Musslewhite, Pamela C., Optimization of Filter Metrics for Mitochondrial DNA Sequence Analysis. Master of Science (Biomedical Sciences, Forensic Genetics) August, 2009, 55 pp., 13 tables, 32 figures, references 11 titles.

Quality metrics translate sequence information into numerical values which allows a software program to filter through data without human intervention. Primer specific settings for the trace score and contiguous read length in Sequence Scanner Software v1.0 (*Applied Biosystems, Foster City, CA*) were established using a calibration dataset of 2,817 sequence traces and validated using a second dataset of 5,617 sequence traces. Prior to optimization 51.7% of the samples required manual intervention while 28.4% require review after optimization. An evaluation of signal intensity and signal to noise ratio variables was performed and no trend was recognized for predictive modeling. Use of quality values per peak to ascertain confidence in the base call was evaluated and found to be a feasible parameter for sample quality assessment and confident base calling.

OPTIMIZATION OF FILTER METRICS FOR MITOCHONDRIAL DNA SEQUENCE ANALYSIS

INTERNSHIP PRACTICUM REPORT

Presented to the Graduate Council of the
Graduate School of Biomedical Sciences
University of North Texas
Health Science Center at Fort Worth
Partial Fulfillment of the Requirements

For the Degree of

MASTER OF SCIENCE

By

Pamela C. Musslewhite, B.S.
Fort Worth, Texas

August 2009

ACKNOWLEDGEMENTS

I would like to express my gratitude to the University of North Texas Center for Human Identification Research & Development Laboratory for allowing me to conduct my internship at their location and for providing an abundance of data. I would also like to thank Dr. Arthur J. Eisenberg, Dr. Bruce Budowle, Dr. John V. Planz, Dr. Joseph E. Warren, Dr. Rhonda K. Roby, and Dr. Suzanne Gonzalez for their exemplary support and guidance. In addition, I would like to acknowledge Ms. Jennifer Thomas, Ms. Nicole Phillips, and my classmates for their assistance and helpful discussions. Last but not least, I wish to extend many thanks to my loving family for their faith in me during academic endeavors.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER I

BACKGROUND

Forensic DNA testing encompasses several types of genetic markers which include: autosomal short tandem repeats (STRs), Y chromosome STRs, and mitochondrial DNA (mtDNA). When nuclear DNA (nDNA) is unavailable, insufficient, or degraded for the standard nuclear STR analysis, mtDNA analysis may be more successful. Typical samples for mtDNA analysis may include hair shafts, hair lacking roots/tissue, bones (long bones and teeth), and decayed soft tissue samples. (1) Typical cases include skeletal remains for missing persons cases and hair found at the scene of the crime (such as pubic hairs during a sexual assault, hairs found in a victim's hand from defensive fighting, or even discarded/shed hair on clothes or masks) (2).

Similar to nuclear DNA, mtDNA is found in all cell types except erythrocytes. However, differences exist between nDNA and mtDNA. A brief side-by-side comparison is detailed in Table 1. The mitochondrial genome (mtGenome) is a small, double-stranded, covalently closed circular molecule. This genome spans approximately 16,569 base pairs and contains 37 essential genes necessary for the oxidative phosphorylation process. Twenty-two of these genes code for transfer RNAs (tRNAs), 13 code for proteins, and 2 code for ribosomal RNAs (rRNAs). The composition of the genome is characterized by asymmetric nucleotide distribution which results in light and heavy strands (3). The outer strand, termed the heavy (H) strand, is composed of a large number of guanines which are the heaviest in molecular weight of the four nucleotides (4). The complementary strand, termed the light (L) strand, is cytosine rich. The genome has a

approximate 1.3 kb control region, also known as the displacement loop, or more commonly the D-loop. The control region does not code for any gene products however does house the origin site for heavy strand replication.

Table 1 - Comparison of Nuclear DNA and Mitochondrial DNA Characteristics

Characteristic	**Nuclear DNA (nDNA)**	**Mitochondrial DNA (mtDNA)**
Genome Size	> 3 billion bp	~16,569 bp
No. copies per cell	2 (one from each parent)	can be > 1000
Genome Structure	Linear, chromosomal packaging	Circular
Inheritance pattern	Mendelian (from mother and father)	Non-Mendelian (from mother only)
Chromosomal Pairing	Diploid	Haploid
Generational Recombination	Yes	No
Replication Repair	Yes	No
Unique	Yes (except for identical twins)	No (shared among maternal relatives)
Mutation Rate	Low	5-10x higher than that of nDNA

The forensic testing community focuses its attention on the DNA sequences of the control region. Within the control region are two hypervariable regions, commonly referred to as HV1 and HV2. This region contains high degree of variability among individuals. Hypervariable region 1 (HV1) is approximately 341 bp in length while hypervariable region 2 (HV2) is approximately 267 bp long (5). (Figure 1)

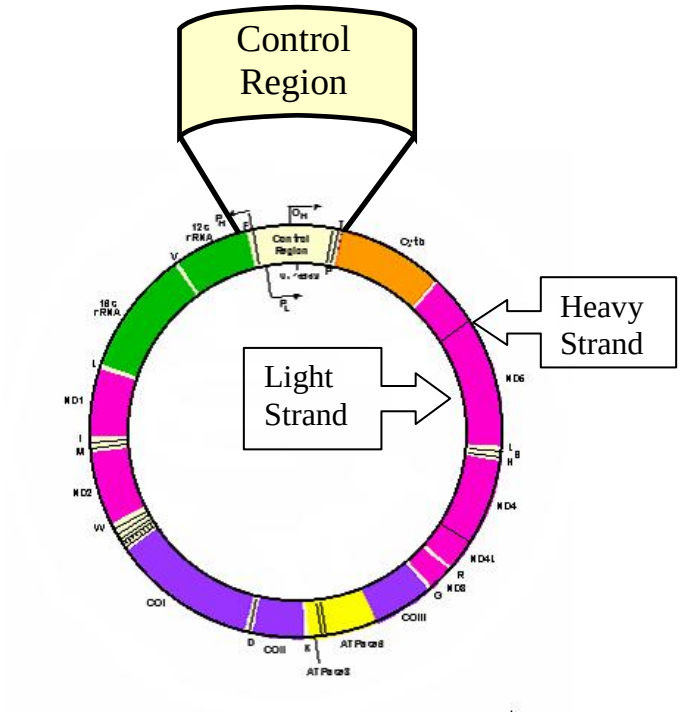


Figure 1 – Human Mitochondrial DNA Genome Map (6)

Mitochondrial DNA follows a non-Mendelian pattern of inheritance and is maternally transmitted. The mitochondria contained in a spermatozoan are found in the tail which typically do not enter the oocyte. In the event that paternal mitochondria enter a fertilized egg, they are selectively destroyed due to a ubiquitin tag that is added during spermatogenesis. This tag targets sperm mitochondria for degradation by the newly formed embryo's cellular machinery, such as proteasomes and lysosomes. (7) Compared to nuclear DNA analysis, mtDNA analysis has limited distinguishing power with respect to discriminating maternally related individuals by haplotype, unless a mutation has occurred. An individual's mtDNA type is known as a haplotype due to the haploid nature of inheritance. (Figure 2)

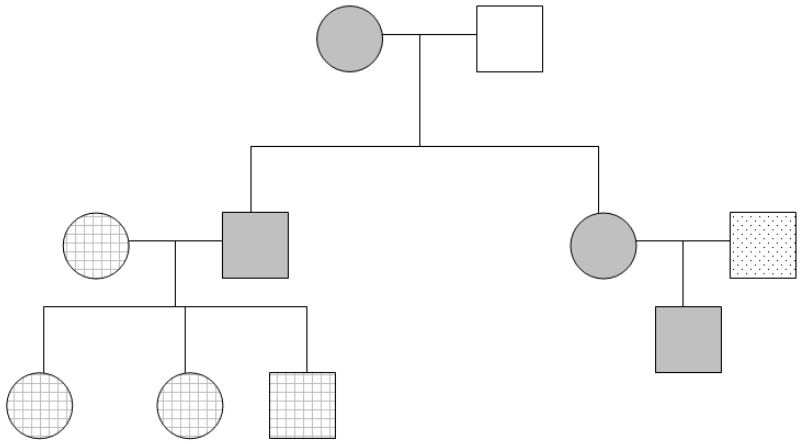


Figure 2 – Maternal Inheritance Pedigree. Circles represent females and squares represent males. Shading/patterns represent shared mitochondrial DNA sequences. All maternal relatives share the same mtDNA.

A cell contains hundreds of mitochondria and each mitochondrion carries four to five mtDNA molecules resulting in thousands of copies of the genome (8). The naturally abundant number of copies of mtDNA is one of the reasons it is more likely to survive harsh environmental degradation factors such as heat and humidity. Another protective feature is that the genome is circular, allowing protection from exonucleases. Finally, the fact that the genome is encapsulated within a double walled organelle provides additional protection.

Unlike nDNA, mtDNA does not undergo recombination between generations, however has limited DNA repair capability which results in higher observed mutation rates. The forensic advantage of having a higher mutation rate is that it facilitates the development of variability among maternal relatives which would otherwise not be observed.

When an individual carries more than one mtDNA haplotype, they are said to exhibit heteroplasmy. Heteroplasmy may be observed between different tissues, between different cells of the same tissue, or between different mitochondrion within a single cell. Heteroplasmy is most commonly seen in hair samples. There are two types of heteroplasmy: 1) sequence heteroplasmy and 2) length heteroplasmy. Sequence heteroplasmy is defined as the observance of two alternate nucleotides at a single sequence position, which appear as superimposed peaks on an electropherogram. There are "hot spots" for heteroplasmy that have been documented within both HV1 and HV2 (Figure 3a). C-stretches are homopolymeric cytosine sequences that may be found in both HV1 and HV2. One hypothesis for why these C-stretches occur is slippage at the C-stretch during replication. Another possible explanation is the mixture of length variants in the cells. (Figure 3b) A length heteroplasmy generates sequencing difficulties downstream of the homopolymeric regions which generates sequences out of register. The end result usually is unresolved sequences 3' to the length heteroplasmy. One way to sequence downstream is to utilize alternative (forward and reverse) sequencing primers. (1) Another solution is to utilize two sequencing primers in the same direction from two separate amplifications to provide confirmation of the sequence.

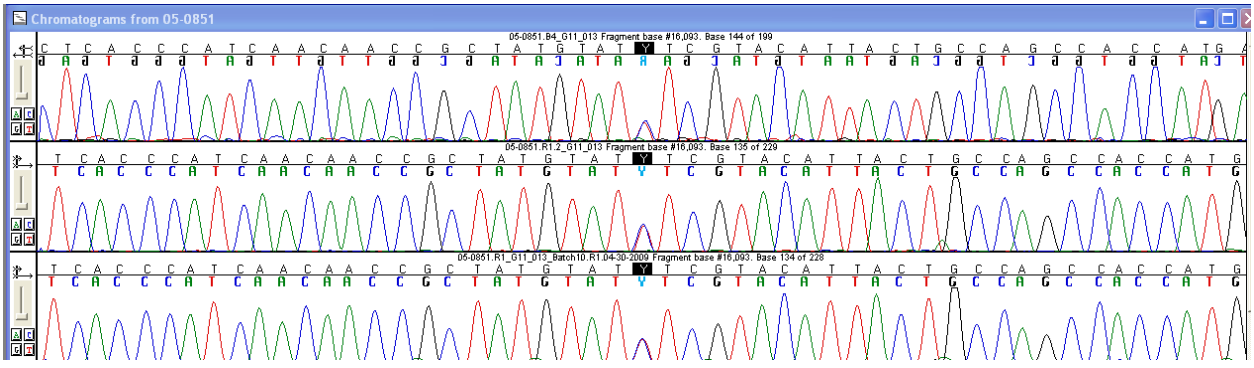


Figure 3a – Confirmed Sequence Heteroplasmy at Position 16093

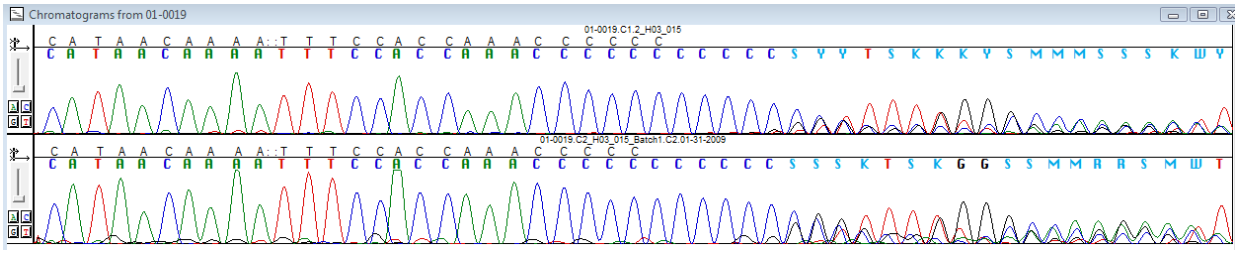


Figure 3b – Confirmed Length Heteroplasmy in HV2

CHAPTER II

INTRODUCTION

The University of North Texas Center for Human Identification (UNTCHI) consists of the Laboratory of Forensic Anthropology and the Laboratory for Molecular Identification. The Laboratory for Molecular Identification is housed on the University of North Texas Health Science Center's campus in Fort Worth, Texas. Collocated within this forensic identification laboratory is the Research & Development Laboratory (RDL) for the Center for Human Identification. The RDL was tasked with building a population database. For this high throughput processing project, 1,000 male buccal swabs were processed. These samples were broken into 12 batches. Within each batch there were 86 samples and accompanying controls. An autosomal short tandem repeat (STR), Y-STR, and mitochondrial DNA (mtDNA) profile was generated for each sample.

Laboratories typically perform two amplifications to minimally obtain sequence information from hypervariable region 1 (HV1) and hypervariable region 2 (HV2). However, the RDL performed a single amplification of a 1.1kb fragment that encompassed both HV1 and HV2. (Figure 4) This single large amplicon was generated using primers R1 (forward) and R2 (reverse).

Table 2 - Cycle Sequencing Primers with 5'-3' Sequence and Base Positions. Arrow indicates forward (→) or reverse(←) sequencing direction.

Primer	**Sequence (5' to 3')**	**Base Positions**
R1 →	CACCAGTCTTGTAAACCGGAGA	15910-15931
B1 ←	GAGGATGGTGGTCAAGGGAC	16391 - 16500
C1 →	CTCACGGGAGCTCTCCATGC	29-48
R2 ←	CTTTGGGGTTTGGTTGGTTC	545 - 564
A4 →	CCCCATGCTTACAAGCAAGT	16190-16209
B4 ←	TTTGATGTGGATTGGGTTT	16164 - 16182
C2 →	TTATTTATCGCACCTACGTTCAAT	154-177
D2 ←	GGGGTTTGGTGGAAATTTTTTG	285 - 306

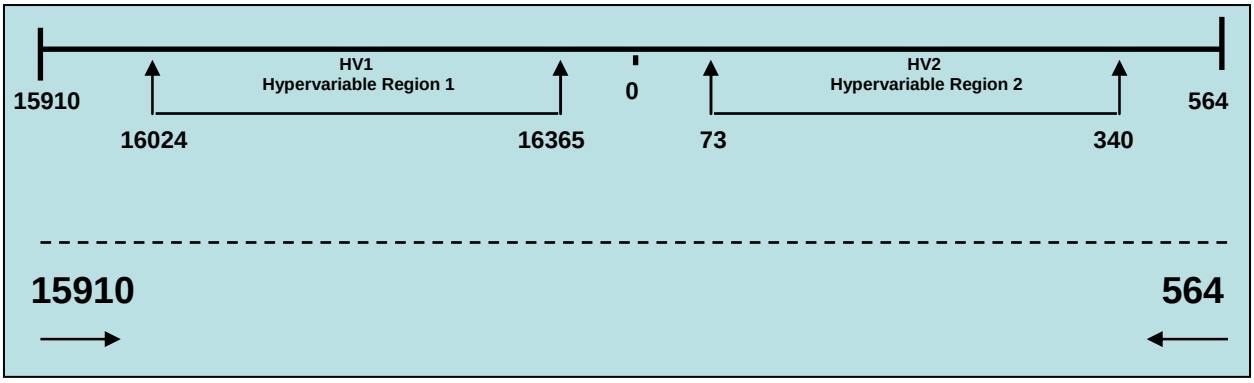


Figure 4 – Diagram of Control Region. Minimally laboratories try to obtain sequence information from sequence positions 16024 to 16365 (HV1) and 73 to 340 (HV2). The RDL performs a single amplification to obtain sequence information from positions 15910 to 564.

Upon initial evaluation of their data, the RDL team determined that 30% of the samples from the population being analyzed contained a homopolymeric stretch in HV1 (16184-16193) and 60% of the samples contained a length heteroplasmy in HV2 (303-315). Cycle sequencing was performed with eight sequencing primers to obtain coverage of the entire amplicon (Figure 5).

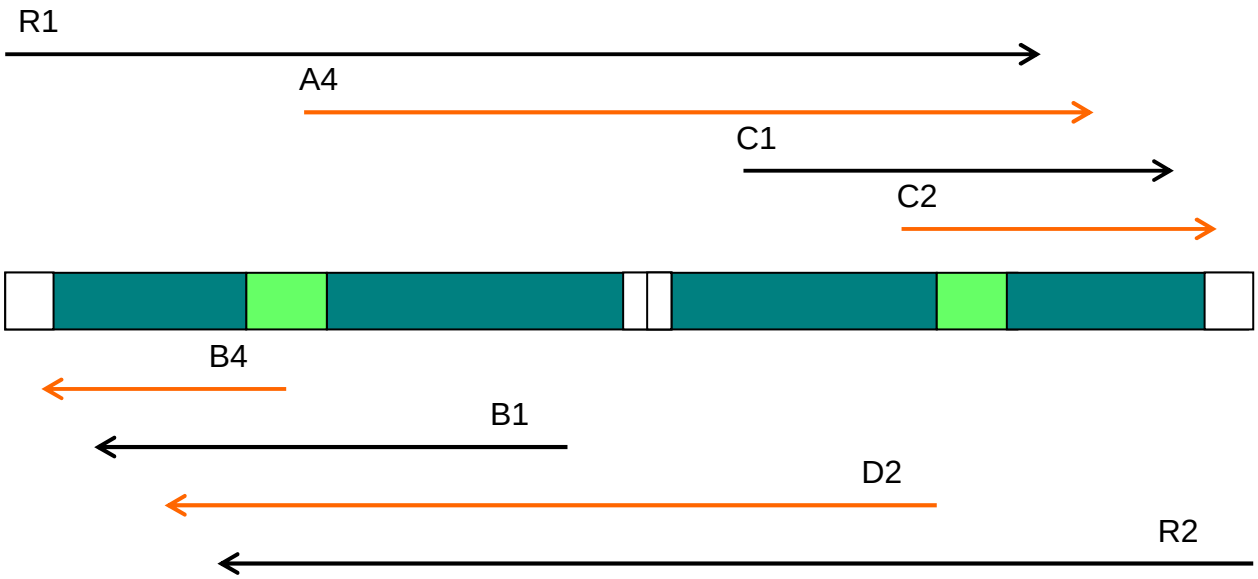


Figure 5 – Cycle Sequencing Primers. Green regions represent HV1 and HV2 while the green regions represent the homopolymeric stretch in HV1 and the length heteroplasmy in HV2. White area is the extra information obtained by performing a single large amplification.

Quality metrics translate sequencing information into numerical values which allows a software program to filter through data for you. Currently, the forensic DNA testing community is evaluating software tools, such as expert systems, to reduce backlogs. An expert system for nuclear DNA (nDNA) is defined by the forensic community as a software program or set of software programs that performs the following functions without any human intervention: identifies peaks/bands, assigns alleles, ensures data meet laboratory defined criteria, describes rationale behind decisions, and makes no incorrect calls. An expert system applies "if-then" statements to make decisions on the quality of data, automate allele calling, and must not make any incorrect allele calls. The software must provide justification for each decision. (9) Typical examples of the parameters used in STR analysis include allele number (AN), peak height imbalance (PHR), and off-scale data (OS). These parameters are laboratory-defined criteria. The

rule firings are of a specific shape and are color-coded. If the sample yields good quality data and meets all the thresholds set by the user, a green square can be seen for each particular parameter. For example, if data do not meet the user defined threshold, a rule is fired drawing the analyst's attention to that particular sample or locus via a yellow triangle. If a locus or sample fails, this is signified with a red octagon.

Expert systems have the potential to streamline data analysis and reduce backlogs within laboratories. An expert system for sequence analysis would reduce the amount of time an analyst must spend reviewing sequence data and therefore increase the throughput of a laboratory. Expert systems may also reduce the potential for human error, as the process is automated, consistent, and accurate. Implementation of expert systems within a laboratory reduces analysis time; therefore, freeing the analyst for other duties. The UNTCHI RDL is trying to define quality metrics for sequence data so that a software program can be developed to follow similar rules.

Software programs are used by analysts to align and analyze mtDNA sequence data. However, before analysis of the sequence data is performed, the RDL analysts employ a software program called Sequence Scanner Software v1.0 (Applied Biosystems) for quick quality assessment of their data. Sequence Scanner Software is a free software program that allows the user to display, edit, trim, export, and generate quality assessments of sequencing files: specifically those files generated by ABI PRISM® instruments. The files imported into the Trace Manager of Sequence Scanner Software are processed sequence *.ab1* files. Within this software program, the user can set trace score, contiguous read length (CRL), quality value (QV), and window size thresholds as well as color coordination. There are several views a user can select to make quick quality assessments. The software produces several different types of reports such as the Plate Report, Quality Control (QC) Report, and Signal Strength Report. Each of these reports

contains a hyperlink to the trace file which allows the user to view the sequence in a trace viewer. Also in Sequence Scanner Software, the user can export files of several types (.txt, .pdf, .xls, .jpg, etc). Quality assessments can be made quickly with quality metrics defined by the user as poor, medium, and good. The color coordination capability enhances the user's ability to quickly assess the quality of a sequence and decide whether to review or automatically allow ("pass") or reject ("fail") the data.

Sequence Scanner Software may prove to be a useful tool when developing a sequence expert system. Sample sequences, also known as trace files, are imported into the Trace Manager following a base calling process. Within the Trace Manager view, each trace file has a hyperlink to the electropherogram sequence window where the user can scan the sequence and manually trim, edit, and review the sequence. Once trace files have been imported into Sequence Scanner's Trace Manager, the user can view reports. One useful report for quick quality assessment is the Plate Report. It provides a "thumbnail" view of each sample and contains a hyperlink to the electropherogram. There are seven reports produced by Sequence Scanner Software: Quality Control, Plate, Trace Score, CRL, CRL Distribution, QV20+, and Signal Strength reports. In addition, the user can modify settings to show the signal to noise ratio rather than the signal strength. Within the electropherogram window, a user can view: analyzed trace file, raw data, analyzed + raw, annotation, sequence (as text), or electrophoresis data that includes voltage, current, and temperature of that individual sample.

High trace score values with low CRL values typically indicate that a good sequence has a stretch of poor base quality which is often observed with homopolymeric regions. During Dr. Roby's evaluation, the trace score thresholds were set at 42 and 44 and the CRL thresholds were set at 300 and 400. (10) The RDL staff believed these thresholds could be less stringent. When

the analysts began reviewing these sequence data, the analysts modified the thresholds until arriving at the current settings of 20 and 34 for the trace score and 200 and 400 for the CRL. This software can be used in an expert-system-like manner; however, no expert system currently exists for mtDNA sequence analysis.

The RDL has implemented the aforementioned filter metrics. Within Sequence Scanner Software v1.0 there are two quality metrics that Dr. Roby proposed and the laboratory is currently reviewing to assess the sequence quality of a sample: trace score and contiguous read length (CRL). The trace score is the average base call quality value of bases in the post-trim sequence and this value ranges from zero to 100. A user can set two thresholds with color coordination. Currently, for a good quality sequence that requires no analyst intervention, the color green is used and the value is 35 to 100. For a poor quality sequence this trace score is colored red and the value is defined as zero to 20. The middle range, 21 to 34, is colored yellow which allows the analyst to quickly assess the particular sequence to be reviewed and check for quality before continuing to analyze the sample. The second quality metric is the CRL which is the longest uninterrupted stretch of bases with a quality value equal to or greater than a specified limit. To calculate the CRL, the software takes into account not only the quality value of a single base but of those bases adjacent to it that make up a specified window size. Window size and quality value thresholds are set at the default setting of 20. The CRL, as a quality metric, also allows the user to set two thresholds. At present, good quality sequence is coded green and set to be above 400, the poor quality sequence is colored red and the value ranges from zero to 200, while the middle range is between 201 and 400 and colored yellow.

Analysts will first launch the Plate Report in Sequence Scanner. (Figure 6) This report is a bird's eye view of the entire 96-well plate and allows quick assessment of the overall quality of

the plate. Each thumbnail represents a single well. The colored bar on a thumbnail signifies that sequence's trace score value. The trace score legend can be seen at the top left on the report. A green trace score is high quality sequence and is greater than or equal to 35, a yellow trace score is medium quality sequence that should be reviewed and is between 21 and 34, and the red trace score is low quality sequence and is less than or equal to 20. Another benefit of the Plate Report is that it allows the analyst to review the controls to ensure they performed as expected as well as scanning the plate for electrophoretic injection problems.

Figure 6 – Plate Report from Sequence Scanner Software v1.0. A Plate Report provides a thumbnail view of an entire 96-well plate for quick quality assessment of the plate. Colored bars of each thumbnail represent the trace score value for that well.

Another view is the QC (quality control) Report generated by Sequence Scanner. (Figure 7) This report displays both the trace score legend and the CRL legend. Similar to the trace score, the CRL is color coded. High CRL is colored green and has a value greater than or equal to 401, medium CRL is yellow and has a value between 201 and 400, and a low CRL is red and is less than or equal to 200. Analysts will commonly use these color codes to refer to the sample quality. (Table 3) For example, a sample with a green trace score and a green CRL is called GG ("green green"). A GG sample does not require manual intervention as the sample is considered good quality and passes. A GY sample has a green trace score and a yellow CRL. This color code does require manual review by an analyst. A poor quality failing sample has a red trace score and a red CRL, RR ("red red"). The lead analyst at the RDL would print the QC Report for each plate and manually review any necessary samples noting comments about each requiring review. (Figure 7)

Table 3 – Current Actions Taken Based on Quality Assessment. A "GY" color code signifies a green trace score and a yellow CRL which alerts the analyst to review that sequence.

Color Code	Action Taken
GG	Pass
GY	Review
GR	Review
YG	Review
YY	Review
YR	Review
RG	Review
RY	Review
RR	Fail

Applied Biosystems

Printed on Apr 05, 2009
Sequence Scanner 1.0

Legend

Trace Score: 20 (low / med) 34 (med / high)

CRL: 200 (short / med) 400 (med / long)

Trace File Name	Well#	Cap#	Trace Score	CRL	QV20+	Signal Intensity A	C	G	T	Comments
04-2009.ab1										
03-0444.C1.2_C03_005_Batch6.C1.2.04-04-2009.ab1	C3	5	35	174	223	460	259	537	523	reinj
03-0445.C1.2_D03_007_Batch6.C1.2.04-04-2009.ab1	D3	7	27	259	381	316	177	301	325	C
03-0446.C1.2_E03_009_Batch6.C1.2.04-04-2009.ab1	E3	9	13	0	51	14	10	12	13	reinj
03-0447.C1.2_F03_011_Batch6.C1.2.04-04-2009.ab1	F3	11	34	499	488	475	265	447	455	
03-0448.C1.2_G03_013_Batch6.C1.2.04-04-2009.ab1	G3	13	38	257	384	253	153	235	288	C
03-0449.C1.2_H03_015_Batch6.C1.2.04-04-2009.ab1	H3	15	34	488	454	303	180	270	320	
03-0450.C1.2_A04_002_Batch6.C1.2.04-04-2009.ab1	A4	2	27	246	381	126	70	112	132	C
03-0451.C1.2_B04_004_Batch6.C1.2.04-04-2009.ab1	B4	4	34	481	486	220	123	203	225	
03-0452.C1.2_C04_006_Batch6.C1.2.04-04-2009.ab1	C4	6	31	461	447	41	26	36	41	C
03-0453.C1.2_D04_008_Batch6.C1.2.04-04-2009.ab1	D4	8	36	491	472	522	305	438	510	
03-0454.C1.2_E04_010_Batch6.C1.2.04-04-2009.ab1	E4	10	36	261	264	261	134	257	262	reinj. primer
03-0455.C1.2_F04_012_Batch6.C1.2.04-04-2009.ab1	F4	12	34	494	467	219	120	206	226	
03-0456.C1.2_G04_014_Batch6.C1.2.04-04-2009.ab1	G4	14	30	492	430	220	119	204	226	C
03-0457.C1.2_H04_016_Batch6.C1.2.04-04-2009.ab1	H4	16	33	256	405	138	83	126	140	C
Ladder_A03_001_Batch6.C1.2.04-04-2009.ab1	A3	1	N/A	N/A	N/A	11	10	8	12	
03-0458.C1.2_A05_001_Batch6.C1.2.04-04-2009.ab1	A5	1	36	252	369	188	110	174	194	C
03-0459.C1.2_B05_003_Batch6.C1.2.04-04-2009.ab1	B5	3	29	488	423	199	111	177	200	C
03-0460.C1.2_C05_005_Batch6.C1.2.04-04-2009.ab1	C5	5	32	191	235	416	211	465	430	reinj
03-0461.C1.2_D05_007_Batch6.C1.2.04-04-2009.ab1	D5	7	34	257	386	130	69	116	128	C
03-0462.C1.2_F05_011_Batch6.C1.2.04-04-2009.ab1	F5	11	36	501	485	562	327	527	621	
03-0463.C1.2_G05_013_Batch6.C1.2.04-04-2009.ab1	G5	13	33	481	448	93	53	89	96	
03-0464.C1.2_H05_015_Batch6.C1.2.04-04-2009.ab1	H5	15	29	275	431	165	91	140	170	reinj
03-0465.C1.2_A06_002_Batch6.C1.2.04-04-2009.ab1	A6	2	33	474	460	564	320	459	572	
03-0466.C1.2_B06_004_Batch6.C1.2.04-04-2009.ab1	B6	4	34	497	484	284	156	233	279	

Page 2 of 5

Figure 7 – Example QC Report with Analyst's Comments. Based on the color code (Table 2) the analyst reviewing the sequences may choose to reinject a partial sequence to attempt to obtain a

more complete sequence or if a sequence has a C-stretch and is acceptable quality the analyst can choose to pass that sequence noting the C-stretch with a "C✓".

An example of a GG ("green green") sequence from the QC Report (Figure 7) can be seen below in Figure 8a. The primer was able to sequence through the region where length heteroplasmy is commonly observed and therefore that trace has a longer CRL. The GY ("green yellow") example (Figure 8b) has a shorter CRL due to the length heteroplasmy observed in the sequence.

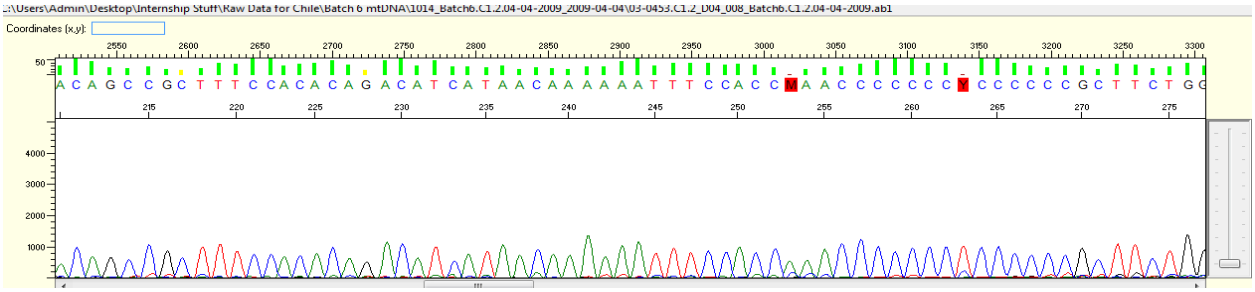


Figure 8a – Example of GG Trace. This sequence is acceptable quality and the DNA polymerase was capable of extending through the C-stretch region.

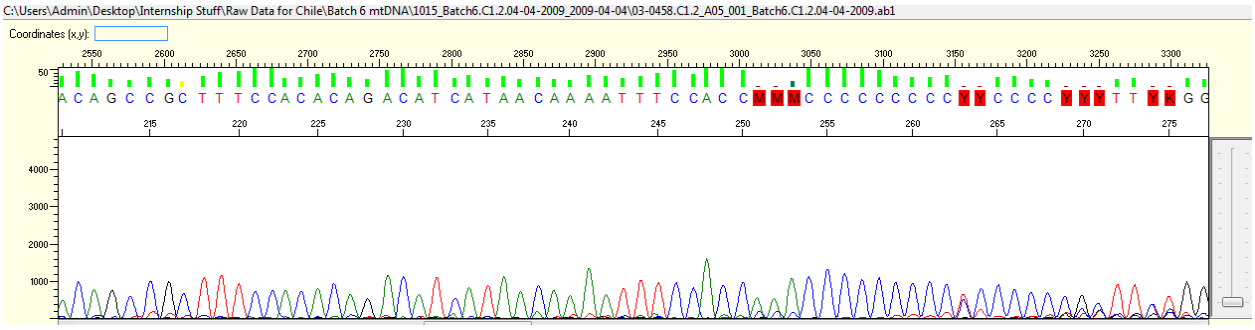


Figure 8b – Example of GY Trace. This sequence is acceptable quality until it becomes out of register due to the length heteroplasmy. The red "Ms" and "Ys" indicate poor quality, ambiguous bases.

Specific Aims

The Research & Development team proposed that the trace score and CRL parameters could be further refined based upon homopolymeric stretch and length heteroplasmy data as well as individual primers. The specific aims using this mtDNA sequence data for this internship project include:

1) Optimization of primer specific quality metrics (Phase 1)

2) Evaluation of signal intensities for possible pattern recognition and predictive modeling (Phase 2)

3) Use of quality values per peak to ascertain confidence in the base call (Phase 3).

CHAPTER III

PHASE 1 – OPTIMIZATION OF PRIMER SPECIFIC FILTER METRICS

Using mitochondrial DNA (mtDNA) sequence data, optimization of primer specific quality metrics, specifically the trace score and contiguous read length (CRL), was performed. Optimization is the process of making the filter metrics as fully effective as possible. Calibration, validation, and a concordance check are performed for optimization of primer specific trace score and CRL settings. Quality metrics translate sequence information into numerical values which allows a software program to filter through the data for you. The trace score and CRL settings can be manually adjusted in the preferences window of Sequence Scanner Software. (Figure 9) A model of the current settings can also be seen in Figure 9.

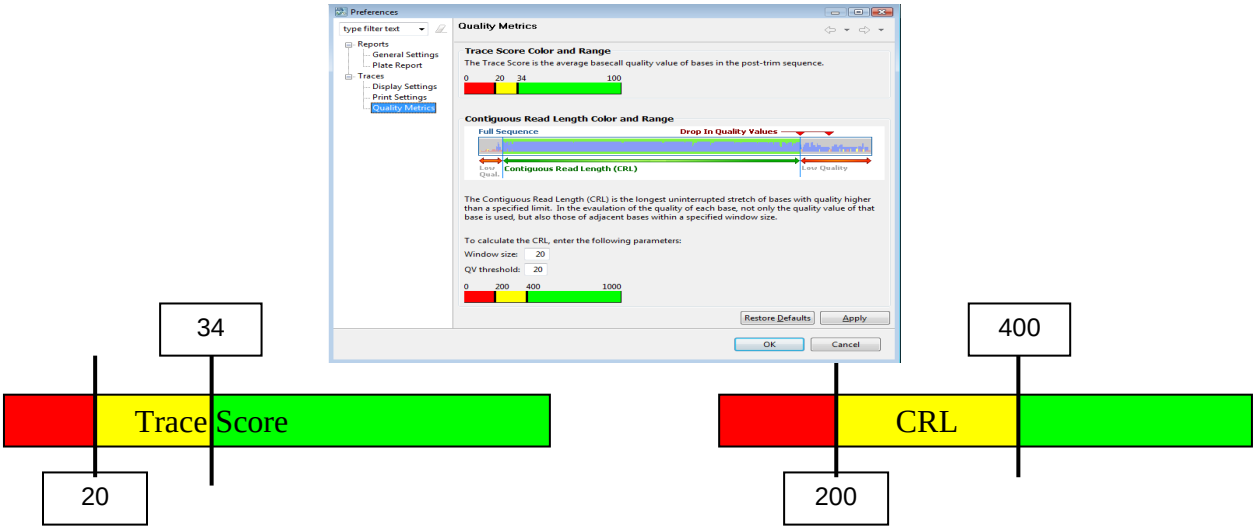


Figure 9 - Preferences Window Displaying Current Settings

Materials & Methods

Using the 12 batches of mtDNA sequence data generated by the University of North Texas Center for Human Identification Research & Development Laboratory (UNTCHI RDL), two datasets were formed. The first dataset contained sequence data from Batches 1-4 (2,817 total sequences) and was designated for primer specific calibration. The second dataset contained sequence data from Batches 5-12 (5,617 total sequences). This second dataset was designated for validation and concordance of the optimized settings. The QC (quality control) Report from Sequence Scanner Software was exported and any comments the analyst had made were noted. Microsoft Access databases were formed which contained all of this numerical information as well as analyst comments. These databases allow for quick querying of potential settings.

Using the revised Cambridge Reference Sequence (rCRS), a standard used for comparison, the number of bases from where a primer binds to where it could potentially read was counted (Appendix A, Table 4)

Table 4 – Primer Specific Read Lengths. Maximum read length is the number of bases from where a primer binds to the end of the sequence. The C-stretch read length is the number of bases from where a primer binds to where that sequence could meet either a homopolymeric stretch, length heteroplasmy, or both.

Primer	**Maximum Read Length (# of bases)**	**Read Length until it Encounters C-stretch (# of bases)**
R1	1183	253 (HV1), 941 (HV2)
B1	460	198 (HV1)
C1	497	255 (HV2)
R2	1183	921 (HV1), 230 (HV2)
C2	368	126 (HV2)

Note: Only primers with potential to sequence an HV1 homopolymeric stretch or an HV2 length heteroplasmy or both are shown here.

Primer R1 (a forward primer), for example, has a maximum potential read length of 1,183 bases; however, if a sample contains an HV1 homopolymeric stretch the read is shortened to 253 bases. If a sample does not contain a homopolymeric stretch in HV1 but does contain a length heteroplasmy is HV2, primer R1 could sequence through HV1 and stop at 941 bases in HV2 due to the length heteroplasmy. Primer R2 (a reverse primer) has the maximum potential to read 1, 183 bases. If a sample contains a length heteroplasmy in HV2, the read length for primer R2 stops at 230 bases. If primer R2 is able to sequence through HV2 and into HV1 but stops at 921 bases, an educated assumption can be made that the sample contains a homopolymeric stretch in HV1. Primer B1, on the other hand, does not have the capability of sequencing HV2. Its maximum potential read length is 460 bases. However, if a sample has a high trace score value and a CRL of approximately 198, it can be assumed that sample contains a homopolymeric stretch in HV1. (Table 4, Figure 5) Using these read length values, potential new settings were determined. Calibration databases were queried and any sample that was not previously ruled as GG, passing, was manually reviewed to ensure it was good quality sequence. Once settings were optimized, validation databases were queried for concordance and any newly ruled GG sample was manually reviewed.

Results

After optimization, there was an overall reduction in the number of samples requiring manual intervention. For primer R1, under the current utilized settings 39.5% of the samples required manual review; after optimization and utilizing the primer specific settings 31.5%

would require manual review. The current settings (trace score of 20, 34 and a CRL of 200, 400) were optimized for primer R1 which explains why there is not a large reduction in the number of samples requiring review after optimization. These current trace score (20, 34) and CRL (200, 400) settings were being applied to the remaining primers. For primer B1, under the current settings 51.6% of the samples required manual review; utilizing primer specific settings 36.0% of the samples would require manual review. For primer C1, under the current settings 79.0% of the samples required manual intervention; after optimization 32.6% of the samples would require human intervention. For primer R2, under the current settings 61.7% of the samples required manual review; after optimization 41.8% would require review. For primer A4, under the current settings 16.7% of the samples required manual review; utilizing primer specific settings 5.5% of the samples would require manual review. For primer B4, under the current settings 91.5% of the samples required human intervention; after optimization only 26.1% of the samples would require human intervention. For primer C2, under the current settings 93.6% of the samples would require manual review; after optimization 46.7% of the samples would require review. For primer D2, under the current settings 24.0% of the samples required manual intervention; utilizing primer specific settings 7.2% of the samples would require manual intervention. (Figure 10)

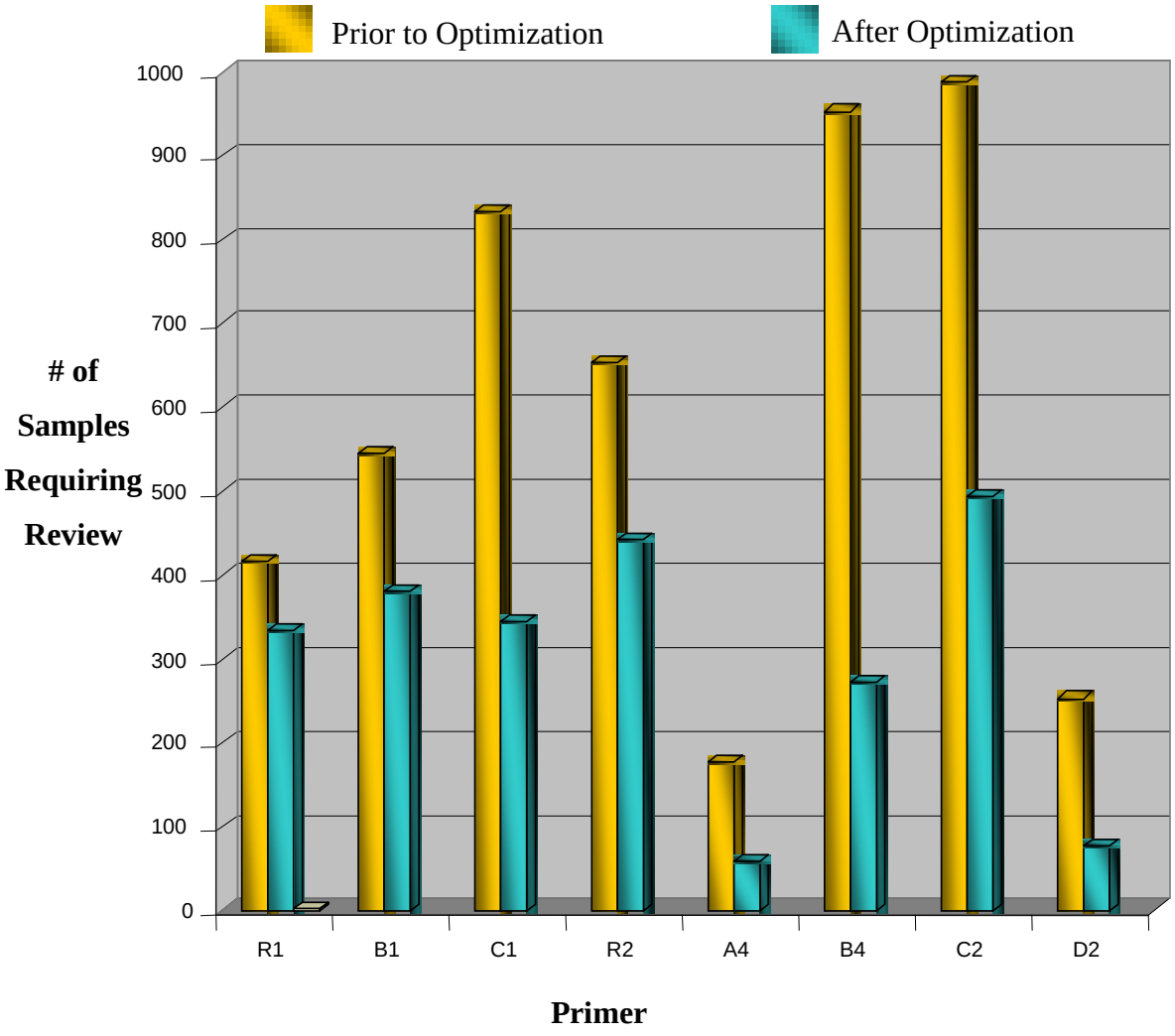


Figure 10 – Number of Samples Requiring Review per Primer. The yellow bars represent the number of samples requiring manual review under the current settings. The blue bars represent the number of samples requiring review if primer specific settings had been utilized.

Another view of the improvements can be seen in Figures 11a and 11b. The first page of a QC Report for a single batch displaying 16 samples shows that under the current settings 13 of

those samples required manual review. However under the proposed settings, only five of those 16 samples would have required manual review.

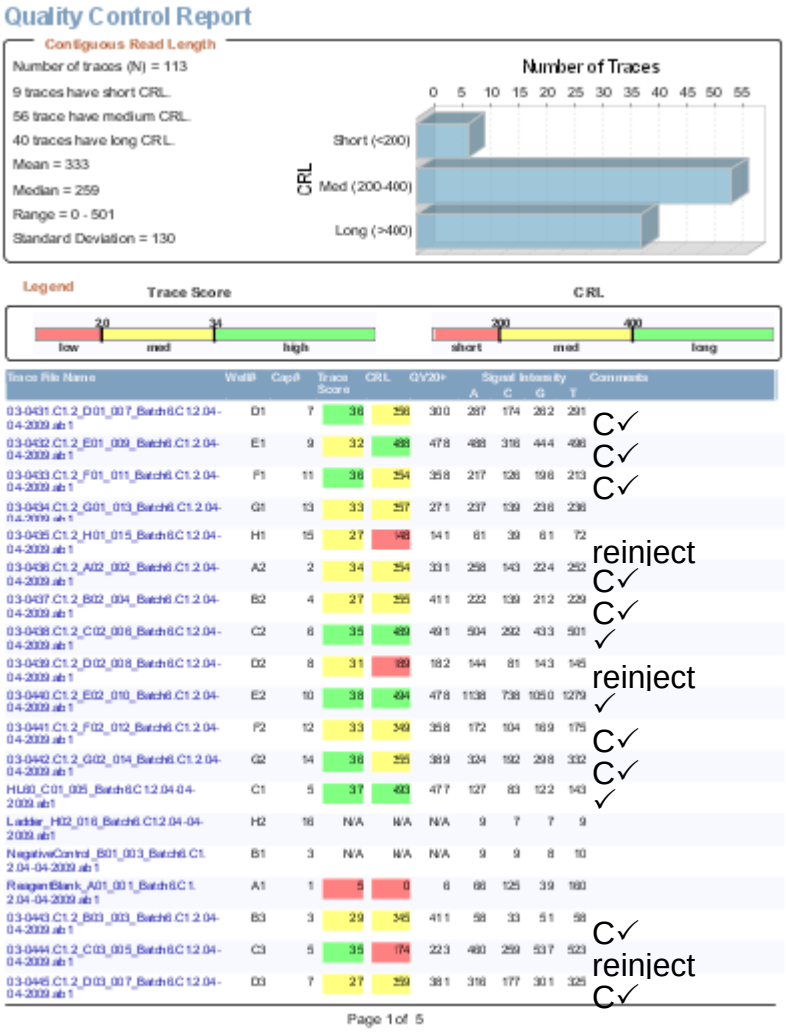


Figure 11a – QC Report with Current Settings. Thirteen of the sixteen samples on this page required review.

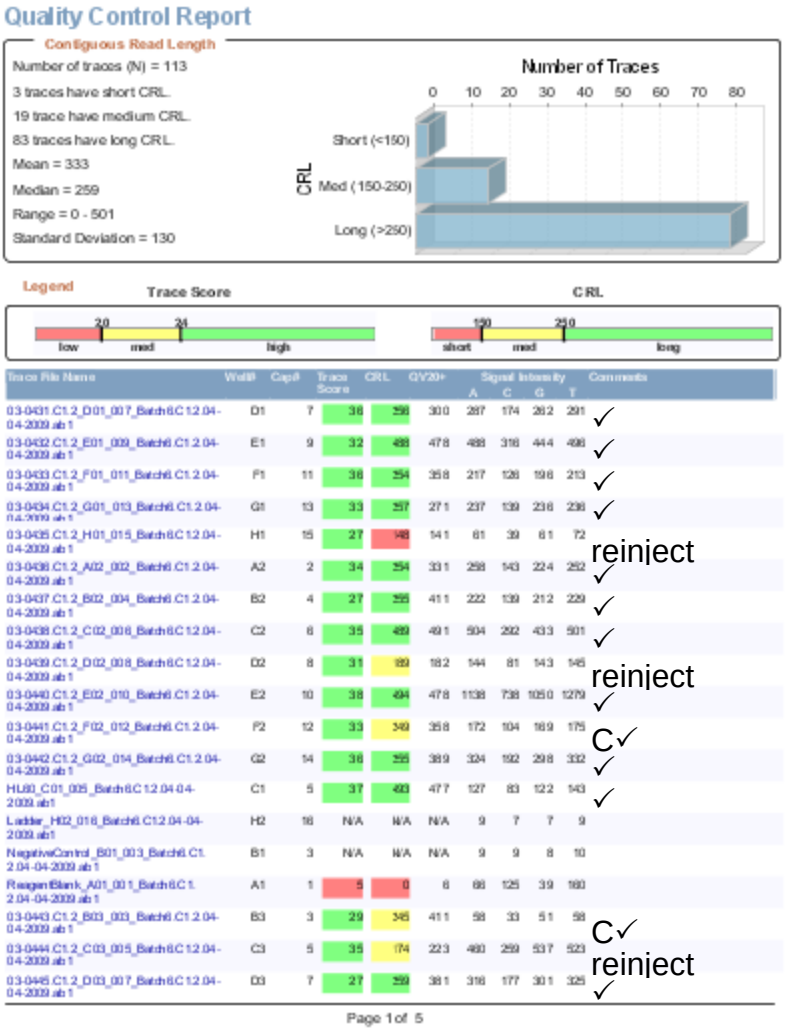


Figure 11b – QC Report with Proposed Settings. Five of the sixteen samples now require review.

Conclusions

The UNTCHI RDL has shown that filter metrics do work and can be applied. By optimizing filter metrics to specific sequencing primers, there was an overall decrease in the number of samples requiring manual intervention. Prior to optimization 57.1% of all sequences required manual review and after optimization only 28.4% of all sequences would have required manual review. The proposed primer specific settings have been submitted to the UNTCHI RDL

for implementation. Future software developments could automate these settings and this would be beneficial for implementation. The UNTCHI RDL team has been asked to consider taking new actions based on the color codes (Table 5). Upon evaluation of sequences for Phase 1, it was noticed that the majority of GR and YR samples were being re-injected as the samples were partial sequences and re-injection was attempted to try to obtain a more complete sequence.

Table 5 – Proposed Actions

Color Code	Action Taken
GG	Pass
GY	Review
GR	Reinject
YG	Review
YY	Review
YR	Reinject
RG	Review
RY	Review
RR	Fail

CHAPTER IV

PHASE 2 – SIGNAL STRENGTH EVALUATION

Evaluation of signal intensities for possible pattern recognition and predictive modeling of sample quality was performed. Two variables, the signal intensity (SI) and the signal to noise ratio (SNR), provided by Sequence Scanner Software were evaluated for possible pattern recognition and predictive modeling of sample quality. Signal intensity is defined as the average raw signal intensity.

Materials and Methods

Sequence Scanner Software displays the average raw signal intensity for the four bases in the QC (quality control) Report by default. The average raw signal to noise ratio for the four bases can be selected to be displayed in place of the signal intensities. Proprietary algorithms are being applied to the raw data for both parameters. Evaluation of using the signal intensity or signal to noise ratio parameters as a means of assessing sample quality was approached in several ways.

The average signal intensities for 90 samples were analyzed for observance of a pattern. At first glance, it appeared that the A/T signal intensities were higher than the C/G signal intensities, but this potential trend was not always reliable. Initial evaluation of the signal to noise ratio seemed to show that all poor quality samples, negative controls, and reagent blanks had extremely low value (<10). A query with Signal to Noise Ratio of less than 11 for all four

nucleotides was run on a database of randomly chosen samples (N = 426) and 34 samples met these criteria. Of those 34 samples, 2 were good quality, passing samples and 7 reagent blanks and negative controls failed to meet these criteria; suggesting that use of the signal to noise ratio of less than or equal to 10 is not a good measure for filtering out poor quality data.

Several calculations using these parameters were performed. A calculation of Signal Intensity over Signal to Noise Ratio was performed and no patterns were observed. For a small dataset of thirteen samples, the (A+T)/(C+G) calculation also revealed no trend.

Applying the (A+T)/(C+G) calculation to eight negative controls and reagent blanks did not provide validation of a trend. Means and standard deviations were calculated for 64 negative controls and reagent blank samples and the average was used as a cutoff for poor quality samples for use in querying samples from a Phase 1 optimization database. The average signal to noise ratio for (A) was 74.6 ± 123.9, (C) was 137.3 ± 237.8, (G) was 46.5 ± 76.4, and (T) was 80.6 ± 140.9. Thirty six samples met the criteria (A<75, C<138, G<47, and T<81) and 18 of the 36 were good quality, passing samples. Analysis of negative controls and reagent blanks to assess background noise supported the idea that signal intensity values provided by Sequence Scanner Software are based on an algorithm that is applied to the raw sequence data and therefore include non-sequence nucleotide peaks visible in electropherograms at the beginning of raw sequence data.

Entire sequences for eight samples were used to calculate the number of As, Ts, Cs, and Gs to evaluate the possibility of a trend with the number of peaks and its average raw signal intensity. This analysis also did not reveal any obvious distributional patterns.

Other software programs for signal intensity and/or signal to noise ratio were also employed to evaluate signal intensities and signal to noise ratios for pattern recognition and

predictive modeling. FinchTV v1.4.0 is a freeware program with multi-pane, scalable views from Geospiza, Inc. (Seattle, WA) which allows the user to view raw DNA sequences and their reverse complement traces, perform BLAST searches, and edit raw sequences. Sequencing Analysis Software v5.2 from Applied Biosystems allows its user to analyze, display, edit, save, and print sample files generated with Applied Biosystems genetic analyzer platforms. This software program also has the ability to perform base calling and provides per base estimate of the base call accuracy.

Sequences were imported into FinchTV v1.4.0 and trimmed so that only the best quality sequence was available for analysis. These trimmed sequences were imported back into Sequence Scanner Software and the signal intensity and signal to noise values remained unchanged. Trimmed sequences were also imported into the base calling program, Sequencing Analysis v5.2, to see if the sequence that was trimmed would be recognized. Again, the signal intensity and signal to noise values were unchanged.

Mutation Surveyor™ v3.25 is a software program by SoftGenetics, LLC. (State College, PA) allowing assessment of DNA variation such as single nucleotide polymorphisms (SNPs) using Sanger sequencing traces. The software can align sequence from the same source and conduct single or bi-directional analysis. Mutation Surveyor™ is capable of performing tasks, such as base calling, genome assembly, insertion/deletion detection, mutation detection, mutation quantification, and methylation analysis. Multiple traces were imported into Mutation Surveyor™ and analyzed under default settings. An Output Trace Data report (.txt file) was exported for each sample. In Microsoft® Excel, the average base height, average green (A) height, average blue (C) height, average black (G) height, average red (T) height, and average Phred score were calculated for each sequence. The average base heights were plotted and no

trend was observed. (Figure 12) The aforementioned trimmed sequences were imported in Mutation Surveyor™ and analyzed. Output Trace Data files were exported, average values were calculated, and average base heights were plotted with no apparent pattern observed.

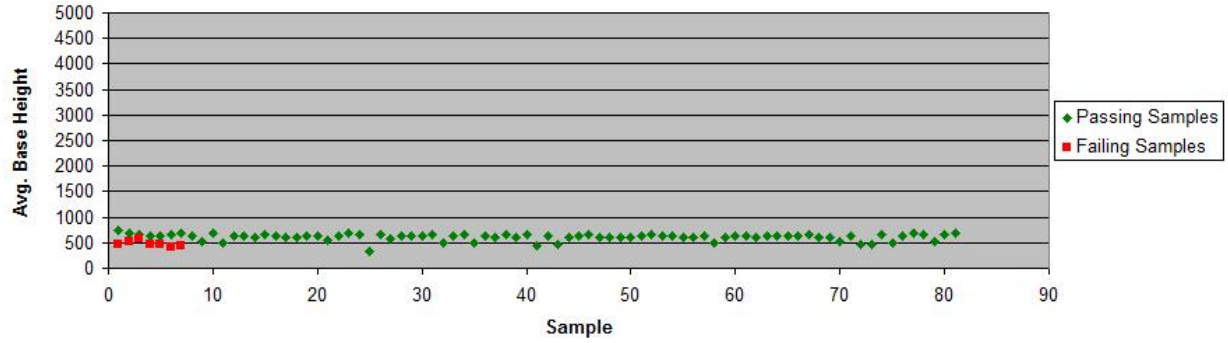


Figure 12 – Average Base Height for 88 Passing and Failing Samples

Results

No patterns were recognized for predictive modeling of sample quality. Mutation Surveyor™ v3.25 software provides a Signal Factor Intensity and a Signal Factor Deviation value per base for a sequence in the Output Trace Data report. Ambiguous base calls have a lower Signal Factor Intensity value (usually below 0.70) and good quality bases typically have higher values (most are greater than 0.85). The technical support staff at SoftGenetics clarified these definitions (where *i* is the nucleotide base called or base of interest).

Signal Factor Intensity = Height(*i*)/(HeightA + HeightT + HeightC + HeightG)

Signal Factor Deviation = Dev(*i*)/(DevA + DevT + DevC + DevG)

Dev(*i*) = 2.0*Height(*i*) - Height(*i*L) – Height(*i*R)

Height(*i*L) = peak height 4 bases to the left

Height(*i*R) = peak height 4 bases to the right

Conclusions

The various ratios attempted did not reveal any evidence of a trend for signal intensity or signal to noise ratio for predictive modeling. The most potential lies with the Signal Factor Intensity value provided by Mutation Surveyor™ and may prove beneficial when building a sequence expert system. Collection of sequence data at a later point to omit the nucleotide collection observed at the beginning on sequence data should be studied further in hopes of providing a truer SI or SNR value for a sample.

CHAPTER V

PHASE 3 – QUALITY VALUE PER PEAK FOR BASE CALLING CONFIDENCE

The third objective was to use quality values (QV) per peak to ascertain confidence in the base call. Currently, the forensic DNA typing community requires either a forward and reverse sequencing primer to confirm a base or two sequencing primers in the same direction from two separate amplifications to confirm the base. Sequencher™ (Gene Codes Corporation, Ann Arbor, MI) assigns a light blue color to high quality bases and darker blue colors to lesser quality bases. The good quality sequence is light blue and is displayed from Sequence Scanner Software and the poor quality sequence is displayed below the Sequencher™ sequence. (Figure 14) The quality values per peak can be obtained from the software. In the example is Figure 13, the high quality sequence has QVs per peak of 48, 48, 48, 55, 55, 55, 55 etc. On the other hand, the low quality sequence has QVs per peak of 1, 1, 1, 1, 1, 17, 1, etc. The goal is to change sequence information into numerical values to aid in software development. Sequence Scanner Software considers a QV per base of greater than or equal to 20 as good quality peaks. The software uses the quality per peak, evaluates the overlap of fluorescent signals, and measures a Gaussian fit to determine a peak's quality value.

$$QV = -10\log_{10} Pe$$

Pe = probability of error

If a peak has a QV = 1, it has a Pe = 79%. That means it has a 79% chance of being the wrong base call. If a peak has a QV = 17, it has a Pe = 2%. If a peak has a QV = 48, it has a Pe = 0.0016%. If a peak has a QV = 55, it has a Pe = 0.0003%.

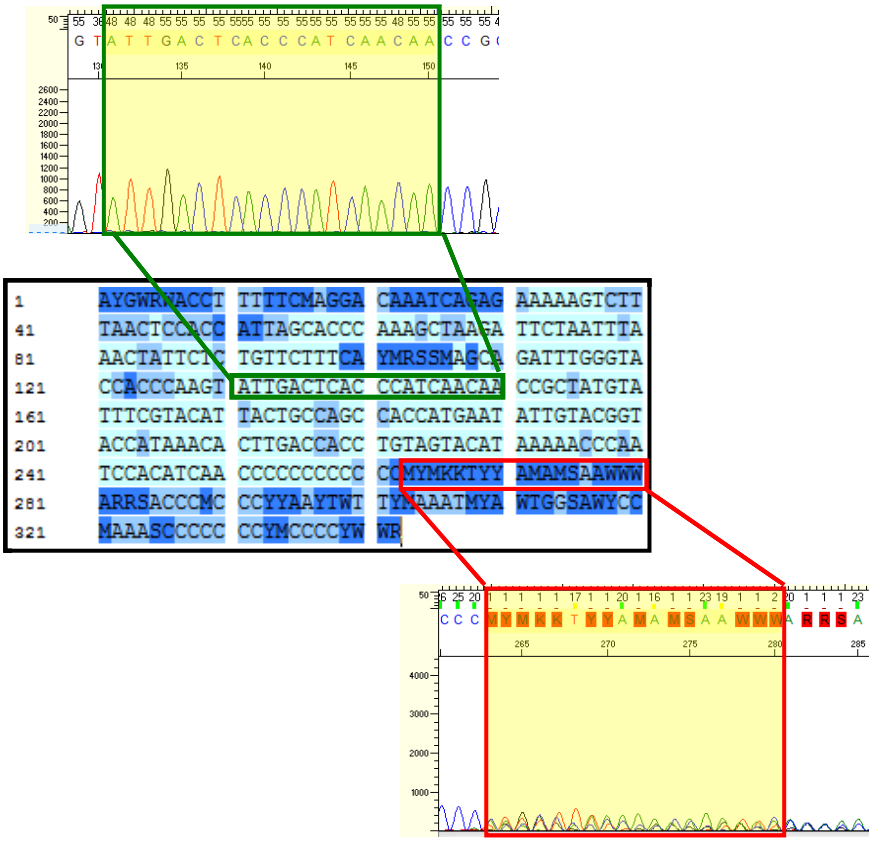


Figure 13 – Peak Quality. The blue sequence box is generated by Sequencher™. Light blue bases are high quality peaks and dark blue bases are poor quality. The two electropherograms with quality values per peak are generated by Sequence Scanner Software. The top electropherogram displays good quality sequence with high quality values per peak. On the other hand, the bottom electropherogram displays poor quality sequence with low quality values per peak and even some mixed/ambiguous bases.

Materials and Methods

Quality (.qual) and sequence (.seq) files were exported from Sequence Scanner Software. Sequencher™ project files (.spf) previously assembled by the RDL analysts were used to review samples for QV per base. Three specific scenarios were evaluated: 1) unconfirmed, good quality bases; 2) confirmed, sequence heteroplasmy; and 3) unacceptable bases with a QV ≥ 20. The third scenario, an example of QV ≥ 20, was not observed.

Results

According to Sequence Scanner, "high quality pure bases are generally assigned a QV between 20 and 50." Hence, high quality values indicate a low Pe (probability of error).

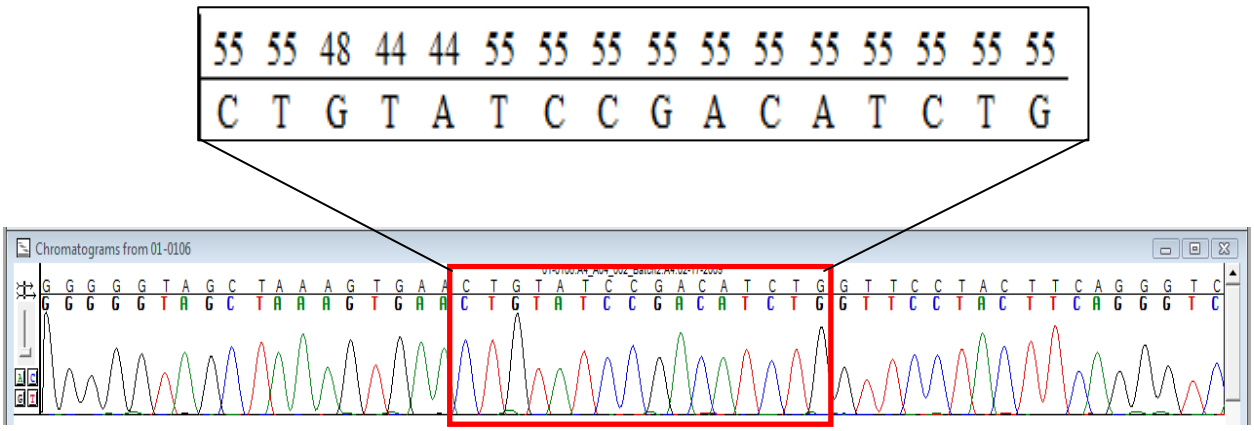


Figure 14 – Unconfirmed, Good Quality Bases. High quality values per peak indicate low probabilities of error, thus providing confidence in the base call.

Sequence Scanner Software's Quality Values Chart states, "high quality mixed bases are generally assigned a QV between 10 and 50." Sequence Scanner Software recommends reviewing pure bases with a QV below 20 and mixed bases with a QV below 10. In Figure 15, the neighboring peaks of the sequence heteroplasmy have high quality values and would not

require review. In this example, only primer B4's point heteroplasmy position would require review since the mixed base QV is below 10. (Figure 15)

Primer	C	T	A	T	G	T	A	T	**Y**	T	C	G	T	A	C	A	T
B4	40	55	55	55	55	55	31	55	**5**	55	36	55	40	36	48	44	36
R1.2	36	55	55	55	55	55	55	55	**17**	55	55	55	55	55	55	55	55
R1	55	36	55	55	55	55	55	55	**16**	55	55	55	55	55	55	55	55

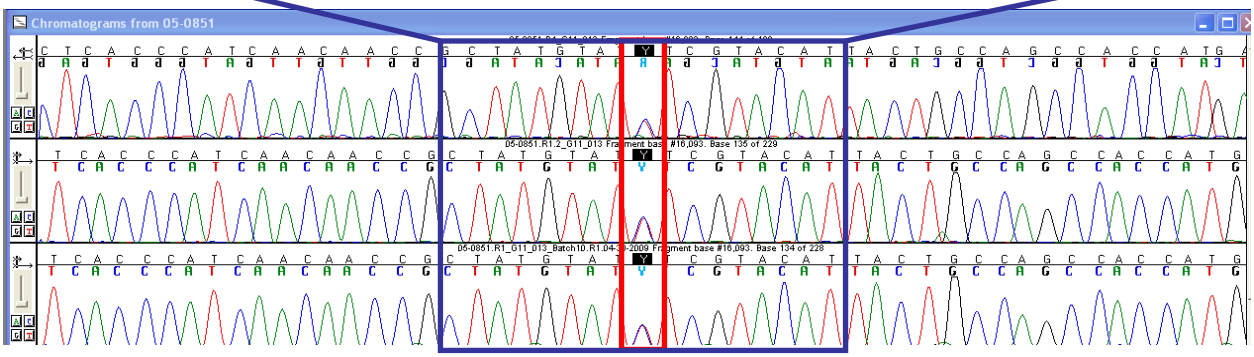


Figure 15 – Quality Value of Sequence Heteroplasmy. The neighboring bases have high quality values indicating low probability of error. The point heteroplasmy has lowered quality value; however, Sequence Scanner Software suggests reviewing mixed bases if the quality value is below 10.

Conclusions

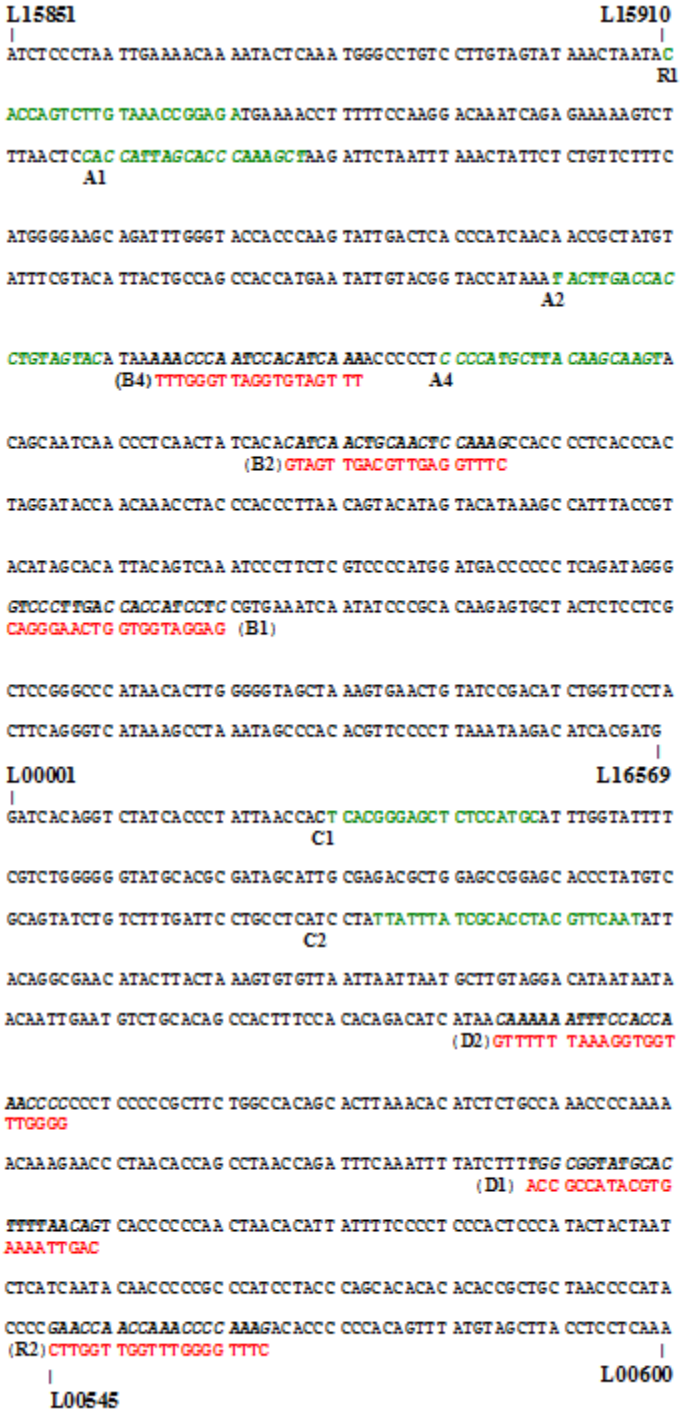
Ultimately, a software system with rule firings that could flag bases with low quality values could prove to be extremely beneficial to the forensic community. A sequence expert system rule could be modeled after Sequence Scanner Software's recommendations for reviewing bases. The ability of using QV per peak for confident base calling is feasible and would be of great assistance in building a sequence expert system. A future study of a possible

window filter is suggested. It is suggested that the window filter look at 10 bases to the left of the base of interest and 10 bases to the right and if no more than five bases have a QV below 20 then that sample should be considered good quality sequence.

APPENDIX A

mtDNA CONTROL REGION & PRIMERS

Figure 16 - rCRS Control Region & Primers



APPENDIX B

PRIMER SPECIFIC SETTINGS & DATA FROM CALIBRATION BATCHES

Table 6 - Primer R1 Data

Color Code	Total # Samples Before Optimization	Total # Samples After Optimization
GG	175	197
GY	94	91
GR	13	16
YG	14	11
YY	17	1
YR	22	19
RG	0	0
RY	0	0
RR	17	17

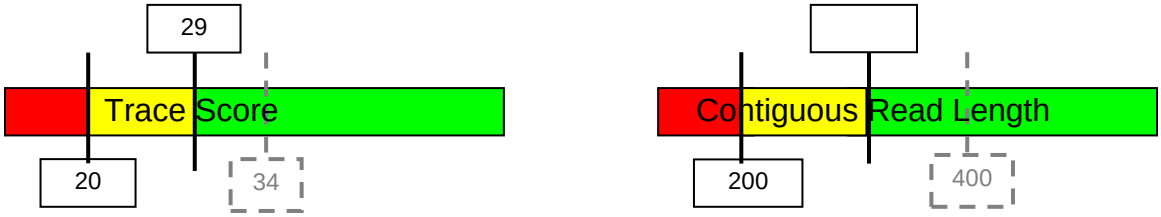


Figure 17 - Primer R1 Settings

Table 7 - Primer B1 Data

Color Code	Total # Samples Before Optimization	Total # Samples After Optimization
GG	137	182
GY	7	86
GR	66	9
YG	19	15
YY	34	10
YR	57	18
RG	0	0
RY	0	1
RR	32	31

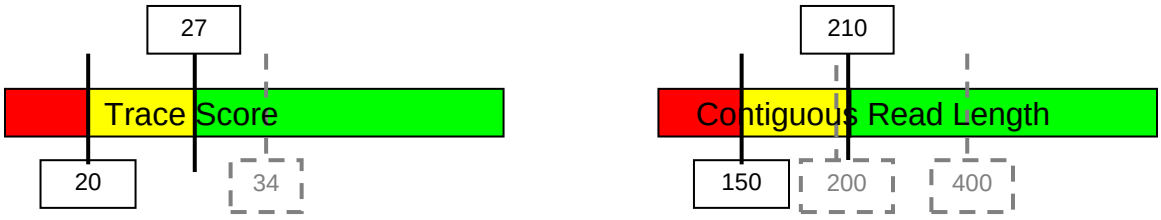


Figure 18 - Primer B1 Settings

Table 8 - Primer C1 Data

Color Code	Total # Samples Before Optimization	Total # Samples After Optimization
GG	98	194
GY	68	93
GR	2	6
YG	21	2
YY	91	12
YR	44	17
RG	0	0
RY	0	0
RR	28	28

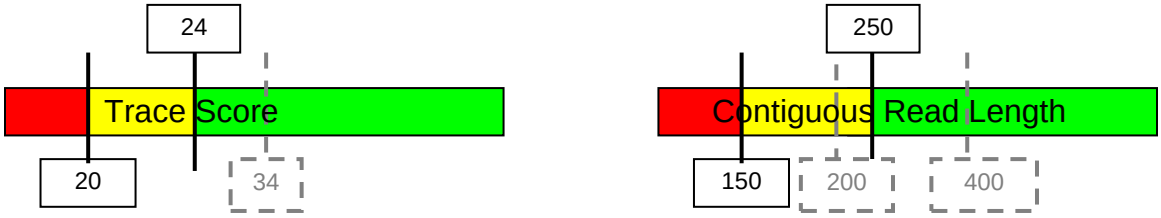


Figure 19 - Primer C1 Settings

Table 9 - Primer R2 Data

Color Code	Total # Samples Before Optimization	Total # Samples After Optimization
GG	101	147
GY	87	158
GR	9	4
YG	14	2
YY	80	1
YR	25	4
RG	0	0
RY	0	0
RR	36	36

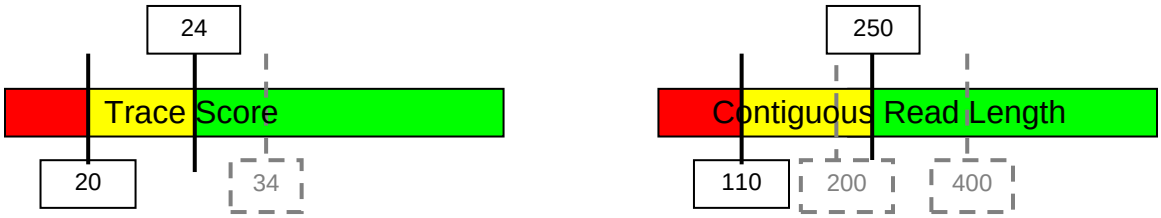


Figure 20 - Primer R2 Settings

Table 10 - Primer A4 Data

Color Code	Total # Samples Before Optimization	Total # Samples After Optimization
GG	244	319
GY	26	6
GR	2	9
YG	24	3
YY	34	0
YR	15	8
RG	0	0
RY	0	0
RR	8	8

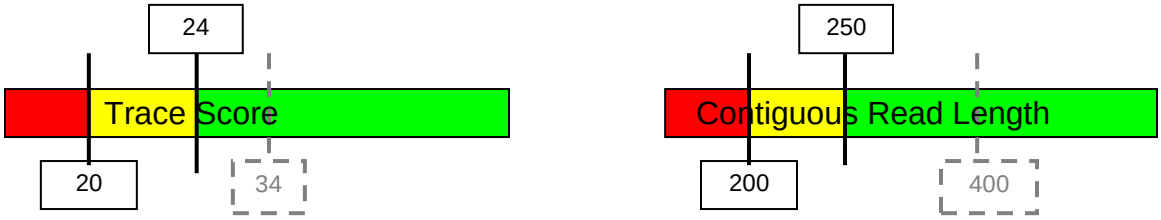


Figure 21 - Primer A4 Settings

Table 11 - Primer B4 Data

Color Code	Total # Samples Before Optimization	Total # Samples After Optimization
GG	0	222
GY	176	36
GR	26	44
YG	2	0
YY	44	2
YR	70	14
RG	0	0
RY	0	1
RR	34	33

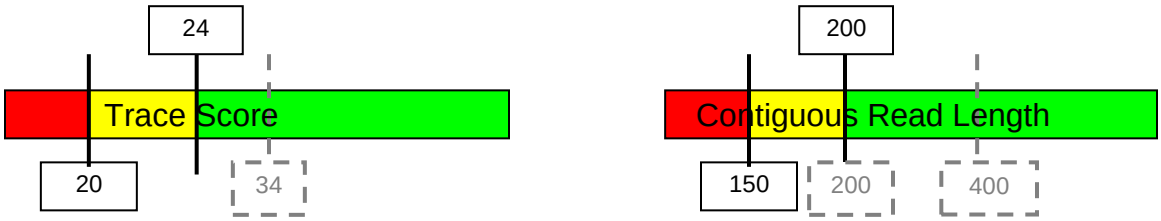


Figure 22 - Primer B4 Settings

Table 12 - Primer C2 Data

Color Code	Total # Samples Before Optimization	Total # Samples After Optimization
GG	1	153
GY	96	135
GR	58	19
YG	0	4
YY	42	17
YR	142	11
RG	0	1
RY	1	0
RR	12	12

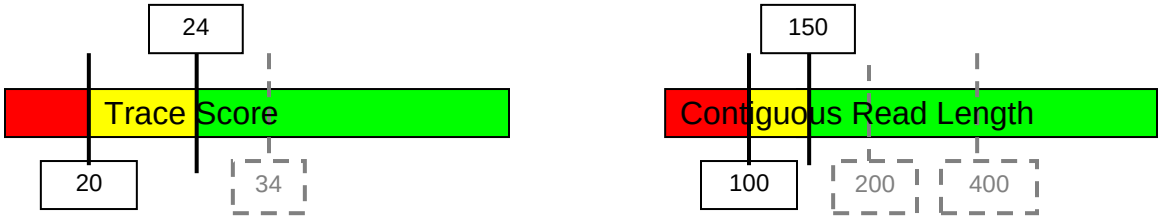


Figure 23 - Primer C2 Settings

Table 13 - Primer D2 Data

Color Code	Total # Samples Before Optimization	Total # Samples After Optimization
GG	176	220
GY	7	1
GR	6	7
YG	23	5
YY	14	3
YR	18	8
RG	0	0
RY	0	0
RR	108	108

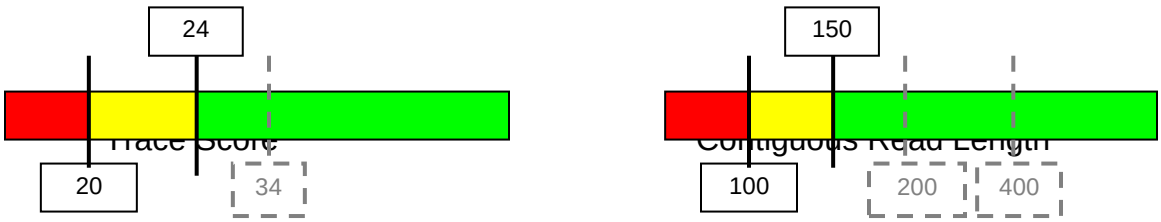


Figure 24 - Primer D2 Setting

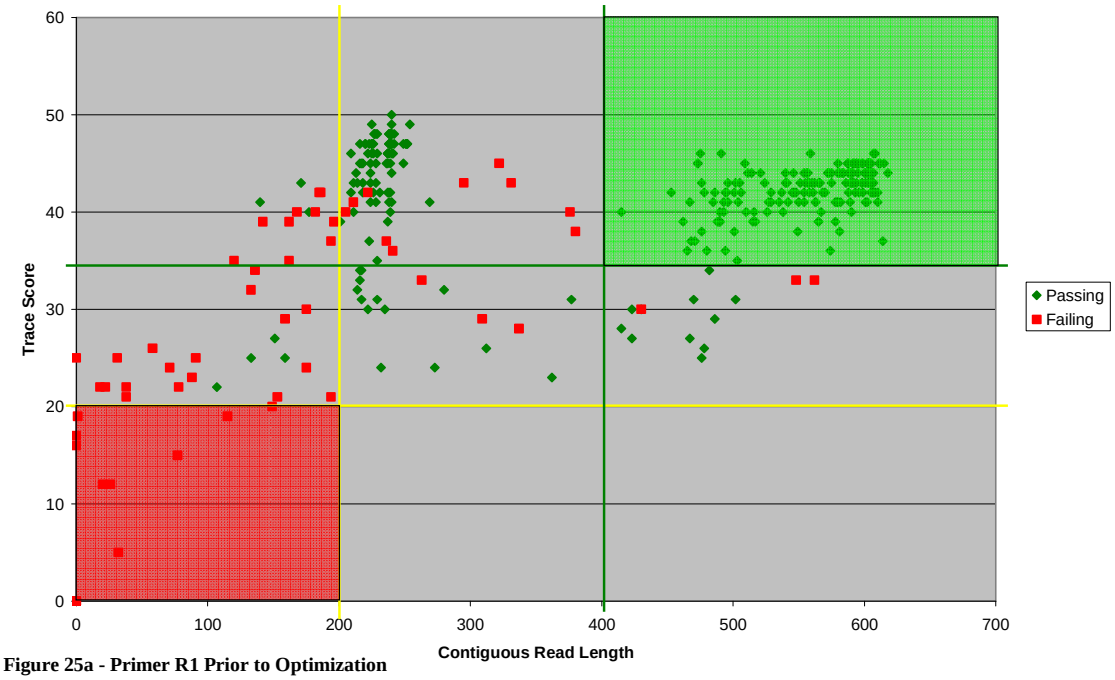


Figure 25a - Primer R1 Prior to Optimization

Figure 25b - Primer R1 After Optimization

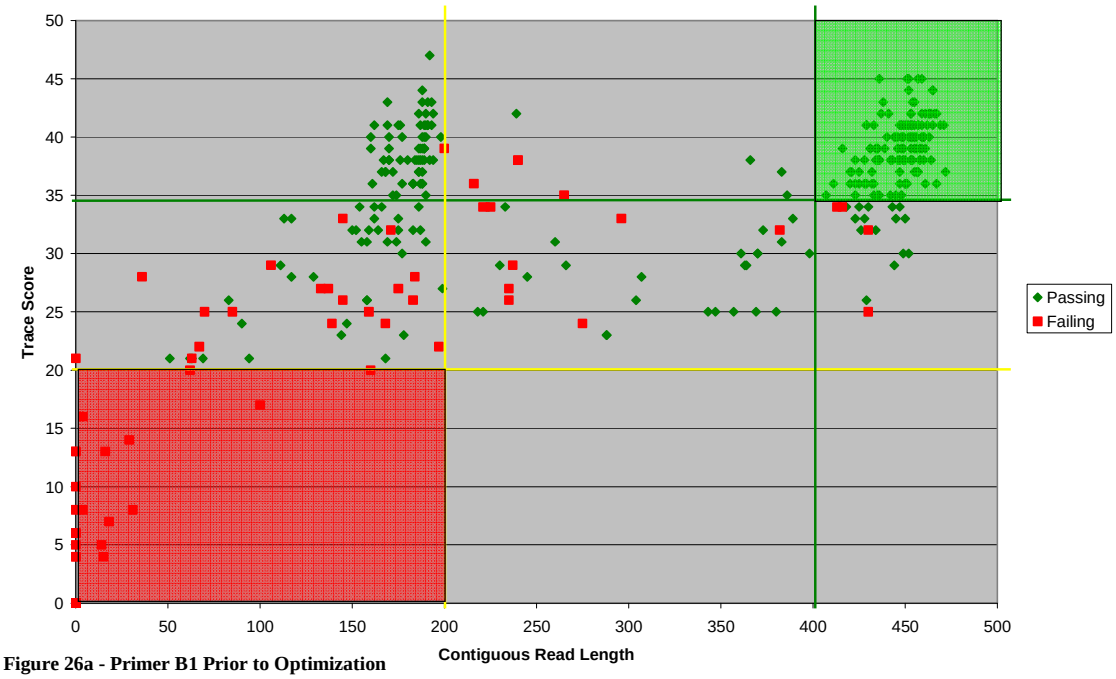


Figure 26a - Primer B1 Prior to Optimization

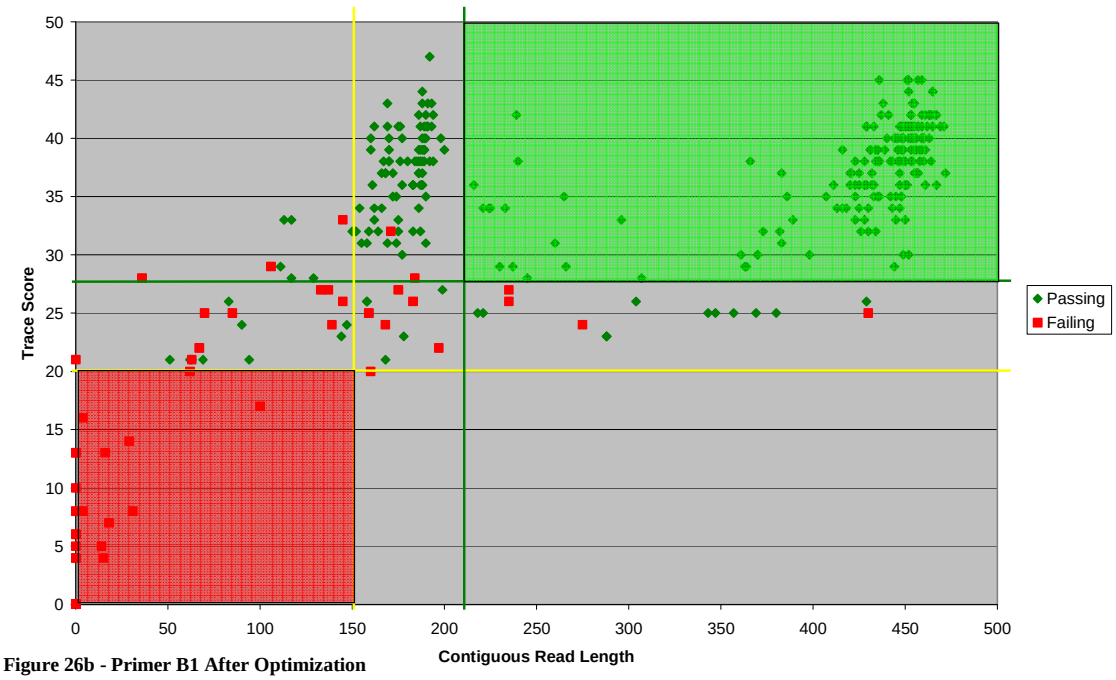


Figure 26b - Primer B1 After Optimization

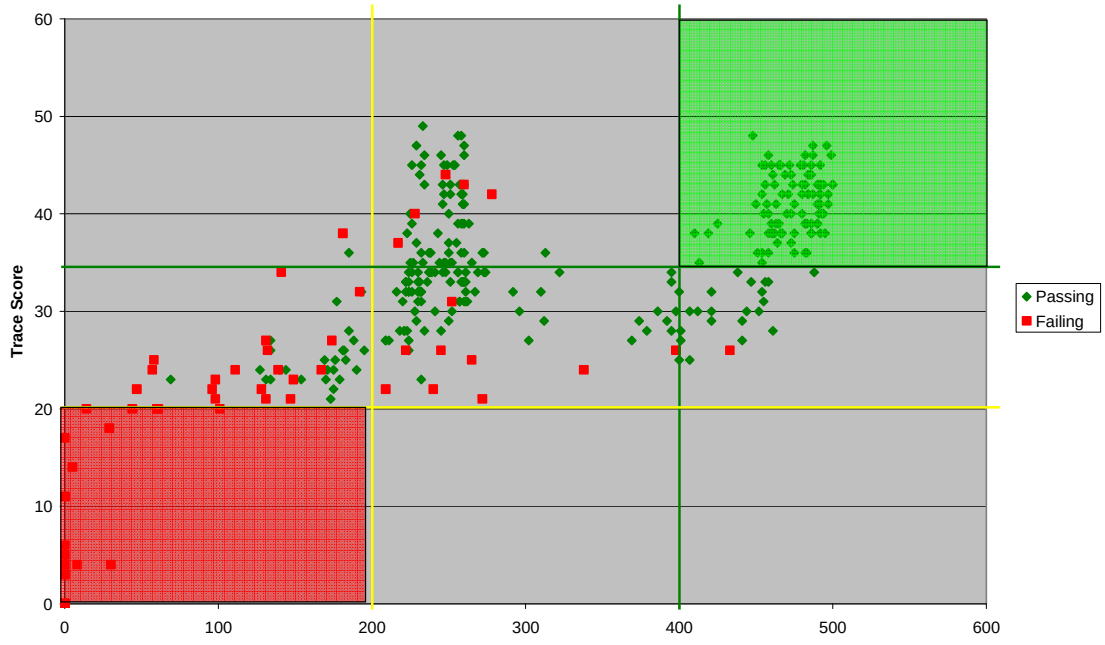


Figure 27a - Primer C1 Prior to Optimization

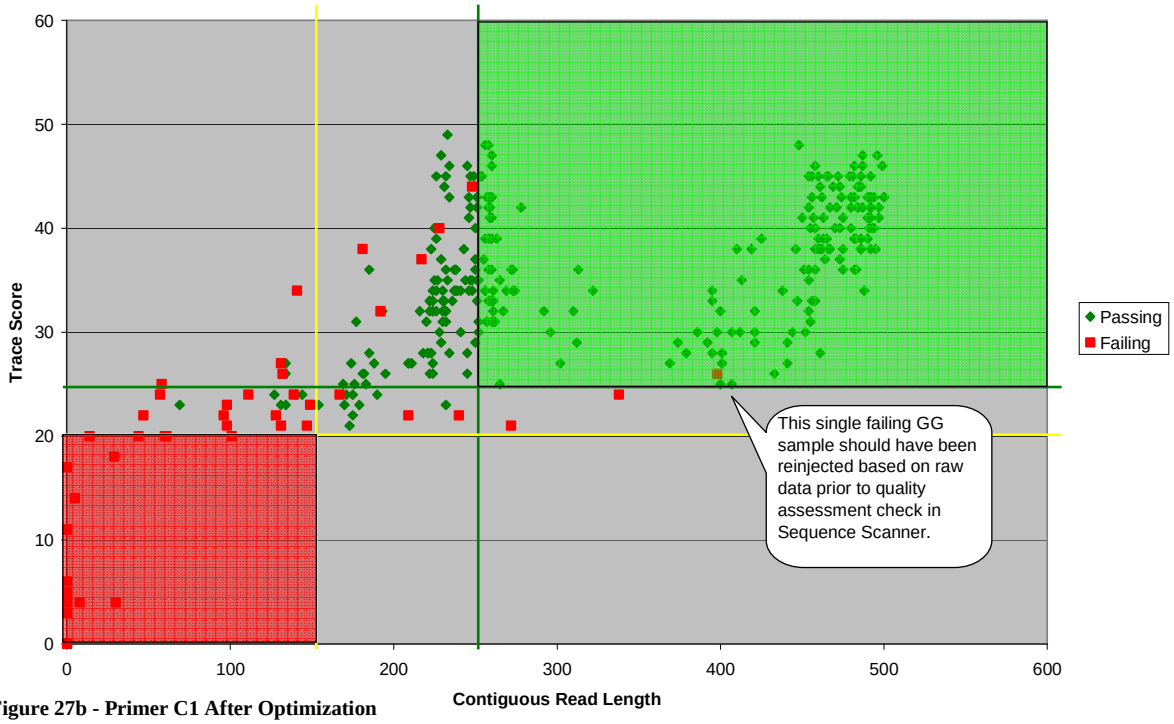


Figure 27b - Primer C1 After Optimization

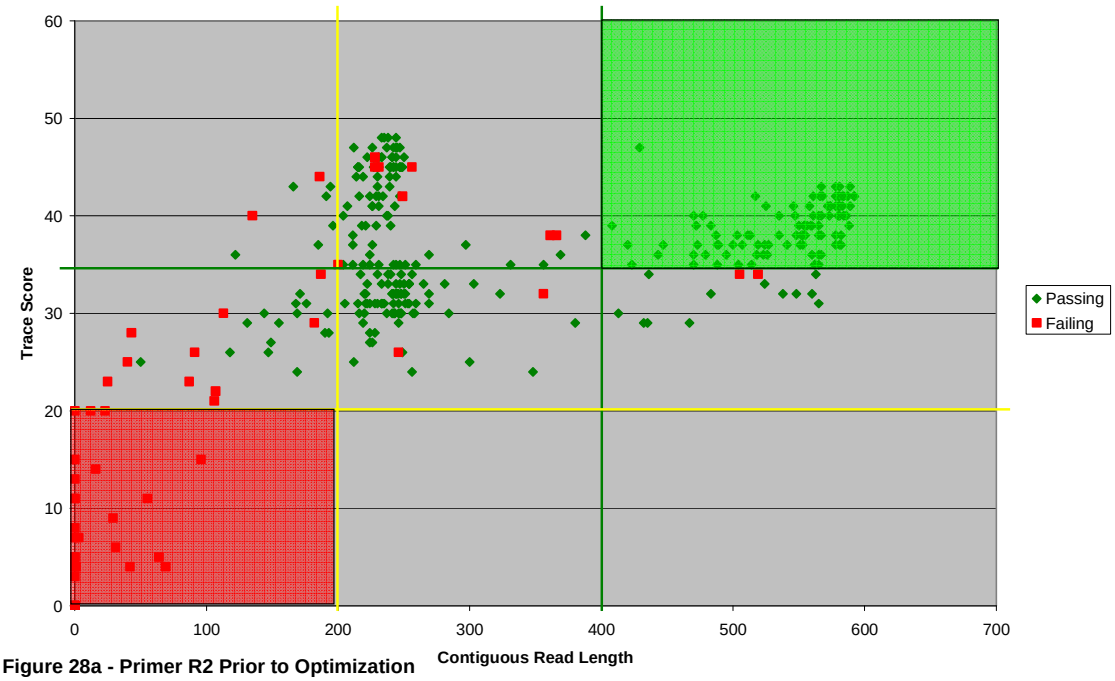


Figure 28a - Primer R2 Prior to Optimization

Figure 28b - Primer R2 After Optimization

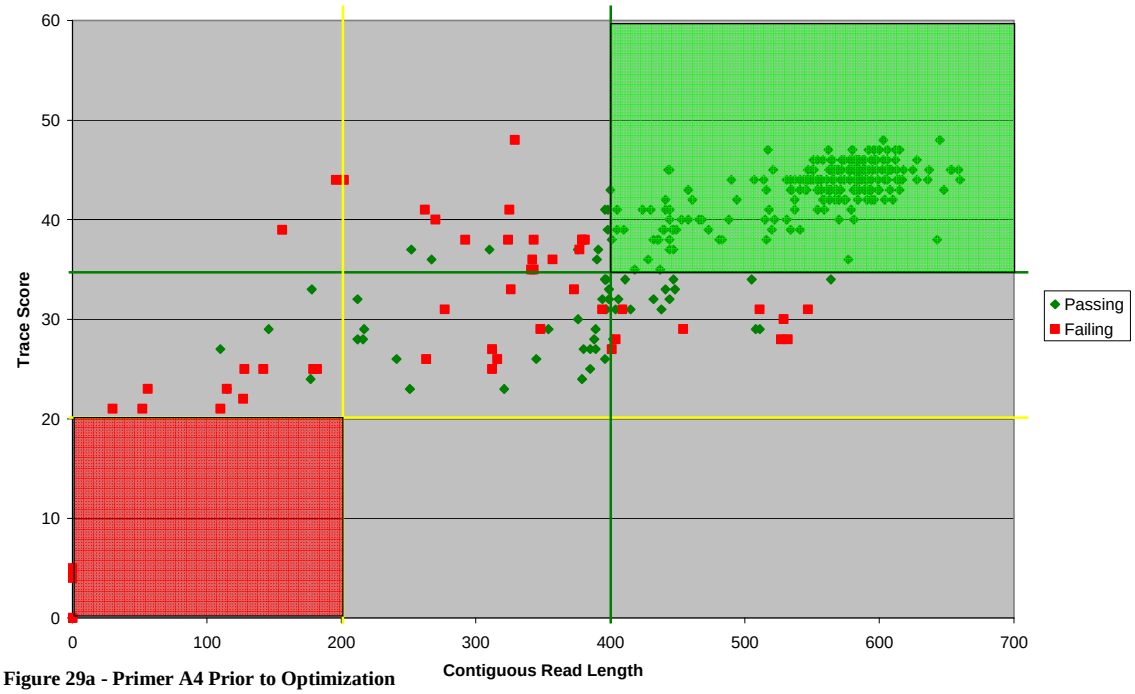


Figure 29a - Primer A4 Prior to Optimization

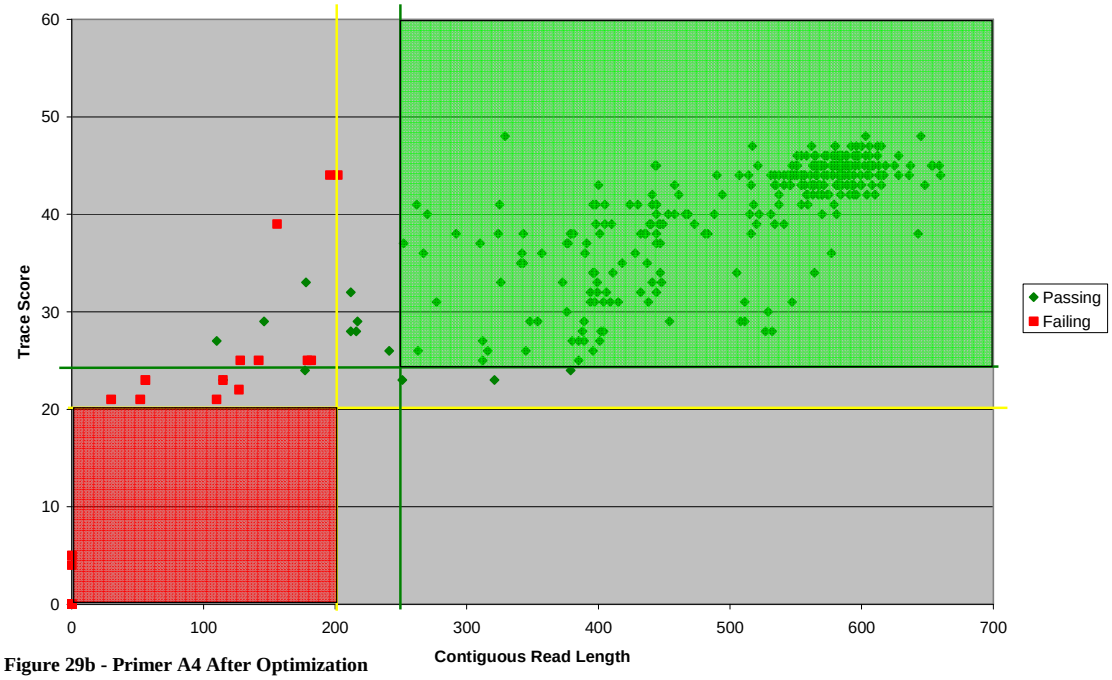


Figure 29b - Primer A4 After Optimization

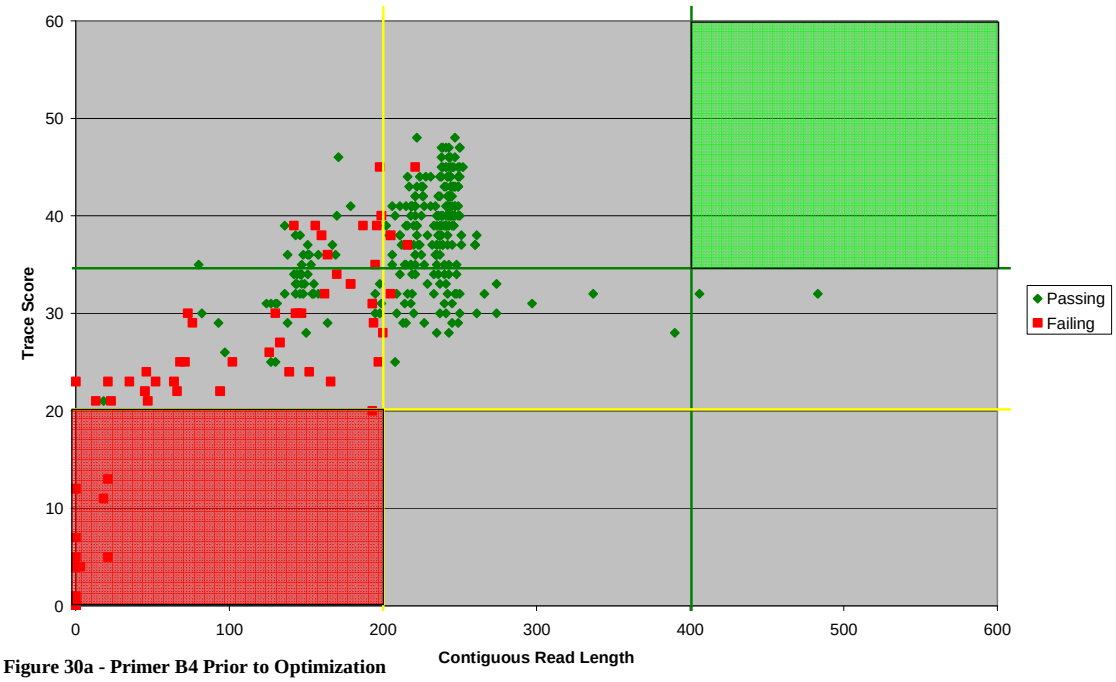


Figure 30a - Primer B4 Prior to Optimization

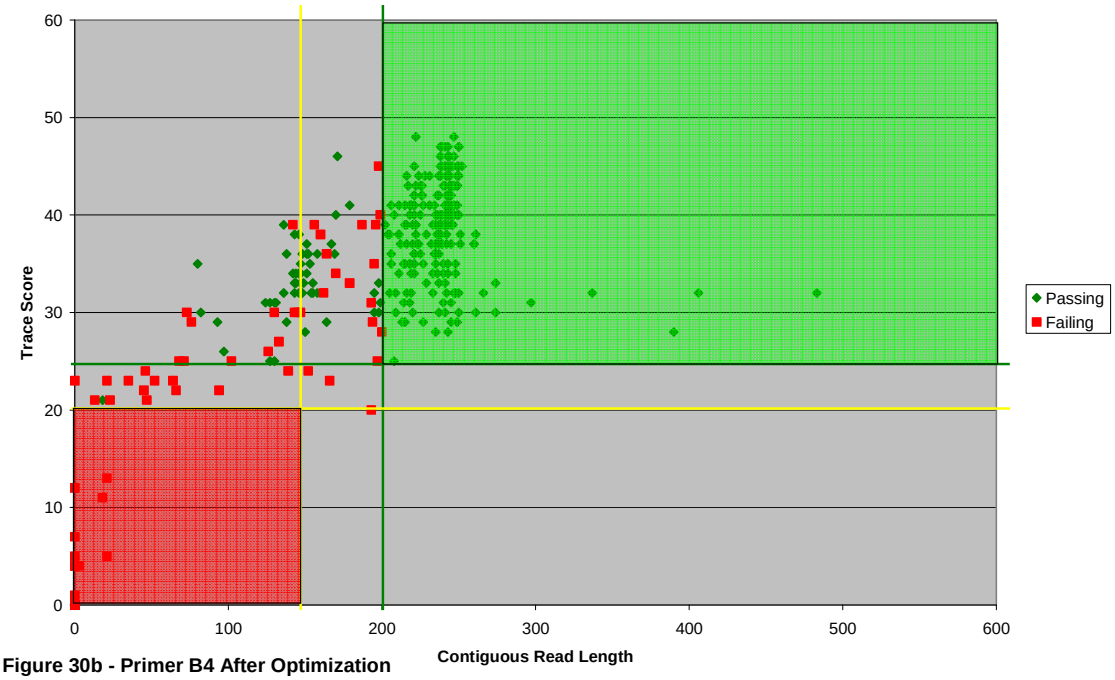


Figure 30b - Primer B4 After Optimization

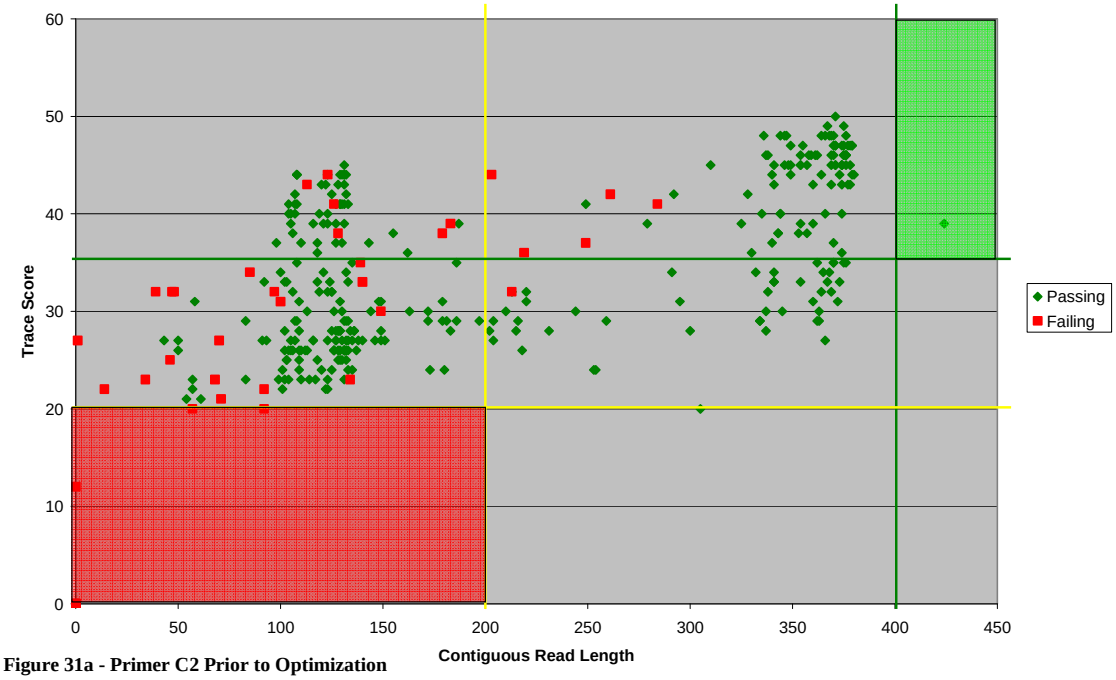


Figure 31a - Primer C2 Prior to Optimization

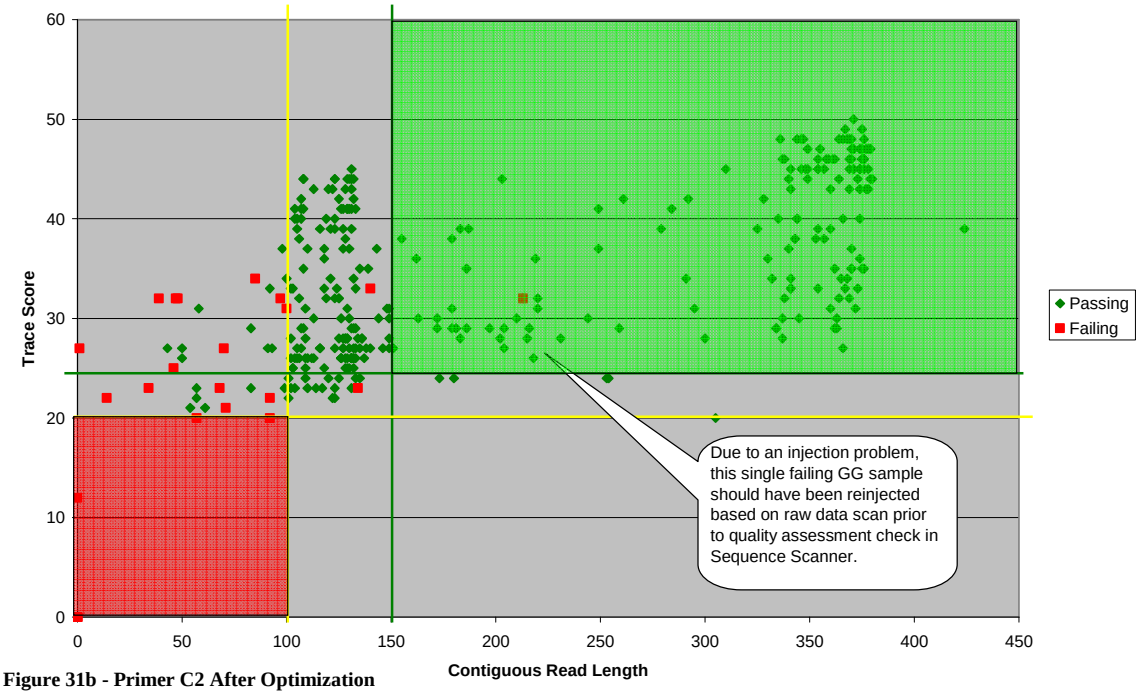


Figure 31b - Primer C2 After Optimization

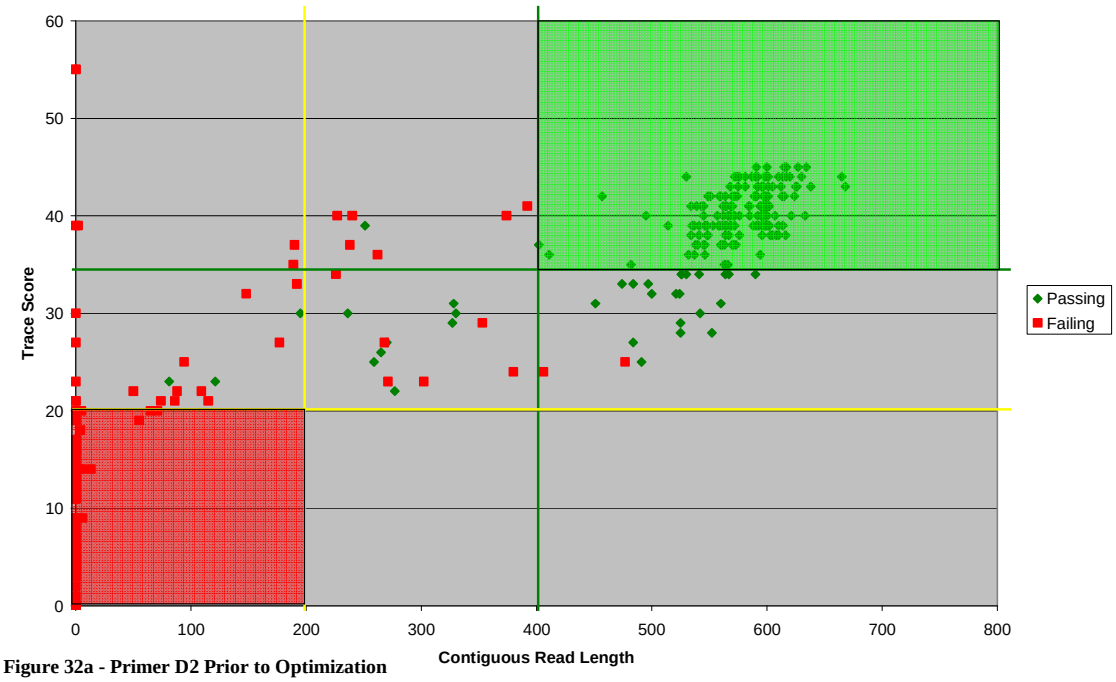


Figure 32a - Primer D2 Prior to Optimization

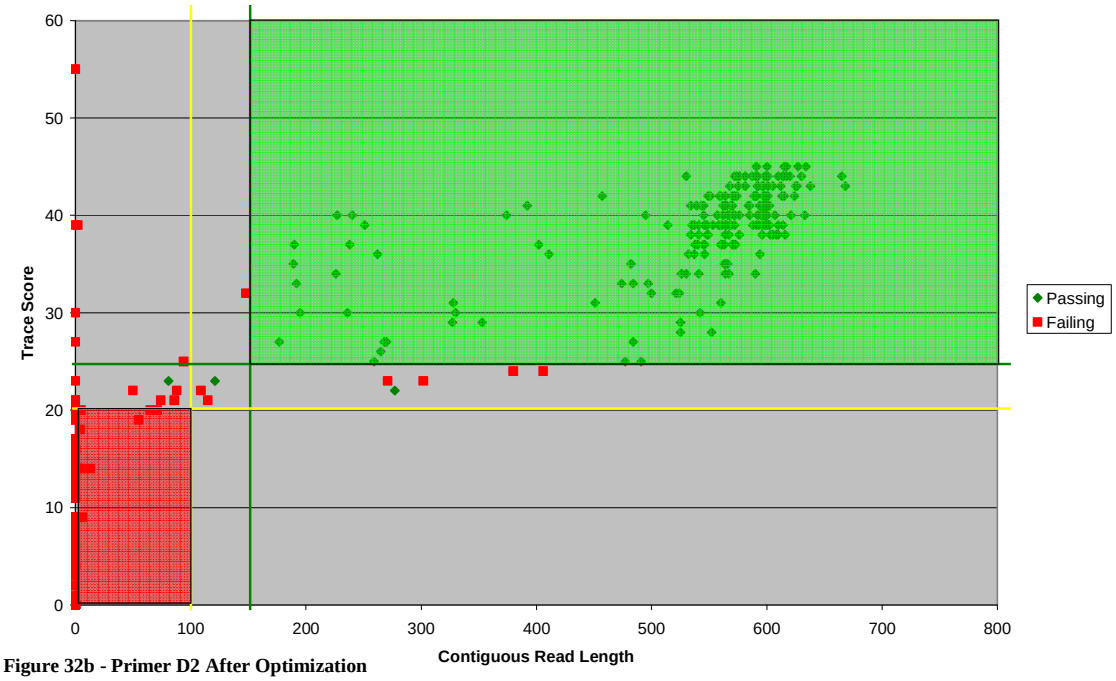


Figure 32b - Primer D2 After Optimization

APPENDIX C

GLOSSARY OF TERMS

Contiguous Read Length (CRL): the longest, uninterrupted stretch of bases with a quality value of greater than or equal to 20

Filter Metric (FM): sequence biometric calculation of FM = (QV20+/CRL)*Trace Score; criteria applied to sample to filter out bad data and return good data

Pe: probability of error

Quality Metric: parameter used to assess quality of a sequence

QV20+ value: total number of bases in an entire trace which have a base call quality value greater than or equal to 20

QV per base: quality value per base/peak which uses neighboring peaks quality and overlap as well as a Gaussian fit to determine the Pe when calculating the QV per base; $-10\log_{10}Pe$

Signal Intensity: the average raw signal intensity

Signal to Noise Ratio: the average raw signal to noise ratio

Trace: sequence, sample

Trace Score (TS): the average base call quality value of the post-trim sequence

REFERENCES

1. Li, Richard. *Forensic Biology*. Florida: CRC Press, 2008.

2. Mitotyping Technologies, LLC. Accessed November 8, 2008. <http://www.mitotyping.com/mitotyping/site/default.asp>

3. Scheffler, I.E. *Mitochondria*. New York: Wiley-Liss, 1999.

4. Fermentas Life Sciences; Accessed May 26, 2009. < http://www.fermentas.com/catalog/nucleotides/utp.htm >

5. Allard, Mark; *et al*. Characterization of human control region sequences of the African American SWGDAM forensic mtDNA data set. *Forensic Science International* 148 (2005) 169–179.

6. http://www.mitomap.org/images/mito2arcs.gif

7. Sutovsky, Peter; *et al*. Ubiquitin tag for sperm mitochondria. *Nature*. 402, 371-372. 25 November 1999.

8. Satoh, M. and Kuroiwa, T. (1991) *Experimental Cell Research*. 196, 137-140.

9. FBI Laboratory. (2005) NDIS DNA Data Acceptance Standards Appendix B. 19-20. Accessed May 27, 2009. <http://www.nlada.org/Defender/forensics/for_lib/Documents/1132070952.06/RF_GN_13_NDIS_Data_Standards%252005_31_05.pdf>

10. Roby, Rhonda. Higher Throughput Software Analysis. *High Throughput Mitochondrial DNA Analysis*. 2008.

11. L-strand of the rCRS AC_000021.2. Accessed April 17, 2009. <http://www.mitomap.org/mitoseq.html>