múltiple de los productos de PCR de las regiones HV1 y HV2 en el Agilent 2100 Bioanalyzer pueden reducir los costes y el tiempo en el proceso de cuantificación.

5. Se ha diseñado una reacción múltiple con la Amelogenina y una repetición de dinucleótidos en el ADNmt, para su aplicación en distintos casos como: análisis de restos, de manchas y de mezclas.

6. Los sistemas expertos están siendo evaluados actualmente para la comunidad forense en bases de datos de ADN nuclear. Las tasas de calidad están disponibles para los datos de la secuencia y podrían ser tenidos en cuenta en el desarrollo de un sistema experto para los resultados del ADNmt.

7. Los marcos de lectura de tres bases fueron sub-clasificados en patrones. Estos patrones se evaluaron para cada uno de los 64 marcos de lectura utilizando el test de la Chi-cuadrado y se observó que los marcos de lectura no están distribuidos de forma uniformemente aleatoria a través de los seis patrones.

8. Cuando comparamos los datos obtenidos en dos laboratorios diferentes, observamos que algunos marcos de lectura no estaban distribuidos de forma uniformemente aleatoria y que estos datos pueden ser útiles para un laboratorio con un punto de optimización inicial a la hora de definir los umbrales para un sistema de experto.

9. Diferentes químicas de fluorocromos y tecnologías producen diferentes modelos y no pueden ser comparados.

10. Con el conocimiento de los datos de la secuencia del ADNmt y las herramientas de secuenciación disponibles en la comunidad, se ha propuesto un diseño para el desarrollo y la integración de un sistema experto para el análisis de la secuencia del ADNmt.

Appendices

Appendix 1

The tables in this appendix are automatically generated from the bioinformatic tools designed for this thesis. The first row is the frame obtained from the sequence. The second row is the pattern from each frame for the first injection or sample. The third row is the pattern from each frame for the second injection or sample. The Consensus row displays the consensus pattern for the sequences compared.

A dash (-) in the first row signifies those frames that have been deleted due to sequence variation in the basecaller program between samples in the second and third rows. Differences in the same sample are usually due to base sequence differences or dye background or other sequencing anomalies that occur with dye chemistry and capillary electrophoresis. A triple dash (---) in the first row is data that have been deleted due to frames exhibiting polymeric stretches of a particular base and cannot be considered independent. An asterisk (*) in the Consensus row signifies a different sequence between the samples/injections in the frame or a **minor** pattern difference when comparing Rows 2 and 3. A number sign (#) in the Consensus row signifies a **major** pattern difference when comparing the frames, Row 2, and Row 3.

With the sequencing protocol held constant (i.e., BigDye™ v1.1, BetterBuffer, 3130*xl* Genetic Analyzer, POP-6), the following comparisons have been made between two (2) injections. Each legend describes the comparisons performed with sequence metrics. Overwhelmingly, consensus of patterns is achieved.

Table A. The same sample sequenced twice with the A1 primer electrophoresed on the same instrument in the same run. The frames displayed in this example are from p16,019 to p16,398.

Frame	1		-	CCT	CTC	TCT	CTG	TGT	GTT	TTC	TCT	CTT	TTT	-	-	-
Inj.1	1		A	A	D	C	A	D	E	B	C	A	A	D	E	A
Inj.2	1		B	C	D	C	A	D	E	B	C	A	A	D	C	A
Consensus	1		*	*	D	C	A	D	E	B	C	A	A	*	*	*

Frame	15	TGG	GGG	---	GGA	GAA	AAG	AGC	GCA	CAG	AGA	GAT	ATT	TTT	TTG	TGG
Inj.1	15	BETA	F	-	A	A	D	E	B	F	C	B	F	F	C	B
Inj.2	15	D	F	-	A	A	D	E	B	F	C	B	F	F	C	B
Consensus	15	*	F	-	A	A	D	E	B	F	C	B	F	F	C	B

Frame	30	GGG	GGT	GTA	TAC	ACC	CCA	CAC	ACC	CCC	CCA	CAA	AAG	AGT	GTA	TAT
Inj.1	30	F	F	C	A	D	F	F	C	B	F	F	E	A	B	C
Inj.2	30	F	F	C	A	D	F	F	C	B	F	F	E	A	B	C
Consensus	30	F	F	C	A	D	F	F	C	B	F	F	E	A	B	C

Frame	45	ATT	TTG	TGA	GAC	ACT	CTT	TTA	TAC	ACC	CCC	CCA	CAT	ATC	TCA	CAA
Inj.1	45	B	C	B	E	A	D	C	D	E	D	C	B	F	E	D
Inj.2	45	B	C	B	E	B	F	C	D	E	D	C	A	D	E	D
Consensus	45	B	C	B	E	*	*	C	D	E	D	C	*	*	E	D

Frame	60	AAC	ACA	CAA	AAC	ACC	CCG	CGC	GCT	CTA	TAT	ATG	TGT	GTA	TAT	ATT
Inj.1	60	C	B	C	B	E	A	D	E	B	C	A	D	F	C	B
Inj.2	60	C	B	C	B	E	A	D	E	B	C	A	D	F	C	B
Consensus	60	C	B	C	B	E	A	D	E	B	C	A	D	F	C	B

Frame	75	TTT	TTC	TCG	CGT	GTA	TAC	ACA	CAT	ATT	TTA	TAC	ACT	CTG	TGC	GCC
Inj.1	75	F	F	E	A	A	A	B	F	F	C	B	E	A	D	E
Inj.2	75	F	F	E	A	A	A	D	E	D	C	B	E	A	D	E
Consensus	75	F	F	E	A	A	A	*	*	*	C	B	E	A	D	E

Frame	90	CCA	CAG	AGC	GCC	CCA	CAC	ACC	CCA	CAT	ATG	TGA	GAA	AAT	ATA	TAT
Inj.1	90	D	F	C	A	B	F	E	A	D	C	B	E	D	E	D
Inj.2	90	D	F	C	A	B	F	E	A	D	C	B	E	D	E	D
Consensus	90	D	F	C	A	B	F	E	A	D	C	B	E	D	E	D

Frame	105	ATT	TTG	TGT	GTA	TAC	ACG	CGG	GGT	GTA	TAC	ACC	CCA	CAT	ATA	TAA
Inj.1	105	F	C	D	E	A	B	C	D	E	A	B	F	E	A	D
Inj.2	105	F	C	D	E	A	B	C	D	E	A	B	F	E	A	D
Consensus	105	F	C	D	E	A	B	C	D	E	A	B	F	E	A	D

Frame	120	AAA	AAT	ATA	TAC	ACT	CTT	TTG	TGA	GAC	ACC	CCA	CAC	ACC	CCT	CTG
Inj.1	120	C	B	E	A	D	E	A	B	F	C	A	D	F	C	A
Inj.2	120	C	B	E	A	D	E	A	B	F	C	A	D	F	C	A
Consensus	120	C	B	E	A	D	E	A	B	F	C	A	D	F	C	A

Frame	135	TGT	GTA	TAG	AGT	GTA	TAC	ACA	CAT	ATA	TAA	AAA	---	---	AAC	ACC
Inj.1	135	D	F	F	C	A	A	B	F	C	D	F	－	－	C	B
Inj.2	135	D	F	F	C	A	A	A	D	C	D	E	－	－	C	B
Consensus	135	D	F	F	C	A	A	＊	＊	C	D	＊	－	－	C	B

Frame	150	CCC	CCA	CAA	AAT	ATC	TCC	CCA	CAC	ACA	CAT	ATC	TCA	CAA	AAA	---
Inj.1	150	F	E	B	E	A	B	C	D	C	B	F	E	B	C	－
Inj.2	150	F	E	B	E	A	B	C	D	C	B	F	E	D	C	－
Consensus	150	F	E	B	E	A	B	C	D	C	B	F	E	＊	C	－

Frame	165	AAC	ACC	CCC	---	---	CCT	CTC	TCC	CCC	CCT	CTA	TAT	ATG	TGC	GCT
Inj.1	165	E	A	D	－	－	A	B	E	D	C	A	B	C	B	F
Inj.2	165	E	A	D	－	－	C	B	E	D	C	A	B	C	B	F
Consensus	165	E	A	D	－	－	＊	B	E	D	C	A	B	C	B	F

Frame	180	CTT	TTA	TAC	ACA	CAA	AAG	AGC	GCA	CAA	AAG	AGT	GTA	TAC	ACA	CAG
Inj.1	180	F	C	B	F	E	B	C	D	E	B	E	D	C	B	F
Inj.2	180	F	C	B	F	E	D	C	D	E	B	E	D	C	B	F
Consensus	180	F	C	B	F	E	＊	C	D	E	B	E	D	C	B	F

Frame	195	AGC	GCA	CAA	AAT	ATC	TCA	CAA	AAC	ACC	CCC	CCT	CTC	TCA	CAA	AAC
Inj.1	195	F	E	B	C	B	E	A	A	A	D	E	D	E	B	C
Inj.2	195	F	E	D	C	B	E	A	A	A	D	E	D	E	B	E
Consensus	195	F	E	＊	C	B	E	A	A	A	D	E	D	E	B	＊

Frame	210	ACT	CTA	TAT	ATC	TCA	CAC	ACA	CAC	ACA	CAT	ATC	TCA	CAA	AAC	ACT
Inj.1	210	A	B	C	D	E	B	C	A	D	E	D	E	D	C	A
Inj.2	210	A	B	C	D	E	B	C	A	D	E	D	E	D	C	A
Consensus	210	A	B	C	D	E	B	C	A	D	E	D	E	D	C	A

Frame	225	CTG	TGC	GCA	CAA	AAC	ACT	CTC	TCC	CCA	CAA	AAA	AAG	AGC	GCC	CCA
Inj.1	225	A	D	F	E	A	A	D	E	D	E	A	D	C	A	D
Inj.2	225	A	D	F	E	A	A	D	E	D	E	A	D	C	A	D
Consensus	225	A	D	F	E	A	A	D	E	D	E	A	D	C	A	D

Frame	240	CAC	ACC	CCC	---	CCT	CTC	TCA	CAC	ACC	CCC	CCA	CAC	ACT	CTA	TAG
Inj.1	240	E	A	B	－	C	D	E	B	C	B	F	E	A	A	D
Inj.2	240	E	A	B	－	C	B	E	B	C	B	F	E	A	A	D
Consensus	240	E	A	B	－	C	＊	E	B	C	B	F	E	A	A	D

Frame	255	AGG	GGA	GAT	ATA	TAT	ATC	TCA	CAA	AAC	ACA	CAA	AAA	AAC	ACC	CCT
Inj.1	255	E	A	D	E	B	F	E	B	C	D	E	A	B	F	C
Inj.2	255	E	A	D	E	B	F	E	B	C	D	E	A	B	F	C
Consensus	255	E	A	D	E	B	F	E	B	C	D	E	A	B	F	C

Frame	270	CTA	TAC	ACC	CCC	CCA	CAC	ACC	CCC	CCT	CTT	TTA	TAA	AAC	ACA	CAG
Inj.1	270	D	C	B	F	E	A	A	D	E	A	A	D	E	A	D
Inj.2	270	D	C	B	F	E	A	A	D	E	A	B	F	E	A	D
Consensus	270	D	C	B	F	E	A	A	D	E	A	＊	＊	E	A	D

Frame	285	AGT	GTA	TAC	ACA	CAT	ATA	TAG	AGT	GTA	TAC	ACA	CAT	ATA	TAA	AAA
Inj.1	285	F	F	C	B	E	A	D	E	A	A	A	B	F	F	E
Inj.2	285	F	F	C	B	E	A	D	E	A	A	A	B	F	F	E
Consensus	285	F	F	C	B	E	A	D	E	A	A	A	B	F	F	E

Frame	300	AAG	AGC	GCC	CCA	CAT	ATT	TTT	TTA	TAC	ACC	CCG	CGT	GTA	TAC	ACA
Inj.1	300	A	A	A	D	E	B	F	E	A	B	E	D	E	A	D
Inj.2	300	A	A	A	D	E	B	F	E	A	D	E	D	E	A	D

Consensus	300	A	A	A	D	E	B	F	E	A	*	E	D	E	A	D

Frame	315	CAT	ATA	TAG	AGC	GCA	CAC	ACA	CAT	ATT	TTA	TAC	ACA	CAG	AGT	GTC
Inj.1	315	E	B	F	E	A	B	C	B	F	E	B	F	F	E	A
Inj.2	315	E	B	F	E	A	B	C	B	F	E	A	D	F	F	C
Consensus	315	E	B	F	E	A	B	C	B	F	E	*	*	F	*	*

Frame	330	TCA	CAA	AAA	AAT	ATC	TCC	CCC	CCT	CTT	TTC	TCT	CTC	TCG	CGC	GCC
Inj.1	330	D	C	B	E	B	C	D	C	A	D	E	D	E	A	A
Inj.2	330	D	C	B	C	B	C	D	C	A	D	E	D	E	A	A
Consensus	330	D	C	B	*	B	C	D	C	A	D	E	D	E	A	A

Frame	345	CCC	---	---	CCA	CAT	ATG	TGG	GGA	GAT	ATG	TGA	GAC	ACC	CCC	---
Inj.1	345	B	-	-	D	C	A	D	E	D	C	A	D	E	D	-
Inj.2	345	B	-	-	D	C	A	BETA	E	D	C	A	D	E	D	-
Consensus	345	B	-	-	D	C	A	*	E	D	C	A	D	E	D	-

Frame	360	---	---	CCT	CTC	TCA	CAG	AGA	GAT	ATA	TAG	AGG	GGG	---	GGT	GTC
Inj.1	360	-	-	C	D	E	D	C	B	E	D	C	B	-	F	ALFA
Inj.2	360	-	-	A	D	E	D	C	B	E	D	C	B	-	F	F
Consensus	360	-	-	*	D	E	D	C	B	E	D	C	B	-	F	*

Frame	375	TCC	CCC	CCT	CTT	TTG
Inj.1	375	PSI	D	E	A	A
Inj.2	375	C	D	E	A	A
Consensus	375	*	D	E	A	A

Table B. Two (2) different samples sequenced with the A1 primer electrophoresed on the same instrument in the same run. The frames displayed in this example are from p16,043 to p16,186.

Frame	1		ATT	TTT	TTG	TGG	GGG	GGT	GTA	TAC	ACC	CCA	CAC	ACC	CCC	CCA
Sample 1	1		F	ALFA	PSI	B	F	F	C	A	D	E	D	C	B	E
Sample 2	1		F	F	C	B	F	F	C	A	D	E	D	C	B	C
Consenus	1		F	*	*	B	F	F	C	A	D	E	D	C	B	*

Frame	15	CAA	AAG	AGT	GTA	TAT	ATT	TTG	TGA	GAC	ACT	CTC	TCA	CAC	ACC	CCC
Sample 1	15	D	E	A	B	C	B	C	B	E	A	D	E	B	C	B
Sample 2	15	D	F	C	B	C	D	C	B	E	B	F	E	B	E	B
Consensus	15	D	*	*	B	C	*	C	B	E	*	*	E	B	*	B

Frame	30	CCA	CAT	ATC	TCA	CAA	AAC	ACA	CAA	AAC	ACC	CCG	CGC	GCT	CTA	TAT
Sample 1	30	E	A	D	C	B	E	B	C	B	E	A	D	E	B	C
Sample 2	30	C	A	D	C	B	E	B	C	D	E	A	D	E	A	B
Consensus	30	*	A	D	C	B	E	B	C	*	E	A	D	E	*	#

Frame	45	ATG	TGT	GTA	TAT	ATT	TTT	TTC	TCG	CGT	GTA	TAC	ACA	CAT	ATT	TTA
Sample 1	45	A	D	E	A	D	F	F	E	A	A	A	B	F	F	C
Sample 2	45	C	D	E	A	D	F	F	E	A	A	A	D	F	F	C
Consensus	45	*	D	E	A	D	F	F	E	A	A	A	*	F	F	C

Frame	60	TAC	ACT	CTG	TGC	GCC	CCA	CAG	AGC	GCC	CCA	CAC	ACC	CCA	CAT	ATG
Sample 1	60	B	E	A	D	E	D	F	C	A	B	F	E	A	D	C
Sample 2	60	B	E	A	D	E	D	F	C	A	B	F	F	C	D	C
Consensus	60	B	E	A	D	E	D	F	C	A	B	F	*	*	D	C

Frame	75	TGA	GAA	AAT	ATA	TAT	ATT	TTG	–	–	–	ACG	CGG	GGT	GTA	TAC
Sample 1	75	B	E	D	E	D	F	C	D	E	A	B	C	D	E	A
Sample 2	75	B	E	D	E	D	F	C	D	E	A	D	C	D	E	A
Consensus	75	B	E	D	E	D	F	C	*	*	*	*	C	D	E	A

Frame	90	ACC	CCA	CAT	ATA	TAA	AAA	AAT	ATA	TAC	ACT	CTT	TTG	TGA	GAC	ACC
Sample 1	90	B	F	C	A	D	C	B	E	A	D	E	A	D	F	C
Sample 2	90	B	F	E	A	D	C	B	E	A	D	E	A	D	F	C
Consensus	90	B	F	*	A	D	C	B	E	A	D	E	A	D	F	C

Frame	105	CCA	CAC	ACC	CCT	CTG	TGT	GTA	TAG	AGT	GTA	TAC	ACA	CAT	ATA	TAA
Sample 1	105	A	D	F	C	A	D	F	F	C	A	A	B	F	C	D
Sample 2	105	A	D	F	C	A	D	F	F	C	A	A	A	D	C	D
Consensus	105	A	D	F	C	A	D	F	F	C	A	A	*	*	C	D

Frame	120	AAA	---	---	AAC	ACC	CCC	CCA	CAA	AAT	ATC	TCC	CCA	CAC	ACA	CAT
Sample 1	120	E	–	–	C	B	F	E	D	E	A	B	C	D	E	A
Sample 2	120	E	–	–	C	B	F	E	B	F	C	B	C	D	E	B
Consensus	120	E	–	–	C	B	F	E	*	*	*	B	C	D	E	*

Frame	135	ATC	TCA	CAA	AAA	---	AAC	ACC	CCC	–	–	–	–	–
Sample 1	135	D	E	D	C	–	C	A	D	–	–	A	B	E
Sample 2	135	F	E	B	C	–	E	A	D	E	B	E	D	–
Consensus	135	*	E	*	C	–	*	A	D	*	*	#	*	*

Table C. The same sample sequenced with the B1 primer electrophoresed on two (2) different instruments in same laboratory. The frames displayed in this example are from p16,367 to p16,034.

Frame	1		–	–	–	–	–	–	–	–	---	---	–	–	–	–
Inst. 1	1		–	–	–	–	–	–	–	–	–	–	–	–	–	–
Inst. 2	1		A	B	F	C	A	B	F	E	–	–	D	C	B	F
Consensus	1		*	*	*	*	*	*	*	*	–	–	*	*	*	*
Frame	15	–	–	–	–	TGG	GGG	---	GGA	GAC	ACG	CGA	GAG	AGA	GAA	AAG
Inst. 1	15	–	–	–	–	B	F	–	A	A	D	E	D	E	B	E
Inst. 2	15	E	A	A	B	F	F	–	A	A	D	E	D	E	B	E
Consensus	15	*	*	*	*	#	F	–	A	A	D	E	D	E	B	E
Frame	30	AGG	GGG	GGA	GAT	ATT	TTT	TTG	TGA	GAC	ACT	CTG	TGT	GTA	TAA	AAT
Inst. 1	30	A	D	E	B	F	F	E	A	A	A	A	D	F	E	A
Inst. 2	30	A	D	E	A	D	F	E	A	A	A	A	D	F	E	A
Consensus	30	A	D	E	*	*	F	E	A	A	A	A	D	F	E	A
Frame	45	ATG	TGT	GTG	TGC	GCT	CTA	TAT	ATG	TGT	GTA	TAC	ACG	CGA	GAT	ATG
Inst. 1	45	A	D	E	A	D	F	C	A	D	F	C	B	C	B	E
Inst. 2	45	A	D	E	A	D	F	C	A	D	F	C	B	C	B	E
Consensus	45	A	D	E	A	D	F	C	A	D	F	C	B	C	B	E
Frame	60	TGA	GAA	AAT	ATG	TGG	GGC	GCT	CTT	TTT	TTA	TAT	ATG	TGT	GTA	TAC
Inst. 1	60	A	A	D	C	B	F	E	B	F	F	C	B	F	F	C
Inst. 2	60	A	A	D	C	D	F	E	B	F	F	C	A	B	F	C
Consensus	60	A	A	D	C	*	F	E	B	F	F	C	*	#	F	C
Frame	75	ACT	CTA	TAT	ATG	TGT	GTA	TAC	ACT	CTG	TGT	GTT	TTG	TGA	GAG	AGG
Inst. 1	75	B	E	B	E	D	F	C	B	C	D	E	D	E	D	C
Inst. 2	75	B	E	B	E	D	F	C	B	C	D	E	D	E	D	C
Consensus	75	B	E	B	E	D	F	C	B	C	D	E	D	E	D	C
Frame	90	GGA	GAT	ATG	TGG	GGG	GGT	GTA	TAG	AGG	GGT	GTT	TTT	TTG	TGT	GTT
Inst. 1	90	B	F	C	B	F	E	B	F	C	B	F	F	C	D	E
Inst. 2	90	B	F	C	B	F	E	D	F	C	B	F	F	C	D	E
Consensus	90	B	F	C	B	F	E	*	F	C	B	F	F	C	D	E
Frame	105	TTG	TGG	GGT	GTA	TAT	ATC	TCC	CCT	CTA	TAG	AGT	GTG	TGG	GGG	GGT
Inst. 1	105	A	B	F	F	C	B	E	D	C	D	E	A	D	E	B
Inst. 2	105	A	B	F	F	C	B	E	B	C	D	E	A	D	E	B
Consensus	105	A	B	F	F	C	B	E	*	C	D	E	A	D	E	B
Frame	120	GTG	TGA	GAG	AGG	GGG	---	GGT	GTG	TGG	GGC	GCT	CTT	TTT	TTG	TGG
Inst. 1	120	C	A	D	C	B	–	D	C	B	F	C	B	F	C	D
Inst. 2	120	C	A	D	C	B	–	D	C	B	F	C	B	F	C	D
Consensus	120	C	A	D	C	B	–	D	C	B	F	C	B	F	C	D
Frame	135	GGA	GAG	AGT	GTT	TTG	TGC	GCA	CAG	AGT	GTT	TTG	TGA	GAT	ATG	TGT
Inst. 1	135	E	B	C	B	F	F	F	F	E	A	B	C	D	C	B
Inst. 2	135	E	B	C	B	F	F	F	F	E	A	A	A	D	C	B
Consensus	135	E	B	C	B	F	F	F	F	E	A	*	*	D	C	B
Frame	150	GTG	TGT	GTG	TGA	GAT	ATA	TAG	AGT	GTT	TTG	TGA	GAA	AAG	AGG	GGT
Inst. 1	150	E	D	C	A	D	E	D	E	A	B	C	B	F	C	B
Inst. 2	150	E	D	C	A	D	E	D	E	A	B	C	BETA	F	C	B
Consensus	150	E	D	C	A	D	E	D	E	A	B	C	*	F	C	B
Frame	165	GTT	TTG	TGA	GAT	ATT	TTG	TGC	GCT	CTG	TGT	GTA	TAC	ACT	CTT	TTG
Inst. 1	165	F	C	A	D	E	A	D	E	A	D	F	C	A	A	A
Inst. 2	165	F	C	A	D	E	A	D	E	A	D	F	C	A	A	A
Consensus	165	F	C	A	D	E	A	D	E	A	D	F	C	A	A	A
Frame	180	TGC	GCT	CTT	TTG	TGT	GTA	TAA	AAG	AGC	GCA	CAT	ATG	TGG	GGG	---
Inst. 1	180	D	E	B	C	D	F	C	B	C	B	E	A	D	F	–
Inst. 2	180	D	E	B	C	D	F	C	B	C	B	E	A	D	F	–
Consensus	180	D	E	B	C	D	F	C	B	C	B	E	A	D	F	–

Frame	195	GGA	GAG	AGG	GGG	---	---	GGT	GTT	TTT	---	TTG	TGA	GAT	ATG	TGT
Inst. 1	195	A	B	C	B	－	－	D	F	F	－	C	A	D	C	B
Inst. 2	195	A	A	A	B	－	－	D	F	F	－	C	B	F	C	B
Consensus	195	A	＊	＊	B	－	－	D	F	F	－	C	＊	＊	C	B

Frame	210	GTG	TGG	GGA	GAT	ATT	TTG	TGG	GGG	GGT	GTT	TTT	---	---	TTA	TAT
Inst. 1	210	ALFA	GAMMA	C	B	F	C	D	E	A	A	D	－	－	C	B
Inst. 2	210	E	D	C	B	F	C	D	E	A	A	D	－	－	C	B
Consensus	210	＊	#	C	B	F	C	D	E	A	A	D	－	－	C	B

Frame	225	ATG	TGT	GTA	TAC	ACT	CTA	TAC	ACA	CAG	AGG	GGT	GTG	TGG	GGT	GTC
Inst. 1	225	F	F	F	C	A	A	D	F	F	C	B	C	B	F	E
Inst. 2	225	F	F	F	C	B	C	D	F	F	C	B	C	B	F	E
Consensus	225	F	F	F	C	＊	＊	D	F	F	C	B	C	B	F	E

Frame	240	TCA	CAA	AAG	AGT	GTA	TAT	ATT	TTT	TTA	TAT	ATG	TGG	GGT	GTA	TAC
Inst. 1	240	A	A	D	C	B	C	B	F	C	A	D	E	D	E	A
Inst. 2	240	A	A	D	C	B	E	B	F	C	A	D	E	D	E	A
Consensus	240	A	A	D	C	B	＊	B	F	C	A	D	E	D	E	A

Frame	255	ACC	CCG	CGT	GTA	TAC	ACA	CAA	AAT	ATA	TAT	ATT	TTC	TCA	CAT	ATG
Inst. 1	255	B	F	E	B	C	B	F	C	A	B	F	F	E	A	A
Inst. 2	255	B	F	E	B	C	B	F	E	A	B	F	F	E	A	A
Consensus	255	B	F	E	B	C	B	F	＊	A	B	F	F	E	A	A

Frame	270	TGG	GGT	GTG	TGG	GGC	GCT	CTG	TGG	GGC	GCA	CAG	AGT	GTA	TAA	AAT
Inst. 1	270	D	F	C	B	F	C	A	D	F	E	D	E	B	C	A
Inst. 2	270	B	F	C	B	F	C	A	D	F	F	F	C	B	C	A
Consensus	270	＊	F	C	B	F	C	A	D	F	＊	＊	＊	B	C	A

Frame	285	ATG	TGT	GTA	TAC	ACG	CGA	GAA	AAA	AAT	ATA	TAC	ACA	CAT	ATA	TAG
Inst. 1	285	A	D	E	A	D	C	B	E	B	C	B	F	C	B	F
Inst. 2	285	A	D	E	A	D	C	B	E	B	C	B	F	C	B	F
Consensus	285	A	D	E	A	D	C	B	E	B	C	B	F	C	B	F

Frame	300	AGC	GCG	CGG	GGT	GTT	TTG	TGT	GTT	TTG	TGA	GAT	ATG	TGG	GGG	GGT
Inst. 1	300	E	A	A	D	C	A	D	E	A	A	D	E	B	F	E
Inst. 2	300	E	A	B	F	C	A	D	E	A	A	D	C	B	F	E
Consensus	300	E	A	＊	＊	C	A	D	E	A	A	D	＊	B	F	E

Frame	315	GTG	TGA	GAG	AGT	GTC	TCA	CAA	AAT	ATA	TAC	ACT	CTT	TTG	TGG	GGG
Inst. 1	315	A	A	D	E	B	C	B	E	D	E	A	A	D	F	E
Inst. 2	315	A	A	D	E	B	C	B	E	B	E	A	A	D	F	E
Consensus	315	A	A	D	E	B	C	B	E	＊	E	A	A	D	F	E

Frame	330	GGT	GTG	TGG	GGT	GTA	TAC	ACC	CCC	CCA	CAA	AAA	AAT	ATC	TCT	CTG
Inst. 1	330	A	A	B	F	E	A	B	F	F	F	C	A	D	C	A
Inst. 2	330	A	A	B	F	CHI	A	B	F	F	F	C	A	D	C	A
Consensus	330	A	A	B	F	＊	A	B	F	F	F	C	A	D	C	A

Frame	345	TGC	GCT	CTT	TTC	TCC	CCC
Inst. 1	345	D	E	A	D	E	B
Inst. 2	345	D	E	B	F	E	D
Consensus	345	D	E	＊	＊	E	＊

Appendix 2

The following tables are examples of the processed data using the bioinformatic tools designed to statistically evaluate each of the 64 frames using Chi-square analysis. Each of the four (4) primers, A1, B1, C1, and D1, were evaluated to characterize the different patterns for each frame. The white cells are the observed number of occurrences of each pattern. The peach cells are the expected number of occurrences under the null hypothesis that each of the six (6) patterns is equally likely. The Chi-square contribution for each pattern is in the corresponding purple cell. The totals correspond to the total number of observations and the total Chi-square. The p-value is the probability of observing a Chi-square greater than or equal to that observed under the null hypothesis of a uniform distribution with each cell having a relative frequency of 1/6 with 5 degrees of freedom.

Table A. Laboratory A data compiled for results obtained from Primer A1.

	A	B	C	D	E	F	TOTALS	
AAA	199.00	87.00	264.00	1.00	159.00	85.00	795.00	
Exp.	132.50	132.50	132.50	132.50	132.50	132.50		p-value
X^2	33.38	15.62	130.51	130.51	5.30	17.03	**332.34**	1.10576E-69
AAC	296.00	271.00	353.00	26.00	217.00	75.00	1238.00	
Exp.	206.33	206.33	206.33	206.33	206.33	206.33		p-value
X^2	38.97	20.27	104.25	157.61	0.55	83.60	**405.24**	2.19786E-85
ACA	174.00	330.00	414.00	380.00	56.00	204.00	1558.00	
Exp.	259.67	259.67	259.67	259.67	259.67	259.67		p-value
X^2	28.26	19.05	91.73	55.76	159.74	11.93	**366.48**	4.94025E-77
ACC	269.00	503.00	386.00	318.00	288.00	268.00	2032.00	
Exp.	338.67	338.67	338.67	338.67	338.67	338.67		p-value
X^2	14.33	79.74	6.62	1.26	7.58	14.75	**124.27**	3.90081E-25
ACT	445.00	58.00	98.00	104.00	112.00	2.00	819.00	
Exp.	136.50	136.50	136.50	136.50	136.50	136.50		p-value
X^2	697.23	45.14	10.86	7.74	4.40	132.53	**897.90**	7.5747E-192
ACG	1.00	98.00	0.00	11.00	0.00	1.00	111.00	
Exp.	18.50	18.50	18.50	18.50	18.50	18.50		p-value
X^2	16.55	341.64	18.50	3.04	18.50	16.55	**414.78**	1.9299E-87
AAT	13.00	120.00	81.00	66.00	162.00	110.00	552.00	
Exp.	92.00	92.00	92.00	92.00	92.00	92.00		p-value
X^2	67.84	8.52	1.32	7.35	53.26	3.52	**141.80**	7.39751E-29
ATA	185.00	94.00	148.00	5.00	445.00	18.00	895.00	
Exp.	149.17	149.17	149.17	149.17	149.17	149.17		p-value
X^2	8.61	20.40	0.01	139.33	586.71	115.34	**870.40**	6.7716E-186
ATT	249.00	199.00	26.00	70.00	238.00	82.00	864.00	
Exp.	144.00	144.00	144.00	144.00	144.00	144.00		p-value
X^2	76.56	21.01	96.69	38.03	61.36	26.69	**320.35**	4.21498E-67
ATC	93.00	222.00	53.00	231.00	13.00	170.00	782.00	
Exp.	130.33	130.33	130.33	130.33	130.33	130.33		p-value
X^2	10.69	64.47	45.89	77.75	105.63	12.07	**316.51**	2.82496E-66
ATG	193.00	0.00	267.00	0.00	3.00	0.00	463.00	
Exp.	77.17	77.17	77.17	77.17	77.17	77.17		p-value
X^2	173.88	77.17	467.00	77.17	71.28	77.17	**943.66**	9.4623E-202
AAG	73.00	176.00	26.00	219.00	54.00	61.00	609.00	
Exp.	101.50	101.50	101.50	101.50	101.50	101.50		p-value
X^2	8.00	54.68	56.16	136.02	22.23	16.16	**293.26**	2.82071E-61

AGA	1.00	0.00	143.00	0.00	0.00	0.00	144.00	
Exp.	24.00	24.00	24.00	24.00	24.00	24.00		p-value
X^2	22.04	24.00	590.04	24.00	24.00	24.00	**708.08**	8.7789E-151
AGG	0.00	0.00	46.00	0.00	101.00	0.00	147.00	
Exp.	24.50	24.50	24.50	24.50	24.50	24.50		p-value
X^2	24.50	24.50	18.87	24.50	238.87	24.50	**355.73**	1.01943E-74
AGC	96.00	32.00	281.00	17.00	240.00	92.00	758.00	
Exp.	126.33	126.33	126.33	126.33	126.33	126.33		p-value
X^2	7.28	70.44	189.35	94.62	102.27	9.33	**473.30**	4.6226E-100
AGT	75.00	0.00	186.00	1.00	305.00	9.00	576.00	
Exp.	96.00	96.00	96.00	96.00	96.00	96.00		p-value
X^2	4.59	96.00	84.38	94.01	455.01	78.84	**812.83**	1.936E-173
CAA	75.00	261.00	216.00	427.00	617.00	81.00	1677.00	
Exp.	279.50	279.50	279.50	279.50	279.50	279.50		p-value
X^2	149.63	1.22	14.43	77.84	407.54	140.97	**791.63**	7.4958E-169
CAC	126.00	481.00	30.00	270.00	213.00	222.00	1342.00	
Exp.	223.67	223.67	223.67	223.67	223.67	223.67		p-value
X^2	42.65	296.07	167.69	9.60	0.51	0.01	**516.52**	2.1631E-109
CCA	145.00	120.00	313.00	422.00	276.00	716.00	1992.00	
Exp.	332.00	332.00	332.00	332.00	332.00	332.00		p-value
X^2	105.33	135.37	1.09	24.40	9.45	444.14	**719.78**	2.599E-153
CCC	18.00	366.00	7.00	520.00	8.00	378.00	1297.00	
Exp.	216.17	216.17	216.17	216.17	216.17	216.17		p-value
X^2	181.67	103.86	202.39	427.05	200.46	121.16	**1236.59**	3.4899E-265
CCT	98.00	29.00	312.00	37.00	275.00	36.00	787.00	
Exp.	131.17	131.17	131.17	131.17	131.17	131.17		p-value
X^2	8.39	79.58	249.31	67.60	157.72	69.05	**631.65**	2.9344E-134
CCG	93.00	12.00	25.00	4.00	78.00	19.00	231.00	
Exp.	38.50	38.50	38.50	38.50	38.50	38.50		p-value
X^2	77.15	18.24	4.73	30.92	40.53	9.88	**181.44**	2.63361E-37
CAT	43.00	368.00	217.00	245.00	555.00	71.00	1499.00	
Exp.	249.83	249.83	249.83	249.83	249.83	249.83		p-value
X^2	171.23	55.89	4.31	0.09	372.76	128.01	**732.30**	5.091E-156
CTA	212.00	194.00	4.00	55.00	3.00	47.00	515.00	
Exp.	85.83	85.83	85.83	85.83	85.83	85.83		p-value
X^2	185.45	136.31	78.02	11.08	79.94	17.57	**508.37**	1.2478E-107
CTT	282.00	11.00	19.00	0.00	145.00	76.00	533.00	
Exp.	88.83	88.83	88.83	88.83	88.83	88.83		p-value
X^2	420.04	68.20	54.90	88.83	35.51	1.85	**669.33**	2.099E-142
CTC	1.00	175.00	2.00	528.00	3.00	46.00	755.00	
Exp.	125.83	125.83	125.83	125.83	125.83	125.83		p-value
X^2	123.84	19.21	121.87	1285.34	119.90	50.65	**1720.81**	4.074E-370
CTG	462.00	1.00	34.00	2.00	0.00	0.00	499.00	
Exp.	83.17	83.17	83.17	83.17	83.17	83.17		p-value
X^2	1725.63	81.18	29.07	79.21	83.17	83.17	**2081.42**	2.681E-448
CAG	0.00	0.00	0.00	176.00	50.00	327.00	553.00	
Exp.	92.17	92.17	92.17	92.17	92.17	92.17		p-value
X^2	92.17	92.17	92.17	76.25	19.29	598.34	**970.38**	1.552E-207
CGA	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
Exp.	N/A	N/A	N/A	N/A	N/A	N/A		p-value
X^2	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

CGG	1.00	0.00	110.00	1.00	1.00	0.00	113.00	
Exp.	18.83	18.83	18.83	18.83	18.83	18.83		p-value
X^2	16.89	18.83	441.31	16.89	16.89	18.83	**529.64**	3.1915E-112
CGC	0.00	1.00	1.00	96.00	0.00	16.00	114.00	
Exp.	19.00	19.00	19.00	19.00	19.00	19.00		p-value
X^2	19.00	17.05	17.05	312.05	19.00	0.47	**384.63**	6.08251E-81
CGT	86.00	15.00	93.00	89.00	5.00	17.00	305.00	
Exp.	50.83	50.83	50.83	50.83	50.83	50.83		p-value
X^2	24.33	25.26	34.98	28.66	41.33	22.52	**177.07**	2.26494E-36
TAA	6.00	82.00	6.00	376.00	16.00	26.00	512.00	
Exp.	85.33	85.33	85.33	85.33	85.33	85.33		p-value
X^2	73.76	0.13	73.76	990.08	56.33	41.26	**1235.31**	6.5893E-265
TAC	878.00	152.00	408.00	88.00	26.00	71.00	1623.00	
Exp.	270.50	270.50	270.50	270.50	270.50	270.50		p-value
X^2	1364.35	51.91	69.89	123.13	221.00	147.14	**1977.42**	9.513E-426
TCA	9.00	1.00	210.00	83.00	608.00	44.00	955.00	
Exp.	159.17	159.17	159.17	159.17	159.17	159.17		p-value
X^2	141.68	157.17	16.23	36.45	1265.66	83.33	**1700.52**	1.019E-365
TCC	28.00	84.00	175.00	29.00	259.00	16.00	591.00	
Exp.	98.50	98.50	98.50	98.50	98.50	98.50		p-value
X^2	50.46	2.13	59.41	49.04	261.53	69.10	**491.67**	5.014E-104
TCT	5.00	0.00	287.00	6.00	163.00	1.00	462.00	
Exp.	77.00	77.00	77.00	77.00	77.00	77.00		p-value
X^2	67.32	77.00	572.73	65.47	96.05	75.01	**953.58**	6.7143E-204
TCG	0.00	0.00	3.00	1.00	91.00	96.00	191.00	
Exp.	31.83	31.83	31.83	31.83	31.83	31.83		p-value
X^2	31.83	31.83	26.12	29.86	109.97	129.34	**358.96**	2.06179E-75
TAT	63.00	89.00	201.00	69.00	198.00	11.00	631.00	
Exp.	105.17	105.17	105.17	105.17	105.17	105.17		p-value
X^2	16.91	2.49	87.33	12.44	81.95	84.32	**285.42**	1.36177E-59
TTA	153.00	41.00	221.00	45.00	41.00	91.00	592.00	
Exp.	98.67	98.67	98.67	98.67	98.67	98.67		p-value
X^2	29.92	33.70	151.68	29.19	33.70	0.60	**278.79**	3.6208E-58
TTT	111.00	184.00	5.00	139.00	1.00	24.00	464.00	
Exp.	77.33	77.33	77.33	77.33	77.33	77.33		p-value
X^2	14.66	147.13	67.66	49.17	75.35	36.78	**390.74**	2.93475E-82
TTC	3.00	78.00	0.00	198.00	2.00	185.00	466.00	
Exp.	77.67	77.67	77.67	77.67	77.67	77.67		p-value
X^2	71.78	0.00	77.67	186.44	73.72	148.33	**557.94**	2.4656E-118
TTG	239.00	1.00	233.00	0.00	0.00	0.00	473.00	
Exp.	78.83	78.83	78.83	78.83	78.83	78.83		p-value
X^2	325.41	76.85	301.49	78.83	78.83	78.83	**940.25**	5.1754E-201
TAG	0.00	0.00	0.00	251.00	19.00	187.00	457.00	
Exp.	76.17	76.17	76.17	76.17	76.17	76.17		p-value
X^2	76.17	76.17	76.17	401.31	42.91	161.28	**834.00**	5.1027E-178
TGA	65.00	157.00	3.00	212.00	0.00	1.00	438.00	
Exp.	73.00	73.00	73.00	73.00	73.00	73.00		p-value
X^2	0.88	96.66	67.12	264.67	73.00	71.01	**573.34**	1.1614E-121
TGG	0.00	223.00	0.00	37.00	0.00	0.00	260.00	
Exp.	43.33	43.33	43.33	43.33	43.33	43.33		p-value
X^2	43.33	744.93	43.33	0.93	43.33	43.33	**919.18**	1.8747E-196
TGC	7.00	99.00	0.00	237.00	0.00	0.00	343.00	

Exp.	57.17	57.17	57.17	57.17	57.17	57.17		p-value
X^2	44.02	30.61	57.17	565.71	57.17	57.17	**811.85**	3.1577E-173
TGT	0.00	17.00	0.00	387.00	2.00	1.00	407.00	
Exp.	67.83	67.83	67.83	67.83	67.83	67.83		p-value
X^2	67.83	38.09	67.83	1501.73	63.89	65.85	**1805.23**	2.040E-388
GAA	90.00	5.00	3.00	4.00	49.00	58.00	209.00	
Exp.	34.83	34.83	34.83	34.83	34.83	34.83		p-value
X^2	87.37	25.55	29.09	27.29	5.76	15.41	**190.47**	3.09427E-39
GAC	0.00	1.00	2.00	58.00	142.00	116.00	319.00	
Exp.	53.17	53.17	53.17	53.17	53.17	53.17		p-value
X^2	53.17	51.19	49.24	0.44	148.43	74.26	**376.72**	3.08355E-79
GCA	142.00	108.00	37.00	94.00	90.00	104.00	575.00	
Exp.	95.83	95.83	95.83	95.83	95.83	95.83		p-value
X^2	22.24	1.54	36.12	0.04	0.36	0.70	**60.99**	7.58872E-12
GCC	209.00	66.00	31.00	4.00	115.00	1.00	426.00	
Exp.	71.00	71.00	71.00	71.00	71.00	71.00		p-value
X^2	268.23	0.35	22.54	63.23	27.27	69.01	**450.62**	3.61072E-95
GCT	0.00	0.00	32.00	7.00	78.00	100.00	217.00	
Exp.	36.17	36.17	36.17	36.17	36.17	36.17		p-value
X^2	36.17	36.17	0.48	23.52	48.39	112.66	**257.39**	1.42841E-53
GCG	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
Exp.	N/A	N/A	N/A	N/A	N/A	N/A		p-value
X^2	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
GAT	0.00	145.00	0.00	147.00	0.00	17.00	309.00	
Exp.	51.50	51.50	51.50	51.50	51.50	51.50		p-value
X^2	51.50	169.75	51.50	177.09	51.50	23.11	**524.46**	4.194E-111
GTA	329.00	173.00	49.00	176.00	266.00	229.00	1222.00	
Exp.	203.67	203.67	203.67	203.67	203.67	203.67		p-value
X^2	77.13	4.62	117.46	3.76	19.08	3.15	**225.19**	1.14924E-46
GTT	0.00	0.00	0.00	0.00	90.00	1.00	91.00	
Exp.	15.17	15.17	15.17	15.17	15.17	15.17		p-value
X^2	15.17	15.17	15.17	15.17	369.23	13.23	**443.13**	1.48831E-93
GTC	85.00	73.00	1.00	12.00	0.00	37.00	208.00	
Exp.	34.67	34.67	34.67	34.67	34.67	34.67		p-value
X^2	73.08	42.39	32.70	14.82	34.67	0.16	**197.81**	8.36294E-41
GTG	1.00	0.00	5.00	1.00	2.00	3.00	12.00	
Exp.	2.00	2.00	2.00	2.00	2.00	2.00		p-value
X^2	N/A	N/A	N/A	N/A	N/A	N/A	**N/A**	N/A
GAG	0.00	0.00	0.00	5.00	0.00	0.00	5.00	
Exp.	0.83	0.83	0.83	0.83	0.83	0.83		p-value
X^2	N/A	N/A	N/A	N/A	N/A	N/A	**N/A**	N/A
GGA	193.00	4.00	1.00	0.00	48.00	13.00	259.00	
Exp.	43.17	43.17	43.17	43.17	43.17	43.17		p-value
X^2	520.08	35.54	41.19	43.17	0.54	21.08	**661.59**	9.8681E-141
GGG	14.00	29.00	0.00	1.00	0.00	197.00	241.00	
Exp.	40.17	40.17	40.17	40.17	40.17	40.17		p-value
X^2	17.05	3.10	40.17	38.19	40.17	612.37	**751.04**	4.5026E-160
GGC	0.00	0.00	0.00	0.00	2.00	0.00	2.00	
Exp.	0.33	0.33	0.33	0.33	0.33	0.33		p-value
X^2	N/A	N/A	N/A	N/A	N/A	N/A	**N/A**	N/A
GGT	1.00	1.00	15.00	111.00	91.00	40.00	259.00	
Exp.	43.17	43.17	43.17	43.17	43.17	43.17		p-value

X^2	41.19	41.19	18.38	106.60	53.00	0.23	260.59	2.93217E-54

Table B. The p-values for each of the frames and each of the primers are provided. Any p-value < 7.8125E-4 (0.05/64) is considered statistically significant with an overall α = 0.05 for that frame. All p-values exceeding 7.8125E-4 are black reversed and justified; the null hypothesis is not to be rejected. Frames are not distributed uniformly random across the six (6) patterns. All N/A correspond to no observations for that frame.

Laboratory A

	p-value			
Frame	A1	B1	C1	D1
AAA	1.10576E-69	1.97613E-16	7.3393E-121	1.12829E-10
AAC	2.19786E-85	2.21986E-08	1.8946E-117	5.98581E-08
ACA	4.94025E-77	3.49871E-18	6.0921E-227	7.98679E-37
ACC	3.90081E-25	3.50665E-33	2.5634E-42	1.59481E-15
ACT	7.5747E-192	1.32776E-67	6.09762E-67	5.86839E-09
ACG	1.9299E-87	1.33474E-49	5.449E-153	4.71333E-08
AAT	7.39751E-29	8.498E-35	2.2824E-194	7.33427E-07
ATA	6.7716E-186	5.61172E-11	4.0275E-119	6.84257E-11
ATT	4.21498E-67	3.29153E-24	4.0832E-286	8.60482E-08
ATC	2.82496E-66	2.79912E-27	1.6738E-118	1.47321E-11
ATG	9.4623E-202	2.93389E-62	2.8625E-201	9.20277E-21
AAG	2.82071E-61	3.82785E-63	7.71945E-35	4.22105E-48
AGA	8.7789E-151	1.11052E-35	4.8407E-206	8.05448E-85
AGG	1.01943E-74	2.2631E-106	1.3894E-181	3.5831E-42
AGC	4.6226E-100	1.16187E-27	6.3073E-203	4.8062E-43
AGT	1.936E-173	2.79175E-96	9.97307E-67	1.52373E-35
CAA	7.4958E-169	1.90125E-06	1.9824E-119	3.7819E-13
CAC	2.1631E-109	N/A	3.4639E-109	0.000455274
CCA	2.599E-153	0.041940545	4.4458E-198	9.2056E-32
CCC	3.4899E-265	3.985E-29	3.4124E-218	3.63988E-05
CCT	2.9344E-134	0.000322535	6.49566E-54	2.05827E-09
CCG	2.63361E-37	1.05168E-24	6.86065E-43	9.49811E-07
CAT	5.091E-156	3.53958E-22	1.0019E-73	5.81374E-24
CTA	1.2478E-107	1.04622E-19	4.00094E-87	1.23556E-06
CTT	2.099E-142	1.33634E-56	1.57331E-90	6.95553E-18
CTC	4.074E-370	1.89699E-09	1.82576E-65	1.1002E-24
CTG	2.681E-448	6.54255E-38	1.6213E-223	2.34536E-24
CAG	1.552E-207	1.70036E-47	2.8300E-370	1.22861E-93
CGA	N/A	1.44346E-42	1.1667E-162	8.58593E-20
CGG	3.1915E-112	1.935E-08	2.81304E-48	6.04351E-32
CGC	6.08251E-81	N/A	4.44939E-71	1.14478E-09
CGT	2.26494E-36	9.19214E-17	1.54625E-91	2.88296E-13
TAA	6.5893E-265	1.68872E-31	6.0223E-238	6.16486E-08
TAC	9.513E-426	8.22043E-88	4.2511E-148	2.2579E-14
TCA	1.019e-365	8.54231E-19	2.9824E-217	1.08796E-22
TCC	5.014E-104	1.65804E-36	4.67588E-79	2.91618E-25
TCT	6.7143E-204	3.1875E-13	1.31091E-95	2.5745E-10
TCG	2.06179E-75	0.415880232	4.01349E-83	4.33358E-33
TAT	1.36177E-59	2.09234E-36	1.9361E-130	1.23074E-25
TTA	3.6208E-58	1.56213E-13	3.4131E-182	3.43921E-42
TTT	2.93475E-82	1.3947E-81	6.6018E-230	1.0613E-29
TTC	2.4656E-118	1.06789E-18	2.4707E-244	2.13543E-22
TTG	5.1754E-201	8.63259E-42	2.6052E-177	3.64868E-65
TAG	5.1027E-178	3.76452E-83	4.75852E-90	2.79722E-66
TGA	1.1614E-121	4.58887E-79	1.1671E-140	1.28784E-06
TGG	1.8747E-196	5.271E-137	6.65475E-85	4.48611E-84
TGC	3.1577E-173	7.34499E-14	6.9029E-218	1.69061E-32
TGT	2.040E-388	5.9162E-142	<1.0048E-499	5.1127E-112
GAA	3.09427E-39	6.61348E-35	1.1894E-100	3.04136E-24
GAC	3.08355E-79	1.08369E-54	6.5042E-125	1.31739E-51
GCA	7.58872E-12	1.7476E-13	2.90678E-57	2.21434E-13
GCC	3.61072E-95	N/A	4.5238E-85	1.06769E-15
GCT	1.42841E-53	3.3107E-21	3.79508E-79	0.000609876
GCG	N/A	4.48654E-27	2.216E-119	6.1411E-12
GAT	4.194E-111	1.31856E-59	1.5307E-107	1.78362E-30
GTA	1.14924E-46	4.17879E-51	5.65438E-59	7.79196E-20
GTT	1.48831E-93	1.05207E-35	5.02487E-90	4.17155E-38

GTC	8.36294E-41	4.37348E-12	8.3542E-171	2.26187E-16
GTG	N/A	1.15875E-58	8.8089E-104	5.51488E-43
GAG	N/A	1.5569E-121	1.56937E-88	1.11912E-14
GGA	9.8681E-141	1.49611E-24	8.47539E-40	0.033005783
GGG	4.5026E-160	8.68089E-75	1.62366E-61	3.58748E-25
GGC	N/A	1.21759E-56	2.96643E-71	1.6383E-11
GGT	2.93217E-54	3.99124E-22	1.06385E-85	1.56661E-30

Laboratory B

	p-value			
Frame	A1	B1	C1	D1
AAA	2.4015E-19	5.66898E-41	5.94755E-39	4.27164E-07
AAC	1.47379E-40	4.66112E-21	1.20691E-67	1.41557E-08
ACA	5.81954E-28	1.07704E-31	1.9407E-70	2.62816E-46
ACC	4.5398E-08	4.15116E-55	4.08269E-18	2.22103E-32
ACT	1.99246E-88	4.5485E-97	2.35893E-18	2.83706E-12
ACG	7.72488E-40	6.43297E-41	8.07106E-43	2.37484E-07
AAT	1.57917E-10	2.29438E-54	6.33869E-56	1.81591E-14
ATA	3.08042E-42	2.18098E-11	1.91897E-21	1.10885E-19
ATT	1.42826E-49	9.18013E-20	2.39761E-61	4.76423E-09
ATC	4.51011E-41	3.05379E-35	8.81307E-56	1.23928E-12
ATG	1.98679E-77	4.09257E-32	2.90208E-57	7.84509E-24
AAG	2.33415E-25	9.3961E-61	0.01188574	1.7849E-48
AGA	8.42606E-64	2.09949E-45	3.99923E-82	5.5188E-100
AGG	9.59874E-25	3.2893E-137	5.23552E-41	1.18655E-41
AGC	2.92461E-30	5.01139E-34	1.75198E-42	1.23857E-47
AGT	4.72293E-47	1.5905E-149	2.26133E-15	6.06819E-39
CAA	9.77873E-50	3.16667E-16	8.02374E-32	8.94217E-18
CAC	1.08711E-28	N/A	3.10063E-27	3.51349E-12
CCA	4.58102E-35	3.27143E-08	1.5628E-46	2.26276E-31
CCC	4.4424E-111	1.02667E-16	1.09157E-62	1.24518E-07
CCT	1.27408E-58	2.31622E-08	6.53436E-21	1.35044E-30
CCG	1.72067E-22	8.05057E-30	1.68456E-15	0.000978458
CAT	2.77056E-44	2.27913E-37	1.0623E-24	3.06673E-34
CTA	2.99906E-42	1.40316E-08	5.59292E-13	7.5511E-12
CTT	4.10768E-42	3.46188E-91	2.27655E-11	2.0618E-44
CTC	2.2497E-117	0.415880232	5.28302E-13	4.73764E-40
CTG	2.103E-150	4.04892E-43	8.80036E-65	1.20559E-33
CAG	9.45356E-65	1.19815E-55	2.592E-101	1.21807E-87
CGA	N/A	2.33428E-27	2.95234E-42	2.87206E-26
CGG	4.35835E-35	9.75139E-07	5.45893E-20	2.21247E-35
CGC	2.55874E-30	N/A	9.26587E-19	1.31116E-14
CGT	3.97319E-25	1.87527E-23	1.21979E-16	4.45073E-07
TAA	2.06078E-64	3.90672E-39	5.5774E-55	1.91901E-05
TAC	9.5631E-174	2.29264E-85	2.48687E-48	1.40833E-24
TCA	1.2769E-273	7.14309E-21	4.42298E-39	7.07389E-25
TCC	2.21154E-30	5.90659E-63	1.18035E-13	2.94546E-14
TCT	2.05411E-48	1.15443E-20	1.30145E-33	4.08338E-36
TCG	5.76146E-60	N/A	7.24764E-29	2.26474E-45
TAT	2.15884E-38	7.51926E-78	1.18758E-51	1.26286E-21
TTA	1.72504E-47	1.89386E-33	2.09194E-57	8.29455E-51
TTT	1.07355E-32	3.1084E-163	7.04152E-45	5.66438E-36
TTC	5.81169E-38	4.22246E-45	5.3141E-65	1.05579E-18
TTG	5.90393E-84	1.43254E-68	5.30689E-65	8.99449E-72
TAG	1.51082E-57	1.8319E-98	4.408E-43	1.78966E-69
TGA	5.65214E-44	1.1231E-130	8.36594E-69	6.69995E-09
TGG	1.05602E-42	3.9485E-158	2.22348E-29	5.38629E-98
TGC	7.56442E-60	5.28634E-27	2.26696E-53	1.88197E-54
TGT	1.164E-140	5.8607E-244	2.5967E-187	7.3112E-157
GAA	2.40547E-20	1.43418E-51	3.45816E-26	4.96895E-28
GAC	2.36184E-29	8.7428E-50	9.84828E-33	1.29791E-70
GCA	0.020564676	7.24901E-15	5.24932E-13	5.5385E-15
GCC	3.34127E-56	N/A	1.11361E-23	3.59097E-23
GCT	1.08487E-26	3.77338E-35	3.81025E-25	7.69814E-24
GCG	N/A	4.20543E-24	8.17806E-31	1.4237E-10
GAT	2.12048E-44	1.98722E-58	1.94885E-34	1.24928E-65
GTA	4.86965E-19	3.02542E-81	2.0574E-18	4.31008E-07
GTT	4.09733E-22	5.73242E-46	1.77089E-19	9.06237E-88
GTC	2.99389E-06	1.83455E-23	1.93039E-40	7.57482E-51
GTG	N/A	5.33664E-93	3.01566E-34	8.51628E-76
GAG	N/A	2.61428E-91	1.6611E-16	6.39698E-20
GGA	7.96391E-45	2.13174E-26	7.5511E-12	1.76546E-08
GGG	4.40048E-44	2.74661E-65	2.4203E-15	1.05373E-44
GGC	N/A	2.03016E-96	6.27076E-24	2.85272E-14
GGT	5.36344E-22	4.60842E-40	4.17504E-31	5.84254E-89

Appendix 3

The following table is an example of the processed data using the bioinformatic tools to statistically evaluate each of the 64 frames using a 2 x 6 Chi-square analysis comparing the results from Laboratory A to Laboratory B. For those frames with expected values less than that required in more than 25% of the cells, a Kolmogorov-Smirnov test was performed. The tables are not presented here; the p-values calculated with the Kolmogorov-Smirnov test statistic were not different from the Chi-square.

These results demonstrate that the data obtained from the two (2) laboratories are comparable. Each of the four (4) primers, A1, B1, C1, and D1, were evaluated to characterize the different patterns for each frame.

Table A. 2 x 6 Chi-square analysis comparing the results from Laboratory A to Laboratory B for Primer A1.

	A	B	C	D	E	F	TOTAL	
AAA	199.00	87.00	264.00	1.00	159.00	85.00	795.00	
AAA	71.00	31.00	76.00	0.00	58.00	29.00	265.00	
TOTAL	270.00	118.00	340.00	1.00	217.00	114.00	1060.00	
Exp.	202.50	88.50	255.00	0.75	162.75	85.50		p-value
Exp.	67.50	29.50	85.00	0.25	54.25	28.50		0.805545483
X^2	0.06	0.03	0.32	0.08	0.09	0.00	**0.58**	
X^2	0.18	0.08	0.95	0.25	0.26	0.01	**1.73**	
AAC	296.00	271.00	353.00	26.00	217.00	75.00	1238.00	
AAC	92.00	74.00	152.00	15.00	76.00	11.00	420.00	
TOTAL	388.00	345.00	505.00	41.00	293.00	86.00	1658.00	
Exp.	289.71	257.61	377.07	30.61	218.78	64.21		p-value
Exp.	98.29	87.39	127.93	10.39	74.22	21.79		0.001683344
X^2	0.14	0.70	1.54	0.70	0.01	1.81	**4.89**	
X^2	0.40	2.05	4.53	2.05	0.04	5.34	**14.42**	
ACA	174.00	330.00	414.00	380.00	56.00	204.00	1558.00	
ACA	87.00	134.00	114.00	131.00	8.00	55.00	529.00	
TOTAL	261.00	464.00	528.00	511.00	64.00	259.00	2087.00	
Exp.	194.84	346.39	394.17	381.47	47.78	193.35		p-value
Exp.	66.16	117.61	133.83	129.53	16.22	65.65		0.000246578
X^2	2.23	0.78	1.00	0.01	1.42	0.59	**6.01**	
X^2	6.57	2.28	2.94	0.02	4.17	1.73	**17.70**	
ACC	269.00	503.00	386.00	318.00	288.00	268.00	2032.00	
ACC	126.00	142.00	153.00	81.00	128.00	77.00	707.00	
TOTAL	395.00	645.00	539.00	399.00	416.00	345.00	2739.00	
Exp.	293.04	478.51	399.87	296.01	308.62	255.95		p-value
Exp.	101.96	166.49	139.13	102.99	107.38	89.05		3.28633E-05
X^2	1.97	1.25	0.48	1.63	1.38	0.57	**7.29**	
X^2	5.67	3.60	1.38	4.70	3.96	1.63	**20.94**	
ACT	445.00	58.00	98.00	104.00	112.00	2.00	819.00	
ACT	172.00	26.00	8.00	37.00	42.00	0.00	285.00	
TOTAL	617.00	84.00	106.00	141.00	154.00	2.00	1104.00	
Exp.	457.72	62.32	78.64	104.60	114.24	1.48		p-value
Exp.	159.28	21.68	27.36	36.40	39.76	0.52		0.000552211
X^2	0.35	0.30	4.77	0.00	0.04	0.18	**5.65**	
X^2	1.02	0.86	13.70	0.01	0.13	0.52	**16.23**	
ACG	1.00	98.00	0.00	11.00	0.00	1.00	111.00	
ACG	1.00	40.00	0.00	0.00	0.00	0.00	41.00	
TOTAL	2.00	138.00	0.00	11.00	0.00	1.00	152.00	
Exp.	1.46	100.78	0.25	8.03	0.25	0.73		p-value
Exp.	0.54	37.22	0.25	2.97	0.25	0.27		0.385636227
X^2	0.15	0.08	0.25	1.10	0.25	0.10	**1.92**	

X^2	0.39	0.21	0.25	2.97	0.25	0.27	4.34	

AAT	13.00	120.00	81.00	66.00	162.00	110.00	552.00	
AAT	5.00	42.00	46.00	31.00	54.00	16.00	194.00	
TOTAL	18.00	162.00	127.00	97.00	216.00	126.00	746.00	
Exp.	13.32	119.87	93.97	71.77	159.83	93.23		p-value
Exp.	4.68	42.13	33.03	25.23	56.17	32.77		0.001045754
X^2	0.01	0.00	1.79	0.46	0.03	3.02	5.31	
X^2	0.02	0.00	5.10	1.32	0.08	8.58	15.10	

ATA	185.00	94.00	148.00	5.00	445.00	18.00	895.00	
ATA	68.00	34.00	49.00	5.00	139.00	27.00	322.00	
TOTAL	253.00	128.00	197.00	10.00	584.00	45.00	1217.00	
Exp.	186.06	94.13	144.88	7.35	429.48	33.09		p-value
Exp.	66.94	33.87	52.12	2.65	154.52	11.91		8.30891E-06
X^2	0.01	0.00	0.07	0.75	0.56	6.88	8.27	
X^2	0.02	0.00	0.19	2.09	1.56	19.13	22.99	

ATT	249.00	199.00	26.00	70.00	238.00	82.00	864.00	
ATT	23.00	99.00	6.00	32.00	7.00	113.00	280.00	
TOTAL	272.00	298.00	32.00	102.00	245.00	195.00	1144.00	
Exp.	205.43	225.06	24.17	77.03	185.03	147.27		p-value
Exp.	66.57	72.94	7.83	24.97	59.97	47.73		1.97104E-48
X^2	9.24	3.02	0.14	0.64	15.16	28.93	57.13	
X^2	28.52	9.31	0.43	1.98	46.78	89.27	176.29	

ATC	93.00	222.00	53.00	231.00	13.00	170.00	782.00	
ATC	38.00	74.00	7.00	118.00	0.00	56.00	293.00	
TOTAL	131.00	296.00	60.00	349.00	13.00	226.00	1075.00	
Exp.	95.29	215.32	43.65	253.88	9.46	164.40		p-value
Exp.	35.71	80.68	16.35	95.12	3.54	61.60		0.000665719
X^2	0.06	0.21	2.00	2.06	1.33	0.19	5.85	
X^2	0.15	0.55	5.35	5.50	3.54	0.51	15.60	

ATG	193.00	0.00	267.00	0.00	3.00	0.00	463.00	
ATG	86.00	1.00	102.00	0.00	2.00	0.00	191.00	
TOTAL	279.00	1.00	369.00	0.00	5.00	0.00	654.00	
Exp.	197.52	0.71	261.23	0.25	3.54	0.25		p-value
Exp.	81.48	0.29	107.77	0.25	1.46	0.25		0.624053924
X^2	0.10	0.71	0.13	0.25	0.08	0.25	1.52	
X^2	0.25	1.72	0.31	0.25	0.20	0.25	2.97	

AAG	73.00	176.00	26.00	219.00	54.00	61.00	609.00	
AAG	25.00	63.00	6.00	75.00	44.00	2.00	215.00	
TOTAL	98.00	239.00	32.00	294.00	98.00	63.00	824.00	
Exp.	72.43	176.64	23.65	217.29	72.43	46.56		p-value
Exp.	25.57	62.36	8.35	76.71	25.57	16.44		9.05752E-07
X^2	0.00	0.00	0.23	0.01	4.69	4.48	9.42	
X^2	0.01	0.01	0.66	0.04	13.28	12.68	26.68	

AGA	1.00	0.00	143.00	0.00	0.00	0.00	144.00	
AGA	0.00	0.00	61.00	0.00	0.00	0.00	61.00	
TOTAL	1.00	0.00	204.00	0.00	0.00	0.00	205.00	
Exp.	0.70	0.25	143.30	0.25	0.25	0.25		p-value
Exp.	0.30	0.25	60.70	0.25	0.25	0.25		0.994592443
X^2	0.13	0.25	0.00	0.25	0.25	0.25	1.13	
X^2	0.30	0.25	0.00	0.25	0.25	0.25	1.30	

AGG	0.00	0.00	46.00	0.00	101.00	0.00	147.00	
AGG	1.00	1.00	24.00	0.00	37.00	0.00	63.00	
TOTAL	1.00	1.00	70.00	0.00	138.00	0.00	210.00	
Exp.	0.70	0.70	49.00	0.25	96.60	0.25		p-value
Exp.	0.30	0.30	21.00	0.25	41.40	0.25		0.311414387
X^2	0.70	0.70	0.18	0.25	0.20	0.25	2.28	
X^2	1.63	1.63	0.43	0.25	0.47	0.25	4.66	

AGC	96.00	32.00	281.00	17.00	240.00	92.00	758.00	
AGC	47.00	3.00	91.00	6.00	71.00	32.00	250.00	
TOTAL	143.00	35.00	372.00	23.00	311.00	124.00	1008.00	
Exp.	107.53	26.32	279.74	17.30	233.87	93.25		p-value

Exp.	35.47	8.68	92.26	5.70	77.13	30.75		0.057880438
X^2	1.24	1.23	0.01	0.01	0.16	0.02	2.65	
X^2	3.75	3.72	0.02	0.02	0.49	0.05	8.04	
AGT	75.00	0.00	186.00	1.00	305.00	9.00	576.00	
AGT	56.00	4.00	28.00	1.00	105.00	15.00	209.00	
TOTAL	131.00	4.00	214.00	2.00	410.00	24.00	785.00	
Exp.	96.12	2.94	157.02	1.47	300.84	17.61		p-value
Exp.	34.88	1.06	56.98	0.53	109.16	6.39		1.05449E-12
X^2	4.64	2.94	5.35	0.15	0.06	4.21	17.34	
X^2	12.79	8.09	14.74	0.41	0.16	11.60	47.79	
CAA	75.00	261.00	216.00	427.00	617.00	81.00	1677.00	
CAA	29.00	110.00	72.00	152.00	202.00	31.00	596.00	
TOTAL	104.00	371.00	288.00	579.00	819.00	112.00	2273.00	
Exp.	76.73	273.72	212.48	427.18	604.25	82.63		p-value
Exp.	27.27	97.28	75.52	151.82	214.75	29.37		0.582330595
X^2	0.04	0.59	0.06	0.00	0.27	0.03	0.99	
X^2	0.11	1.66	0.16	0.00	0.76	0.09	2.78	
CAC	126.00	481.00	30.00	270.00	213.00	222.00	1342.00	
CAC	50.00	143.00	5.00	92.00	75.00	80.00	445.00	
TOTAL	176.00	624.00	35.00	362.00	288.00	302.00	1787.00	
Exp.	132.17	468.61	26.28	271.85	216.28	226.80		p-value
Exp.	43.83	155.39	8.72	90.15	71.72	75.20		0.387273169
X^2	0.29	0.33	0.53	0.01	0.05	0.10	1.30	
X^2	0.87	0.99	1.58	0.04	0.15	0.31	3.94	
CCA	145.00	120.00	313.00	422.00	276.00	716.00	1992.00	
CCA	81.00	47.00	90.00	152.00	93.00	221.00	684.00	
TOTAL	226.00	167.00	403.00	574.00	369.00	937.00	2676.00	
Exp.	168.23	124.31	299.99	427.28	274.68	697.50		p-value
Exp.	57.77	42.69	103.01	146.72	94.32	239.50		0.003573017
X^2	3.21	0.15	0.56	0.07	0.01	0.49	4.48	
X^2	9.34	0.44	1.64	0.19	0.02	1.43	13.06	
CCC	18.00	366.00	7.00	520.00	8.00	378.00	1297.00	
CCC	6.00	150.00	0.00	211.00	1.00	93.00	461.00	
TOTAL	24.00	516.00	7.00	731.00	9.00	471.00	1758.00	
Exp.	17.71	380.69	5.16	539.31	6.64	347.49		p-value
Exp.	6.29	135.31	1.84	191.69	2.36	123.51		0.002298114
X^2	0.00	0.57	0.65	0.69	0.28	2.68	4.87	
X^2	0.01	1.59	1.84	1.95	0.78	7.54	13.71	
CCT	98.00	29.00	312.00	37.00	275.00	36.00	787.00	
CCT	41.00	20.00	127.00	7.00	108.00	3.00	306.00	
TOTAL	139.00	49.00	439.00	44.00	383.00	39.00	1093.00	
Exp.	100.09	35.28	316.10	31.68	275.77	28.08		p-value
Exp.	38.91	13.72	122.90	12.32	107.23	10.92		0.008383757
X^2	0.04	1.12	0.05	0.89	0.00	2.23	4.34	
X^2	0.11	2.88	0.14	2.30	0.01	5.74	11.17	
CCG	93.00	12.00	25.00	4.00	78.00	19.00	231.00	
CCG	40.00	1.00	6.00	1.00	21.00	0.00	69.00	
TOTAL	133.00	13.00	31.00	5.00	99.00	19.00	300.00	
Exp.	102.41	10.01	23.87	3.85	76.23	14.63		p-value
Exp.	30.59	2.99	7.13	1.15	22.77	4.37		0.040836662
X^2	0.86	0.40	0.05	0.01	0.04	1.31	2.67	
X^2	2.89	1.32	0.18	0.02	0.14	4.37	8.93	
CAT	43.00	368.00	217.00	245.00	555.00	71.00	1499.00	
CAT	48.00	144.00	72.00	63.00	184.00	23.00	534.00	
TOTAL	91.00	512.00	289.00	308.00	739.00	94.00	2033.00	
Exp.	67.10	377.52	213.09	227.10	544.89	69.31		p-value
Exp.	23.90	134.48	75.91	80.90	194.11	24.69		1.25314E-07
X^2	8.65	0.24	0.07	1.41	0.19	0.04	10.61	
X^2	24.29	0.67	0.20	3.96	0.53	0.12	29.77	
CTA	212.00	194.00	4.00	55.00	3.00	47.00	515.00	
CTA	82.00	52.00	1.00	27.00	1.00	2.00	165.00	

TOTAL	294.00	246.00	5.00	82.00	4.00	49.00	680.00	
Exp.	222.66	186.31	3.79	62.10	3.03	37.11		p-value
Exp.	71.34	59.69	1.21	19.90	0.97	11.89		0.003385636
X^2	0.51	0.32	0.01	0.81	0.00	2.64	4.29	
X^2	1.59	0.99	0.04	2.54	0.00	8.23	13.38	

CTT	282.00	11.00	19.00	0.00	145.00	76.00	533.00	
CTT	98.00	8.00	4.00	3.00	58.00	43.00	214.00	
TOTAL	380.00	19.00	23.00	3.00	203.00	119.00	747.00	
Exp.	271.14	13.56	16.41	2.14	144.84	84.91		p-value
Exp.	108.86	5.44	6.59	0.86	58.16	34.09		0.008916897
X^2	0.44	0.48	0.41	2.14	0.00	0.93	4.40	
X^2	1.08	1.20	1.02	5.33	0.00	2.33	10.96	

CTC	1.00	175.00	2.00	528.00	3.00	46.00	755.00	
CTC	0.00	62.00	4.00	174.00	0.00	15.00	255.00	
TOTAL	1.00	237.00	6.00	702.00	3.00	61.00	1010.00	
Exp.	0.75	177.16	4.49	524.76	2.24	45.60		p-value
Exp.	0.25	59.84	1.51	177.24	0.76	15.40		0.220444513
X^2	0.09	0.03	1.38	0.02	0.26	0.00	1.77	
X^2	0.25	0.08	4.08	0.06	0.76	0.01	5.23	

CTG	462.00	1.00	34.00	2.00	0.00	0.00	499.00	
CTG	144.00	0.00	2.00	0.00	0.00	0.00	146.00	
TOTAL	606.00	1.00	36.00	2.00	0.00	0.00	645.00	
Exp.	468.83	0.77	27.85	1.55	0.25	0.25		p-value
Exp.	137.17	0.25	8.15	0.45	0.25	0.25		0.214183285
X^2	0.10	0.07	1.36	0.13	0.25	0.25	2.16	
X^2	0.34	0.25	4.64	0.45	0.25	0.25	6.18	

CAG	0.00	0.00	0.00	176.00	50.00	327.00	553.00	
CAG	0.00	0.00	1.00	60.00	27.00	113.00	201.00	
TOTAL	0.00	0.00	1.00	236.00	77.00	440.00	754.00	
Exp.	0.25	0.25	0.73	173.09	56.47	322.71		p-value
Exp.	0.25	0.25	0.27	62.91	20.53	117.29		0.312788018
X^2	0.25	0.25	0.73	0.05	0.74	0.06	2.08	
X^2	0.25	0.25	2.02	0.13	2.04	0.16	4.85	

CGA	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
CGA	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
TOTAL	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
Exp.	N/A	N/A	N/A	N/A	N/A	N/A		p-value
Exp.	N/A	N/A	N/A	N/A	N/A	N/A		N/A
X^2	N/A	N/A	N/A	N/A	N/A	N/A	N/A	
X^2	N/A	N/A	N/A	N/A	N/A	N/A	N/A	

CGG	1.00	0.00	110.00	1.00	1.00	0.00	113.00	
CGG	1.00	0.00	38.00	0.00	2.00	0.00	41.00	
TOTAL	2.00	0.00	148.00	1.00	3.00	0.00	154.00	
Exp.	1.47	0.25	108.60	0.73	2.20	0.25		p-value
Exp.	0.53	0.25	39.40	0.27	0.80	0.25		0.630549383
X^2	0.15	0.25	0.02	0.10	0.66	0.25	1.42	
X^2	0.41	0.25	0.05	0.27	1.81	0.25	3.03	

CGC	0.00	1.00	1.00	96.00	0.00	16.00	114.00	
CGC	10.00	1.00	1.00	42.00	0.00	1.00	55.00	
TOTAL	10.00	2.00	2.00	138.00	0.00	17.00	169.00	
Exp.	6.75	1.35	1.35	93.09	0.25	11.47		p-value
Exp.	3.25	0.65	0.65	44.91	0.25	5.53		5.53538E-05
X^2	6.75	0.09	0.09	0.09	0.25	1.79	9.06	
X^2	13.98	0.19	0.19	0.19	0.25	3.71	18.51	

CGT	86.00	15.00	93.00	89.00	5.00	17.00	305.00	
CGT	47.00	5.00	5.00	30.00	0.00	2.00	89.00	
TOTAL	133.00	20.00	98.00	119.00	5.00	19.00	394.00	
Exp.	102.96	15.48	75.86	92.12	3.87	14.71		p-value
Exp.	30.04	4.52	22.14	26.88	1.13	4.29		3.63629E-06
X^2	2.79	0.02	3.87	0.11	0.33	0.36	7.47	
X^2	9.57	0.05	13.27	0.36	1.13	1.22	25.60	

TAA	6.00	82.00	6.00	376.00	16.00	26.00	512.00	281+1
TAA	1.00	12.00	0.00	102.00	3.00	36.00	154.00	
TOTAL	7.00	94.00	6.00	478.00	19.00	62.00	666.00	
Exp.	5.38	72.26	4.61	367.47	14.61	47.66		p-value
Exp.	1.62	21.74	1.39	110.53	4.39	14.34		5.93505E-10
X^2	0.07	1.31	0.42	0.20	0.13	9.85	11.98	
X^2	0.24	4.36	1.39	0.66	0.44	32.74	39.82	

TAC	878.00	152.00	408.00	88.00	26.00	71.00	1623.00	
TAC	335.00	59.00	130.00	37.00	9.00	8.00	578.00	
TOTAL	1213.00	211.00	538.00	125.00	35.00	79.00	2201.00	
Exp.	894.46	155.59	396.72	92.17	25.81	58.25		p-value
Exp.	318.54	55.41	141.28	32.83	9.19	20.75		0.015386859
X^2	0.30	0.08	0.32	0.19	0.00	2.79	3.69	
X^2	0.85	0.23	0.90	0.53	0.00	7.83	10.35	

TCA	9.00	1.00	210.00	83.00	608.00	44.00	955.00	
TCA	1.00	3.00	19.00	26.00	320.00	7.00	376.00	
TOTAL	10.00	4.00	229.00	109.00	928.00	51.00	1331.00	
Exp.	7.18	2.87	164.31	78.21	665.85	36.59		p-value
Exp.	2.82	1.13	64.69	30.79	262.15	14.41		8.99632E-15
X^2	0.46	1.22	12.71	0.29	5.03	1.50	21.21	
X^2	1.18	3.09	32.27	0.75	12.76	3.81	53.86	

TCC	28.00	84.00	175.00	29.00	259.00	16.00	591.00	
TCC	6.00	29.00	73.00	16.00	73.00	5.00	202.00	
TOTAL	34.00	113.00	248.00	45.00	332.00	21.00	793.00	
Exp.	25.34	84.22	184.83	33.54	247.43	15.65		p-value
Exp.	8.66	28.78	63.17	11.46	84.57	5.35		0.172686462
X^2	0.28	0.00	0.52	0.61	0.54	0.01	1.97	
X^2	0.82	0.00	1.53	1.80	1.58	0.02	5.75	

TCT	5.00	0.00	287.00	6.00	163.00	1.00	462.00	
TCT	1.00	0.00	68.00	0.00	36.00	0.00	105.00	
TOTAL	6.00	0.00	355.00	6.00	199.00	1.00	567.00	
Exp.	4.89	0.25	289.26	4.89	162.15	0.81		p-value
Exp.	1.11	0.25	65.74	1.11	36.85	0.25		0.908547618
X^2	0.00	0.25	0.02	0.25	0.00	0.04	0.57	
X^2	0.01	0.25	0.08	1.11	0.02	0.25	1.72	

TCG	0.00	0.00	3.00	1.00	91.00	96.00	191.00	
TCG	1.00	0.00	1.00	0.00	67.00	6.00	75.00	
TOTAL	1.00	0.00	4.00	1.00	158.00	102.00	266.00	
Exp.	0.72	0.25	2.87	0.72	113.45	73.24		p-value
Exp.	0.28	0.25	1.13	0.28	44.55	28.76		2.5424E-08
X^2	0.72	0.25	0.01	0.11	4.44	7.07	12.60	
X^2	1.83	0.25	0.01	0.28	11.31	18.01	31.70	

TAT	63.00	89.00	201.00	69.00	198.00	11.00	631.00	
TAT	36.00	35.00	110.00	29.00	17.00	1.00	228.00	
TOTAL	99.00	124.00	311.00	98.00	215.00	12.00	859.00	
Exp.	72.72	91.09	228.45	71.99	157.93	8.81		p-value
Exp.	26.28	32.91	82.55	26.01	57.07	3.19		2.71471E-11
X^2	1.30	0.05	3.30	0.12	10.16	0.54	15.48	
X^2	3.60	0.13	9.13	0.34	28.13	1.50	42.83	

TTA	153.00	41.00	221.00	45.00	41.00	91.00	592.00	
TTA	21.00	9.00	108.00	5.00	22.00	26.00	191.00	
TOTAL	174.00	50.00	329.00	50.00	63.00	117.00	783.00	
Exp.	131.56	37.80	248.75	37.80	47.63	88.46		p-value
Exp.	42.44	12.20	80.25	12.20	15.37	28.54		4.08815E-07
X^2	3.50	0.27	3.09	1.37	0.92	0.07	9.23	
X^2	10.83	0.84	9.59	4.25	2.86	0.23	28.60	

TTT	111.00	184.00	5.00	139.00	1.00	24.00	464.00	
TTT	20.00	11.00	1.00	24.00	7.00	76.00	139.00	
TOTAL	131.00	195.00	6.00	163.00	8.00	100.00	603.00	
Exp.	100.80	150.05	4.62	125.43	6.16	76.95		p-value
Exp.	30.20	44.95	1.38	37.57	1.84	23.05		8.64189E-46
X^2	1.03	7.68	0.03	1.47	4.32	36.43	50.97	

X^2	3.44	25.64	0.11	4.90	14.42	121.62	170.13	

TTC	3.00	78.00	0.00	198.00	2.00	185.00	466.00	
TTC	1.00	24.00	0.00	67.00	0.00	56.00	148.00	
TOTAL	4.00	102.00	0.00	265.00	2.00	241.00	614.00	
Exp.	3.04	77.41	0.25	201.12	1.52	182.91		p-value
Exp.	0.96	24.59	0.25	63.88	0.48	58.09		0.966048267
X^2	0.00	0.00	0.25	0.05	0.15	0.02	0.48	
X^2	0.00	0.01	0.25	0.15	0.48	0.08	0.98	

TTG	239.00	1.00	233.00	0.00	0.00	0.00	473.00	
TTG	67.00	0.00	114.00	0.00	0.00	0.00	181.00	
TOTAL	306.00	1.00	347.00	0.00	0.00	0.00	654.00	
Exp.	221.31	0.72	250.96	0.25	0.25	0.25		p-value
Exp.	84.69	0.28	96.04	0.25	0.25	0.25		0.071437393
X^2	1.41	0.11	1.29	0.25	0.25	0.25	3.56	
X^2	3.69	0.28	3.36	0.25	0.25	0.25	8.08	

TAG	0.00	0.00	0.00	251.00	19.00	187.00	457.00	
TAG	0.00	0.00	0.00	92.00	14.00	58.00	164.00	
TOTAL	0.00	0.00	0.00	343.00	33.00	245.00	621.00	
Exp.	0.25	0.25	0.25	252.42	24.29	180.30		p-value
Exp.	0.25	0.25	0.25	90.58	8.71	64.70		0.377102578
X^2	0.25	0.25	0.25	0.01	1.15	0.25	2.16	
X^2	0.25	0.25	0.25	0.02	3.20	0.69	4.67	

TGA	65.00	157.00	3.00	212.00	0.00	1.00	438.00	
TGA	28.00	74.00	0.00	67.00	0.00	0.00	169.00	
TOTAL	93.00	231.00	3.00	279.00	0.00	1.00	607.00	
Exp.	67.11	166.69	2.16	201.32	0.25	0.72		p-value
Exp.	25.89	64.31	0.84	77.68	0.25	0.28		0.322424272
X^2	0.07	0.56	0.32	0.57	0.25	0.11	1.88	
X^2	0.17	1.46	0.84	1.47	0.25	0.28	4.46	

TGG	0.00	223.00	0.00	37.00	0.00	0.00	260.00	
TGG	0.00	59.00	0.00	29.00	0.00	0.00	88.00	
TOTAL	0.00	282.00	0.00	66.00	0.00	0.00	348.00	
Exp.	0.25	210.69	0.25	49.31	0.25	0.25		p-value
Exp.	0.25	71.31	0.25	16.69	0.25	0.25		0.010371514
X^2	0.25	0.72	0.25	3.07	0.25	0.25	4.79	
X^2	0.25	2.13	0.25	9.08	0.25	0.25	12.21	

TGC	7.00	99.00	0.00	237.00	0.00	0.00	343.00	
TGC	4.00	34.00	0.00	85.00	0.00	0.00	123.00	
TOTAL	11.00	133.00	0.00	322.00	0.00	0.00	466.00	
Exp.	8.10	97.89	0.25	237.01	0.25	0.25		p-value
Exp.	2.90	35.11	0.25	84.99	0.25	0.25		0.987543905
X^2	0.15	0.01	0.25	0.00	0.25	0.25	0.91	
X^2	0.41	0.03	0.25	0.00	0.25	0.25	1.20	

TGT	0.00	17.00	0.00	387.00	2.00	1.00	407.00	
TGT	0.00	1.00	0.00	135.00	0.00	1.00	137.00	
TOTAL	0.00	18.00	0.00	522.00	2.00	2.00	544.00	
Exp.	0.25	13.47	0.25	390.54	1.50	1.50		p-value
Exp.	0.25	4.53	0.25	131.46	0.50	0.50		0.399630721
X^2	0.25	0.93	0.25	0.03	0.17	0.16	1.79	
X^2	0.25	2.75	0.25	0.10	0.50	0.49	4.34	

GAA	90.00	5.00	3.00	4.00	49.00	58.00	209.00	
GAA	35.00	0.00	0.00	0.00	29.00	10.00	74.00	
TOTAL	125.00	5.00	3.00	4.00	78.00	68.00	283.00	
Exp.	92.31	3.69	2.22	2.95	57.60	50.22		p-value
Exp.	32.69	1.31	0.78	1.05	20.40	17.78		0.015633318
X^2	0.06	0.46	0.28	0.37	1.29	1.21	3.66	
X^2	0.16	1.31	0.78	1.05	3.63	3.40	10.34	

GAC	0.00	1.00	2.00	58.00	142.00	116.00	319.00	
GAC	0.00	1.00	0.00	26.00	57.00	41.00	125.00	
TOTAL	0.00	2.00	2.00	84.00	199.00	157.00	444.00	
Exp.	0.25	1.44	1.44	60.35	142.98	112.80		p-value

Exp.	0.25	0.56	0.56	23.65	56.02	44.20		0.859124216
X^2	0.25	0.13	0.22	0.09	0.01	0.09	0.79	
X^2	0.25	0.34	0.56	0.23	0.02	0.23	1.63	

GCA	142.00	108.00	37.00	94.00	90.00	104.00	575.00	
GCA	33.00	37.00	13.00	36.00	33.00	36.00	188.00	
TOTAL	175.00	145.00	50.00	130.00	123.00	140.00	763.00	
Exp.	131.88	109.27	37.68	97.97	92.69	105.50		p-value
Exp.	43.12	35.73	12.32	32.03	30.31	34.50		0.504545459
X^2	0.78	0.01	0.01	0.16	0.08	0.02	1.06	
X^2	2.37	0.05	0.04	0.49	0.24	0.07	3.25	

GCC	209.00	66.00	31.00	4.00	115.00	1.00	426.00	
GCC	97.00	16.00	3.00	0.00	41.00	2.00	159.00	
TOTAL	306.00	82.00	34.00	4.00	156.00	3.00	585.00	
Exp.	222.83	59.71	24.76	2.91	113.60	2.18		p-value
Exp.	83.17	22.29	9.24	1.09	42.40	0.82		0.009146875
X^2	0.86	0.66	1.57	0.41	0.02	0.64	4.16	
X^2	2.30	1.77	4.21	1.09	0.05	1.72	11.14	

GCT	0.00	0.00	32.00	7.00	78.00	100.00	217.00	
GCT	0.00	0.00	1.00	4.00	40.00	34.00	79.00	
TOTAL	0.00	0.00	33.00	11.00	118.00	134.00	296.00	
Exp.	0.25	0.25	24.19	8.06	86.51	98.24		p-value
Exp.	0.25	0.25	8.81	2.94	31.49	35.76		0.021404569
X^2	0.25	0.25	2.52	0.14	0.84	0.03	4.03	
X^2	0.25	0.25	6.92	0.39	2.30	0.09	10.19	

GCG	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
GCG	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
TOTAL	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
Exp.	N/A	N/A	N/A	N/A	N/A	N/A		p-value
Exp.	N/A	N/A	N/A	N/A	N/A	N/A		N/A
X^2	N/A	N/A	N/A	N/A	N/A	N/A	N/A	
X^2	N/A	N/A	N/A	N/A	N/A	N/A	N/A	

GAT	0.00	145.00	0.00	147.00	0.00	17.00	309.00	
GAT	0.00	63.00	0.00	54.00	0.00	6.00	123.00	
TOTAL	0.00	208.00	0.00	201.00	0.00	23.00	432.00	
Exp.	0.25	148.78	0.25	143.77	0.25	16.45		p-value
Exp.	0.25	59.22	0.25	57.23	0.25	6.55		0.985301073
X^2	0.25	0.10	0.25	0.07	0.25	0.02	0.94	
X^2	0.25	0.24	0.25	0.18	0.25	0.05	1.22	

GTA	329.00	173.00	49.00	176.00	266.00	229.00	1222.00	
GTA	124.00	43.00	24.00	61.00	98.00	92.00	442.00	
TOTAL	453.00	216.00	73.00	237.00	364.00	321.00	1664.00	
Exp.	332.67	158.63	53.61	174.05	267.31	235.73		p-value
Exp.	120.33	57.38	19.39	62.95	96.69	85.27		0.19388313
X^2	0.04	1.30	0.40	0.02	0.01	0.19	1.96	
X^2	0.11	3.60	1.10	0.06	0.02	0.53	5.42	

GTT	0.00	0.00	0.00	0.00	90.00	1.00	91.00	
GTT	0.00	0.00	0.00	0.00	22.00	0.00	22.00	
TOTAL	0.00	0.00	0.00	0.00	112.00	1.00	113.00	
Exp.	0.25	0.25	0.25	0.25	90.19	0.81		p-value
Exp.	0.25	0.25	0.25	0.25	21.81	0.25		0.999971899
X^2	0.25	0.25	0.25	0.25	0.00	0.05	1.05	
X^2	0.25	0.25	0.25	0.25	0.00	0.25	1.25	

GTC	85.00	73.00	1.00	12.00	0.00	37.00	208.00	
GTC	22.00	10.00	9.00	4.00	3.00	24.00	72.00	
TOTAL	107.00	83.00	10.00	16.00	3.00	61.00	280.00	
Exp.	79.49	61.66	7.43	11.89	2.23	45.31		p-value
Exp.	27.51	21.34	2.57	4.11	0.77	15.69		9.78592E-09
X^2	0.38	2.09	5.56	0.00	2.23	1.53	11.79	
X^2	1.11	6.03	16.07	0.00	6.44	4.41	34.05	

GTG	1.00	0.00	5.00	1.00	2.00	3.00	12.00
GTG	0.00	0.00	0.00	0.00	0.00	0.00	0.00

TOTAL	1.00	0.00	5.00	1.00	2.00	3.00	12.00	
Exp.	N/A	N/A	N/A	N/A	N/A	N/A		p-value
Exp.	N/A	N/A	N/A	N/A	N/A	N/A		N/A
X^2	N/A	N/A	N/A	N/A	N/A	N/A	**N/A**	
X^2	N/A	N/A	N/A	N/A	N/A	N/A	**N/A**	

GAG	0.00	0.00	0.00	5.00	0.00	0.00	5.00	
GAG	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
TOTAL	0.00	0.00	0.00	5.00	0.00	0.00	5.00	
Exp.	N/A	N/A	N/A	N/A	N/A	N/A		p-value
Exp.	N/A	N/A	N/A	N/A	N/A	N/A		N/A
X^2	N/A	N/A	N/A	N/A	N/A	N/A	**N/A**	
X^2	N/A	N/A	N/A	N/A	N/A	N/A	**N/A**	

GGA	193.00	4.00	1.00	0.00	48.00	13.00	259.00	
GGA	66.00	1.00	0.00	0.00	22.00	5.00	94.00	
TOTAL	259.00	5.00	1.00	0.00	70.00	18.00	353.00	
Exp.	190.03	3.67	0.73	0.25	51.36	13.21		p-value
Exp.	68.97	1.33	0.27	0.25	18.64	4.79		0.914555051
X^2	0.05	0.03	0.10	0.25	0.22	0.00	**0.65**	
X^2	0.13	0.08	0.27	0.25	0.61	0.01	**1.34**	

GGG	14.00	29.00	0.00	1.00	0.00	197.00	241.00	
GGG	1.00	30.00	0.00	0.00	0.00	62.00	93.00	
TOTAL	15.00	59.00	0.00	1.00	0.00	259.00	334.00	
Exp.	10.82	42.57	0.25	0.72	0.25	186.88		p-value
Exp.	4.18	16.43	0.25	0.28	0.25	72.12		0.000729675
X^2	0.93	4.33	0.25	0.11	0.25	0.55	**6.41**	
X^2	2.42	11.21	0.25	0.28	0.25	1.42	**15.83**	

GGC	0.00	0.00	0.00	0.00	2.00	0.00	2.00	
GGC	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
TOTAL	0.00	0.00	0.00	0.00	2.00	0.00	2.00	
Exp.	N/A	N/A	N/A	N/A	N/A	N/A		p-value
Exp.	N/A	N/A	N/A	N/A	N/A	N/A		N/A
X^2	N/A	N/A	N/A	N/A	N/A	N/A	**N/A**	
X^2	N/A	N/A	N/A	N/A	N/A	N/A	**N/A**	

GGT	1.00	1.00	15.00	111.00	91.00	40.00	259.00	
GGT	0.00	0.00	2.00	41.00	21.00	39.00	103.00	
TOTAL	1.00	1.00	17.00	152.00	112.00	79.00	362.00	
Exp.	0.72	0.72	12.16	108.75	80.13	56.52		p-value
Exp.	0.28	0.28	4.84	43.25	31.87	22.48		0.000114644
X^2	0.11	0.11	0.66	0.05	1.47	4.83	**7.24**	
X^2	0.28	0.28	1.66	0.12	3.71	12.14	**18.20**	

Table B. The p-values for the comparison between Laboratory A and Laboratory B for each frame and each of the primers are provided. Any p-value < 7.8125E-4 (0.05/64) is considered statistically significant with an overall α = 0.05 for that frame given the null hypothesis that the distribution of patterns is the same. All p-values exceeding 7.8125E-4 are black reversed; the null hypothesis is not to be rejected. All N/A correspond to no observations for that frame.

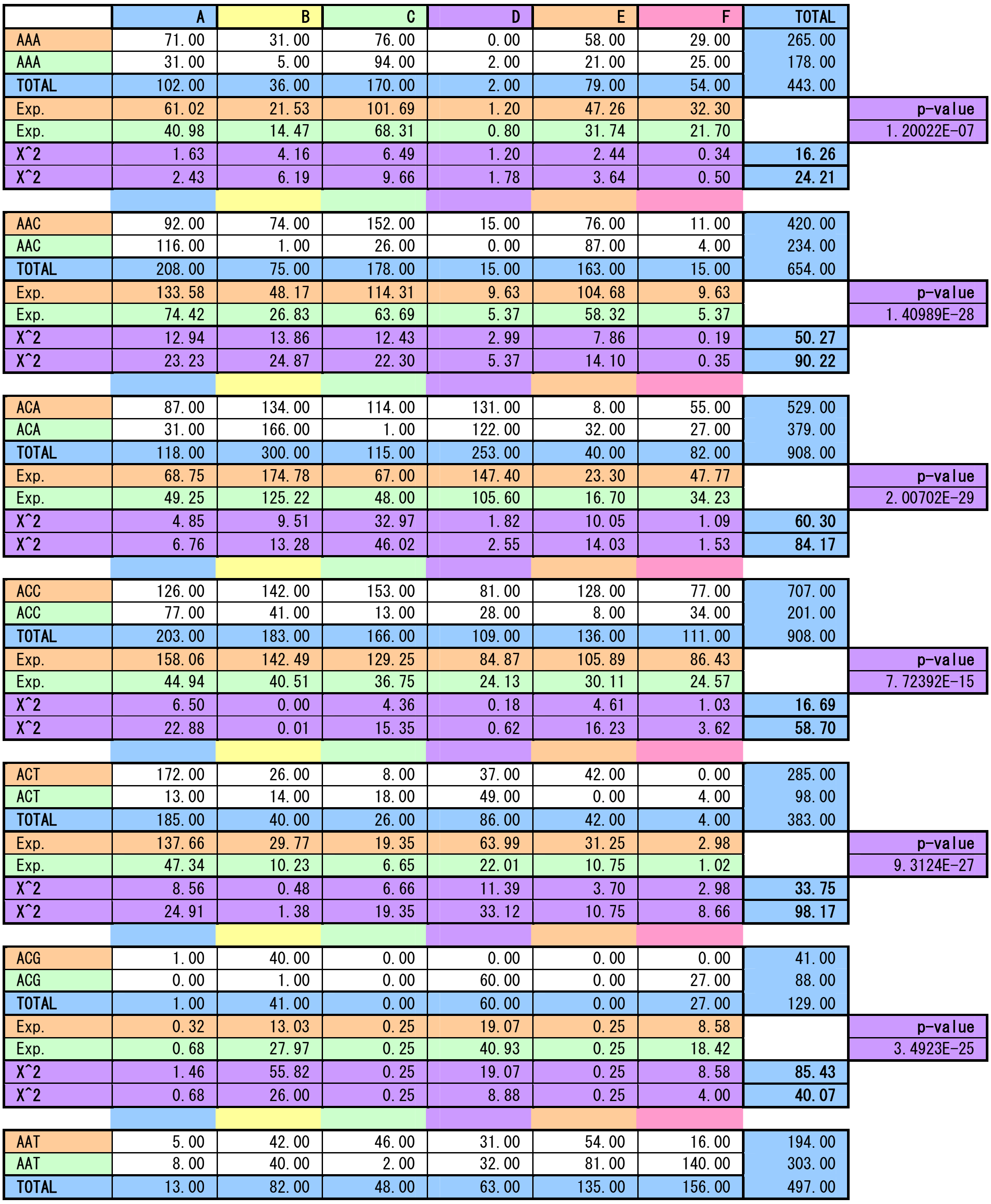
	p-value			
Frame	A1	B1	C1	D1
AAA	0.805545483	0.021028387	0.15527684	0.041154751
AAC	0.001683344	0.853187729	7.8697E-08	0.173902987
ACA	0.000246578	0.190484114	0.37142707	0.211463948
ACC	3.28633E-05	0.056276928	5.38378E-11	0.787496426
ACT	0.000552211	0.048671752	0.349082834	0.915116185
ACG	0.385636227	0.914056021	0.667375479	0.104641466
AAT	0.001045754	0.001146891	0.707583754	0.193459763
ATA	8.30891E-06	0.072346356	0.000309195	1.18322E-07
ATT	1.97104E-48	3.44426E-13	3.5702E-22	2.28456E-06
ATC	0.000665719	0.014231848	3.38954E-09	0.000634051
ATG	0.624053924	0.00031568	0.798578346	1.46476E-06
AAG	9.05752E-07	0.046995814	1.44323E-05	0.56795021
AGA	0.994592443	0.181656286	0.029574993	0.774247395
AGG	0.311414387	3.13192E-05	0.897599027	0.918656398
AGC	0.057880438	0.799939444	0.178391345	0.634620207
AGT	1.05449E-12	0.007445644	6.34677E-14	0.800142041
CAA	0.582330595	0.008446692	0.813047819	0.988619265
CAC	0.387273169	N/A	0.110483163	0.424859978
CCA	0.003573017	3.82091E-05	8.09899E-12	0.921384507
CCC	0.002298114	0.287932574	0.342186048	0.986800968
CCT	0.008383757	2.43351E-08	5.5433E-05	0.021033552
CCG	0.040836662	0.99929933	0.61175165	0.085796601
CAT	1.25314E-07	0.003129742	1.25507E-08	0.983603337
CTA	0.003385636	5.24768E-06	1.57037E-06	0.203329534
CTT	0.008916897	0.001110402	1.37616E-10	5.0813E-06
CTC	0.220444513	0.162783656	0.163109874	0.063484446
CTG	0.214183285	0.053631141	0.626347062	3.58232E-06
CAG	0.312788018	0.781062	0.79481545	0.254238506
CGA	N/A	0.142909021	0.039815033	0.817066519
CGG	0.630549383	0.00139454	3.89586E-15	0.850444487
CGC	5.53538E-05	N/A	0.457782718	0.857987337
CGT	3.63629E-06	0.749491785	0.000302216	0.000155967
TAA	5.93505E-10	0.297752907	0.003310072	0.297105881
TAC	0.015386859	1.5669E-05	0.001140743	0.966998964
TCA	8.99632E-15	0.727390027	2.93957E-08	0.96389772
TCC	0.172686462	1.68345E-06	7.64322E-17	0.044870703
TCT	0.908547618	0.00299369	3.80678E-10	1.50691E-09
TCG	2.5424E-08	N/A	0.736008048	0.836821379
TAT	2.71471E-11	3.07764E-05	1.52526E-17	2.03963E-11
TTA	4.08815E-07	0.082337003	0.020072433	0.291726754
TTT	8.64189E-46	0.000881087	2.50269E-36	0.074728074
TTC	0.966048267	1.2737E-05	2.12592E-27	0.047191801
TTG	0.071437393	0.002322039	0.000666125	0.002146119
TAG	0.377102578	0.108353468	0.000398975	0.999233686
TGA	0.322424272	0.087144226	0.099409151	0.192119423
TGG	0.010371514	2.17339E-06	5.88995E-05	0.435826546
TGC	0.987543905	0.001843282	0.62181951	0.234658592
TGT	0.399630721	1.89014E-06	0.01555385	3.82337E-06
GAA	0.015633318	0.264539721	0.000577676	0.990734443
GAC	0.859124216	0.036105683	0.855986024	0.99986273
GCA	0.504545459	0.236261123	0.38989923	0.242388731
GCC	0.009146875	N/A	0.128223065	0.35637003
GCT	0.021404569	0.011618732	0.407221429	3.46422E-10
GCG	N/A	0.255054645	0.211557491	0.637831307
GAT	0.985301073	0.008431232	0.411397128	0.007719256
GTA	0.19388313	0.009689778	1.17453E-28	0.000158686
GTT	0.999971899	0.080287238	0.00187769	1.43503E-11
GTC	9.78592E-09	0.000734651	0.416546495	0.000952275
GTG	N/A	0.008175723	4.60639E-05	0.001591149
GAG	N/A	0.001022539	3.30067E-09	0.104400885

GGA	0.914555051	0.003192933	1.44908E-05	0.000128548
GGG	0.000729675	0.224596561	0.816549782	0.00248417
GGC	N/A	0.062701476	0.171480285	2.51252E-05
GGT	0.000114644	0.203993288	0.180395055	1.02674E-10

Appendix 4

The following table is an example of the processed data using the bioinformatic tools to statistically evaluate each of the 64 frames using a 2 x 6 Chi-square analysis comparing the results from data from different primers produced in the same laboratory. These results demonstrate that the data obtained from the different primers are not comparable. Comparisons were made between Primers A1 and C1; Primers B1 and D1; and Primers A1 and B1.

Table A. 2 x 6 Chi-square analysis comparing the results from Primer A1 to Primer C1.



	A	B	C	D	E	F	TOTAL	
AAA	71. 00	31. 00	76. 00	0. 00	58. 00	29. 00	265. 00	
AAA	31. 00	5. 00	94. 00	2. 00	21. 00	25. 00	178. 00	
TOTAL	102. 00	36. 00	170. 00	2. 00	79. 00	54. 00	443. 00	
Exp.	61. 02	21. 53	101. 69	1. 20	47. 26	32. 30		p-value
Exp.	40. 98	14. 47	68. 31	0. 80	31. 74	21. 70		1. 20022E-07
X^2	1. 63	4. 16	6. 49	1. 20	2. 44	0. 34	**16. 26**	
X^2	2. 43	6. 19	9. 66	1. 78	3. 64	0. 50	**24. 21**	
AAC	92. 00	74. 00	152. 00	15. 00	76. 00	11. 00	420. 00	
AAC	116. 00	1. 00	26. 00	0. 00	87. 00	4. 00	234. 00	
TOTAL	208. 00	75. 00	178. 00	15. 00	163. 00	15. 00	654. 00	
Exp.	133. 58	48. 17	114. 31	9. 63	104. 68	9. 63		p-value
Exp.	74. 42	26. 83	63. 69	5. 37	58. 32	5. 37		1. 40989E-28
X^2	12. 94	13. 86	12. 43	2. 99	7. 86	0. 19	**50. 27**	
X^2	23. 23	24. 87	22. 30	5. 37	14. 10	0. 35	**90. 22**	
ACA	87. 00	134. 00	114. 00	131. 00	8. 00	55. 00	529. 00	
ACA	31. 00	166. 00	1. 00	122. 00	32. 00	27. 00	379. 00	
TOTAL	118. 00	300. 00	115. 00	253. 00	40. 00	82. 00	908. 00	
Exp.	68. 75	174. 78	67. 00	147. 40	23. 30	47. 77		p-value
Exp.	49. 25	125. 22	48. 00	105. 60	16. 70	34. 23		2. 00702E-29
X^2	4. 85	9. 51	32. 97	1. 82	10. 05	1. 09	**60. 30**	
X^2	6. 76	13. 28	46. 02	2. 55	14. 03	1. 53	**84. 17**	
ACC	126. 00	142. 00	153. 00	81. 00	128. 00	77. 00	707. 00	
ACC	77. 00	41. 00	13. 00	28. 00	8. 00	34. 00	201. 00	
TOTAL	203. 00	183. 00	166. 00	109. 00	136. 00	111. 00	908. 00	
Exp.	158. 06	142. 49	129. 25	84. 87	105. 89	86. 43		p-value
Exp.	44. 94	40. 51	36. 75	24. 13	30. 11	24. 57		7. 72392E-15
X^2	6. 50	0. 00	4. 36	0. 18	4. 61	1. 03	**16. 69**	
X^2	22. 88	0. 01	15. 35	0. 62	16. 23	3. 62	**58. 70**	
ACT	172. 00	26. 00	8. 00	37. 00	42. 00	0. 00	285. 00	
ACT	13. 00	14. 00	18. 00	49. 00	0. 00	4. 00	98. 00	
TOTAL	185. 00	40. 00	26. 00	86. 00	42. 00	4. 00	383. 00	
Exp.	137. 66	29. 77	19. 35	63. 99	31. 25	2. 98		p-value
Exp.	47. 34	10. 23	6. 65	22. 01	10. 75	1. 02		9. 3124E-27
X^2	8. 56	0. 48	6. 66	11. 39	3. 70	2. 98	**33. 75**	
X^2	24. 91	1. 38	19. 35	33. 12	10. 75	8. 66	**98. 17**	
ACG	1. 00	40. 00	0. 00	0. 00	0. 00	0. 00	41. 00	
ACG	0. 00	1. 00	0. 00	60. 00	0. 00	27. 00	88. 00	
TOTAL	1. 00	41. 00	0. 00	60. 00	0. 00	27. 00	129. 00	
Exp.	0. 32	13. 03	0. 25	19. 07	0. 25	8. 58		p-value
Exp.	0. 68	27. 97	0. 25	40. 93	0. 25	18. 42		3. 4923E-25
X^2	1. 46	55. 82	0. 25	19. 07	0. 25	8. 58	**85. 43**	
X^2	0. 68	26. 00	0. 25	8. 88	0. 25	4. 00	**40. 07**	
AAT	5. 00	42. 00	46. 00	31. 00	54. 00	16. 00	194. 00	
AAT	8. 00	40. 00	2. 00	32. 00	81. 00	140. 00	303. 00	
TOTAL	13. 00	82. 00	48. 00	63. 00	135. 00	156. 00	497. 00	

Exp.	5.07	32.01	18.74	24.59	52.70	60.89		p-value
Exp.	7.93	49.99	29.26	38.41	82.30	95.11		9.02911E-26
X^2	0.00	3.12	39.67	1.67	0.03	33.10	77.59	
X^2	0.00	2.00	25.40	1.07	0.02	21.19	49.68	
ATA	68.00	34.00	49.00	5.00	139.00	27.00	322.00	
ATA	70.00	10.00	71.00	11.00	34.00	22.00	218.00	
TOTAL	138.00	44.00	120.00	16.00	173.00	49.00	540.00	
Exp.	82.29	26.24	71.56	9.54	103.16	29.22		p-value
Exp.	55.71	17.76	48.44	6.46	69.84	19.78		6.74572E-13
X^2	2.48	2.30	7.11	2.16	12.45	0.17	26.67	
X^2	3.66	3.39	10.50	3.19	18.39	0.25	39.39	
ATT	23.00	99.00	6.00	32.00	7.00	113.00	280.00	
ATT	30.00	10.00	53.00	4.00	156.00	96.00	349.00	
TOTAL	53.00	109.00	59.00	36.00	163.00	209.00	629.00	
Exp.	23.59	48.52	26.26	16.03	72.56	93.04		p-value
Exp.	29.41	60.48	32.74	19.97	90.44	115.96		1.99252E-55
X^2	0.01	52.51	15.63	15.92	59.23	4.28	147.61	
X^2	0.01	42.13	12.54	12.78	47.52	3.44	118.42	
ATC	38.00	74.00	7.00	118.00	0.00	56.00	293.00	
ATC	1.00	6.00	0.00	81.00	0.00	62.00	150.00	
TOTAL	39.00	80.00	7.00	199.00	0.00	118.00	443.00	
Exp.	25.79	52.91	4.63	131.62	0.25	78.05		p-value
Exp.	13.21	27.09	2.37	67.38	0.25	39.95		2.65293E-13
X^2	5.78	8.40	1.21	1.41	0.25	6.23	23.28	
X^2	11.28	16.42	2.37	2.75	0.25	12.16	45.23	
ATG	86.00	1.00	102.00	0.00	2.00	0.00	191.00	
ATG	79.00	1.00	35.00	0.00	0.00	0.00	115.00	
TOTAL	165.00	2.00	137.00	0.00	2.00	0.00	306.00	
Exp.	102.99	1.25	85.51	0.25	1.25	0.25		p-value
Exp.	62.01	0.75	51.49	0.25	0.75	0.25		0.004046262
X^2	2.80	0.05	3.18	0.25	0.45	0.25	6.98	
X^2	4.66	0.08	5.28	0.25	0.75	0.25	11.27	
AAG	25.00	63.00	6.00	75.00	44.00	2.00	215.00	
AAG	8.00	4.00	13.00	17.00	4.00	8.00	54.00	
TOTAL	33.00	67.00	19.00	92.00	48.00	10.00	269.00	
Exp.	26.38	53.55	15.19	73.53	38.36	7.99		p-value
Exp.	6.62	13.45	3.81	18.47	9.64	2.01		2.91673E-12
X^2	0.07	1.67	5.56	0.03	0.83	4.49	12.65	
X^2	0.29	6.64	22.12	0.12	3.30	17.89	50.35	
AGA	0.00	0.00	61.00	0.00	0.00	0.00	61.00	
AGA	9.00	0.00	91.00	0.00	0.00	2.00	102.00	
TOTAL	9.00	0.00	152.00	0.00	0.00	2.00	163.00	
Exp.	3.37	0.25	56.88	0.25	0.25	0.75		p-value
Exp.	5.63	0.25	95.12	0.25	0.25	1.25		0.216618383
X^2	3.37	0.25	0.30	0.25	0.25	0.75	5.16	
X^2	2.01	0.25	0.18	0.25	0.25	0.45	3.39	
AGG	1.00	1.00	24.00	0.00	37.00	0.00	63.00	
AGG	9.00	0.00	49.00	0.00	0.00	0.00	58.00	
TOTAL	10.00	1.00	73.00	0.00	37.00	0.00	121.00	
Exp.	5.21	0.52	38.01	0.25	19.26	0.25		p-value
Exp.	4.79	0.48	34.99	0.25	17.74	0.25		3.61857E-10
X^2	3.40	0.44	5.16	0.25	16.33	0.25	25.83	
X^2	3.69	0.48	5.61	0.25	17.74	0.25	28.01	
AGC	47.00	3.00	91.00	6.00	71.00	32.00	250.00	
AGC	0.00	0.00	84.00	2.00	37.00	42.00	165.00	
TOTAL	47.00	3.00	175.00	8.00	108.00	74.00	415.00	
Exp.	28.31	1.81	105.42	4.82	65.06	44.58		p-value
Exp.	18.69	1.19	69.58	3.18	42.94	29.42		2.24005E-09
X^2	12.33	0.79	1.97	0.29	0.54	3.55	19.47	
X^2	18.69	1.19	2.99	0.44	0.82	5.38	29.51	
AGT	56.00	4.00	28.00	1.00	105.00	15.00	209.00	

AGT	2.00	10.00	6.00	12.00	37.00	1.00	68.00	
TOTAL	58.00	14.00	34.00	13.00	142.00	16.00	277.00	
Exp.	43.76	10.56	25.65	9.81	107.14	12.07		p-value
Exp.	14.24	3.44	8.35	3.19	34.86	3.93		4.9304E-13
X^2	3.42	4.08	0.21	7.91	0.04	0.71	16.38	
X^2	10.52	12.53	0.66	24.31	0.13	2.18	50.34	
CAA	29.00	110.00	72.00	152.00	202.00	31.00	596.00	
CAA	0.00	51.00	0.00	13.00	68.00	64.00	196.00	
TOTAL	29.00	161.00	72.00	165.00	270.00	95.00	792.00	
Exp.	21.82	121.16	54.18	124.17	203.18	71.49		p-value
Exp.	7.18	39.84	17.82	40.83	66.82	23.51		1.00751E-31
X^2	2.36	1.03	5.86	6.24	0.01	22.93	38.43	
X^2	7.18	3.12	17.82	18.97	0.02	69.73	116.84	
CAC	50.00	143.00	5.00	92.00	75.00	80.00	445.00	
CAC	49.00	45.00	85.00	0.00	98.00	24.00	301.00	
TOTAL	99.00	188.00	90.00	92.00	173.00	104.00	746.00	
Exp.	59.05	112.14	53.69	54.88	103.20	62.04		p-value
Exp.	39.95	75.86	36.31	37.12	69.80	41.96		2.70263E-47
X^2	1.39	8.49	44.15	25.11	7.70	5.20	92.04	
X^2	2.05	12.55	65.27	37.12	11.39	7.69	136.08	
CCA	81.00	47.00	90.00	152.00	93.00	221.00	684.00	
CCA	7.00	0.00	34.00	108.00	16.00	42.00	207.00	
TOTAL	88.00	47.00	124.00	260.00	109.00	263.00	891.00	
Exp.	67.56	36.08	95.19	199.60	83.68	201.90		p-value
Exp.	20.44	10.92	28.81	60.40	25.32	61.10		1.71467E-17
X^2	2.68	3.30	0.28	11.35	1.04	1.81	20.46	
X^2	8.84	10.92	0.94	37.50	3.43	5.97	67.60	
CCC	6.00	150.00	0.00	211.00	1.00	93.00	461.00	
CCC	0.00	5.00	0.00	91.00	0.00	42.00	138.00	
TOTAL	6.00	155.00	0.00	302.00	1.00	135.00	599.00	
Exp.	4.62	119.29	0.25	232.42	0.77	103.90		p-value
Exp.	1.38	35.71	0.25	69.58	0.25	31.10		1.58567E-09
X^2	0.41	7.91	0.25	1.97	0.07	1.14	11.76	
X^2	1.38	26.41	0.25	6.60	0.25	3.82	38.71	
CCT	41.00	20.00	127.00	7.00	108.00	3.00	306.00	
CCT	15.00	36.00	6.00	15.00	39.00	76.00	187.00	
TOTAL	56.00	56.00	133.00	22.00	147.00	79.00	493.00	
Exp.	34.76	34.76	82.55	13.66	91.24	49.03		p-value
Exp.	21.24	21.24	50.45	8.34	55.76	29.97		4.30325E-44
X^2	1.12	6.27	23.93	3.24	3.08	43.22	80.86	
X^2	1.83	10.25	39.16	5.31	5.04	70.72	132.32	
CCG	40.00	1.00	6.00	1.00	21.00	0.00	69.00	
CCG	3.00	32.00	16.00	0.00	2.00	32.00	85.00	
TOTAL	43.00	33.00	22.00	1.00	23.00	32.00	154.00	
Exp.	19.27	14.79	9.86	0.45	10.31	14.34		p-value
Exp.	23.73	18.21	12.14	0.55	12.69	17.66		6.55286E-23
X^2	22.31	12.85	1.51	0.68	11.10	14.34	62.79	
X^2	18.11	10.43	1.23	0.55	9.01	11.64	50.97	
CAT	48.00	144.00	72.00	63.00	184.00	23.00	534.00	
CAT	36.00	84.00	21.00	3.00	66.00	21.00	231.00	
TOTAL	84.00	228.00	93.00	66.00	250.00	44.00	765.00	
Exp.	58.64	159.15	64.92	46.07	174.51	30.71		p-value
Exp.	25.36	68.85	28.08	19.93	75.49	13.29		4.77169E-08
X^2	1.93	1.44	0.77	6.22	0.52	1.94	12.82	
X^2	4.46	3.34	1.79	14.38	1.19	4.48	29.63	
CTA	82.00	52.00	1.00	27.00	1.00	2.00	165.00	
CTA	20.00	2.00	40.00	16.00	45.00	9.00	132.00	
TOTAL	102.00	54.00	41.00	43.00	46.00	11.00	297.00	
Exp.	56.67	30.00	22.78	23.89	25.56	6.11		p-value
Exp.	45.33	24.00	18.22	19.11	20.44	4.89		1.28144E-34
X^2	11.33	16.13	20.82	0.41	23.59	2.77	75.05	
X^2	14.16	20.17	26.03	0.51	29.49	3.46	93.81	

CTT	98.00	8.00	4.00	3.00	58.00	43.00	214.00	
CTT	56.00	47.00	24.00	0.00	35.00	32.00	194.00	
TOTAL	154.00	55.00	28.00	3.00	93.00	75.00	408.00	
Exp.	80.77	28.85	14.69	1.57	48.78	39.34		p-value
Exp.	73.23	26.15	13.31	1.43	44.22	35.66		3.10181E-12
X^2	3.67	15.07	7.78	1.29	1.74	0.34	**29.89**	
X^2	4.05	16.62	8.58	1.43	1.92	0.38	**32.97**	

CTC	0.00	62.00	4.00	174.00	0.00	15.00	255.00	
CTC	0.00	12.00	1.00	24.00	0.00	24.00	61.00	
TOTAL	0.00	74.00	5.00	198.00	0.00	39.00	316.00	
Exp.	0.25	59.72	4.03	159.78	0.25	31.47		p-value
Exp.	0.25	14.28	0.97	38.22	0.25	7.53		6.30261E-10
X^2	0.25	0.09	0.00	1.27	0.25	8.62	**10.47**	
X^2	0.25	0.37	0.00	5.29	0.25	36.04	**42.20**	

CTG	144.00	0.00	2.00	0.00	0.00	0.00	146.00	
CTG	115.00	0.00	45.00	0.00	28.00	4.00	192.00	
TOTAL	259.00	0.00	47.00	0.00	28.00	4.00	338.00	
Exp.	111.88	0.25	20.30	0.25	12.09	1.73		p-value
Exp.	147.12	0.25	26.70	0.25	15.91	2.27		1.23144E-13
X^2	9.22	0.25	16.50	0.25	12.09	1.73	**40.05**	
X^2	7.01	0.25	12.55	0.25	9.20	1.31	**30.57**	

CAG	0.00	0.00	1.00	60.00	27.00	113.00	201.00	
CAG	2.00	2.00	0.00	35.00	8.00	139.00	186.00	
TOTAL	2.00	2.00	1.00	95.00	35.00	252.00	387.00	
Exp.	1.04	1.04	0.52	49.34	18.18	130.88		p-value
Exp.	0.96	0.96	0.48	45.66	16.82	121.12		0.000214204
X^2	1.04	1.04	0.44	2.30	4.28	2.44	**11.55**	
X^2	1.12	1.12	0.48	2.49	4.63	2.64	**12.48**	

CGA	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
CGA	3.00	0.00	61.00	0.00	25.00	0.00	89.00	
TOTAL	3.00	0.00	61.00	0.00	25.00	0.00	89.00	
Exp.	N/A	N/A	N/A	N/A	N/A	N/A		p-value
Exp.	N/A	N/A	N/A	N/A	N/A	N/A		N/A
X^2	N/A	N/A	N/A	N/A	N/A	N/A	**N/A**	
X^2	N/A	N/A	N/A	N/A	N/A	N/A	**N/A**	

CGG	1.00	0.00	38.00	0.00	2.00	0.00	41.00	
CGG	11.00	2.00	1.00	0.00	31.00	0.00	45.00	
TOTAL	12.00	2.00	39.00	0.00	33.00	0.00	86.00	
Exp.	5.72	0.95	18.59	0.25	15.73	0.25		p-value
Exp.	6.28	1.05	20.41	0.25	17.27	0.25		6.69545E-14
X^2	3.90	0.95	20.26	0.25	11.99	0.25	**37.59**	
X^2	3.55	0.87	18.46	0.25	10.92	0.25	**34.30**	

CGC	10.00	1.00	1.00	42.00	0.00	1.00	55.00	
CGC	1.00	1.00	53.00	52.00	32.00	33.00	172.00	
TOTAL	11.00	2.00	54.00	94.00	32.00	34.00	227.00	
Exp.	2.67	0.48	13.08	22.78	7.75	8.24		p-value
Exp.	8.33	1.52	40.92	71.22	24.25	25.76		2.99927E-16
X^2	20.19	0.55	11.16	16.23	7.75	6.36	**62.23**	
X^2	6.45	0.18	3.57	5.19	2.48	2.03	**19.90**	

CGT	47.00	5.00	5.00	30.00	0.00	2.00	89.00	
CGT	0.00	0.00	21.00	0.00	21.00	0.00	42.00	
TOTAL	47.00	5.00	26.00	30.00	21.00	2.00	131.00	
Exp.	31.93	3.40	17.66	20.38	14.27	1.36		p-value
Exp.	15.07	1.60	8.34	9.62	6.73	0.64		1.23753E-22
X^2	7.11	0.76	9.08	4.54	14.27	0.30	**36.06**	
X^2	15.07	1.60	19.24	9.62	30.23	0.64	**76.40**	

TAA	1.00	12.00	0.00	102.00	3.00	36.00	154.00	
TAA	35.00	110.00	5.00	121.00	0.00	37.00	308.00	
TOTAL	36.00	122.00	5.00	223.00	3.00	73.00	462.00	
Exp.	12.00	40.67	1.67	74.33	1.00	24.33		p-value
Exp.	24.00	81.33	3.33	148.67	2.00	48.67		2.45186E-15

X^2	10.08	20.21	1.67	10.30	4.00	5.59	51.85	
X^2	5.04	10.10	0.83	5.15	2.00	2.80	25.92	
TAC	335.00	59.00	130.00	37.00	9.00	8.00	578.00	
TAC	70.00	2.00	58.00	1.00	3.00	1.00	135.00	
TOTAL	405.00	61.00	188.00	38.00	12.00	9.00	713.00	
Exp.	328.32	49.45	152.40	30.81	9.73	7.30		p-value
Exp.	76.68	11.55	35.60	7.19	2.27	1.70		1.45071E-06
X^2	0.14	1.84	3.29	1.25	0.05	0.07	6.64	
X^2	0.58	7.90	14.10	5.33	0.23	0.29	28.44	
TCA	1.00	3.00	19.00	26.00	320.00	7.00	376.00	
TCA	14.00	0.00	67.00	0.00	34.00	0.00	115.00	
TOTAL	15.00	3.00	86.00	26.00	354.00	7.00	491.00	
Exp.	11.49	2.30	65.86	19.91	271.09	5.36		p-value
Exp.	3.51	0.70	20.14	6.09	82.91	1.64		4.16722E-48
X^2	9.57	0.21	33.34	1.86	8.83	0.50	54.32	
X^2	31.30	0.70	109.00	6.09	28.85	1.64	177.59	
TCC	6.00	29.00	73.00	16.00	73.00	5.00	202.00	
TCC	3.00	27.00	18.00	21.00	56.00	14.00	139.00	
TOTAL	9.00	56.00	91.00	37.00	129.00	19.00	341.00	
Exp.	5.33	33.17	53.91	21.92	76.42	11.26		p-value
Exp.	3.67	22.83	37.09	15.08	52.58	7.74		9.76724E-06
X^2	0.08	0.52	6.76	1.60	0.15	3.48	12.60	
X^2	0.12	0.76	9.83	2.32	0.22	5.05	18.31	
TCT	1.00	0.00	68.00	0.00	36.00	0.00	105.00	
TCT	17.00	23.00	57.00	4.00	79.00	0.00	180.00	
TOTAL	18.00	23.00	125.00	4.00	115.00	0.00	285.00	
Exp.	6.63	8.47	46.05	1.47	42.37	0.25		p-value
Exp.	11.37	14.53	78.95	2.53	72.63	0.25		7.79446E-08
X^2	4.78	8.47	10.46	1.47	0.96	0.25	26.40	
X^2	2.79	4.94	6.10	0.86	0.56	0.25	15.50	
TCG	1.00	0.00	1.00	0.00	67.00	6.00	75.00	
TCG	0.00	0.00	2.00	1.00	33.00	42.00	78.00	
TOTAL	1.00	0.00	3.00	1.00	100.00	48.00	153.00	
Exp.	0.49	0.25	1.47	0.49	49.02	23.53		p-value
Exp.	0.51	0.25	1.53	0.51	50.98	24.47		1.00597E-07
X^2	0.53	0.25	0.15	0.49	6.60	13.06	21.08	
X^2	0.51	0.25	0.14	0.47	6.34	12.56	20.27	
TAT	36.00	35.00	110.00	29.00	17.00	1.00	228.00	
TAT	73.00	102.00	10.00	16.00	2.00	8.00	211.00	
TOTAL	109.00	137.00	120.00	45.00	19.00	9.00	439.00	
Exp.	56.61	71.15	62.32	23.37	9.87	4.67		p-value
Exp.	52.39	65.85	57.68	21.63	9.13	4.33		1.91224E-30
X^2	7.50	18.37	36.47	1.36	5.15	2.89	71.74	
X^2	8.11	19.85	39.41	1.46	5.57	3.12	77.52	
TTA	21.00	9.00	108.00	5.00	22.00	26.00	191.00	
TTA	104.00	26.00	113.00	5.00	9.00	13.00	270.00	
TOTAL	125.00	35.00	221.00	10.00	31.00	39.00	461.00	
Exp.	51.79	14.50	91.56	4.14	12.84	16.16		p-value
Exp.	73.21	20.50	129.44	5.86	18.16	22.84		5.84873E-12
X^2	18.30	2.09	2.95	0.18	6.53	5.99	36.04	
X^2	12.95	1.48	2.09	0.13	4.62	4.24	25.50	
TTT	20.00	11.00	1.00	24.00	7.00	76.00	139.00	
TTT	1.00	41.00	0.00	81.00	0.00	71.00	194.00	
TOTAL	21.00	52.00	1.00	105.00	7.00	147.00	333.00	
Exp.	8.77	21.71	0.42	43.83	2.92	61.36		p-value
Exp.	12.23	30.29	0.58	61.17	4.08	85.64		5.91723E-13
X^2	14.40	5.28	0.81	8.97	5.69	3.49	38.65	
X^2	10.32	3.78	0.58	6.43	4.08	2.50	27.69	
TTC	1.00	24.00	0.00	67.00	0.00	56.00	148.00	
TTC	49.00	0.00	1.00	12.00	14.00	115.00	191.00	

TOTAL	50.00	24.00	1.00	79.00	14.00	171.00	339.00	
Exp.	21.83	10.48	0.44	34.49	6.11	74.65		p-value
Exp.	28.17	13.52	0.56	44.51	7.89	96.35		1.37724E-28
X^2	19.87	17.45	0.44	30.64	6.11	4.66	79.18	
X^2	15.40	13.52	0.34	23.75	4.74	3.61	61.35	

TTG	67.00	0.00	114.00	0.00	0.00	0.00	181.00	
TTG	62.00	0.00	87.00	0.00	0.00	0.00	149.00	
TOTAL	129.00	0.00	201.00	0.00	0.00	0.00	330.00	
Exp.	70.75	0.25	110.25	0.25	0.25	0.25		p-value
Exp.	58.25	0.25	90.75	0.25	0.25	0.25		0.981600137
X^2	0.20	0.25	0.13	0.25	0.25	0.25	1.33	
X^2	0.24	0.25	0.16	0.25	0.25	0.25	1.40	

TAG	0.00	0.00	0.00	92.00	14.00	58.00	164.00	
TAG	0.00	0.00	1.00	49.00	2.00	3.00	55.00	
TOTAL	0.00	0.00	1.00	141.00	16.00	61.00	219.00	
Exp.	0.25	0.25	0.75	105.59	11.98	45.68		p-value
Exp.	0.25	0.25	0.25	35.41	4.02	15.32		0.000171764
X^2	0.25	0.25	0.75	1.75	0.34	3.32	6.66	
X^2	0.25	0.25	2.23	5.21	1.01	9.91	18.87	

TGA	28.00	74.00	0.00	67.00	0.00	0.00	169.00	
TGA	0.00	74.00	0.00	7.00	0.00	0.00	81.00	
TOTAL	28.00	148.00	0.00	74.00	0.00	0.00	250.00	
Exp.	18.93	100.05	0.25	50.02	0.25	0.25		p-value
Exp.	9.07	47.95	0.25	23.98	0.25	0.25		5.0687E-10
X^2	4.35	6.78	0.25	5.76	0.25	0.25	17.64	
X^2	9.07	14.15	0.25	12.02	0.25	0.25	35.99	

TGG	0.00	59.00	0.00	29.00	0.00	0.00	88.00	
TGG	0.00	33.00	0.00	44.00	0.00	4.00	81.00	
TOTAL	0.00	92.00	0.00	73.00	0.00	4.00	169.00	
Exp.	0.25	47.91	0.25	38.01	0.25	2.08		p-value
Exp.	0.25	44.09	0.25	34.99	0.25	1.92		0.014598301
X^2	0.25	2.57	0.25	2.14	0.25	2.08	7.54	
X^2	0.25	2.79	0.25	2.32	0.25	2.26	8.13	

TGC	4.00	34.00	0.00	85.00	0.00	0.00	123.00	
TGC	22.00	65.00	1.00	91.00	0.00	0.00	179.00	
TOTAL	26.00	99.00	1.00	176.00	0.00	0.00	302.00	
Exp.	10.59	40.32	0.41	71.68	0.25	0.25		p-value
Exp.	15.41	58.68	0.59	104.32	0.25	0.25		0.019495242
X^2	4.10	0.99	0.41	2.47	0.25	0.25	8.47	
X^2	2.82	0.68	0.28	1.70	0.25	0.25	5.98	

TGT	0.00	1.00	0.00	135.00	0.00	1.00	137.00	
TGT	0.00	21.00	0.00	200.00	0.00	0.00	221.00	
TOTAL	0.00	22.00	0.00	335.00	0.00	1.00	358.00	
Exp.	0.25	8.42	0.25	128.20	0.25	0.38		p-value
Exp.	0.25	13.58	0.25	206.80	0.25	0.62		0.025445335
X^2	0.25	6.54	0.25	0.36	0.25	1.00	8.64	
X^2	0.25	4.05	0.25	0.22	0.25	0.62	5.64	

GAA	35.00	0.00	0.00	0.00	29.00	10.00	74.00	
GAA	1.00	43.00	0.00	2.00	5.00	34.00	85.00	
TOTAL	36.00	43.00	0.00	2.00	34.00	44.00	159.00	
Exp.	16.75	20.01	0.25	0.93	15.82	20.48		p-value
Exp.	19.25	22.99	0.25	1.07	18.18	23.52		1.85645E-21
X^2	19.87	20.01	0.25	0.93	10.97	5.36	57.39	
X^2	17.30	17.42	0.25	0.81	9.55	4.67	50.00	

GAC	0.00	1.00	0.00	26.00	57.00	41.00	125.00	
GAC	11.00	57.00	0.00	28.00	0.00	1.00	97.00	
TOTAL	11.00	58.00	0.00	54.00	57.00	42.00	222.00	
Exp.	6.19	32.66	0.25	30.41	32.09	23.65		p-value
Exp.	4.81	25.34	0.25	23.59	24.91	18.35		1.4373E-32
X^2	6.19	30.69	0.25	0.64	19.33	12.73	69.83	
X^2	7.98	39.55	0.25	0.82	24.91	16.41	89.91	

GCA	33.00	37.00	13.00	36.00	33.00	36.00	188.00	
GCA	38.00	65.00	46.00	23.00	1.00	45.00	218.00	
TOTAL	71.00	102.00	59.00	59.00	34.00	81.00	406.00	
Exp.	32.88	47.23	27.32	27.32	15.74	37.51		p-value
Exp.	38.12	54.77	31.68	31.68	18.26	43.49		2.3865E-11
X^2	0.00	2.22	7.51	2.76	18.91	0.06	31.45	
X^2	0.00	1.91	6.47	2.38	16.31	0.05	27.13	

GCC	97.00	16.00	3.00	0.00	41.00	2.00	159.00	
GCC	18.00	32.00	39.00	0.00	64.00	1.00	154.00	
TOTAL	115.00	48.00	42.00	0.00	105.00	3.00	313.00	
Exp.	58.42	24.38	21.34	0.25	53.34	1.52		p-value
Exp.	56.58	23.62	20.66	0.25	51.66	1.48		4.09952E-19
X^2	25.48	2.88	15.76	0.25	2.85	0.15	47.37	
X^2	26.31	2.98	16.27	0.25	2.95	0.15	48.90	

GCT	0.00	0.00	1.00	4.00	40.00	34.00	79.00	
GCT	0.00	31.00	51.00	0.00	1.00	32.00	115.00	
TOTAL	0.00	31.00	52.00	4.00	41.00	66.00	194.00	
Exp.	0.25	12.62	21.18	1.63	16.70	26.88		p-value
Exp.	0.25	18.38	30.82	2.37	24.30	39.12		1.00929E-23
X^2	0.25	12.62	19.22	3.45	32.53	1.89	69.96	
X^2	0.25	8.67	13.21	2.37	22.35	1.30	48.14	

GCG	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
GCG	0.00	2.00	13.00	23.00	1.00	57.00	96.00	
TOTAL	0.00	2.00	13.00	23.00	1.00	57.00	96.00	
Exp.	N/A	N/A	N/A	N/A	N/A	N/A		p-value
Exp.	N/A	N/A	N/A	N/A	N/A	N/A		N/A
X^2	N/A	N/A	N/A	N/A	N/A	N/A	N/A	
X^2	N/A	N/A	N/A	N/A	N/A	N/A	N/A	

GAT	0.00	63.00	0.00	54.00	0.00	6.00	123.00	
GAT	0.00	42.00	0.00	0.00	0.00	42.00	84.00	
TOTAL	0.00	105.00	0.00	54.00	0.00	48.00	207.00	
Exp.	0.25	62.39	0.25	32.09	0.25	28.52		p-value
Exp.	0.25	42.61	0.25	21.91	0.25	19.48		5.93954E-16
X^2	0.25	0.01	0.25	14.97	0.25	17.78	33.50	
X^2	0.25	0.01	0.25	21.91	0.25	26.04	48.71	

GTA	124.00	43.00	24.00	61.00	98.00	92.00	442.00	
GTA	25.00	9.00	2.00	0.00	41.00	6.00	83.00	
TOTAL	149.00	52.00	26.00	61.00	139.00	98.00	525.00	
Exp.	125.44	43.78	21.89	51.36	117.02	82.51		p-value
Exp.	23.56	8.22	4.11	9.64	21.98	15.49		1.96631E-07
X^2	0.02	0.01	0.20	1.81	3.09	1.09	6.23	
X^2	0.09	0.07	1.08	9.64	16.47	5.82	33.18	

GTT	0.00	0.00	0.00	0.00	22.00	0.00	22.00	
GTT	0.00	28.00	0.00	5.00	2.00	32.00	67.00	
TOTAL	0.00	28.00	0.00	5.00	24.00	32.00	89.00	
Exp.	0.25	6.92	0.25	1.24	5.93	7.91		p-value
Exp.	0.25	21.08	0.25	3.76	18.07	24.09		1.26492E-15
X^2	0.25	6.92	0.25	1.24	43.52	7.91	60.08	
X^2	0.25	2.27	0.25	0.41	14.29	2.60	20.06	

GTC	22.00	10.00	9.00	4.00	3.00	24.00	72.00	
GTC	10.00	0.00	0.00	0.00	34.00	66.00	110.00	
TOTAL	32.00	10.00	9.00	4.00	37.00	90.00	182.00	
Exp.	12.66	3.96	3.56	1.58	14.64	35.60		p-value
Exp.	19.34	6.04	5.44	2.42	22.36	54.40		2.53594E-13
X^2	6.89	9.23	8.31	3.69	9.25	3.78	41.16	
X^2	4.51	6.04	5.44	2.42	6.06	2.48	26.94	

GTG	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
GTG	10.00	0.00	44.00	0.00	54.00	0.00	108.00	
TOTAL	10.00	0.00	44.00	0.00	54.00	0.00	108.00	
Exp.	N/A	N/A	N/A	N/A	N/A	N/A		p-value
Exp.	N/A	N/A	N/A	N/A	N/A	N/A		N/A
X^2	N/A	N/A	N/A	N/A	N/A	N/A	N/A	

X^2	N/A	N/A	N/A	N/A	N/A	N/A	N/A	

GAG	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
GAG	8.00	20.00	0.00	36.00	0.00	36.00	100.00	
TOTAL	8.00	20.00	0.00	36.00	0.00	36.00	100.00	
Exp.	N/A	N/A	N/A	N/A	N/A	N/A		p-value
Exp.	N/A	N/A	N/A	N/A	N/A	N/A		N/A
X^2	N/A	N/A	N/A	N/A	N/A	N/A	N/A	
X^2	N/A	N/A	N/A	N/A	N/A	N/A	N/A	

GGA	66.00	1.00	0.00	0.00	22.00	5.00	94.00	
GGA	37.00	22.00	0.00	2.00	19.00	13.00	93.00	
TOTAL	103.00	23.00	0.00	2.00	41.00	18.00	187.00	
Exp.	51.78	11.56	0.25	1.01	20.61	9.05		p-value
Exp.	51.22	11.44	0.25	0.99	20.39	8.95		3.57942E-06
X^2	3.91	9.65	0.25	1.01	0.09	1.81	16.72	
X^2	3.95	9.75	0.25	1.02	0.09	1.83	16.89	

GGG	1.00	30.00	0.00	0.00	0.00	62.00	93.00	
GGG	1.00	1.00	0.00	0.00	0.00	18.00	20.00	
TOTAL	2.00	31.00	0.00	0.00	0.00	80.00	113.00	
Exp.	1.65	25.51	0.25	0.25	0.25	65.84		p-value
Exp.	0.35	5.49	0.25	0.25	0.25	14.16		0.209268032
X^2	0.25	0.79	0.25	0.25	0.25	0.22	2.02	
X^2	1.18	3.67	0.25	0.25	0.25	1.04	6.64	

GGC	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
GGC	0.00	29.00	0.00	1.00	0.00	32.00	62.00	
TOTAL	0.00	29.00	0.00	1.00	0.00	32.00	62.00	
Exp.	N/A	N/A	N/A	N/A	N/A	N/A		p-value
Exp.	N/A	N/A	N/A	N/A	N/A	N/A		N/A
X^2	N/A	N/A	N/A	N/A	N/A	N/A	N/A	
X^2	N/A	N/A	N/A	N/A	N/A	N/A	N/A	

GGT	0.00	0.00	2.00	41.00	21.00	39.00	103.00	
GGT	0.00	0.00	0.00	33.00	0.00	2.00	35.00	
TOTAL	0.00	0.00	2.00	74.00	21.00	41.00	138.00	
Exp.	0.25	0.25	1.49	55.23	15.67	30.60		p-value
Exp.	0.25	0.25	0.51	18.77	5.33	10.40		7.94153E-06
X^2	0.25	0.25	0.17	3.67	1.81	2.30	8.45	
X^2	0.25	0.25	0.51	10.79	5.33	6.78	23.91	

Table B. The p-values for the comparisons between Primers A1 and C1; Primers B1 and D1; and Primers A1 and B1. Any p-value < 7.8125E-4 (0.05/64) is considered statistically significant with an overall α = 0.05 for that frame given the null hypothesis that the distribution of patterns is the same. In this example, all p-values exceed 7.8125E-4; the null hypothesis is not to be rejected. Comparisons cannot be made between primers.

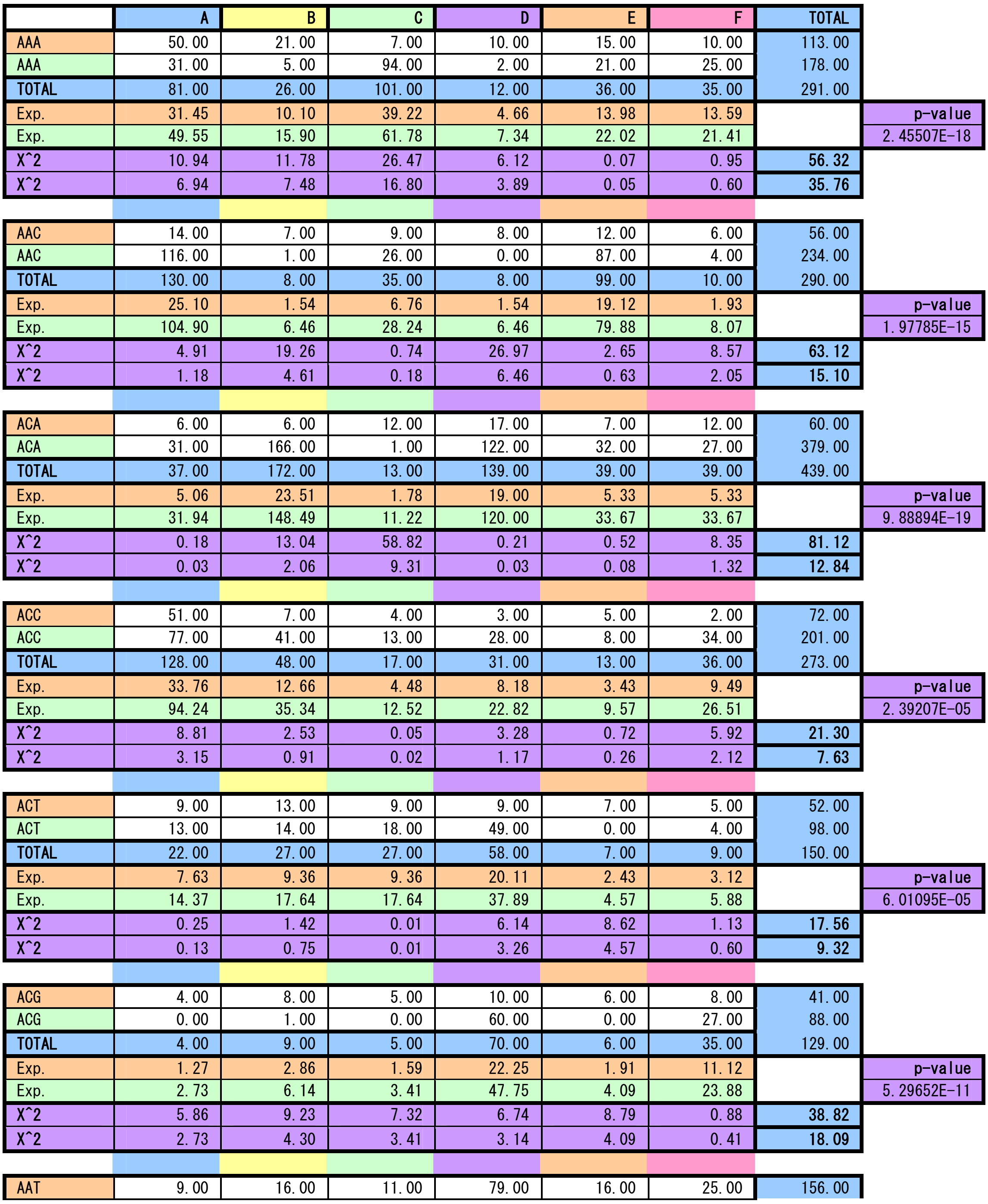
	p-value		
Frame	A1 - C1	B1 - D1	A1 - B1
AAA	1.20022E-07	7.013E-09	6.45651E-12
AAC	1.40989E-28	4.53575E-05	1.99847E-15
ACA	2.00702E-29	3.12266E-06	4.73195E-36
ACC	7.72392E-15	0.934557717	5.09304E-33
ACT	9.3124E-27	7.84194E-25	2.11203E-22
ACG	3.4923E-25	1.03049E-08	3.9667E-08
AAT	9.02911E-26	7.98438E-12	1.32109E-29
ATA	6.74572E-13	5.27502E-28	1.39138E-15
ATT	1.99252E-55	9.96064E-11	7.82831E-10
ATC	2.65293E-13	0.004300712	1.45441E-07
ATG	0.004046262	0.002898016	2.08447E-25
AAG	2.91673E-12	2.26737E-17	2.85657E-14
AGA	0.216618383	2.67454E-10	2.01334E-11
AGG	3.61857E-10	1.79521E-06	2.97852E-14
AGC	2.24005E-09	6.06573E-12	4.60009E-07
AGT	4.9304E-13	1.81423E-10	1.29992E-21
CAA	1.00751E-31	1.71861E-07	1.95454E-43
CAC	2.70263E-47	N/A	N/A
CCA	1.71467E-17	2.59221E-14	3.9681E-10
CCC	1.58567E-09	1.25396E-06	5.11253E-06
CCT	4.30325E-44	3.14998E-07	8.43966E-25
CCG	6.55286E-23	0.000342836	9.20901E-19
CAT	4.77169E-08	5.12204E-11	1.49321E-36
CTA	1.28144E-34	1.59877E-15	1.77731E-36
CTT	3.10181E-12	1.56409E-11	8.90771E-45
CTC	6.30261E-10	0.082287561	0.011600491
CTG	1.23144E-13	1.39157E-07	3.18246E-19
CAG	0.000214204	1.02616E-05	0.000668257
CGA	N/A	0.160445616	N/A
CGG	6.69545E-14	1.30627E-07	9.33045E-13
CGC	2.99927E-16	N/A	N/A
CGT	1.23753E-22	1.32679E-05	2.95144E-21
TAA	2.45186E-15	1.54798E-28	2.62831E-60
TAC	1.45071E-06	3.5057E-16	3.97032E-21
TCA	4.16722E-48	6.27949E-10	9.52412E-45
TCC	9.76724E-06	3.30645E-10	2.96373E-15
TCT	7.79446E-08	0.574033387	0.083747644
TCG	1.00597E-07	N/A	N/A
TAT	1.91224E-30	1.51128E-08	5.68855E-11
TTA	5.84873E-12	3.7283E-15	4.46194E-06
TTT	5.91723E-13	3.35272E-21	6.07473E-11
TTC	1.37724E-28	0.001968687	4.12383E-08
TTG	0.981600137	4.65451E-15	4.29383E-23
TAG	0.000171764	0.068379023	0.0008092
TGA	5.0687E-10	8.24932E-10	4.14736E-60
TGG	0.014598301	1.49799E-09	2.83406E-07
TGC	0.019495242	1.35218E-37	4.4806E-27
TGT	0.025445335	1.87482E-10	2.07463E-11
GAA	1.85645E-21	2.66916E-11	2.57306E-30
GAC	1.4373E-32	1.76059E-24	8.6831E-34
GCA	2.3865E-11	2.21577E-12	5.24235E-06
GCC	4.09952E-19	N/A	N/A
GCT	1.00929E-23	6.53521E-16	2.35488E-26
GCG	N/A	4.30287E-10	N/A
GAT	5.93954E-16	3.21042E-21	1.04213E-08
GTA	1.96631E-07	2.77654E-10	1.58E-28
GTT	1.26492E-15	8.6537E-36	2.48183E-09
GTC	2.53594E-13	2.733E-25	3.57506E-17
GTG	N/A	0.000976928	N/A
GAG	N/A	4.56534E-33	N/A

GGA	3.57942E-06	4.99242E-20	7.60976E-13
GGG	0.209268032	4.38836E-12	3.8279E-10
GGC	N/A	3.09797E-28	N/A
GGT	7.94153E-06	1.48212E-28	1.90916E-08

Appendix 5

The following table is an example of the processed data using the bioinformatic tools to statistically evaluate each of the 64 frames using a 2 x 6 Chi-square analysis comparing the results from data for the same samples using different dye chemistries produced in the same laboratory. These results demonstrate that the data obtained from the different dye chemistries are not comparable. Comparisons were made between dRhodamine and BigDye v1.1.

Table A. 2 x 6 Chi-square analysis comparing the results between dRhodamine and BigDye v1.1 for Primer C1.



	A	B	C	D	E	F	TOTAL	
AAA	50.00	21.00	7.00	10.00	15.00	10.00	113.00	
AAA	31.00	5.00	94.00	2.00	21.00	25.00	178.00	
TOTAL	81.00	26.00	101.00	12.00	36.00	35.00	291.00	
Exp.	31.45	10.10	39.22	4.66	13.98	13.59		p-value
Exp.	49.55	15.90	61.78	7.34	22.02	21.41		2.45507E-18
X^2	10.94	11.78	26.47	6.12	0.07	0.95	**56.32**	
X^2	6.94	7.48	16.80	3.89	0.05	0.60	**35.76**	
AAC	14.00	7.00	9.00	8.00	12.00	6.00	56.00	
AAC	116.00	1.00	26.00	0.00	87.00	4.00	234.00	
TOTAL	130.00	8.00	35.00	8.00	99.00	10.00	290.00	
Exp.	25.10	1.54	6.76	1.54	19.12	1.93		p-value
Exp.	104.90	6.46	28.24	6.46	79.88	8.07		1.97785E-15
X^2	4.91	19.26	0.74	26.97	2.65	8.57	**63.12**	
X^2	1.18	4.61	0.18	6.46	0.63	2.05	**15.10**	
ACA	6.00	6.00	12.00	17.00	7.00	12.00	60.00	
ACA	31.00	166.00	1.00	122.00	32.00	27.00	379.00	
TOTAL	37.00	172.00	13.00	139.00	39.00	39.00	439.00	
Exp.	5.06	23.51	1.78	19.00	5.33	5.33		p-value
Exp.	31.94	148.49	11.22	120.00	33.67	33.67		9.88894E-19
X^2	0.18	13.04	58.82	0.21	0.52	8.35	**81.12**	
X^2	0.03	2.06	9.31	0.03	0.08	1.32	**12.84**	
ACC	51.00	7.00	4.00	3.00	5.00	2.00	72.00	
ACC	77.00	41.00	13.00	28.00	8.00	34.00	201.00	
TOTAL	128.00	48.00	17.00	31.00	13.00	36.00	273.00	
Exp.	33.76	12.66	4.48	8.18	3.43	9.49		p-value
Exp.	94.24	35.34	12.52	22.82	9.57	26.51		2.39207E-05
X^2	8.81	2.53	0.05	3.28	0.72	5.92	**21.30**	
X^2	3.15	0.91	0.02	1.17	0.26	2.12	**7.63**	
ACT	9.00	13.00	9.00	9.00	7.00	5.00	52.00	
ACT	13.00	14.00	18.00	49.00	0.00	4.00	98.00	
TOTAL	22.00	27.00	27.00	58.00	7.00	9.00	150.00	
Exp.	7.63	9.36	9.36	20.11	2.43	3.12		p-value
Exp.	14.37	17.64	17.64	37.89	4.57	5.88		6.01095E-05
X^2	0.25	1.42	0.01	6.14	8.62	1.13	**17.56**	
X^2	0.13	0.75	0.01	3.26	4.57	0.60	**9.32**	
ACG	4.00	8.00	5.00	10.00	6.00	8.00	41.00	
ACG	0.00	1.00	0.00	60.00	0.00	27.00	88.00	
TOTAL	4.00	9.00	5.00	70.00	6.00	35.00	129.00	
Exp.	1.27	2.86	1.59	22.25	1.91	11.12		p-value
Exp.	2.73	6.14	3.41	47.75	4.09	23.88		5.29652E-11
X^2	5.86	9.23	7.32	6.74	8.79	0.88	**38.82**	
X^2	2.73	4.30	3.41	3.14	4.09	0.41	**18.09**	
AAT	9.00	16.00	11.00	79.00	16.00	25.00	156.00	

AAT	8.00	40.00	2.00	32.00	81.00	140.00	303.00	
TOTAL	17.00	56.00	13.00	111.00	97.00	165.00	459.00	
Exp.	5.78	19.03	4.42	37.73	32.97	56.08		p-value
Exp.	11.22	36.97	8.58	73.27	64.03	108.92		1.65255E-25
X^2	1.80	0.48	9.80	45.16	8.73	17.22	**83.20**	
X^2	0.93	0.25	5.05	23.25	4.50	8.87	**42.83**	

ATA	22.00	14.00	11.00	10.00	13.00	6.00	76.00	
ATA	70.00	10.00	71.00	11.00	34.00	22.00	218.00	
TOTAL	92.00	24.00	82.00	21.00	47.00	28.00	294.00	
Exp.	23.78	6.20	21.20	5.43	12.15	7.24		p-value
Exp.	68.22	17.80	60.80	15.57	34.85	20.76		0.000108357
X^2	0.13	9.80	4.91	3.85	0.06	0.21	**18.96**	
X^2	0.05	3.42	1.71	1.34	0.02	0.07	**6.61**	

ATT	79.00	19.00	41.00	17.00	30.00	12.00	198.00	
ATT	30.00	10.00	53.00	4.00	156.00	96.00	349.00	
TOTAL	109.00	29.00	94.00	21.00	186.00	108.00	547.00	
Exp.	39.46	10.50	34.03	7.60	67.33	39.09		p-value
Exp.	69.54	18.50	59.97	13.40	118.67	68.91		1.02567E-31
X^2	39.63	6.89	1.43	11.62	20.69	18.78	**99.04**	
X^2	22.49	3.91	0.81	6.59	11.74	10.65	**56.19**	

ATC	17.00	10.00	19.00	12.00	10.00	7.00	75.00	
ATC	1.00	6.00	0.00	81.00	0.00	62.00	150.00	
TOTAL	18.00	16.00	19.00	93.00	10.00	69.00	225.00	
Exp.	6.00	5.33	6.33	31.00	3.33	23.00		p-value
Exp.	12.00	10.67	12.67	62.00	6.67	46.00		4.86079E-26
X^2	20.17	4.08	25.33	11.65	13.33	11.13	**85.69**	
X^2	10.08	2.04	12.67	5.82	6.67	5.57	**42.85**	

ATG	10.00	9.00	25.00	12.00	70.00	18.00	144.00	
ATG	79.00	1.00	35.00	0.00	0.00	0.00	115.00	
TOTAL	89.00	10.00	60.00	12.00	70.00	18.00	259.00	
Exp.	49.48	5.56	33.36	6.67	38.92	10.01		p-value
Exp.	39.52	4.44	26.64	5.33	31.08	7.99		8.44338E-33
X^2	31.50	2.13	2.09	4.26	24.82	6.38	**71.19**	
X^2	39.45	2.67	2.62	5.33	31.08	7.99	**89.14**	

AAG	55.00	21.00	58.00	7.00	18.00	16.00	175.00	
AAG	8.00	4.00	13.00	17.00	4.00	8.00	54.00	
TOTAL	63.00	25.00	71.00	24.00	22.00	24.00	229.00	
Exp.	48.14	19.10	54.26	18.34	16.81	18.34		p-value
Exp.	14.86	5.90	16.74	5.66	5.19	5.66		4.9977E-07
X^2	0.98	0.19	0.26	7.01	0.08	0.30	**8.82**	
X^2	3.16	0.61	0.84	22.72	0.27	0.97	**28.57**	

AGA	7.00	65.00	16.00	66.00	8.00	16.00	178.00	
AGA	9.00	0.00	91.00	0.00	0.00	2.00	102.00	
TOTAL	16.00	65.00	107.00	66.00	8.00	18.00	280.00	
Exp.	10.17	41.32	68.02	41.96	5.09	11.44		p-value
Exp.	5.83	23.68	38.98	24.04	2.91	6.56		1.54464E-40
X^2	0.99	13.57	39.78	13.78	1.67	1.81	**71.60**	
X^2	1.73	23.68	69.43	24.04	2.91	3.17	**124.96**	

AGG	11.00	37.00	8.00	48.00	10.00	38.00	152.00	
AGG	9.00	0.00	49.00	0.00	0.00	0.00	58.00	
TOTAL	20.00	37.00	57.00	48.00	10.00	38.00	210.00	
Exp.	14.48	26.78	41.26	34.74	7.24	27.50		p-value
Exp.	5.52	10.22	15.74	13.26	2.76	10.50		8.85743E-31
X^2	0.83	3.90	26.81	5.06	1.05	4.00	**41.66**	
X^2	2.19	10.22	70.26	13.26	2.76	10.50	**109.18**	

AGC	8.00	18.00	11.00	88.00	9.00	16.00	150.00
AGC	0.00	0.00	84.00	2.00	37.00	42.00	165.00
TOTAL	8.00	18.00	95.00	90.00	46.00	58.00	315.00

Exp.	3.81	8.57	45.24	42.86	21.90	27.62		p-value
Exp.	4.19	9.43	49.76	47.14	24.10	30.38		1.03736E-39
X^2	4.61	10.37	25.91	47.55	7.60	4.89	100.93	
X^2	4.19	9.43	23.56	43.23	6.91	4.44	91.76	

AGT	13.00	8.00	11.00	4.00	10.00	8.00	54.00	
AGT	2.00	10.00	6.00	12.00	37.00	1.00	68.00	
TOTAL	15.00	18.00	17.00	16.00	47.00	9.00	122.00	
Exp.	6.64	7.97	7.52	7.08	20.80	3.98		p-value
Exp.	8.36	10.03	9.48	8.92	26.20	5.02		2.9264E-06
X^2	6.09	0.00	1.61	1.34	5.61	4.05	18.70	
X^2	4.84	0.00	1.27	1.07	4.46	3.22	14.85	

CAA	45.00	12.00	8.00	4.00	14.00	17.00	100.00	
CAA	0.00	51.00	0.00	13.00	68.00	64.00	196.00	
TOTAL	45.00	63.00	8.00	17.00	82.00	81.00	296.00	
Exp.	15.20	21.28	2.70	5.74	27.70	27.36		p-value
Exp.	29.80	41.72	5.30	11.26	54.30	53.64		1.05113E-25
X^2	58.40	4.05	10.38	0.53	6.78	3.93	84.07	
X^2	29.80	2.07	5.30	0.27	3.46	2.00	42.89	

CAC	5.00	14.00	14.00	8.00	48.00	9.00	98.00	
CAC	49.00	45.00	85.00	0.00	98.00	24.00	301.00	
TOTAL	54.00	59.00	99.00	8.00	146.00	33.00	399.00	
Exp.	13.26	14.49	24.32	1.96	35.86	8.11		p-value
Exp.	40.74	44.51	74.68	6.04	110.14	24.89		4.0599E-08
X^2	5.15	0.02	4.38	18.54	4.11	0.10	32.29	
X^2	1.68	0.01	1.42	6.04	1.34	0.03	10.51	

CCA	9.00	9.00	6.00	15.00	6.00	53.00	98.00	
CCA	7.00	0.00	34.00	108.00	16.00	42.00	207.00	
TOTAL	16.00	9.00	40.00	123.00	22.00	95.00	305.00	
Exp.	5.14	2.89	12.85	39.52	7.07	30.52		p-value
Exp.	10.86	6.11	27.15	83.48	14.93	64.48		6.64381E-15
X^2	2.90	12.90	3.65	15.21	0.16	16.55	51.38	
X^2	1.37	6.11	1.73	7.20	0.08	7.83	24.32	

CCC	5.00	17.00	3.00	15.00	4.00	15.00	59.00	
CCC	0.00	5.00	0.00	91.00	0.00	42.00	138.00	
TOTAL	5.00	22.00	3.00	106.00	4.00	57.00	197.00	
Exp.	1.50	6.59	0.90	31.75	1.20	17.07		p-value
Exp.	3.50	15.41	2.10	74.25	2.80	39.93		1.40869E-12
X^2	8.19	16.45	4.92	8.83	6.55	0.25	45.20	
X^2	3.50	7.03	2.10	3.78	2.80	0.11	19.32	

CCT	9.00	22.00	13.00	39.00	7.00	13.00	103.00	
CCT	15.00	36.00	6.00	15.00	39.00	76.00	187.00	
TOTAL	24.00	58.00	19.00	54.00	46.00	89.00	290.00	
Exp.	8.52	20.60	6.75	19.18	16.34	31.61		p-value
Exp.	15.48	37.40	12.25	34.82	29.66	57.39		6.30199E-13
X^2	0.03	0.10	5.79	20.48	5.34	10.96	42.69	
X^2	0.01	0.05	3.19	11.28	2.94	6.03	23.51	

CCG	1.00	4.00	6.00	12.00	6.00	7.00	36.00	
CCG	3.00	32.00	16.00	0.00	2.00	32.00	85.00	
TOTAL	4.00	36.00	22.00	12.00	8.00	39.00	121.00	
Exp.	1.19	10.71	6.55	3.57	2.38	11.60		p-value
Exp.	2.81	25.29	15.45	8.43	5.62	27.40		1.54692E-08
X^2	0.03	4.20	0.05	19.90	5.51	1.83	31.52	
X^2	0.01	1.78	0.02	8.43	2.33	0.77	13.35	

CAT	11.00	10.00	11.00	8.00	14.00	8.00	62.00	
CAT	36.00	84.00	21.00	3.00	66.00	21.00	231.00	
TOTAL	47.00	94.00	32.00	11.00	80.00	29.00	293.00	
Exp.	9.95	19.89	6.77	2.33	16.93	6.14		p-value
Exp.	37.05	74.11	25.23	8.67	63.07	22.86		2.74849E-05
X^2	0.11	4.92	2.64	13.82	0.51	0.57	22.57	

X^2	0.03	1.32	0.71	3.71	0.14	0.15	6.06	
CTA	3.00	7.00	5.00	11.00	8.00	7.00	41.00	
CTA	20.00	2.00	40.00	16.00	45.00	9.00	132.00	
TOTAL	23.00	9.00	45.00	27.00	53.00	16.00	173.00	
Exp.	5.45	2.13	10.66	6.40	12.56	3.79		p-value
Exp.	17.55	6.87	34.34	20.60	40.44	12.21		1.47031E-05
X^2	1.10	11.11	3.01	3.31	1.66	2.71	22.90	
X^2	0.34	3.45	0.93	1.03	0.51	0.84	7.11	
CTT	24.00	23.00	14.00	11.00	9.00	9.00	90.00	
CTT	56.00	47.00	24.00	0.00	35.00	32.00	194.00	
TOTAL	80.00	70.00	38.00	11.00	44.00	41.00	284.00	
Exp.	25.35	22.18	12.04	3.49	13.94	12.99		p-value
Exp.	54.65	47.82	25.96	7.51	30.06	28.01		2.6684E-05
X^2	0.07	0.03	0.32	16.20	1.75	1.23	19.60	
X^2	0.03	0.01	0.15	7.51	0.81	0.57	9.09	
CTC	6.00	10.00	11.00	14.00	15.00	8.00	64.00	
CTC	0.00	12.00	1.00	24.00	0.00	24.00	61.00	
TOTAL	6.00	22.00	12.00	38.00	15.00	32.00	125.00	
Exp.	3.07	11.26	6.14	19.46	7.68	16.38		p-value
Exp.	2.93	10.74	5.86	18.54	7.32	15.62		1.42705E-07
X^2	2.79	0.14	3.84	1.53	6.98	4.29	19.57	
X^2	2.93	0.15	4.03	1.61	7.32	4.50	20.53	
CTG	4.00	5.00	4.00	10.00	5.00	53.00	81.00	
CTG	115.00	0.00	45.00	0.00	28.00	4.00	192.00	
TOTAL	119.00	5.00	49.00	10.00	33.00	57.00	273.00	
Exp.	35.31	1.48	14.54	2.97	9.79	16.91		p-value
Exp.	83.69	3.52	34.46	7.03	23.21	40.09		5.34485E-41
X^2	27.76	8.34	7.64	16.67	2.34	77.01	139.76	
X^2	11.71	3.52	3.22	7.03	0.99	32.49	58.96	
CAG	17.00	14.00	54.00	5.00	11.00	8.00	109.00	
CAG	2.00	2.00	0.00	35.00	8.00	139.00	186.00	
TOTAL	19.00	16.00	54.00	40.00	19.00	147.00	295.00	
Exp.	7.02	5.91	19.95	14.78	7.02	54.32		p-value
Exp.	11.98	10.09	34.05	25.22	11.98	92.68		3.9521E-43
X^2	14.19	11.07	58.10	6.47	2.26	39.49	131.57	
X^2	8.31	6.48	34.05	3.79	1.32	23.14	77.10	
CGA	3.00	8.00	6.00	6.00	10.00	10.00	43.00	
CGA	3.00	0.00	61.00	0.00	25.00	0.00	89.00	
TOTAL	6.00	8.00	67.00	6.00	35.00	10.00	132.00	
Exp.	1.95	2.61	21.83	1.95	11.40	3.26		p-value
Exp.	4.05	5.39	45.17	4.05	23.60	6.74		2.96881E-13
X^2	0.56	11.16	11.48	8.37	0.17	13.96	45.70	
X^2	0.27	5.39	5.54	4.05	0.08	6.74	22.08	
CGG	19.00	39.00	3.00	10.00	5.00	8.00	84.00	
CGG	11.00	2.00	1.00	0.00	31.00	0.00	45.00	
TOTAL	30.00	41.00	4.00	10.00	36.00	8.00	129.00	
Exp.	19.53	26.70	2.60	6.51	23.44	5.21		p-value
Exp.	10.47	14.30	1.40	3.49	12.56	2.79		3.08516E-13
X^2	0.01	5.67	0.06	1.87	14.51	1.50	23.62	
X^2	0.03	10.58	0.11	3.49	27.08	2.79	44.08	
CGC	4.00	6.00	10.00	3.00	11.00	11.00	45.00	
CGC	1.00	1.00	53.00	52.00	32.00	33.00	172.00	
TOTAL	5.00	7.00	63.00	55.00	43.00	44.00	217.00	
Exp.	1.04	1.45	13.06	11.41	8.92	9.12		p-value
Exp.	3.96	5.55	49.94	43.59	34.08	34.88		3.01308E-07
X^2	8.47	14.25	0.72	6.19	0.49	0.39	30.51	
X^2	2.22	3.73	0.19	1.62	0.13	0.10	7.98	
CGT	51.00	1.00	11.00	5.00	2.00	9.00	79.00	

CGT	0.00	0.00	21.00	0.00	21.00	0.00	42.00	
TOTAL	51.00	1.00	32.00	5.00	23.00	9.00	121.00	
Exp.	33.30	0.65	20.89	3.26	15.02	5.88		p-value
Exp.	17.70	0.35	11.11	1.74	7.98	3.12		4.96619E-16
X^2	9.41	0.18	4.68	0.92	11.28	1.66	**28.15**	
X^2	17.70	0.35	8.81	1.74	21.22	3.12	**52.94**	

TAA	8.00	12.00	6.00	14.00	15.00	8.00	63.00	
TAA	35.00	110.00	5.00	121.00	0.00	37.00	308.00	
TOTAL	43.00	122.00	11.00	135.00	15.00	45.00	371.00	
Exp.	7.30	20.72	1.87	22.92	2.55	7.64		p-value
Exp.	35.70	101.28	9.13	112.08	12.45	37.36		1.53755E-18
X^2	0.07	3.67	9.14	3.47	60.88	0.02	**77.25**	
X^2	0.01	0.75	1.87	0.71	12.45	0.00	**15.80**	

TAC	5.00	5.00	4.00	4.00	9.00	8.00	35.00	
TAC	70.00	2.00	58.00	1.00	3.00	1.00	135.00	
TOTAL	75.00	7.00	62.00	5.00	12.00	9.00	170.00	
Exp.	15.44	1.44	12.76	1.03	2.47	1.85		p-value
Exp.	59.56	5.56	49.24	3.97	9.53	7.15		5.26601E-17
X^2	7.06	8.79	6.02	8.57	17.26	20.39	**68.09**	
X^2	1.83	2.28	1.56	2.22	4.47	5.29	**17.65**	

TCA	8.00	51.00	12.00	24.00	14.00	9.00	118.00	
TCA	14.00	0.00	67.00	0.00	34.00	0.00	115.00	
TOTAL	22.00	51.00	79.00	24.00	48.00	9.00	233.00	
Exp.	11.14	25.83	40.01	12.15	24.31	4.56		p-value
Exp.	10.86	25.17	38.99	11.85	23.69	4.44		7.94784E-27
X^2	0.89	24.53	19.61	11.54	4.37	4.33	**65.27**	
X^2	0.91	25.17	20.12	11.85	4.49	4.44	**66.97**	

TCC	18.00	53.00	16.00	6.00	13.00	12.00	118.00	
TCC	3.00	27.00	18.00	21.00	56.00	14.00	139.00	
TOTAL	21.00	80.00	34.00	27.00	69.00	26.00	257.00	
Exp.	9.64	36.73	15.61	12.40	31.68	11.94		p-value
Exp.	11.36	43.27	18.39	14.60	37.32	14.06		3.05193E-10
X^2	7.24	7.21	0.01	3.30	11.02	0.00	**28.78**	
X^2	6.15	6.12	0.01	2.80	9.35	0.00	**24.43**	

TCT	15.00	9.00	21.00	17.00	26.00	8.00	96.00	
TCT	17.00	23.00	57.00	4.00	79.00	0.00	180.00	
TOTAL	32.00	32.00	78.00	21.00	105.00	8.00	276.00	
Exp.	11.13	11.13	27.13	7.30	36.52	2.78		p-value
Exp.	20.87	20.87	50.87	13.70	68.48	5.22		2.11567E-08
X^2	1.35	0.41	1.39	12.87	3.03	9.78	**28.82**	
X^2	0.72	0.22	0.74	6.86	1.62	5.22	**15.37**	

TCG	5.00	9.00	6.00	4.00	3.00	8.00	35.00	
TCG	0.00	0.00	2.00	1.00	33.00	42.00	78.00	
TOTAL	5.00	9.00	8.00	5.00	36.00	50.00	113.00	
Exp.	1.55	2.79	2.48	1.55	11.15	15.49		p-value
Exp.	3.45	6.21	5.52	3.45	24.85	34.51		3.22366E-11
X^2	7.69	13.84	5.01	3.88	5.96	3.62	**40.00**	
X^2	3.45	6.21	2.25	1.74	2.67	1.62	**17.95**	

TAT	22.00	53.00	34.00	22.00	25.00	21.00	177.00	
TAT	73.00	102.00	10.00	16.00	2.00	8.00	211.00	
TOTAL	95.00	155.00	44.00	38.00	27.00	29.00	388.00	
Exp.	43.34	70.71	20.07	17.34	12.32	13.23		p-value
Exp.	51.66	84.29	23.93	20.66	14.68	15.77		8.54536E-16
X^2	10.51	4.44	9.66	1.26	13.06	4.56	**43.48**	
X^2	8.81	3.72	8.11	1.05	10.96	3.83	**36.48**	

TTA	10.00	37.00	6.00	42.00	21.00	27.00	143.00	
TTA	104.00	26.00	113.00	5.00	9.00	13.00	270.00	
TOTAL	114.00	63.00	119.00	47.00	30.00	40.00	413.00	
Exp.	39.47	21.81	41.20	16.27	10.39	13.85		p-value

Exp.	74. 53	41. 19	77. 80	30. 73	19. 61	26. 15		6. 21845E-40
X^2	22. 01	10. 57	30. 08	40. 67	10. 84	12. 49	126. 65	
X^2	11. 65	5. 60	15. 93	21. 54	5. 74	6. 61	67. 08	
TTT	96. 00	101. 00	46. 00	24. 00	17. 00	19. 00	303. 00	
TTT	1. 00	41. 00	0. 00	81. 00	0. 00	71. 00	194. 00	
TOTAL	97. 00	142. 00	46. 00	105. 00	17. 00	90. 00	497. 00	
Exp.	59. 14	86. 57	28. 04	64. 01	10. 36	54. 87		p-value
Exp.	37. 86	55. 43	17. 96	40. 99	6. 64	35. 13		1. 35895E-47
X^2	22. 98	2. 40	11. 50	25. 01	4. 25	23. 45	89. 59	
X^2	35. 89	3. 76	17. 96	39. 07	6. 64	36. 62	139. 93	
TTC	15. 00	13. 00	45. 00	32. 00	20. 00	17. 00	142. 00	
TTC	49. 00	0. 00	1. 00	12. 00	14. 00	115. 00	191. 00	
TOTAL	64. 00	13. 00	46. 00	44. 00	34. 00	132. 00	333. 00	
Exp.	27. 29	5. 54	19. 62	18. 76	14. 50	56. 29		p-value
Exp.	36. 71	7. 46	26. 38	25. 24	19. 50	75. 71		4. 67497E-31
X^2	5. 54	10. 03	32. 85	9. 34	2. 09	27. 42	87. 26	
X^2	4. 12	7. 46	24. 42	6. 94	1. 55	20. 39	64. 88	
TTG	6. 00	16. 00	9. 00	26. 00	20. 00	77. 00	154. 00	
TTG	62. 00	0. 00	87. 00	0. 00	0. 00	0. 00	149. 00	
TOTAL	68. 00	16. 00	96. 00	26. 00	20. 00	77. 00	303. 00	
Exp.	34. 56	8. 13	48. 79	13. 21	10. 17	39. 14		p-value
Exp.	33. 44	7. 87	47. 21	12. 79	9. 83	37. 86		1. 16604E-51
X^2	23. 60	7. 61	32. 45	12. 37	9. 52	36. 64	122. 19	
X^2	24. 39	7. 87	33. 54	12. 79	9. 83	37. 86	126. 29	
TAG	4. 00	7. 00	11. 00	9. 00	19. 00	39. 00	89. 00	
TAG	0. 00	0. 00	1. 00	49. 00	2. 00	3. 00	55. 00	
TOTAL	4. 00	7. 00	12. 00	58. 00	21. 00	42. 00	144. 00	
Exp.	2. 47	4. 33	7. 42	35. 85	12. 98	25. 96		p-value
Exp.	1. 53	2. 67	4. 58	22. 15	8. 02	16. 04		1. 42781E-17
X^2	0. 94	1. 65	1. 73	20. 11	2. 79	6. 55	33. 78	
X^2	1. 53	2. 67	2. 80	32. 54	4. 52	10. 60	54. 66	
TGA	5. 00	8. 00	6. 00	6. 00	17. 00	5. 00	47. 00	
TGA	0. 00	74. 00	0. 00	7. 00	0. 00	0. 00	81. 00	
TOTAL	5. 00	82. 00	6. 00	13. 00	17. 00	5. 00	128. 00	
Exp.	1. 84	30. 11	2. 20	4. 77	6. 24	1. 84		p-value
Exp.	3. 16	51. 89	3. 80	8. 23	10. 76	3. 16		1. 95185E-16
X^2	5. 45	16. 23	6. 54	0. 32	18. 54	5. 45	52. 54	
X^2	3. 16	9. 42	3. 80	0. 18	10. 76	3. 16	30. 49	
TGG	15. 00	46. 00	7. 00	46. 00	6. 00	43. 00	163. 00	
TGG	0. 00	33. 00	0. 00	44. 00	0. 00	4. 00	81. 00	
TOTAL	15. 00	79. 00	7. 00	90. 00	6. 00	47. 00	244. 00	
Exp.	10. 02	52. 77	4. 68	60. 12	4. 01	31. 40		p-value
Exp.	4. 98	26. 23	2. 32	29. 88	1. 99	15. 60		1. 93418E-07
X^2	2. 47	0. 87	1. 15	3. 32	0. 99	4. 29	13. 09	
X^2	4. 98	1. 75	2. 32	6. 68	1. 99	8. 63	26. 35	
TGC	5. 00	7. 00	5. 00	8. 00	4. 00	6. 00	35. 00	
TGC	22. 00	65. 00	1. 00	91. 00	0. 00	0. 00	179. 00	
TOTAL	27. 00	72. 00	6. 00	99. 00	4. 00	6. 00	214. 00	
Exp.	4. 42	11. 78	0. 98	16. 19	0. 65	0. 98		p-value
Exp.	22. 58	60. 22	5. 02	82. 81	3. 35	5. 02		2. 01504E-15
X^2	0. 08	1. 94	16. 46	4. 14	17. 11	25. 67	65. 39	
X^2	0. 02	0. 38	3. 22	0. 81	3. 35	5. 02	12. 79	
TGT	3. 00	4. 00	23. 00	16. 00	72. 00	49. 00	167. 00	
TGT	0. 00	21. 00	0. 00	200. 00	0. 00	0. 00	221. 00	
TOTAL	3. 00	25. 00	23. 00	216. 00	72. 00	49. 00	388. 00	
Exp.	1. 29	10. 76	9. 90	92. 97	30. 99	21. 09		p-value
Exp.	1. 71	14. 24	13. 10	123. 03	41. 01	27. 91		1. 04523E-65
X^2	2. 26	4. 25	17. 34	63. 72	54. 27	36. 93	178. 77	
X^2	1. 71	3. 21	13. 10	48. 15	41. 01	27. 91	135. 09	

GAA	29.00	13.00	61.00	9.00	48.00	57.00	217.00	
GAA	1.00	43.00	0.00	2.00	5.00	34.00	85.00	
TOTAL	30.00	56.00	61.00	11.00	53.00	91.00	302.00	
Exp.	21.56	40.24	43.83	7.90	38.08	65.39		p-value
Exp.	8.44	15.76	17.17	3.10	14.92	25.61		1.49219E-22
X^2	2.57	18.44	6.73	0.15	2.58	1.08	**31.54**	
X^2	6.56	47.07	17.17	0.39	6.59	2.75	**80.53**	

GAC	9.00	4.00	5.00	3.00	9.00	7.00	37.00	
GAC	11.00	57.00	0.00	28.00	0.00	1.00	97.00	
TOTAL	20.00	61.00	5.00	31.00	9.00	8.00	134.00	
Exp.	5.52	16.84	1.38	8.56	2.49	2.21		p-value
Exp.	14.48	44.16	3.62	22.44	6.51	5.79		2.94479E-14
X^2	2.19	9.79	9.49	3.61	17.08	10.39	**52.55**	
X^2	0.84	3.74	3.62	1.38	6.51	3.96	**20.05**	

GCA	9.00	6.00	11.00	3.00	53.00	10.00	92.00	
GCA	38.00	65.00	46.00	23.00	1.00	45.00	218.00	
TOTAL	47.00	71.00	57.00	26.00	54.00	55.00	310.00	
Exp.	13.95	21.07	16.92	7.72	16.03	16.32		p-value
Exp.	33.05	49.93	40.08	18.28	37.97	38.68		1.58114E-30
X^2	1.76	10.78	2.07	2.88	85.31	2.45	**105.24**	
X^2	0.74	4.55	0.87	1.22	36.00	1.03	**44.41**	

GCC	13.00	8.00	11.00	11.00	2.00	7.00	52.00	
GCC	18.00	32.00	39.00	0.00	64.00	1.00	154.00	
TOTAL	31.00	40.00	50.00	11.00	66.00	8.00	206.00	
Exp.	7.83	10.10	12.62	2.78	16.66	2.02		p-value
Exp.	23.17	29.90	37.38	8.22	49.34	5.98		4.52845E-14
X^2	3.42	0.44	0.21	24.35	12.90	12.28	**53.60**	
X^2	1.16	0.15	0.07	8.22	4.36	4.15	**18.10**	

GCT	3.00	6.00	6.00	2.00	5.00	3.00	25.00	
GCT	0.00	31.00	51.00	0.00	1.00	32.00	115.00	
TOTAL	3.00	37.00	57.00	2.00	6.00	35.00	140.00	
Exp.	0.54	6.61	10.18	0.36	1.07	6.25		p-value
Exp.	2.46	30.39	46.82	1.64	4.93	28.75		1.63107E-08
X^2	11.34	0.06	1.72	7.56	14.40	1.69	**36.76**	
X^2	2.46	0.01	0.37	1.64	3.13	0.37	**7.99**	

GCG	14.00	7.00	28.00	8.00	45.00	12.00	114.00	
GCG	0.00	2.00	13.00	23.00	1.00	57.00	96.00	
TOTAL	14.00	9.00	41.00	31.00	46.00	69.00	210.00	
Exp.	7.60	4.89	22.26	16.83	24.97	37.46		p-value
Exp.	6.40	4.11	18.74	14.17	21.03	31.54		4.91079E-20
X^2	5.39	0.91	1.48	4.63	16.06	17.30	**45.78**	
X^2	6.40	1.09	1.76	5.50	19.08	20.55	**54.37**	

GAT	12.00	6.00	13.00	7.00	54.00	8.00	100.00	
GAT	0.00	42.00	0.00	0.00	0.00	42.00	84.00	
TOTAL	12.00	48.00	13.00	7.00	54.00	50.00	184.00	
Exp.	6.52	26.09	7.07	3.80	29.35	27.17		p-value
Exp.	5.48	21.91	5.93	3.20	24.65	22.83		1.42774E-27
X^2	4.60	15.47	4.99	2.68	20.71	13.53	**61.98**	
X^2	5.48	18.41	5.93	3.20	24.65	16.11	**73.78**	

GTA	20.00	8.00	38.00	13.00	8.00	18.00	105.00	
GTA	25.00	9.00	2.00	0.00	41.00	6.00	83.00	
TOTAL	45.00	17.00	40.00	13.00	49.00	24.00	188.00	
Exp.	25.13	9.49	22.34	7.26	27.37	13.40		p-value
Exp.	19.87	7.51	17.66	5.74	21.63	10.60		2.86269E-14
X^2	1.05	0.24	10.98	4.54	13.71	1.58	**32.08**	
X^2	1.33	0.30	13.89	5.74	17.34	1.99	**40.58**	

GTT	93.00	31.00	8.00	17.00	6.00	4.00	159.00
GTT	0.00	28.00	0.00	5.00	2.00	32.00	67.00

TOTAL	93.00	59.00	8.00	22.00	8.00	36.00	226.00	
Exp.	65.43	41.51	5.63	15.48	5.63	25.33		p-value
Exp.	27.57	17.49	2.37	6.52	2.37	10.67		1.10057E-22
X^2	11.62	2.66	1.00	0.15	0.02	17.96	**33.41**	
X^2	27.57	6.31	2.37	0.36	0.06	42.62	**79.29**	

GTC	46.00	15.00	14.00	7.00	6.00	2.00	90.00	
GTC	10.00	0.00	0.00	0.00	34.00	66.00	110.00	
TOTAL	56.00	15.00	14.00	7.00	40.00	68.00	200.00	
Exp.	25.20	6.75	6.30	3.15	18.00	30.60		p-value
Exp.	30.80	8.25	7.70	3.85	22.00	37.40		3.98886E-28
X^2	17.17	10.08	9.41	4.71	8.00	26.73	**76.10**	
X^2	14.05	8.25	7.70	3.85	6.55	21.87	**62.26**	

GTG	2.00	4.00	7.00	9.00	9.00	5.00	36.00	
GTG	10.00	0.00	44.00	0.00	54.00	0.00	108.00	
TOTAL	12.00	4.00	51.00	9.00	63.00	5.00	144.00	
Exp.	3.00	1.00	12.75	2.25	15.75	1.25		p-value
Exp.	9.00	3.00	38.25	6.75	47.25	3.75		5.25983E-12
X^2	0.33	9.00	2.59	20.25	2.89	11.25	**46.32**	
X^2	0.11	3.00	0.86	6.75	0.96	3.75	**15.44**	

GAG	32.00	13.00	74.00	6.00	18.00	15.00	158.00	
GAG	8.00	20.00	0.00	36.00	0.00	36.00	100.00	
TOTAL	40.00	33.00	74.00	42.00	18.00	51.00	258.00	
Exp.	24.50	20.21	45.32	25.72	11.02	31.23		p-value
Exp.	15.50	12.79	28.68	16.28	6.98	19.77		1.10443E-26
X^2	2.30	2.57	18.15	15.12	4.42	8.44	**51.00**	
X^2	3.63	4.06	28.68	23.89	6.98	13.33	**80.57**	

GGA	17.00	24.00	18.00	25.00	27.00	133.00	244.00	
GGA	37.00	22.00	0.00	2.00	19.00	13.00	93.00	
TOTAL	54.00	46.00	18.00	27.00	46.00	146.00	337.00	
Exp.	39.10	33.31	13.03	19.55	33.31	105.71		p-value
Exp.	14.90	12.69	4.97	7.45	12.69	40.29		2.37162E-19
X^2	12.49	2.60	1.89	1.52	1.19	7.05	**26.74**	
X^2	32.77	6.82	4.97	3.99	3.13	18.49	**70.16**	

GGG	4.00	9.00	9.00	9.00	61.00	67.00	159.00	
GGG	1.00	1.00	0.00	0.00	0.00	18.00	20.00	
TOTAL	5.00	10.00	9.00	9.00	61.00	85.00	179.00	
Exp.	4.44	8.88	7.99	7.99	54.18	75.50		p-value
Exp.	0.56	1.12	1.01	1.01	6.82	9.50		0.001994726
X^2	0.04	0.00	0.13	0.13	0.86	0.96	**2.11**	
X^2	0.35	0.01	1.01	1.01	6.82	7.61	**16.80**	

GGC	7.00	4.00	11.00	10.00	6.00	17.00	55.00	
GGC	0.00	29.00	0.00	1.00	0.00	32.00	62.00	
TOTAL	7.00	33.00	11.00	11.00	6.00	49.00	117.00	
Exp.	3.29	15.51	5.17	5.17	2.82	23.03		p-value
Exp.	3.71	17.49	5.83	5.83	3.18	25.97		1.52492E-10
X^2	4.18	8.54	6.57	4.51	3.58	1.58	**28.97**	
X^2	3.71	7.58	5.83	4.00	3.18	1.40	**25.70**	

GGT	44.00	5.00	20.00	5.00	5.00	6.00	85.00	
GGT	0.00	0.00	0.00	33.00	0.00	2.00	35.00	
TOTAL	44.00	5.00	20.00	38.00	5.00	8.00	120.00	
Exp.	31.17	3.54	14.17	26.92	3.54	5.67		p-value
Exp.	12.83	1.46	5.83	11.08	1.46	2.33		2.92027E-18
X^2	5.28	0.60	2.40	17.85	0.60	0.02	**26.75**	
X^2	12.83	1.46	5.83	43.34	1.46	0.05	**64.97**	

Table B. The p-values for the comparisons between dRhodamine and BigDye v1.1 dye chemistries for each of the primers, Primers A1, B1, C1, and D1. Any p-value < 7.8125E-4 (0.05/64) is considered statistically significant with an overall α = 0.05 for that frame given the null hypothesis that the distribution of patterns is the same. In this example, most p-values exceed 7.8125E-4; the null hypothesis is not to be rejected. Comparisons between dye chemistries cannot be made.

	p-value			
Frame	A1	B1	C1	D1
AAA	6.16016E-17	2.45218E-07	2.45507E-18	5.68865E-05
AAC	7.42701E-14	2.53476E-12	1.97785E-15	6.67162E-07
ACA	1.74519E-05	2.65634E-09	9.88894E-19	2.80704E-27
ACC	0.898580318	2.53989E-19	2.39207E-05	8.55863E-17
ACT	2.40572E-12	9.36644E-47	6.01095E-05	9.84718E-06
ACG	2.39647E-09	8.44096E-28	5.29652E-11	3.47416E-10
AAT	8.55696E-15	2.99609E-25	1.65255E-25	8.31074E-06
ATA	0.00014045	2.02584E-06	0.000108357	8.39582E-11
ATT	9.231E-50	2.88039E-28	1.02567E-31	0.003788378
ATC	4.77926E-22	1.5425E-31	4.86079E-26	1.22799E-15
ATG	1.42378E-33	2.30898E-11	8.44338E-33	2.49644E-12
AAG	1.05149E-39	1.26433E-30	4.9977E-07	2.28566E-26
AGA	8.90613E-55	5.64318E-17	1.54464E-40	2.21826E-62
AGG	1.1577E-39	3.22146E-70	8.85743E-31	6.58819E-60
AGC	9.93676E-38	5.9707E-29	1.03736E-39	2.32074E-38
AGT	2.87085E-28	1.31455E-69	2.9264E-06	1.43778E-12
CAA	1.42852E-21	1.33925E-09	1.05113E-25	1.09032E-13
CAC	0.495224527	N/A	4.0599E-08	1.87137E-08
CCA	0.372296673	5.75097E-11	6.64381E-15	2.36207E-09
CCC	6.62206E-30	1.71613E-12	1.40869E-12	3.32305E-06
CCT	2.0189E-18	0.00022521	6.30199E-13	5.02676E-34
CCG	9.40118E-16	8.44796E-07	1.54692E-08	0.047958378
CAT	1.37468E-17	1.51086E-31	2.74849E-05	1.4824E-17
CTA	1.14762E-08	1.35099E-08	1.47031E-05	3.00249E-05
CTT	2.12123E-25	2.72968E-30	2.6684E-05	1.06678E-28
CTC	6.88923E-16	0.273186906	1.42705E-07	4.80945E-26
CTG	5.62689E-60	7.98395E-39	5.34485E-41	1.84013E-30
CAG	7.88052E-42	2.74303E-31	3.9521E-43	9.19128E-30
CGA	N/A	8.40556E-13	2.96881E-13	2.78319E-19
CGG	5.90774E-12	7.48635E-07	3.08516E-13	1.16316E-15
CGC	3.71598E-10	N/A	3.01308E-07	1.22487E-07
CGT	1.3526E-07	6.73872E-13	4.96619E-16	0.024679084
TAA	1.17761E-09	3.51214E-27	1.53755E-18	0.001062979
TAC	4.2101E-31	1.1553E-09	5.26601E-17	1.74571E-11
TCA	1.77551E-33	0.010295351	7.94784E-27	6.76191E-13
TCC	1.00373E-25	4.51852E-45	3.05193E-10	6.76137E-11
TCT	2.41946E-18	1.38466E-10	2.11567E-08	7.92172E-24
TCG	2.09432E-14	N/A	3.22366E-11	7.06554E-12
TAT	7.42205E-23	1.15031E-48	8.54536E-16	1.55971E-27
TTA	1.51034E-32	1.32654E-21	6.21845E-40	9.22466E-22
TTT	1.27486E-24	2.42045E-55	1.35895E-47	1.66901E-21
TTC	1.17765E-14	3.16892E-35	4.67497E-31	2.49665E-09
TTG	1.94128E-72	4.90151E-16	1.16604E-51	9.80693E-74
TAG	1.62758E-29	1.45973E-28	1.42781E-17	4.78699E-20
TGA	1.71049E-65	8.46536E-79	1.95185E-16	1.92539E-09
TGG	1.95282E-30	1.05512E-50	1.93418E-07	2.03096E-65
TGC	4.64701E-28	0.002893949	2.01504E-15	2.26355E-55
TGT	6.72914E-45	4.37992E-28	1.04523E-65	5.08113E-63
GAA	0.002901578	8.04619E-34	1.49219E-22	2.7239E-19
GAC	1.25675E-14	2.00355E-20	2.94479E-14	1.68173E-27
GCA	2.41945E-12	1.38709E-19	1.58114E-30	6.60941E-19
GCC	3.33371E-15	N/A	4.52845E-14	0.000411963
GCT	4.18256E-22	3.13126E-13	1.63107E-08	8.34279E-12
GCG	N/A	3.92678E-10	4.91079E-20	9.31409E-20
GAT	1.77492E-44	9.9718E-06	1.42774E-27	7.42707E-40
GTA	5.26794E-08	4.02333E-52	2.86269E-14	1.01273E-09
GTT	6.73137E-37	2.92847E-07	1.10057E-22	1.24276E-86
GTC	0.002355522	9.12813E-20	3.98886E-28	4.84494E-42
GTG	N/A	1.91843E-11	5.25983E-12	3.44158E-51
GAG	N/A	9.15203E-41	1.10443E-26	1.76721E-25

GGA	5.01032E-13	5.02106E-17	2.37162E-19	0.008376152
GGG	1.05882E-18	8.02414E-31	0.001994726	2.55739E-44
GGC	N/A	1.11532E-15	1.52492E-10	1.68975E-13
GGT	8.69175E-25	2.74521E-23	2.92027E-18	2.45509E-79

REFERENCES

Anonymous 2008, USDA-ARS, Mars and IBM to Sequence the Cocoa Genome . 2008, pp. 1.

Anonymous 2007a, CODIS—NDIS Statistics. **2008**, pp. 1.

Anonymous 2007b, National DNA Index System. **2008**, pp. 1.

Anonymous 2007c, What is CODIS. **2008**, pp. 1.

Anonymous Phred - Quality Base Calling. **2008**, pp. 1.

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

Anderson, T.D., J.P. Ross, R.K. Roby, D.A. Lee, and M.M. Holland, 1999, A validation study for the extraction and analysis of DNA from human nail material and its application to forensic casework. *Journal of forensic sciences,* **44**, pp. 1053-1056.

Andreasson, H., U. Gyllensten, and M. Allen, 2002, Real-time DNA quantification of nuclear and mitochondrial DNA in forensic analysis. *BioTechniques,* **33**, pp. 402-4, 407-11.

Andrews, R.M., I. Kubacka, P.F. Chinnery, R.N. Lightowlers, D.M. Turnbull, and N. Howell, 1999, Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nature genetics,* **23**, pp. 147.

Association for the Advancement of Artificial Intelligence, 2008, Expert systems. **July 19, 2008**, pp. 1.

Barber, R., 1992: *BONES : an expert system for diagnosis with fault models.* E. Horwood, 224 pp.

Bini, C., S. Ceccardi, D. Luiselli, G. Ferri, S. Pelotti, C. Colalongo, M. Falconi, and G. Pappalardo, 2003, Different informativeness of the three hypervariable mitochondrial DNA regions in the population of Bologna (Italy). *Forensic science international,* **135**, pp. 48-52.

Brown, W.M., M. George Jr, and A.C. Wilson, 1979, Rapid evolution of animal mitochondrial DNA. *Proceedings of the National Academy of Sciences of the United States of America,* **76**, pp. 1967-1971.

Budowle, B., M.R. Wilson, J.A. DiZinno, C. Stauffer, M.A. Fasano, M.M. Holland, and K.L. Monson, 1999, Mitochondrial DNA regions HVI and HVII population data. *Forensic science international,* **103**, pp. 23-35.

Budowle, B., J. Smith, T. Moretti, and J. DiZinno, 2000: *DNA Typing Protocols: Molecular Biology and Forensic Analysis.* Eaton Publishing,

Cann, R.L., M. Stoneking, and A.C. Wilson, 1987, Mitochondrial DNA and human evolution. *Nature,* **325**, pp. 31-36.

Chan, E.Y., 2005, Advances in sequencing technology. *Mutation research,* **573**, pp. 13-40.

Collins, P.J., L.K. Hennessy, C.S. Leibelt, R.K. Roby, D.J. Reeder, and P.A. Foxall, 2004, Developmental validation of a single-tube amplification of the 13 CODIS STR loci, D2S1338, D19S433, and amelogenin: the AmpFlSTR Identifiler PCR Amplification Kit. *Journal of forensic sciences,* **49**, pp. 1265-1277.

Copeland, W.C., 2002, Mitochondrial DNA. Methods and protocols. Introduction. *Methods in molecular biology (Clifton, N.J.),* **197**, pp. v-vi.

Dabrowski, C. E., and E. N. Fong, 1991: *Guide to Expert System Building Tools for Microcomputers.* National Institute of Standards and Technology, U.S. Department of Commerce, 141 pp.

DNA Advisory Board, cited 2000: Quality Assurance Standards for Forensic DNA Testing Laboratories and for Convicted Offender DNA Databasing Laboratories. [Available online at http://www.fbi.gov/hq/lab/fsc/backissu/july2000/codispre.htm.]

Edwards, M.C., and R.A. Gibbs, 1994, Multiplex PCR: advantages, development, and applications. *PCR methods and applications,* **3**, pp. S65-75.

Engelmore, R.S., and E. Feigenbaum, 1993, Knowledge-Bases Systems in Japan: Introduction. **2008**

Englund, P.T., 1972, The 3'-terminal nucleotide sequences of T7 DNA. *Journal of Molecular Biology,* **66**, pp. 209-224.

Englund, P.T., 1971, Analysis of nucleotide sequences at 3' termini of duplex deoxyribonucleic acid with the use of the T4 deoxyribonucleic acid polymerase. *The Journal of biological chemistry,* **246**, pp. 3269-3276.

Ewing, B., L. Hillier, M.C. Wendl, and P. Green, 1998, Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome research,* **8**, pp. 175-185.

França, L.T.C., E. Carrilho, and T.B.L. Kist, 2002, A review of DNA sequencing techniques. *Quarterly Reviews of Biophysics,* **35**, pp. 169.

Giles, R.E., H. Blanc, H.M. Cann, and D.C. Wallace, 1980, Maternal inheritance of human mitochondrial DNA. *Proceedings of the National Academy of Sciences of the United States of America,* **77**, pp. 6715-6719.

Gray, M.W., 1992, The endosymbiont hypothesis revisited. *International review of cytology,* **141**, pp. 233-357.

Grivell, L.A., 1997, Mitochondria. In *McGraw-Hill Encyclopedia of Science & Technology,* Anonymous (New York: McGraw-Hill, 1997).

Hunt, V. D., 1986: *Artificial intelligence & expert systems sourcebook.* Chapman and Hall, 315 p. pp.

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

Lutz, S., H. Wittig, H.J. Weisser, J. Heizmann, A. Junge, N. Dimo-Simonin, W. Parson, J. Edelmann, K. Anslinger, S. Jung, and C.

Augustin, 2000, Is it possible to differentiate mtDNA by means of HVIII in samples that cannot be distinguished by sequencing the HVI and HVII regions? *Forensic science international,* **113**, pp. 97-101.

Melton, T., S. Clifford, M. Kayser, I. Nasidze, M. Batzer, and M. Stoneking, 2001, Diversity and heterogeneity in mitochondrial DNA of North American populations. *Journal of forensic sciences,* **46**, pp. 46-52.

Nickerson, D.A., V.O. Tobe, and S.L. Taylor, 1997, PolyPhred: automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing. *Nucleic acids research,* **25**, pp. 2745-2751.

Patel, V. L., and G. J. Groen, 1991, Real versus artificial expertise: The development of cognitivie models of clinical reasoning. In *Proceedings of the Third Conference on Artificial Intelligence in Medicine: 1991,* Maastricht, Limburg, Netherlands, (Berlin:Springer-Verlag),

Ralston, A., and and Reilly, Edwin D. (Eds.), 1993, *Encyclopedia of computer science,* pp. 1558 (New York, New York: Van Nostrand Reinhold Co.).

Rieder, M.J., S.L. Taylor, V.O. Tobe, and D.A. Nickerson, 1998, Automating the identification of DNA variations using quality-based fluorescence re-sequencing: analysis of the human mitochondrial genome. *Nucleic acids research,* **26**, pp. 967-973.

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

Rosenthal, A., and D.S. Charnock-Jones, 1992, New protocols for DNA sequencing with dye terminators. *DNA sequence : the journal of DNA sequencing and mapping,* **3**, pp. 61-64.

Sanger, F., and A.R. Coulson, 1975, A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of Molecular Biology,* **94**, pp. 441-448.

Sanger, F., S. Nicklen, and A.R. Coulson, 1977, DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America,* **74**, pp. 5463-5467.

Shortliffe, E. H., 1976: *MYCIN: Computer-based Medical Consultations.* Elsevier,

Smith, L.M., J.Z. Sanders, R.J. Kaiser, P. Hughes, C. Dodd, C.R. Connell, C. Heiner, S.B. Kent, and L.E. Hood, 1986, Fluorescence detection in automated DNA sequence analysis. *Nature,* **321**, pp. 674-679.

Sullivan, K.M., A. Mannucci, C.P. Kimpton, and P. Gill, 1993, A rapid and quantitative DNA sex test: fluorescence-based PCR analysis of X-Y homologous gene amelogenin. *BioTechniques,* **15**, pp. 636-8, 640-1.

Timken, M.D., K.L. Swango, C. Orrego, and M.R. Buoncristiani, 2005, A duplex real-time qPCR assay for the quantification of human nuclear and mitochondrial DNA in forensic samples: implications for quantifying DNA in degraded samples. *Journal of forensic sciences,* **50**, pp. 1044-1060.

Tracy, T.E., and L.S. Mulcahy, 1991, A simple method for direct automated sequencing of PCR fragments. *BioTechniques,* **11**, pp. 68-75.

Venter, J.C., M.D. Adams, E.W. Myers, P.W. Li, R.J. Mural, G.G. Sutton, H.O. Smith, M. Yandell, C.A. Evans, R.A. Holt, J.D. Gocayne, P. Amanatides, R.M. Ballew, D.H. Huson, J.R. Wortman, Q. Zhang, C.D. Kodira, X.H. Zheng, L. Chen, M. Skupski, G. Subramanian, P.D. Thomas, J. Zhang, G.L. Gabor Miklos, C. Nelson, S. Broder, A.G. Clark, J. Nadeau, V.A. McKusick, N. Zinder, A.J. Levine, R.J. Roberts, M. Simon, C. Slayman, M. Hunkapiller, R. Bolanos, A. Delcher, I. Dew, D. Fasulo, M. Flanigan, L. Florea, A. Halpern, S. Hannenhalli, S. Kravitz, S.

Levy, C. Mobarry, K. Reinert, K. Remington, J. Abu-Threideh, E. Beasley, K. Biddick, V. Bonazzi, R. Brandon, M. Cargill, I. Chandramouliswaran, R. Charlab, K. Chaturvedi, Z. Deng, V. Di Francesco, P. Dunn, K. Eilbeck, C. Evangelista, A.E. Gabrielian, W. Gan, W. Ge, F. Gong, Z. Gu, P. Guan, T.J. Heiman, M.E. Higgins, R.R. Ji, Z. Ke, K.A. Ketchum, Z. Lai, Y. Lei, Z. Li, J. Li, Y. Liang, X. Lin, F. Lu, G.V. Merkulov, N. Milshina, H.M. Moore, A.K. Naik, V.A. Narayan, B. Neelam, D. Nusskern, D.B. Rusch, S. Salzberg, W. Shao, B. Shue, J. Sun, Z. Wang, A. Wang, X. Wang, J. Wang, M. Wei, R. Wides, C. Xiao, C. Yan, A. Yao, J. Ye, M. Zhan, W. Zhang, H. Zhang, Q. Zhao, L. Zheng, F. Zhong, W. Zhong, S. Zhu, S. Zhao, D. Gilbert, S. Baumhueter, G. Spier, C. Carter, A. Cravchik, T. Woodage, F. Ali, H. An, A. Awe, D. Baldwin, H. Baden, M. Barnstead, I. Barrow, K. Beeson, D. Busam, A. Carver, A. Center, M.L. Cheng, L. Curry, S. Danaher, L. Davenport, R. Desilets, S. Dietz, K. Dodson, L. Doup, S. Ferriera, N. Garg, A. Gluecksmann, B. Hart, J. Haynes, C. Haynes, C. Heiner, S. Hladun, D. Hostin, J. Houck, T. Howland, C. Ibegwam, J. Johnson, F. Kalush, L. Kline, S. Koduru, A. Love, F. Mann, D. May, S. McCawley, T. McIntosh, I. McMullen, M. Moy, L. Moy, B. Murphy, K. Nelson, C. Pfannkoch, E. Pratts, V. Puri, H. Qureshi, M. Reardon, R. Rodriguez, Y.H. Rogers, D. Romblad, B. Ruhfel, R. Scott, C. Sitter, M. Smallwood, E. Stewart, R. Strong, E. Suh, R. Thomas, N.N. Tint, S. Tse, C. Vech, G. Wang, J. Wetter, S. Williams,

M. Williams, S. Windsor, E. Winn-Deen, K. Wolfe, J. Zaveri, K. Zaveri, J.F. Abril, R. Guigo, M.J. Campbell, K.V. Sjolander, B. Karlak, A. Kejariwal, H. Mi, B. Lazareva, T. Hatton, A. Narechania, K. Diemer, A. Muruganujan, N. Guo, S. Sato, V. Bafna, S. Istrail, R. Lippert, R. Schwartz, B. Walenz, S. Yooseph, D. Allen, A. Basu, J. Baxendale, L. Blick, M. Caminha, J. Carnes-Stine, P. Caulk, Y.H. Chiang, M. Coyne, C. Dahlke, A. Mays, M. Dombroski, M. Donnelly, D. Ely, S. Esparham, C. Fosler, H. Gire, S. Glanowski, K. Glasser, A. Glodek, M. Gorokhov, K. Graham, B. Gropman, M. Harris, J. Heil, S. Henderson, J. Hoover, D. Jennings, C. Jordan, J. Jordan, J. Kasha, L. Kagan, C. Kraft, A. Levitsky, M. Lewis, X. Liu, J. Lopez, D. Ma, W. Majoros, J. McDaniel, S. Murphy, M. Newman, T. Nguyen, N. Nguyen, M. Nodell, S. Pan, J. Peck, M. Peterson, W. Rowe, R. Sanders, J. Scott, M. Simpson, T. Smith, A. Sprague, T. Stockwell, R. Turner, E. Venter, M. Wang, M. Wen, D. Wu, M. Wu, A. Xia, A. Zandieh, and X. Zhu, 2001, The sequence of the human genome. *Science (New York, N.Y.),* **291**, pp. 1304-1351.

Whitaker, J.P., T.M. Clayton, A.J. Urquhart, E.S. Millican, T.J. Downes, C.P. Kimpton, and P. Gill, 1995, Short tandem repeat typing of bodies from a mass disaster: high success rate and characteristic amplification patterns in highly degraded samples. *BioTechniques,* **18**, pp. 670-677.

Zakeri, H., G. Amparo, S.M. Chen, S. Spurgeon, and P.Y. Kwok, 1998, Peak height pattern in dichloro-rhodamine and energy transfer dye terminator sequencing. *BioTechniques,* **25**, pp. 406-10, 412-4.