Sturm, Sarah A., A Novel Multiplex Assay for an Ancestry-Informative Marker (AIM) Panel of INDELs. Master of Science (Forensic Genetics), April 2016, pp. 28, 3 tables, 10 illustrations, 31 references.

The current standard for forensic laboratories in criminal casework is to use Short Tandem Repeat (STR) markers to develop an evidentiary profile. Commercially available STR amplification kits yield amplicons 100 to 500 base pairs (bp) in length. Commonly, forensic DNA samples are highly degraded to approximately 180-200 bps in length, resulting in incomplete STR profiles. Therefore, markers that can be generated with smaller amplicons may be better suited for degraded DNA samples. Additionally, there are cases where no STR match was obtained through a DNA database search and thus no investigative lead is obtained. The bioancestry of a sample donor could aid law enforcement in such cases.

A class of markers that could provide investigative value from degraded DNA samples is Ancestry-Informative Marker (AIM) Insertion/Deletions (INDELs). INDELs are polymorphisms that can be amplified from degraded samples due to their smaller amplicon size. AIMs have the ability provide bioancestry information. This project tested the hypothesis that a multiplex PCR-based assay of INDELs can be developed, and subsequently be analyzed by capillary electrophoresis for population identity testing applications. The use of this assay would require no additional tools or machinery than what already is in standard forensic laboratories. To test this hypothesis, a previously developed panel of AIM-INDEL markers was used to develop this multiplex assay.

A NOVEL MULTIPLEX ASSAY FOR

AN ANCESTRY-INFORMATIVE

MARKER (AIM) PANEL OF INDELS

Sarah Sturm, B.S.

APPROVED:

_____

Major Professor

_____

Committee Member

_____

Committee Member

_____

University Member

_____

Chair, Department of Molecular and Medical Genetics

_____

Dean, Graduate School of Biomedical Science

A NOVEL MULTIPLEX ASSAY FOR AN ANCESTRY-INFORMATIVE MARKER (AIM) PANEL OF INDELS

THESIS

Presented to the Graduate Council

of the Graduate School of Biomedical Sciences

University of North Texas

Health Science Center at Fort Worth

In Partial Fulfillment of the Requirements

For the Degree of

MASTER OF SCIENCE

By

Sarah A. Sturm, B.S.

Fort Worth, TX

April 2016

ACKNOWLEDGEMENTS

I would first like to thank Dr. Bobby LaRue for his encouragement and motivation to help me produce a project I am proud of. I would also like to thank the members of my committee, Drs. Bruce Budowle, Raghu Krishnamoorthy, and Rong Ma for their invaluable input. Excessive thanks are due to Jonathan King, M.S. for his contributions to my work, which include answering endless questions, ordering supplies, and then some. Thank you to the FGEN class of 2016, especially Kelly Sage, my partner in crime, without whom I do not think I would have survived. Finally, my deepest gratitude to my parents and my sister for believing in me always.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

CHAPTER 1

INTRODUCTION

Deoxyribonucleic Acid (DNA) typing has been considered the gold standard of forensic science for the past few decades. Scientists routinely use Short Tandem Repeat (STR) markers to differentiate individuals primarily due to their highly polymorphic nature, meaning the variation in the number of repeats at each locus among persons [1]. The Polymerase Chain Reaction (PCR) is used to amplify the repeat regions [1, 2], which are then separated by capillary electrophoresis (CE), and finally the data produced are analyzed *en silica* [3, 4]. The significance of the genetic profiles is determined using standard statistical calculations based on allele frequencies from reference population databases [5].

The process of PCR involves primers, or short oligonucleotide sequences, and a polymerase that mimics DNA duplication to generate exponentially the number of copies of the desired DNA target sequence [2]. Typically, the STR amplicons produced are between 100-500 base pairs (bp) in length. Each group of primer pairs has a different fluorescent dye attached, which allows the amplified products to be detected during CE. The amplified DNA is electrokinetically injected into the capillary and migrates from the cathode to the anode when voltage is applied. The DNA is separated by fragment size while migrating through the capillary due a sieving effect of the polymer, larger fragments move slower than smaller ones [3, 6]. As the fragments pass by a window in the capillary, a laser excites

the different fluorescent dyes; the emissions are captured by a camera which detects the specific labeled loci fragments. A program such as Gene Mapper ID-X creates a visual profile based on these data (Figure 1).

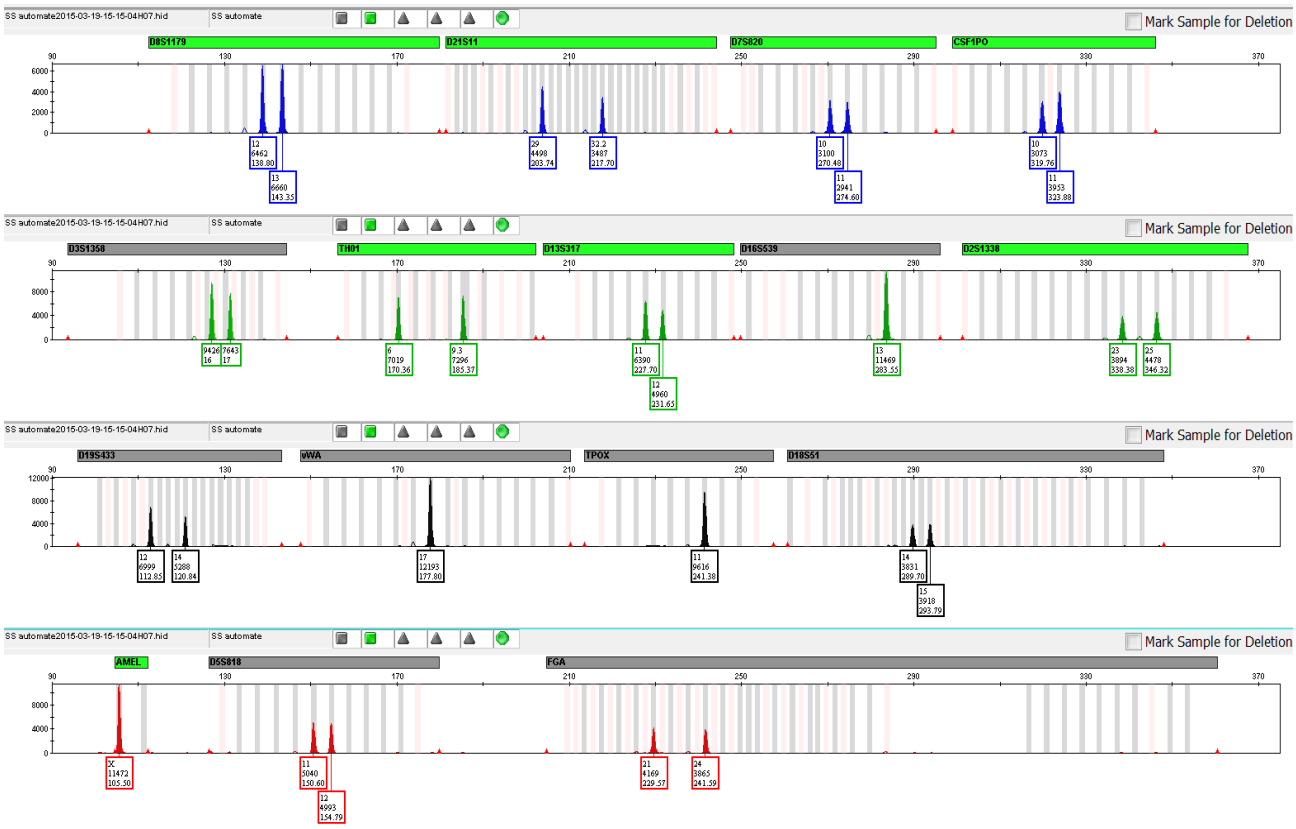


Figure 1: Genetic profile produced by Gene Mapper ID-X v1.2 of STR markers amplified with reagents from the Identifiler Plus kit on a 3500xL Genetic Analyzer. The different fluorescent dyes produce the different colors depicted.

To place significance on an evidence profile various statistical approaches are used. For single source profiles, a Random Match Probability (RMP) is calculated. This statistic is the probability that a random individual in a given population would have the same DNA profile as that from the evidence. The RMP is computed from allele frequencies in a population database and by multiplying the genotypic frequencies of each locus together [7, 8].

DNA, when exposed to the environment, can degrade. These environmental insults can damage the DNA and result in incomplete STR profiles, such as allelic dropout and other challenges [9, 10]. An example of an incomplete STR profile produced from a degraded bone (circa 1890) can be seen in Figure 2. Essentially little or no DNA was available to generate a profile.

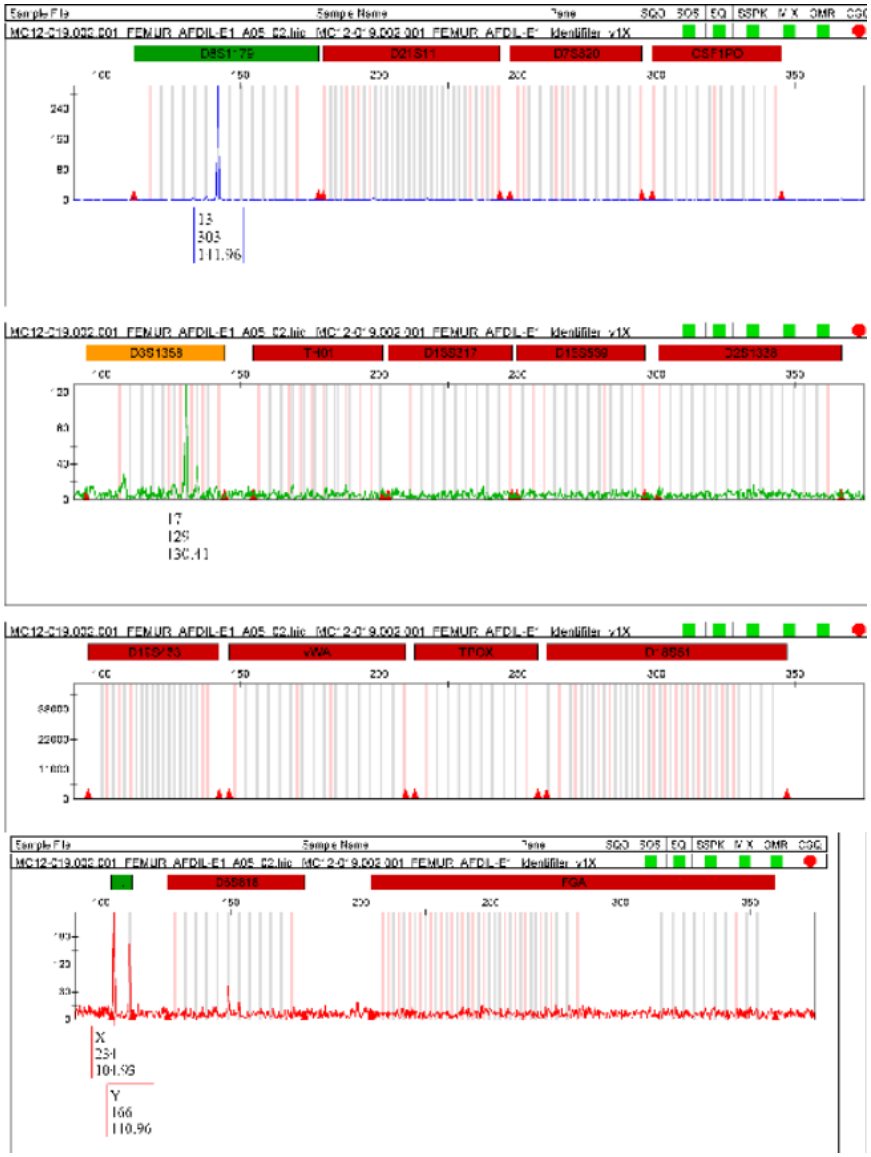


Figure 2: STR profile of 19th century bone. The green and yellow bars show the STR loci that could be typed. The blue, green, and red peaks indicate the alleles. Due to degradation of the bone, dropout occurred at most of the loci.

Many other options have been explored to attempt to analyze degraded samples. Typing of bi-allelic Single Nucleotide Polymorphisms (SNPs) is one possibility. SNPs are single variations in the DNA sequence at specific locations in the genome, and as a result can be captured in short amplicons. Short amplicons tend to yield results from degraded DNA more so than longer amplicons. However, SNP typing has its drawbacks [11, 12]. These single base pair changes in DNA sequence require complex sequencing techniques and expensive instrumentation to detect them. Unfortunately, these requirements do not make SNPs a current viable option for most forensic casework laboratories.

Another type of polymorphism found in DNA is an insertion/deletion (INDEL) of a sequence of DNA [13]. Some INDEL systems have been developed that are very useful for forensic casework [14-18]. Unlike STRs, the amplicon size of INDELs can be as small as 55 bps, making them ideal for analysis of degraded DNA [15]. Additionally, INDELs are non-repeating polymorphisms, therefore artifacts produced during PCR such as stutter do not occur. Another benefit of using INDELs is that the amplified product can be separated by fragment size by CE, equipment that is common to forensic laboratories [13]. Since the equipment currently used for CE in forensic laboratories can be used to genotype INDELs, there is no need for additional tools. It is anticipated that INDELs will provide results in some cases where STR typing was unsuccessful. The same bone sample from Figure 2 was amplified using an Insertion and Null Allele (INNUL) multiplex system, which is similar to an INDEL multiplex (Figure 3).

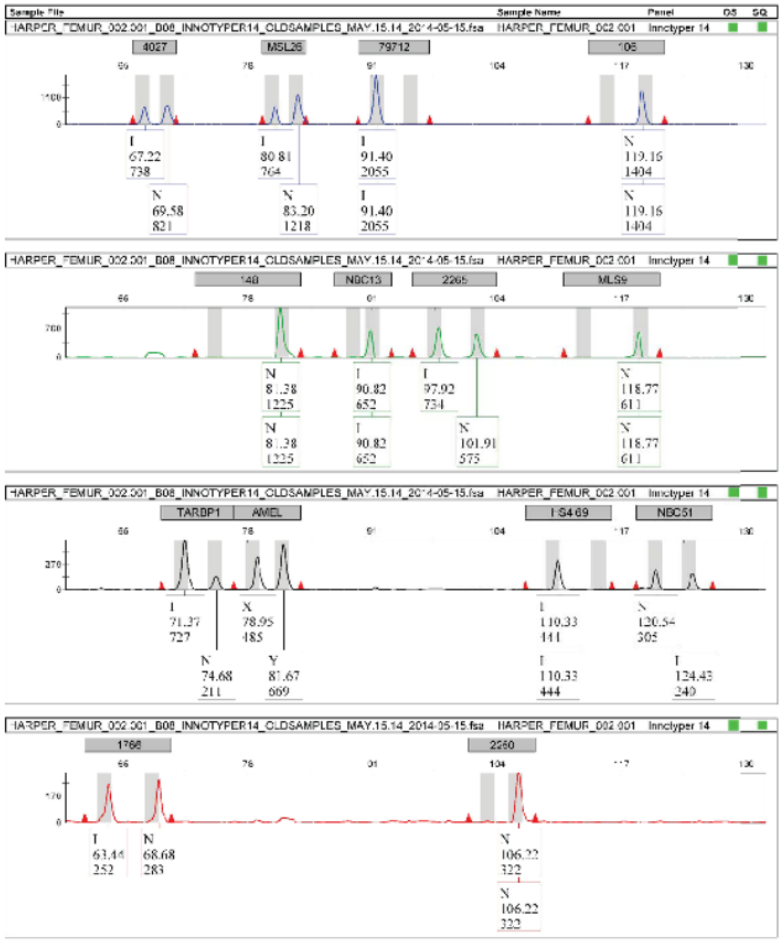


Figure 3: The bone (circa 1890) from Figure 2 amplified using an INNUL system. The smaller amplicons were able to type the degraded bone much more so than the STR system.

INDELs have recently been used to develop panels of markers for human identification (HID) purposes, as well as to amplify degraded DNA samples [14, 18]. HID markers are important in forensics for comparing suspect samples to evidentiary samples. Unfortunately, there often are cases where no suspect exists, and HID markers are limited in their ability to provide information in such situations. Ancestry-Informative Markers (AIMs) are polymorphisms at various loci in the genome that exist at different frequencies between populations. AIMs have the capability to provide not only genotypic information, but also population affinity information [19-24]. Being able to determine the population an

individual belongs to can help provide indirectly phenotypic information to investigators. This capability will be ground breaking in solving crimes.

Thompson [26] successfully developed an AIM-INDEL panel of 59 markers to genetically differentiate Caucasian, African, and East Asian populations using genetic data from the 1000 Genomes Project. All INDELs were chosen to be between 3-6 bp in length and have high F_{ST} values. F_{ST} is a measure of pairwise population substructure. High F_{ST} values were desired in this study to separate the overall population into subpopulations. Population-specific allele frequencies were also used to distinguish between the populations at each marker. High frequency divergence was desired in order to do this. For example, at marker rs139570718 a high frequency of the insertion is present in the Caucasian population, but there is a low frequency of insertion in the African and East Asian populations. Looking at this marker is therefore a good indication of Caucasian descent. The same is true for all other markers chosen in the panel, where either an insertion or deletion may be indicative of ancestry. The resolution of the Caucasian, African, and East Asian populations makes this panel of markers useful for a new multiplex assay for forensic casework. Table 1 lists the markers and allele frequencies by population.

Table 1: AIM-INDEL markers chosen for 3 different populations and their descriptive statistics.

CAUCASIAN								
rs#	Chrom.	Position	Sequence	Frequency[1] African	Caucasian	East Asian	Pairwise Fst[2] v.AFR	v.EAS
rs139570718	1	214397853	CCCAG (Ins)	0.0352564	0.727459	0.223333	0.640101	0.400736
rs3831920	1	1227664	TGAG (Del)	0.375	0.913934	0.293333	0.508888	0.600295
rs141516305	2	13725708	AGCTTT (Del)	0.865385	0.278689	0.65	0.504749	0.244201
rs67934853	2	74943887	TAAC (Del)	0.923077	0.258197	0.81	0.603647	0.460751
rs139220746	2	200205694	TATC (Del)	0.826923	0.227459	0.673333	0.52312	0.338786
rs140498743	3	139232513	TGTC (Del)	0.842949	0.360656	0.95	0.37517	0.517759
rs5864437	[1]4	178146869	CTAT (Del)	0.839744	0.192623	0.803333	0.585954	0.542572
rs149676649	5	28495386	GATT (Ins)	0.349359	0.79918	0.106667	0.350045	0.637831
rs57237250	6	110263002	GAGT (Ins)	0.826923	0.260246	0.903333	0.479728	0.574514
rs1160871	7	28168745	TCTT (Del)	0.217949	0.788934	0.0233333	0.491182	0.72318
rs72404898	8	122272251	ATAGAG (Del)	0.855769	0.381148	0.996667	0.368124	0.561435
rs67538813	9	30471814	CAGA (Del)	0.958333	0.383197	0.696667	0.507485	0.17589
rs10651200	10	69800907	TAACAA (Ins)	0.939103	0.334016	0.83	0.525682	0.389713
rs140507887	10	28470438	AATA (Del)	0.74359	0.348361	0.996667	0.266669	0.596017
rs11576045	12	111799524	TGT (Del)	0.762821	0.235656	0.936667	0.433617	0.646023
rs35779249	13	43964476	TAA (Ins)	0.961538	0.29713	0.82	0.607878	0.422731
rs6145374	14	65368820	CTTGA (Del)	0.910256	0.209016	0.63	0.648534	0.314874
rs138439822	15	35537968	TAACTC (Del)	0.858974	0.270492	0.713333	0.506957	0.327046
rs10528149	16	69989686	TGAT (Del)	0.0769231	0.721311	0.36	0.578944	0.233118
rs138814632	17	79605107	ATTAA (Del)	0.304487	0.657787	0.003333333	0.219042	0.602563

[1]Allele frequency is the frequency of an allele, in this case the insertion or deletion in the sequence column, at a locus (rs#) in each population.

[2]Pairwise F_{ST} is the measure of the population substructure versus each of the other two populations. The abbreviations are: AFR for African, EAS for East Asian, and CAU for Caucasian.

EAST ASIAN								
rs#	Chrom.	Position	Sequence	Frequency African	Frequency Caucasian	Frequency East Asian	Pairwise Fst v.AFR	Pairwise Fst v.CAU
rs141933116	1	8189066	AAGT (Del)	0.701923	0.956967	0.39	0.176461	0.579729
rs5839799	2	241417278	GTCT (Del)	0.88141	0.694672	0.286667	0.533799	0.282733
rs72375069	3	27427821	AATT (Del)	0.980769	0.657787	0.256667	0.7167	0.273236
rs33915414	4	21762063	CATGTT (Del)	0.080128	0.385246	0.803333	0.693429	0.295226
rs1610951	5	108999835	TTGG (Del)	0.971154	0.868852	0.336667	0.61792	0.475083
rs147268567	6	21621169	TTAA (Del)	0.285256	0.284836	0.89	0.544839	0.527296
rs151280400	7	125249166	AATC (Del)	0.910256	0.659836	0.35	0.504593	0.17317
rs10581451	8	73854660	TGAG (Del)	0.894231	0.965164	0.18	0.677952	0.799592
rs150560593	9	95478810	TGCA (Del)	0.865385	0.739754	0.283333	0.514276	0.344585
rs150244296	10	94941566	TTGAC (Del)	0.971154	0.885246	0.1333333	0.831329	0.724881
rs143873637	11	97893598	TTGA (Del)	0.823718	0.866803	0.243333	0.504557	0.57738
rs66693708	12	77398405	TAAG (Del)	0.974359	0.805328	0.326667	0.633852	0.386204
rs10587399	13	37776954	TACT (Del)	0.887821	0.717213	0.243333	0.594295	0.362933
rs141122561	14	49242955	TTAGT (Del)	0.996795	0.963115	0.37	0.627839	0.612742
rs71964979	15	102264144	GCAGG (Del)	0.714744	0.702869	0.13	0.515882	0.486211
rs35968516	17	5328978	TTTA (Del)	0.852564	0.719262	0.18	0.622449	0.443694
rs143394724	18	52716306	ATGTC (Del)	0.983974	0.786885	0.376667	0.598112	0.301386
rs33965072	19	266759	GAAAG (Ins)	0.86859	0.63729	0.14	0.692713	0.393972
rs11474791	20	19234875	GGACT (Ins)	0.221154	0.1086	0.79	0.487299	0.652616
rs3074939	21	43422429	CAGT (Del)	0.205128	0.364754	0.836667	0.569109	0.361085

AFRICAN								
rs#	Chrom.	Position	Sequence	Frequency African	Frequency Caucasian	Frequency East Asian	Pairwise Fst v.EAS	Pairwise Fst v.CAU
rs150866650	1	16367160	AAGG (Ins)	0.314103	0.821721	0.99	0.66481	0.424857
rs202017686	1	248818535	AAGAT (Del)	0.689103	0.0881148	0.24666	0.325636	0.575336
rs11277277	2	11273217	CACAG (Del)	0.339744	0.987705	0.93666	0.552569	0.687956
rs137858080	2	178513061	GTTT (Del)	0.875	0.256148	0.263333	0.551826	0.545406
rs148921522	3	85588405	TAAC (Ins)	0.160256	0.625	0.86	0.656259	0.354217
rs112191273	3	7351968	GCTT (Ins)	0.657051	0.0266393	0.0433333	0.580653	0.656176
rs72228292	4	106669965	AGTT (Del)	0.916667	0.243852	0.12	0.776951	0.613038
rs72255563	5	176226827	ACTT (Del)	0.772436	0.114754	0.136667	0.576689	0.622541
rs150723104	6	155859718	CCAA(Ins)	0.75	0.239754	0.156667	0.521703	0.412464
rs35379320	7	79883089	AGAT (Ins)	0.894231	0.354508	0.106667	0.764883	0.450782
rs56767439	8	12977501	TTAC (Del)	0.810897	0.204918	0.156667	0.59818	0.534393
rs113043680	9	126640635	TAAG (Ins)	0.708333	0.139344	0.0966667	0.556619	0.511686
rs113501732	10	128948642	CCTGT (Ins)	0.272436	0.911885	0.763333	0.386335	0.616993
rs139666905	11	5270343	AAAG (Del)	0.746795	0.30123	0.15	0.526348	0.326999
rs74499778	11	129941381	AGCT (Del)	0.375	0.952869	0.62	0.110407	0.583277
rs2307553	14	80121686	TGAC (Ins)	0.884615	0.252049	0.38	0.430009	0.562808
rs138123572	15	72786235	TGAC (Ins)	0.185897	0.959016	0.946667	0.738709	0.782133
rs66913380	17	42191379	GCCA (Del)	0.195513	0.786885	0.85	0.598891	0.515387
rs149016222	20	59105205	CTTC (Del)	0.272436	0.75	0.87	0.531245	0.371308

Multiplex assays are commonplace for DNA typing systems, such as the kits used to amplify STR markers. A multiplex assay is the simultaneous amplification of multiple genetic markers for characterization and interpretation. One of the most important aspects of designing a multiplex is primer design, since they must capture the desired sequence [27]. Important characteristics that impact primer design are the melting temperature, Gibbs free energy change (ΔG), GC content, primer length, and sequence length. It is also important that the primers do not form dimers. Primer-dimers deplete the multiplex reaction of primers when they bind to each other, and therefore interfere with amplification of the target DNA. Using these criteria, the above panel of AIM-INDEL markers was developed into a usable multiplex assay.

CHAPTER 2

RESEARCH DESIGN AND METHODOLOGIES

Primer Selection

Thirty markers were chosen from Thompson's panel [26] to develop primer pairs for a multiplex assay. The FASTA, or nucleotide, sequence for each marker was acquired using the dbSNP page through the National Center for Biotechnology Information (NCBI) website. If the sequence was not found though this method, the University of California Santa Cruz (UCSC) Genome Browser was used instead. The program Primer-BLAST [28] was used to design primers *en silica* by inputting the FASTA, or nucleotide, sequence of each marker. Primer pairs were checked for potential dimerization using MPprimer [29].

Unlabeled Primers

Unlabeled primers were obtained from Invitrogen™ and reconstituted with nuclease free water to a concentration of 100 µM. Each marker was amplified in singlet with DNA from a known sample. Each amplification reaction contained 5.5 µL water, 2.5 µL 10X buffer, 2.5 µL bovine serum albumin (BSA; 10 mg/mL), 2.0 µL magnesium chloride ($MgCl_2$; 25 mM), 1.0 µL deoxynucleotide triphosphate mixture (dNTPs), 0.5 µL Taq Gold® polymerase (5 U/ µL), 0.5 µL forward primer (10 µM), 0.5 µL reverse primer (10 µM), and 10 µL template DNA (1 ng/µL). The samples were amplified on the Applied Biosystems® GeneAmp® PCR System 9700 thermal cycler under the following parameters: 95°C for 11 minutes, 36 cycles of 95°C for 10 seconds, 61°C for 30 seconds, 72°C for 30 seconds, and a

final extension at 70°C for 10 minutes. The amplified product was analyzed on the Agilent© 2200 TapeStation [30] using 2 µL of TapeStation buffer and 2 µL of sample in each well. All primer pairs successfully amplified DNA, except number 18. Therefore, 29 of the 30 primer pairs were arranged into six multiplexes. Five sets of five primer pairs, and one set of four primer pairs, were amplified using the Qiagen® Multiplex PCR Plus Kit [31] on the thermal cycler in a multiplex fashion with the same known sample of DNA under the following parameters: 95°C for 5 minutes, 35 cycles of 95°C for 30 seconds, 60°C for 90 seconds, 72°C for 90 seconds, and a final extension at 68°C for 10 minutes. The amplified products were run on the TapeStation to assess whether successful simultaneous amplification occurred. After successfully completing this multiplex trial, the primer pairs were rearranged into the five dye channels used with the GlobalFiler™ amplification kit (Table 2) with at least 10 bps between each marker. Since three of the product lengths were similar and could potentially overlap, they were assigned to dye channels and considered as alternates (highlighted below) in the case that the first marker does not work properly.

Table 2: Primer pairs arranged into dye channels and the expected sequence lengths of each (bps).

Dye Channel	Primer Pair	Alleles (Expected bp sequence lengths)
Blue (6-FAM)	4	59, 64
	28	60, 64
	25	69, 73
	13	83, 86
	23	114, 118
	15	134, 139
	1	151, 155
	10	171, 175
Green (VIC)	16	61, 65
	29	61, 65
	19	78, 82
	12	92, 97
	27	132, 136
	2	142, 147
Yellow (NED)	9	60, 64
	30	60, 64
	22	74, 78
	5	107, 112
	17	137, 143
	11	151, 155
Red (TAZ)	6	59, 64
	3	72, 76
	26	83, 87
	8	128, 133
	14	140, 144
Purple (SID)	24	59, 63
	21	65, 70
	20	94, 98
	7	136, 140

Fluorescently Labeled Primers

Fluorescently labeled forward primers (20 μM) were attained from Applied Biosystems®. A 10 μM working solution was prepared of each primer, and the primer pairs were tested in singlet with a known sample by CE on the Applied Biosystems® 3500xL Genetic Analyzer [6]. Each amplification reaction contained 5.5 μL water, 2.5 μL buffer, 2.5 μL BSA, 2.0 μL $MgCl_2$, 1.0 μL dNTPs, 0.5 μL Taq Gold® polymerase, 0.5 μL labeled forward primer, 0.5 μL unlabeled reverse primer, and 10 μL template DNA (0.5 ng/μL). Each sample well for the CE contained 9.6 μL of HiDi Formamide, 0.4 μL of LIZ 600 Size Standard, and 1 μL of sample. Results were analyzed using GeneMapper ID-X software (v. 1.2). After successfully amplifying and separating the markers individually, primer pairs with the same fluorophore were amplified together in a multiplex using the Qiagen Multiplex PCR Plus Kit. The results of the singleplexes and multiplexes were oversaturated. The amplified products of the five multiplexes were diluted 1:10, 1:20, 1:50, and 1:100, and run by CE again to reduce oversaturation. The 1:100 dilution produced the clearest results; thus a 1:100 dilution of the DNA sample was prepared. The primer sets were used to amplify the DNA in singlet and multiplex once again, and subsequently analyzed.

CHAPTER 3

RESULTS AND DISCUSSION

Primer pairs were chosen to have a sequence length, within the range of 60-160 base pairs, as shown in Figure 4 under Primer Parameters. Primer-BLAST produced potential primer sequences, lengths, and the associated melting temperatures, and G-C content percentage (Figure 5). Once primer pairs were chosen, they were checked for potential dimerization using MPprimer. Each forward primer is compared to each reverse primer, and the output gives matches, an alignment score, 3'-3' dimer check, and ΔG (kcal/mol). Any primers with alignment scores of 5 or greater, or ΔG of -7 or less were discarded, and different primer pairs for those markers were chosen. The final 30 primer pairs are shown in Table 3.

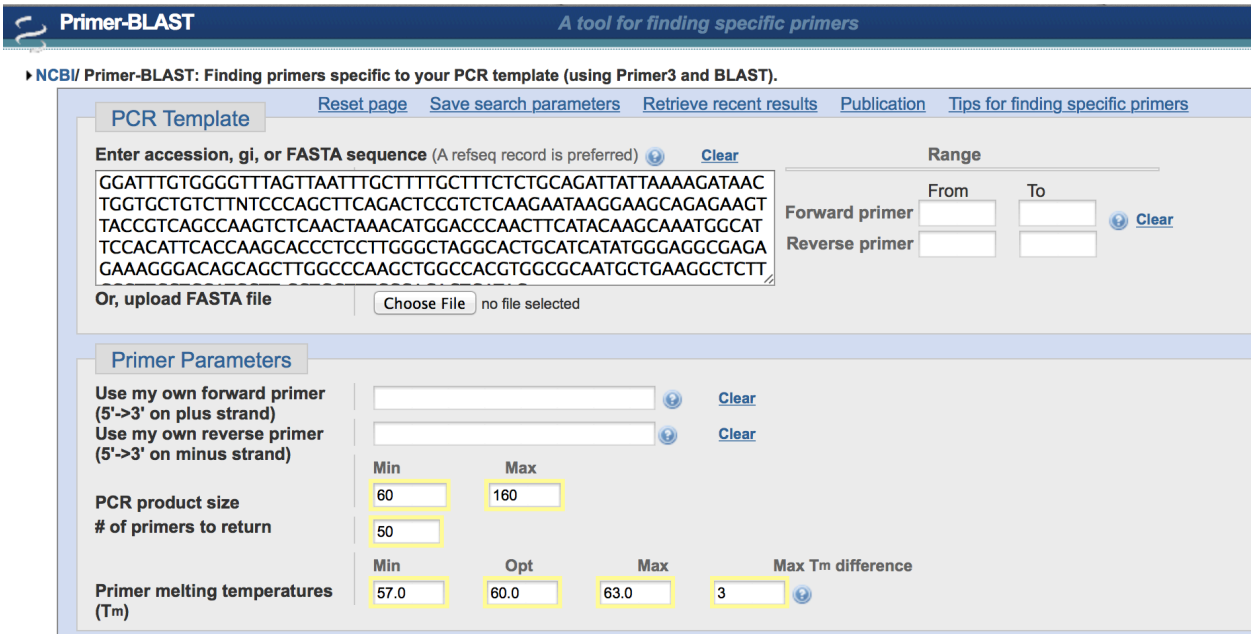
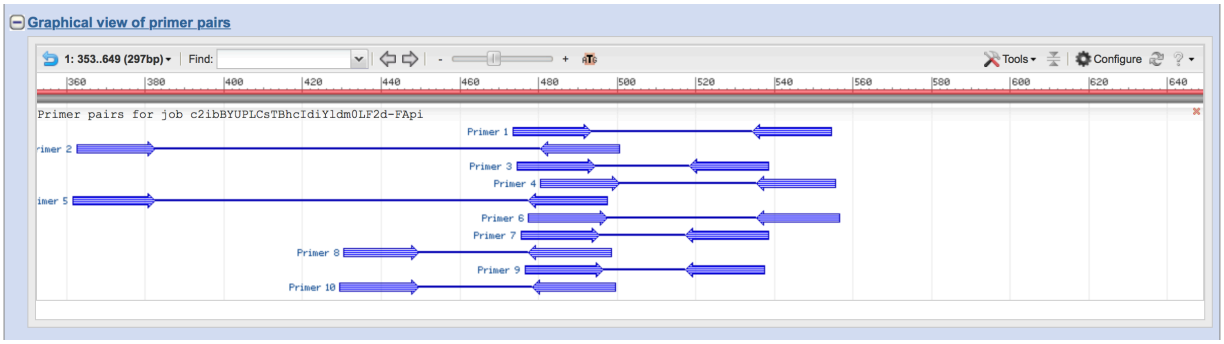


Figure 4: Screen capture of Primer-BLAST input including box for FASTA sequence, and parameters used for primer design.

A)



B)

Primer pair 9

	Sequence (5'->3')	Template strand	Length	Start	Stop	Tm	GC%	Self complementarity	Self 3' complementarity
Forward primer	AGCTCCCTAGCATTGGACAG	Plus	20	477	496	59.16	55.00	4.00	1.00
Reverse primer	GGGGTATTCACAGAGGGTCT	Minus	20	537	518	58.12	55.00	3.00	3.00
Product length	61								

Figure 5: Screen capture of Primer-BLAST output for rs1160871.
A) Graphical representation of the potential primer pairs; **B)** Data output of primer pairs.

Table 3: Top 30 markers and selected forward and reverse primers.

Order #	rs#	Chrom	Position	Forward Primer	Reverse Primer	Seq Length	INDEL
1	rs138123572	15	72786235	GCTTTTCTCCATAACCTCAGA	TTTGTGCTTTTTGAATTTGAACC	151	TGAC (Ins)
2	rs139570718	1	214397853	CACTTCTAGGGATTTGTGGGGT	AGTTGAGACTTGGCTGACGG	147	CCCAG (Ins)
3	rs67934853	2	74943887	ACCAGTACTGCAAGACAAAGAGT	GCAAGTGGGACGGAGTGTAA	72	TAAC (Del)
4	rs370096890	14	65368820	ACCAAATGCTTGGAAGTCTTGA	AACTGGGGCCAGGTGTTAAT	59	CTTGA (Del)
5	rs113501732	10	128948642	TCAATCCCCATTGCTCACCC	CTGTGTGATTCTGCCCTGGT	106	CCTGT (Ins)
6	rs67205569	10	94941566	CCAGGGTCTAAACAGAGGCA	TGACCCAGAATCCTGTGACTT	64	TTGAC (Del)
7	rs35625334	7	79883089	AGCAACATGGCCTTAGGTTTT	AGCTTGTTTGTGATCCCACG	136	AGAT (Ins)
8	rs10668859	19	266759	CAGGAGTAGCCCATCATGAACA	CCCTAAGCTGGACTGTCTCC	128	GAAAG (Ins)
9	rs1160871	7	28168745	AGCTCCCTAGCATTGGACAG	GGGGTATTCACAGAGGGTCT	60	TCTT (Del)
10	rs149676649	5	28495386	TTGTTTGTCCCTGTATTTAACAGAA	ATTGCATTGTGCATTTTTGTCATGT	171	GATT (Ins)
11	rs10581451	8	73854660	ATGAAGTGATTTTCCAAAGAACTGT	AGGAAAGACAACCCATAACCTCA	151	TGAG (Del)
12	rs11474791	20	19234875	TCCCACAGAGTGACATTGCC	GAACCCCTGGACCATGTGAG	92	GGACT (Ins)
13	rs35779249	13	43964476	TTGCACCAGATGGCTGTGT	TTTGCAGGCATTCTCCTTGAT	83	TAA (Ins)
14	rs72375069	3	27427821	TAAATCCCTTGCACTACGCA	AGGTACTCTAATGTATTGCTGAAGA	140	AATT (Del)
15	rs55885844	17	79605107	ACCAGGAAACCGGAAGACTAAA	GGCACCCTGAGCAAACTAATAC	134	ATTAA (Del)
16	rs66913380	17	42191379	CAGCATGGCCTGGGAGC	GAGAGGGTTCAGCCAACACC	61	GCCA (Del)
17	rs33915414	4	21762063	CGCCTACAAATTCATGCTGCT	GTCTCTAAAACCCATAATTTGCCTG	143	CATGTT (Del)
18	rs72255563	5	176226827	ACACGCACACTCAGCACAC	GGAGACACACGTCTCCATGC	65	ACTT (Del)
19	rs148921522	3	85588405	AGTAGACTGACACATAAGGCTGTA	ACACTTTGAACTCTTGAGAAATGTT	78	TAAC (Ins)
20	rs3831920	1	1227664	TGAGCCGGGTAGCACTCA	GGGCATCAGGACCCAGATTT	94	TGAG (Del)
21	rs11277277	2	11273217	CCTTTCCTAGGAGCTGTCCG	AGTTTCGTTTTGAACTCCCGC	65	CACAG (Del)
22	rs59385244	1	16367160	AAATCACCACCCTGCCTGAG	AAGTGCAGCAGGAAAAGCTC	73	AAGG (Ins)
23	rs71991275	10	28470438	TGCCACAACTTGAGCTGACT	TCGTGGGGCACGATAATAGA	114	AATA (Del)
24	rs5864438	4	178146869	CTGAACCTGGACGTGGTCAT	CCAGAGTGGATGCACCATAGAC	59	CTAT (Del)
25	rs57237250	6	110263002	TGCTGTTCTCATTCCACGTAT	AGTTAGCCATGGGAAGCACA	69	GAGT (Ins)
26	rs1610951	5	108999835	ATGTCAAGCACCGTGCCA	CTGTGTGACCTCTCTGAGC	83	TTGG (Del)
27	rs367799178	6	21621169	TTGCATTATGGCCAAAAATCATGT	CAGTTCCAACACAAAGGTAGCA	136	TTAA (Del)
28	rs10549914	17	5328978	AGCAATCAGTTCTCTTTGTCAAC	ACAGATACAGAATGTCAGGGTC	60	TTTA (Del)
29	rs112191273	3	7351968	TGGTGATGATTTTCAAATGGGACT	ACATTGCAGATTTAACTCATGAACC	61	GCTT (Ins)
30	rs56767439	8	12977501	ATGCCATAGTGAGAGAAGGAACA	ACCTGTCTTGCAGGAAGAACC	59	TTAC (Del)

The TapeStation, which was used to analyze the amplicon results, is an automated electrophoresis system that uses a ScreenTape matrix similar to agarose gel. The samples absorb an intercalating dye, are separated by size, and then fluorescence is captured by a camera (Figure 6). Using a ladder, a reference of bp length, the approximate size of the amplified product can be determined. An electropherogram is then produced by the program to give a graphic representation of the sequence lengths (Figure 7). The observed amplicon size was close to that predicted for all primer pairs. Primer pair 18 produced no product after multiple attempts, and was therefore removed from further testing. The TapeStation results from the initial multiplex trial are show in Figure 8A-H.

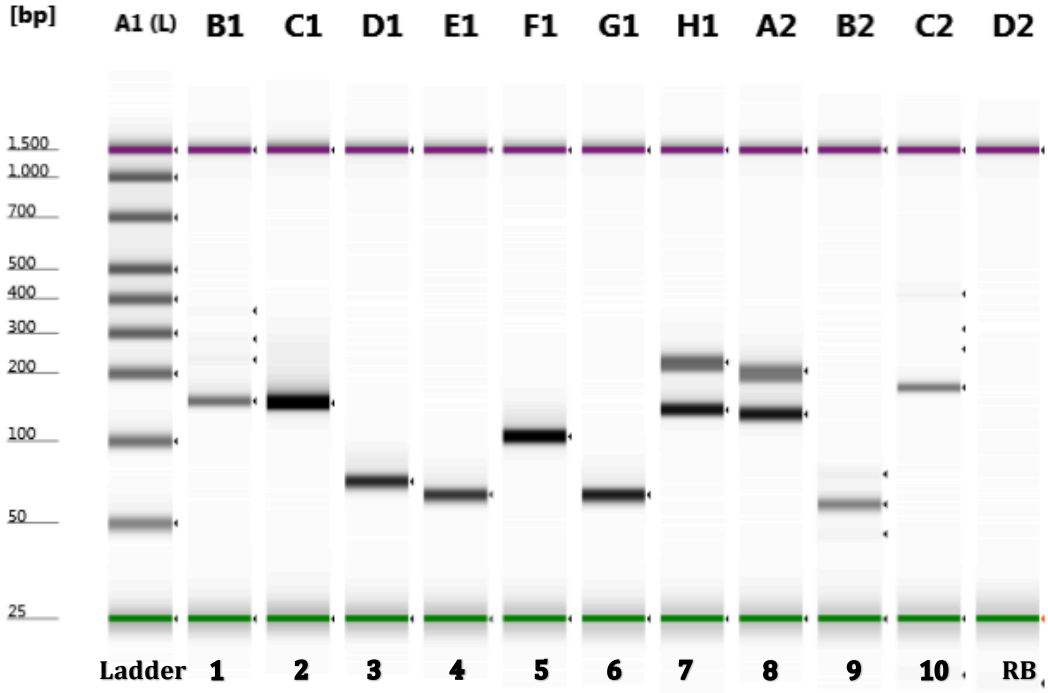


Figure 6: Electrophoresis results of primer pairs 1-10 on the Agilent© 2200 TapeStation.

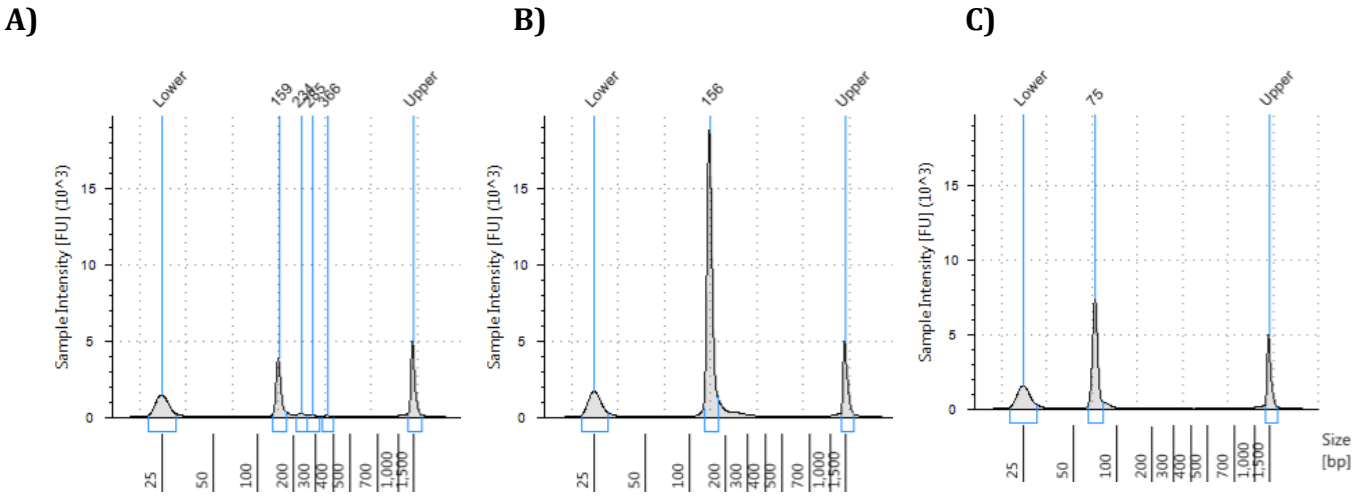
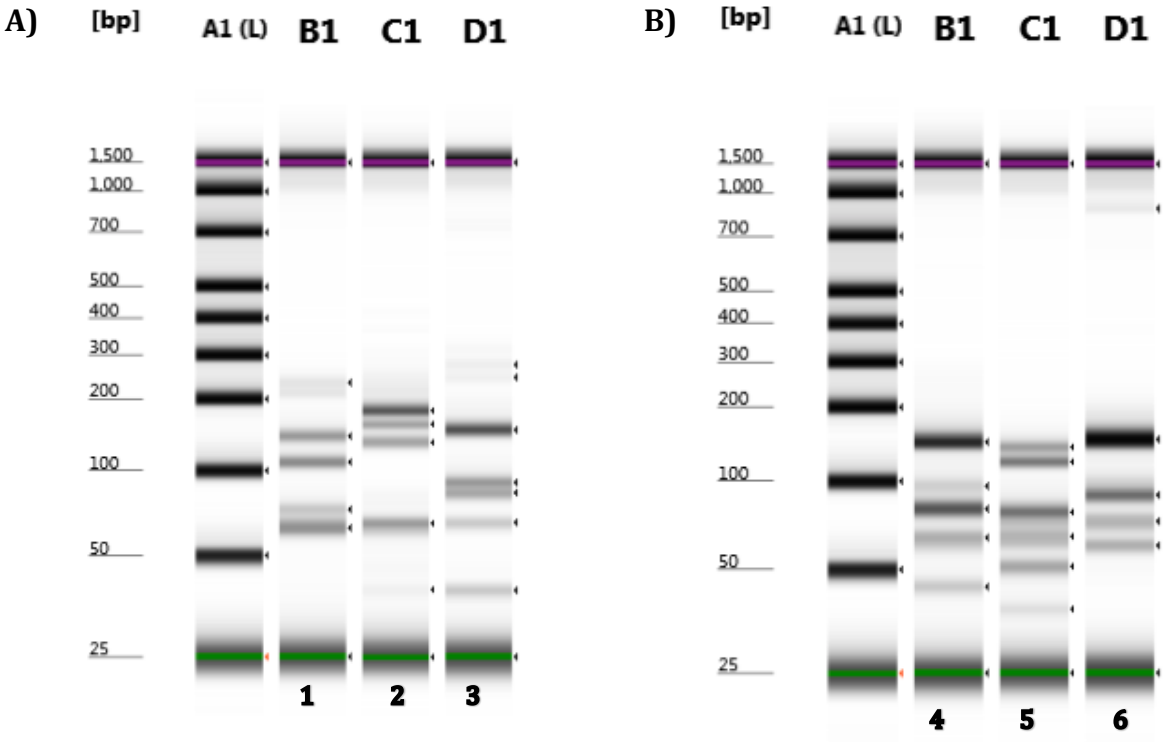


Figure 7: Electropherograms produced by electrophoresis of primer pairs 1 **(A)**, 2 **(B)**, and 3 **(C)** on the Agilent© 2200 TapeStation.



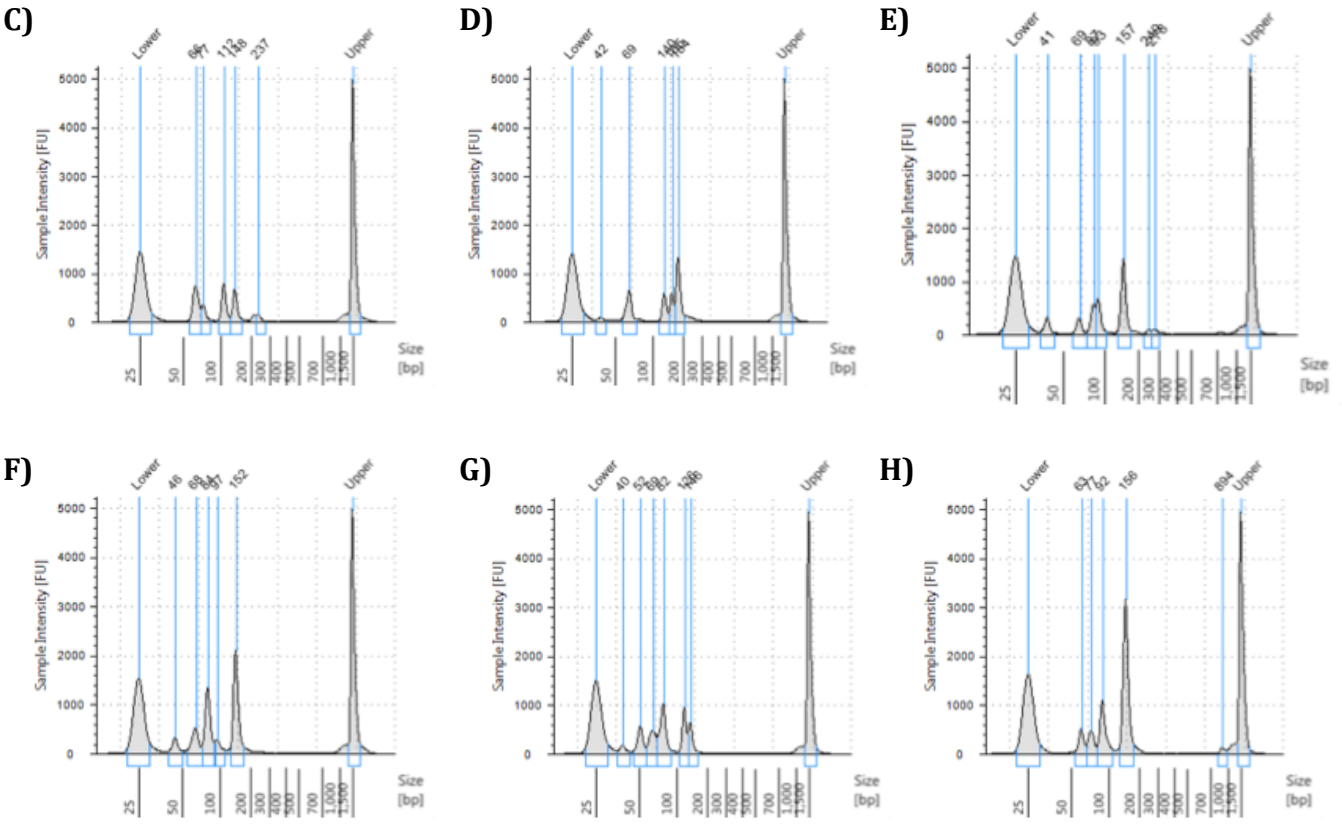


Figure 8: Electrophoresis results of multiplex trials 1-3 **(A)** and 4-6 **(B)** and electropherogram results **(C-H)** on the Agilent© 2200 TapeStation.

Fluorescently labeled primers were used to amplify DNA in singlet as well as multiplex by dye channel, and both resulted in oversaturation. When a large amount of amplified DNA is present, it may overwhelm the instrument's ability to measure the results; this is known as oversaturation. A 1:100 dilution of the DNA sample was made, and then the primer sets were used to amplify the DNA in singlet (Figure 9) as well as multiplexes of the same fluorophore (Figure 10A-E). Each single amplicon peak matched its respective location within the multiplex by a difference of no more than 1 base pair.

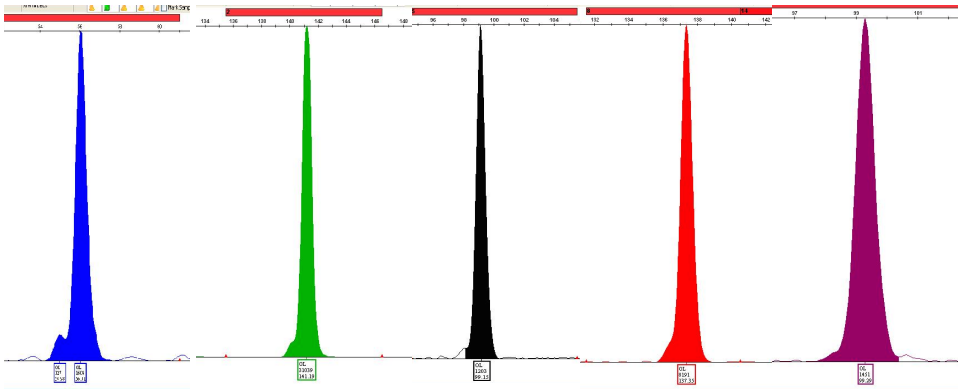
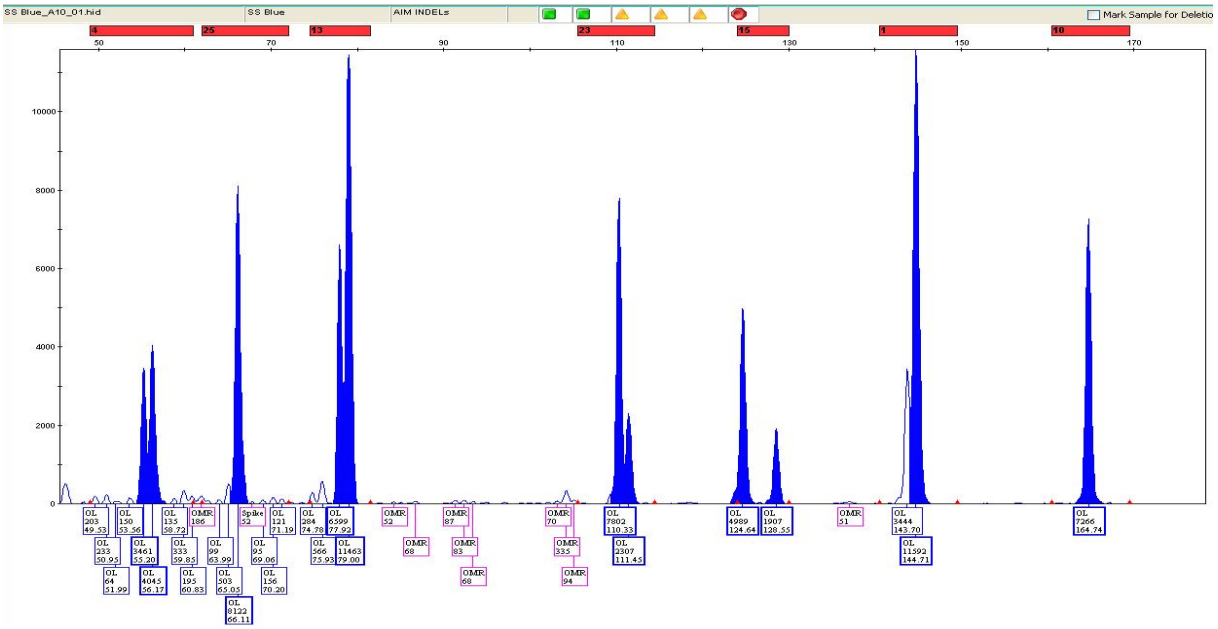
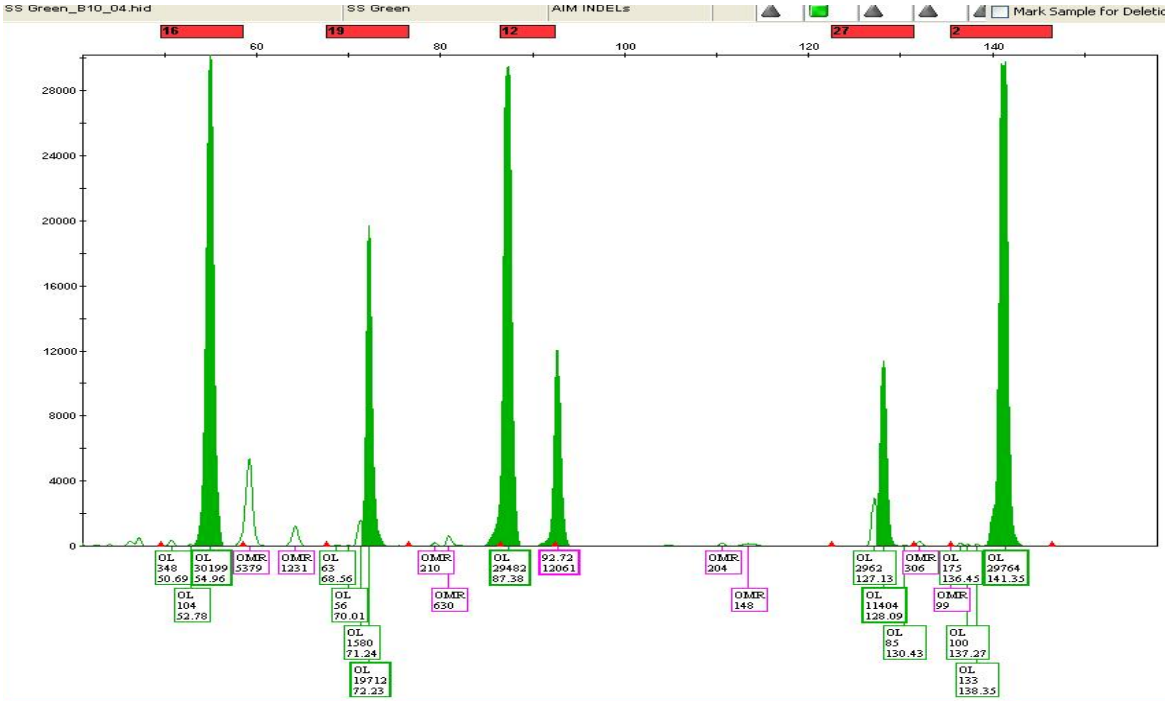


Figure 9: CE results of a single amplicon from each colored fluorophore: Markers 4 (blue), 2 (green), 5 (yellow), 8 (red), and 20 (purple).

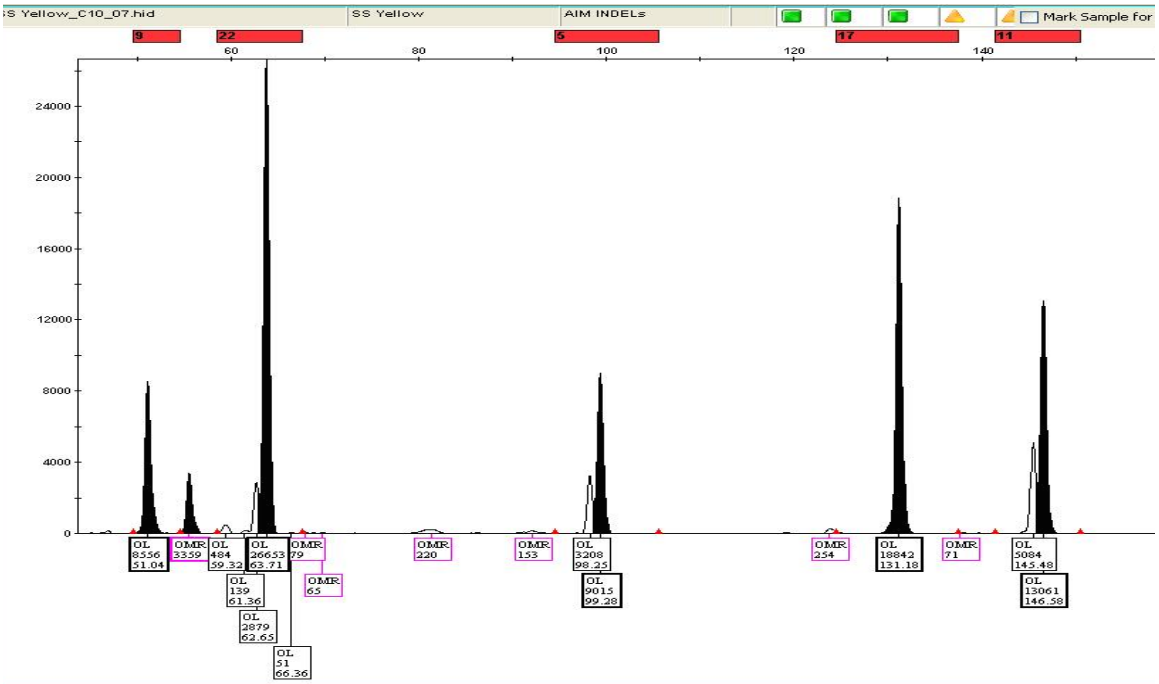
A)



B)


SS Green_B10_04.hid
SS Green
AIM INDELs
Mark Sample for Deletic
16
19
12
27
2
60
80
100
120
140
28000
24000
20000
16000
12000
8000
4000
0
OL 348 50.69
OL 30199 54.96
OMR 5379
OMR 1231
OL 63 68.56
OL 104 52.78
OL 56 70.01
OL 1580 71.24
OL 19712 72.23
OMR 210
OMR 630
OL 29482 87.38
92.72 12061
OMR 204
OMR 148
OL 2962 127.13
OL 11404 128.09
OL 85 130.43
OMR 306
OL 175 136.45
OMR 99
OL 100 137.27
OL 133 138.35
OL 29764 141.35

C)


S Yellow_C10_07.hid
SS Yellow
AIM INDELs
Mark Sample for
9
22
5
17
11
60
80
100
120
140
24000
20000
16000
12000
8000
4000
0
OL 8556 51.04
OMR 3359
OL 484 59.32
OL 139 61.36
OL 2879 62.65
OL 26653 63.71
OL 51 66.36
OMR 79
OMR 65
OMR 220
OMR 153
OL 3208 98.25
OL 9015 99.28
OMR 254
OL 18842 131.18
OMR 71
OL 5084 145.48
OL 13061 146.58

D)

E)

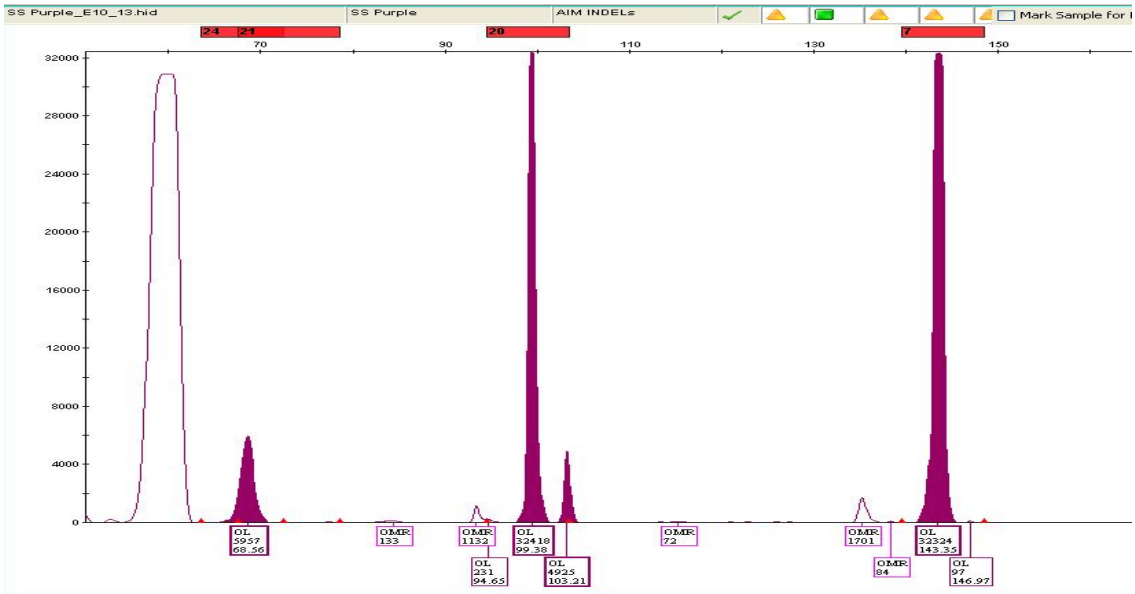


Figure 10: Multiplex CE results of markers 4, 25, 13, 23, 15, 1, and 10 in the blue channel **(A)**, markers 16, 19, 12, 27, and 2 in the green channel **(B)**, markers 9, 22, 5, 17, and 11 in the yellow channel **(C)**, markers 6, 3, 26, 8, and 14 in the red channel **(D)**, and markers 24, 21, 20, and 7 in the purple channel **(E)** with diluted DNA sample.

It is important to note that the initial quantity of DNA used would be ideal for STR typing. However, this amount resulted in oversaturating the camera with fluorescence

from the amplified AIM INDEL markers. As previously stated, this INDEL assay is intended for use with degraded DNA, where such initial quantities will not be possible. Being able to dilute the DNA sample by a factor of 100 and still produce strong results is a demonstration to how sensitive and useful these markers and this multiplex assay are.

After comparing the CE results from the single amplicons to the multiplexes, it was determined that size overlap was occurring between markers 24 and 21 in the purple dye channel. The results from marker 24 alone were not always clear or consistent, as the other markers were, so a future improvement of this assay will be to remove marker 24 (rs5864438). An example of this behavior can be see in Figure 10E, where the first peak on the left, presumably marker 24, is characteristically different than the other markers.

Further modifications necessary for this multiplex may include decreasing the number of PCR cycles from 35 to 32. By doing this, the potential for contamination and oversaturation is reduced, and peak height ratios for heterozygotic markers will be more balanced. Adding 20%, or 18 seconds, to the extension time during PCR may also improve results. In addition to these steps, the concentrations of the following primer pairs can be reduced to attain balanced peak heights as well: primer pairs for markers 16 and 2 in the green channel, 22 in the yellow channel, 6 and 26 in the red channel, and 20 and 7 in the purple channel. After these measures have been taken, all 26 primer pairs might be able to be combined into a single primer mixture for use with the multiplex PCR kit.

The final steps required for this multiplex assay to become operational within forensic laboratories are conducting developmental validation studies of the multiplex assay, followed by population studies with DNA samples with known ancestry.

CHAPTER 4

CONCLUSION

Primer pairs were designed to amplify 30 AIM INDEL markers from the panel developed by Thompson [26] that can distinguish between the three major population groups: Caucasian, African, and East Asian. With the use of publically available online programs, the optimal PCR parameters were achieved such as amplicon size, G-C content, melting temperature, and Gibbs' free energy change, and MPprimer was used to reduce the chance that dimerization would occur between the selected primers. Successful amplification of DNA occurred with 29 of the 30 primer sets, and the amplicons were checked using the Agilent© 2200 TapeStation. Due to overlapping amplicon lengths of 3 markers, only 26 of the 29 primer pairs were arranged into a multiplex of 5 dye channels. Fluorescently labeled primers were used to amplify these 26 AIM INDELs, which were then separated by CE on the Applied Biosystems 3500xL Genetic Analyzer, and analyzed using GeneMapper ID-X v.1.2. In response to oversaturation, a 1:100 dilution of the DNA sample was made and used with the primers. This managed to reduce the peak heights to a readable level.

Additional modifications are required to make this multiplex assay practical for forensic laboratory use. Using fewer PCR cycles, adding extension time during PCR, and lowering the concentrations of some of the primer pairs in the primer mixture for amplification are some of the required next steps. Once peak balance is achieved within each dye channel, all primer pairs will be combined into a single primer mix for simultaneous amplification of all markers.

In conclusion, primers were successfully designed for 26 AIM-INDEL markers that can distinguish among the three major populations: African, Caucasian, and East Asian. Though some fine-tuning is still needed, the use of this assay should greatly benefit forensic casework without the need for supplementary lab equipment. Using this system in addition to STR typing in forensic laboratories will be beneficial in cases with degraded DNA, or no investigative leads. The primers designed for this multiplex of AIM-INDEL markers successfully amplified DNA, and produced the predicted results.

REFERENCES

1. Edwards A, Civitello A, Hammond HA, Caskey CT. DNA typing and genetic mapping with trimeric and tetrameric tandem repeats. Am J Hum Genet 1991 Oct;49(4):746-56.

2. Mullis KB, Faloona FA. Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction. Meth Enzymol 1987;155(0):335-50.

3. Wang Y, Ju J, Carpenter BA, Atherton JM, Sensabaugh GF, Mathies RA. Rapid sizing of short tandem repeat alleles using capillary array electrophoresis and energy-transfer fluorescent primers. Anal Chem 1995;67(7):1197-203.

4. Buel E, Schwartz MB, LaFountain. Capillary electrophoresis STR analysis: comparison to gel-based systems. J Forensic Sci 1998;43(1):164-70.

5. Budowle B, Shea B, Niezgoda S, Chakraborty R. CODIS STR loci data from 41 sample populations. J Forensic Sci 2001;46(3):453-89.

6. Applied Biosystems. 2010. Applied Biosystems 3500/3500xL Genetic Analyzer User Guide.

7. National Research Council. The evaluation of forensic DNA evidence. Washington D.C.: National Academy Press; 1996. Report No.: 2.

8. Hammond HH, Jin L, Zhong Y, Caskey CT, Chakraborty R. Evaluation of 13 short tandem repeat loci for use in personal identification applications. Am J Hum Genet 1994;55(1):175-89.

9. Burger J, Hummel S, Herrmann B, Henke W. DNA preservation: a microsatellite-DNA study on ancient skeletal remains. Electrophoresis 1999;20(8):1722-8.

10. Golenberg EM, Bickel A, Weihs P. Effect of highly fragmented DNA on PCR. Nucleic Acids Res 1996;24(24):5026-33.

11. Kidd KK, Pakstis AJ, Speed WC, Grigorenko EL, Kajuna SLB, Karoma NJ, et al. Developing a SNP panel for forensic identification of individuals. Forensic Sci Int 2006 Dec;164(1):20-32.

12. Pakstis AJ, Speed WC, Kidd JR, Kidd KK. Candidate SNPs for a universal individual identification panel. Hum Genet 2007;121:305-17.

13. Pereira R, Phillips C, Alves C, Amorim A, Carracedo Á, Gusmão L. A new multiplex for human identification using insertion/deletion polymorphisms. Electrophoresis. 2009;30(21):3682-90.

14. LaRue BL, Lagacé R, Chang C, Holt A, Hennessy L, Ge J, et al. Characterization of 114 insertion/deletion (INDEL) polymorphisms, and selection for a global INDEL panel for human identification. Leg Med 2014 Jan;16(1):26-32.

15. Fondevila M, Phillips C, Santos C, Pereira R, Gusmao L, Carracedo A, Butler JM, Lareu MV, Vallone PM. Forensic performance of two insertion-deletion marker assays. Int J Legal Med 2012;126:725-37.

16. Oka K, Asari M, Omura T, Yoshida M, Maseda C, Yajima D, et al. Genotyping of 38 insertion/deletion polymorphisms for human identification using universal fluorescent PCR. Mol Cell Probes 2014 Feb;28(1):13-8.

17. Wei Y, Qin C, Dong H, Jia J, Li C. A validation study of a multiplex INDEL assay for forensic use in four Chinese populations. Forensic Sci Int Genet 2014 Mar;9(0):e22-5.

18. Seong KM, Park JH, Hyun YS, Kang PW, Choi DH, Han MS, et al. Population genetics of insertion–deletion polymorphisms in South Koreans using investigator DIPplex kit. Forensic Sci Int Genet 2014 Jan;8(1):80-3.

19. Galanter JM, Fernandez-Lopez J, Gignoux CR, Barnholtz-Sloan J, Fernandez-Rozadilla C, Via M, et al. Development of a panel of genome-wide ancestry informative markers to study admixture throughout the Americas. PLoS Genetics 2012 Mar;8(3):1-16.

20. Jia J, Wei Y, Qin C, Hu L, Wan L, Li C. Developing a novel panel of genome-wide ancestry informative markers for bio-geographical ancestry estimates. For Sci Int Genet 2014 Jan;8(1):187-94.

21. Nievergelt CM, Maihofer AX, Shekhtman T, Libiger O, Wang X, Kidd KK, et al. Inference of human continental origin and admixture proportions using a highly discriminative ancestry informative 41-SNP panel. Invest Genet 2013 Aug;4(1):1-16.

22. Phillips C, Fondevila M, Vallone PM, Carla S, Freire-Aradas A, Butler JM, Lareu MV, Carrecedo A. Characterization of U.S. population samples using a 34plex ancestry informative SNP multiplex. Forensic Sci Int Genet 2011;3:e182-3.

23. Kidd KK, Speed WC, Pakstis AJ, Furtado MR, Fang R, Madbouly A, et al. Progress toward an efficient panel of SNPs for ancestry inference. Forensic Sci Int Genet 2014 May;10(0):23-32.

24. Frudakis TN. Molecular photofitting: Predicting ancestry and phenotype using DNA. Burlington, MA: Elsevier; 2009.

25. Kidd KK, Speed WC, Pakstis AJ, Furtado MR, Fang R, Madbouly A, et al. Progress toward an efficient panel of SNPs for ancestry inference. Forensic Sci Int Genet 2014 May;10(0):23-32.

26. Thompson, Lindsey M. April 2015. *Selection of an Ancestry-Informative Marker (AIM) Panel of INDELs* (Master's Thesis). Retrieved from Gibson D. Lewis Library, University of North Texas Health Science Center.

27. Edwards MC & Gibbs RA 1994 Multiplex PCR: advantages, development, and applications. *PCR Methods and Applications* **3** S65–S75.

28. Ye J, Coulouris G, Zaretskaya I, Cutcutache I, Rozen S, Madden TL: Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. BMC Bioinformatics 2012, 13:134.

29. Shen Z, Qu W, Wang W, Lu Y, Wu Y, Li Z, Hang X, Wang X, Zhao D, Zhang C. MPprimer: a program for reliable multiplex PCR primer design. BMC Bioinformatics 2010;11:143.

30. Agilent Technologies. 2013. Agilent 2200 TapeStation User Manual.

31. QIAGEN. 2011. QIAGEN Multiplex PCR *Plus* Kit User Manual.