5-1-2015

Ancestry Informativeness of Alu Markers in Four Populations Relevant for the United States

Aislinn G. D'Auben
University of North Texas Health Science Center at Fort Worth, adauben@gmail.com

D'Auben, Aislinn. Ancestry Informativeness of Alu Markers in Four Populations Relevant for the United States. Master of Science (Biomedical Sciences, Forensic Genetics). April 2015. 36 Pages, 8 tables, 2 figures, 17 references.

Determination of ancestry using DNA markers is an important issue in DNA forensics. The ability to identify an individual's ancestry could narrow down the pool of possible individuals involved in a crime. Several types of ancestry informative markers (AIMs) have been suggested in the literature. For this study, Alu markers were used for investigating their utility for Caucasian versus African and Caucasian versus Asian ancestry determinations. Three measures of AIMs were calculated for 42 Alu markers. Rank correlations of these three measures were used for investigating if a smaller number of top-ranked loci can improve ancestry determination. The Alu markers chosen for this study were less informative than anticipated but did show potential for ancestry estimation when all 42 markers were used together.

ANCESTRY INFORMATIVENESS OF ALU MARKERS IN FOUR POPULATIONS RELEVANT FOR THE UNITED STATES

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

Aislinn D'Auben, B.S.

Fort Worth, TX

April 2015

ACKNOWLEDGEMENTS

I would like to thank my major professor, Dr. Ranajit Chakraborty, for all of the assistance and guidance throughout my educational career at the University of North Texas Health Science Center. I would also like to thank all of the professors who have provided leadership and support of my educational endeavors. A special thanks to my committee members: Dr. Bobby LaRue, Dr. Robert Barber, and Dr. Rosalie Uht for their support and guidance. InnoGenomics Technologies Inc, who provided my data and corroborated with throughout the project. I would like to thank my fellow Forensic Genetics colleagues for their friendship and support over the course of this program. Ms. Judy Schulte, who taught me the love of learning and helped shape who I am today. I want to thank to my mother for always pushing me to follow my dreams and supporting me throughout my educational career. Lastly, to my family and loved ones for always supporting and believing in me.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

The forensic community has long since been using DNA analysis not only to identify suspects and victims but also to associate unidentified human remains with families of missing persons. Frequently though, the suspect(s) are not clear and so while the DNA analyst may already have a profile for the alleged suspect the profile cannot go into real use until the suspect pool has been narrowed down. Identification of missing persons can also be simplified if the pool of families can be narrowed down by incorporating additional information, preferably obtained by selecting appropriate panel of markers. In the past many different types of markers have been studied to hopefully be able to infer ancestral population origin, including mitochondria, Y-chromosome, microsatellites or, short tandem repeats (STRs), and single nucleotide polymorphisms (SNPs). (1, 2, 3)

The majority of forensic DNA analyses is done with STRs which are multiallelic polymorphic markers, consequently making them very informative for personal identification, missing person identification, as well as DNA mixture analyses. At present, the forensic laboratories of continental U.S. most commonly use the panel of 15 DNA markers for human identification, called the Identifiler STR panel. (4)

However, if for instance a DNA sample is degraded then a full STR profile will not likely be seen and the individualizing power of the profile may be substantially reduced. (5)

Because of the nature of degraded samples, genome wide studies have been done to find other genetic markers that can preserve a high individualizing power of the sample. Markers that have been found to obtain such goals are single nucleotide polymorphisms (SNPs), Indels, and Alu markers. (6, 7, 8)

SNPs are biallelic single base change differences in the genome and because of this SNPs have some limitations. SNPs can be merely identical by state, may have arisen as a result of an independent parallel forward or backward mutation resulting in genotype homoplasy. (1) In addition, for the SNP sites often the state of the progenitor and mutant alleles are unknown. Due to these limitations SNPs are not always an accurate way to show ancestry.

INDELs, or insertion deletion polymorphsims, are also biallelic and are of varying lengths from 1 to 10,000bp. Like the SNPs, they are also highly abundant in the human genome. (9) The limitation of these markers is that location of the deletion events in the genome related to the indel sites are unknown, and only their current sites where they currently reside can be mapped. (8)

This leads us to the type of markers that is the focus of this study- the Alu markers. Alu markers, or Alu insertion elements, are the most abundant class of short interspersed elements (SINEs) in the human genome. They are dimeric 300 bp sequences that propagate by retrotransposition into new chromosomal locations. Alu markers along with other SINE elements are highly informative markers for evolutionary and phylogenetic studies because they have a unique mutations mechanism, an absence of back mutation, and a lack of recurrent forward mutation (10, 11, 12, 13). Because of these a specific Alu marker will be identical by descent in all individuals in whom they occur (10). This allows sets of related chromosome regions marked by an Alu marker to be distinguished from a pool of ancestral chromosomes that lack the

element. These features give each locus genetic polarity that allows the independent assignment of an ancestral state and a root for phylogenetic analyses. (14)

A previous study of 100 Alu markers suggested reliable ancestry determinations for 18 selected individuals of various ancestry based on their pedigree information. (1) In this study 42 of these markers were chosen and examined to investigate which Alu markers are most informative for a comparison between populations: Caucasians versus Africans and Caucasians versus Asians.

CHAPTER II

METHODS

Data

The data for this study was received from our collaborations with InnoGenomics Technologies through a NSF-funded project for novel genetic marker development for human diversity studies. Anonymized genotype data on 733 individuals from 4 populations (155 Caucasians, 118 Africans, 365 Asian Indians (which from now on will be referred to as Indians), and 77 Asians) along with the same 18 individuals of the previous study (1) with known pedigree ancestry were used in this research.

Allele frequency, Hardy-Weinberg equilibrium, linkage disequilibrium

The software GDA (Genetics Data Analysis, downloaded from the web-site http://hydrodictyon.eeb.uconn.edu/people/plewis/software.php) was used for calculations of allele frequencies, conducting tests for Hardy-Weinberg equilibrium (for checking random association of alleles within a locus to form genotypes), and linkage equilibrium between all pairs of loci (for checking random association of alleles between pairs of loci). In these computations allele frequency estimates used the gene counting method, and significance testing was done by empirical determination of significance by permutation tests. (16)

Three measures of ancestry informativeness notation

For these calculations consider populations i=1,2,…,K with K ≥ 2 and a locus with N = 2 alleles. Let p_{ij} denote the frequency of allele j, j= 1,2,…,N, in population i. Let p_j denote the average frequency of allele j over the K populations, for example $p_j = (p_{1\ caucasians} + p_{1africans})/2$.

The following measures of ancestry informativeness were calculated by the comparisons of Caucasian population versus African population and Caucasian population versus Asian population. This was done because these population comparisons are the most "extreme" and least likely to show admixture.

Absolute allele frequency difference (delta, δ)

Delta, is the absolute frequency difference of a particular allele observed in two ancestral populations. A marker with the δ=1 has perfect information in relation to the ancestry, since δ=1 implies that one of the two alleles is totally fixed in one ancestral population with the alternative allele fixed in the other ancestral population. In contrast a δ=0 has no information in regard to ancestry, since for such a locus both populations will have identical allele frequencies. (16)For a biallelic locus:

$$\delta = |p_{11} - p_{21}|$$

Delta only tells a limited amount of information in relation to ancestry so should not be used alone but in conjunction with multiple other measures of ancestry informativeness calculations.

F statistics (F_{ST})

F_{ST} is the proportion of the total genetic variance of a locus contributed by the genetic variances between subpopulations. "When only two parental populations and markers with only

two alleles are considered, this informativeness (F_{ST}) for ancestry includes the differences and sum of the reference allele frequencies in the two parental populations."(16)

$$F_{ST} = \frac{(p_{1j} - p_{2j})^2}{(p_{1j} + p_{2j})(2 - (p_{1j} + p_{2j}))}.$$

A high F_{ST} value implies a large degree of differentiation between populations. In other words, the above formulation of F_{ST} measures the genetic distance between any two populations. This calculation has recently been used as a criterion for selecting markers for ancestry estimation (i.e. the ones with high F_{ST} values).

Informativeness for assignment (I_n)

I_n is a mutual information-based statistics that takes into account self-reported ancestry information from the sampled individuals. Following Ding et al. (16) the informativeness for assignment can be defined as:

$$I_n = \sum_{j=1}^{N} (-p_j \log_2 p_j + \sum_{i=1}^{K} \frac{p_{ij} \log_2 p_{ij}}{K}).$$

This formula is a generalization for more than two populations. From a likelihood perspective, it gives the expected logarithm of the likelihood ratio that n allele is assigned to one of the populations compared with a hypothetical average population whose allele frequencies equal the mean allele frequency across the K populations. The smaller the value the more similar the allele frequencies are in all populations. (16) For this particular study the formula above could more easily be defined as follows:

$$I_n = [(-p_j \log_2 p_j) + (p_{0caucasian} \log_2 p_{0caucasians} + p_{0african} \log_2 p_{0african})/2] + [(-p_j \log_2 p_j) + (p_{1caucasian} \log_2 p_{1caucasians} + p_{1african} \log_2 p_{1african})/2]$$

Comparison of measures

These three measures were calculated for each of the 42 markers by writing Excel functions for the two contrasts of Caucasians versus Africans and Caucasians versus Asians. Two approaches were taken to compare these three measures of informativeness. First, within each of the two contracts, the values of the three measures were ranked for the 42 markers (i.e. a rank of 1 assigned to the markers with the highest value, and a rank of 42 assigned to the marker with the lowest observed value). Correlations of these ranks were computed for delta versus F_{ST}, delta versus I_n, and F_{ST} versus I_n to examine the degree of congruence of ancestry informativeness across the three measures for a specific contrast of ancestral populations. Second, for each measure, rankings of the loci for the two contrasts were checked to examine if the same set of markers can be used for both contrasts to select a smaller selection of markers for ancestry determination.

STRUCTURE

The STRUCTURE software was used to further analyze the loci within each of the populations to better determine the individualized ancestry of each individual sampled. (parameters used: admixture model, length of burn_in period= 100,000, number of MCMC reps after burn_in- 10,000) (17) In all cases of STRUCTURE analyses, 42-locus genotypes of all 733 individuals were used, grouping the individuals adjacently as being to population 0= individuals of known ancestry (n=18), 1= Africans (n=118), 2= Asians (n=77), 3= Caucasians (n=155), and 4=Indians (n=365). First, all 42 loci were used to examine how each individual's ancestry compared with their stated population originally before eliminating loci. An average cluster percentage was also obtained for individuals sampled from the four populations (Caucasians, Africans, Asians, and Indians). Next, the comparisons were analyzed with their top 15 (for the

contrast of Caucasian versus African) or top 16 (for the contrast of Caucasian versus Asian) loci. Each comparison had their average cluster percentages calculated per ancestry and these numbers were then compared to the 42 loci cluster percentages to examine if the reduced set of markers provide any improvement of ancestry inference since they include the loci with higher degree of ancestry informativeness.

Validation with known ancestry samples

Finally, the 18 individuals who have a known pedigree based ancestry were cross checked with the original 42 informative loci to assess if these loci can indeed pinpoint the ancestry of an individual accurately. This was done with using the results of the same STRUCTURE software analyses (by using the STRUCTURE percentages for the first set of 18 individuals with population designation of 0) and calculating their cluster percentages and comparing to their known ancestries.

CHAPTER III

RESULTS

Allele frequencies of the markers and Allelic independence within loci (Hardy-Weinberg Equilibrium) and between pairs of loci (Linkage Equilibrium)

Allele frequencies (obtained by the gene count method) for the 42 Alu markers in four populations (Caucasians with n=155 individuals, Africans with n=118, Asians with n=77, and Indians with n=365) are shown in Appendix Table A1. Of the 42 markers, the HS4.75 locus had the Alu insertion allele (designation 1) fixed (i.e., had a frequency of 1.0) in populations Caucasian and Asian, the Ya5NBC132 locus had the Alu insertion allele (designation 1) fixed (i.e. had the frequency of 1) in populations Caucasians, Asian, and Indians, the Ya5NBC150 locus had the Alu insertion allele (designation 1) fixed (i.e., had a frequency of 1.0) in population Asian, the Ya5NBC157 locus had the Alu insertion allele (designation 1) fixed (i.e., had a frequency of 1.0) in populations Caucasians, Asians, and Indians, the Ya5NBC159 locus had the Alu insertion allele (designation 1) fixed (i.e., had a frequency of 1.0) in population Caucasian, the Ya5NBC212 locus had the Alu insertion allele (designation 1) fixed (i.e., had a frequency of 1.0) in populations Caucasians and Asians, the Yb8NBC450 locus had the Alu insertion allele (designation 1) fixed (i.e., had a frequency of 1.0) in populations Asians and Indians.

List of loci with the observed p-values from the GDA software which showed deviation from Hardy-Weinberg equilibrium (HWE) are shown in Appendix table A2. At the nominal level

of significance of 5%, the number of loci that deviated from HWE were: 1 locus in Africans, 2 in Asians, 1 in Caucasians, and 4 in Indians but after the Bonferroni correction of adjusted level of significance (0.05/42= 0.0012) no locus was observed to deviate from HWE. Likewise, none of the pairs of loci showing deviation from linkage equilibrium (with a nominal level of significance of 5%), described in Appendix table A3, had p-values that are lower than 5.8×10^{-5} (=0.05/861). In other words, overall it was concluded that deviations from HWE or that from LE were not significant in this dataset of 42 Alu markers in the four populations examined in this study.

Measures of Ancestry informativeness and their correlations/congruence

As these 42 markers in general exhibited statistically independent genotype distributions, their ancestry informativeness were investigated with the three measures: delta, F_{ST}, and I_n.

These three measures were calculated using all 42 markers. Each measure was then ranked on a scale of one to forty-two (one being more informative in relation to ancestry). Table 1 shows how each marker ranked in regards to delta, F_{ST}, and I_n. For ease of comparisons, the markers are arranged in the order of their informativeness ranking with respect to delta, which readily helps to examine if the same set of markers appear in any list of top-ranked markers by any of the other two measure of informativeness.

			Cau-Afa			
Marker	delta	delta ra	Fst	Fst rar	In	In rar
Ya5NBC157	0.94218	1	0.04683	28	0.36712	2
HS4.75	0.77113	2	0.20573	15	0.18829	5
Ya5NBC212	0.71	3	0.25668	13	0.14909	9
Ya5NBC241	0.70614	4	0.70858	1	0.44852	1
Ya5NBC159	0.7028	5	0.27306	11	0.14499	10
Ya5NBC132	0.66552	6	0.30602	9	0.12524	13
Yb8NBC547	0.44165	7	0.33386	4	0.15929	6
Ya5NBC150	0.44001	8	0.37079	3	0.20115	4
Ya5NBC45	0.4341	9	0.39453	2	0.22641	3
Ya5NBC351	0.43385	10	0.32413	5	0.14399	11
Ya5NBC208	0.43012	11	0.32267	6	0.151	7
Yb8NBC466	0.40136	12	0.30817	7	0.15077	8
TPA25	0.38951	13	0.27806	10	0.11933	14
Yc1NBC53	0.37526	14	0.3062	8	0.13301	12
Yb8NBC568	0.3419	15	0.25864	12	0.10831	15
Yb8NBC405	0.34153	16	0.22156	14	0.09208	16
Yb8NBC148	0.31331	17	0.19258	19	0.08309	19
Yb8NBC485	0.29668	18	0.1947	18	0.07874	21
Ya5NBC354	0.27113	19	0.1962	17	0.08045	20
Yb9NBC10	0.25271	20	0.11719	24	0.04667	25
Ya5NBC221	0.22069	21	0.14247	23	0.0634	23
Ya5NBC347	0.2202	22	0.19741	16	0.0832	18
Yb8NBC157	0.20631	23	0.14461	22	0.05733	24
COL3A1	0.20225	24	0.1542	21	0.07533	22
Yb8NBC450	0.19475	25	0.16407	20	0.08756	17
Yb8NBC596	0.16951	26	0.0577	25	0.02363	26
Yb8NBC479	0.15852	27	0.05179	27	0.02106	28
Yb8NBC419	0.15834	28	0.05219	26	0.02117	27
Yb8NBC5	0.15105	29	0.04197	30	0.01699	30
Ya5NBC311	0.14625	30	0.04656	29	0.01887	29
Yb8NBC480	0.14429	31	0.04108	31	0.0165	31
Ya5NBC148	0.08944	32	0.01772	32	0.00786	32
Yb8NBC576	0.08664	33	0.01601	33	0.0071	33
PV92	0.08124	34	0.01236	34	0.00602	34
Yb8NBC201	0.05954	35	0.0037	35	0.00259	35
Ya5NBC51	0.05486	36	0.0022	36	0.00218	36
Yb9NBC50	0.03621	37	9.5E-05	38	0.00128	38
Yb8NBC125	0.03198	38	0.00151	37	0.00203	37
Ya5NBC345	0.02594	39	-0.003	39	0.0005	39
Yb8NBC106	0.01238	40	-0.0039	40	0.00011	40
B65	0.00767	41	-0.0045	42	4.32E-05	41
Yb8NBC636	0.00196	42	-0.0043	41	2.87E-06	42

Table 1a: Caucasian versus African comparison - rank of 1 being the most informative marker and rank of 42 being the least informative marker. The Top ranking markers for this population comparison are the top 15 markers.

			cau- asa			
Marker	delta	delta ran	Fst	Fst rar	In	In rar
Ya5NBC159	0.99324	1	0.00309	31	0.47413	1
Yb8NBC450	0.9913	2	0.00129	34	0.46828	2
Ya5NBC150	0.9537	3	0.03354	21	0.38638	3
PV92	0.61921	4	0.54616	1	0.30088	4
Yb8NBC419	0.45882	5	0.35018	2	0.1587	6
Yb8NBC480	0.37994	6	0.30282	3	0.16117	5
Ya5NBC51	0.35995	7	0.2439	4	0.11418	7
Yb8NBC576	0.33606	8	0.20085	7	0.08343	9
Yb8NBC157	0.33377	9	0.20161	6	0.0832	10
Ya5NBC345	0.31941	10	0.2045	5	0.09575	8
Yb8NBC485	0.30505	11	0.16775	8	0.07115	11
Yb8NBC5	0.2779	12	0.13964	9	0.05653	12
Ya5NBC354	0.27178	13	0.13321	10	0.05396	13
Ya5NBC241	0.22785	14	0.10265	13	0.04023	16
Ya5NBC148	0.22085	15	0.10826	11	0.04184	14
Yb8NBC547	0.20253	16	0.10444	12	0.04058	15
Yb8NBC106	0.19013	17	0.06668	15	0.02793	18
Yb9NBC50	0.1868	18	0.07499	14	0.02918	17
TPA25	0.18535	19	0.06114	16	0.02494	19
Yb8NBC596	0.16671	20	0.05656	17	0.0229	20
Ya5NBC347	0.15762	21	0.05159	18	0.02052	21
Ya5NBC208	0.13292	22	0.04815	19	0.01898	22
B65	0.12685	23	0.02626	22	0.01164	24
Yb8NBC568	0.11806	24	0.02215	25	0.01051	27
Yb8NBC479	0.11058	25	0.02446	24	0.01067	25
Ya5NBC351	0.08814	26	0.01189	28	0.00655	30
Yb8NBC466	0.08159	27	0.02449	23	0.01059	26
Ya5NBC311	0.0785	28	0.01462	27	0.00764	29
Yb8NBC201	0.07287	29	0.00564	30	0.00387	32
Ya5NBC221	0.06919	30	0.02176	26	0.00946	28
COL3A1	0.0634	31	0.0358	20	0.01415	23
Yb9NBC10	0.06289	32	0.00301	32	0.00308	33
Yb8NBC636	0.06081	33	0.00153	33	0.00271	34
Yb8NBC125	0.04547	34	0.01124	29	0.00648	31
Yb8NBC148	0.03989	35	-0.0002	39	0.00187	35
Yb8NBC405	0.02162	36	-0.0045	41	0.00034	37
Ya5NBC45	0.00981	37	-0.0043	40	0.0015	36
Yc1NBC53	0.00779	38	-0.006	42	4.38E-05	38
HS4.75	0	39	0	35	0	39
Ya5NBC132	0	40	0	36	0	40
Ya5NBC157	0	41	0	37	0	41
Ya5NBC212	0	42	0	38	0	42

Table 1b: Caucasian versus Asian comparison - rank of 1 being the most informative marker and rank of 42 being the least informative marker. The top ranking markers for this population comparison are the top 16 markers.

The rankings from Table 1 were then used to compute their correlations between the three measures of AIMS to examine how closely correlated the three measures were to one

another. A correlation closer to one showing high correlation which for this study would mean the three measures for a particular marker have similar rankings. This also can make choosing the top markers more easily.

The Caucasian versus African comparison showed high correlations across the three measures while the Caucasian versus Asian comparison showed similar correlation for delta- F_{ST} but the delta-I_n being much lower than anticipated.

Cau/Afa			
	delta	F_{ST}	I_n
delta			
F_{ST}	0.974		
I_n	0.848	0.9238	

Table 2a: Caucasian versus African comparison- Correlation calculations of the three measures of informativeness. Rankings by Delta and F_{ST} showing the highest correlation of 0.974.

Cau/Asa			
	delta	F_{ST}	I_n
delta			
F_{ST}	0.934		
I_n	0.61	0.789	

Table 2b: Caucasian versus Asian comparison- correlation calculations of the three measures of informativeness. Rankings of Delta and F_{ST} showing the highest correlation of 0.934.

Though the trend of correlations between three measures of informativeness is the same for both populations contrasts (namely, correlation between delta and F_{ST} is the largest, and that between delta and I_n the poorest), all three measures are more strongly correlated for the Caucasians versus Africans as opposed to Caucasians versus Asians. In others words, these three measures are likely to give better resolution of African versus Caucasian ancestry as compared to Caucasian versus Asian ancestry. This is also reflected in the observation that the top-ranked Alu

markers are not necessarily the same for these two population contracts. For example, among the 15 top ranked markers for the Caucasians versus Africans contrast, only four appear in the list of 16 top ranked markers of the Caucasians versus Asians contrast.

Distributions of the three measures of marker informativeness for both comparisons (Caucasians versus Africans and Caucasians versus Asians) can be seen in Figure 1 (six panels). First, note that although the markers in each panel (Y-axis) are arranged from top ranked (at the bottom) to bottom ranked (at the top), the order of the markers are different in each panel (as reflected in tables 1a and 1b). Nonetheless, all six distributions show that the majority of the markers have informativeness values lower than 0.5. In contrast, markers having a strong relation to ancestry should have a value closer to one. In other words, this panel of 42 Alu markers is not particularly a rich set of informative markers for either Caucasians versus Africans or Caucasians versus Asians ancestry discrimination.

To better detail this point Table 3 shows the mean, standard deviation, minimum value, and maximum value for the three measures of informativeness using all forty-two markers and the same descriptive statistics for the 15 or 16 top-ranked markers for the same two population contrasts (Caucasians versus Africans and Caucasians versus Asians). Several observations from these calculations are worthy to note in relation to the utility of these markers for ancestry investigations. First, for the 42 markers in aggregate, the mean informativeness of the measures for both populations' contrasts is at best 0.2264 (delta for Caucasians versus Africans contrast). The maximum also does not exceed 0.7086 (for F_{ST} in Caucasians versus Africans contrast). In other words, as a panel for ancestry informativeness, they are far from being ideal for either Caucasians versus Africans or Caucasians versus Asians contrasts. Selection of top ranked markers (15 top ranked or 16 top ranked) does not improve the effectiveness much. Though the

means are substantially raised by elevating their minimum values, the improvement may not be effective since the reduction of standard deviation is not substantial. Further, as a panel of ancestry informative markers, these markers are likely to be less effective for Caucasian versus Asian ancestry as compared to Caucasian versus African ancestry.

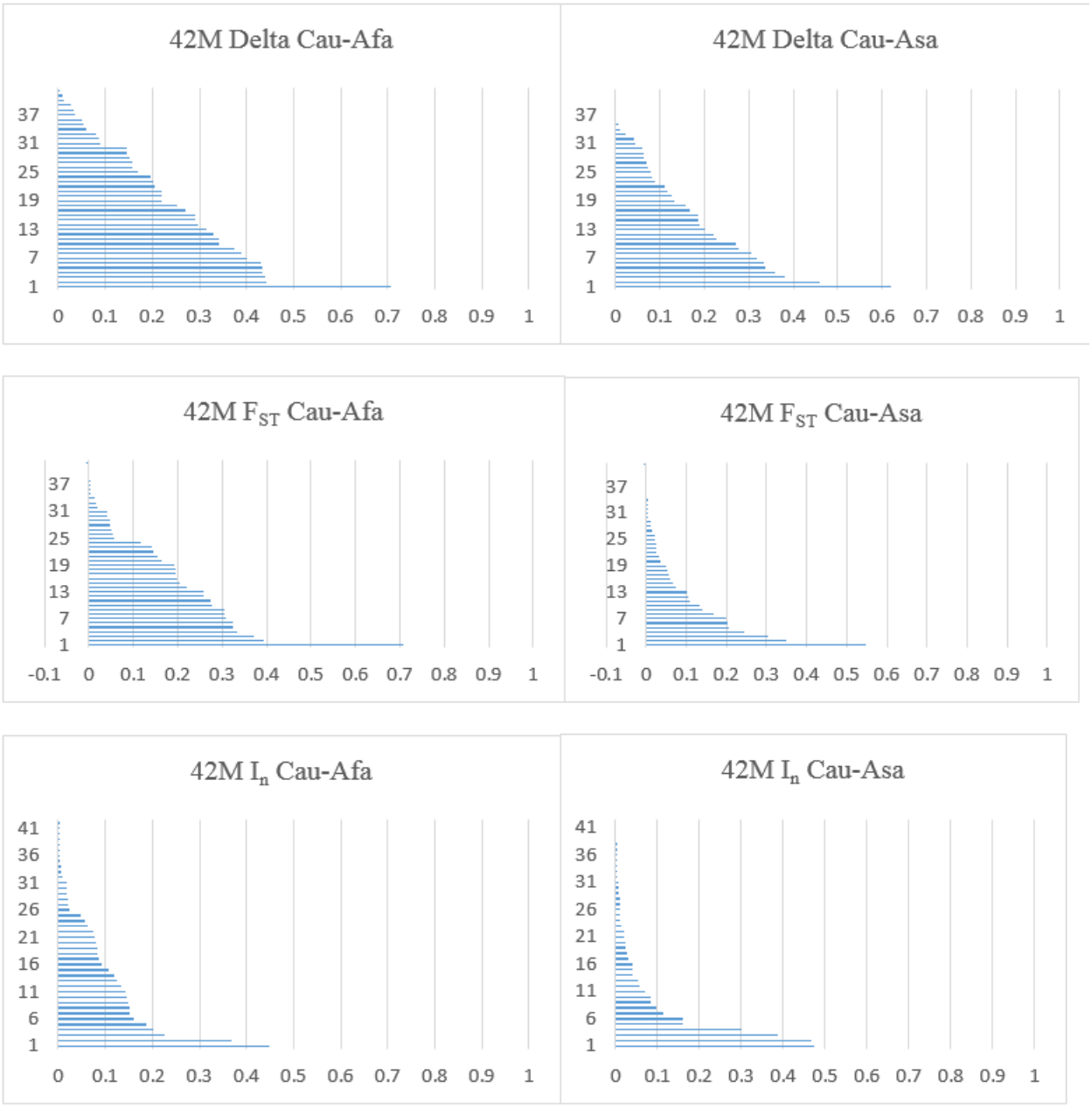

Figure 1: (left) Caucasian to African comparison- bar graph showing the delta values (top), F_{ST} values (middle) and I_n values (bottom). Good ancestry information for all of these measures would be a number close to one. None of the values are near one. (right) Caucasian to Asian comparison- bar graph showing the delta values (top), F_{ST} values (middle) and I_n values (bottom). This comparison also showed values not nearing one similar to the Caucasian to African comparison.

	Caucasians vs. Africans			Caucasians vs. Asians		
	delta	F_{ST}	I_n	delta	F_{ST}	I_n
All 42 markers						
mean	0.2264	0.1607	0.0884	0.1521	0.0807	0.068
s.d.	0.1565	0.1516	0.0963	0.1441	0.1141	0.119
minimum	0.0020	-0.0045	2.86E-06	0	-0.006	0
maximum	0.7061	0.7086	0.4485	0.619	0.5462	0.474
Top ranked* panel of markers						
Mean	0.3864	0.3095	0.1533	0.305	0.188	0.865
s.d	0.1149	0.1339	0.0922	0.1118	0.1229	0.0692
minimum	0.1443	0.0411	0.0165	0.1853	0.0611	0.0249
maximum	0.7061	0.7086	0.4485	0.6192	0.5462	0.3009

Table 3: Calculations of mean, standard deviation, minimum, and maximum of the three measures of informativeness using all forty-two markers for both comparisons of Caucasian to African and Caucasian to Asian along with the top ranked panel of markers for each contrast.

Effectiveness of Ancestry Determination of the 42 Alu markers

Finally, for an explicit evaluation of effectiveness of ancestry determination by these markers, the STRUCTURE software was used to create Figure 2 which shows all 733 individuals arranged in adjacent order of their designated population (Group 0= individuals with known ancestry, 1=Africans, 2=Asians, 3= Caucasians, and 4=Indians). Each vertical line in this

figure represents an individual's ancestral make-up. For example, an individual with two colors in a vertical line would represent having been derived from two ancestral populations (under the admixture model of STRUCTURE), or can be assigned to ancestry in one of the two populations (with probability corresponding to the length of the two colors). Though the number of presumed populations, K, was chosen for these results, initially the colors (green, yellow, blue, and red) were not necessarily assigned to any of the known populations. Nonetheless, some observations are instructive as far as the effectiveness of these 42 Alu markers for ancestry determination. For example, the individuals in group 1 (which are sampled as Africans) are predominantly of green color, with some indication of red color for several individuals within this group. To a lesser degree, individuals in this group also occasionally showed blue or yellow colors. This situation is considerably more complex for the other colors. For example, while yellow is prominent for individuals sampled as Asians (group 2), influences of red color is also seen in a substantial number of them. The color blue is again predominant for group 3 (Caucasians) individuals, though several of them also have yellow and red color influences. The individuals of the fourth group (Indians) are most mosaic in structure with considerable mixing of red and yellow and occasionally blue. Taken together, except for green versus the other colors, these 42 markers do not appear to be of great confidence in their use for distinction of the four populations examined in this study.

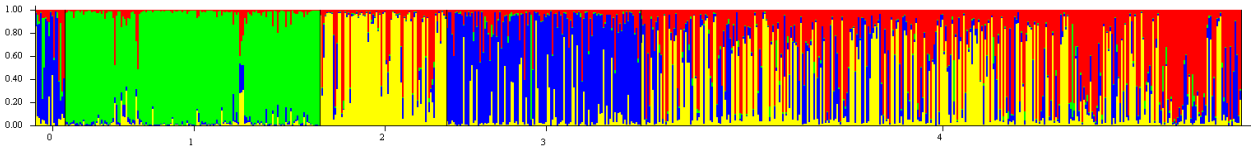

Figure 2: STRUCTURE bar graph of all 733 individuals with K=4 assumed, grouped into their designated population (0= known ancestry, 1= African, 2=Asian, 3= Caucasian, 4= Indian). Each color represents a population and every line represents an individual.

These observations were quantitatively assessed by using other components of the STRUCTURE software outputs. This software also calculated the cluster assignment proportions for the four populations for each individual of the study. Table 4 presents average values of these cluster assignments within the four groups of the present study. In other words, for the 733 individuals studied, the entries of Table 4 reflect the estimates of their assignments in the four presumed populations (K=4) of the STRUCTURE software. Based on these numbers each cluster (green, yellow, blue, and red) can be assigned a population. For instance, cluster two (green) had the highest (0.905) average in individuals of population one (Africans) and hence so cluster two (green) may be inferred as the African cluster. With the same logic, upon doing this with the other clusters and populations it can be said that cluster three (blue) is the Caucasian cluster and cluster one and four (yellow and red) together represent the Asians and Indian clusters.

	Inferred Clusters				
Given pop	1 (Asian/India)	2 (Africa)	3 (Caucasian)	4 (Indian/Asian)	n
0 (with known pedigree)	0.171	0.122	0.587	0.121	18
1 (Africans)	0.038	0.905	0.027	0.03	155
2 (Asians)	0.196	0.009	0.068	0.727	77
3 (Caucasians	0.106	0.019	0.741	0.134	118
4 (Indians)	0.431	0.019	0.193	0.357	365

Table 4: Average cluster assignment for K=4 using all 42 markers.

Upon choosing the top-ranked markers (15- or 16-) the cluster assignment performance in general decreased in contrast to the expectation of an overall increase. This overall decrease is most likely due to the fact of choosing the top makers did not actually increase the empirical

values of the measures drastically and by decreasing the numbers of markers the cumulative informativeness of the multilocus genotypes decreased.

Validation by replication study of clustering assignment with the use of persons of known ancestry

The reliability of inferred cluster assignments by using these 42 Alu markers could finally be tested by considering the cluster assignment probabilities of each of the 18 individuals with known pedigree information from the same run of the STRUCTURE software. Table 5 presents the results of the cluster assignment estimates for these known samples. The most likely cluster assignment of individuals is denoted in bold in entries of this table. Though the degree of confidence is not ideal, in general, the results of this validation study is promising. Of the 11 individuals of known European ancestry (including the Greek subject), 8 could be assigned to European cluster with probabilities exceeding 0.75. The most likely assignment of the three African- Americans were also in either European or African cluster. Grossly inaccurate clustering assignment was found in one (SUB001) of the 18 individuals of this validation study.

Known Persons		Estimated chance of Cluster assignment in			
ID	Population ancestry	1 (Asian/Indian	2 (African)	3 (Caucasian)	4 (Indian/Asian)
SUB001	European	0.019	0.003	0.12	**0.857**
SUB002	European	0.036	0.091	**0.789**	0.085
SUB003	European	0.098	0.017	**0.82**	0.066
SUB004	European	0.016	0.006	**0.958**	0.019
SUB005	African-American	0.066	**0.702**	0.18	0.053
SUB006	African-American	0.071	0.427	**0.458**	0.044
SUB007	European	0.06	0.013	**0.908**	0.018
SUB008	European	0.097	0.003	**0.892**	0.008
SUB009	Jamaican	0.101	0.126	**0.516**	0.257
SUB010	Greek	0.01	0.005	**0.979**	0.006
SUB011	European	0.047	0.009	**0.507**	0.437
SUB012	European	0.409	0.021	**0.544**	0.026
SUB013	European	0.008	0.004	**0.983**	0.005
SUB014	European	0.018	0.048	**0.927**	0.006
SUB015	Venezuelan	**0.61**	0.008	0.322	0.059
SUB016	African-American	0.114	**0.665**	0.164	0.057
SUB017	Indian	**0.705**	0.043	0.151	0.101
SUB018	Chinese	**0.586**	0.005	0.34	0.07

Table 5: Cluster assignment of known individuals. Entries with the highest values within each row are denoted in bold, indicating the most likely cluster assignment.

In summation, therefore, the prospect of ancestry determination by these 42 Alu markers appears quite promising. It should be noted that the cluster assignment of these 18 individuals shown in Table 5 were in general congruent (15 out of 18) with those reported by Ray et al (1) who used a total of 100 Alu markers from which the 42 markers were used in this study.

CHAPTER IV

DISCUSSION/CONCLUSIONS

Estimating an individual's ancestry for the use in forensic cases could be a powerful tool for law enforcement. However, the ability to use ancestry in these types of cases relies on the ability of finding ideally ancestry informative markers. As discussed earlier in the introduction section, there are many genetic markers that researchers are examining in hopes of being able to infer ancestry and this research started with the presumption that Alu markers were believed to be the most promising of all of these markers.

Several measures have been suggested for ancestral informativeness of markers which would hopefully allow selection of a panel of such markers. The decision to choose a given measure to be used should depend on the efficiency of each measure to select the most ancestry informative marker. Currently, there is no consensus as to which of these measures should be used to select a marker for ancestry informativeness. Generally though, selecting a marker with large allele frequency differences between the ancestral populations should give a good indication of a marker that will be useful for an ancestry study.

In this study, three different analytical tools were used to evaluate the Alu markers to find the top markers for Caucasians versus Africans and Caucasians versus Asians contrasts of ancestry distinctions. Each of the three measures (delta, F_{ST}, and I_n) showed relatively low empirical values, compared to the ideal value of 1.0 which would infer perfect information in

relation to ancestry. Due to these low numbers for all forty-two markers this already shows that the markers will not be as useful as originally anticipated. Selection of top-ranked markers did not help in increasing the efficiency of ancestry inference. Upon taking the top 15 markers for Caucasians versus Africans and the top 16 markers for Caucasians versus Asians contrasts, the values of the three measures did increase somewhat, but the means were still below 0.5, with wide variation (i.e. large enough standard deviation). This result shows that the original forty-two markers for this study are overall not as informative with regard to ancestry as a panel of markers should be. Given that the correlations between the informative measures were relatively high, with the exception of a relatively low correlation between delta and I_n the Caucasians versus Asians contrast had good concordance between the three measures in relation to ancestry informativeness.

Even though the majority of the markers did not exhibit values above 0.5 across all three measures, a few markers surpass 0.5 and may be more useful for ancestry estimation. These markers include: Ya5NBC241 (delta= 0.706, F_{ST}= 0.7086, I_n= 0.4485) and PV92 (delta= 0.619208, F_{ST}= 0.5462, I_n= 0.301). Even these markers that have higher values for delta and F_{ST} still fall slightly short for the I_n measure but could still be useful for further study.

While the top markers for both population comparisons did not prove to be more informative than the original 42 markers this does not mean these markers are not useful. The lack of improvement by selection of the top-ranked markers (in comparison to the initial 42 markers) is mainly due to a lower degree of informativeness of 15- (or 16-) locus genotypes in comparison to 42-locus genotypes. Upon analyzing the eighteen known ancestry individuals using all 42 markers only one individual was grossly categorized incorrectly and two were

questionable which gives a total of thirteen individuals who were put into the correct population cluster. This is an accuracy of 72.2%.

Some limitations of this study should be noted for any generalization of the results presented here. First, the choice of the study samples is not clearly ideal for ancestry determinations of individuals of continental USA. The Caucasian verses Asian contrast considered here is not a perfect surrogate of distinctions of Caucasian and Native American ancestry.

The Alu markers used for this study were not the most informative in relation to ancestry but Alu markers as a whole show true potential in the future for correctly categorizing an individual into the correct ancestral population. This can be deduced based on the 72.2% accuracy of the known individuals. If further Alu markers can be found that show higher values (closer to one) in the absolute allele frequency difference (delta), F statistics (F_{ST}), and the Informativeness for assignment measure (I_n) then the accuracy of ancestry testing will increase and can be made part of the DNA analysis testing for forensic cases.

APPENDIX

Allele Frequencies	Caucasian		African		Asian		Indian	
Marker	n	frequency	n	frequency	n	frequency	n	frequency
B85	118	0.560185	155	0.567857	77	0.433333	365	0.45961
COL3A1	118	0.025641	155	0.227891	77	0.089041	365	0.063712
HS4.73	118	1	155	0.771127	77	1	365	0.993094
PV92	118	0.232143	155	0.31338	77	0.851351	365	0.467967
TPA25	118	0.582609	155	0.193103	77	0.39726	365	0.595714
Ya5NBC132	118	1	155	0.665517	77	1	365	1
Ya5NBC148	118	0.199153	155	0.288591	77	0.42	365	0.232044
Ya5NBC150	118	0.953704	155	0.513699	77	1	365	0.968056
Ya5NBC157	118	1	155	0.942177	77	1	365	1
Ya5NBC159	118	1	155	0.702797	77	0.993243	365	0.980226
Ya5NBC208	118	0.849138	155	0.419014	77	0.716216	365	0.879213
Ya5NBC212	118	1	155	0.71	77	1	365	0.948895
Ya5NBC221	118	0.931034	155	0.710345	77	0.861842	365	0.940278
Ya5NBC241	118	0.734513	155	0.028369	77	0.506667	365	0.512712
Ya5NBC311	118	0.782609	155	0.636364	77	0.861111	365	0.743478
Ya5NBC345	118	0.56087	155	0.586806	77	0.880282	365	0.693642
Ya5NBC347	118	0.752212	155	0.972414	77	0.594595	365	0.588235
Ya5NBC351	118	0.644737	155	0.210884	77	0.732877	365	0.627143
Ya5NBC354	118	0.360169	155	0.089041	77	0.631944	365	0.347222
Ya5NBC45	118	0.983051	155	0.548951	77	0.992857	365	0.983146
Ya5NBC51	118	0.508475	155	0.563333	77	0.868421	365	0.688187
Yb8NBC106	118	0.473913	155	0.461538	77	0.283784	365	0.409091
Yb8NBC125	118	0.08547	155	0.11745	77	0.04	365	0.070442
Yb8NBC148	118	0.830435	155	0.517123	77	0.790541	365	0.662465
Yb8NBC157	118	0.725664	155	0.931973	77	0.391892	365	0.617898
Yb8NBC201	118	0.413793	155	0.473333	77	0.486667	365	0.494382
Yb8NBC403	118	0.452174	155	0.793706	77	0.430556	365	0.271588
Yb8NBC419	118	0.234513	155	0.392857	77	0.693333	365	0.412429
Yb8NBC450	118	0.991304	155	0.796552	77	1	365	1
Yb8NBC466	118	0.908257	155	0.506897	77	0.826667	365	0.780899
Yb8NBC479	118	0.761261	155	0.60274	77	0.650685	365	0.763231
Yb8NBC480	118	0.426606	155	0.282313	77	0.046667	365	0.180556
Yb8NBC483	118	0.444954	155	0.148276	77	0.75	365	0.598854
Yb8NBC5	118	0.342105	155	0.493151	77	0.62	365	0.351389
Yb8NBC547	118	0.853211	155	0.411565	77	0.650685	365	0.564607
Yb8NBC568	118	0.458333	155	0.116438	77	0.340278	365	0.281337
Yb8NBC576	118	0.300926	155	0.214286	77	0.636986	365	0.525496
Yb8NBC596	118	0.245455	155	0.414966	77	0.412162	365	0.342618
Yb8NBC636	118	0.594595	155	0.596552	77	0.533784	365	0.697479
Yb9NBC10	118	0.396226	155	0.648936	77	0.333333	365	0.517143
Yb9NBC50	118	0.227679	155	0.263889	77	0.414474	365	0.291176
Yc1NBC53	118	0.527523	155	0.902778	77	0.519737	365	0.74212

Table A1: Allele frequencies of 42 markers for four populations

Locus	African	Asian	Caucasian	Indian
Ya5NBC351	0.0442			
PV92		0.0475		
Yb9NBC50		0.0343		
Ya5NBC45			0.0263	
COL3A				0.0014
TPA25				0.0344
Yb8NBC405				0.0052
Yb8NBC547				0.0372

Note: None of these loci showed deviation from HWE at the p-value of 0.0012 after Bonferroni adjustment.

Table A2: Deviations from Hardy-Weinberg Equilibrium at the 5% level of significance.

Population	No. of pairs of loci deviating from LE (with $p < 0.05$) in 861 LE tests in each population	p-value	
		Minimum	Maximum
African	119	$< 10^{-5}$	0.0499
Asian	64	$< 10^{-5}$	0.0486
Caucasian	37	$< 10^{-5}$	0.0467
Indian	99	$< 10^{-5}$	0.0482

Table A3: Summary Results of Deviations from Linkage Equilibrium at the 5% level of significance.

- Of the 119 deviations from linkage equilibrium in the African population 18 consist of the Ya5NBC351 loci also significantly (with p, 0.05) deviated from HWE.
- Similar comparisons were also seen between deviations from HWE and LE in the other populations: Asians- 18 with PV92, 24 with Yb9NBC50; Caucasians-17 with Ya5NBC45; Indians-41 with COL3A, 11 with TPA25, 21 with Yb8NBC405, 9 with Yb8NBC547
- None of these pairwise tests of LE showed below 5.8×10^{-5} after Bonferroni adjustment.

REFERENCES

1. Ray, David A., et al. "Inference of human geographic origins using Alu insertion polymorphisms." *Forensic Science International* 153.2 (2005): 117-124.
2. Smith, Michael W., et al. "Markers for mapping by admixture linkage disequilibrium in African American and Hispanic populations." *The American Journal of Human Genetics* 69.5 (2001): 1080-1094.
3. Smith, Michael W., et al. "A high-density admixture map for disease gene discovery in African Americans." *The American Journal of Human Genetics* 74.5 (2004): 1001-1013.
4. Butler, John M. *Forensic DNA typing: biology, technology, and genetics of STR markers*. Academic Press, 2005.
5. Clayton, T. M., et al. "Analysis and interpretation of mixed forensic stains using DNA STR profiling." *Forensic Science International* 91.1 (1998): 55-70.
6. LaRue, Bobby L., et al. "Characterization of 114 insertion/deletion (INDEL) polymorphisms, and selection for a global INDEL panel for human identification." *Legal Medicine* 16.1 (2014): 26-32.
7. Turakulov, Rust, and Simon Easteal. "Number of SNPS loci needed to detect population structure." *Human heredity* 55.1 (2002): 37-45.
8. Batzer, Mark A., and Prescott L. Deininger. "Alu repeats and human genomic diversity." *Nature Reviews Genetics* 3.5 (2002): 370-379.

9. Mullaney, Julienne M., et al. "Small insertions and deletions (INDELs) in human genomes." *Human molecular genetics* 19.R2 (2010): R131-R136.
10. Okada, Norihiro. "SINEs." *Current opinion in genetics & development* 1.4 (1991): 498-504.
11. Batzer, Mark A., et al. "African origin of human-specific polymorphic Alu insertions." *Proceedings of the National Academy of Sciences* 91.25 (1994): 12288-12292.
12. Hamdi, Hamdi, et al. "Origin and phylogenetic distribution of Alu DNA repeats: irreversible events in the evolution of primates." *Journal of molecular biology*289.4 (1999): 861-871.
13. Roy-Engel, Astrid M., et al. "Alu insertion polymorphisms for the study of human genomic diversity." *Genetics* 159.1 (2001): 279-290.
14. Watkins, W. Scott, et al. "Genetic variation among world populations: inferences from 100 Alu insertion polymorphisms." *Genome research* 13.7 (2003): 1607-1618.
15. Weir, Bruce S. *Genetic data analysis. Methods for discrete population genetic data*. Sinauer Associates, Inc. Publishers, 1990.
16. Ding, Lili, et al. "Comparison of measures of marker informativeness for ancestry and admixture mapping." *BMC genomics* 12.1 (2011): 622.
17. Hubisz, Melissa J., et al. "Inferring weak population structure with the assistance of sample group information." *Molecular ecology resources* 9.5 (2009): 1322-1332.