Thompson, Lindsey M., Selection of an Ancestry-Informative Marker (AIM) Panel of INDELs. Master of Science (Forensic Genetics), April 2015, pp. 30, 5 tables, 9 illustrations, 37 references.

Short Tandem Repeat (STR) loci are commonly used for forensic identification purposes. Most commercially available STR kits yield amplified fragments with lengths between 100 and 600 base pairs (bp). However, the genomic DNA of forensic samples can be highly degraded, yielding incomplete STR profiles. Small insertion/deletion polymorphisms (INDELs) in the intergenic regions of the genome, are viable options for typing degraded samples. Furthermore, when there are no suspects for comparison, ancestry-informative markers (AIMs) are useful for developing investigative leads. This project tested the hypothesis that using publicly available genome data, a panel of AIM-INDELs can be selected for the purposes of distinguishing the Caucasian, East Asian and African population groups. To test this hypothesis, the data from the 1000 Genomes Project were mined to select a panel of AIMs that can be used for the purposes of providing ancestry information as an investigative lead to law enforcement.

SELECTION OF 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

Lindsey M. Thompson, B.S.

Fort Worth, TX

April 2015

ACKNOWLEDGMENTS

I would like to express my appreciation to Dr. Bobby LaRue for his guidance and support. His enthusiasm for research and education has inspired me to produce a project of which I am proud. I would also like to thank my committee members, Drs. Lisa Hodge, Michael Oglesby, and Ranajit Chakraborty for their valuable comments and advice which contributed to the quality of my work. Thanks also to Xiangpei Zeng M.D. and Jonathan King, M.S. for their contributions to my work in the laboratory. I would like to thank the girls of the FGEN class of 2015, especially Jenny King. Without them, I would not have made it through graduate school. To Kelly Sage and Sarah Sturm, without their help in the laboratory, I would not have been able to get everything done. Finally, many thanks to my parents and family for their support of my education and enthusiasm for my success.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

CHAPTER 1
INTRODUCTION

Short Tandem Repeat (STR) analysis has been the generally accepted method of DNA analysis in the forensic community for several decades. These markers have been described to be highly polymorphic, with respect to variations in the number of repeats between individuals (1). Analysis of STRs is accomplished through amplification of the repeat region by a process known as the Polymerase Chain Reaction (PCR) (1, 2) followed by capillary electrophoresis (CE) fragment separation and genotype analysis *en silica* (3, 4) . Population databases have been developed that allow analysts to provide statistical support for their genetic conclusions (5).

The process of amplification via PCR allows an analyst to generate hundreds of thousands of copies of a specific DNA sequence. Primers, or short oligonucleotide sequences, hybridize to a specific location in the DNA and an enzyme copies the DNA sequence by adding complementary nucleotides. The result is an exponential increase in the amount of the specific target sequence (2). The primers designed for STR analysis produce amplicons approximately 100-600 bp in length. The amplicons can vary in length within a specific STR region based on the number of repeats observed in that sample. Each primer set has a fluorescently-labelled dye that becomes incorporated in the amplicons and is important for downstream analysis (1).

Once amplified, the PCR amplicons are analyzed by CE. DNA molecules are electrokinetically injected into a capillary through which a voltage is applied causing the

negatively charged DNA fragments to move through a liquid polymer. As they migrate, the fragments are separated by size, as smaller fragments migrate more quickly than larger ones (3, 6). When the amplicons move through the capillary, they reach a laser that excites the fluorescent dye and a camera captures the wavelength and corresponding time of passage. The amplicons can then be sized based on the internal size standard (6). The wavelengths of the dyes are important because they allow for the analysis of several PCR amplicons in tandem. Using a software system like Gene Mapper ID-X, the peaks can be visualized (Figure 1).

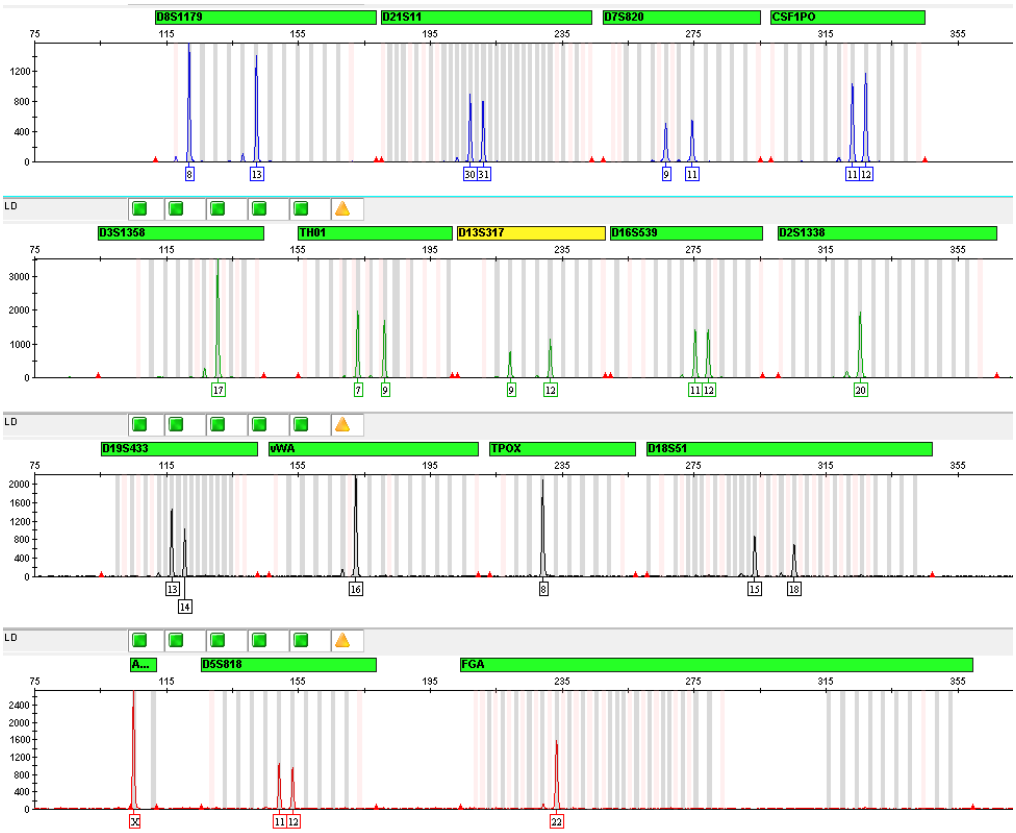


Figure 1. Image from GeneMapper ID-X v1.2 of an STR profile amplified with Identifiler Plus on a 3500xL Genetic Analyzer. The various colors represent the different fluorescent dyes.

Finally, a Random Match Probability (RMP) is calculated from the profile generated, which is a statistic that indicates how often the profile is expected to be observed. Specifically, it is the probability that a randomly selected individual in a population will have the same genotype as the evidentiary profile. The RMP is calculated for each locus individually, using allele frequencies obtained from a population database, adjusting for population substructure effect, if needed. The probabilities are then multiplied across loci to achieve a RMP for the entire DNA profile (7, 8).

One issue frequently encountered with this marker system is caused by sample degradation. With time and exposure to environmental elements, DNA in a biological sample will begin to degrade. These samples present a challenge to forensic analysts because phenomena such as allelic drop-out occur which leads to incomplete STR profiles (9, 10). When events such as drop-out occur, the RMP for the profile may not be sufficiently small to be individualizing. Figure 2 shows the profile of a Civil War bone typed for STR markers. Due to degradation, only three loci were successfully typed.

There have been several marker systems that have been proposed to handle degraded samples. One such system is bi-allelic Single Nucleotide Polymorphisms (SNPs), which examine single base-pair changes in a nucleotide sequence (11, 12). The advantages of this system are that amplicons can be designed as small as allowed by the constraints of PCR primers. Additionally, as they are not repeat sequences, polymerase slippage artifacts such as stutter are avoided. Stutter peaks, are reproducible PCR products that make STR profiles slightly more difficult to analyze, especially when looking at mixtures. However advantageous, SNPs require the detection of a single base substitution which necessitates cumbersome chemistries or instrumentation that is not readily available in most forensic casework laboratories.

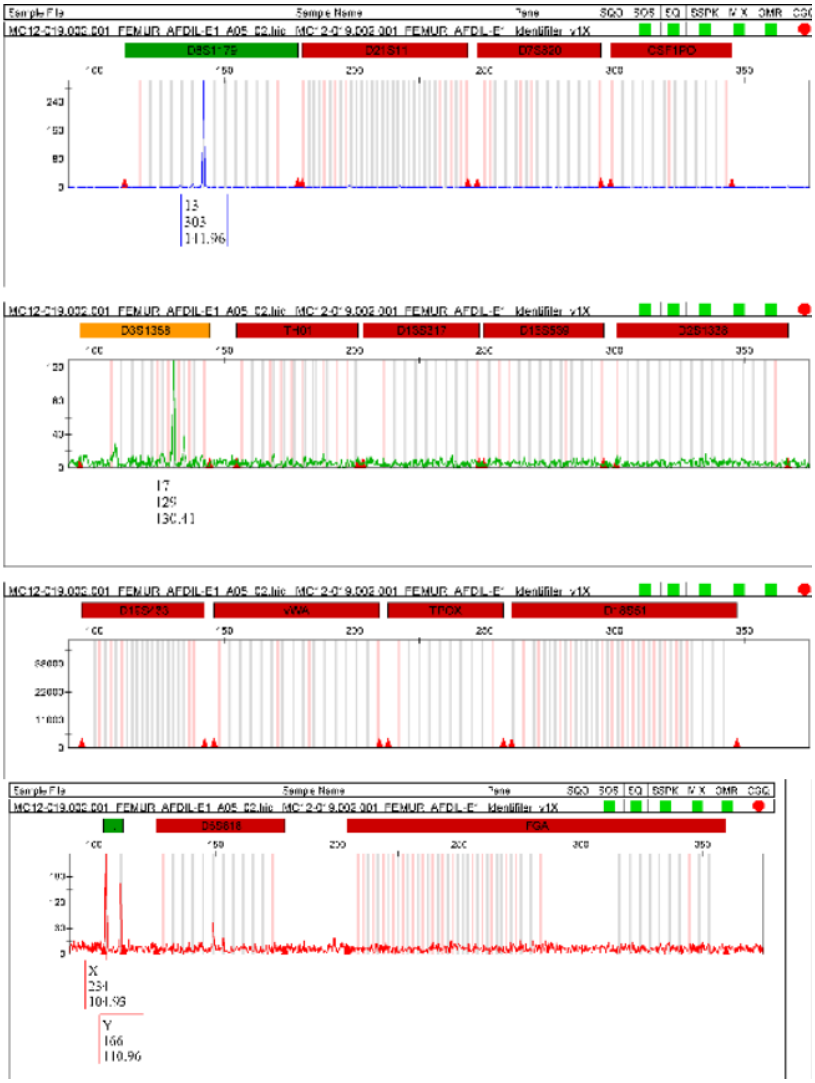


Figure 2. Civil War bones typed with STR markers.

A similar type of marker for typing degraded samples is insertion/deletion polymorphisms (INDELs). INDELs are insertions or deletions of a DNA sequence, as determined by comparison to a known consensus sequence. These polymorphisms arise when a genetic sequence is either inserted or deleted from the genome (13). Several INDEL panels have been developed for human identification purposes (14-18). INDELs provide a number of

benefits when dealing with forensic samples. Similar to SNPs, amplicons can be designed to be very small (~55bp as limited by primer design), and as a non-repeating polymorphism, slippage artifacts are non-existent. Unlike SNPs, INDEL alleles have amplification products that differ in length which facilitates genotyping based on fragment length electrophoresis (13). As this is the present method for determining STR genotypes, the laboratory infrastructure required to genotype INDELs is pre-existing in forensic laboratories. These types of markers perform very well with degraded samples (15). Figure 3 shows the same Civil War bone sample from Figure 2 amplified with an INDEL system.

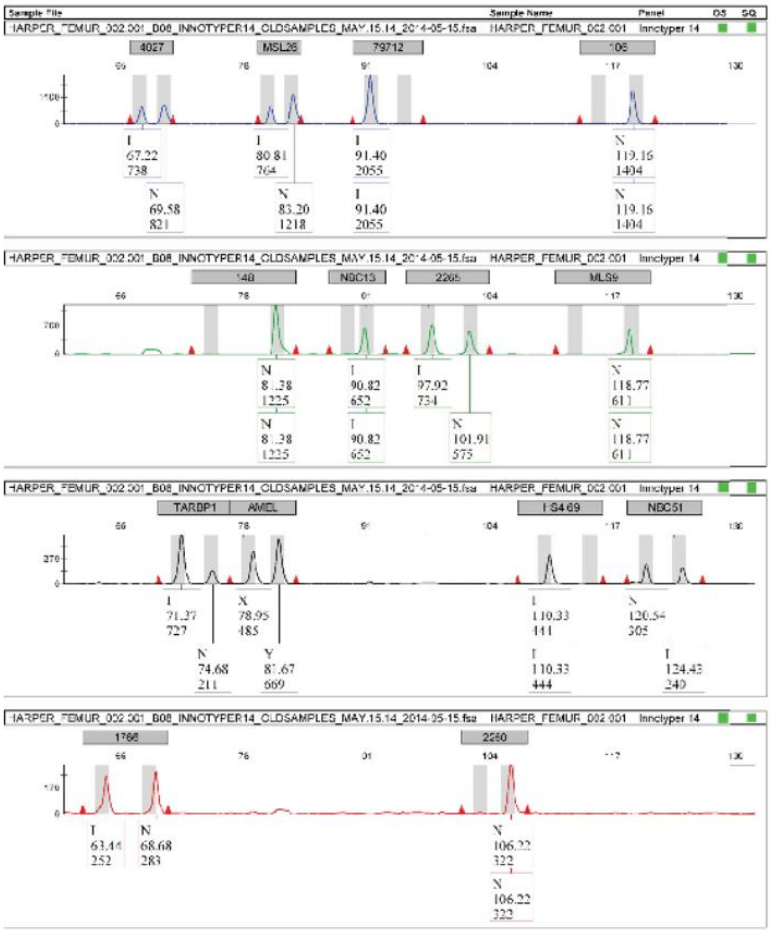


Figure 3. Civil War bone sample amplified with an INNUL marker system.

INDELs have been proven to be successful for human identification (HID) purposes. LaRue *et al*. 2014 presents the validation of a panel of 38 INDELs with a RMP of at least 10^{-16} (14). Seong *et al*. 2014 describes the population genetics of a South Korean population the Investigator DIPplex (Qiagen) with a RMP of 2.84 x 10^{-11} (18). An example a degraded sample amplified using an INDEL system is shown in Figure 4 (14).

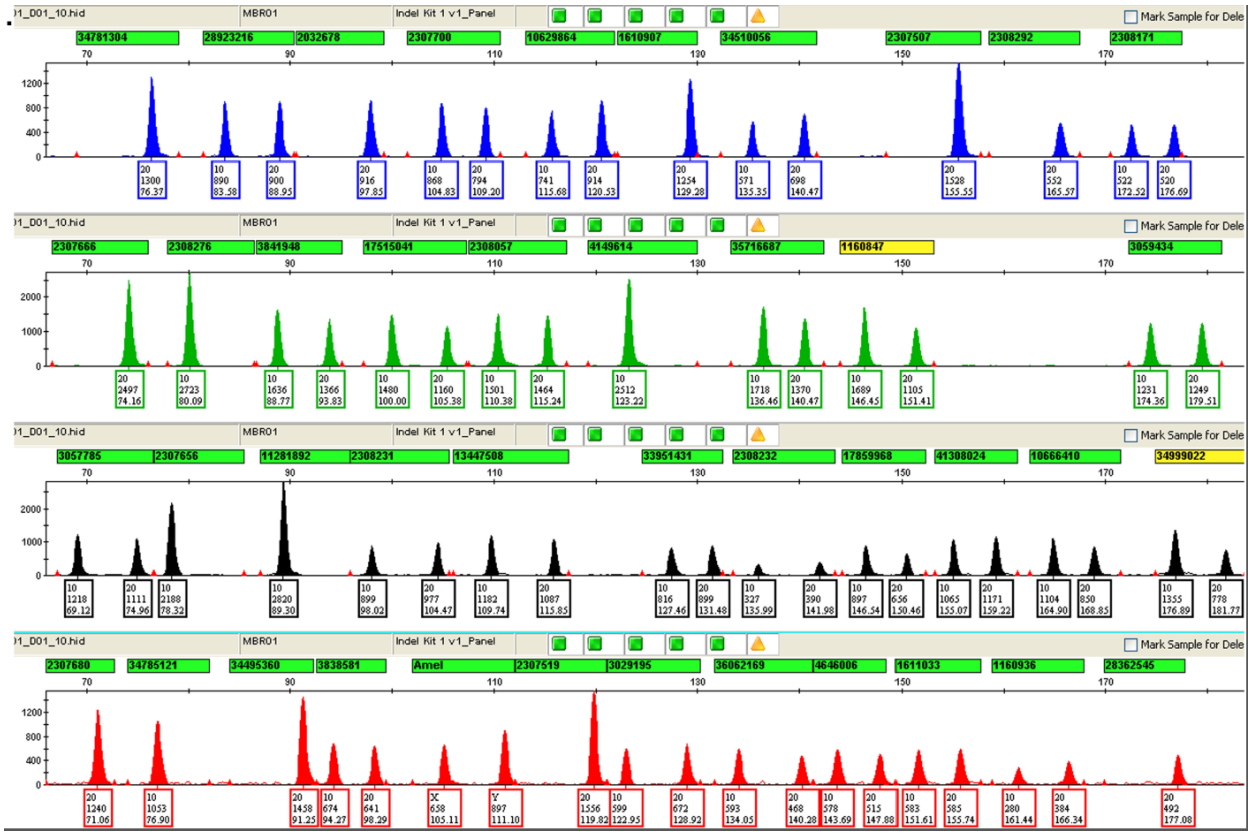


Figure 4. DNA amplified using an INDEL system and analyzed by capillary electrophoresis (14).

HID markers are useful in cases where a suspect has been identified because the profile obtained from the evidentiary sample can be compared to the profile obtained from the suspect's reference sample. However, in cases where there is no suspect for DNA typing, an HID profile cannot give investigators any descriptive information about the perpetrator such as ancestral

origin. In such cases, Ancestry-Informative Markers (AIMs) have been a topic of discussion as a method to provide additional information to investigators for samples of unknown origin, especially regarding SNPs (19-23). Several AIM-INDEL panels are currently in development (24, 25). Pereira *et al.* 2012 published a panel of 46 AIM-INDELs that can distinguish between African, East Asian, European, and Native American samples. Additionally, a set of 48 AIM-INDELs was used to determine the parental admixture proportions of Brazilian samples. However, to date, a panel of AIM-INDELs has not been widely adopted for use in forensic laboratories.

AIM-SNPs have been used to generate leads that have been critical in the resolution of difficult forensic cases, including a serial murder case in Louisiana (26). Kidd *et al*. 2014 presents a small, robust panel of AIM-SNPs for use in forensics for the purposes of determining the ancestral origin of unknown samples. Publicly available population databases were used to identify SNPs with high pairwise allele frequency differences. Additionally, SNPs were further refined to include only those with large pairwise F_{ST} values, a measure of population substructure (27). F_{ST} is a value that ranges from zero to one. An F_{ST} of zero means there is no substructure and an F_{ST} of one means the population is highly substructured. For ancestry markers, the desired F_{ST} should be high, indicating the overall population is substructured into population groups. For example, the final Kidd panel included 55 SNPs with an average F_{ST} value of 0.436 (28). The markers were balanced to include SNPs capable of distinguishing between the most populations.

Since SNPs and INDELs are both bi-allelic marker systems, their selection criteria are relatively similar and can be translated into the selection criteria for a panel of AIM-INDELs. This project tested the hypothesis that publically available data can be used to select a panel of

INDELs that can distinguish between the major global populations on degraded samples. Genome data from the 1000 Genomes Project, a collaborative effort to sequence over 1000 human genomes, was mined to select a robust panel of markers that can not only handle degraded samples, but also give investigators additional ancestry information about the perpetrator that may be critical in the investigation of many forensic cases.

CHAPTER 2

RESEARCH DESIGN AND METHODOLOGY

Samples

Variant call files for chromosomes 1 through 22 were downloaded from the 1000 Genomes Project website (http://www.1000genomes.org/data). These files contain the autosomal genome data for 3500 individuals. When possible, the populations were binned into their associated major global population group (Table 1). Of the 3500 individuals, 550 individuals (Caucasian, N=244; African, N=156; and East Asian, N=150) were chosen comprising the training set used for marker selection for differentiating Caucasian, African, and East Asian ancestry.

Table 1. Populations from 1000 Genomes Project binned into three major global population.

African	Caucasian	East Asian
Yoruba in Ibadah, Nigeria	British in England and Scottland	Southern Han Chinese, China
African ancestry in southwest U.S.	Finnish in Finland	Han Chinese in Bejing, China
Luhya in Webuye, Kenya	Utah resident with Western Europe ancestry	Japanese in Tokyo, Japan
	Toscani in Italy	

Marker selection

Using the Linux-based software program, VCFtools, chromosomal data was filtered to include only INDEL markers (29). Pairwise population substructure, F_{ST}, was calculated for each INDEL. The output file was sorted in Excel® and a new text file containing only the positions with F_{ST} greater than 0.5 in at least one pairwise comparison was created. From this new file, population-specific allele frequencies were calculated. The markers were then manually sorted by length and markers of length 3-6 base pairs were considered for selection. This is so that they are large enough to resolve the alleles by CE, but small enough to conserve real estate in the CE system. Next, to verify that these markers are true INDELs, and part of a repeat region, the UCSC Genome Browser was used to eliminate markers that showed a presence of repeat or proto-repeat sequences near the INDEL by enabling the Repeat Masker algorithm. From this reduced list of markers, 20 per population group were selected based on high allele frequency divergence, or delta value, and genetic distance. Markers on the same chromosome were selected to be more than 1 Mb from its nearest neighbor in order to minimize potential for selection linked loci.

Statistical Analysis

Using VCFtools, genotype data for the 550 training set individuals were pulled out of the variant call files. The panel of 60 AIMs was evaluated by Principal Component Analysis using the software program Past3 to determine if the populations would cluster separately. Additionally, known populations were added to the analysis to determine if they would cluster with the expected population group (30). Next, the 60 AIMs were evaluated for Hardy-

Weinberg Equilibrium (HWE) and linkage disequilibrium (LD). Using the software package, Genetic Data Analysis (GDA), exact tests for HWE and LD were performed. The AIMs panel was then evaluated for ancestry admixture using the program STRUCTURE v.2.3.4.

Table 2. Linux commands used for VCFtools.

	Command	Input File Format	Output Format
Pairwise F_{ST} calculations	vcftools --gzvcf ~/*chromosome1*.vcf.gz --keep-only-indels --weir-fst-pop ~/*pop1*.txt --weir-fst-pop ~/*pop2*.txt --out ~/*outfile*	*chromosome1*.vcf.gz- variant call file downloaded from 1000 Genomes Project website *pop1*.txt/ *pop2*.txt- text file containing 2 columns: (1) sample name (2) population	*outfile*.weir.fst
Allele frequency calculations	vcftools --gzvcf ~/*chromosome1*.vcf.gz --positions ~/*fst*.txt --keep ~/*pop1*.txt --freq --out ~/*outfile*	*fst*.txt- text file containing the positions of INDEL markers with Fst >0.5 *pop1*.txt- text file containing a single column with one sample per line	*outfile*.frq
Genotype Data	vcftools --gzvcf ~/*chromosome1*.vcf.gz --keep ~/pop1.txt --positions ~/*markers1* --recode --recode-INFO-all --out ~/*outfile*	*pop1*.txt- text file containing a single column with one sample name per line *markers1*- gedit file containing 2 columns: (1) chromosome number (2) position	*outfile*.recode.vcf

Italicized words indicate file names

Additional Populations

Southwest Hispanic (SWH; N=243) and Southwest Asian (SWA; N=489) samples were compiled from the 1000 Genomes Project data. The genotype data for the 59 AIMs for these individuals were retrieved using VCFtools and added to the PCA of the original training set. Using Past3, these two population groups were also compared pairwise in PCA.

Table 3. Populations in 1000 Genomes Project binned into additional population groups.

Southwest Hispanic	Southwest Asian
Colombian in Medellin, Colombia	Gujarati Indian in Houtson, TX
Peruvian in Lima, Peru	Bengali in Bangladesh
Mexican Ancestry in Los Angeles, California	Sri Lankan Tamil in the UK
	Indian Telugu in the UK

CHAPTER 3

RESULTS

Marker Selection

All INDEL markers identified in VCFtools were filtered based on pairwise F_{ST} comparisons of the three major global population groups, Caucasian, African, and East Asian. Those with at least one pairwise F_{ST} value greater than 0.5 were included. These markers were subsequently filtered to include only INDELs of length three to six base pairs. Next, the markers with allele frequency divergence in one of the three population groups were selected. A summary of the number of INDELs that meet these criteria can be seen in Table 3.

From the remaining INDELs, 60 markers, 20 for each population group, were selected as potential AIMs. These were chosen based on physical distance and allele frequency divergence. The allele frequency difference, or delta (δ) value was calculated between each population group. Markers with high delta value (> 0.5) were included in the panel as potential AIMs. Additionally, all syntenic markers, markers on the same chromosome, were selected to have a physical distance of at least 1Mb from its nearest neighbor.

Table 4. Summary of AIM-INDELs identified using VCFtools.

Chromosome	Caucasian	East Asian	African	Total
1	7	35	58	**100**
2	11	40	107	**158**
3	7	14	72	**93**
4	5	32	117	**154**
5	7	46	56	**109**
6	8	22	38	**68**
7	7	23	38	**68**
8	6	5	21	**32**
9	7	11	43	**61**
10	9	19	45	**73**
11	3	12	44	**59**
12	0	8	15	**23**
13	0	11	7	**18**
14	1	2	4	**7**
15	13	15	24	**52**
16	1	6	8	**15**
17	2	5	18	**25**
18	1	6	10	**17**
19	0	6	10	**16**
20	1	5	8	**14**
21	0	3	3	**6**
22	1	3	10	**14**
Total	**97**	**329**	**756**	**1182**

Statistical Analysis

To test whether these marker would cluster the population groups correctly, Principal Component Analysis (PCA) was performed using the software program Past3 (Figure 5A). Samples from each population group were labelled with a different color to show the distinct clusters among the training set. The first two principal components (PC1 and PC2) explained 40.3% of the variation. To further test the markers capacity to separate the major global population groups, populations from the 1000 Genomes Project that were not in the original training set were added to the PCA (Figure 5B-D). A population from Gambia in the Western

Division, an Iberian population from Spain, and a Kinh population in Ho Chi Minh City, Vietnam were selected to represent the African, Caucasian, and East Asian population groups, respectively.

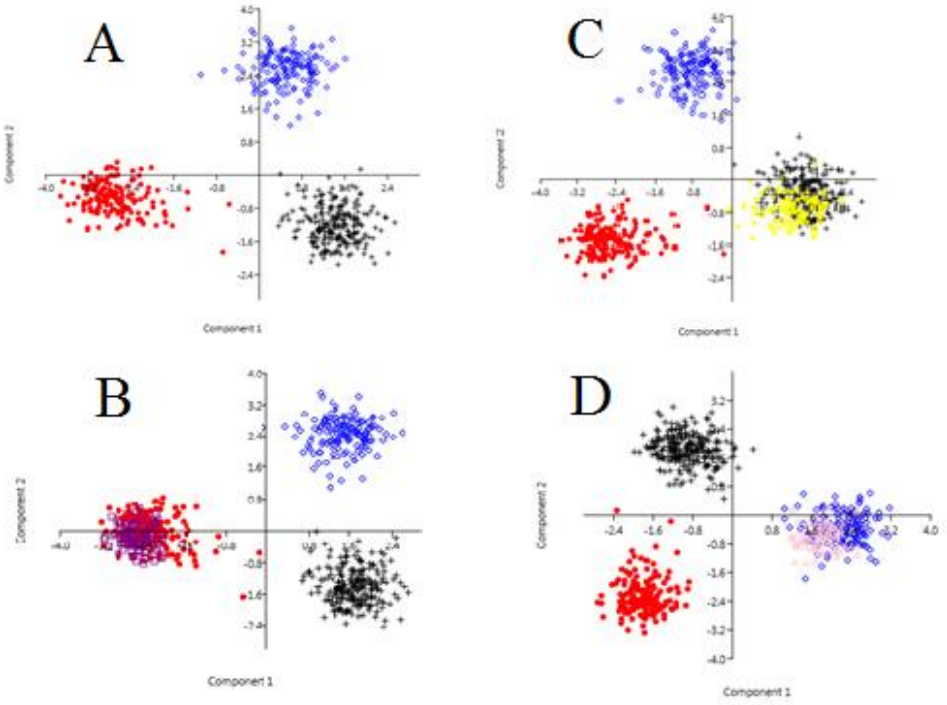


Figure 5. Principal Component Analysis (PCA) of 60 Ancestry Informative Markers (AIMs) using the software package, Past3. A) Original training set of 550 Individuals; Caucasian (Black Plus), East Asian (Blue Diamond) and African (Red Dot). B) Original training set with additional samples from Gambia in the Western Division (Purple Square). C) Original training set with additional Iberian samples from Spain (Yellow Star). D) Original training set with additional samples from Kinh in Ho Chi Minh City, Vietnam (Pink Triangle).

Additional statistical analysis was performed on the 60 markers using the software package Genetic Data Analysis (GDA). Exact tests for Hardy-Weinberg Equilibrium (HWE) were performed on the set of 60 AIMs. Of the 60 markers, five in the African population group,

seven in the Caucasian population group, and three in the East Asian population group showed departure from HWE at a significance level of 0.05. After Bonferroni correction for multiple comparisons (α=0.05/60), only one marker showed significant departure from HWE in all three population groups. INDEL rs78981054 showed a p-value of less than 8.33×10^{-4} in all three population groups.

The remaining 59 markers were evaluated for Linkage Disequilibrium (LD) to determine if there was an observable pattern of inheritance between any of the marker combinations. Of the 1710 combinations per population group, 222 in the African population group, 145 in the Caucasian population group, and 130 combinations in the East Asian population group showed LD. After Bonferroni correction for multiple comparisons (α= 0.05/1710), only three in the African population group, two in the Caucasian population group, and one combination in the East Asian population group showed significant LD. The marker combinations that showed significant LD in the African population group included, rs59009450/ rs67344973, rs67344973/rs10651200, and rs113043680/rs10528149. In the Caucasian population group, they were rs59009450/rs113501732 and rs59009450/rs35779249. Finally, in the East Asian population group, the markers that showed significant LD were rs141933116/ rs74515961. The 59 AIMs are described in Table 4.

The panel of 59 AIMs was analyzed for ancestry admixture in the software program STRUCTURE v.2.3.4 (31-34). After 20 simulations for 10 values of K, the *ad hoc* statistic, ΔK, was calculated (35). Figure 6A describes the distribution of ΔK for K values 1 through 10. ΔK is maximized when K=3. Figure 6 describes the STRUCTURE output for each individual (6C) and the population groups as a whole (6B).

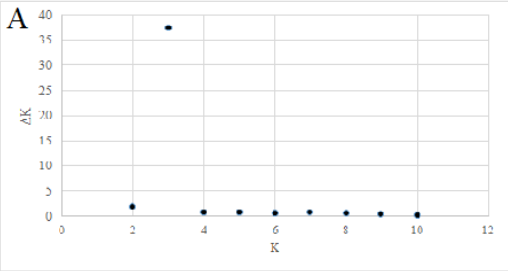
Table 5. Descriptive Statistics of the 59 Ancestry-Informative Markers. (He and Ho refer to expected and observed heterozygosity, respectively).

CAUCASIAN												
				Frequency of Insertion			Delta		Pairwise Fst		Heterozygosity	
rs#	Chrom.	Position	Sequence	African	Caucasian	East Asian	v.AFR	v.EAS	v.AFR	v.EAS	He	Ho
rs139570718	1	214397853	-/CCCAG	0.0352564	0.727459	0.223333	0.6922026	0.504126	0.640101	0.400736	0.477808	0.296364
rs3831920	1	1227664	-/TGAG	0.375	0.913934	0.293333	0.538934	0.620601	0.508888	0.600295	0.483578	0.289091
rs70958016	2	13725708	-/AGTTT	0.865385	0.278689	0.65	0.586696	0.371311	0.504749	0.244201	0.496152	0.372727
rs67934853	2	74943887	-/TAAC	0.923077	0.258197	0.81	0.66488	0.551803	0.603647	0.460751	0.481514	0.3
rs139220746	2	200205694	-/TATC	0.826923	0.227459	0.673333	0.599464	0.445874	0.52312	0.338786	0.499725	0.365455
rs140498743	3	139232513	-/TGTC	0.842949	0.360656	0.95	0.482293	0.589344	0.37517	0.517759	0.450366	0.287273
rs5864438	4	178146869	-/CTAT	0.839744	0.192623	0.803333	0.647121	0.61071	0.585954	0.542572	0.4968	0.303636
rs149676649	5	28495386	-/GATT	0.349359	0.79918	0.106667	0.449821	0.692513	0.350045	0.637831	0.499858	0.343636
rs57237250	6	110263002	-/GAGT	0.826923	0.260246	0.903333	0.566677	0.643087	0.479728	0.574514	0.481866	0.312727
rs1160871	7	28168745	-/TCTT	0.217949	0.788934	0.0233333	0.570985	0.7656007	0.491182	0.72318	0.487054	0.272727
rs55855642	8	122272251	-/ATAGAG	0.855769	0.381148	0.996667	0.474621	0.615519	0.368124	0.561435	0.432949	0.287273
rs67538813	9	30471814	-/CAGA	0.958333	0.383197	0.696667	0.575136	0.31347	0.507485	0.17589	0.465671	0.354545
rs10651200	10	69800907	-/TAACAA	0.939103	0.334016	0.83	0.605087	0.495984	0.525682	0.389713	0.460708	0.336364
rs71991275	10	28470438	-/AATA	0.74359	0.348361	0.996667	0.395229	0.648306	0.266669	0.596017	0.462733	0.329091
rs11576045	12	111799524	-/TGT	0.762821	0.235656	0.936667	0.527165	0.701011	0.433617	0.646023	0.488782	0.298182
rs35779249	13	43964476	-/TAA	0.961538	0.297131	0.82	0.664407	0.522869	0.607878	0.422731	0.467564	0.289091
rs370096890	14	65368820	-/CTTGA	0.910256	0.209016	0.63	0.70124	0.420984	0.648534	0.314874	0.499421	0.307273
rs138439822	15	35537968	-/TAACTC	0.858974	0.270492	0.713333	0.588482	0.442841	0.506957	0.327046	0.493679	0.345455
rs10528149	16	69989686	-/TGAT	0.0769231	0.721311	0.36	0.6443879	0.361311	0.578944	0.233118	0.493248	0.323636
rs55885844	17	79605107	-/ATTAA	0.304487	0.657787	0.00333333	0.3533	0.6544537	0.219042	0.602563	0.47119	0.321818

EAST ASIAN												
				Frequency of Insertion			Delta		Pairwise Fst		Heterozygosity	
rs#	Chrom.	Position	Sequence	African	Caucasian	East Asian	v.AFR	v.CAU	v.AFR	v.CAU	He	Ho
rs141933116	1	8189066	-/AAGT	0.701923	0.956967	0.39	0.311923	0.566967	0.176461	0.579729	0.394559	0.310909
rs5839799	2	241417278	-/GTCT	0.88141	0.694672	0.286667	0.594743	0.408005	0.533799	0.282733	0.463231	0.352727
rs72375069	3	27427821	-/AATT	0.980769	0.657787	0.256667	0.724102	0.40112	0.7167	0.273236	0.461219	0.330909
rs33915414	4	21762063	-/CATGTT	0.0801282	0.385246	0.803333	0.7232048	0.418087	0.693429	0.295226	0.485208	0.334545
rs1610951	5	108999835	-/TTGG	0.971154	0.868852	0.336667	0.634487	0.532185	0.61792	0.475083	0.372597	0.232727
rs367799178	6	21621169	-/TTAA	0.285256	0.284836	0.89	0.604744	0.605164	0.544839	0.527296	0.49545	0.38
rs151280400	7	125249166	-/AATC	0.910256	0.659836	0.35	0.560256	0.309836	0.504593	0.17317	0.457571	0.358182
rs10581451	8	73854660	-/TGAG	0.894231	0.965164	0.18	0.714231	0.785164	0.677952	0.799592	0.39372	0.163636
rs150560593	9	95478810	-/TGCA	0.865385	0.739754	0.283333	0.582052	0.456421	0.514276	0.344585	0.454866	0.334545
rs67205569	10	94941566	-/TTGAC	0.971154	0.885246	0.1333333	0.8378207	0.7519127	0.831329	0.724881	0.416701	0.165455
rs143873637	11	97893598	-/TTGA	0.823718	0.866803	0.243333	0.580385	0.62347	0.504557	0.57738	0.432279	0.274545
rs66693708	12	77398405	-/TAAG	0.974359	0.805328	0.326667	0.647692	0.478661	0.633852	0.386204	0.40115	0.270909
rs10587399	13	37776954	-/TACT	0.887821	0.717213	0.243333	0.644488	0.47388	0.594295	0.362933	0.463231	0.341818
rs141122561	14	49242955	-/TTAGT	0.996795	0.963115	0.37	0.626795	0.593115	0.627839	0.612742	0.30695	0.174545
rs200047010	15	102264144	-/GCAGG	0.714744	0.702869	0.13	0.584744	0.572869	0.515882	0.486211	0.49545	0.336364
rs10549914	17	5328978	-/TTTA	0.852564	0.719262	0.18	0.672564	0.539262	0.622449	0.443694	0.476233	0.332727
rs74515961	18	52716306	-/ATGTC	0.983974	0.786885	0.376667	0.607307	0.410218	0.598112	0.301386	0.39372	0.269091
rs10668859	19	266759	-/GAAAG	0.86859	0.637295	0.14	0.72859	0.497295	0.692713	0.393972	0.491395	0.341818
rs11474791	20	19234875	-/GGACT	0.221154	0.108607	0.79	0.568846	0.681393	0.487299	0.652616	0.440101	0.278182
rs3074939	21	43422429	-/CAGT	0.205128	0.364754	0.836667	0.631539	0.471913	0.569109	0.361085	0.49508	0.358182

AFRICAN												
				Frequency of Insertion			Delta		Pairwise Fst		Heterozygosity	
rs#	Chrom.	Position	Sequence	African	Caucasian	East Asian	v.EAS	v.CAU	v.EAS	v.CAU	He	Ho
rs59385244	1	16367160	-/AAGG	0.314103	0.821721	0.99	0.675897	0.507618	0.66481	0.424857	0.400337	0.261818
rs59009450	1	248818535	-/AAGAT	0.689103	0.0881148	0.246667	0.442436	0.6009882	0.325636	0.575336	0.421831	0.24
rs11277277	2	11273217	-/CACAG	0.339744	0.987705	0.936667	0.596923	0.647961	0.552569	0.687956	0.332102	0.176364
rs67344973	2	178513061	-/GTTT	0.875	0.256148	0.263333	0.611667	0.618852	0.551826	0.545406	0.491639	0.325455
rs148921522	3	85588405	-/TAAC	0.160256	0.625	0.86	0.699744	0.464744	0.656259	0.354217	0.493889	0.354545
rs112191273	3	7351968	-/GCTT	0.657051	0.0266393	0.0433333	0.6137177	0.6304117	0.580653	0.656176	0.332102	0.165455
rs70941213	4	106669965	-/AGTT	0.916667	0.243852	0.12	0.796667	0.672815	0.776951	0.613038	0.480799	0.256364
rs72255563	5	176226827	-/ACTT	0.772436	0.114754	0.136667	0.635769	0.657682	0.576689	0.622541	0.4261	0.229091
rs60234845	6	155859718	-/CCAA	0.75	0.239754	0.156667	0.593333	0.510246	0.521703	0.412464	0.462232	0.349091
rs35625334	7	79883089	-/AGAT	0.894231	0.354508	0.106667	0.787564	0.539723	0.764883	0.450782	0.493248	0.312727
rs56767439	8	12977501	-/TTAC	0.810897	0.204918	0.156667	0.65423	0.605979	0.59818	0.534393	0.463231	0.287273
rs113043680	9	126640635	-/TAAG	0.708333	0.139344	0.0966667	0.6116663	0.568989	0.556619	0.511686	0.411409	0.265455
rs113501732	10	128948642	-/CCTGT	0.272436	0.911885	0.763333	0.490897	0.639449	0.386335	0.616993	0.428189	0.249091
rs74499778	11	129941381	-/AGCT	0.375	0.952869	0.62	0.245	0.577869	0.110407	0.583277	0.421831	0.309091
rs2307553	14	80121686	-/TGAC	0.884615	0.252049	0.38	0.504615	0.632566	0.430009	0.562808	0.49819	0.383636
rs138123572	15	72786235	-/TGAC	0.185897	0.959016	0.946667	0.76077	0.773119	0.738709	0.782133	0.388618	0.149091
rs66913380	17	42191379	-/GCCA	0.195513	0.786885	0.85	0.654487	0.591372	0.598891	0.515387	0.463231	0.312727
rs10540310	20	59105205	-/CTTC	0.272436	0.75	0.87	0.597564	0.477564	0.531245	0.371308	0.457037	0.327273
rs10560659	21	17025686	-/CAAT	0.778846	0.17418	0.12	0.658846	0.604666	0.607315	0.54009	0.443219	0.272727

A



B

Population of Origin	Inferred Population Assignment (%)			# of Individuals
	AFR	EAS	CAU	
African	100	0	0	156
East Asian	0	100	0	150
Caucasian	0	0	100	244

C

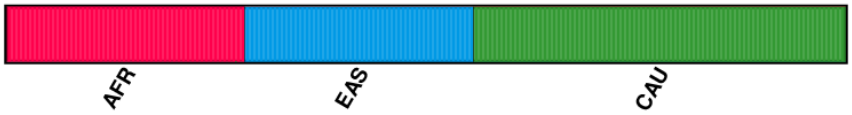


Figure 6. STRUCTURE v.2.3.4 Analysis of 59 AIMs. A) Graphical representation of the *ad hoc* statistic, ΔK. B) Table describing the overall population assignment of the training set samples for the 20 simulations at K=3. C) STRUCTURE plot for African (AFR), East Asian (EAS) and Caucasian (CAU) population groups compiled in CLUMPP v1.1.2 (36) and graphically displayed in *distruct v1.1* (37).

Additional Populations

When the Southwest Hispanic and Southwest Asian samples were added to the PCA, an additional cluster appeared in the plot (Figure 7A-B). Both population groups tended to cluster between the African and Caucasian population groups with complete separation from the East Asian population group. Both SWH and SWA samples cluster more closely with the Caucasian population group than the other two. The first three principal components for the SWH and SWA PCA explained 38.2% and 33.8% variance, respectively. To determine if the SWH and SWA population groups could be separated from each other using these 59 AIMs, PCA was performed on these two population groups (Figure 7C). The PCA showed significant overlap between to two population groups with PC1 and PC2 explaining only 8.5% of the variance.

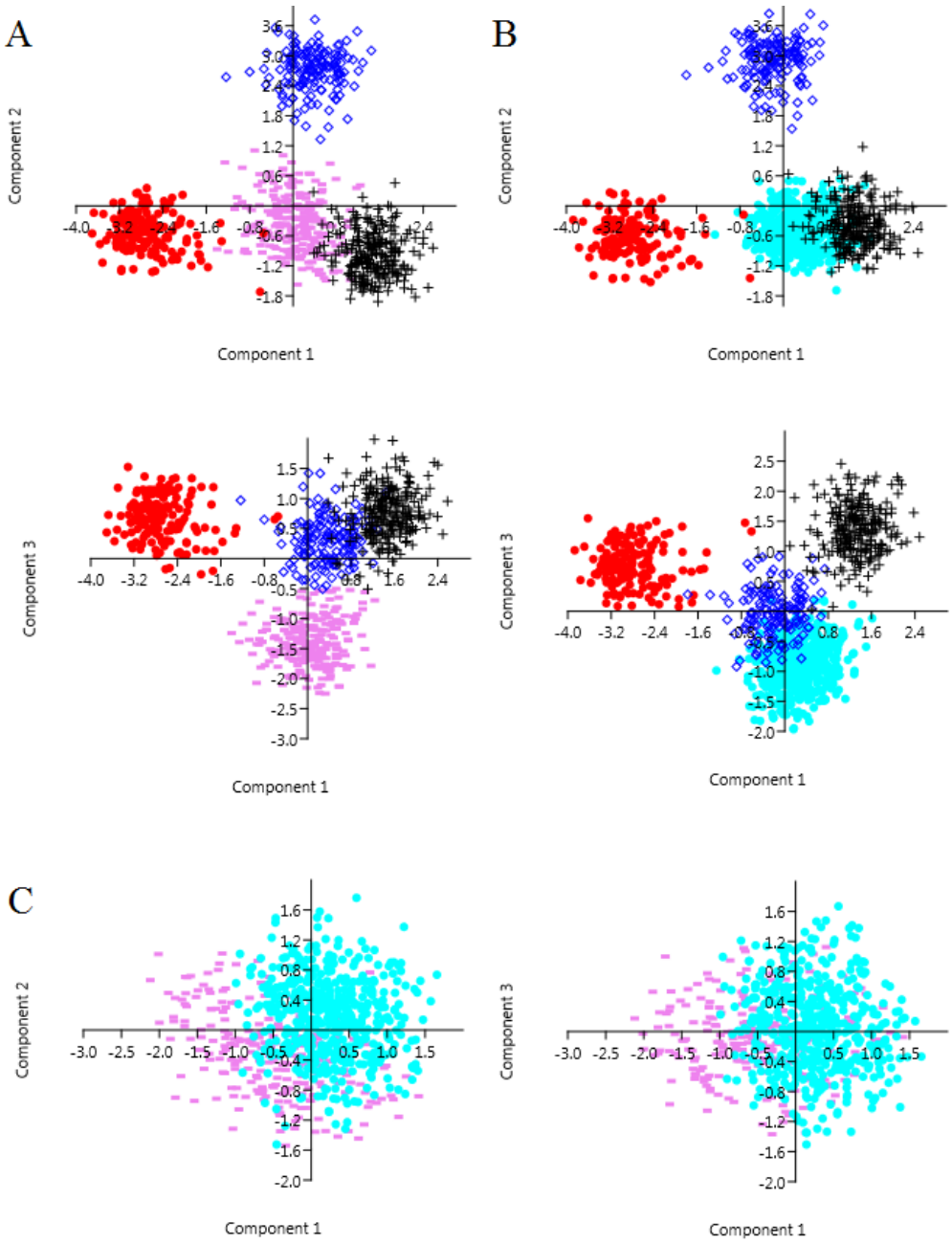


Figure 7. Principal Component Analysis (PCA) with Additional Population Groups Using Past3. A) Original training set of 550 Individuals; Caucasian (Black Plus), East Asian (Blue Diamond) and African (Red Dot) with Southwest Hispanic individuals (Pink Bar). B) Original training set individuals with Southwest Asian individuals (Turquoise Dot). C) PCA of SWH and SWA individuals.

CHAPTER 4

DISCUSSION

In this project, the hypothesis that publically available genome data could be used to select a panel of Ancestry-Informative Markers that can distinguish between Caucasian, East Asian, and African population groups was tested. A panel of 59 AIMs were selected using the genome data available through the 1000 Genomes Project. The Linux-based program called VCFtools was used to parse through the genome data. This program was used to calculate pairwise F_{ST} values, population-specific allele frequencies, and to obtain the genotype data for the 550 individuals in the training set. The initial screening of AIMs based on INDEL length, F_{ST}, and allele frequency divergence yielded approximately 1182 potential AIMs. These were then evaluated for strength based on their allele frequency differences, or delta value. In theory, the ideal delta value would be near 1 in both pairwise comparisons. However, in practice, this is not usually achievable. When choosing the 60 markers, those with the highest delta value were selected, taking into account genetic distance of syntenic markers. The highest and lowest delta values for the final panel of 59 markers were 0.838 and 0.275, respectively. The distribution of all pairwise delta values is described in Figure 8. The mean delta values for each of the three population groups are at least 0.55. Although some delta values are lower than anticipated, this is due to the difficulty to choose markers with high deltas in both pairwise comparisons.

PCA with the Past3 software indicated that the original 60 markers selected are sufficient to separate the three population groups. In Figure 5A, PCA shows three distinct clusters with no overlap between samples in different population groups. Additionally, three test populations from the 1000 Genomes Project cluster with the appropriate population group, shown in Figure 5B-D.

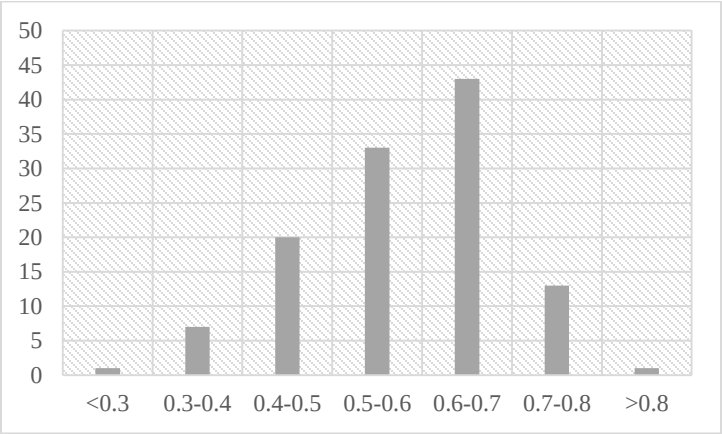


Figure 8. Distribution of Pairwise Allele Frequency Differences, or Delta, for the 59 AIMs.

Of the 60 AIMs initially selected, only one marker, rs78981054, showed significant departure from HWE. Upon closer examination of the sequence surrounding this INDEL, it appears that there may have been misalignment issues during sequencing due to the presence of a pseudo-repeat sequence when the insertion is present (Figure 8). This marker was removed from the panel. The remaining 59 markers were evaluated for LD. Six combinations, 0.35% of the total combinations tested, showed significant linkage disequilibrium. The LD in these marker systems could be attributed to the asymmetric allele frequencies of AIMs. Additionally, at least five of the six combinations involved one or more markers that showed significant departure from HWE prior to correction for multiple comparisons. Therefore, the LD observed in this panel is not more than was expected.

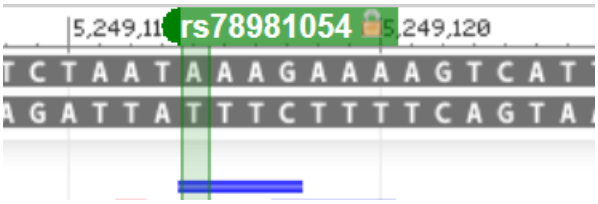


Figure 9. Screen capture from the dbSNP website for INDEL rs78981054.

In STRUCTURE v.2.3.4, the training set of 550 individuals were tested for ancestry admixture. Based on the *ad hoc* statistic, ΔK, calculated from 20 simulations of K 1-10, the 59 AIMs can cluster the samples into 3 population groups (Figure 6). Below and above K=3 introduces admixture within population groups that the 59 AIMs cannot distinguish.

When the individuals from SWH and SWA population groups were added to the PCA, both population groups clustered similarly between the Caucasian and African population groups (Figure 7A-B). Based on the similarity in the PCA, a second PCA was run with just the SWH and SWA population groups to determine if they could be distinguished from each other. Figure 7C describes the results of this PCA. As indicated previously, these two population groups were not able to be distinguished from each other. Additional markers would need to be added to the panel in order to distinguish between these two population groups.

CHAPTER 5

CONCLUSIONS

A panel of 59 Ancestry-Informative Markers were selected that can distinguish between three major global population groups; African, East Asian, and Caucasian. The INDEL markers are short in length, with high allele frequency divergence and population substructure. All markers are in Hardy-Weinberg Equilibrium with six combinations exhibiting significant linkage disequilibrium. Principal Component Analysis of the final panel indicated its ability to completely separate the three population groups as well as correctly cluster three known test populations. Similarly, analysis in STRUCTURE v2.3.4 demonstrates the presence of three separate clusters within the training set samples. These AIM-INDELs have the potential to be multiplexed into a single reaction for quick and simple analysis on a capillary electrophoresis platform. In future studies, the number of AIMs will systematically be decreased to determine the minimum number of INDELs necessary to distinguish between the three population groups. Preliminary results indicate that as little as twelve AIMs is sufficient to separate the three population groups. It would be advisable to first remove markers with the lowest delta and mean delta values in order to retain the most informative markers. A multiplex reaction will then be designed and developmentally validated for use in forensic laboratories.

Due to the limited size of the available data, populations were binned into their associated major global population group. By doing so, this may be underestimating admixture present

within the population groups. Additionally, these markers may be limited in their ability to definitively assign admixed samples to a population group. In future projects where the markers will be multiplexed, we may run into issues while designing primers.

In conclusion, a robust panel of Ancestry-Informative Marker-INDELs was selected with the ability to distinguish between the Caucasian, African and East Asian population groups. This panel will impact the forensic community by providing a simple method for ancestry analysis of samples with an unknown origin. The information gained from the panel can provide information to investigators that may be critical in the progress of forensic cases.

CHAPTER 6

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, and 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 electrophorsis 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. Applied biosystems 3500/3500xL genetic analyzer user guide. Rev. C edLife Technologies Corporation; 2009.

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 Research. 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;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 1;16(1):26-32.

15. Fondevila M, Phillips C, Santos C, Pereira R, Gusmao L, Carracedo A, et al. 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 2;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 3;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 1;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 Genet 2012;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. Forensic Sci Int Genet 2014;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. Investig Genet 2013;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;10(0):23-32.

24. Pereira R, Phillips C, Pinto N, Santos C, dos Santos SE, Amorim A, et al. Straightforward inference of ancestry and admixture proportions through ancestry-informative insertion deletion multiplexing. PLoS One 2012;7(1):e29684.

25. Francez PA, Ribeiro-Rodrigues EM, dos Santos SE. Allelic frequencies and statistical data obtained from 48 AIM INDEL loci in an admixed population from the brazilian amazon. Forensic Sci Int Genet 2012;6(1):132-5.

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

27. Nei M. Analysis of gene diversity in subdivided populations. Proc Natl Acad Sci 1973;70:3321-23.

28. 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;10(0):23-32.

29. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011 Aug;27(15):2156-8.

30. Hammer, Ø., Harper, D.A.T., Ryan, P.D. PAST: Paleontological statistics software package for education and data analysis. Palaeontol Electron 2001;4(1):9.

31. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics 2000 Jun;155(2):945-59.

32. Falush D, Stephens M, and Pritchard J. Inference of population structure using multilocus genotype data: Dominant markers and null alleles. Mol Ecol Notes 2007;7:574-8.

33. Falush D, Stephens M, Pritchard JK. Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. Genetics 2003 Aug;164(4):1567-87.

34. Hubisz MJ, Falush D, Stephens M, and Pritchard JK. Inferring weak population structure with the assistance of sample group information. Mol Ecol Resour 2009;9(5):1322-32.

35. Evanno G, Regnaut S, Goudet J. Detecting the number of clusters of individuals using the software structure: A simulation study. Mol Ecol 2005;14(8):2611-20.

36. Jakobsson M, Rosenberg NA. CLUMPP: A cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. Bioinformatics 2007 Jul;23(14):1801-6.

37. Rosenberg NA. DISTRUCT: A program for the graphical display of population structure. Mol Ecol Notes 2004;4:137-8.