The author(s) shown below used Federal funds provided by the U.S. Department of Justice and prepared the following final report:

Document Title:	**Development of a SNP Assay Panel for Ancestral Origin Inference and Individuals Somatic Traits**
Author(s):	**Daniele Podini**
Document No.:	**249206**
Date Received:	**October 2015**
Award Number:	**2009-DN-BX-K178**

This report has not been published by the U.S. Department of Justice. To provide better customer service, NCJRS has made this federally funded grant report available electronically.

Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

- Final Report -

Submitted to:	The National Institute of Justice
Project Number:	2009-DN-BX-K178
Project Title:	Development of a SNP Assay Panel for Ancestral Origin Inference and Individuals Somatic Traits
PI info:	Dr. Daniele Podini, Assistant Professor Email: podini@gwu.edu Tel: 1-202-242-5766
Submitted by:	Daniele Podini
Submission date:	March 29th 2013
DUNS:	043990498-0003
EIN:	530196584
Recipient:	The George Washington University, 2121 I St. NW, Washington, DC 20052
Recipient Id. #:	31024-1-CCNS20768F
Grant Period:	January 1st 2010 – December 31st 2011, Extended at no cost to October 30th 2014

(Daniele Podini)

Abstract

When an STR DNA profile obtained from crime scene evidence does not match identified suspects or profiles from available databases, further DNA analyses targeted at inferring the possible ancestral origin and phenotypic characteristics of the perpetrator (i.e. hair color, skin color and eye color) could yield valuable information. Single Nucleotide Polymorphisms (SNPs), the most common form of genetic polymorphisms, have alleles associated with specific populations and/or correlated to physical characteristics. The objective of this project was to (1) identify SNPs in the literature that could provide information on ancestry and pigmentation, (2) obtain samples from volunteers collecting data on ancestry and pigmentation, (3) type the selected SNPs, (4) select a sub-panel of SNPs providing the maximum amount of information, and (5) develop an assay that can be successfully used to type forensic samples.

We have used single base primer extension (SBE) technology to develop a 50 SNP assay designed to predict ancestry among the primary U.S. populations (African American, East Asian, European, and Hispanic/Native American), as well as pigmentation phenotype among Europeans. We have optimized this assay to a sensitivity level comparable to current forensic DNA analyses, and shown robust performance on forensic-type samples. In addition, we developed a prediction model for ancestry in the U.S. population, based on the random match probability and likelihood ratio formulas already used in forensic laboratories. Lastly, we evaluated the biogeographic ancestry prediction models using a test set, and we evaluated new and existing models for eye color among Europeans with our U.S. individuals of European origin. Using these models with recommended thresholds, the 50 SNP assay provided accurate ancestry information in 98.6% of the test set samples, and provided accurate eye color information in 61% of the European samples tested (25% were determined inconclusive and 14% were incorrect). The assay, which uses equipment already available in forensic DNA laboratories, is recommended for use in U.S. forensic casework to provide additional information about the donor of a DNA sample when the STR profile has not been linked to an individual.

Table of Contents **Page**

Abstract…….….. 2

Table of Contents…… 3

Executive Summary…….……... 4

1. Introduction…… 15

2. Methods…….. 19

3. Results…… 38

4. Conclusions…… 55

5. References…….. 58

6. Dissemination of Research Findings…….…. 63

7. Work preformed during grant extension: July 2013 – July 2014 …….. 64

8. Appendix ……..…….. 97

Executive Summary

Introduction

A composite profile from a battery of ancestry and phenotype informative single nucleotide polymorphisms can provide an estimate of physical appearance [1], which could be valuable to a criminal investigation [2,3]. The single base primer extension (SBE) technique allows for the simultaneous typing from one to over 30 SNPs [4], and robust results can be obtained from a broad range of typical forensic samples. The objective of this research is to develop an assay for combined ancestry and phenotype inference using SNPs that can be processed with the same equipment currently used in crime laboratories for STR testing.

As previously reported [5], we collected samples from 276 individuals along with ancestry information and phenotype data (including eye color, hair color, and skin spectrophotometer measurement). We then screened these individuals, along with 175 in-house samples (ancestry information only), with 11 SBE assays composed of 103 SNPs found in the literature that were either ancestry informative (optimized for U.S. populations) or phenotype informative, or both (one SNP serving both purposes). Then we added an additional 3,989 samples from available databases with varying SNP coverage, and performed several statistical analyses to identify an efficient SNP panel for ancestry and phenotype inference. These analyses included multinomial logistic regression models (using Stata v.11, College Station, TX) for pigmentation phenotype in Europeans, after the method described by Liu *et.al.* and Walsh *et.al.* [6,7]; Principle Component Analysis for pigmentation phenotype in Europeans (using Statistica 9, Factor Analysis module, Statsoft, Tulsa, Oklahoma); X^2 analysis, pairwise F_{ST} analysis, and web-based Snipper analysis [8] for ancestry. All ancestry analyses were designed to evaluate how well each SNP separated the four primary U.S. populations: African American, East Asian, European, and Native American (with Hispanic primarily being a mixture of the latter two). See Appendix Table 1a through 1d for results of these analyses.

By cross-referencing each of these analyses and paying particular attention to SNPs for which published prediction models already exist [7,9], we defined 50 SNPs that we expect to be most predictive of ancestry in the U.S., pigmentation phenotype in Europeans, or both. The resulting list (Table 1) includes 19 ancestry informative markers (AIMs) and 31 phenotype informative markers (PIMs) for pigmentation, 13 of which also have a strong association to ancestry.

Herein, we describe the optimized 50-SNP SBE assay (composed of three multiplexes) that can be implemented in a crime laboratory setting. Additionally, we present a method which uses a subset of these SNPs for ancestry classification in the four primary U.S. populations, including an evaluation of this method with a test set of individuals from each of these populations. This assay is a tool that can aid investigators by providing ancestral and phenotypic background in cases when an STR profile obtained from evidence collected at a crime scene does not match any of the suspects, or any of the profiles in the available databases.

SNP ID	Designation	Chromo-some	Gene (if applicable)	Population differentiated by SNP (based on pairwise Fst)	SNP ID	Designation	Chromo-some	Gene (if applicable)	Population differentiated by SNP (based on pairwise Fst)
rs10007810	AIM	4	*LIMGH1* int.	African American	rs260690	AIM	2	*EDAR*	European
rs10108270	AIM	8	*CSMD1*	African American	rs26722	PIM	5	*SLC45A2*	
rs1042602	PIM	20	*TSIP*		rs2714758	AIM	15		African American
rs1126809	PIM	11	*TYR*		rs2814778	AIM	1	*DARC* 5'UTR	African American
rs11547464	PIM	16	*MC1R*		rs3737576	AIM	1		Native American
rs12203592	PIM	6	*IRF4*		rs3784230	AIM	14	*BRF1*	African American
rs12821256	PIM	12	*KITLG*		rs3827760	PIM (& AIM)	2	*EDAR*	East Asian/Native American
rs12896399	PIM	14	*SLC24A4*		rs4778138	PIM (& AIM)	15	*OCA2*	European
rs12913832	PIM (& AIM)	15	*HERC2*	European	rs4778241	PIM	15	*OCA2*	
rs1344870	AIM	3		Native American	rs4891825	AIM	18	*RAAN* int.	African American
rs1375164	PIM (& AIM)	15	*OCA2* int.	African American	rs4911414	PIM	20	*ASIP*	
rs1393350	PIM	11	*TYR*		rs4911442	PIM	20	*NC0A6*	
rs1426654	PIM (& AIM)	15	*SLC24A5*	European	rs4918842	AIM	10	*HTBP2*	Native American
rs1540771	PIM	6	*IRF4*		rs6451722	AIM	5		African American
rs1545397	PIM (& AIM)	15	*OCA2*	East Asian	rs6548616	AIM	3	*ROBO1* int.	African American
rs1667394	PIM (& AIM)	15	*HERC2*	European	rs714857	AIM	11		European
rs16891982	PIM (& AIM)	5	*SLC45A2*	European	rs7170852	PIM	15	*HERC2*	
rs1800407	PIM	15	*OCA2*		rs722869	AIM	14	*VRK1*	East Asian/Native American
rs1800414	PIM (& AIM)	15	*OCA2*	East Asian	rs730570	AIM	14		European
rs1805007	PIM	16	*MC1R*		rs735612	AIM	15	*RYR3*	East Asian
rs1805008	PIM	16	*MC1R*		rs7495174	PIM (& AIM)	15	*OCA2*	East Asian
rs1805009	PIM	16	*MC1R*		rs885479	PIM (& AIM)	16	*MC1R*	East Asian/Native American
rs1834640	PIM (& AIM)	15	*SLC24A5*	European	rs896788	PIM	2	*RNF144A*	
rs1876482	AIM	2	*LOG442008*	East Asian	rs916977	PIM (& AIM)	15	*HERC2*	European
rs2065982	AIM	13		East Asian/Native American	rs952718	AIM	2	*TBGT12*	African American

Table 1. List of the 50 best SNPs for ancestry and pigmentation prediction identified in this study.

Materials and Methods

50 SNP Assay Development

The 50 selected SNPs were divided into three multiplexes (A: 16plex, B: 15plex and C: 19plex), based on the compatibility of the primers that were designed during the first phase of this project. See Appendix Table 2a through 2d for information on the SNPs in each multiplex.

Optimization of Protocol

Optimization was performed by comparing varying concentrations of PCR reaction components ($MgCl_2$, dNTPs and DNA polymerase) and cycling parameters. The optimized reaction was compared to the AmpFLSTR® Identifiler® Plus (Applied Biosystems, Foster City, CA) reaction mix and cycling parameters. Low volume purification was optimized such that the entire purification product was used in the SBE reaction, which reduces the cost of reagents and consumables. The SBE reaction was optimized by comparing varying reaction volumes and cycling parameters. Both PCR and SBE primer inputs were optimized to maximize balance in the resulting electropherogram peaks.

Sensitivity was tested ranging from 2.5pg to 10ng of input DNA, using a sample quantified via UV-Vis spectrophotometry (NanoDrop 2000, Thermo Scientific). Additional testing was performed on eight highly heterozygous samples, also quantified with UV-Vis spectrophotometry, at 100pg, 150pg and 200pg. The multiplexes were evaluated for robustness with various types of mock forensic samples, all of which had previously yielded STR profiles with AmpFLSTR® Identifiler® Plus.

Bin sets were also developed for each multiplex in order to facilitate data analysis and interpretation in GeneMarker v. 2.4 (Softgenetics, State College, PA) and GeneMapper v. 4.0 software, (Applied Biosystems) (see Appendix Table 3a – 3c); however, these will require adjustment based on polymer used and other laboratory-specific conditions.

Ancestry Interpretation Model

Prior to performing this analysis, it was necessary to evaluate which SNPs were in linkage disequilibrium (LD), because including linked SNPs would inflate the impact of that gene region on the overall ancestry prediction. Linkage was calculated using WGAviewer software [10], which utilizes HapMap genotype data and SNP information (as available) to generate the two common measures of LD (r^2 and D') between each pair of SNPs occurring on the same chromosome. Six of the 50 SNPs are each found on chromosomes where none of the other 50 SNPs are present; therefore these were not evaluated for linkage. Thirty-six of the remaining 44 SNPs were included in the linkage analysis (the remaining eight were not present in the HapMap data set). A conservative review of the linkage disequilibrium analysis reduced the number of SNPs to be included in the biogeographic ancestry prediction to 32 (see Appendix Table 1 for this subset of SNPs).

Of the available genotypes from a combination of samples (some internally tested and some downloaded from the 1000 genome project [11]), a subset of one thousand samples from the four populations of interest was selected using the web-based application Snipper. Under the "Thorough analysis of population data of a custom Excel file" function in Snipper, a set of up to 1000 samples can be evaluated ("verbose cross-validation analysis" function was used) for the success rate of classifying samples into their known population groups. Samples were removed and added in an iterative fashion to determine a subset of samples that were highly predictive of the correct ancestry group, in order to create the most divergence between population groups. The purpose of this "training set" is to allow for comparison of test sample(s) (in terms of forensic casework, the "training set" is the statistical population database, and the test sample is a forensic unknown). The composition of the training set is 266 Europeans, 250 East Asians, 250 African Americans, and 234 Hispanic/Native Americans (See Supplementary Table 3 for training set). Allele frequencies for each of the 32 loci were then calculated within each population.
These frequencies were used to calculate the random match probability (RMP) in all four populations, for all samples in a test set composed of 40 Europeans, 32 African Americans, 35 Hispanics and 32 East Asians. The majority of these test samples (95) were obtained from the National Institute of Standards and Technology; of the remaining 44, 32 were internally available East Asian samples and 12 were recently collected from volunteers (a combination of European and Hispanic individuals). Aside from the 32 East Asian samples, these test set samples had not previously been used for any purpose in this project (neither selection of the 50 SNP panel, nor the development of the training set). The 32 East Asian samples were used in the 50 SNP selection process, and had been evaluated as candidates for and excluded from the training set. The two possible results of this are 1) inflation of RMP values for the 32 East Asian samples, because these individuals helped inform SNP selection and 2) deflation of RMP values for the 32 East Asian samples because these individuals were less predictive of East Asian ancestry compared to the samples chosen for the training set. We expect the latter factor to have a greater effect on the results; therefore, we expect the results for the East Asian test set to be conservative, or statistically lower than we would expect from true unknown forensic samples of East Asian ancestry.
The LR was calculated for each sample by dividing the highest RMP obtained among the four populations by the other three. The number obtained expresses the likelihood that, given a specific profile, the sample originated from the population in the numerator versus the population in the denominator: LR_1 = highest RMP / second highest RMP.

A threshold of 1000 was empirically chosen above which the LR_1 is considered *significant* for a sample to be classified as belonging to a specific population (the one in the numerator) while LR_1 values below 1000 were defined as *inconclusive* (but still informative) between the two populations with highest and second highest RMPs meaning that the individual most likely belongs to one of the two (or both) populations.

Snipper employs the same frequency based approach to calculate RMP/LR values for a single unknown sample. Because it is far simpler to test a large sample set using in-house developed spreadsheets rather than singularly inputting test samples into Snipper, the website was not used in our current analysis. However, the site would be an easy way for a practitioner to predict the ancestry of a forensic sample. We would expect a practitioner to obtain a success rate of classification similar to that described below, using their unknown sample (assuming it is from one of the four primary U.S. populations) and our U.S.-specific training set with the "Classification with a custom Excel file of populations" function in Snipper . The benefit in using Snipper when testing one unknown sample is a user-friendly interface and a clear report of the results.

Other prediction models were evaluated, one based on multinomial logistic regression and one based on the use of the software STATISTICA. Both methods did not perform as well as the RMP/LR method, perhaps because the number of SNPs used in the prediction is drastically reduced. Description of these methods was not included in this executive summary and can be found in the body of the report: sections 2 and 3.

Eye Color Prediction Model

Once ancestry prediction has been established for a sample, eye color predictions among Europeans using a published model [7] can provide additional investigative information. The six SNPs comprising this eye color model are included in the 50-SNP assay; therefore, we were able to use the supplementary excel-based calculator to evaluate this model on the European samples for which we have eye color information (N=196). The results of this calculator are prediction probabilities for blue, brown, or intermediate eye color (where the sum of the probabilities equals one, and the highest number is the predicted eye color). These prediction probabilities were compiled for each individual, and compared to their reported eye color (self reported and confirmed by the individual collecting the sample). The results were evaluated using probability thresholds of 0.5, 0.7 and 0.9, and the accuracy/error rate (known eye color or incorrect eye color being predicted above threshold) was compared to the sensitivity (number of individuals below threshold, considered inconclusive).

As for ancestry prediction another method for eye color prediction model, based on *Chi-squared Automatic Interaction Detector* (CHAID) using the software STATISTICA platform, was tested on the available samples and is described in detail in sections 2 and 3 of this report.

Results

Optimization Results

The best peak balance with the least background was found in a 25µL reaction volume. Evaluation of PCR reaction mixture components showed that increasing DNA polymerase and dNTP input improved results, while the AmpFLSTR® Identifiler® Plus reaction mix performed poorly in comparsion. The multiplexes performed best with increased PCR cycle number (35), 1 min. incubations for denaturation, annealing and extension; annealing temperature of 58°C (PCR primer T_M range from 52°C to 62°C, with the

majority falling between 55°C-59°C); and extension temperature of 72°C. SBE reaction volume evaluation showed an 8ul reaction best balances sensitivity and background. The optimal SBE parameters were 28 cycles with a 55°C annealing temperature.

Recommended Protocol

PCR reaction components in a 25µL reaction include 1xPCR Buffer Gold® (Applied Biosystems), 2.5mM $MgCl_2$ (Applied Biosystems), 0.22mM dNTPs (Roche Diagnostics, Indianapolis, IN), 0.0568mg/ml BSA (Fisher Scientific, Waltham, MA), 4.375 U AmpliTaq Gold DNA Polymerase® (Applied Biosystems), 2µL multiplex-specific PCR primer mix (Integrated DNA Technologies, Coralville, IA; see Appendix Table 2a-c for primer sequences and reaction concentration), and the remaining volume H_20/DNA extract.

PCR amplification (GeneAmp PCR System 9700, Applied Biosystems) proceeded with an initial incubation step of 95°C for 10 minutes; then 35 cycles of 1) 94°C denaturation for 1 minute, 2) 58°C annealing for 1 minute, and 3) 72°C extension for 1 minute; followed by a final extension at 72°C for 10 minutes, and a 4°C indefinite hold.

Unincorporated primers and dNTPs were removed from 2µL of PCR product by adding 5 U Exonuclease I (Thermo Scientific, Waltham, MA) and 0.5 U Shrimp Alkaline Phosphatase (Affymetrix, Santa Clara, CA), plus 0.25µL H_20, in a final volume of 3µL. The enzymatic reaction (9700) proceeded with a 37°C incubation for 70 minutes, followed by a 70°C incubation for 20 minutes. This entire purified product was then used in the single base extension reaction.

The SBE reaction components were 1µL SNaPshot Reaction Mix® (Applied Biosystems), 1µL multiplex-specific SBE primer mix (Integrated DNA Technologies, see Appendix Table 2d for primer sequences and reaction concentration), 3µL H_20, and 3µL purified product, (to reduce consumables, the SBE reaction components can be added directly to the purification tube/plate). The SBE reaction was performed on the 9700 with the following conditions: 96 °C denaturation for 10 seconds, 28 cycles of 1) 55°C annealing for 5 seconds and 2) 60°C extension for 30 seconds, followed by a 4°C indefinite hold.

To prepare samples for electrophoresis, ten microliters of LIZ 120 size standard (Applied Biosystems) was added to 400µl of Hi-Di formamide (Applied Biosystems), and 1µl of sample was added to 10µl of the Formamide/ILS mixture. Samples were electrophoresed on the 3130 Genetic Analyzer (Applied Biosystems), using a 36cm capillary (Applied Biosystems, refurbished from gelcompany Inc.) and POP-7 polymer (Applied Biosystems), with injection parameters of 1.2kV for 16 seconds.

Initial sensitivity testing showed detection of all 50 SNPs at 100pg of input DNA. Further testing with multiple samples, chosen to maximize heterozygosity, revealed that four SNPs (Multiplex A: rs1805008 and rs65488616; Multiplex C: rs1540771 and rs7495174) often contain background and/or low non-specific peaks, which can cause these SNPs to be mis-typed as heterozygotes at or below 200pg of input DNA. Careful evaluation of results and controls is required at or below this level. To minimize stochastic effects, recommended input is 0.5-2ng DNA per multiplex.

The multiplexes performed well with various types of mock forensic samples, including cigarette butts extracted with DNA IQ® (Promega Corporation, Madison, WI), QIAamp DNA Mini Kit® (Qiagen, Hilden, Germany), and Chelex® 100 Resin (Bio-Rad Laboratories, Hercules, CA); mouth area of bottles extracted with DNA IQ® and QIAamp DNA Mini Kit®; and chewing gum extracted with QIAamp DNA Mini Kit®. See Figure 1 for electropherograms showing multiplex performance on a forensic sample and Appendix Figure 1 for additional example electropherograms from the three multiplexes.

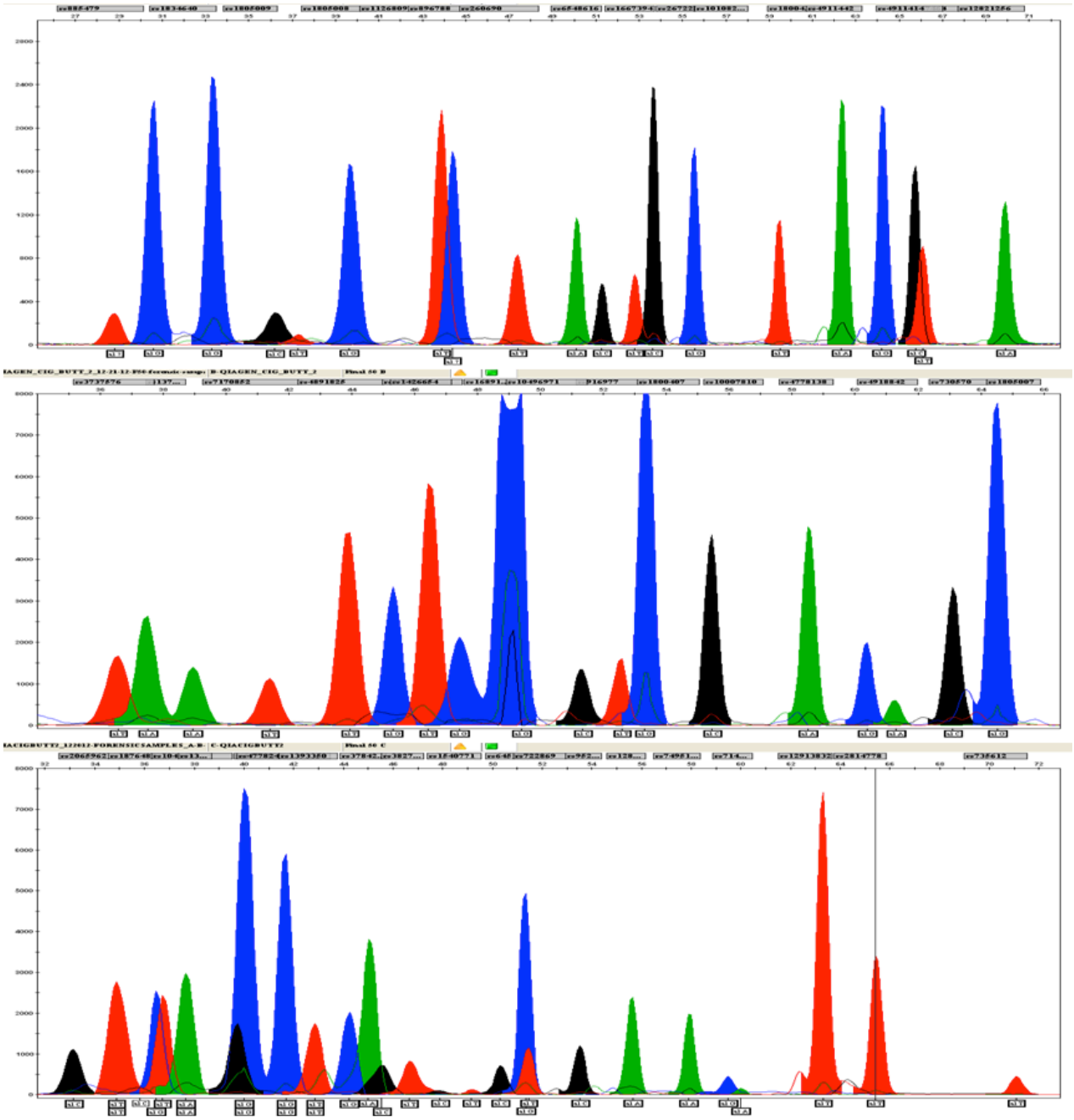


Figure 1. Electropherograms results of the 50 SNP assay (three multiplexes); profile obtained from a cigarette butt.

Ancestry Model Performance

The classification results obtained on the test set are summarized in Figure 2. Out of the 139 samples in the test set, 108 (77.7%) showed a *significant* LR_1 (>1000), and one of these would be predicted incorrectly (classifying as Hispanic/Native American instead of NIST-classified African American). The misclassifying individual has an African mtDNA haplogroup L1c and an African Y chromosome haplogroup E. The remaining 31 (22.3%) individuals had a LR_1 below 1000 and were classified as *inconclusive* between two populations (the highest and second highest RMPs). One of the samples in this category

would be incorrectly predicted as either Hispanic/Native American or European (sample was NIST-classified as African American) because those two populations had the highest two RMPs, while the RMP obtained from the African American population was the third highest. This sample has a mtDNA haplogroup H1a, supporting a maternal European lineage, and a Y chromosome haplogroup E, supporting a paternal African lineage. Overall two individuals out of the 139 would have been incorrectly predicted (1.4%).

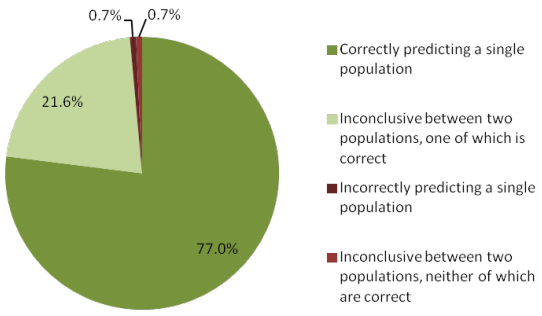


Figure 2. Summary of ancestry model performance. 1.4% of samples were misclassified, 77% were assigned to the correct population, and 21.6% were classified as inconclusive between two populations, where the correct population was one of the two indicated.

Figure 3 graphically summarizes the results obtained from the test sample set. Each data point represents the LR_1 for each sample (Y-axis) sorted low to high on the X-axis. Each color represents a population and the highlighted data-points correspond to the two misclassified individuals, one above and one below the threshold of 1000 (dotted line) both belonging to the African American population.

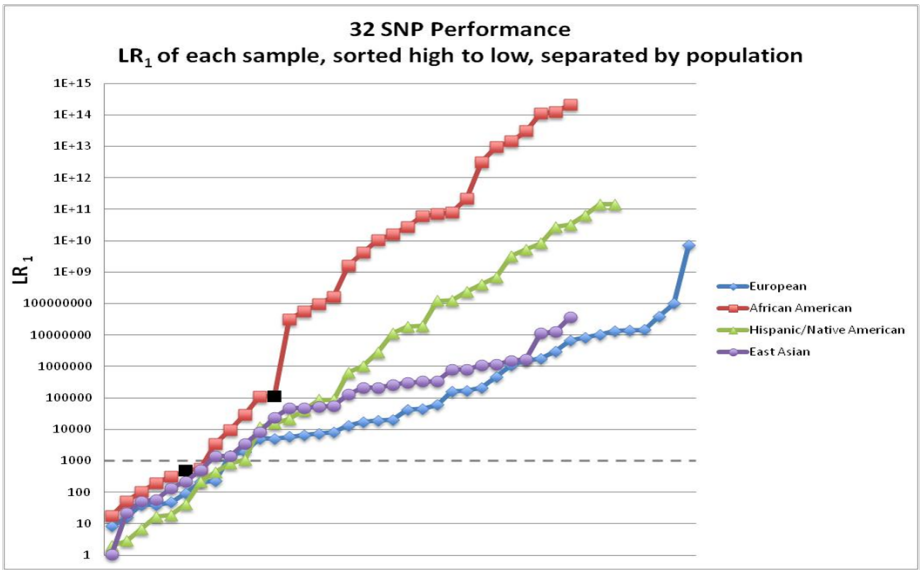


Figure 3: Graphical summary of the results. Each data point represents an individual. The position on the Y-axis corresponds to the LR_1 sorted low to high on the X-axis, the color corresponds to the population of origin and the dotted line corresponds to the 1000 threshold. The highlighted data points represent the two misclassified individuals.

Eye Color Model Performance

Results from testing 196 European individuals for whom eye color information was available in the Irisplex model [7] show an expected trade-off between accuracy and sensitivity, and an overall issue with predicting intermediate eye color. Establishing a threshold below which a prediction probability is inconclusive will aide a practitioner in interpreting and delivering the results of this model. Using a 0.5 threshold, >90% of samples are classified: 96% of individuals with blue eyes and 92% of individuals with brown eyes are correctly classified; however 21% of individuals predicted to have blue eyes actually have an intermediate eye color (green or hazel), and 33% of individuals predicted to have brown eyes actually have blue (N=2) or intermediate (N=20) eye color. At a 0.7 threshold, 75% of samples are classified: 94% of individuals with blue eyes and 67% of individuals with brown eyes are correctly classified; 20% of individuals predicted to have blue eyes actually have an intermediate eye color (67% green and 33% hazel), and 17% of individuals predicted to have brown eyes actually have an intermediate eye color (all hazel). Lastly, at a 0.9 threshold, only 48% of samples are classified: 73% of individuals with blue eyes and 29% of individuals with brown eyes are correctly classified; and the error rates are 17% for blue and 6% for brown. Based on this data set, we do not recommend the 0.5 threshold due to the high error rate, nor do we recommend the 0.9 threshold due to the low sensitivity. The use of a 0.7 threshold allows for eye color prediction in ¾ of European individuals, where 81% of predicted samples are correct and erroneous prediction for blue eyes are most likely be green in color, while erroneous prediction for brown eyes are expected to be hazel. A more conservative option for delivering eye color prediction information to law enforcement would be to define a sample as 'not blue', when predicted to be brown, and 'not brown' when predicted to be blue. With this approach all individuals would be classified correctly with the 0.7 and 0.9 thresholds (Figure 4).

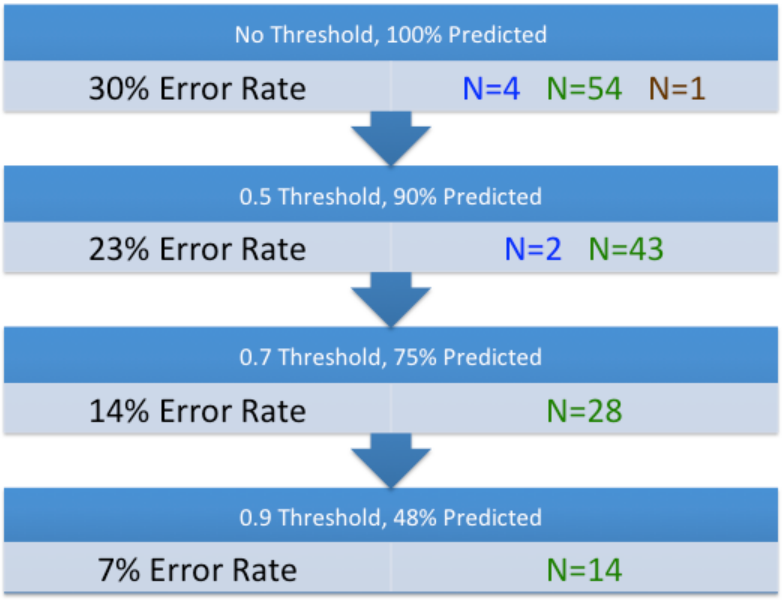


Figure 4. Results for the Irisplex model at various thresholds. The "N" values correspond to individuals with the color-coded known eye color who are erroneously predicted to have a different eye color.

Of note is that the prediction probability for the intermediate eye color never exceeded 0.5, and out of N=56 individuals of known intermediate eye color, the prediction probability was the highest for intermediate in only two individuals. This issue is the primary cause of the error rate in blue/brown prediction, and the same issue was noted in previous work on this model [6], although to a lesser extent. We agree with the authors of the model's hypotheses that this could be due to inconsistencies in phenotype categorization and/or the existence of unidentified variants that could better predict this phenotype.

When evaluating unknown individuals in casework, we recommend this eye color prediction model only be used once an individual is predicted as European (or inconclusive between European and another population). Although we would expect the Irisplex model to predict brown eyes for non-Europeans, and we expect that is true in the vast majority of individuals, the six SNPs in this model were selected based on eye color variation in the European population. It has been shown in the literature that some pigmentation genes show marked divergence from the ancestral genotype in non-European populations where Europeans are largely monomorphic for the ancestral allele [12] and these genes might not have been assessed for this European eye color model; therefore, any non-European individuals with blue or intermediate eye color might not be detected with this model.

Conclusions and Implication for Policy and Practice

In a forensic case where an STR profile has not matched any known individuals or database samples, the unknown sample can be genotyped with this 50 SNP assay to provide predicted likelihood of the four most frequent U.S. populations (African American, East Asian, European, or Hispanic/Native American). By entering the 32 SNP genotypes and the U.S. training set into the web-based application Snipper, a forensic practitioner can quickly generate highly accurate results (employing the aforementioned threshold) in a report format. Additionally, when European ancestry is indicated, the Irisplex model may provide additional eye color information.

This low cost assay can be implemented in any US crime lab as it uses exactly the same technology as conventional STR analysis methods, it generates reliable results with less than 1 ng of template DNA, and it is robust enough to work on typical forensic samples. The information obtained can then be used by investigators, for example, to prioritize suspect processing, corroborate the testimony of a witness to a crime, and overall optimize their resources.

Future Work

In our initial proposal we had planned to collect samples from approximately 200 individuals. During the course of the project we realized that, in order to be able to develop effective prediction models for eye, hair, and skin pigmentation, we needed to significantly increase the number of subjects in the study. After obtaining IRB approval we continued sample collection. We now have collected samples and data from over 300 individuals, which is still insufficient for developing accurate prediction models compared to other studies [7, 9]. Thus, for eye color we relied on a published model [7] and we also continue to collect DNA samples with corresponding ancestry and phenotype information, in order to

test the selected pigmentation phenotype markers in a larger population, develop pigmentation models based on the U.S. population, and evaluate these models in an independent sample. Funding is being sought for a large collection effort of over 3000 individuals. This collection will represent a valuable resource to the forensic science community as it will contain extensive information regarding individuals' ancestry and phenotype, along with skin and hair spectrophotometric measurements (Konica-Minolta CM 2500-d) for melanin content and color determination (see Appendix for data collection tools).

Although the SNP assay published herein contains 50 SNPs, only 35 of those are used in the models presented (32 in the ancestry model, including three eye color model SNPs, plus three additional eye color model SNPs). We are currently evaluating other methods of ancestry prediction, and the possibility of using haplotypes in the ancestry prediction models, to allow for inclusion of linked loci.

In addition, we are investigating the possibility of incorporating a STR-based ancestry likelihood into an overall ancestry model. From the literature it is clear that far less ancestry information is contained in the forensic STR loci compared to AIMs (as these STR loci were chosen for their ability to differentiate individuals, not populations) [13]. However, because the forensic STR profile should already be available by the time an evidence profile is subjected to SNP analysis, it would be worthwhile to incorporate any amount of ancestry association that exists in the STR data.

Works Cited in Executive Summary

[1] M. Kayser, P. de Knijff, Improving human forensics through advances in genetics, genomics and molecular biology, Nat. Rev. 12 (2011) 179–192.

[2] J.M. Butler, B. Budowle, P. Gill, K.K. Kidd, C. Phillips, P.M. Schneider, P.M. Vallone, N. Morling, Report on ISFG SNP panel discussion, Foren. Sci. Int. Supp. Ser. 1 (2008) 471–472.

[3] J.M. Butler, M.D. Coble, P.M. Vallone, STRs vs SNPs: thoughts on the future of forensic DNA testing, Foren. Sci. Med. Pathol. 3 (2007) 200–205.

[4] C. Phillips et al., Inferring ancestral origin using a single multiplex assay of ancestry-informative marker SNPs. Foren. Sci. Int. Genet. 1 (2007) 273-280.

[5] K. Butler, M. Peck, J. Hart, M. Schanfield, D. Podini. Molecular eyewitness: Forensic prediction of phenotype and ancestry, Foren. Sci. Int. Genet. Supp. Ser. 3 (2011), 498-499.

[6] F. Liu, K. van Duijn, J.R. Vingerling, A. Hofman, A.G. Uitterlinden, A.C.J.W. Janssens, M. Kayser, Eye color and the prediction of complex phenotypes from genotypes, Curr. Biol. 19 (2009) R192–R193.

[7] S. Walsh, F. Liu, K.N. Ballantyne, M. van Oven, O. Lao, M. Kayser, IrisPlex: a sensitive DNA tool for accurate prediction of blue and brown eye colour in the absence of ancestry information, Foren. Sci. Int. Genet. 5 (2011) 170–180.

[8] C. Phillips, M. Fondevila, and M.V. Lareau. A 34-plex autosomal SNP single base extension assay for ancestry investigations, DNA Electrophoresis Protocols for Forensic Genetics, Methods in Molecular Biology 830 (2012) 109-126. *URL:* http://mathgene.usc.es/snipper/

[9] W. Branicki, F. Liu, K. van Duijn, J. Draus-Barini, E. Pospiech, S. Walsh, T. Kupiec, A. Wojas-Pelc, M. Kayser, Model-based prediction of human hair color using DNA variants, Hum. Genet. 129 (2011) 443-454.

[10] D. Ge, D. Zhang, A.C. Need, O. Martin, J. Fellay, A. Telenti, D.B. Goldstein. WGAViewer: Software for Genomic Annotation of Whole Genome Association Studies. Gen. Res. 18 (2008) 640-643. URL: http://compute1.lsrc.duke.edu/softwares/WGAViewer

[11] The 1000 Genomes Project Consortium. A map of human genome variation form population-scale sequencing. Nature 467 (2010) 1061-1073. URL: http://browser.1000genomes.org

[12] S. Myles, M. Somel, K. Tang, J. Kelso, M. Stoneking, Identifying genes underlying skin pigmentation differences among human populations, Hum. Genet. 120 (2007) 613-621.

[13] J.S. Barnholtz-Sloan, C.L. Pfaff, R. Chakraborty, J.C. Long, Informativeness of the CODIS STR Loci, J. Foren. Sci. 6 (2005) 1322-1226.

1 Introduction

1.1 Statement of the problem

When a Short Tandem Repeat (STR) DNA profile obtained from evidence collected from a crime scene does not match identified suspects or profiles from available databases it is of no immediate use to the investigators. The objective of this study is to develop a tool that can aid investigators by providing ancestral and phenotypic information in such cases: a Single Nucleotide Polymorphisms (SNPs) assay, able to generate interpretable data on forensic samples, that can be processed with the same equipment currently used in crime laboratories for STR testing. Such a tool can aid investigators in prioritizing suspect processing, corroborating witness testimony, determining the relevance of a piece of evidence to a crime, and ultimately increase the ability to identify individuals related to the crime scene.

1.2. Literature citations and review

Current forensic DNA testing for human identification (HID) purposes is based on the ability to generate a DNA profile from biological samples using STR markers. This has become a routine procedure and is an important tool in criminal investigations (Butler 2005). However, while STRs allow a determination of whether a sample is consistent with an existing profile from a database or an identified suspect, the method is not of use in solving a crime when no matches are found and no suspects have been identified. Further DNA analyses targeted at inferring the ancestral origin and the physical characteristics of the perpetrator or involved individuals can be a valuable investigational tool increasing the ability to identify potential suspects.

1.2.1 Forensic SNPs

The completion of the Human Genome and the International HapMap Project has provided the scientific community with a repository of reference information for the human nuclear genome. Identification and typing of SNPs in the nuclear genome has been performed mainly to aid in studies of genetic diseases, however newly identified SNPs could also be valuable to the field of forensic sciences (Butler 2005). Using Kidd's classification, there are four distinct groups of SNPs which are potentially useful in forensic science applications: Ancestry Informative SNPs (AISNPs); Phenotype Informative SNPs (PISNPs); Lineage Informative SNPs (LISNPs); and Individual Identification SNPs (IISNPs) (Butler 2007). A composite profile from a battery of AISNPs and PISNPs may be able to provide an estimate of ancestry and physical morphology. Such a tool would help prioritizing suspect processing, corroborating witness testimony, and help determining the relevance of a piece of evidence to a crime (Butler 2007).

The existing theories surrounding human evolution and population genetics create the framework to support the idea of using DNA polymorphisms to distinguish one population group from the next (Nelson 2007, Vallone 2004). Typing of specific SNP loci, both on the maternally inherited mitochondrial DNA (mtDNA) and the paternally inherited male-only Y

chromosome is an effective way to infer the ancestral origin of a sample (Nelson 2007, Brión 2005) as both genomes are inherited without recombination, though each only provide information about maternal and paternal lineages. Although autosomal AISNPs are subject to greater variation due to recombination, there are several autosomal AISNPs where markedly different population frequencies occur due to an adaptation to a particular environment or other evolutionary forces. For example, the Duffy (*FY*) blood group phenotype FY (A-B-)(homozygous FY*B*) is lacking the receptor for *P. vivax* malaria reducing susceptibility to malarial infection in Sub-Saharan African populations (Hadley 1986). This adaptation to malarial infection only occurs in Sub-Saharan African populations and is useful as an AISNP. Other researchers have shown that using panels of only 10 and 34 autosomal and X linked AISNPs it is possible to consistently obtain high ancestral group classification probabilities for a set of tested samples (Phillips 2007, Lao 2006). More recently online tools have been developed to facilitate practitioners' access and use of data in forensic AISNPs. An example is FROGkb developed by Dr. Kenneth Kidd's group (http://frog.med.yale.edu, Rajeevan 2012) supported by NIJ. The acronym stands for Forensic Research Reference knowledge based, which is an open access resource designed to enable data retrieval and statistical calculation on several forensic relevant SNP panels. Another example is 'The Snipper" app suite (http://mathgene.usc.es/snipper/), which is a web-based application that uses algorithms similar to *STRUCTURE* to calculate likelihood ratios of inclusion of a questioned sample into populations based on a known data set (of up to 1000 individuals) (Phillips 2012).

Phenotype Informative SNPs (PISNPs), are found by sequencing genes coding for proteins that play an important role in determining individual physical characteristics, such as hair, skin and eye color, and skull morphology (Jackson 2006). For example, an important observable trait is an individual's hair, eye, and skin color which depend on the amount, type, and distribution of melanin in these tissues. Melanin is synthesized in melanocytes which are located in the basal level of the skin, the hair bulb and the iris (Parra 2007). Differences in melanocyte density depending on body location have been described (Whiteman 1999), yet these are not sufficient to explain the differences in body pigmentation among individuals. Two factors better explain these differences: the amount and type of melanin and the shape and distribution of melanosomes. There are two types of melanin: eumelanin, brown/black in color, and phaeomelanin, red/yellow in color. The melanocortin-1 receptor gene (MC1R) is involved in the transfer of both types of melanin affecting human hair and skin color (Beaumont 2005). Other examples are the genes associated with oculocutaneous albinism, and iris colors (including but not limited to OCA2, HERC2, MYO5A, AIM, DCT), which provide information on the eye color of an individual (Frudakis 2003).

A project funded by NIJ, on polymorphisms associated to human pigmentation, concluded that six SNPs in five genes (SLC24A5, OCA2, SLC45A2, MC1R, and ASIP) account for a great proportion of hair, skin, and eye pigmentation variations across populations (Brilliant 2008). Furthermore other researchers demonstrated that the SNP rs12913832 (T/C) on HERC 2 predicted eye color very efficiently: individuals carrying the C/C genotype had only a 1% probability of having brown eyes while T/T carriers had an 80% probability of being brown eyed (Kayser 2008). This is consistent with recent study that showed that the HERC2 rs12913832 (T/C) region functions as an enhancer regulating transcription of OCA2, which encodes for the trans-melanosomal membrane protein "P". In darkly pigmented human melanocytes transcription factors HLTF, LEF1, and MITF were detected

binding to the HERC2 rs12913832 enhancer carrying the T-allele. Long-range chromatin loops between this enhancer and the OCA2 promoter lead to elevated OCA2 expression. Whereas, in lightly pigmented melanocytes carrying the rs12913832 C-allele, chromatin-loop formation, transcription factor recruitment, and OCA2 expression were all reduced (Visser 2012).

Based on these data Walsh 2011 reported a multiplex assay for the analysis of a set 6 eye color predictive SNPs called IrisPlex. The assay was then upgraded to the HIrisPlex (Walsh 2012) with the addition of 18 SNPs for the prediction of hair color for a total of 24 markers.

1.2.2 The Single Base Primer Extension method

Single Base Primer Extension (SBE) technique, also known as "minisequencing" (Syvänen 1990), allows for the simultaneous typing from 1 to over 30 single nucleotide polymorphisms (SNPs) scattered throughout the organism's genome (Phillips 2007). Advantages of this methodology include the possibility of typing tetra-allelic SNPs, sensitivity, specificity, robustness, and amenability to automation (Fiorentino 2003). Once the assay is optimized, it generally allows one to obtain robust results over a broad range of both quantity and quality of genomic DNA template. SBE has been applied in several different applications from single cell analysis for pre-implantation genetic diagnosis (PGD), and prenatal and postnatal molecular diagnosis of monogenic diseases, to forensic mitochondrial DNA analysis on highly degraded human remains, and high throughput SNP screening for population studies (Vallone 2004).

The Single Base Primer Extension method is based on an initial multiplex PCR amplification of fragments that can be small (~50 base pairs) as long as the targeted SNP is included in the amplicon (amplified DNA fragment i.e. in between the primer binding sites but not included in the primer sequence). Generally, the smaller the amplified fragment, the greater the amplification efficiency; this is particularly relevant in the situation where the starting template is at very low copy numbers/concentrations and/or the template is highly degraded (Vallone 2004). After the multiplex PCR amplification is performed the reaction product is purified to eliminate unincorporated PCR primers and dNTPs, using a simple procedure with low likelihood of sample contamination and sample mix-up. The single base primer extension reaction then uses the purified PCR product as a template. SBE primers are designed similarly to a standard sequencing primer. The SBE primer binds in a 5'→ 3' orientation to the PCR amplicon with the 3' end of the primer adjacent to the SNP of interest. The second sequence specific annealing step adds further specificity to the assay. The SNaPshot® reagent kit contains buffer, polymerase, and fluorescently labeled dideoxynucleotides (ddNTP) (one dye for each nucleotide). During thermal cycling the SBE primer binds to the PCR amplicon and the appropriate ddNTP is incorporated at the SNP site (Figure 5). Following the SBE reaction, samples are further purified to eliminate unincorporated labeled ddNTPs that would interfere with data analysis again using technology with a low likelihood of contamination and sample mix-up, and loaded onto a capillary electrophoresis (CE) instrument. The electropherograms generated can then be analyzed using commercially available programs already commonly used in crime laboratories for STR analysis. Customized macros can then be created to facilitate data interpretation, processing, and management. Once optimized the assay is, sensitive, robust, simple to perform, and amenable to automation. Multiplexing of a SBE assay is

accomplished by adding a non-binding tail sequence to the 5' end of the SBE primer. The tail is typically a poly-T or a repeating AGCT sequence. The total length of SBE primers can range from 20 - 80 nucleotides (the length is somewhat defined by limitations in the automated synthesis of DNA oligomers). Each SBE primer is usually separated in size by 3 - 4 nucleotides to ensure resolution on a gel or capillary detection platform. An assay based on PCR amplification followed by SBE methodology can be used to type DNA evidence collected at a crime scene, and can be processed on DNA analysis hardware conventionally used in forensic DNA laboratories. (ABI 310, 3130, and 3500).

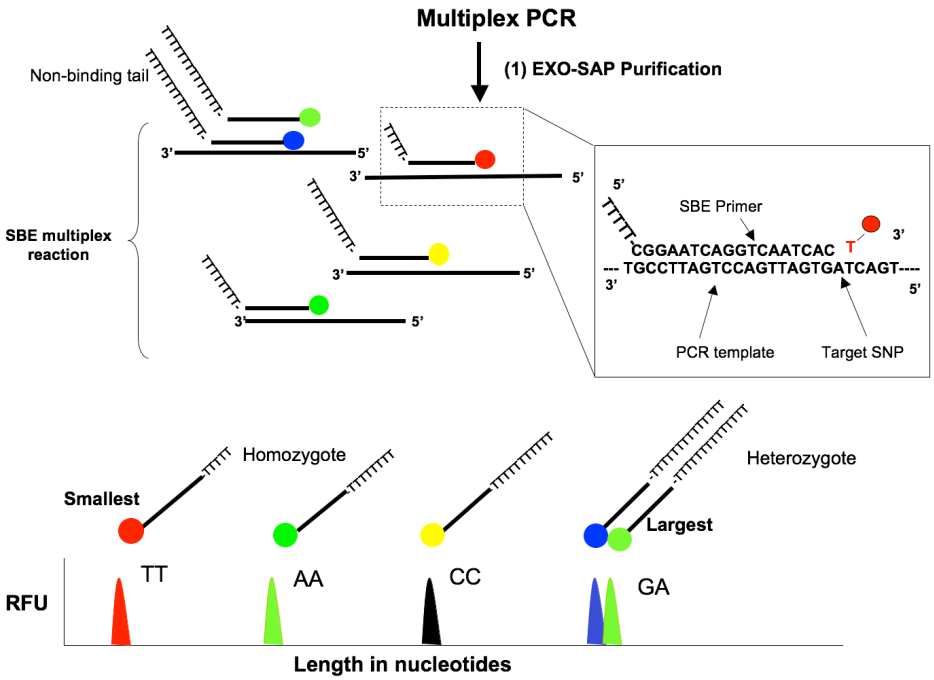


Figure 5: Schematic representation of the SBE assay. Initial multiplex PCR amplification is performed targeting the flanking regions of the SNPs. Following amplification samples are purified to eliminate unincorporated PCR primers and dNTPs. The single base primer extension reaction then uses the purified PCR product as a template. SBE primers bind in a 5'→ 3' orientation to the corresponding PCR amplicon with the 3' end of the primer adjacent to the SNP of interest and the appropriate ddNTP is incorporated at the SNP site. Following the single base extension reaction samples are purified to eliminate unincorporated labeled ddNTPs and then loaded on a CE apparatus. In this example the targets are 4 diploid loci of which the first three (left to right) are homozygous and the forth one (the largest) is a heterozygote (G/A). Note that the migration of the SBE primers is affected by the specific dye attached by the incorporated nucleotide. The two alleles, although having the same number of bases, exhibit different electrophoretic mobility and appear as two separate peaks.

1.3. Statement of hypothesis or rationale for the research

Our hypothesis is that by collecting DNA samples, biogeographic ancestry, and phenotypic information from individuals from the US population (volunteers), and by screening over 100 of the most informative SNPs that can be identified in the literature, it is be possible to identify a subset of SNPs that can provide ancestry and phenotype predictions. Furthermore using the SBE method it is possible to include the selected SNPs in an assay/s that is as robust and sensitive as the commercial STR kits and can be typed on the same CE platforms conventionally used in US crime labs. Such an assay can provide useful information to investigators in cases where a conventional STR profile did not match any of the suspects and did not *hit* other profiles in the available databases.

2 Methods

This project is divided into two major phases: Phase 1 includes the selection of over 100 candidates SNPs for ancestry and phenotype prediction, sample collection and testing, and selection of the final SNP panel; Phase 2 included the development of an robust and sensitive assay for the selected SNPs, that can produce reliable results on forensic evidence, and the development and testing of prediction models. To facilitate the reader the following section (2.1 Phase 1) describes both the materials and methods an the results that allowed to move to Phase 2. Results for Phase 2 and described in the section 3.

2.1 Phase 1

2.1.1 Candidate SNP Selection

As GWAS and other analyses of ancestry and pigmentation-associated SNPs became available, a list of candidate SNPs was selected from the literature (Lao 2006, Duffy 2007, Stokowski 2007, Sulem 2007, Brilliant 2008, Halder 2008, Han 2008, Kidd 2008, Shekar 2008, Bouakaze 2009, Branicki 2009, Iida 2009, Kosoy 2009, Sturm 2009, Mengel-From 2010). One hundred and eight SNPs were selected, which can provide information on phenotype, ancestry or both. Five SNPs were not genotyped due to sequence incompatibility with the typing method or the existence of paralogous gene regions. Forty-three of the remaining 103 SNPs are considered ancestry markers, 53 are phenotype markers associated with pigmentation, and the remaining seven are associated with other physical characteristics such as hair form or baldness.

2.1.2 Sample Collection

From January 2010 to July 2011, 276 samples were collected from anonymous volunteers in the Washington, DC area using a GWU IRB approved protocol, consisting of the following components:
1) After reading an assent form (Appendix Data Collection Tools), volunteers completed a comprehensive questionnaire (Appendix Data Collection Tools) regarding many aspects of their physical appearance (i.e. height, body build, pigmentation, and hair form) and

including ancestry/phenotype information of their parents and grandparents (when known). While much of this information is relevant to the current project, insufficient genetic association information exists to evaluate some of these traits. Overall, this sample set is a repository of DNA samples and phenotype information that can be used now and in the future to allow for more precise and comprehensive inferences of physical traits of individuals.

2) Pigmentation measurements were collected via spectrophotometry (Konica Minolta CM-2500d). Data was collected in duplicate from the inner wrist, inner forearm, inner side above elbow, and inner side below underarm (avoiding hair, moles, or other discolored areas); from the forehead and cheek (noting if makeup is worn); and, because the spectrophotometer also measures hue, from three areas in the hair (attempting to measure natural hair color, and noting if this is not possible). See figure 6 for relative melanin measurements obtained; generally the face measurements were significantly darker than the arm due to increased UV exposure. Due to most samples being collected at the George Washington University and the desire to avoid facultative pigmentation (suntan), sample collection was suspended during the summer months.

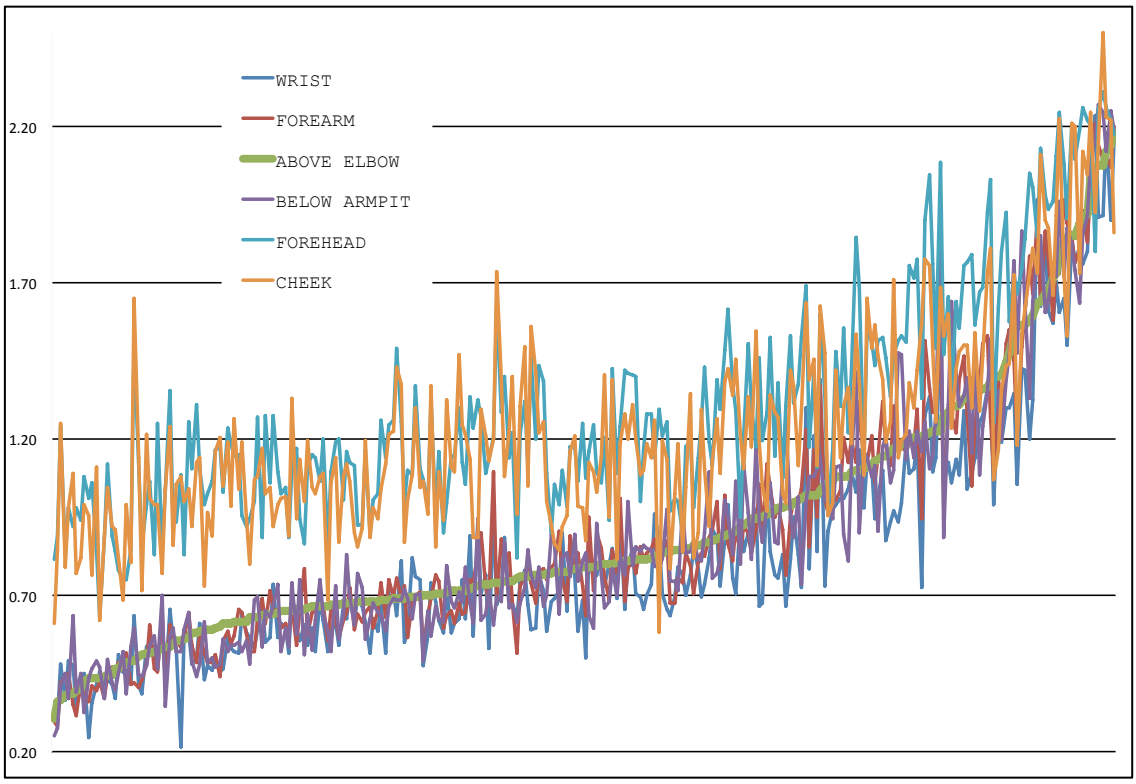


Figure 6. Skin melanin index measurements collected from volunteers, sorted from low to high based on "above elbow" values. The face measurements are consistently higher than arm measurements due to increased UV exposure over time.

3) Three buccal (cheek) swabs were collected.

4) All collected items were labeled with a unique sample code.

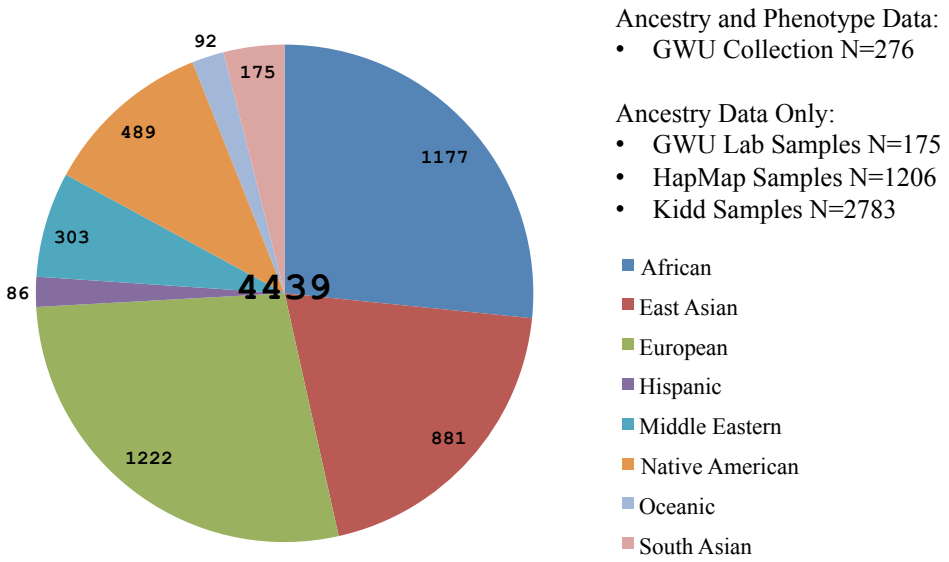
5) The researcher collecting the sample also completed a to ensure complete collection and verify key pieces of self-reported information.

After collection, sample information from questionnaires and spectrophotometer

measurements were entered into a Microsoft Access™ database, facilitated by the creation of an input screen customized to the questionnaire checklist (Appendix data collection tools). In addition, one buccal swab from each sample was extracted with Qiagen® Mini and quantified via Quantifiler™ Human. The remaining two buccal swabs were dried and placed into room temperature storage.

Due to the high proportion of European samples collected from volunteers (71%), additional anonymous DNA samples with known (self-reported) ancestry were obtained from Dr. Moses Schanfield, Department of Forensic Sciences, GWU (samples previously ruled "NOT human subject research" by the GWU IRB). These additional samples (N=175) were a combination of African American, Native American, or East Asian ancestry, and were added to the samples collected, for a total of 451 samples.

To further supplement the ancestry information, genotype data from an additional 2783 samples from varying populations was received for 65 of the 103 candidate SNPs from the laboratory of Dr. Ken Kidd, Yale University. Lastly, all available HapMap data for the 43 ancestry SNPs was downloaded, and this included varying levels of data for 1206 samples from 11 populations. See Figure 7 for complete breakdown of samples sources, and



information on the available data.

Figure 7: Sample breakdown by ethnicity, and sample sources

2.1.3 SNP Genotyping

The selected SNPs were typed with the SBE method described in section 1.1.2.

Eleven SBE multiplexes were developed and optimized for the candidate SNPs, and the combined set of 451 samples was genotyped for 101 SNPs. Two of the candidate SNPs (*rs3829241* and *rs6119471*) failed to genotype and were eliminated during this phase. Figure 8 shows electropherograms of a sample analyzed with five of the multiplexes developed to genotype the candidate SNPs.

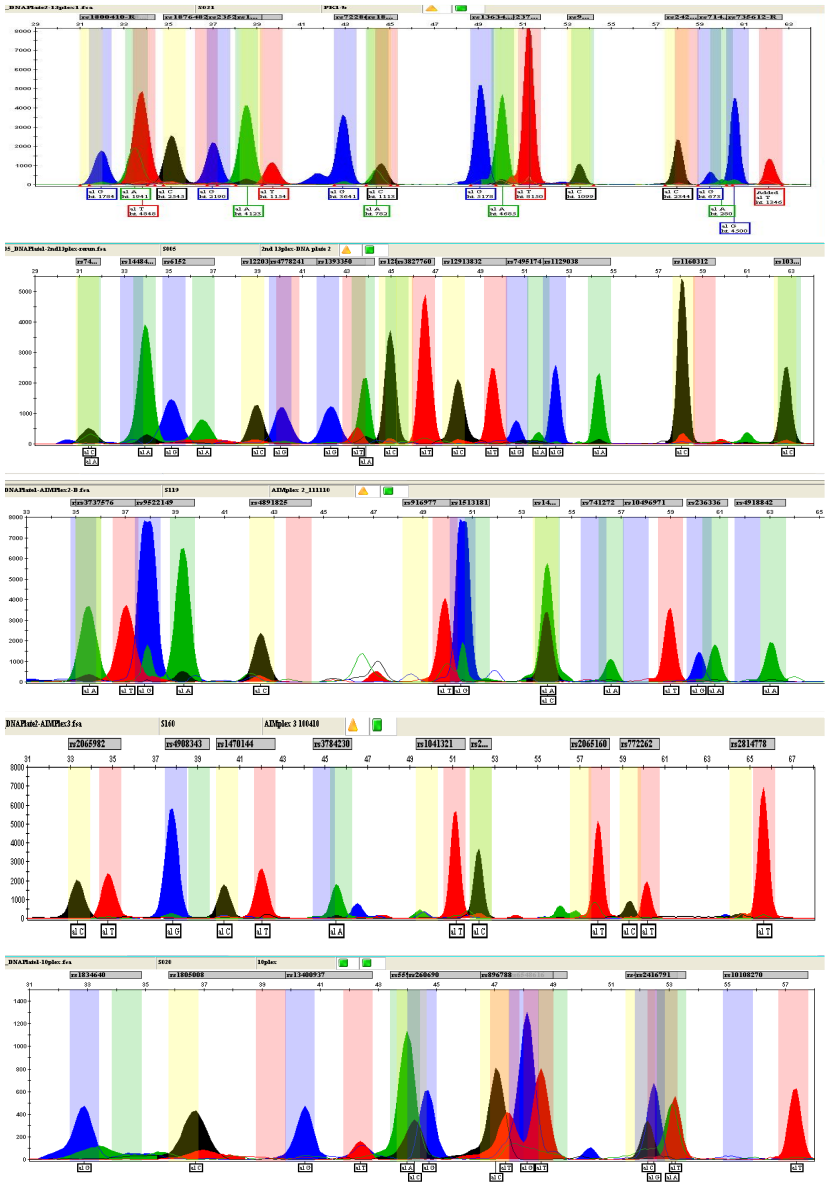


Figure 8: Examples of five SNP multiplexes that were used to screen volunteer samples.

2.1.4 Candidate SNP Evaluation / Reduction

The genotyped SNPs were evaluated for their ability to predict a specific physical trait (or to discern between distinct traits, for example light-colored vs. dark-colored iris) or the ancestral origin of an individual. Referring back to the previously mentioned examples, *rs12913832* shows the expected strong association between the G homozygote genotype and the blue eye phenotype and *rs2814778*, where the C allele represents an adaptation to presence of malaria, occurs predominantly in African or African American individuals (Figure 9). These SNPs are clear choices for the final assay; however, most of the candidate SNPs required a multi-factorial evaluation in order to select a panel that best balance ancestry prediction in the four U.S. populations of interest (African American, East Asian, European, and Hispanic/Native American), and potential phenotype prediction.

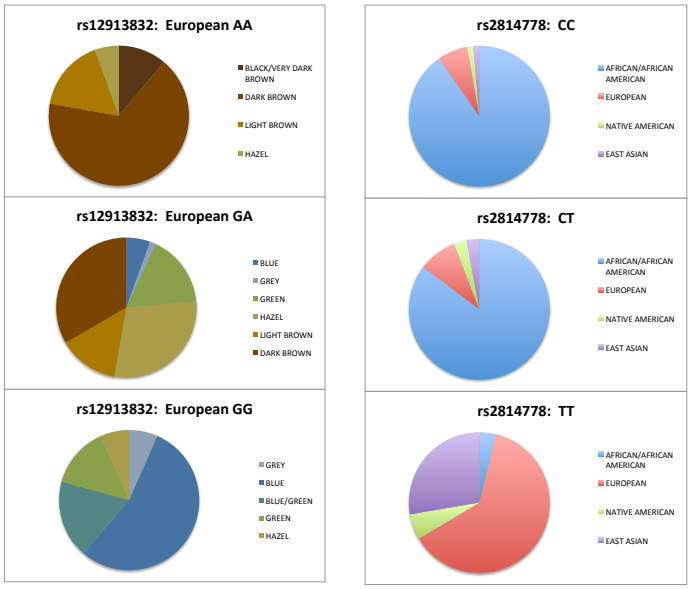


Figure 9. (left) *rs12913832*, Of 196 Europeans with phenotype data available, homozygous A individuals (9%) have brown eyes; whereas homozygous G individuals (54%) have light colored eyes. The remaining individuals (37%) are heterozygous and present both phenotypes. (right) *rs2814778*, the C allele represents an adaptation to malaria, thus the presence of at least one C allele is indicative of sub-Saharan African ancestry or admixture. Out of 395 individuals tested in the four populations of interest, 90% of homozygous C individuals were African American or African, 85% of heterozygous C/T individuals were African American, and only 4% of

Many phenotype SNPs also contain ancestry information; therefore, the ideal SNPs will have a dual role (for example, a genotype can be indicative of both European ancestry and blue eyes). Methods of evaluation for SNP ancestry content included X^2 analysis, Snipper (web-based program) divergence ranking (Phillips 2012), and pairwise F_{ST} analysis. Methods of analysis for pigmentation phenotype included X^2 and principle component analyses for eye, skin and hair color in Europeans. There were an insufficient number of samples with known phenotype to

evaluate pigmentation in non-Europeans or to evaluate the balding phenotype SNPs (*rs6152* and *rs6625163*).

2.1.5 Materials and Methods for best Ancestry SNP selection

X^2 Analysis: This analysis evaluated the 99 remaining SNPs in relation to ancestry for the four populations of interest using a X^2 analysis. To facilitate evaluation of results, the ancestry SNPs were ranked from lowest p-value (most divergent SNP) to highest p-value.
PCA: Another approach was to analyze the data with Principal Component Analysis (*STATISTICA Data Miner* software) in order to identify SNPs accounting for high levels of variance in the data, and eliminate less informative ones. This method determines the best ancestry (or phenotype) SNPs by taking the individual population results and converting them to sample population frequencies, then performing principle components analysis on the array of populations and individual allele frequencies. The analysis generates a series of uncorrelated variables that maximally extract information from all of the data points and between populations. This provides a rapid method to determine if specific alleles are correlated, redundant or non-informative. Further, it will yield information as to which SNP alleles have the highest correlation (factor loading) with the highly informative synthetic variables. This allows for a rapid reduction in the number of SNP that need to be used, and provides significantly more information content than traditional F_{ST} analysis of between and within group variation.

Data for the 43 ancestry SNPs was divided into eight categories for PCA. The placement of smaller ethnic groups into larger categories was verified using *STRUCTURE 2.3.1.* This heuristic algorithm assigns individuals, based on their genotype data, to one or more of a user-defined number of categories (Pritchard 2000).

Snipper Analysis: A web-based application called *Snipper* (Phillips 2012) was also used to aid in narrowing down the SNP list for ancestry prediction, both by ranking all SNPs based on each SNP's divergence level (ability to separate the dataset into the four populations of interest), and by evaluating the frequency of misclassification with different SNP sets. To perform this analysis, samples from the four populations of interest with genotyping results at all 99 loci (N=389) were uploaded. Then, the "verbose cross-validation" function was selected with all SNPs included in the analysis.
F_{ST} Analysis: The SNP data was also evaluated for ancestry content using F statistics. These statistics, based on the theory that subdividing a population leads to a decrease in heterozygosity, use observed and expected heterozygosity levels to estimate genetic differentiation. For all genotyped SNPs, we performed pairwise F_{ST} analysis (pairs included African/African American—European, African/African American—East Asian, East Asian—European, East Asian—Native American, and Native American—European), which compares allele frequencies and levels of heterozygosity in the subpopulation to the total of the two populations. Performing this in a pairwise fashion allows for determining the SNPs that best differentiate any two populations. Significance was evaluated with X^2 testing using the harmonic mean, at α=0.001 with one degree of freedom. Pairwise Euclidian distance was also calculated (simply calculating differences in allele frequencies between populations); and while these results were usually consistent with the F statistic results, the latter calculation is a more informative distance measure.

2.1.6 Materials and Methods – Pigmentation in Europeans

X^2 Analysis: This analysis evaluated the 99 SNPs in relation to the specific phenotypes of eye, skin and hair color in those of European descent using a X^2 analysis. After Bonferroni correction for multiple testing, the p-value for statistical significance was less than 0.01. Table 2 shows the categorization of phenotypes for this analysis.

Eye color	**Skin color**	**Hair color**
Blue, blue/green, grey	Light (melanin index 0.30 – 0.65)	Black
Green/hazel	Medium (melanin index 0.66 – 0.95)	Brown
Brown	Dark (melanin index 0.96 – 1.28)	Blonde
		Red

Table 2. Phenotype categorization in Europeans. Melanin index was measured on inner arm, above elbow.

PCA: Because many of the pigmentation SNPs are also highly associated with ancestry, when grouping and analyzing a diverse data set based on varying pigmentation, PCA may give high levels of significance to SNPs strongly associated with ancestry while these SNPs may have little influence on pigmentation. To overcome this, PCA analyses for pigmentation were performed among all four populations and within European populations only. The latter analyses were used for candidate SNP reduction.

All samples for which phenotype data was available (N = 276) were categorized into hair color groups (black, dark brown, light brown, dark blonde, light blonde, and red/ auburn), eye color groups (brown, blue, other), and skin color groups (melanin indices from inner arm above elbow, where light = minimum–0.89, medium = 0.90–1.49, and dark = 1.50–maximum). Then, samples of European ancestry for which phenotype data was available (N=196) were categorized as before for hair and eye color. The categorized data was subjected to PCA using the 53 pigmentation SNPs.

PHASE: In two gene regions that impact pigmentation, *MC1R* and *OCA2/HERC2,* there were many candidate SNPs that might be linked (10 and 19 SNPs, respectively). To account for this, the program PHASE version 2.1 was used to generate haplotypes from the genotype data (Stephens 2003) and to evaluate recombination (Crawford 2004). All samples with genotype data in these gene regions were divided by ethnicity: European, African/African American, and East Asian. Samples that did not fall into one of these categories were not included in this analysis. PHASE analysis was performed in each population for 1) the 10 *MC1R* SNPs, 2) the first 10 of 19 *OCA2/HERC2* SNPs, 3) the last 10 of 19 *OCA2/HERC2* SNPs, for a total of nine analyses (NOTE: *OCA2/HERC2* SNPs were divided, with one overlapping SNP in each analysis, due to insufficient computational ability to analyze all 19 SNPs together). The analyses included the settings of 10,000 iterations with a 1000 iteration burn-in period, and a thinning interval of 1. The inferred haplotypes within regions where recombination is unlikely were then evaluated to determine which SNPs are definitive of the haplotype and/or appear to be associated with pigmentation.

2.1.7 Results for best Ancestry SNP selection

X^2 results: As seen in Appendix Table, this analysis found (as expected) that all of the 43 ancestry SNPs were strongly associated with ancestry ($p<10^{-10}$). These results were ranked by significance to loosely define those SNPs most predictive of ancestry. Further, X^2 analysis showed that 55 of the 60 phenotype SNPs were also strongly associated with ancestry.

PCA Results: A subset of 25 SNPs with the highest factor loading was selected from the 43 ancestry SNPs. The ability of this subset to diverge the populations of interest was evaluated with STRUCTURE 2.2 software analysis, a population genetics and anthropology software package based on Bayesian statistics, developed to analyze the genetic composition of individuals and populations. Figure 10 shows the results of a STRUCTURE analysis performed initially with the 43 ancestry SNPs. After ranking the SNPs with PCA, the same analysis was performed with the best 25 AIMs, first with K=4 then with K=5. Results indicate that the predominant ethnic groups in the United States (European American, African American, Asian and Hispanic) can still be well-differentiated with the subset of 25 AIMs.

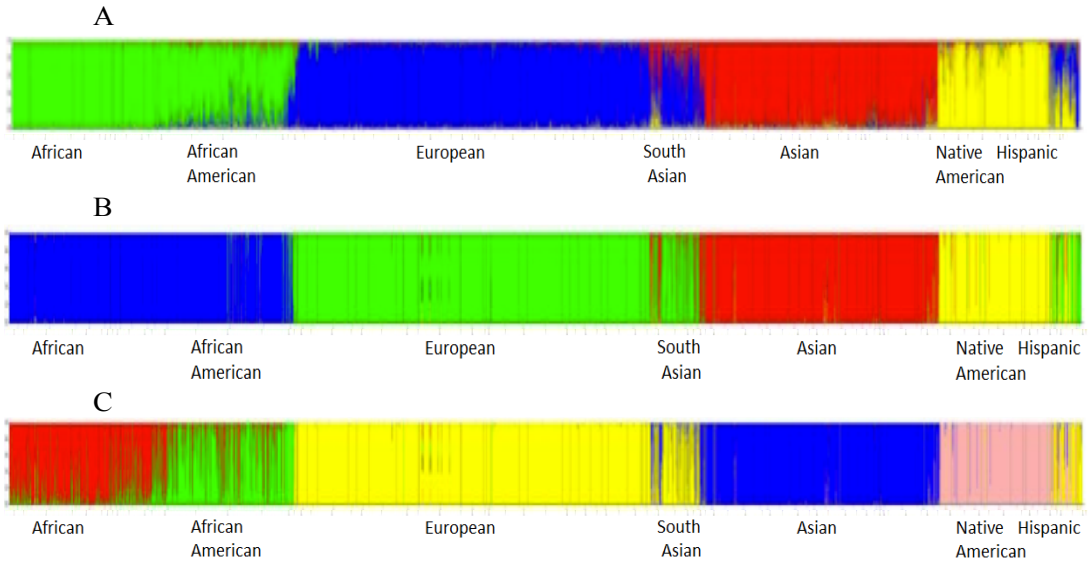


Figure 10: Structure plots (A) 43 AIMs K=4, (B) 25 AIMs K=4, and (C) 25 AIMs K=5 analyzed on 4440 individuals from multiple populations. The 25 AIMs were selected from the 43 with Principal Component Analysis (PCA *STATISTICA Data Miner* software).

Snipper results: This analysis produced a ranked list of divergence for each SNP (1 being the most divergent SNP and 99 being the least divergent), seen in Appendix Table. The output also shows how successful the 99 SNPs are in classifying each sample into its known population. The success rate for African/African American, East Asian, and European are all over 90%; however, the rate is lower for Native Americans (81%). There were three misclassified Native Americans, all classified as Europeans. This could be caused by the small number of Native Americans in the analysis (N=16), a failure to include SNPs that sufficiently distinguish Native Americans from Europeans, or the complicated nature of this admixture (e.g. the self-reported ancestry is Native American but the Native American component of the individual's genome is relatively small).

F_{ST} results: In Appendix Table, the pairwise F_{ST} values are shown. This analysis is very beneficial in choosing a SNP panel because, as opposed to other methods that give general rankings, the pairwise F_{ST} shows which population can be distinguished by each SNP (because these SNPs are biallelic, typically one SNP distinguishes one population from all of the others). Using the previously cited example of *rs2814778*, the pairwise F_{ST} results show this SNP to be excellent at distinguishing African/African Americans from Europeans and from East Asians (0.815 and 0.841, respectively). This analysis is also key in determining which SNPs can distinguish Native American individuals from East Asian individuals. A disproportionate number of candidate SNPs were chosen for this purpose, under the hypothesis that the ability of the final panel to distinguish U.S. Hispanic individuals from the other populations is dependent upon identifying Native American-predictive SNPs. The relatively low pairwise F_{ST} values seen in the East Asian-Native American column of the table (highest value is 0.517), indicates this will be a more difficult separation. It is interesting to note that, for our primary groups of interest (African/European/East Asian), the phenotype markers are more "ancestry informative" than the ancestry markers.

2.1.8 Results for best Pigmentation SNP selection

X^2 results: This analysis showed a significant relationship for European eye color with 17 SNPs, European skin color with 11 SNPs and European hair color with 17 SNP at the α=0.05 significance level. Using a stricter significance criteria of α =0.005, associations remain for 12, 4 and 10 SNPs, respectively. Several SNPs showed weaker evidence of a relationship with a p-value between 0.5 and 0.15. While many SNPs appeared associated with only one of the phenotype, several others showed significance across the board. Specifically, *rs12913832* (previously described) and *rs1129038*, both located in the HERC2 gene, were highly significant for all three phenotypes.

Table 3 lists those SNPs that showed significant associations with ancestry in the entire cohort and with eye, skin and hair color in the European cohort (results for all SNPs evaluated can be found in Appendix Table 1a-1d).

SNP	Category	European Eye Color p-value	European Skin Color p-value	European Hair Color p-value
rs1129038	PIM	$1.630E^{-15}$	0.001	0.008
rs12913832	PIM	$1.430E^{-15}$	0.001	0.016
rs16891982	PIM	$1.440E^{-06}$	0.051	0.002
rs1805008	PIM	0.167	0.011	$4.610E^{-06}$
rs1805009	PIM	0.119	0.006	0.003
rs2238289	PIM	$1.440E^{-13}$	0.048	0.288
rs2352476	AIM	0.026	0.050	0.004
rs26722	PIM	0.018	0.115	0.001
rs7495174	PIM	$1.070E^{-04}$	0.004	0.592

Table 3. SNPs showing significance with more than one phenotype

PCA Results: Analysis of all samples combined showed excellent genetic discrimination of the eye, skin, and hair color groups; however, it was unclear which SNPs were actually associated with pigmentation, as opposed to being indicative of ancestry. Performing the analysis on samples of European ancestry only provided a more informative analysis. The hair color analysis exemplifies this well: as seen in the results for all groups (Figure 11), the black hair color is separated the farthest from all other hair colors but when analyzing Europeans only (Figure 12), the black hair color clusters more closely with the other hair color categories. The difference between these two plots is due to the ancestry component of the SNPs causing increased divergence of individuals of African or Asian descent.

By analyzing the PCA weighting for each SNP within Europeans, a subset of 20 SNPs were selected and the analysis was repeated (Figure 11). These results show that the subset of 20 SNPs is similarly effective at differentiating the groups as 53 SNPs.

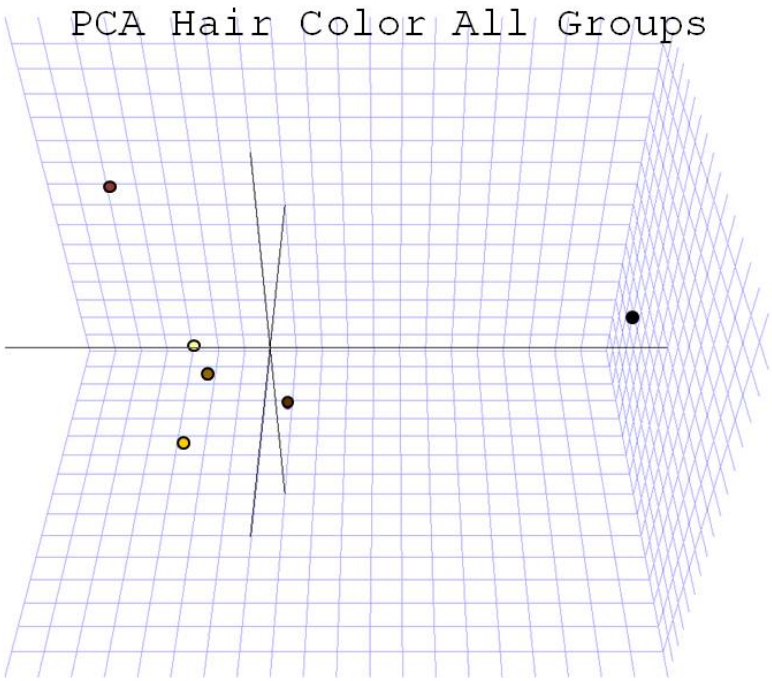
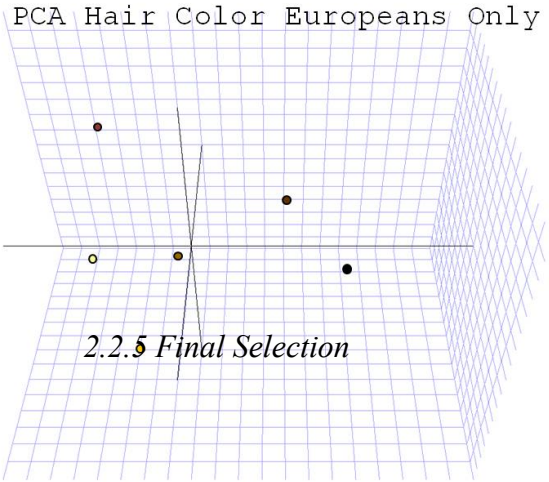
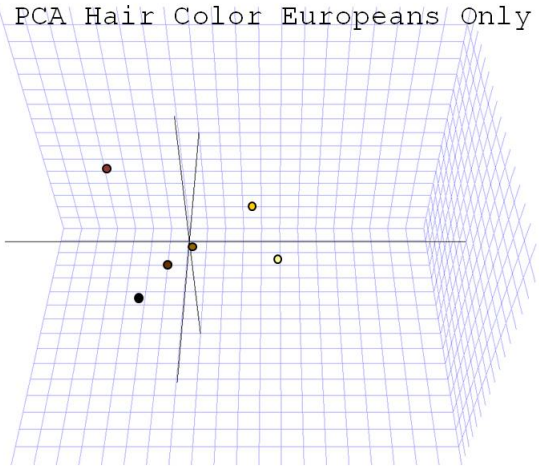


Figure 11. Tridimensional PCA plot of the 53 PISNPs analyzed on all individuals with known phenotype. Individuals were divided in 6 groups based on their hair color represented by the color of the dot: black, dark brown, light brown, dark blonde, light blonde, red/auburn.

Figure 12. (below left) Tridimensional PCA plot of the 51 PISNPs analyzed only on individuals with known phenotype and of European descent. Individuals were divided into six groups based on their hair color represented by the color of the dot: black, dark brown, light brown, dark blonde, light blonde, red/auburn. Two of the 53 SNPs analyzed on all individuals were monomorphic in Europeans; therefore, PCA was performed on 51 SNPs. (below right) Tridimensional PCA plot of the most informative 20 PISNPs analyzed only on individuals with known phenotype and of European descent.





2.2.5 Final Selection

PHASE results- Recombination: The phase test for recombination rate is based on the median values of the probabilities of recombination between each SNP, which are calculated during every iteration. According to the authors, a median value >1.92 is significant, meaning that recombination is likely to be occurring between the two associated SNPs when the median value exceeds 1.92. The *MC1R* data did not reveal any likely recombination for the three populations, which is not surprising as the 10 SNPs analyzed span only 765 bases. The *OCA2/HERC2* region showed slightly varying patterns of likely recombination in the populations, as seen in Table 4.

SNPs	1--2	2--3	3--4	4--5	5--6	6--7	7--8	8--9	9--10
Distance	9265	33147	134	1475	28260	8937	14451	8371	44008
European	0.71	0.55	1.06	1.01	**2.46**	1.48	0.59	0.68	**3.28**
African	0.91	0.82	1.10	1.19	0.71	1.79	1.08	0.75	1.06
Asian	1.12	0.36	0.97	1.07	**2.74**	1.48	0.76	0.89	0.94

10--11	11--12	12--13	13--14	14--15	15--16	16--17	17--18	18--19
2893	5525	12621	8759	21008	41360	25229	60149	16818
1.33	**8.95**	1.29	0.64	0.77	0.77	0.62	0.44	0.76
2.17	**3.32**	1.11	0.77	0.89	0.80	0.63	0.57	0.79
2.11	**4.14**	0.83	0.86	0.78	0.69	0.77	0.98	0.51

Table 4. Recombination likelihood of the *OCA2/HERC2* region in different populations, values in bold indicate recombination is likely.

Based on these analyses of the *OCA2/HERC2* SNPs, recombination is likely between SNPs 5 and 6 in both the European and Asian populations, and between SNPs 9 and 12 in all three populations. These results can be used in candidate SNP reduction and selecting the final SNP panel, by choosing representative SNPs among 1-5, 6-9, and 12-19.

Phase Results- Haplotype: The MC1R haplotype analysis reveals that, consistent with the literature, this region is highly variable among Europeans and more conserved in other population groups. This can be seen in Figure 13, where the first box contains the distribution of haplotypes among the Europeans, the second are African/African Americans, and the third are East Asians (where each chart contains only the samples for which phenotype information was available).

MC1R Haplotype Analysis

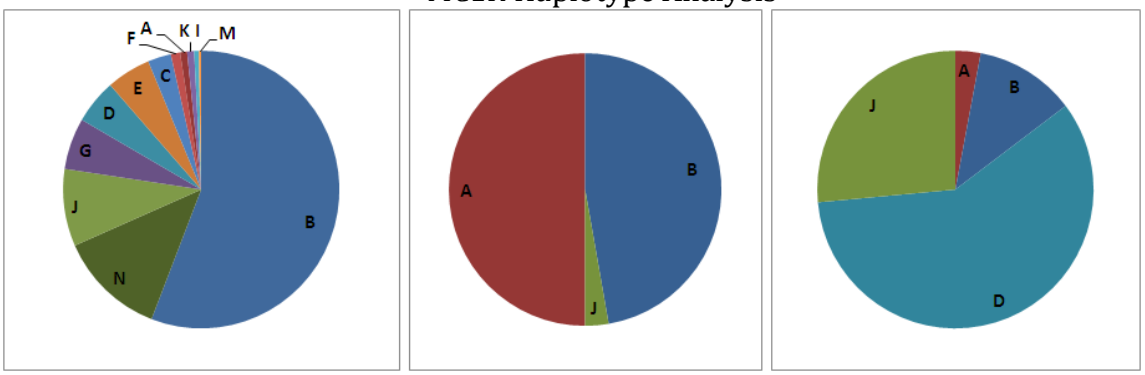


Figure 13. Haplotype distribution in the different populations that were tested (from left to right, European, African/African American, and East Asian, each chart contains only the samples for which phenotype information was available.

Further analysis shows that only the C, E, and G haplotypes appear to be associated with a lighter pigmentation among Europeans (Figure 14). Therefore, the three SNPs that define these three haplotypes (rs1805009, rs1805008, and rs1805007, respectively) are good candidates for the final assay.

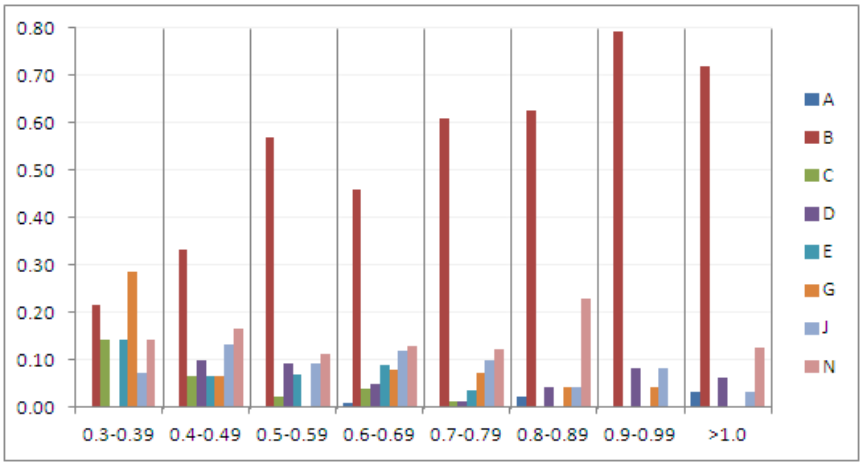


Figure 14. The graph compares melanin index (measured above elbow) on X- axis to frequency of haplotype on Y-axis, among Europeans. Haplotype B increases in frequency and diversity decreases as melanin index increases. The frequency of haplotypes C, E, and G decrease steadily as melanin index increases.

The results for *OCA2/HERC2* haplotype distribution in linked regions were not nearly as informative. Comparing results for European, East Asian and African/African American within the three predetermined linked regions (SNPs 1-5, 6-9, and 12-19), similar patterns of haplotype distribution are seen in each group within each gene region (Figure 15). This difference in results compared to those for *MC1R* could be due to the large size of the *OCA2/HERC2* regions analyzed (the three regions range from approximately 32,000 to 186,000 bases, compared to only 765 bases in the *MC1R* region analyzed), making mutation events much more likely and resulting in a higher number of haplotypes by chance rather than selective forces. The European haplotypes found in linked regions were evaluated for correlation to skin pigmentation. No clear relationship exists between any haplotype and a lighter or darker skin pigmentation (for example, Figure 16). This is not surprising, since literature associates this gene region more strongly with eye color than with skin pigmentation.

OCA2/HERC2 Haplotype Analysis

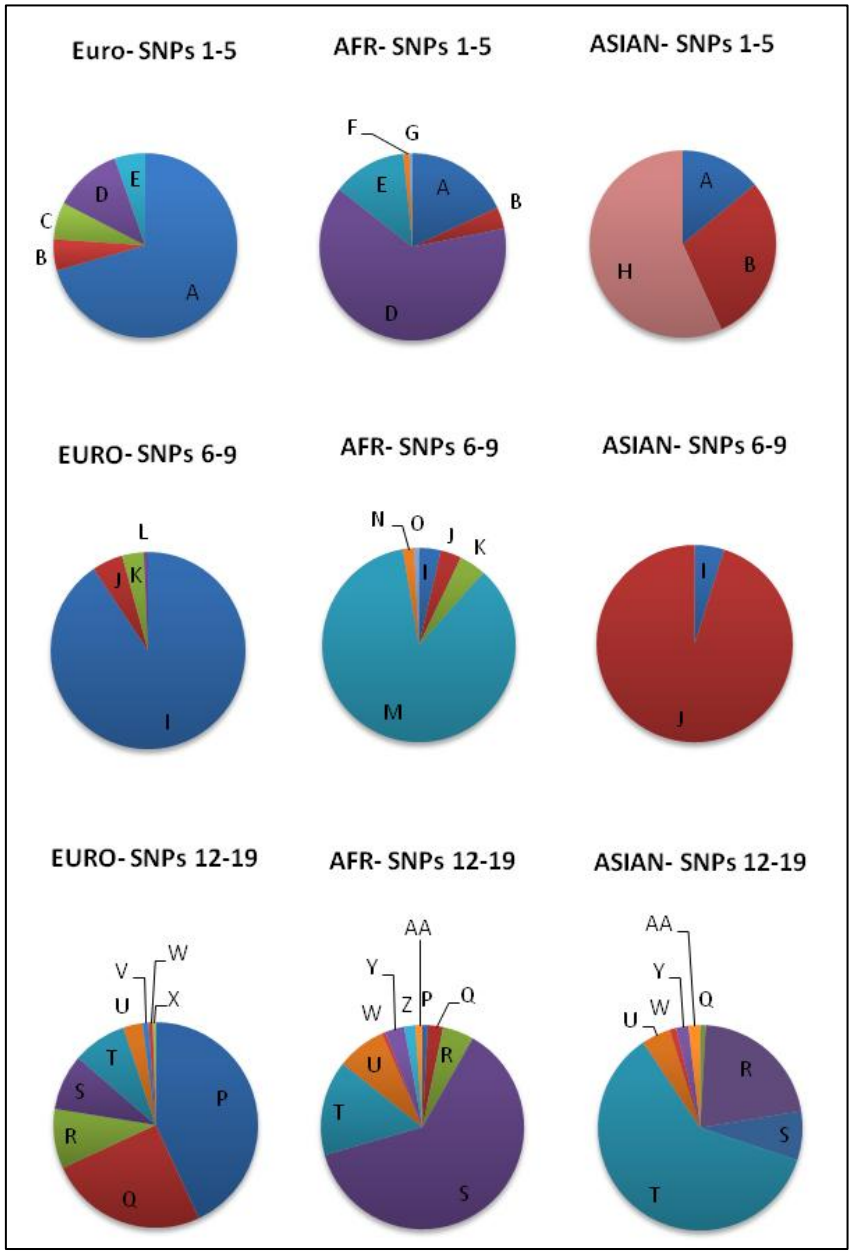


Figure 15. OCA2/HERC2 haplotypes in the three determined linked regions for Europeans (EURO), African/African Americans (AFR) and East Asians (ASIAN). Compared to the haplotype distribution for the MC1R SNPs, these gene regions show more similar number and distribution of haplotypes between the three populations. The large size of this gene region makes chance mutation more likely.

European Melanin Index vs OCA2/HERC2 Haplotype Frequency

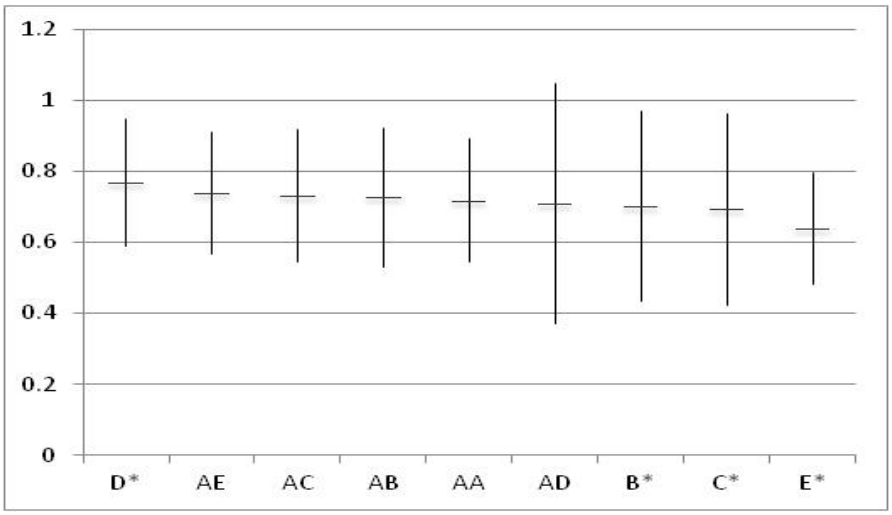


Figure 16. Example of OCA2/HERC2 haplotypes (x-axis) compared to skin melanin content above elbow (y-axis) in Europeans (SNPs 1-5). Two letters indicate an individual's two predicted haplotypes, whereas one letter and "*" indicates one predicted haplotype combined with any other haplotype. The horizontal line is the average and the vertical line is the range. Based on this analysis, no haplotype shows a clear relationship to skin melanin content.

2.1.9 Final SNP selection

By cross-referencing each of these analyses and paying particular attention to SNPs for which published prediction models already exist (Branicki 2011, Walsh 2011), 50 SNPs were selected that are expected to be most predictive of ancestry, specific phenotype traits, or both (see Appendix Tables 1a – 1d for results of each statistical approach). The resulting list includes 19 AIMs and 31 pigmentation PIMs, 13 of which also have a strong association to ancestry.

2.2 Phase 2

The objectives of this phase were three:

1) Optimize the 50 SNPs assay with a minimum necessary amount of starting DNA taking into account that a sample, to be useful in investigations, should first also yield an STR profile. The target amount of DNA to be able to type all the selected SNPs was set to no more that 1ng of DNA.
2) Develop/Test prediction modes for ancestry. Once a DNA profile has been generated with the 50-SNP assay, a statistical model is needed to generate ancestry predictions. The ideal model provides accurate predictions across the populations of interest, is tractable for the forensic science practitioner, and produces comprehendible results for the investigator.

3) Test available phenotype prediction models. Once ancestry prediction has been established for a sample, phenotype predictions can provide additional investigative information. Currently the most accurate phenotype predictions are for eye color among Europeans using a published model (Walsh 2011).

2.2.1 Development / Optimization of 50-SNP Assay

The assay was designed using the previously described SBE method. The 50 selected SNPs were divided into three multiplexes (A: 16plex, B: 15plex and C: 19plex), based on the compatibility of the primers that were designed during the first phase of this project. See Appendix Table for information on the SNPs in each multiplex.

Optimization was performed by comparing varying concentrations of PCR reaction components ($MgCl_2$, dNTPs and Taq DNA polymerase) and cycling parameters. The optimized reaction was compared to the AmpFLSTR® Identifiler® Plus (Applied Biosystems, Foster City, CA) reaction mix and cycling parameters. Low volume purification was optimized such that the entire purification product was used in the SBE reaction, which reduces the cost of reagents and consumables, in addition to reducing the number tube transfers, making the process less prone to contamination and more amenable to automation. Samples were electrophoresed on the 3130 Genetic Analyzer (Applied Biosystems), using a 36cm capillary (Applied Biosystems, refurbished from gelcompany Inc.) and POP-7 polymer (Applied Biosystems), with injection parameters of 1.2kV for 16 seconds.

The SBE reaction was optimized by comparing varying reaction volumes and cycling parameters. Both PCR and SBE primer inputs were optimized to maximize balance in the resulting electropherogram peaks.

Sensitivity was tested ranging from 2.5pg to 10ng of input DNA, using a sample quantified via UV-Vis spectrophotometry (NanoDrop 2000, Thermo Scientific). Additional testing was performed on eight highly heterozygous samples, also quantified with UV-Vis spectrophotometry, at 100pg, 150pg and 200pg. The multiplexes were evaluated for robustness with various types of mock forensic samples, all of which had previously yielded STR profiles with AmpFLSTR® Identifiler® Plus.

Bin sets were also developed for each multiplex in order to facilitate data analysis and interpretation in GeneMarker v. 2.4 (Softgenetics, State College, PA) and GeneMapper v. 4.0 software, (Applied Biosystems) (see Appendix Table 3 a-c); however, these will require adjustment based on polymer used and other laboratory-specific conditions.

2.2.2 Development / Testing of Prediction Models for Ancestry

Linkage Disequilibrium Analysis: Prior to performing this analysis, it was necessary to evaluate which SNPs were in linkage disequilibrium (LD), because including linked SNPs would inflate the impact of that gene region on the overall ancestry prediction. Linkage was calculated using WGAviewer software (Ge 2008), which utilizes HapMap genotype data and SNP information (as available) to generate the two common measures of LD, r^2 and D', between each pair of SNPs occurring on the same chromosome. Also considered were the results of the Phase analysis test for linkage (performed for *MC1R* and *OCA2/HERC2* SNPs) addressed above.

Six of the 50 SNPs are each found on chromosomes where none of the other 50 SNPs are present; therefore, these were not evaluated for linkage. Thirty-six of the remaining 44 SNPs were included in the linkage analysis (the remaining eight were not present in the HapMap data set). A conservative review of the linkage disequilibrium analysis reduced the number of SNPs to be included in the biogeographic ancestry prediction to 32 (see Appendix Table 1a for this subset of SNPs and Appendix Tables 1 b-c for complete results of the linkage evaluation).

Training Set Development: Next, the development of an ancestry model requires the creation of a training set, comprised of known individuals from each of the populations of interest. This training set is used to establish allele frequencies for each SNP in the model, upon which prediction calculations for unknown samples will be based.

Of the available genotypes from a combination of samples (some internally tested and some downloaded from the 1000 genome project (The 1000 Genomes Project Consortium 2008)), a subset of one thousand samples from the four populations of interest was selected using the web-based application Snipper. Under the "Thorough analysis of population data of a custom Excel file" function in Snipper, a set of up to 1000 samples can be evaluated ("verbose cross-validation analysis" function was used) for the success rate of classifying samples into their known population groups. Samples were removed and added in an iterative fashion to determine a subset of samples that were highly predictive of the correct ancestry group, in order to create the most divergence between population groups.

The composition of the training set is 266 Europeans, 250 East Asians, 250 African Americans, and 234 Hispanic/Native Americans. Allele frequencies for each of the 32 loci were then calculated within each population.

The 32 East Asian samples were used in the 50 SNP selection process, and had been evaluated as candidates for, and excluded from, the training set. The two possible results of this are 1) inflation of RMP values for the 32 East Asian samples, because these individuals helped inform SNP selection and 2) deflation of RMP values for the 32 East Asian samples because these individuals were less predictive of East Asian ancestry compared to the samples chosen for the training set. The latter factor is expected to have a greater effect on the results; therefore, the results for the East Asian test set should be conservative, or statistically lower than the expected results from true unknown forensic samples of East Asian ancestry.

Test Set: The samples tested under each ancestry model were composed of 31 Europeans, 32 African Americans, 32 Hispanics and 32 East Asians. The majority of these test samples (European, African American and Hispanic) were obtained from the National Institute of Standards and Technology (NIST); the East Asian samples were internally available. Aside from the East Asian samples, these test set samples had not previously been used for any

purpose in this project (neither selection of the 50 SNP panel, nor the development of the training set).

The LR was calculated for each sample by dividing the highest RMP obtained among the four populations by the other three. The number obtained expresses the likelihood of the profile if the sample originated from the population in the numerator versus if the sample originated from the population in the denominator:
LR1 = highest RMP / second highest RMP

A threshold of 1000 was empirically chosen above which the LR1 is considered significant for a sample to be classified as belonging to a specific population (the one in the numerator) while LR1 values below 1000 were defined as inconclusive (but still informative) between the two populations with highest and second highest RMPs meaning that the individual most likely belongs to one of the two (or both) populations.
Snipper employs the same frequency based approach to calculate RMP/LR values for a single unknown sample. Because it is far simpler to test a large sample set using in-house developed spreadsheets rather than singularly inputting test samples into Snipper, the website was not used in our current analysis. However, the site would be an easy way for a practitioner to predict the ancestry of a forensic sample. We would expect a practitioner to obtain a success rate of classification similar to that described below, using their unknown sample (assuming it is from one of the four primary U.S. populations) and our U.S.-specific training set with the "Classification with a custom Excel file of populations" function in Snipper. The benefit in using Snipper when testing one unknown sample is a user-friendly interface and a clear report of the results.

7 SNP MLR Ancestry Prediction Model: In order to develop a best fitting model, the sample of 1000 subjects were also used to test each of the 50 SNPs individually against ancestry using a multinomial logistic model. Any SNPs showing evidence of a significant association with ancestry (via the pseudo r2 providied by the regression model and the p-values associated with each ancestry level) were retained for the final model. Those retained SNPs were then included in a final model which was iteratively adjusted for inclusioin/exclusion of SNPs until the final model containin 7 SNPs was chosen. Because the ancestry of the 1000 subjects in building dataset were well defined, we were able to simplify our final to only 7 SNPs.

CHAID based 5 SNP Decision Tree Ancestry Prediction Model: The generation of classification trees from large data sets is part of a relatively new area of statistics referred to as "data mining". There are several forms of data mining, one using regression analysis to compartmentalize continuous variables and one using Chi-Square to compartmentalize the categorical data. CHAID is the acronym for *Chi-squared Automatic Interaction Detector*. It is one of the oldest methods and was originally proposed by Kass(1980). CHAID will build non-binary decision trees, based on a relatively simple algorithm. To our knowledge this is the first application of CHAID to a forensic / genetic problem. CHAID uses the Chi-square test to determine the next split at each step. In this case the predictors are the genotypes of the SNPs used and the items being classified are ancestry, eye color, hair color etc. The categorical predictors are discontinuous so they are easily divided. In our case for bi-allelic SNPs there are three states: homozygous for the ancestral allele (defined as the highest frequency allele in Africa), heterozygous for the ancestral allele and derived allele, and homozygous for the derived allele. In practice these were simply coded as 1, 2 or 3. The

algorithm cycles through the predictors to find the predictor that has the lowest probability, which creates the most significant splits, after having eliminated all of the non-significant predictors (similar to a Principle Components Analysis). In our case we used "Exhaustive CHAID" algorithms, which performs a more thorough merging and testing of predictor values, and reduces all decisions until only two categories remain for each predictor. To carry out this analysis Excel spreadsheets were loaded into Statistica (12th edition, 64 bit)(StatSoft, Tulsa, OK), The dependent variable was chose (ancestry, eye color, hair color, etc) the categorical variables were chosen (SNPs) and the algorithm was run. The classification tree is grown, yielding a graph which is an easy way of envisioning the process, in that it tells you what SNP is involved in each split. You can choose to generate the tree with a training set, and test it on unknown samples or use the entire data set and test each sample against algorithm using V-fold testing, which is the equivalent of jack knife or boot strap testing, in that each sample is removed and tested. This can be printed or saved to determine how you misclassification errors occurred. A printout of summary results includes correct and incorrect classification is available.

2.2.3 Testing of Available Prediction Models for Phenotype

MLR (Irisplex): The six SNPs comprising this published eye color model (Walsh 2011) are included in the 50-SNP assay; therefore, the supplementary excel-based calculator was used to evaluate this model on the European samples for which eye color information was available (N=196). The results of this calculator are prediction probabilities for blue, brown, or intermediate eye color (where the sum of the probabilities equals one, and the highest number is the predicted eye color). These prediction probabilities were compiled for each individual, and compared to their reported eye color (self reported and confirmed by the individual collecting the sample). The results were evaluated using probability thresholds of 0.5, 0.7 and 0.9, and the accuracy/error rate (known eye color or incorrect eye color being predicted above threshold) was compared to the sensitivity (number of individuals below threshold, considered inconclusive).

CHAID based 4-SNP Eye Color Decision Tree: This approach is virtually identical to the one described in the previous section simply targeting different SNPs (rs12913832, rs1800407, rs722889, rs1876482) and categorizing individuals based on their eye color (brown, blue, and intermediate).

3 Results

3.1 Development / Optimization of 50-SNP Assay Results

The best peak balance with the least background was found in a 25µL reaction volume. Evaluation of PCR reaction mixture components showed that increasing DNA polymerase and dNTP input improved results, while the AmpFLSTR® Identifiler® Plus reaction mix performed poorly in comparsion. The multiplexes performed best with increased PCR cycle number (35), 1 minute incubation for denaturation, annealing and extension; annealing temperature of 58°C (PCR primer T_M ranged from 52°C to 62°C, with the majority falling between 55°C-59°C); and extension temperature of 72°C. SBE reaction volume evaluation showed an 8ul reaction best balanced sensitivity and background. The optimal SBE parameters were 28 cycles with a 55°C annealing temperature.

Recommended Protocol

PCR reaction components in a 25µL reaction include: 1X PCR Buffer Gold® (Applied Biosystems), 2.5mM $MgCl_2$ (Applied Biosystems), 0.22mM dNTPs (Roche Diagnostics, Indianapolis, IN), 0.0568mg/ml BSA (Fisher Scientific, Waltham, MA), 4.375 U AmpliTaq Gold DNA Polymerase® (Applied Biosystems), 2µL multiplex-specific PCR primer mix (Integrated DNA Technologies, Coralville, IA; see Appendix Tables 2 a-c for primer sequences and reaction concentration), with the remaining volume provided by H_20/DNA extract.

PCR amplification (GeneAmp PCR System 9700, Applied Biosystems) proceeded with an initial incubation step of 95°C for 10 minutes; then 35 cycles of 1) 94°C denaturation for 1 minute, 2) 58°C annealing for 1 minute, and 3) 72°C extension for 1 minute; followed by a final extension at 72°C for 10 minutes, and a 4°C indefinite hold.

Unincorporated primers and dNTPs were removed from 2µL of PCR product by adding 5 U Exonuclease I (Thermo Scientific, Waltham, MA) and 0.5 U Shrimp Alkaline Phosphatase (Affymetrix, Santa Clara, CA), plus 0.25µL H20, in a final volume of 3µL. The enzymatic reaction (9700) proceeded with a 37°C incubation for 70 minutes, followed by a 70°C incubation for 20 minutes. This entire purified product was then used in the SBE reaction.

The SBE reaction components were 1µL SNaPshot Reaction Mix® (Applied Biosystems), 1µL multiplex-specific SBE primer mix (Integrated DNA Technologies, see Appendix Table 2d for primer sequences and reaction concentration), 3µL H_20, and 3µL purified product, (to reduce consumables, the SBE reaction components can be added directly to the purification tube/plate). The SBE reaction was performed on the 9700 with the following conditions: 96°C denaturation for 10 seconds, 28 cycles of 1) 55°C annealing for 5 seconds and 2) 60°C extension for 30 seconds, followed by a 4°C indefinite hold.

To prepare samples for electrophoresis, ten microliters of LIZ 120 size standard (Applied Biosystems) was added to 400µl of Hi-Di formamide (Applied Biosystems), and 1µl of sample was added to 10µl of the Formamide/ILS mixture. Samples were electrophoresed on the 3130 Genetic Analyzer (Applied Biosystems), using a 36cm capillary (Applied Biosystems, refurbished from gelcompany Inc.) and POP-7 polymer (Applied Biosystems), with injection parameters of 1.2kV for 16 seconds. Note that most crime labs use POP 4 for STR analysis and POP 6 for sequencing analysis, both these polymers can be used to separate SBE fragments but parameters such as injection time/voltage and run voltage may need to modified, adapting them to the characteristics of the polymer.

Initial sensitivity testing detected all 50 SNPs at 100pg of input DNA. Further testing with samples chosen to maximize heterozygosity revealed that four SNPs (Multiplex A: rs1805008, rs65488616; Multiplex C: rs1540771, rs7495174) often contain background and/or low non-specific peaks, which can cause these SNPs to be mis-typed as heterozygotes at or below 200pg of input DNA (see Table 5 for evaluation of nine SNPs with relatively low peak heights; SNPs not included in this table were correctly typed to 100pg as heterozygotes). Careful evaluation of results and controls is required at or below this level. To minimize stochastic effects, recommended input range is 0.5-2ng DNA per multiplex; however, the goal of genotyping all 50 SNPs with 1ng of DNA was met, as concordant results would generally be expected with inputs totaling 1ng.

Table 5. Sensitivity test results for SNPs with relatively low peak heights (height listed next to each allele).

Sample	Input DNA (pg)	Multiplex A								Multiplex B	
		rs885479		rs1834640		rs1805008		rs6548616		rs16891982	
1	100	C-40	T-42	G-211	A?-111	C?-69	T?-40	G-646	A?-251	G-613	C-493
	150	C-116	T-98	G-154	A-116	C-157	T?-50	G-564	A-297	G-1174	C-439
	200	C-246	T-206	G-901	A-766	C-428		G-2061	A-768		
2	100	C-160		G-285		C-166		G-522	A?-143	G-848	C-466
	150	C-244		G-821		C?-211		G-1247	A?-144	G-1077	C-356
	200	C-275		G-1370		C-286		G-1948			
3	100	C-62	T-55		A?-171	C-104	T?-33	G-144	A-369	G-1175	C-476
	150	C-112	T-114		A-417	C-160	T?-49	G-225	A-420	G-1139	C-404
	200	C-133	T-209		A-943	C-346			A-635		
4	100	C-258			A-276	C-246		G-443	A-266		C-894
	150	C-416			A-511	C-315		G-389	A-239		C-692
	200	C-418			A-999	C-519		G-1422	A-691		
5	100	C-129			A-389	C?-111		G-112	A-335	G-730	C-349
	150	C-280			A-597	C-195		G-141	A-646	G-1413	C-476
	200	C-281			A-1031	C-227			A-888		
6	100	C-224			A-298	C-141	T-163	G-177	A-390		C-894
	150	C-299			A-971	C?-160	T-206		A-447		C-1072
	200	C-76			A-321	C?-57	T?-116	G?-98	A-214		
7	100	C-103	T-93	G-605		C-161		G?-109	A-307	G?-180	C-360
	150	C-161	T-251	G-986		C-278		G?-107	A-612	G-578	C-321

Table 5 (continued).

Sample	Input DNA	Multiplex C							
		rs3827760		rs1540771		rs7495174		rs735612	
1	100		T-1128	C?-76	T-111		A-96	G-76	T-53
	150		T-1302	C?-91	T-175		A-140	G-105	T-48
	200		T-2029		T-308		A-176	G-111	T-65
2	100		T-2015	C?-100	T-205	G-885	A-113	G-452	
	150		T-1534	C?-77	T-166	G-217	A?-42	G-343	
	200		T-1640	C?-99	T-182	G-217		G-540	
3	100	C-1970	T?-76		T-732		A-182	G-685	
	150	C-1324	T-136		T-493		A-204	G-349	
	200	C-902			T-325		A-91	G-598	
4	100		T-1141		T-239		A-156		T-117
	150		T-1558		T-335		A-154		T-122
	200		T-1676		T-430		A-157		T-169
5	100		T-2886		T-683	G-132	A-124	G-113	T-162
	150		T-2693		T-732	G-108	A?-54	G-291	T-96
	200		T-2248		T-583	G-375	A?-78	G-687	T-293
6	100		T-1428		T-511		A-253		T-154
	150		T-1916		T-578		A-314		T-217
	200		T-2333		T-539		A-169		T-477
7	100	C-595	T-663	C?-126	T-290		A?-51	G-602	
	150	C-727	T-997	C-202	T-304		A-374	G-691	

KEY:

?	An actual allele where a peak is visible but of poor quality (low peak height, bad morphology, or high background)
?	Not an actual allele where a peak is visible but of poor quality (low peak height, bad morphology, or high background)
	Not an actual allele where the peak would be incorrectly called an allele

NOTES:

1. Multiplex A rs1805008, negative control also shows a non-specific T allele. This non-specific T overlaps the C allele; whereas an actual T allele migrates two bases longer than the C allele.

2. Multiplex A rs6548616, negative control also shows a non-specific G allele. In a true G/A heterozygote, the G allele should be significantly greater peak height than the A allele (as seen in samples 1 and 4).

3. Multiplex C rs3827760, in an actual heterozygote CT, the alleles should be similar in peak height. The non-specific T alleles seen in sample 3 at 100pg and 150pg are of lower relative peak height than expected.

4. Multiplex C rs1540771, in an actual CT heterozygote, the alleles should be similar in peak height. The non-specific C alleles are all of poor quality and lower relative peak height than expected.

5. Multiplex C rs7495174, sample 2 at 200pg, A allele completely dropped out; however it was called at 100pg in samples 2 and 5.

The multiplexes performed well with various types of mock forensic samples, including cigarette butts extracted with DNA IQ® (Promega Corporation, Madison, WI), QIAamp DNA Mini Kit® (Qiagen, Hilden, Germany), and Chelex® 100 Resin (Bio-Rad Laboratories, Hercules, CA); mouth area of bottles extracted with DNA IQ® and QIAamp DNA Mini Kit®; and chewing gum extracted with QIAamp DNA Mini Kit®. See Figure 17 for electropherograms showing multiplex performance on a forensic sample.

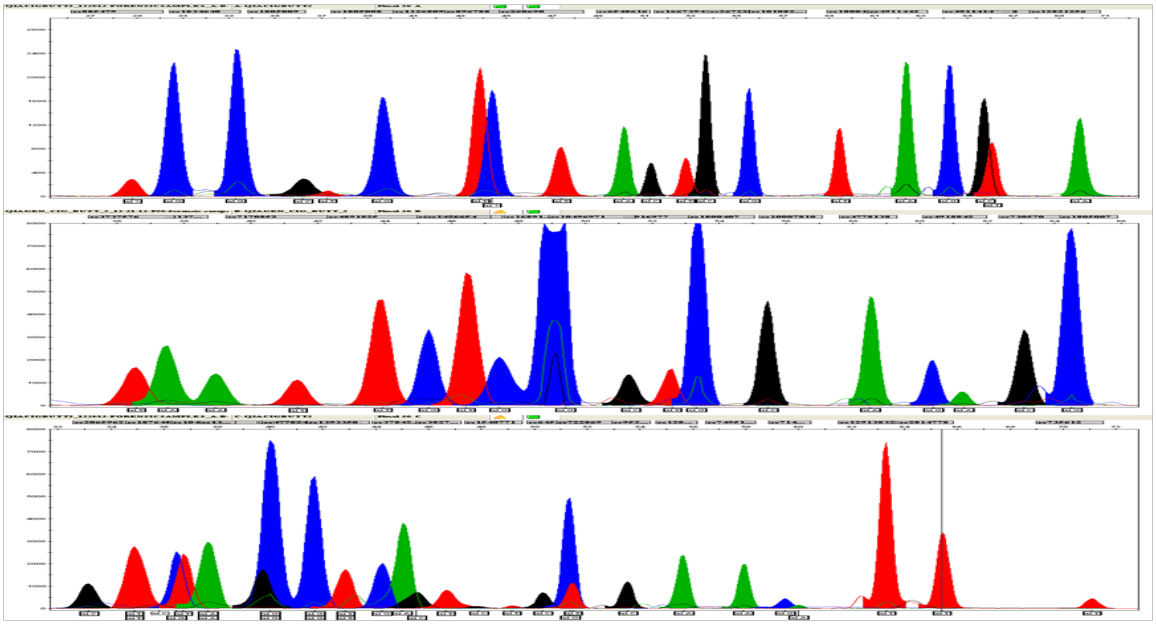


Figure 17. Electropherograms results of the 50 SNP assay (three multiplexes); profile obtained from a cigarette butt.

To facilitate allele call at each locus panels and bins have been developed for both GeneMapper (Applied Biosystems) and GeneMarker (SoftGenetics) software platforms. Examples are shown in figures 18 and 19.

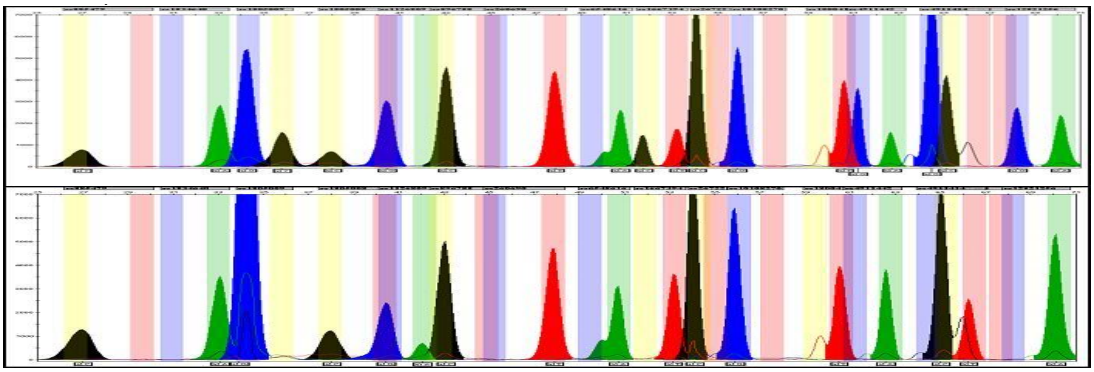


Figure 18. Electropherogram of samples typed with multiplex A, visualized on GeneMarker software, with dedicated panels and bins to facilitate allele call at each locus. A genotype table is created and can be extracted directly into excel format for further manipulation.

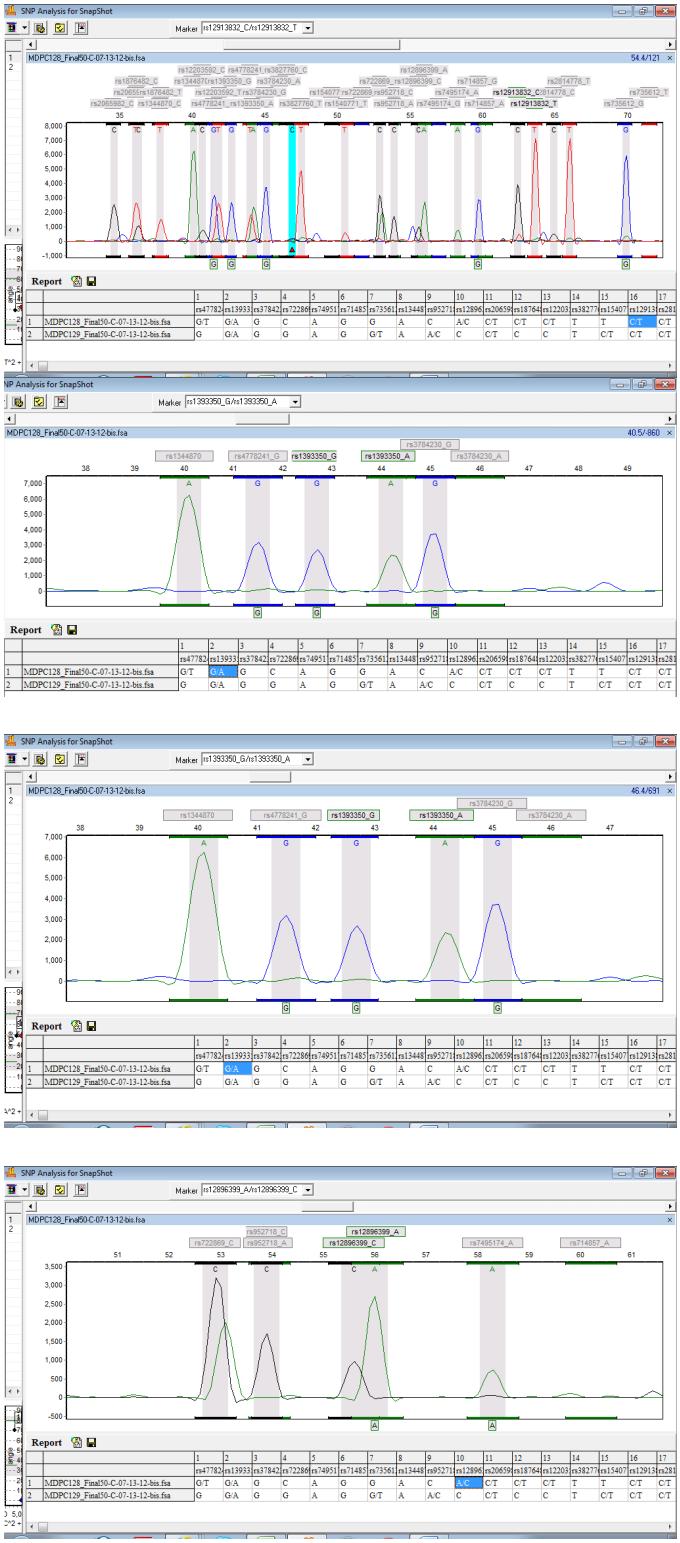


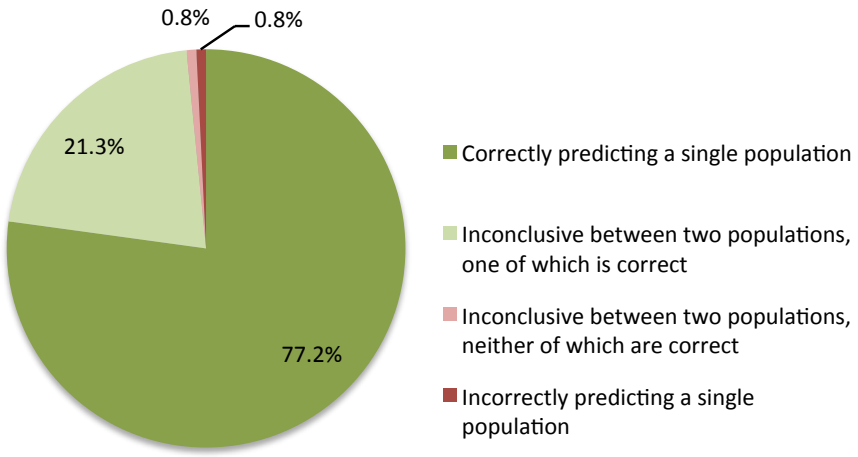
Figure 19. Example of GeneMarker SNaPShot/SNPlex specific application, this application is designed specifically to interpret SNP data and allows creating custom panels and bins tailored to each multiplex. A genotype table is created and can be extracted directly into excel format for further manipulation.

3.2 Development / Testing of Prediction Models for Ancestry Results

32 SNP RMP/LR Ancestry Model Performance: See Figure 20a for a summary of the results. Of the 127 samples in the test set, 99 (78%) showed a significant LR1 (>1000), and one of these would be predicted incorrectly (classifying as Hispanic/Native American instead of NIST-classified African American). The misclassifying individual has an African mtDNA haplogroup L1c and an African Y chromosome haplogroup E. The remaining 28 (22.1%) individuals had a LR1 below 1000 and were classified as inconclusive between two populations (the highest and second highest RMPs). As seen in Figure 20b, the ratio of inconclusive to predicted individuals is consistent across the populations, indicating a balance of highly predictive SNPs for each population. One of the samples in the inconclusive category would be incorrectly predicted as either Hispanic/Native American or European (sample was NIST-classified as African American) because those two populations had the highest two RMPs, while the RMP obtained from the African American population was the third highest. This sample has a mtDNA haplogroup H1a, supporting a maternal European heritage, and a Y chromosome haplogroup E, supporting a paternal African lineage. Overall, two individuals out of the 127 would have been incorrectly predicted (1.6%) and correct information would be relayed to the investigator for 98.4% of individuals.

32-SNP RMP-LR Ancestry Model Performance

a.



b.

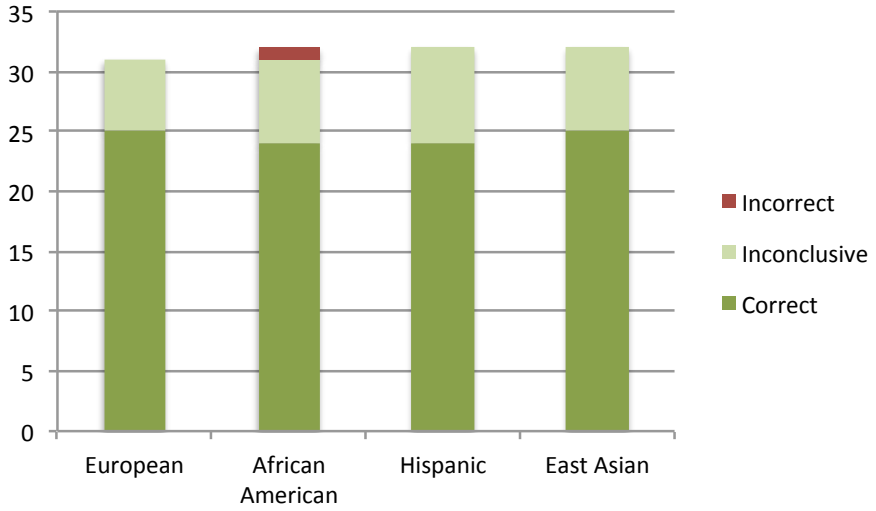
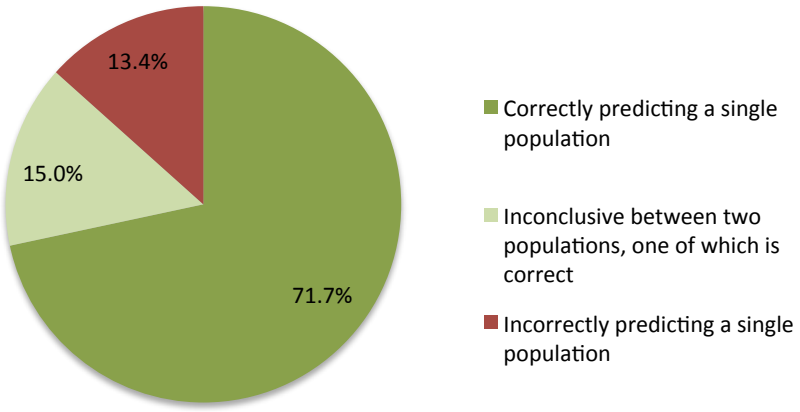


Figure 20. (a) 32-SNP RMP/LR ancestry model performance by population; similar distribution of inconclusive samples seen in each group, incorrect prediction only seen in African American population. (b) Summary of overall 32-SNP RMP/LR ancestry model performance.

7 SNP MLR Ancestry Model Performance: These results were evaluated with prediction probability thresholds of 0.8 and 0.9, meaning if the highest prediction probability did not reach the threshold, the result was considered inconclusive. The two thresholds gave the same percentage of correctly classified individuals; however, the 0.9 threshold was shown to reduce the number of incorrect predictions, and was used to further evaluate the results as described below.

The overall results (Figure 21a) show a significantly higher proportion of individuals (13.4%) would be incorrectly classified when compared to the previous model. This outcome is expected when less SNPs are employed in the prediction model. As seen in Figure b, the incorrect predictions are distributed fairly evenly across the populations. All of the inconclusive results (where the highest prediction probability is less than 0.9) are such that the correct population is one of the highest two predicted; therefore, correct information could still be given to the investigator for these individuals (e.g. the sample came from either a Hispanic or East Asian individual). The proportion of inconclusive results varies widely among the populations (Figure 21b): at the low end, no inconclusive results were seen for the European samples and at the high end, 13 inconclusive results were seen for East Asian samples. This indicates the 7-SNP model contains more or more powerful SNPs for discriminating European individuals and less or less powerful SNPs for discriminating East Asian individuals. Overall, correct information would be given for 86.7% of samples in the test set under this model.

a.



b.

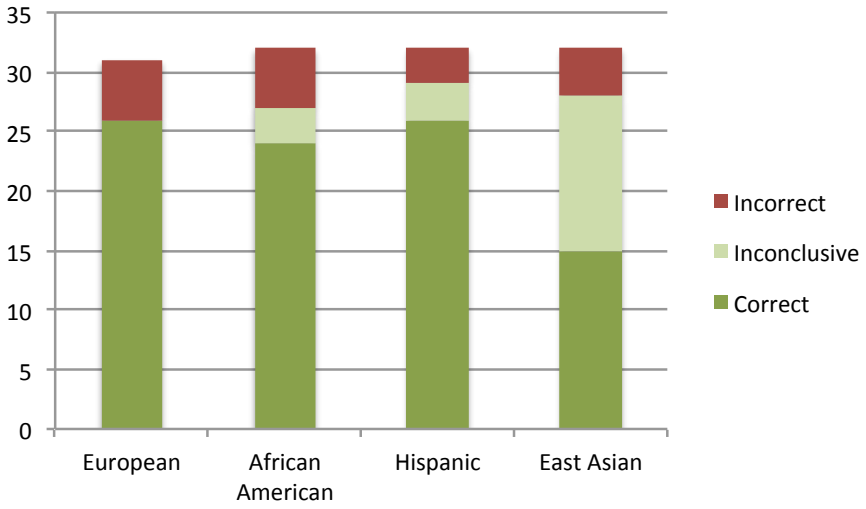


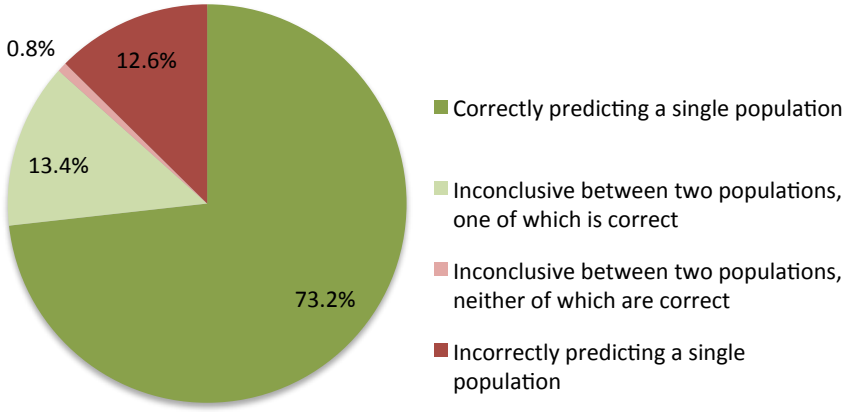
Figure 21. (a) Summary of overall 7-SNP MLR ancestry model performance. (b) 7-SNP MLR ancestry model performance by population; similar distribution of incorrectly predicted samples seen in each group, varying distribution of inconclusive samples among the groups, indicating an imbalance of predictive ability for different populations.

CHAID based 5 SNP Decision Tree Ancestry Prediction Model Performance: As was done for the previous model, these results were also evaluated with prediction probability thresholds of 0.8 and 0.9. The 0.8 threshold yielded 6% more correctly classified individuals, 8% less inconclusive individuals, and 2% more incorrectly classified individuals. This 0.8 threshold was used to further evaluate the results as described below.

The overall results of this 5-SNP model (Figure 22a) show very similar overall results when compared to the 7-SNP model, and a significant increase in incorrect predictions compared to the 32-SNP model. As seen in Figure 22b, the incorrect and inconclusive predictions vary widely in their distribution across the populations. This distribution indicates the 5-SNP model contains more or more powerful SNPs for discriminating African American individuals and less or less powerful SNPs for discriminating Hispanic and East Asian individuals, with European individuals falling somewhere in between. All but one of the inconclusive results (where the highest prediction probability is less than 0.8) are such that the correct population is one of the highest two predicted; therefore, correct information could still be given to the investigator for all but one of these individuals. Overall, correct information would be given for 86.7% of samples in the test set under this model.

CHAID Based 5-SNP Decision Tree Ancestry Model Performance

a.



b.

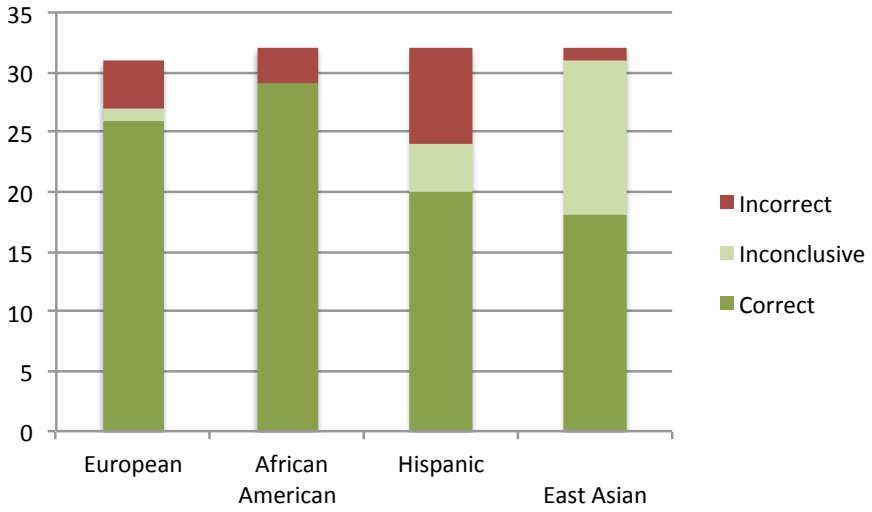
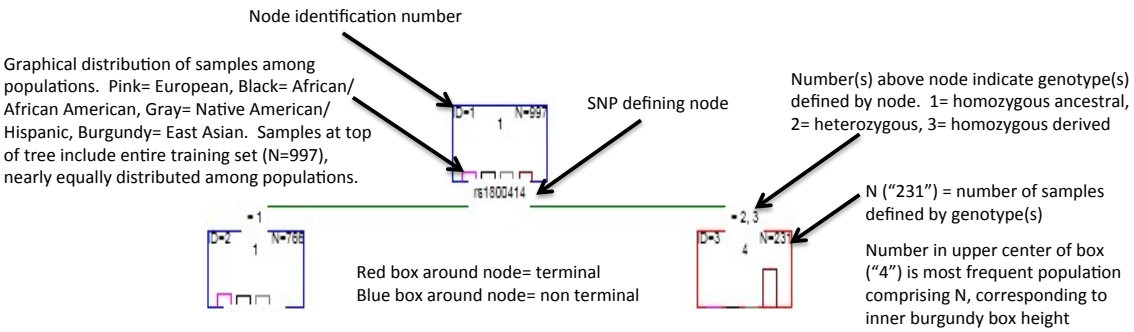


Figure 22. (a) Summary of overall 5-SNP decision tree ancestry model performance. (b) 5-SNP decision tree ancestry model performance by population; imbalanced distribution of incorrectly predicted and inconclusive samples seen in each group, indicating an imbalance of predictive ability for different populations.

Classification trees generated by CHAID can offer several advantages over logistic regression and other methods of decision making. The output as a tree (figure 23 a and b) gives the practitioner a simple way to sort the data for classification purposes, so that any given SNP result can be classified. The method seems to use the smallest number of SNPs to reach a decision comparable to other methods. CHAID is an easily explained algorithm as opposed to logistic regression.

a



b

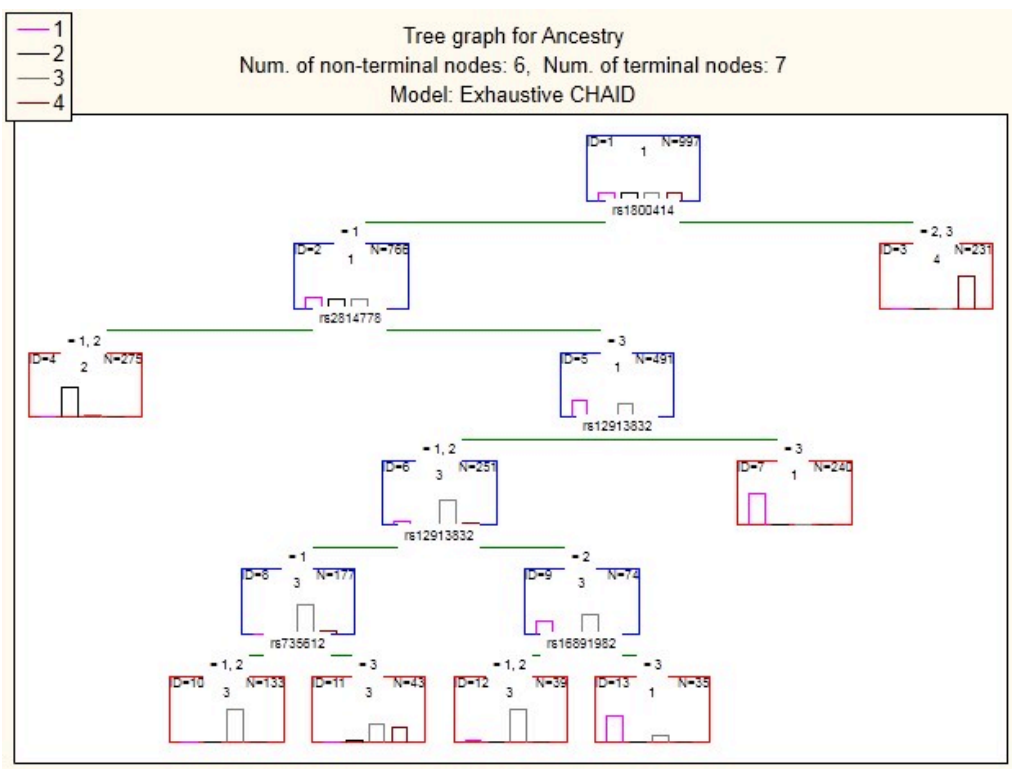


Figure 23. (a) Guide to reading decision tree, (b) decision tree model created with training set samples. Note that a single SNP may appear twice on the tree (as seen with rs12913832) if each of the three possible genotypes are used separately to discriminate the samples.

3.3 Testing of Available Prediction Models for Phenotype Results

Irisplex: As seen in Figure 24, results from testing 196 European individuals for whom eye color information was available in the Irisplex model show an expected trade-off between accuracy and sensitivity, and an overall issue with predicting intermediate eye color.

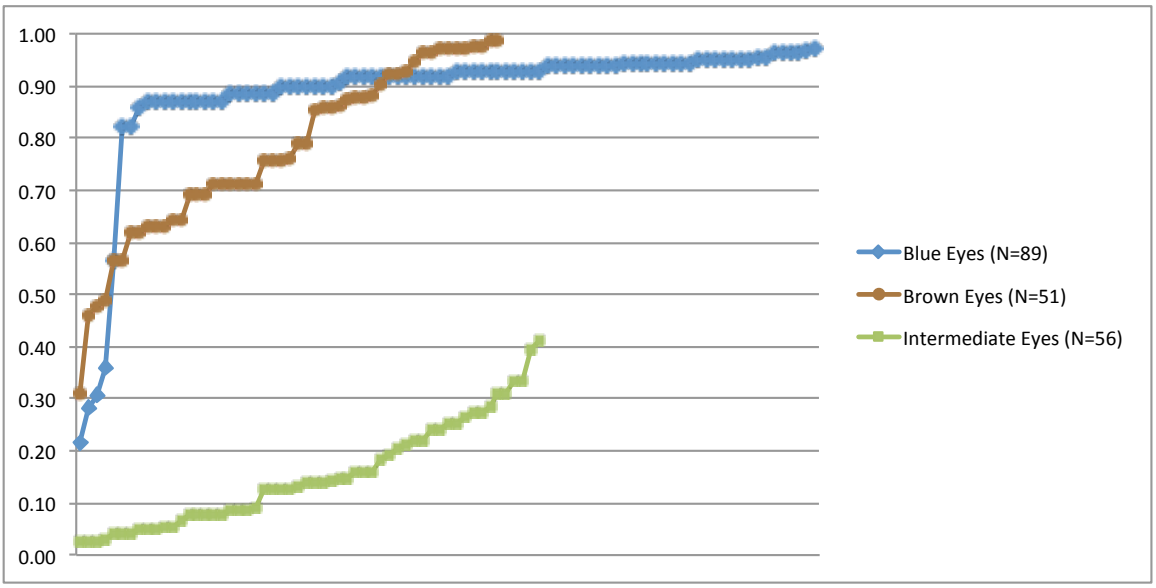


Figure 24. Results from Irisplex model showing the prediction probability (y-axis) for the known eye color of each sample. Red dashed line indicates the level below which the known eye color is not the predicted eye color.

Establishing a threshold below which a prediction probability is inconclusive will aide a practitioner in interpreting and delivering the results of this model. Using a 0.5 threshold, >90% of samples are classified: 96% of individuals with blue eyes and 92% of individuals with brown eyes are correctly classified; however 21% of individuals predicted to have blue eyes actually have an intermediate eye color (green or hazel), and 33% of individuals predicted to have brown eyes actually have blue (N=2) or intermediate (N=20) eye color. At a 0.7 threshold, 75% of samples are classified: 94% of individuals with blue eyes and 67% of individuals with brown eyes are correctly classified; 20% of individuals predicted to have blue eyes actually have an intermediate eye color (67% green and 33% hazel), and 17% of individuals predicted to have brown eyes actually have an intermediate eye color (all hazel). Lastly, at a 0.9 threshold, only 48% of samples are classified: 73% of individuals with blue eyes and 29% of individuals with brown eyes are correctly classified; and the error rates are 17% for blue and 6% for brown (Figure 25). Based on this data set, the 0.5 threshold cannot be recommended due to the high error rate, and the 0.9 threshold cannot be recommended due to the low sensitivity. The use of a 0.7 threshold allows for eye color prediction in ¾ of European individuals, where 81% of predicted samples are correct and erroneous prediction for blue eyes are most likely be green in color, while erroneous prediction for brown eyes are expected to be hazel.

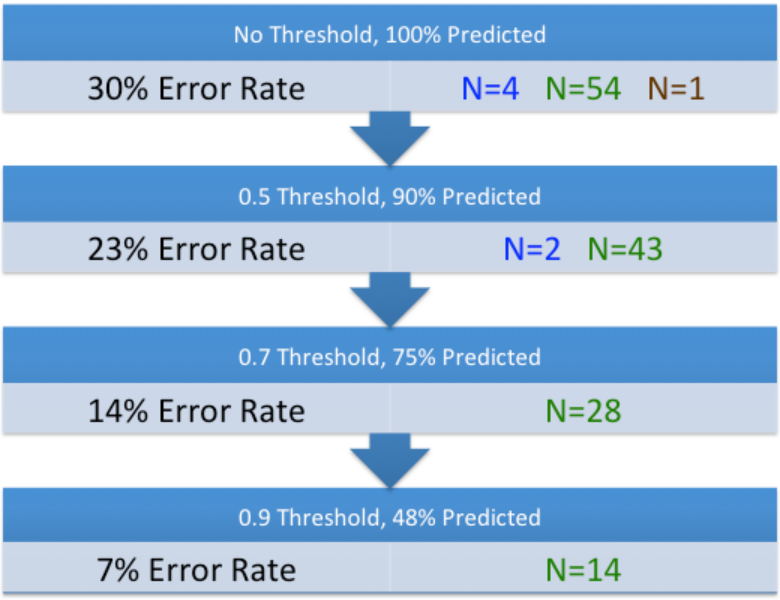


Figure 25. Results for the Irisplex model at various thresholds. The "N" values correspond to individuals with the color-coded known eye color who are erroneously predicted to have a different eye color.

A more conservative option for delivering eye color prediction information to law enforcement would be to define a sample as 'not blue', when predicted to be brown, and 'not brown' when predicted to be blue. With this approach all individuals would be classified correctly with the 0.7 and 0.9 thresholds.

Of note is that the prediction probability for the intermediate eye color never exceeded 0.5, and out of N=56 individuals of known intermediate eye color, the prediction probability was the highest for intermediate in only two individuals. This issue is the primary cause of the error rate in blue/brown prediction, and the same issue was noted in previous work on this model [Liu 2009], although to a lesser extent. We agree with the authors of the model's hypotheses that this could be due to inconsistencies in phenotype categorization and/or the existence of unidentified variants that could better predict this phenotype.

4-SNP Eye Color CHAID Decision Tree: One hundred and eighty seven European individuals for whom eye color and complete genotype information were evaluated with this method, using bootstrapping (*i.e.* all samples were used to build the model then each sample was removed and evaluated using the model). The decision tree model generated from this sample set is shown in Figure 26.

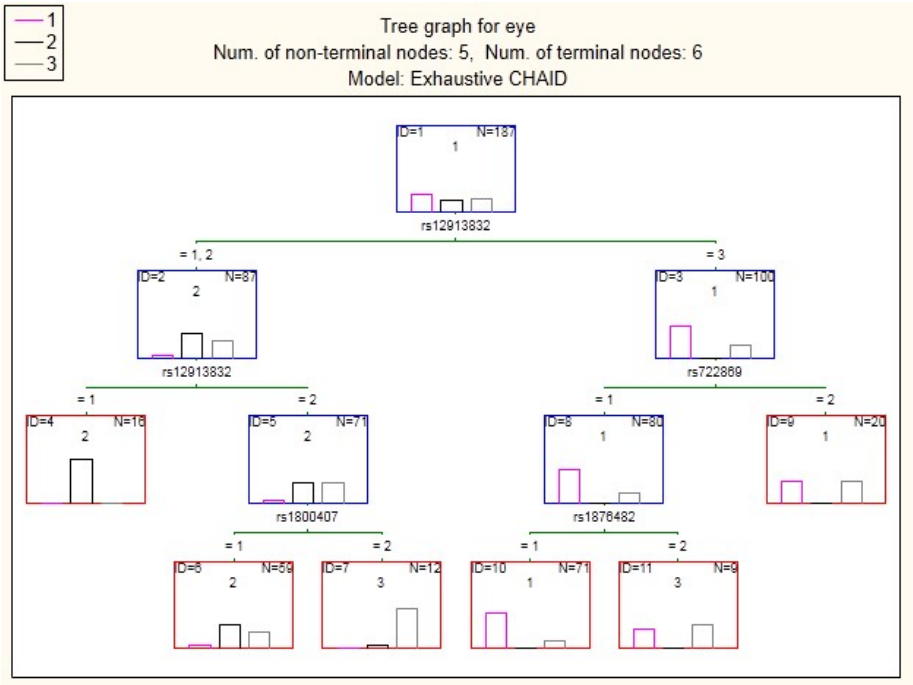


Figure 26. Decision Tree generated from 4-SNP eye color model. See previous guide to reading decision tree.

As was seen with Irisplex, there is again a trade-off between accuracy and sensitivity, and an overall issue with predicting intermediate eye color, although this determination is improved (Figure 27).

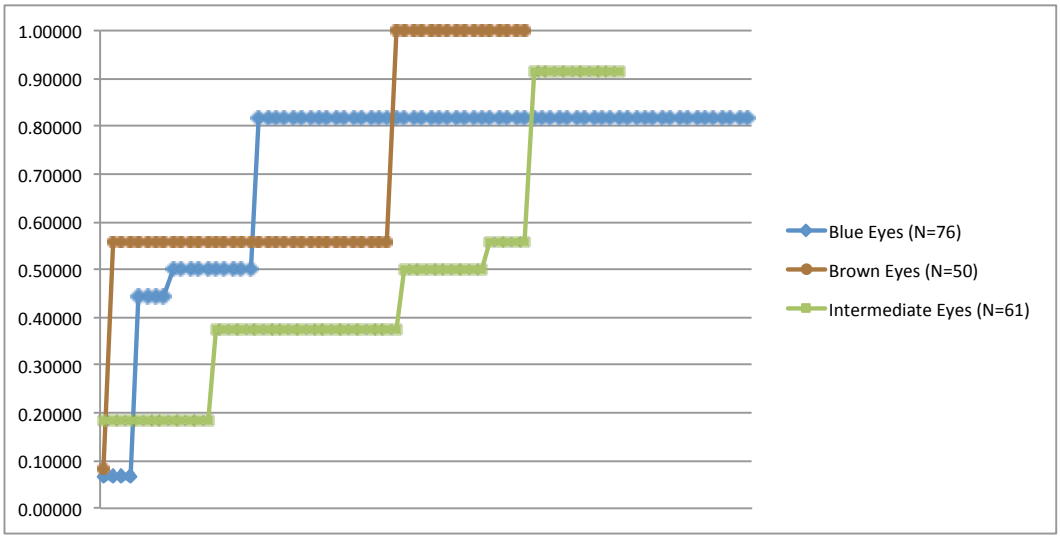


Figure 27. Results from 4-SNP decision tree model showing the prediction probability (y-axis) for the known eye color of each sample. Red dashed line indicates the level below which the known eye color is not the predicted eye color.

Twenty samples have equal probabilities for two eye colors (0.5 probability blue and 0.5 probability brown); therefore these 20 samples are inconclusive with or without a

threshold. Figure 28 summarizes the results at the different thresholds. The same results were obtained using no threshold or a 0.5 threshold: 89% of samples were predicted, with a 26% error rate. At a 0.7 threshold, only 53% of samples are predicted, with a 14% error rate and the 0.9 threshold reduces the number of predicted individuals to an unacceptable 15%. In comparison to the Irisplex model, the error rates are similar; however the percentage of individuals predicted under each threshold is markedly lower.

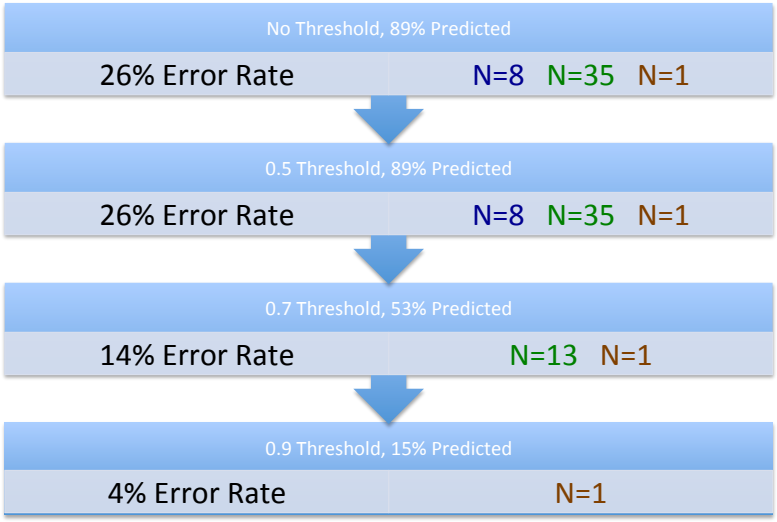


Figure 28. Results for the 4-SNP decision tree model at various thresholds. The "N" values correspond to individuals with the color-coded known eye color who are erroneously predicted to have a different eye color.

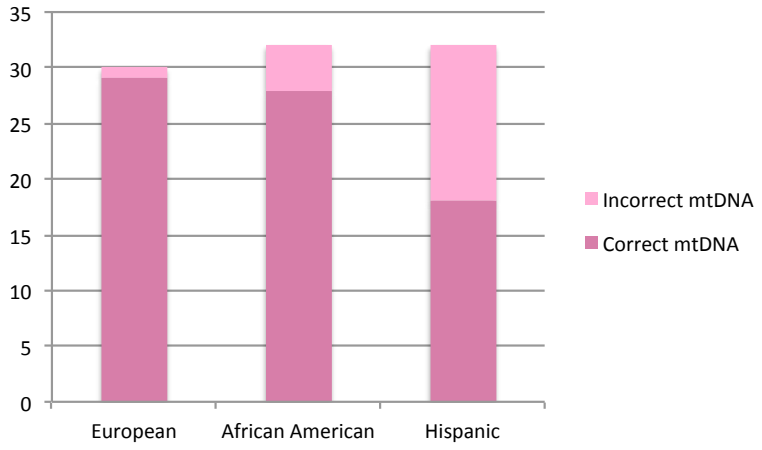
In addition, the previously described conservative option for delivering eye color prediction information to law enforcement (defining a sample as 'not blue', when predicted to be brown, and 'not brown' when predicted to be blue) does not work under this model even with a 0.9 threshold, as one brown-eyed individual is predicted to have intermediate eye color at >0.9 probability.

3.4 Mito and Y analysis

The test set samples from the European, African American, and Hispanic populations from NIST had previously been analyzed and haplotypes assigned for regions of the mitochondrial genome and Y chromosome (all samples were male). Both mtDNA and Y data were missing for one European sample, and Y data was missing for one African American sample.

To evaluate the haplogroup frequencies and whether including the haplogroups would improve the overall analysis, population haplogroup frequency data was gathered for the four populations in the mitochondrial genome (Allard 2002, Allard 2005, Allard 2006, Budowle 2002, Irwin 2007, Irwin 2008, Lee 2006) and the Y chromosome (Willuweit 2007). Then for each test set sample, the population in which the mitochondrial or Y haplogroup was most frequent was determined. Figure 29 shows the results.

a.



b.

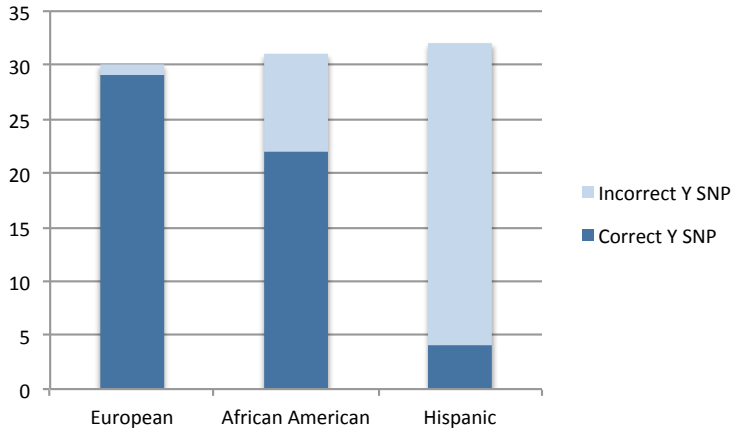


Figure 29. Haplogroup results by known population of test samples. (a) Dark pink portion of column indicates how often the mtDNA haplogroup is consistent with the known ancestry of the individual, light pink indicates the haplogroup is not consistent with the known ancestry (b) Dark blue portion of column indicates how often the Y chromosome haplogroup is consistent with the known ancestry of the individual, light blue indicates the haplogroup is not consistent with the known ancestry.

For European individuals, including the mitochondrial and Y chromosome information would improve the ancestry prediction for the majority of samples. This information would make the ancestry prediction worse for two individuals, one with a predominantly African American mtDNA haplogroup, and the other with a predominantly African American Y chromosome haplogroup.

In African American individuals, the majority of ancestry predictions would be improved by including the mitochondrial haplogroup (87.5% of samples improved) and the Y chromosome haplogroup (71% of samples improved) in the evaluation; however, the overall success rate of the 32-SNP autosomal ancestry model is higher. For the samples that

would be negatively impacted by including these results, the mitochondrial haplogroup is most frequent in Europeans for two of these samples and most frequent in East Asians for an additional samples two samples; whereas, the Y chromosome haplogroup is most frequent in Europeans for all "incorrect" samples (N=9).

The Hispanic individuals would be the most negatively affected by including the mitochondrial and Y chromosome haplogroup information, with only 56% and 12.5% of samples being improved, respectively. The majority of the known Hispanic individuals that would be incorrectly classified as European based on both the mtDNA and Y chromosome haplogroups.

Overall, because the mitochondrial and Y chromosome information are lineage specific and not representative of the entire heritage of an individual, it is not surprising that these results would not consistently improve ancestry prediction. Based on these results, a well-chosen autosomal SNP panel is generally expected to outperform mitochondrial and Y chromosome ancestry predictions, particularly in regions of the world where admixed populations are common.

4 Conclusions

4.1 Discussion of findings and Implications for policy and practice

In a forensic case where an STR profile has not matched any known individuals or database samples, the unknown sample can be genotyped with this 50 SNP assay to provide predicted likelihood of the four most frequent U.S. populations (African American, East Asian, European, or Hispanic/Native American). By entering the 32 SNP genotypes and the U.S. training set into the web-based application Snipper, a forensic practitioner can quickly generate highly accurate results (employing the aforementioned threshold) in a report format. With the RMP/LR method 77.2% of the individuals within the test set were correctly predicted of belonging to one of the four populations tested while 21.3% of the individuals were classified as inconclusive between two populations. The latter prediction although defined as 'inconclusive' still provides information that is potentially useful to an investigation, as two populations are excluded. Additionally, if a sample were to be defined as inconclusive between Hispanic/Native American and East Asian in a geographic region where there are very few from the latter population but there are higher proportions of Europeans, African Americans, and Hispanics/Native Americans, the information combined with local demographics could further guide the investigation.

Using this approach, the misclassification rate was less than 2%. Although low, this number should still be considered when providing the prediction information to investigators. Also, the model has been tested on the four most common populations in the US but remains to be evaluated on other, rarer US populations such as Central/South Asians, Pacific Islanders, etc.

Useful information regarding the paternal and maternal lineage of an individual can also come from the Y chromosome and the mitochondrial DNA (mtDNA) respectively. We are currently evaluating the best approach to incorporate this information in the prediction process. The challenge is how to weigh the ancestral lineage information of the two markers

in the overall prediction. Although useful at times, it could also be very misleading, for example an individual may appear European but have a Native American mtDNA haplogroup that entered the family many generations ago. The same could be true for an African American family with a European Y chromosome haplogroup.

The number of samples collected to date is insufficient to develop effective prediction models for pigmentation, particularly with the MLR approach. Hair color and eye color variation is detected predominantly in Europeans and the SNPs determining these variations have been primarily established among Europeans, thus such predictions should be performed only once European ancestry is indicated. In this report we presented two different models, one based on MLR and one based on CHAID, for eye color prediction. The former was published by Walsh in 2011 and is referred to as the Irisplex model and the latter is a new approach that, to our knowledge, has not yet been tested on this type of data.

Overall the results are comparable between the two methods although they change depending on the thresholds that are set when interpreting the results. These thresholds affect the number of individuals for which a prediction is made and the error in the classification of individuals. It is possible that with a much larger training set for the CHAID approach, more SNPs would be included in the model, potentially increasing the number of individuals correctly predicted.

This low cost assay can be implemented in any US crime lab as it uses the same technology used with conventional STR analysis, it generates reliable results with less than 1 ng of template DNA, and it is robust enough for typical forensic samples. Investigators can use the information obtained to prioritize suspect processing, corroborate the testimony of a witness to a crime, and overall optimize their resources.

4.2 Implications for further research

In our initial proposal we had planned to collect samples from approximately 200 individuals. During the course of the project we realized that, in order to be able to develop effective prediction models for eye, hair, and skin pigmentation, we needed to significantly increase the number of subjects in the study. After obtaining IRB approval we continued sample collection. We now have collected samples and data from over 300 individuals, which is still insufficient for developing accurate prediction models compared to other studies [Walsh 2011, Branicki 2011]. Thus we will continue to collect DNA samples with corresponding ancestry and phenotype information in order to eventually test the selected pigmentation phenotype markers in a larger population, develop pigmentation models based on the U.S. population, and evaluate these models in an independent sample. Funding is being sought for a large collection effort of over 3000 individuals. This collection will represent a valuable resource to the forensic science community, as it will contain extensive information regarding individuals' ancestry and phenotype, along with skin and hair spectrophotometric measurements for melanin content and color determination.

Although the SNP assay published herein contains 50 SNPs, only 35 of those are used in the models presented (32 in the ancestry model, including three eye color model SNPs, plus three additional eye color model SNPs). We are currently evaluating other methods of ancestry prediction, and the possibility of using haplotypes in the ancestry prediction models, to allow for inclusion of linked loci and increase prediction accuracy.

In addition, we are investigating the possibility of incorporating an STR-based ancestry likelihood into an overall ancestry model. From the literature it is clear that far less

ancestry information is contained in the forensic STR loci compared to AIMs (as these STR loci were chosen for their ability to differentiate individuals, not populations) [Barnholtz-Sloan 2005]. However, because the forensic STR profile should already be available by the time an evidence profile is subjected to SNP analysis, it would be worthwhile to incorporate any amount of ancestry association that exists in the STR data.

5 References

The 1000 Genomes Project Consortium. A map of human genome variation form population-scale sequencing. Nature 467 (2010) 1061-1073. URL: http://browser.1000genomes.org

M.W. Allard, D. Polanskey, K. Miller, M.R. Wilson, K.L. Monson, B. Budowle. Characterization of human control region sequences of the African American SWGDAM forensic mtDNA data set. Foren. Sci. Int. 148 (2005) 169-179.

M.W. Allard, D. Polanskey, M.R. Wilson, K.L. Monson, B. Budowle. Evaluation of variation in control region sequences for Hispanic individuals in the SWGDAM mtDNA data set. J. Foren. Sci. 51 (2006) 566-573.

J.S. Barnholtz-Sloan, C.L. Pfaff, R. Chakraborty, J.C. Long, Informativeness of the CODIS STR Loci, J. Foren. Sci. 6 (2005) 1322-1226.

K. A. Beaumont, R.A Newton, D.J. Smit *et al.*, Altered cell surface expression of human MC1R variant receptor alleles associated with red hair and skin cancer. *Hum Mol Genet* **14 (2005)**, 2145.

B. Bertoni, B. Budowle, M. Sans, S.A. Barton, R. Chakraborty, Admixture in Hispanics: distribution of ancestral population contributions in the Continental United States, Hum.Biol. 75 (2003) 1-11.

C. Bouakaze, C. Keyser, E. Crubézy, D. Montagnon, B. Ludes, Pigment phenotype and biogeographical ancestry from ancient skeletal remains: inferences from multiplexed autosomal SNP analysis, Int.J.Legal Med. 123 (2009) 315-325.

W. Branicki, U. Brudnik, A. Wojas-Pelc, Interactions between HERC2, OCA2 and MC1R may influence human pigmentation phenotype, Ann.Hum.Genet. 73 (2009) 160-170.

W. Branicki, F. Liu, K. van Duijn, J. Draus-Barini, E. Pospiech, S. Walsh, T. Kupiec, A. Wojas-Pelc, M. Kayser, Model-based prediction of human hair color using DNA variants, Hum. Genet. 129 (2011) 443-454.

M. H. Brilliant, NIJ Grant# 2002-IJ-CX-K010, final report (Sept. 2008).

M. Brión, J.J. Sanchez, K. Balogh, C. Thacker, A. Blanco-Verea, C. Børsting, et al., Introduction of an single nucleodite polymorphism-based "Major Y-chromosome haplogroup typing kit" suitable for predicting the geographical origin of male lineages, Electrophoresis. 26 (2005) 4411-4420.

A.J. Brookes, The essence of SNPs, Gene. 234 (1999) 177-186.

B. Budowle, M.W. Allard, C.L. Fisher, A.R. Isenberg, K.L. Monson, J.E. Stewart, M.R. Wilson, K.W. Miller. HVI and HVII mitochondrial DNA data in Apaches and Navajos. Int. J. Leg. Med. 116 (2002) 212-215.

J.M. Butler, Forensic DNA typing : biology, technology, and genetics of STR markers, 2nd ed., Elsevier Academic Press, Burlington, MA, 2005.J. M. Butler *et al.*, *Progress in Forensic Genetics 12*, Copenhagen 2007.

J.M. Butler, M.D. Coble, P.M. Vallone, STRs vs SNPs: thoughts on the future of forensic DNA testing, Foren. Sci. Med. Pathol. 3 (2007) 200–205.

J.M. Butler, B. Budowle, P. Gill, K.K. Kidd, C. Phillips, P.M. Schneider, P.M. Vallone, N. Morling, Report on ISFG SNP panel discussion, Foren. Sci. Int. Supp. Ser. 1 (2008) 471–472.

K. Butler, M. Peck, J. Hart, M. Schanfield, D. Podini. Molecular eyewitness: Forensic prediction of phenotype and ancestry, Foren. Sci. Int. Genet. Supp. Ser. 3 (2011), 498-499.

D.L. Duffy, G.W. Montgomery, W. Chen, Z.Z. Zhao, L. Le, M.R. James, et al., A three-single-nucleotide polymorphism haplotype in intron 1 of OCA2 explains most human eye-color variation, Am.J.Hum.Genet. 80 (2007) 241-252..

D. Falush, M. Stephens, J.K. Pritchard, Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies, Genetics. 164 (2003) 1567-1587.

F. Fiorentino, MC Magli, D Podini *et al*., The minisequencing method: an alternative strategy for preimplantation genetic diagnosis of single gene disorders. *Mol. Hum. Reprod.* 9 (2003), 399.

M. Fondevila, C. Phillips, N. Naveran, L. Fernandez, M. Cerezo, A. Salas, et al., Case report: identification of skeletal remains using short-amplicon marker analysis of severely degraded DNA extracted from a decomposed and charred femur, Forensic Sci Int Genet. 2 (2008) 212-218.

T. Frudakis, M. Thomas, Z. Gaskin, K Venkateswarlu, KS Chandra, S Ginjupalli, S Gunturi, S Natrajan, VK Ponnuswamy, KN Ponnuswamy. Sequences Associated with human pigmentation, *Genetics* 165 (2003), 2071-83.

D. Ge, D. Zhang, A.C. Need, O. Martin, J. Fellay, A. Telenti, D.B. Goldstein. WGAViewer: Software for Genomic Annotation of Whole Genome Association Studies. Gen. Res. 18 (2008) 640-643. URL: http://compute1.lsrc.duke.edu/softwares/WGAViewer

R.J. Haasl, B.A. Payseur, Multi-locus inference of population structure: a comparison between single nucleotide polymorphisms and microsatellites, Heredity. 106 (2011) 158-171.

I. Halder, M. Shriver, M. Thomas, J.R. Fernandez, T. Frudakis, A panel of ancestry informative markers for estimating individual biogeographical ancestry and admixture from four continents: utility and applications, Hum.Mutat. 29 (2008) 648-658.

T. Hadley, F. Koltz, and L. Miller, Invasion of erythrocytes by malaria parasites: a cellular a molecular overview. *Annu Rev Microbiol* **40** (1986), 451-477.

J. Han, P. Kraft, H. Nan, Q. Guo, C. Chen, A. Qureshi, et al., A genome-wide association study identifies novel alleles associated with hair color and skin pigmentation, PLoS Genet. 4 (2008) e1000074-e1000074.

N. Homer, S. Szelinger, M. Redman, D. Duggan, W. Tembe, J. Muehling, et al., Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays, PLoS Genet. 4 (2008) e1000167-e1000167.

R. Iida, M. Ueki, H. Takeshita, J. Fujihara, T. Nakajima, Y. Kominato, et al., Genotyping of five single nucleotide polymorphisms in the OCA2 and HERC2 genes associated with blue-brown eye color in the Japanese population, Cell Biochem.Funct. 27 (2009) 323-327.

J.A. Irwin, J.L. Saunier, P. Beh, K.M. Strouss, C.D. Paintner, T.J. Parsons. Mitochondrial DNA control region variation in a population sample from Hong Kong, China. Foren. Sci. Int. Genet. 3 (2009) e119-e125.

J.A. Irwin, J.L. Saunier, K.M. Strouss, T.M. Diegoli, K.A. Sturk, J.E. O'Callaghan, C.D. Paintner, C. Hohoff, B. Brinkmann, T.J. Parsons. Mitochondrial control region sequences from a Vietnamese population sample. Int. J. Leg. Med. 122 (2008) 257-259.

I. J. Jackson. Pigmentation diversity identifying the genes causing human diversity. Eur J Hum Genet 14, 979 (Sep, 2006).

M. Kayser, F. Liu, A.C. Janssens, F. Rivadeneira, O. Lao, K. van Duijn, et al., Three genome-wide association studies and a linkage analysis identify HERC2 as a human iris color gene, Am.J.Hum.Genet. 82 (2008) 411-423.

M. Kayser, P. de Knijff, Improving human forensics through advances in genetics, genomics and molecular biology, Nat. Rev. 12 (2011) 179–192.

K.K. Kidd, NIJ Grant # 2004-DN-BX-K025, final report (Sept. 2008).

J.R. Kidd, F.R. Friedlaender, W.C. Speed, A.J. Pakstis, D.L. Vega, K.K. Kidd, Analyses of a set of 128 ancestry informative single-nucleotide polymorphisms in a global set of 119 population samples, Investig Genet. 2 (2011) 1-1.R. Kosoy et al., Hum Mut 30, 69 (2009).

R. Kosoy, R. Nassir, C. Tian, P.A. White, L.M. Butler, G. Silva, et al., Ancestry informative marker sets for determining continental origin and admixture proportions in common populations in America, Hum.Mutat. 30 (2009) 69-78.

O. Lao, K. van Duijn, P. Kersbergen, P. de Knijff, M. Kayser, Proportioning whole-genome single-nucleotide-polymorphism diversity for the identification of geographic population structure and genetic ancestry, Am.J.Hum.Genet. 78 (2006) 680-690.

H.Y. Lee, J.E. Yoo, M.J. Park, U. Chung, C.Y. Kim, K.J. Shin. East Asian mtDNA haplogroup determination in Koreans: haplogroup-level coding region SNP analysis and subhaplogroup-level control region sequence analysis. Electrophoresis 27 (2006) 4408-4418.

F. Liu, K. van Duijn, J.R. Vingerling, A. Hofman, A.G. Uitterlinden, A.C.J.W. Janssens, M. Kayser, Eye color and the prediction of complex phenotypes from genotypes, Curr. Biol. 19 (2009) R192–R193.

J. Mengel-From, C. Børsting, J.J. Sanchez, H. Eiberg, N. Morling, Human eye colour and HERC2, OCA2 and MATP, Forensic Sci Int Genet. 4 (2010) 323-328.

D.A. Merriwether, S. Huston, S. Iyengar, R. Hamman, J.M. Norris, S.M. Shetterly, et al., Mitochondrial versus nuclear admixture estimates demonstrate a past history of directional mating, Am.J.Phys.Anthropol. 102 (1997) 153-159.

M.L. Metzker, Sequencing technologies - the next generation, Nat.Rev.Genet. 11 (2010) 31-46.

S. Myles, M. Somel, K. Tang, J. Kelso, M. Stoneking, Identifying genes underlying skin pigmentation differences among human populations, Hum. Genet. 120 (2007) 613-621.

T.M. Nelson, R.S. Just, O. Loreille, M.S. Schanfield, D. Podini, Development of a multiplex single base extension assay for mitochondrial DNA haplogroup typing, Croat.Med.J. 48 (2007) 460-472. E.J. Parra *et al., Am J Hum Genet* **63**, 1839 (1998).

E. J. Parra, Human pigmentation variation: evolution, genetic basis, and implications for public health. *Yearbook of Physical Anthropology*, **50**, 85 (2007).

E.J. Parra, A. Marcini, J. Akey, J. Martinson, M.A. Batzer, R. Cooper, et al., Estimating African American admixture proportions by use of population-specific alleles, Am.J.Hum.Genet. 63 (1998) 1839-1851.

T. Pastinen, A. Kurg, A. Metspalu, L. Peltonen, A.C. Syva¨nen, Minisequencing: a specific tool for DNA analysis and diagnostics on oligonucleotide arrays, Genome Res. 7 (1997) 606–614.

C. Phillips et al., Inferring ancestral origin using a single multiplex assay of ancestry-informative marker SNPs. Foren. Sci. Int. Genet. 1 (2007) 273-280.

C. Phillips, A. Salas, Sánchez J.J., M. Fondevila, A. Gómez-Tato, J. Alvarez-Dios, et al., Inferring ancestral origin using a single multiplex assay of ancestry-informative marker SNPs, Forensic Sci Int Genet. 1 (2007) 273-280.

C. Phillips, M. Fondevila, and M.V. Lareau. A 34-plex autosomal SNP single base extension assay for ancestry investigations, DNA Electrophoresis Protocols for Forensic Genetics, Methods in Molecular Biology 830 (2012) 109-126. URL: http://mathgene.usc.es/snipper/

A.L. Price, N. Patterson, F. Yu, D.R. Cox, A. Waliszewska, G.J. McDonald, et al., A genomewide admixture map for Latino populations, Am.J.Hum.Genet. 80 (2007) 1024-1036.

J.K. Pritchard, M. Stephens, P. Donnelly, Inference of population structure using multilocus genotype data, Genetics 155 (2000) 945–959.

H. Rajeevan, U. Soundararajan, A. Pakstis, and K. Kidd, Introducing the Forensic Research/Reference on Genetics knowledge base, FROG-kb, Investig Genet. 2012 Sep 1;3(1):18 doi: 10.1186/2041-2223-3-18.

B.P. Sokolov, Primer extension technique for the detection of single nucleotide in genomic DNA, Nucleic Acids Res. 18 (1990) 3671-3671.

S.N. Shekar, D.L. Duffy, T. Frudakis, R.A. Sturm, Z.Z. Zhao, G.W. Montgomery, et al., Linkage and association analysis of spectrophotometrically quantified hair color in Australian adolescents: the effect of OCA2 and HERC2, J.Invest.Dermatol. 128 (2008) 2807-2814.

J. Shlens, A Tutorial on Principle Component Analysis, (2009) 1-12.

G. N. Stamatas, *et al., Pigment Cell Res* **17**, 618 (2004).

M. Stephens, P. Scheet, Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation, Am.J.Hum.Genet. 76 (2005) 449-462.

R.P. Stokowski, P.V.K. Pant, T. Dadd, A. Fereday, D.A. Hinds, C. Jarman, et al., A genomewide association study of skin pigmentation in a South Asian population, Am.J.Hum.Genet. 81 (2007) 1119-1132.

R.A. Sturm, D.L. Duffy, Z.Z. Zhao, F.P.N. Leite, M.S. Stark, N.K. Hayward, et al., A single SNP in an evolutionary conserved region within intron 86 of the HERC2 gene determines human blue-brown eye color, Am.J.Hum.Genet. 82 (2008) 424-431.

R.A. Sturm, Molecular genetics of human pigmentation diversity, Hum.Mol.Genet. 18 (2009) R9-R17.

P. Sulem, D.F. Gudbjartsson, S.N. Stacey, A. Helgason, T. Rafnar, K.P. Magnusson, et al.,

Genetic determinants of hair, eye and skin pigmentation in Europeans, Nat.Genet. 39 (2007) 1443-1452.

Syvänen A.C., From gels to chips: "minisequencing" primer extension for analysis of point mutations and single nucleotide polymorphisms, Hum.Mutat. 13 (1999) 1-10.

K.J. Travers, C. Chin, D.R. Rank, J.S. Eid, S.W. Turner, A flexible and efficient template format for circular consensus sequencing and SNP detection, Nucleic Acids Res. 38 (2010) e159-e159.

P.M. Vallone, J.M. Butler, Y-SNP typing of U.S. African American and Caucasian samples using allele-specific hybridization and primer extension, J.Forensic Sci. 49 (2004) 723-732.

M. Visser, M. Kayser, R-J Palstra, *HERC2* rs12913832 modulates human pigmentation by attenuating chromatin-loop formation between a long-range enhancer and the *OCA2* promoter, *Genome Research* 3, 446-455 (2012).

S. Walsh, F. Liu, K.N. Ballantyne, M. van Oven, O. Lao, M. Kayser, IrisPlex: a sensitive DNA tool for accurate prediction of blue and brown eye colour in the absence of ancestry information, Foren. Sci. Int. Genet. 5 (2011) 170–180.

S. Walsh, F Liu, A. Wollstein, L Kovatsi, A. Ralf, A. Kosiniak-Kamysz, W. Branicki, M Kayser The HIrisPlex system for simultaneous prediction of hair and eye colour from DNA. Forensic Sci. Int. Genet. (2012), http://dx.doi.org/10.1016/j.fsigen.2012.07.005

D. C. Whiteman, P.G. Parsons, A.C. Green. Determinants of melanocyte density in adult human skin. *Arch Dermatol Res* 291 (1999), 511-516.

S. Willuweit and L. Roewer. Y chromosome haplotype reference database (YHRD): Update. Foren. Sci. Int. Genet., 1 (2007) 83-87. URL: http//www.yhrd.org

6 Dissemination of Research Findings

Results from this research, as it was ongoing, have been presented at national (AAFS, ISHI) and international (ISFG) conferences. Preliminary results were published in FSI Genetics and in February 2013 a manuscript was submitted to FSI Genetics. The paper was accepted and published in 2014: Gettings KB, Lai R, Johnson JL, Peck MA, Hart JA, Gordish-Dressman H, et al. A 50-SNP assay for biogeographic ancestry and phenotype prediction in the U.S. population. Foren. Sci. Int. Genet. 2014; 8: 101-108.

Given the quantity of data generated with this project we anticipate being able to produce at least another two manuscripts for publication: one discussing the inclusion of STR data and autosomal haplotypes to the prediction model, and one comparing the use of FROGkb and the prediction models presented in this report.

7. Work preformed during grant extension: July 2013 – July 2014

As mentioned at the top of this report the first 6 chapters are part of this project's final report submitted in 2013. The work was also published on FSI: Genetics in early January 2014 (Gettings et al. 2014). Upon discussion with the program officer an extension request to expand upon the work already conducted was submitted to NIJ. The request was accepted and this chapter summarizes the work conducted during the year extension.

7.1 Assay optimization efforts

In order to improve the yield and reduce the amplification time, different combinations of polymerases and PCR reaction enhancers were evaluated. The reagents evaluated include KAPA2G Fast Multiplex PCR kit (KAPA Biosystems), ExTaq Polymerase (Takara Bio, Inc.), SpeedSTAR DNA Polymerase (Takara Bio, Inc.), and Prep-n-Go™ buffer (Applied Biosystems). Best results were obtained using KAPA2G Fast Multiplex PCR kit. KAPA2G is a second generation polymerase which is significantly faster than conventional Taq enzymes, also the solution already contains KCl and $MgCl_2$, at concentrations specifically optimized for multiplex reactions. These allow for balanced amplification of all targets. It is a 2X solution that also simplifies PCR set up. The optimized protocol requires 12.5µL of KAPA2G Fast Multiplex PCR, 2.0 µL of primer mix (primer concentrations reported in the appendix and in Gettings e al. 2014), 5.5 µL of H_2O and 5 µL of 0.2 ng/µL of template DNA. The thermocycler protocol is 2 min at 95 °C followed by 35 cycles of 94 °C for 30 sec, 58 °C for 30 sec, and 72 °C for 30 sec, followed by a final extension at 72 °C for 10 min, and a 4 °C indefinite hold. Figure 30 shows examples of electropherograms obtained with the newly optimized protocol.

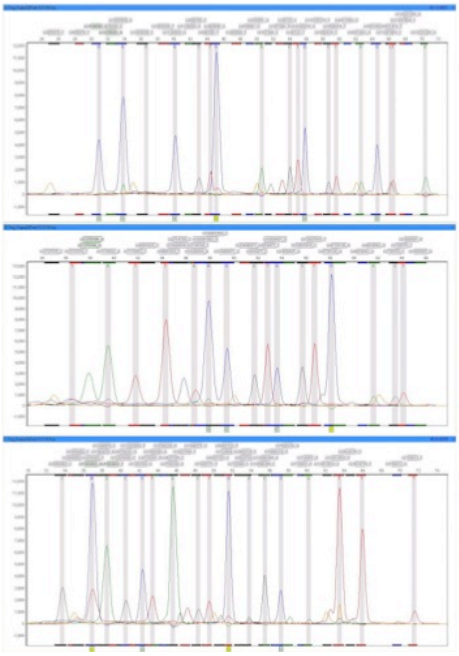


Figure 30 Examples of electropherograms obtained with the newly optimized protocol. With 1 ng of input DNA peaks are higher and more balanced than with the conventional (Taq Gold – based) protocol published in Gettings et al 2014.

It is likely that, given the results obtained as part of a different project (data not shown), amplification time can further be reduced down to a total of 30 minutes.

7.2 Sample Collection and Analysis

The original goal of this project was to collect DNA samples with ancestry and phenotype data of the donor from 200 individuals. Once the number was achieved IRB approval was obtained to increase the number of individuals in order to generate better prediction models and a total of 276 samples were collected. Another 13 samples were collected and used to evaluate the prediction models described in the previous chapters. Given the availability of funds a further extension was requested to collect another 100 samples to further evaluate the ancestry predictive power of the assay and eventually develop better prediction models for pigmentation. Sample collection was limited to the winter moths to minimize the environmental effects of sun exposure of the donors, which increase the melanin index above the basal level.

Several collection sessions were scheduled in various parts of the University: library, cafeteria, study areas, etc. Given the approved IRB collection protocol nothing can be offered to subjects to incentivize their participation to the study and the collection process takes approximately 10 minutes given that the participant has to fill in a questionnaire, donate a buccal swab, and have the melanin index measured. Thus recruiting individuals was not as successful as planned; most students denied the request to donate. Only 60 more samples were collected for a total 349 samples. This is disappointing but allowed us to identify, for future studies, possible ways to increase recruitment. Specifically, after speaking to an IRB case worker, a possible (IRB approvable) 'reward' would be to offer a gift card (for example a Starbucks card) to participants, it is likely that even a low amount, such as $5, would be a sufficient incentive to increase participation.

The limited number of samples collected are not sufficient for the purpose of improving the pigmentation prediction capabilities of the assay. Given the low cost of the process, sample collection will continue after this grant is closed and further funding will be sought to collect even more.

Focus was placed on ancestry prediction also taking advantage of a set of anonymous DNA samples already available to our lab, and IRB approved, in addition to the buccal swabs collected as part of this extension. Table 6 summarizes the samples tested during this last year and figure 31 summarizes sample ancestry in a pie chart.

Specifically 94 DNA samples (24 US-Europeans, 32 African American, and 38 self reported Hispanic).

	GWU BUCCAL	GWU DNA	TOT
EUROPEAN	38	24	62
ASIAN	5		5
AFRIC/AM	2	32	34
HISP/NA	3	38	41
MIDDLE EST	3		3
SOUT ASIA	5		5
MIX	4		4
TOT	60	94	154

Table 6. Breakdown of the samples tested during the last year of this project, GWU BUCCAL are samples collected by the PI and his staff for which both self-declared ancestry and phenotype data is available, GWU DNA are extracted anonymous samples that were collected by Dr. Schanfiled in the 80s and 90s.

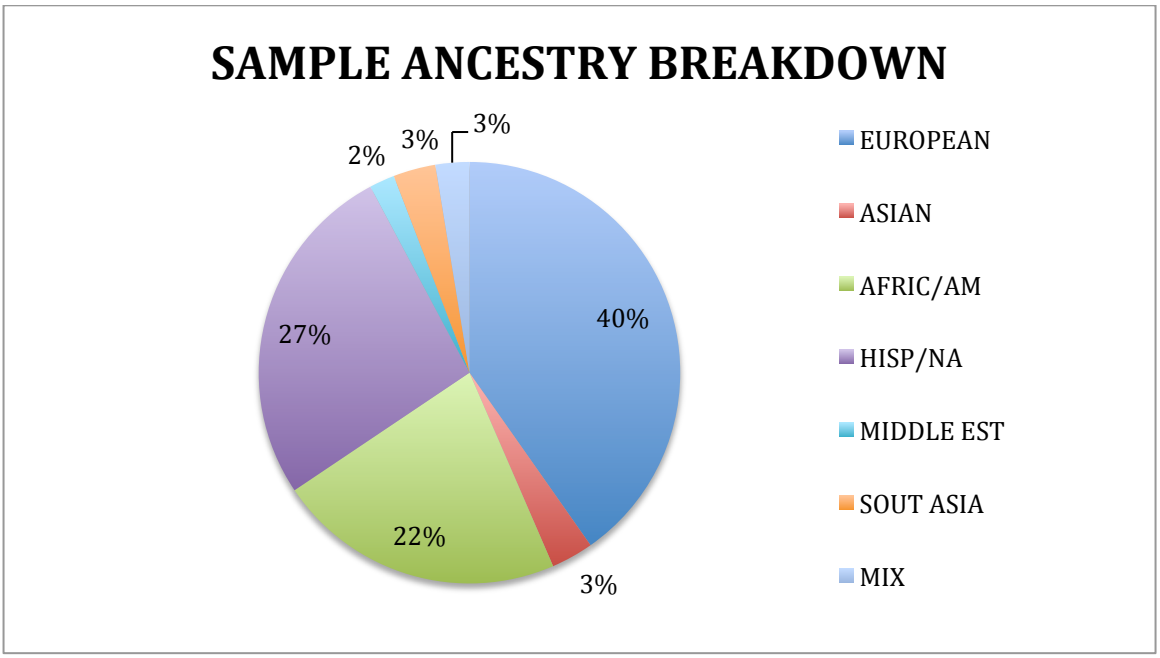


Figure 31. Pie chart representing sample ethnicity breakdown of the samples collected and tested as part of the one-year extension described in this chapter.

All samples were typed with AmpflSTR® Identifiler™ Plus (Life Technologies) and with the three SNP assays. Of the 154 samples four (4) did not generate interpretable SNP and STR data and a total of eight (8), including the previous four, did not generate interpretable STR data.

As described in the methods chapter of this report (see section 2.2.2) Random Match Probability (RMP) was calculated in the four major US populations, of the unlinked subset of 32 SNPs. The LR was calculated for each sample by dividing the highest RMP obtained

among the four populations by the other three. The number obtained expresses the likelihood of the profile if the sample originated from the population in the numerator versus if the sample originated from the population in the denominator:

LR = highest RMP / second highest RMP.

A threshold of 1000 was empirically chosen above which the LR is considered significant for a sample to be classified as belonging to a specific population (the one in the numerator) while LR values below 1000 were defined as inconclusive (but still informative) between the two populations with highest and second highest RMPs meaning that the individual most likely belongs to one of the two (or both) populations.

The STR RMP for each sample was also calculated in each of the four populations (Table 7) and then factored to the SNP RMP. The LR was then recalculated after including the STR data and the accuracy of the prediction was reevaluated. Results are summarized in figures 32 and 33. The 12 samples that did not belong to any of the four populations (4 mixed, 5 South Asian, 3 Middle Eastern) were initially excluded from the analysis given that population specific allele frequencies are not available for these individuals.

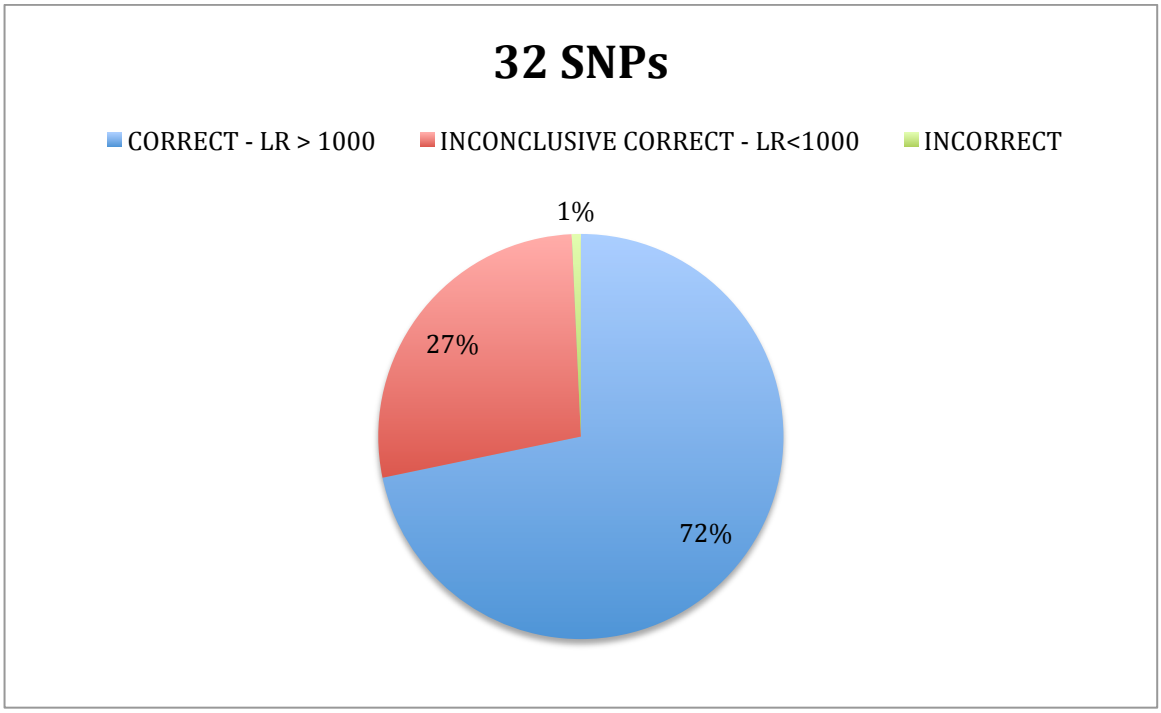


Figure 32. Pie chart summarizing prediction accuracy of the assay with 32 SNPs, 138 individuals are represented in this data set.

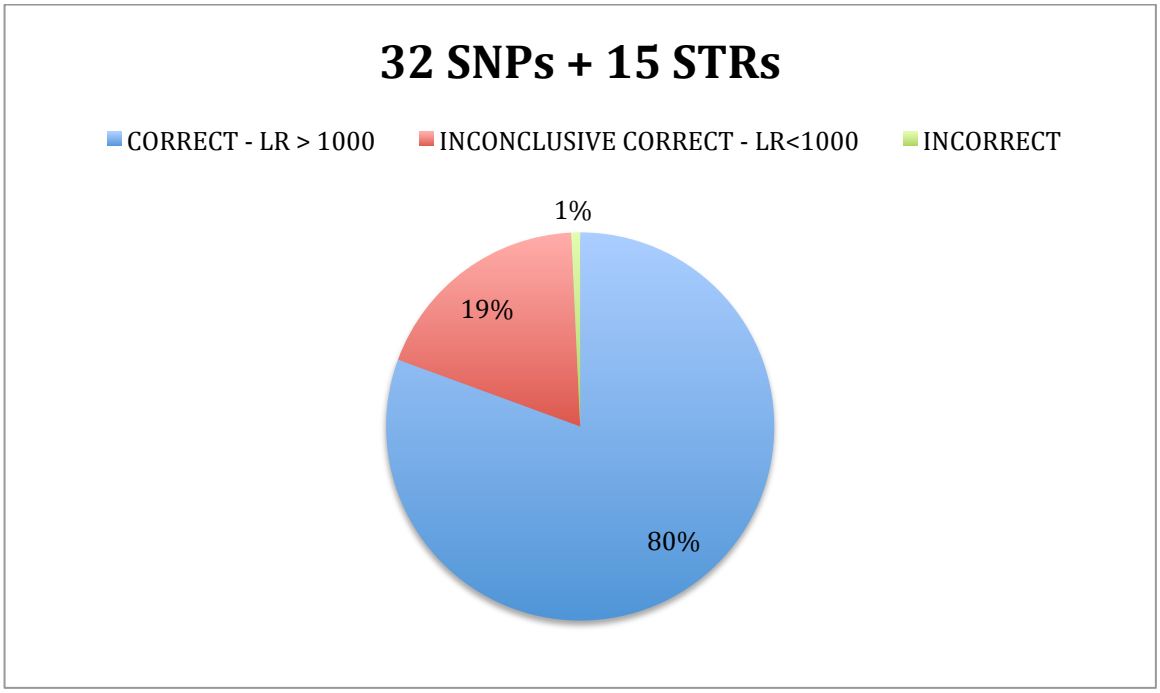


Figure 33. Pie chart summarizing prediction accuracy of the assay with 32 SNPs and 15 STRs, 134 individuals are represented in this data set.

In Both analyses the same individual (and only one individual) was misclassified to be Hispanic/Native American but self reported African American with an overall error rate that can be conservatively approximated to 1%. Unfortunately the misclassified individual is not one of the recently collected buccal swabs, which would have allowed reviewing the questionnaire filled in by the subject together with notes from the collector to verify possible data transfer errors. Other than the self-reported ancestry there is no other information on this sample.

With only the SNP data approximately 72% of the individuals were classified correctly and 27% resulted inconclusive between two populations where one of the two is the correct population of origin of the individual. When including the STR data the accuracy of the prediction increases by 8% which is statistically significant at a 95% confidence (Z-score= - 1.7252154; p=0.0422) but not at a 99%. This indicates that, although STRs were not selected to provide ancestry information, they can improve the prediction of a SNP based assay, thus STR data should be included if available.

Of the 12 samples that did not belong to any of the four major US populations all five South Asians were classified as Hispanic-Native American, three of which with LR < 1000 and the second population being European. Two of the Middle Eastern subjects were classified as European and one as Hispanic-Native American, the latter with a LR < 1000 and the second population being European. Of the mixed individuals two were European/South Asians and both classified as Hispanic-Native American and one of the two had a LR<1000 with the second population being European; one was European/East Asian and classified as Hispanic-Native American, lastly one was European/African American classified as Hispanic-Native American. Upon review of what reported by the last individual it was interesting to notice that the paternal grandparent was Native American.

Results demonstrate that the ancestry prediction power of the assay is robust when individuals are from the main four population groups but that it has limitations when individuals are from different populations (for example South Asia) or of mixed ancestry. Also it is important to note that in the sample set analyzed, as part of this last one-year extension, there is a very limited amount of individual of East Asian ancestry on top of the 32 previously tested. It would be sound to verify the prediction accuracy on a greater number of subjects from that population. Furthermore, to broaden assay capabilities, it would be useful to obtain SNP allele frequencies from other populations (again for example South Asia) and evaluate RMPs and LRs in a similar manner.

An important concern, common to many bio-geographic ancestry studies, is that the ancestry of the samples is self-identified by the donor. This may not truly represent the actual genetic ancestry of individuals but rather their perception, which may be based on the social environment in which they grew up and/or live. This should be taken into account when incorporating the prediction information into an investigation.

A limitation of this study is in the grouping the Native American and the Hispanic population into one. First of all the term Hispanic doesn't refer to a genetically uniform population but rather to Spanish speaking individuals in the US, particularly those of Latin America. Thus it is a very heterogeneous and admixed population, the genetic make up of which may differ significantly in different parts of the US. For example South West Hispanic individuals are mostly a mix between European and Native American ancestries, where as individuals from Puerto Rico have a significant African contribution together with the other two. During the preliminary work it was decided that, given the significant Native American genetic contribution to the majority of individuals that define themselves as Hispanic, and given the limited number of SNPs suitable for a SNaPShot-based forensic assay, it was not possible to distinguish between US-Hispanics and Native American thus they were grouped together. To best address the diversity of this group different populations were chosen for the training set: CLM (Columbians from Medellin, Columbia), MXL (Mexican population, Los Angeles, USA), PUR (Puerto Ricans from Puerto Rico) from the 1000 Genome project data (total 162), a diverse set of self-identified Hispanics collected as part of this project (9 total), and Native Americans from multiple US regions part of the Dr. Schanfiled's sample set (63 total).

The prediction results obtained for self described 'Hispanics' and Native Americans support the fact that this approach could be a practical solution to correctly classify individuals from these populations. Nonetheless, when communicating results of a 'Hispanic / Native American' prediction to interested parties, the limitation of grouping the two populations together needs to be acknowledged together with the fact that the term 'Hispanic' doesn't technically refer to a biogeographic ancestry although, in this context, it is used to indicate the average admixed ancestry individual from Central and South American.

EUROPEAN

D21S11	
<24.2	0.0075
24.2	0.0075
24.3	0.0075
25	0.0075
25.2	0.0075
26	0.0075
26.2	0.0075
27	0.0359
28	0.1633
28.2	0.0075
29	0.19995
29.2	0.0075
29.3	0.0075
30	0.26505
30.2	0.0305
30.3	0.0075
31	0.0773
31.1	0.0075
31.2	0.0968
32	0.01115
32.1	0.0075
32.2	0.0778
32.3	0.0075
<9	0.01745
9	0.00975
10	0.0992
11	0.0716
12	0.1627
13	0.3151
14	0.18975
15	0.10645
16	0.0291
17	0.0075
>17	0.0075

D7S820	
6	0.0075
6.3	0.0075
7	0.01545
8	0.1579
8.1	0.0075
8.2	0.0075
9	0.1766
9.1	0.0075
9.3	0.0075
10	0.2576
10.1	0.0075
10.3	0.0075
11	0.19375
11.3	0.0075
12	0.1568
13	0.0361
14	0.0076
15	0.0075

CSF1PO	
<6	0.0075
7	0.0075
8	0.0075
9	0.0075
10	0.22955
10.3	0.0075
11	0.30955
11.1	0.0075
12	0.34455
12.1	0.0075
13	0.08455
14	0.01115
15	0.0075

D3S1358	
<12	0.0075
12	0.0075
13	0.0146
14	0.1303
15	0.2578
16	0.2404
17.1	0.0075
18	0.1584
19	0.011
>19	0.0075

D13S317	
<8	0.0075
8	0.1174
8.1	0.0075
9	0.0762
10	0.0477
11	0.3185
12	0.278
13	0.11785
13.3	0.0075
14	0.0426
15	0.0075

D16S539	
<8	0.0075
9	0.1088
9.3	0.0075
10	0.05595
11	0.32025
11.3	0.0075
12	0.31415
13	0.1568
14	0.02505
15	0.0075

THO1	
<5	0.0075
6	0.21845
7	0.2039
8	0.0993
9	0.13795
9.3	0.0075
10	0.0075
>10	0.0075

D2S1338	
15	0.0076
16	0.04015
17	0.1777
18	0.071
19	0.12575
20	0.14605
21	0.0334
22	0.03905
23	0.1163
24	0.12025
25	0.0995
26	0.0286
27	0.0076
28	0.0076

D19S433	
9	0.0076
10	0.0076
11	0.0076
11.2	0.0076
12	0.0792
12.2	0.0076
13	0.2712
13.2	0.0126
14	0.355
14.2	0.0133
15	0.1548
15.2	0.0311
16	0.21675
17	0.263
17.2	0.0161
18	0.21245
18.2	0.0076

VWA	
11	0.0075
12	0.0075
13	0.0075
14	0.08855
15	0.1121
16	0.5534
17	0.1175
18	0.0495
19	0.09355
20	0.09975
21	0.0075
>21	0.0075

TPOX	
<6	0.0075
6	0.0075
7	0.0075
8	0.0075
9	0.0075
10	0.0075
11	0.25115
12	0.04415
13	0.0075
>13	0.0075

D18S51	
<11	0.0083
11	0.01425
12	0.133
13	0.1269
13.2	0.0075
14	0.1523
14.2	0.0075
15	0.14755
15.2	0.0075
16	0.13755
16.2	0.0075
17	0.1246
18	0.0767
19	0.0412
20	0.0196
20.2	0.0075
21	0.009
21.2	0.0075
22	0.0075
>22	0.0075

D8S1179	
<7	0.0075
7	0.0075
8	0.0075
9	0.04575
9.2	0.0075
10	0.0527
11	0.3768
12	0.3682
13	0.14785
14	0.0075
15	0.0075
>15	0.0075

FGA	
<18	0.0075
18	0.0266
18.2	0.0075
19	0.0573
19.2	0.0075
19.3	0.0075
20	0.133
20.2	0.0075
21	0.17705
21.2	0.0075
22	0.19405
22.2	0.01245
22.3	0.0075
23	0.14295
23.1	0.0075
23.2	0.0075
23.3	0.0075
24	0.13675
24.2	0.0075
24.3	0.0075
25	0.0785
25.1	0.0075
25.2	0.0075
25.3	0.0075
26	0.0251
26.2	0.0075
27	0.0076
27.2	0.0075
28	0.0075
29	0.0075
29.2	0.0075
30	0.0075
>30	0.0075

ASIAN

D21S11	
<24.2	0.003
24.2	0.005
24.3	0.005
25	0.005
25.2	0.005
26	0.005
26.2	0.005
27	0.004
28	0.0376
28.2	0.0065
29	0.26225
29.2	0.003
29.3	0.005
30	0.3119
30.2	0.00905
30.3	0.005
31	0.1043
31.1	0.005
31.2	0.0632
32	0.02005
32.1	0.005
32.2	0.1284
32.3	0.005
<9	0.005
9	0.0035
10	0.1269
11	0.1003
12	0.11185
13	0.2231
14	0.20405
15	0.149
16	0.07475
17	0.00905
>17	0.005

D7S820	
6	0.005
6.3	0.005
7	0.005
8	0.14145
8.1	0.005
8.2	0.005
9	0.04365
9.1	0.005
9.3	0.004
10	0.18895
10.1	0.005
10.3	0.005
11	0.3631
11.3	0.005
12	0.21765
13	0.0411
14	0.0025
>14	0.005

CSF1PO	
<6	0.005
7	0.0065
8	0.005
9	0.0431
10	0.2418
10.3	0.005
11	0.2157
11.1	0.005
12	0.4087
12.1	0.005
13	0.06565
14	0.01705
15	0.0015

D3S1358	
<12	0.005
12	0.005
13	0.005
14	0.0291
15	0.37155
15.2	0.1695
16	0.26665
17	0.12165
17.1	0.0065
18	0.034
19	0.005
>19	0.0035

D13S317	
<8	0.003
8	0.27685
8.1	0.005
9	0.15245
10	0.1189
11	0.23325
12	0.1664
13	0.0381
13.3	0.005
14	0.01055
15	0.003

D16S539	
<8	0.005
9	0.3063
9.3	0.003
10	0.15835
11	0.23335
11.3	0.005
12	0.2067
13	0.0803
14	0.00955
15	0.001

THO1	
<5	0.005
6	0.005
7	0.27385
8	0.0637
9	0.43895
9.3	0.0035
10	0.0035
>10	0.0035

D2S1338	
15	0.003
16	0.01055
17	0.0872
18	0.12575
19	0.1981
20	0.10135
21	0.0266
22	0.0642
23	0.16755
24	0.1274
25	0.0647
26	0.02105
27	0.004
28	0.0035

D19S433	
9	0.005
10	0.0045
11	0.004
11.2	0.005
12	0.04965
12.2	0.007
13	0.2984
13.2	0.03965
14	0.30425
14.2	0.0888
15	0.0532
15.2	0.1189
16	0.00755
17	0.2567
17.2	0.0251
18	0.1981
18.2	0.0035

VWA	
11	0.005
12	0.0296
13	0.2016
14	0.0247
15	0.1989
16	0.07425
17	0.0336
18	0.005
>21	0.005

TPOX	
<6	0.0035
6	0.005
7	0.005
8	0.47355
9	0.12785
10	0.0271
11	0.3334
12	0.0336
13	0.002
>13	0.0015

D18S51	
<11	0.0015
11	0.0035
12	0.04915
13	0.17805
13.2	0.005
14	0.20915
14.2	0.005
15	0.1795
15.2	0.005
16	0.1304
16.2	0.005
17	0.08025
18	0.0471
19	0.04665
20	0.0276
20.2	0.005
21	0.02155
21.2	0.005
22	0.01505
>22	0.0105

D8S1179	
<7	0.005
7	0.01105
8	0.0075
9	0.0887
9.2	0.005
10	0.20055
11	0.30155
12	0.2257
13	0.1489
14	0.01405
15	0.0035
>15	0.0035

FGA	
<18	0.005
18	0.0286
18.2	0.005
19	0.0757
19.2	0.005
19.3	0.005
20	0.07315
20.2	0.005
21	0.1368
21.2	0.0045
22	0.17905
22.2	0.0035
22.3	0.005
23	0.2021
23.1	0.005
23.2	0.0035
23.3	0.005
24	0.162
24.2	0.005
24.3	0.005
25	0.0788
25.1	0.005
25.2	0.0045
25.3	0.005
26	0.03765
26.2	0.0035
27	0.007
27.2	0.003
28	0.0015
29	0.0035
29.2	0.003
30	0.005
>30	0.0035

AFRICAN AMERICAN

D21S11	
<24.2	0.0085
24.2	0.0085
24.3	0.0085
25	0.0085
25.2	0.0085
26	0.0085
26.2	0.0085
27	0.0642
28	0.24385
28.2	0.0085
29	0.19565
29.2	0.0085
29.3	0.0085
30	0.17315
30.2	0.012
30.3	0.0085
31	0.0804
31.1	0.0085
31.2	0.0634
32	0.0106
32.1	0.0085
32.2	0.0584
32.3	0.0085
<9	0.0085
9	0.0085
10	0.0264
11	0.0421
12	0.13705
13	0.22475
14	0.30055
15	0.19265
16	0.0581
17	0.0106
>17	0.0085

D7S820	
6	0.0085
6.3	0.0085
7	0.0115
8	0.21185
8.1	0.0085
8.2	0.0085
9	0.12115
9.1	0.0085
9.3	0.0085
10	0.33775
10.1	0.0085
10.3	0.0085
11	0.20095
11.3	0.0085
12	0.0974
13	0.0147
14	0.0085
>14	0.0085

CSF1PO	
<6	0.0085
7	0.0496
8	0.0678
9	0.0374
10	0.26785
11	0.30365
11.1	0.0085
12	0.22745
12.1	0.0085
13	0.31035
14	0.19985
15	0.1371
16	0.0636
17	0.0092
>17	0.0085

D3S1358	
<12	0.0085
12	0.0085
13	0.0085
14	0.1047
15	0.30365
15.2	0.0085
16	0.23035
17	0.12165
17.1	0.0065
18	0.0085
19	0.0092
>19	0.0085

D13S317	
<8	0.0085
8	0.0319
8.1	0.0085
9	0.0291
10	0.0304
11	0.27555
12	0.4431
13	0.14955
13.3	0.0085
14	0.0392
15	0.0085

D16S539	
<8	0.0085
9	0.0356
9.3	0.0085
10	0.1126
11	0.31655
11.3	0.0085
12	0.19185
13	0.13325
14	0.0162
15	0.0085

THO1	
<5	0.0085
6	0.0085
7	0.4248
8	0.20065
9	0.1546
9.3	0.0085
10	0.0099
>10	0.0085

D2S1338	
15	0.0085
16	0.0556
17	0.1034
18	0.0475
19	0.14475
20	0.0816
21	0.14205
22	0.13085
23	0.1094
24	0.089
25	0.0766
26	0.0158
27	0.0085
28	0.0085

D19S433	
9	0.085
10	0.0127
11	0.0667
11.2	0.0085
12	0.1109
12.2	0.049
13	0.27215
13.2	0.0547
14	0.21655
14.2	0.0605
15	0.0628
15.2	0.0468
16	0.0169
16.2	0.0254
17	0.0085
17.2	0.0085
18	0.0085
18.2	0.0085

VWA	
11	0.0085
12	0.074
13	0.0482
14	0.0703
15	0.17785
16	0.0085
17	0.16105
18	0.121
19	0.0796
20	0.0565
21	0.0155
22	0.0085
>22	0.0085

TPOX	
<6	0.0085
6	0.0847
7	0.0197
8	0.36665
9	0.19475
10	0.0907
11	0.21665
12	0.0259
13	0.0085
>13	0.0085

D18S51	
<11	0.0085
11	0.0085
12	0.0074
13	0.0482
13.2	0.0085
14	0.0703
14.2	0.0085
15	0.17785
15.2	0.0085
16	0.16165
16.2	0.0085
17	0.16705
18	0.121
19	0.0796
20	0.0565
20.2	0.0085
21	0.0155
21.2	0.0085
22	0.0085
>22	0.0085

D8S1179	
<7	0.0085
7	0.0085
8	0.0513
9	0.0279
9.2	0.0085
10	0.0686
11	0.24395
12	0.35855
13	0.22685
14	0.0199
15	0.0085
>15	0.0085

FGA	
<18	0.0085
18	0.0085
18.2	0.013
19	0.0646
19.2	0.0085
19.3	0.0085
20	0.063
20.2	0.0085
21	0.12245
21.2	0.0085
22	0.20585
22.2	0.0085
22.3	0.0085
23	0.16045
23.1	0.0085
23.2	0.0085
23.3	0.0085
24	0.14855
24.2	0.0085
24.3	0.0085
25	0.1019
25.1	0.0085
25.2	0.0085
25.3	0.0085
26	0.058
26.2	0.0085
27	0.0206
27.2	0.0085
28	0.013
29	0.0085
29.2	0.0085
30	0.0085
>30	0.0092

HISPANIC/NA

D21S11	
<24.2	0.01333333
24.2	0.01333333
24.3	0.01333333
25	0.01333333
25.2	0.01333333
26	0.01333333
26.2	0.01333333
27	0.02036667
28	0.0834
28.2	0.01333333
29	0.1932
29.2	0.01333333
29.3	0.01333333
30	0.299
30.2	0.02886667
30.3	0.01333333
31	0.06893333
31.1	0.01333333
31.2	0.12856667
32	0.0155
32.1	0.01333333
32.2	0.1184
32.3	0.01333333
<9	0.01333333
9	0.01333333
10	0.0772
11	0.04986667
12	0.1263
13	0.32386667
14	0.2728
15	0.1106
16	0.0216
17	0.01333333
>17	0.01333333

D7S820	
6	0.01333333
6.3	0.01333333
7	0.01606667
8	0.12303333
8.1	0.01333333
8.2	0.01333333
9	0.08476667
9.1	0.01333333
9.3	0.01333333
10	0.2623333
10.1	0.01333333
10.3	0.01333333
11	0.27763333
11.3	0.01333333
12	0.20216667
13	0.03483333
14	0.01333333
>14	0.01333333

CSF1PO	
<6	0.01333333
7	0.01433333
8	0.01333333
9	0.03793333
10	0.2573
11	0.36026667
11.1	0.01333333
12	0.0573
12.1	0.01333333
13	0.0495
14	0.01333333
15	0.01333333

D3S1358	
<12	0.01333333
12	0.01333333
13	0.01333333
14	0.1546
15	0.36426667
15.2	0.01333333
16	0.29433333
17	0.1546
17.1	0.01333333
18	0.0995
19	0.01376667
>19	0.01333333

D13S317	
<8	0.01333333
8	0.01333333
8.1	0.01333333
9	0.18306667
10	0.10916667
11	0.23683333
12	0.22
13	0.0994
13.3	0.01333333
14	0.0602
15	0.01333333

D16S539	
<8	0.01333333
9	0.11836667
9.3	0.01333333
10	0.1413333
11	0.29053333
11.3	0.01333333
12	0.27543333
13	0.13626667
14	0.01986667
15	0.01333333

THO1	
<5	0.01333333
6	0.0891
7	0.184
8	0.1236
9	0.0285
9.3	0.01333333
10	0.0542
10.3	0.01333333
11	0.0266
11.1	0.01333333
12	0.0242
12.1	0.01333333
13	0.0092
13.3	0.01333333
14	0.0162
15	0.01333333
>15	0.01333333
>10	0.01333333

D2S1338	
15	0.01333333
16	0.02876667
17	0.1692
18	0.0707
19	0.23413333
20	0.1236
21	0.0285
22	0.09563333
23	0.1092
24	0.078
25	0.0527
26	0.01323333
27	0.01323333
28	0.01323333

D19S433	
9	0.01323333
10	0.01323333
11	0.01323333
11.2	0.01323333
12	0.0525
12.2	0.0167
13	0.1961
13.2	0.0909
14	0.31363333
14.2	0.0432
15	0.12966667
15.2	0.07706667
16	0.03446667
16.2	0.0242
17	0.01323333
17.2	0.01846667
18	0.01323333
18.2	0.01323333

VWA	
11	0.01333333
12	0.01333333
13	0.01333333
14	0.0665
15	0.1129
16	0.3123
17	0.2562
18	0.17
19	0.0656
20	0.01493333
21	0.01333333
>21	0.01333333

TPOX	
<6	0.01333333
6	0.01333333
7	0.01333333
8	0.44906667
9	0.07276667
10	0.03753333
11	0.31336667
12	0.12046667
13	0.01333333
>13	0.01333333

D18S51	
<11	0.01333333
11	0.01436667
12	0.12353333
13	0.1158
13.2	0.01333333
14	0.18793333
14.2	0.01333333
15	0.1583
15.2	0.01333333
16	0.12016667
16.2	0.01333333
17	0.139
18	0.0538
19	0.0399
20	0.02423333
20.2	0.01333333
21	0.0138
21.2	0.01333333
22	0.01333333
>22	0.01333333

D8S1179	
<7	0.01333333
7	0.0891
8	0.01333333
9	0.05163333
9.2	0.01333333
10	0.05153333
11	0.38413333
12	0.29203333
13	0.11503333
14	0.01333333
15	0.01333333
>15	0.01333333

FGA	
<18	0.01333333
18	0.01336667
18.2	0.01333333
19	0.07893333
19.2	0.01333333
19.3	0.01333333
20	0.09536667
20.2	0.01333333
21	0.14763333
21.2	0.01333333
22	0.14236667
22.2	0.01333333
22.3	0.01333333
23	0.13823333
23.1	0.01333333
23.2	0.01333333
23.3	0.01333333
24	0.1535
24.2	0.01333333
24.3	0.01333333
25	0.1346
25.1	0.01333333
25.2	0.01333333
25.3	0.01333333
26	0.05583333
26.2	0.01333333
27	0.0267
27.2	0.01333333
28	0.01333333
29	0.01333333
29.2	0.01333333
30	0.01333333
>30	0.01333333

Table 7. STR allele frequency used for RMP calculation

7.3 Evaluation of Forensic Resource/Reference on Genetics Knowledge Base (FROG-kb) for the Prediction of Individual Biogeographic Ancestry (25 SNP panel)

Currently, forensic investigations most commonly employ Short Tandem Repeat (STR) analysis of evidence and compare the resulting profile to known profiles or databases such as CODIS. However, in forensic investigations when a DNA profile derived from the evidence does not match identified suspects or profiles from available databases, additional DNA analyses, such as those targeted at inferring the possible ancestral origin of the perpetrator, could yield valuable information. In recent years there have been many proposed Ancestry Informative Marker (AIM) sets for use in predicting biogeographic ancestry [Bouakaze et al. 2009, Gettings et al. 2013, Gettings et al. 2014 , Halder 2008, Kidd 2011, Lowe et al. 2001, Nassir 2009]. There are a range of DNA polymorphisms available with potential to be used as AIMs including autosomal and Y-chromosome STRs, mitochondrial sequence variation (mtDNA) and Single Nucleotide Polymorphisms (SNPs). SNPs are the most common form of genetic polymorphism; while they do not have the same power of discrimination as STRs for individual identification, some known as Ancestry Informative SNPs (AISNPs) have alleles associated with specific populations and/or correlated with phenotypic characteristics which can be helpful in forensic investigations when STR profiles fail to yield an identification.

The purpose of this project was to evaluate the Forensic Resource/Reference on Genetics Knowledge Base (FROG-kb - frog.med.yale.edu/FrogKB/) as a tool for the prediction of an individual's biogeographic ancestry. FROG-kb is a freely available online tool with the primary objective of providing a web interface with the data housed in the already extensively used and referenced Allele FREquency Database (http://alfred.med.yale.edu/) making it more suitable for forensic purposes [Rajeevan et al. 2011]. FROG-kb provides the ability to display the ALFRED data in an organized manner as well as computational tools that use the underlying allele frequencies with user-provided data. Multiple Individual Identification SNP (IISNP), AISNP, and Phenotype Informative SNP (PISNP) panels are available on FROG-kb. Figure 34 is a screenshot of the FROG-kb website. This project specifically evaluates the "Daniele Podini's list of 32 AISNPs", a panel consisting of 25 AISNPs based on the 50-SNP assay developed by Gettings et al. and designed to predict ancestry among the primary U.S. populations. Although the 50-SNP assay used to derive the "Daniele Podini's list of 32 AISNPs" panel contained 32 AISNPs, only 25 AISNPs are used in the panel for computation. All of these 25 SNPs have data on all of the populations listed and are therefore appropriate for calculations of likelihoods. The remaining seven SNPs have data on fewer and diverse populations. What data are available for those seven are in ALFRED. When they have complete data, they will be included in the panel of SNPs available for analysis. The accuracy of this panel of AISNPs has been demonstrated by Gettings et al., and its utilization can aid in the evaluation of the quality of FROG-kb as an ancestry prediction tool.

Figure 34. Screenshot of FROG-kb website and the AISNP sets available including the "Daniele Podini's List of 32 AISNPs" panel evaluated in this project.

While the literature has described the use of general or proprietary software and more-tedious researcher-executed calculations [Bouakaze et al. 2009, Evett et al. 1992, Gettings et al. 2013, Phillips et al. 2007] for the prediction of biogeographic ancestry from a DNA profile, FROG-kb is an easily accessed tool designed specifically for use by the forensic community and can efficiently compute the likelihood probabilities of 89 populations from multiple geographic regions. Unique from much of the literature, instead of simply classifying samples based on the major U.S. populations (African American, East Asian, European, and Admixed (Hispanic/Native American)), the 89 populations are categorized by 8 geographic regions: Africa, Asia, East Asia, Europe, North America, Oceania, Siberia, and South America, which could potentially provide a greater wealth of information in forensic

investigations as it is not only informative for origins from major geographic regions but also informative for distinguishing relationships within several of those regions.

Ancestry Informative Markers and AISNP Assay: This project utilized the 50 SNP assay developed by Gettings et al. 2013. The author considered 103 candidate SNPs chosen from the relevant literature. Selected from the candidate SNPs, the 50 SNP assay consists of 18 Phenotype Informative SNPs (PISNPs) and 32 AISNPs. Of the 32 AISNPs, only 25 were analyzed using FROG-kb due to limited data on fewer and diverse populations Table 8 lists the 25 AISNPs that were analyzed in FROG-kb using the "Daniele Podini's List of 32 AISNPs" panel. Table 9 lists the 89 populations which were evaluated in FROG-kb.

Samples: The known, self-reported ancestries for all sample tested fall into the classification of the 4 major U.S. populations: African American, East Asia, European, and Hispanic/Native American.

I. The test set of 127 samples used to develop the FROG-kb result evaluation criteria consisted of 31 European American, 32 African American, 32 Hispanic American, and 32 East Asian samples. The majority of these test samples (European American, African American, and Hispanic samples) were standards obtained from the National Institute of Standards and Technology (NIST). The East Asian samples were internally available. The known population origins of the samples were originally based on individual self-identification, and have been consistently tested and referenced as standards for research purposes.
II. Further testing was conducted using genotypes downloaded from the 1000 Genomes Project [http://browser.1000genomes.org]. These 200 sample profiles were tested blind of the known, self-reported population ancestry.
III. Samples were also collected from anonymous volunteers in the Washington, D.C. area using a George Washington University (GWU) IRB approved protocol, consisting of a comprehensive questionnaire regarding multiple aspects of their ancestry information and three buccal swabs collected after the volunteer read an assent form. Other information was collected regarding phenotypic characteristics that were for assessment on another project. These samples were tested blind of the known, self-reported population ancestry.
IV. Additional anonymous DNA samples with known, self-reported ancestry were obtained from Dr. Moses Schanfield, GWU Department of Forensic Sciences (samples previously ruled "NOT human subjects research by the GWU IRB).These samples were tested blind of the known ancestry.

TABLE 8: List of the 25 SNPs analyzed in the AISNP Panel used in FROG-kb		
Panel: Daniele Podini's List of 32 AISNPs		
dbSNP rs Number	**Chromosome**	**Chromosome Position**
rs10007810	4	41,554,364
rs10108270	8	4,190,793
rs1042602	11	88,911,696
rs10496971	2	145,769,943
rs12821256	12	89,328,335
rs12896399	14	92,773,663
rs12913832	15	28,365,618
rs1344870	3	21,307,401
rs1426654	15	48,426,484
rs16891982	5	33,951,693
rs1876482	2	17,362,568
rs2065982	13	34,864,240
rs2814778	1	159,174,683
rs3737576	1	101,709,563
rs3784230	14	105,679,055
rs3827760	2	109,513,601
rs4891825	18	67,867,663
rs4918842	10	115,306,802
rs6451722	5	43,711,378
rs6548616	3	79,399,575
rs714857	11	15,974,389
rs722869	14	97,277,005
rs730570	14	101,142,890
rs896788	2	7,149,155
rs952718	2	215,888,624

Table 8. List of the 25 AISNPs analyzed in the "Daniele Podini's List of 32 AISNPs" panel in FROG-kb. These are the SNPs that have data on all the populations tested and are therefore appropriate for calculations of likelihoods. The remaining seven SNPs have data on fewer and diverse populations. What data are available for those seven are in ALFRED. When they have complete data, they will be included in the SNPs available for analysis.

TABLE 9: Populations included in computing match probability in FROG-kb for the "Daniele Podini's List of 32 AISNPs" panel		
Populations	**Geographic Region**	**Sample Size (2N)**
African American	Africa	182
Biaka	Africa	140
Chagga	Africa	90
Ethiopian Jews	Africa	64
Hausa	Africa	78
Ibo	Africa	96
Mandenka, HGDP-CEPH	Africa	48
Masai	Africa	44
Mbuti	Africa	78
Mbuti, HGDP-CEPH	Africa	30
Mozabite, HGDP-CEPH	Africa	60
San, HGDP-CEPH	Africa	14
Sandawe	Africa	80
Yoruba	Africa	156
Yoruba, HGDP-CEPH	Africa	50
Zaramo	Africa	80
Balochi, HGDP-CEPH	Asia	50
Brahui, HGDP-CEPH	Asia	50
Burusho, HGDP-CEPH	Asia	50
Druze	Asia	212
Hazara, HGDP-CEPH	Asia	50
Kalash, HGDP-CEPH	Asia	50
Keralite	Asia	60
Khanty	Asia	100
Komi-Zyrian	Asia	94
Kuwaiti	Asia	32
Lao Long	Asia	238
Makrani, HGDP-CEPH	Asia	50
Mongola, HGDP-CEPH	Asia	20
Oroqen, HGDP-CEPH	Asia	20
Palestinian, HGDP-CEPH	Asia	102
Sindhi, HGDP-CEPH	Asia	50
Yemenite Jews	Asia	146
Ami	East Asia	80
Atayal	East Asia	84
Cambodian	East Asia	52
Cambodian, HGDP-CEPH	East Asia	22

Dai, HGDP-CEPH	East Asia	20
Daur, HGDP-CEPH	East Asia	20
Hakka	East Asia	86
Hezhen, HGDP-CEPH	East Asia	20
Japanese	East Asia	112
Korean	East Asia	132
Lahu, HGDP-CEPH	East Asia	20
Miaozu, HGDP-CEPH	East Asia	20
Naxi, HGDP-CEPH	East Asia	20
San Francisco Chinese	East Asia	124
She, HGDP-CEPH	East Asia	20
Taiwanese Han	East Asia	100
Tu, HGDP-CEPH	East Asia	20
Tujia, HGDP-CEPH	East Asia	20
Uygur, HGDP-CEPH	East Asia	20
Xibo, HGDP-CEPH	East Asia	18
Yizu, HGDP-CEPH	East Asia	20
Adygei	Europe	108
Adygei, HGDP-CEPH	Europe	34
Ashkenazi Jews	Europe	166
Basques, HGDP-CEPH	Europe	48
Chuvash	Europe	84
Danes	Europe	102
Finns	Europe	72
French, HGDP-CEPH	Europe	58
Irish	Europe	232
Mixed Europeans	Europe	190
Orcadians, HGDP-CEPH	Europe	32
Russians	Europe	96
Russians, Archangel'sk	Europe	68
Russians, HGDP-CEPH	Europe	50
Samaritans	Europe	82
Sardinian, HGDP-CEPH	Europe	56
Tuscan, HGDP-CEPH	Europe	16
Arizona Pima	North America	104
Cheyenne	North America	112
Maya, HGDP-CEPH	North America	50
Maya, Yucatec	North America	106
Mexican Pima	North America	106
Pima, HGDP-CEPH	North America	50

Micronesia	Oceania	78
Nasioi	Oceania	48
Papuan, HGDP-CEPH	Oceania	34
Yakut	Siberia	102
Yakut, HGDP-CEPH	Siberia	50
Amerindian, HGDP-CEPH	South America	26
Karitiana	South America	114
Karitiana, HGDP-CEPH	South America	48
Peruvian Quechuan	South America	44
Surui, HGDP-CEPH	South America	42
Surui, Rondonia	South America	100
Ticuna	South America	134

Table 9. Table of FROG-kb populations included in computing match probability for the "Daniele Podini's List of 32 AISNPs" panel. These are the populations for which all SNPs in the panel have allele frequency data; the populations are sorted by the biogeographic region.

FROG-kb: FROG-kb offers two options for users to enter genotype data for a selected panel: "Selection by Radio Button" and "File Upload." Both data entry options result in identical profiles. Sample genotypes for each AISNP were entered into FROG-kb. Data was entered for the "Daniele Podini's List of 32 AISNPs" panel. Table 10 is an example of a data input template for the "File Upload" option.

TABLE 10: FROG-kb Data Input Template for File Upload for "Daniele Podini's List of 32 AISNPs" panel (Sample # 25OAS)

ai32 Podini's list of 32 AISNPs

ALFRED_UID	dbSNP_rsnumber	chrom	chrom_pos	alleles	genotype
SI014380Q	rs10007810	4	41554364	A/G	GG
SI014484V	rs10108270	8	4190793	A/C	AC
SI018380U	rs1042602	11	88911696	A/C	CC
SI014477X	rs10496971	2	145769943	G/T	GG
SI166188E	rs12821256	12	89328335	C/T	TT
SI168220T	rs12896399	14	92773663	G/T	GT
SI007119S	rs12913832	15	28365618	A/G	AA
SI007821S	rs1344870	3	21307401	A/C	AC
SI007419V	rs1426654	15	48426484	A/G	GG
SI003963V	rs16891982	5	33951693	C/G	CC
SI011374Q	rs1876482	2	17362568	C/T	TT
SI007623S	rs2065982	13	34864240	C/T	CC
SI007627W	rs2814778	1	159174683	C/T	TT
SI014459X	rs3737576	1	101709563	A/G	AA
SI014399A	rs3784230	14	105679055	C/T	TT
SI663326A	rs3827760	2	109513601	C/T	CC

SI014465U	rs4891825	18	67867663	A/G	AA
SI014409S	rs4918842	10	115306802	C/T	CC
SI014405O	rs6451722	5	43711378	A/G	AG
SI014471R	rs6548616	3	79399575	C/T	AG
SI001818S	rs714857	11	15974389	C/T	TT
SI003730N	rs722869	14	97277005	C/G	GG
SI008734W	rs730570	14	101142890	A/G	GG
SI008732U	rs896788	2	7149155	A/G	GG
SI004800M	rs952718	2	215888624	A/C	CC

Table 10. Example file upload template for "Daniele Podini's List of 32 AISNPs" panel into FROG-kb. The genotype sample depicted is sample #25OAS.

FROG-kb then outputs results of probability calculations. The calculations start by assuming Hardy-Weinberg proportions of the genotypes based on the allele frequencies available for each SNP in each population. This probability of each genotype is stored after being pre-calculated from the allele frequencies in ALFRED. Thus, the probability of each genotype at one locus is given as:

P_r (homozygous allele 1) = (frequency allele 1) $_2$ = p_2

P_r (heterozygous) = 2*(frequency allele 1)*(frequency allele 2) = 2pq

P_r (homozygous allele 2) = (frequency allele 2) $_2$ = q_2

The probability of a specific multi-locus genotype in a specific population is the product across all loci of the locus-specific genotype probabilities. The program implements the product rule by using the genotype probabilities corresponding to the genotypes entered by the user. The calculation is repeated for each of the 89 populations and the population-specific probabilities are provided in an output ordered by the probabilities from highest to lowest. Table 11 is an example of a result output table for a sample evaluated using the "Daniele Podini's List of 32 AISNPs" panel.

The output table consists of the population tested including the biogeographic region, the probability of the entered genotype in each population tested, and the likelihood ratio which is the ratio between the probability of the profile in the population in which it is highest over the probability of the profile in the specified population. The likelihood ratio expresses how much more likely it is to observe that sample profile if it originated from the population at the numerator versus if it originated from the population at the denominator:

$$LR_1 = RMP_{Highest} / RMP_{Specified}$$

The resulting outputs for all samples in the test set were evaluated and the likelihood calculations were examined at the first, second, and third highest orders of magnitude and compared to the known populations to the sample. Evaluation criteria and a prediction model were then established with the goal to generate a prediction of ancestry inferring that a sample (1) originated from an individual population, (2) was admixed between multiple populations, or (3) was inconclusive. The prediction model was then tested on the 1000 Genomes Project downloaded and volunteer samples.

7.3.1 Evaluation Criteria and Prediction Model Determination

Based on the initial output results of the test set (Sample set I), it was determined that the number of populations from each geographic region would be evaluated as a percentage of the total number of populations with likelihood ratios within the first through third orders of magnitude from the population with the highest probability. The total number of populations was considered the best approach since there were no uniform numbers of sample sizes per population or populations per region. An individual was assigned to a biogeographic region using the following methodology (described below and in Figure 35):

A. If a single biogeographic region's populations composed of 55% or more of the populations in the first through third orders or magnitude, then the sample was classified as originating from said region. = Identification, Region 1
B. If a single region's populations do not reach 55%, then the two most numerous regions' populations were added together.
    1. If the total percentage of the two most numerous regions' populations composed of 80% or more of the populations, then the sample was classified as an admixture of those two regions. = Admixture, Region 1 and Region 2
        i. If an individual sample is an admixture it is classified as Hispanic/Native American since the Hispanic/Native American major U.S. population classification is actually an admixed population.
    2. If there was a tie resulting in three regions' populations being considered (e.g. 50% Africa, 20% Europe, and 20% East Asia) then the next level of evaluation was used.
C. If the two most numerous regions' populations did not reach 80%, then the three most numerous regions' populations were added together.
    1. If the total percentage of the three most numerous regions' populations composed of 90% or more of the populations, then the sample was classified as an admixture of those three regions. = Admixture, Region 1, Region 2, and Region 3
        i. If an individual sample is an admixture it is classified as Hispanic/Native American since the Hispanic/Native American major U.S. population classification is actually an admixed population.
    2. If there was a tie resulting in more than three regions' populations being considered (e.g. 50% Africa, 20% Europe, 10% East Asia, 10% Siberia) then the sample was considered inconclusive. = Inconclusive
D. If the three most numerous regions' populations did not reach 90%, then the sample was classified as inconclusive. = Inconclusive

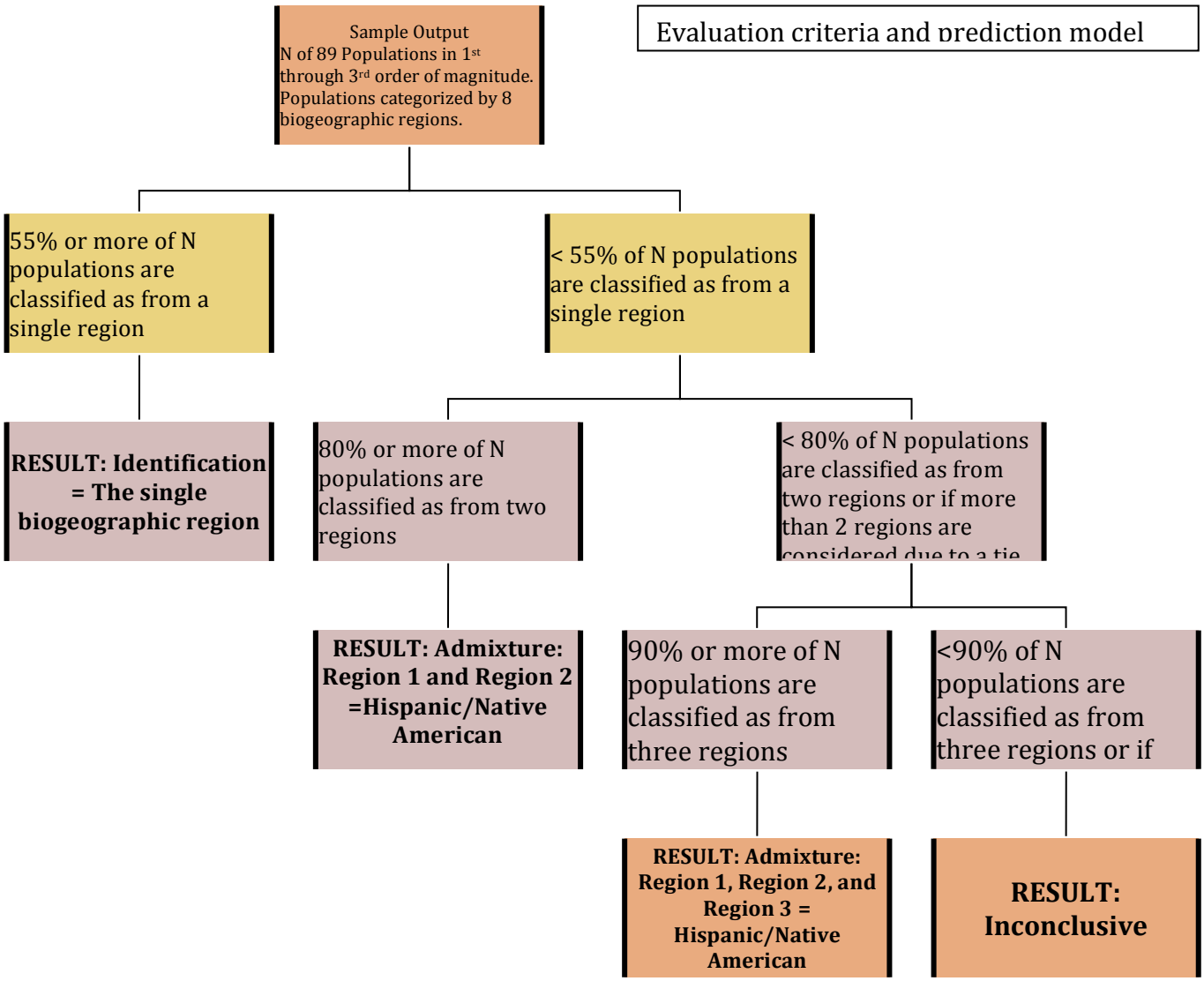


Figure 35. Evaluation criteria and prediction model for FROG-kb results based on the number of populations (N) in the first through third orders of magnitude of likelihood.

Evaluation of FROG-kb with Sample Set II: Table 12 displays the criteria for the interpretation of the FROG-kb results compared to the samples' known population based on the prediction model. Sample set II consisting of 200 profiles downloaded from the 1000 Genomes Project was evaluated using the prediction model previously described. Figure 36 depicts the summary of the performance of FROG-kb in the classification of samples from each of the four major U.S. populations separately. 100% correct identification was observed for known African American, East Asian, and European samples. The FROG-kb predictive model results for Hispanic/Native American samples were 74% correctly identified, 15% misclassified, and 11% were classified as inconclusive.

TABLE 11: FROG-kb Result Output Table (Sample #25OAS)		
Population (Region, Sample Size 2N)	**Probability of Genotype in each Population**	**Likelihood Ratio**
Daur_ HGDP-CEPH (EastAsia,**20**)	0.000057	
Tujia_ HGDP-CEPH (EastAsia,**20**)	0.000052	1.1
Atayal (EastAsia,**84**)	0.000045	1.3
Korean (EastAsia,**132**)	0.000024	2.4
Taiwanese Han (EastAsia,**100**)	0.000019	3
Hakka (EastAsia,**86**)	0.000016	3.7
Miaozu_ HGDP-CEPH (EastAsia,**20**)	0.000014	4.2
San Francisco Chinese (EastAsia,**124**)	0.000011	5.1
Yizu_ HGDP-CEPH (EastAsia,**20**)	0.0000089	6.4
Japanese (EastAsia,**112**)	0.0000069	8.3
Cambodian_ HGDP-CEPH (EastAsia,**22**)	0.0000056	10
Tu_ HGDP-CEPH (EastAsia,**20**)	0.000005	11
Lao Long (Asia,**238**)	0.0000043	13
Yakut (Siberia,**102**)	0.0000027	21
Cambodian (EastAsia,**52**)	0.0000026	22
Oroqen_ HGDP-CEPH (Asia,**20**)	0.0000025	22
Mongola_ HGDP-CEPH (Asia,**20**)	0.0000022	26
Dai_ HGDP-CEPH (EastAsia,**20**)	0.0000021	27
Ami (EastAsia,**80**)	0.0000019	30
She_ HGDP-CEPH (EastAsia,**20**)	0.0000019	30
Xibo_ HGDP-CEPH (EastAsia,**18**)	0.0000018	32
Micronesia (Oceania,**78**)	0.0000012	49
Lahu_ HGDP-CEPH (EastAsia,**20**)	0.000001	57
Hezhen_ HGDP-CEPH (EastAsia,**20**)	8.10E-07	70
Naxi_ HGDP-CEPH (EastAsia,**20**)	1.10E-07	5.00E+02
Yakut_ HGDP-CEPH (Siberia,**50**)	5.00E-08	1.10E+03
Nasioi (Oceania,**48**)	1.80E-08	3.10E+03
Maya_ Yucatec (NorthAmerica,**106**)	6.40E-09	8.90E+03
Maya_ HGDP-CEPH (NorthAmerica,**50**)	4.10E-09	1.40E+04
Cheyenne (NorthAmerica,**112**)	7.10E-10	8.00E+04
Ticuna (SouthAmerica,**134**)	6.80E-10	8.40E+04
Surui_ Rondonia (SouthAmerica,**100**)	1.20E-10	4.90E+05
Papuan_ HGDP-CEPH (Oceania,**34**)	5.40E-11	1.00E+06
Arizona Pima (NorthAmerica,**104**)	4.40E-11	1.30E+06
Peruvian Quechuan (SouthAmerica,**44**)	2.20E-11	2.60E+06
Mexican Pima (NorthAmerica,**106**)	8.40E-12	6.80E+06
Surui_ HGDP-CEPH (SouthAmerica,**42**)	2.30E-12	2.50E+07
Amerindian_ HGDP-CEPH (SouthAmerica,**26**)	2.30E-12	2.50E+07

Pima_ HGDP-CEPH (NorthAmerica,**50**)	1.80E-12	3.10E+07
Uygur_ HGDP-CEPH (EastAsia,**20**)	1.80E-13	3.10E+08
Khanty (Asia,**100**)	1.20E-13	4.90E+08
Karitiana (SouthAmerica,**114**)	5.50E-15	1.00E+10
Karitiana_ HGDP-CEPH (SouthAmerica,**48**)	4.00E-16	1.40E+11
Keralite (Asia,**60**)	3.60E-17	1.60E+12
Hazara_ HGDP-CEPH (Asia,**50**)	2.80E-17	2.00E+12
Balochi_ HGDP-CEPH (Asia,**50**)	5.10E-18	1.10E+13
Mozabite_ HGDP-CEPH (Africa,**60**)	1.20E-18	4.70E+13
Burusho_ HGDP-CEPH (Asia,**50**)	1.10E-19	5.40E+14
Chuvash (Europe,**84**)	5.50E-21	1.00E+16
Brahui_ HGDP-CEPH (Asia,**50**)	2.80E-21	2.00E+16
Makrani_ HGDP-CEPH (Asia,**50**)	1.20E-22	4.60E+17
Kalash_ HGDP-CEPH (Asia,**50**)	1.60E-23	3.50E+18
Ethiopian Jews (Africa,**64**)	3.30E-24	1.70E+19
African American (Africa,**182**)	1.80E-24	3.20E+19
Kuwaiti (Asia,**32**)	1.80E-24	3.20E+19
Tuscan_ HGDP-CEPH (Europe,**16**)	2.30E-25	2.50E+20
Adygei_ HGDP-CEPH (Europe,**34**)	9.70E-26	5.90E+20
Adygei (Europe,**108**)	7.30E-26	7.80E+20
Sardinian_ HGDP-CEPH (Europe,**56**)	4.20E-26	1.30E+21
Sindhi_ HGDP-CEPH (Asia,**50**)	1.20E-26	4.90E+21
Komi-Zyrian (Asia,**94**)	1.10E-27	5.20E+22
Palestinian_ HGDP-CEPH (Asia,**102**)	8.40E-28	6.80E+22
Finns (Europe,**72**)	6.60E-28	8.70E+22
San_ HGDP-CEPH (Africa,**14**)	6.00E-28	9.50E+22
Russians_ HGDP-CEPH (Europe,**50**)	1.30E-28	4.30E+23
Russians (Europe,**96**)	9.10E-29	6.20E+23
Yemenite Jews (Asia,**146**)	4.70E-29	1.20E+24
Ashkenazi Jews (Europe,**166**)	1.70E-29	3.40E+24
Samaritans (Europe,**82**)	5.10E-30	1.10E+25
Russians_ Archangel'sk (Europe,**68**)	4.40E-30	1.30E+25
Sandawe (Africa,**80**)	3.30E-30	1.70E+25
Druze (Asia,**212**)	1.10E-30	5.00E+25
Mandenka_ HGDP-CEPH (Africa,**48**)	2.40E-31	2.40E+26
Orcadians_ HGDP-CEPH (Europe,**32**)	1.80E-31	3.20E+26
French_ HGDP-CEPH (Europe,**58**)	1.70E-31	3.30E+26
Mbuti_ HGDP-CEPH (Africa,**30**)	1.50E-31	3.70E+26
Danes (Europe,**102**)	6.60E-32	8.60E+26
Chagga (Africa,**90**)	1.70E-32	3.50E+27
Mixed Europeans (Europe,**190**)	1.10E-32	5.30E+27
Masai (Africa,**44**)	6.20E-33	9.10E+27

Hausa (Africa,**78**)	4.80E-33	1.20E+28
Irish (Europe,**232**)	7.70E-34	7.40E+28
Zaramo (Africa,**80**)	7.40E-34	7.80E+28
Basques_ HGDP-CEPH (Europe,**48**)	3.70E-34	1.50E+29
Mbuti (Africa,**78**)	4.20E-35	1.40E+30
Yoruba_ HGDP-CEPH (Africa,**50**)	7.20E-36	8.00E+30
Ibo (Africa,**96**)	2.10E-36	2.70E+31
Yoruba (Africa,**156**)	1.20E-38	4.70E+33
Biaka (Africa,**140**)	6.50E-40	8.80E+34

Table 11. Example FROG-kb result output table for sample #25OAS. Output table displays the FROG-kb populations being tested, as well as their geographic region and sample size, in order of highest probability of the genotype in that population and the likelihood ratio for each subsequent population compared to the most probable population.

TABLE 12: FROG-kb Result Interpretation Criteria		
Known Population Based on the Conventional 4 Major U.S. Populations	**Prediction of Ancestry in FROG-kb Based on the 8 FROG-kb Biogeographic Regions**	**Interpretation of Result**
African American	Africa	Correct
	Other single region	Misclassified
	Admixture of 2 or 3 population regions	Misclassified
	Inconclusive	Inconclusive
East Asian	East Asia	Correct
	Other single region	Misclassified
	Admixture of 2 or 3 population regions	Misclassified
	Inconclusive	Inconclusive
European	Europe	Correct
	Other single region	Misclassified
	Admixture of 2 or 3 population regions	Misclassified
	Inconclusive	Inconclusive
Hispanic/Native American	Africa	Misclassified
	East Asia	Misclassified
	Europe	Misclassified
	Other single region	Correct
	Admixture of 2 or 3 population regions	Correct
	Admixture of more than 3 population regions	Inconclusive
	Inconclusive	Inconclusive

Table 12. Interpretation criteria for FROG-kb results; interpretation of the predicted biogeographic region based on FROG-kb results compared to the known major U.S. population. "Other single region" refers to the other biogeographic regions in FROG-kb that do not correspond to the major U.S. populations (Asia, North America, Oceania, Siberia, and South America).

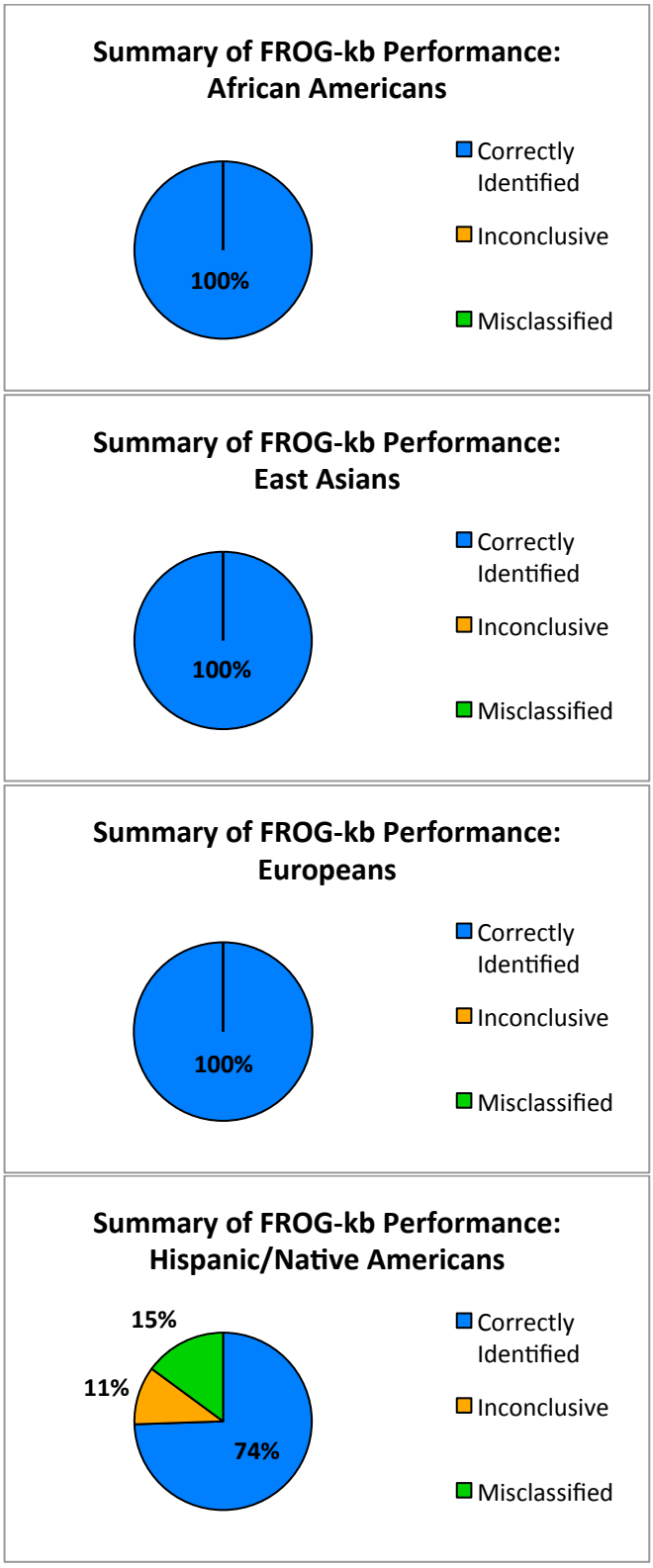


Figure 36. Summary of FROG-kb prediction model performance for the four major U.S. populations. 100% of African Americans, East Asians, and Europeans were correctly identified. 74% of Hispanic/Native Americans were correctly identified, 15% were misclassified, and 11% were classified as inconclusive.

Figure 37 depicts an additional summary of the classification of Hispanic/Native American samples showing that of the 74% correctly identified, 38% were correctly identified as an admixture of two or three biogeographic regions, and 36% were correctly identified as a result of identification of a single "other" biogeographic region (Asia, North America, Oceania, Siberia, or South America).

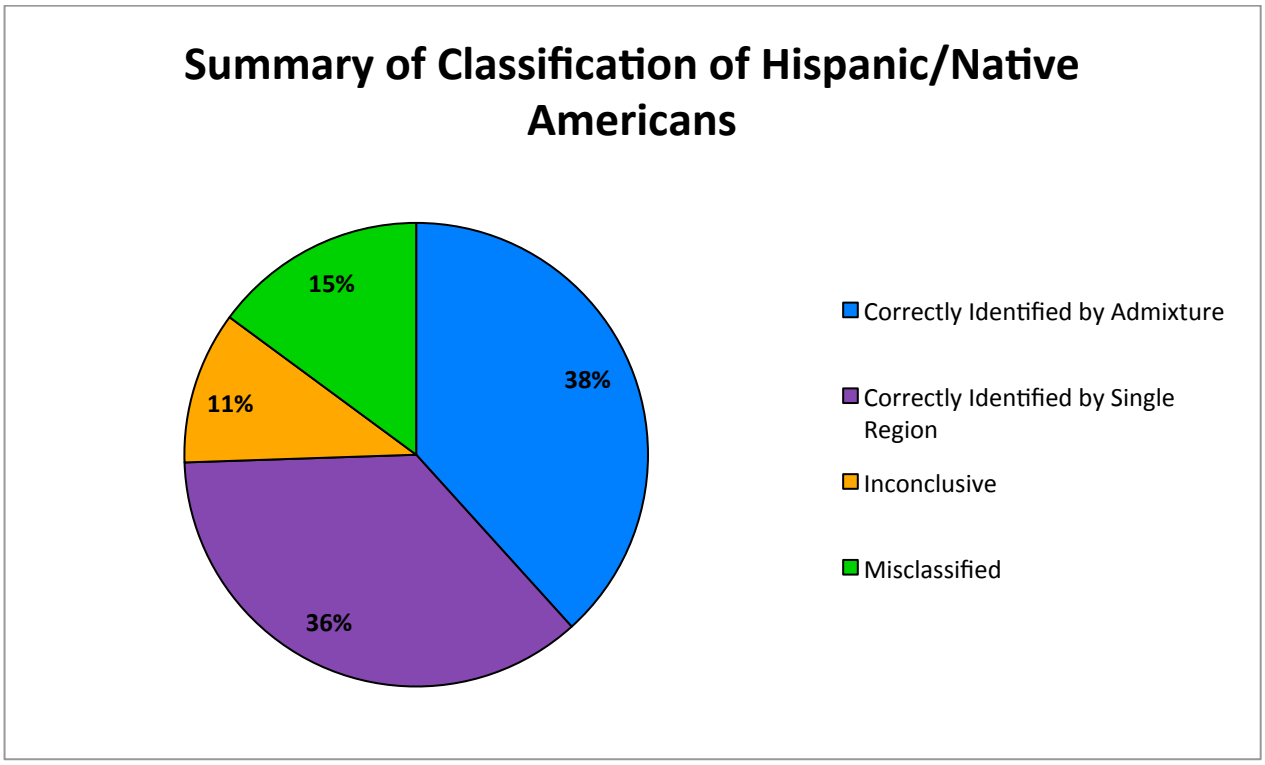


Figure 37. Summary of FROG-kb prediction model classification of Hispanic/Native American samples. 74% of Hispanic/Native Americans were correctly identified with 38% correctly identified as an admixture of two or three biogeographic regions and 36% identified as a single region not corresponding to a major U.S. population (Asia, North America, Oceania, Siberia, or South America), 15% were misclassified, and 11% were classified as inconclusive.

Figure 38 displays the overall summary of FROG-kb performance across all samples. 94% were correctly identified, 3.5% were misclassified, and 2.5% were classified as inconclusive.

At this time, results have not yet been collected for Sample sets III and IV and analysis will be part of future research.

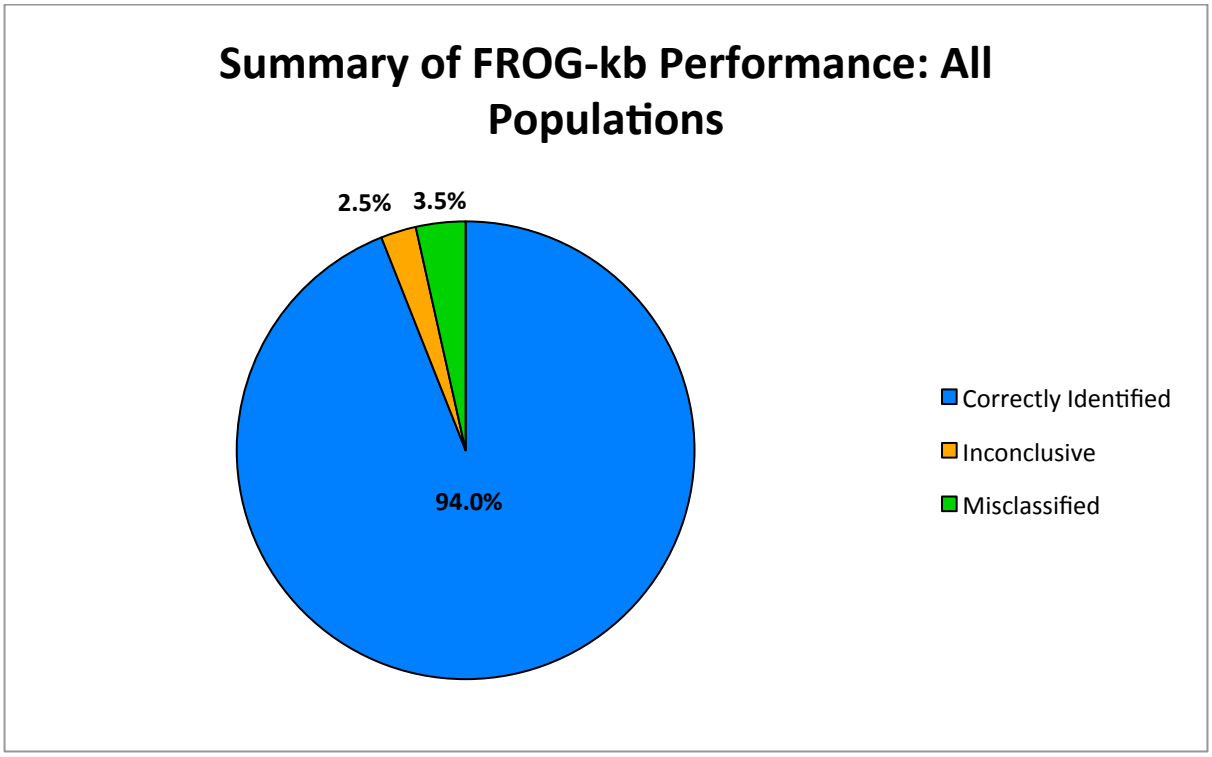


Figure 38. Summary of FROG-kb prediction model performance. 94% were assigned to the correct major U.S. population, 3.5% were misclassified, and 2.5% were classified as inconclusive.

FROG-kb classifies populations into eight biogeographic regions (Africa, Asia, East Asia, Europe, North America, Oceania, Siberia, or South America). Due to the conventionally accepted classification method of four major U.S. population ancestries utilized in forensic investigations (African American, East Asian, and Hispanic/Native American), the FROG-kb prediction model was adapted as shown in Table 12 to reflect these major populations. African American, East Asian, and European samples were considered correctly identified if the FROG-kb prediction classified the sample as African, East Asian, or European, respectively. Since the major category of Hispanic/Native American is an admixed population, those samples were considered correctly identified if the FROG-kb prediction classified the sample as an admixture between two or three biogeographic regions or if it was assigned to a single biogeographic region that was not Africa, East Asia, or Europe (Asia, North America, Oceania, Siberia, or South America).

For the prediction model developed, it was originally intended that a sample would be classified as Hispanic/Native American only if an admixture between two or three biogeographic regions was determined. The admixture determination was limited to only two or three regions to prevent blanket admixture of all eight regions since likelihood calculations were conducted for all populations and regions and there were not uniform numbers of populations and sample sizes per region. The classification of Hispanic/Native American was adjusted to include single-region identifications of Asia, North America, Oceania, Siberia, and South America since predictions of those regions would not have been misclassification of the sample as one of the other major U.S. populations (African American, East Asian, or European).

Of the 200 sample profiles tested, 94% were classified to the correct major U.S. population. 3.5% (n=7) were misclassified, and 2.5% (n=5) were classified as inconclusive. These 12 were Hispanic/Native American sample profiles. All sample profiles classified as inconclusive would have been correctly identified as an admixture of at least three regions if the criteria threshold for three regions was reduced to 75% or there was accommodation for ties of greater than three regions. However, adjustments to the evaluation criteria that minimized misclassifications can also result in increasing the rate of inconclusive classifications.

Of those misclassified, 2 were incorrectly identified as East Asian and 5 were incorrectly identified as European. Although those 7 sample profiles were incorrectly classified, it is important to consider that these profiles are based on individuals self-reporting their ancestry based on the four major U.S. populations. It is possible that although an individual self-identifies as Hispanic, their genetic biogeographic ancestry may actually originate primarily from European populations. While it could be argued that this is a limitation of any AISNP panel and the limitations of classifications using the four major U.S. populations, it is important to understand the error rates and limitations of any tool. The advantage of the FROG-kb output table is that it provides a ranking and breakdown of the probabilities of the genotype in all 89 populations evaluated, thus allowing investigators to maximize the amount of information that may be obtained from the profile with or without an identification based on the major U.S. populations.

Compared to the predictive performance of the AISNP panel developed by Gettings et al without a similar tool, FROG-kb has demonstrated to be a more precise and accurate tool. Gettings et al. reported 98.6% accuracy based on a 77% rate of correctly predicting a single U.S. population, and a 21.6% rate of classification of inconclusive between two major U.S. populations, where the correct population was one of the two indicated. The subsequently reported rate of misclassification was 1.4% based on a 0.7% rate of incorrectly predicting a single population, and a 0.7% rate of classification of inconclusive between two populations, where neither were the correct population. It is important to note that the prediction model developed by Gettings et al. was limited to the allele and subsequent genotype frequencies for the major U.S. populations, rather than more thorough frequency data of 89 populations offered by FROG-kb as well as unable to make predictions of population admixture for Hispanic/Native American individuals. Evaluation using FROG-kb offers a more precise accuracy with a rate of 94% being identified to the correct major U.S. population. FROG-kb also allows for proper classification of admixture compared to the determination of inconclusive between two populations. Being more conservative, the Gettings et al. method has a lower misclassification rate, and it also uses more AISNPs. These additional AISNPs can be included in FROG-kb analysis if data becomes available for all 89 populations.

Future research initiatives would need to increase the number of samples tested, particularly admixed population or Hispanic/Native American samples. Additional testing of samples of known mixed ancestral origin, such as biracial individuals, would also elucidate additional valuable information on the advantages and limitations of FROG-kb.

Although STR analysis is most commonly employed due to their higher power of identification, in forensic investigations when a DNA profile derived from the evidence does not match identified suspects or profiles from available databases, additional DNA analyses, such as those targeted at inferring the possible ancestral origin of the perpetrator, could yield valuable information. Ancestry Informative SNPs (AISNPs) have alleles associated

with specific populations which can be helpful in forensic investigations when STR profiles fail to yield an identification. This study has demonstrated that the Forensic Resource/Reference on Genetics Knowledge Base (FROG-kb) is a convenient, precise, and accurate tool for the prediction of an individual's biogeographic ancestry and can be a value tool in forensic investigations. FROG-kb has been demonstrated to increase the efficacy of the AISNP panel developed by Gettings et al. Even with inherent limitations resulting from individual self-identification, the extensive evaluation of 89 populations from 8 geographic regions offers a greater wealth of information to investigators, information that is unbiased and more reliable compared to other investigative sources of ancestry inference such as eyewitness testimony.

7.4 Preliminary Evaluation of Biogeographic Ancestry Prediction Using a Newly Developed NGS Assay

ThermoFisher Life Technologies has recently released a beta test version of an AISNP panel for the Ion Torrent PGM™, an instrument designed for next generation sequencing (NGS). This prototype version of the HID-Ion AmpliSeq™ Ancestry Panel (to be released soon) contains 40 of the final 55 *Kidd* SNPs (data unpublished, SNP list available at http://frog.med.yale.edu/FrogKB) and all of the 128 *Seldin* SNPs (Kosoy 2009); seven SNPs that overlap between the two panels are included, along with nine additional SNPs for a total of 170 AISNPs. This chemistry, as well as the PGM instrument, is available at NIST. A set of samples from the GWU sample collection was brought to NIST for analysis. One nanogram of input DNA was used for each sample.

Sequencing libraries for the PGM were prepared according to the *Ion AmpliSeq™ Library Preparation User Guide* (Publication Number MAN0006735, Revision 5.0) and the *Ion PGM™ Template OneTouch™ 2 200 Kit User Guide* (Publication Number MAN0007220, Revision 5.0), incorporating the additional recommendations found in the *Procedure Guidelines for using the Ion AmpliSeq™ Library Kit 2.0 User Guide with HID-SNP Identity Panel v2.3* (with the exception of using the Ancestry Panel v3.0 amplification primers). Thirty-two DNA samples were included in one barcoded library pool (17 from GWU, 15 from NIST), which was loaded on an Ion 318™ chip. The PGM instrument was prepared and run according to the *Ion PGM™ Sequencing 200 Kit v2 Quick Reference* (Publication Number MAN0007360, Revision 1.0), with the exception that the chip was loaded using the *Ion PGM™ Chip Loading with the Ion PGM™ Weighted Chip Bucket User Bulletin* (Publication Number MAN0007517, Revision 1.0).

Data was analyzed on the Torrent Server using the following plugins: Alignment (v4.0-r77189) to the hg19 human reference genome from UCSC and HID SNP Genotyper Plugin (beta v4.3). Figure 38 below shows coverage per SNP for two samples (S057=blue, S069=red). Four SNPs show consistently low results and are frequently not callable; however, this assay has not yet been released and is still being optimized by the manufacturer. Considering the maximum throughput of SNaPshot technology is approximately 30 SNPs for a single sample, the ability to genotype over 160 SNPs for over 30 samples at a time is a significant improvement. In addition, more barcodes are available than were used in this experiment and an average coverage of >600X speaks to the potential of including more samples in one sequencing run.

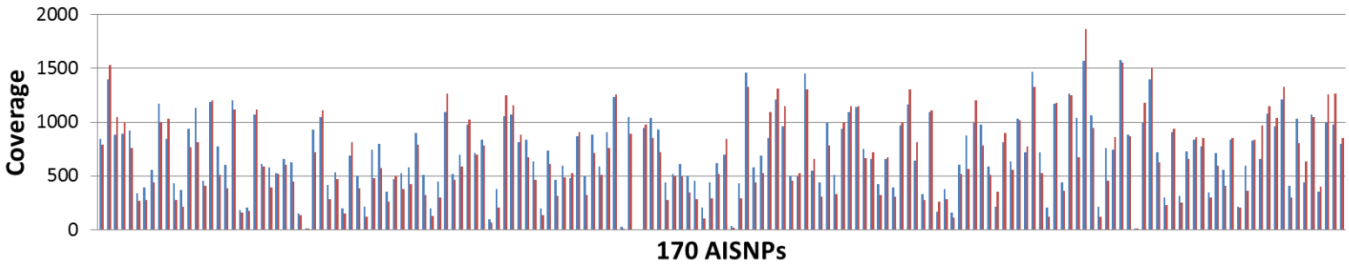


Figure 38. SNP sequence coverage for two samples (S057=blue, S069=red).

The genotyper v4.3 (beta) results output includes calculations of random match probability for the sample in each population where complete data exists in Frog-KB. An example of this output for a West African individual is (S277) is seen figure 39. By dividing one population RMP by another, likelihood ratios can be developed to test the likelihood of the profile if the individual came from one population versus another (similar methods have been illustrated elsewhere in this project).

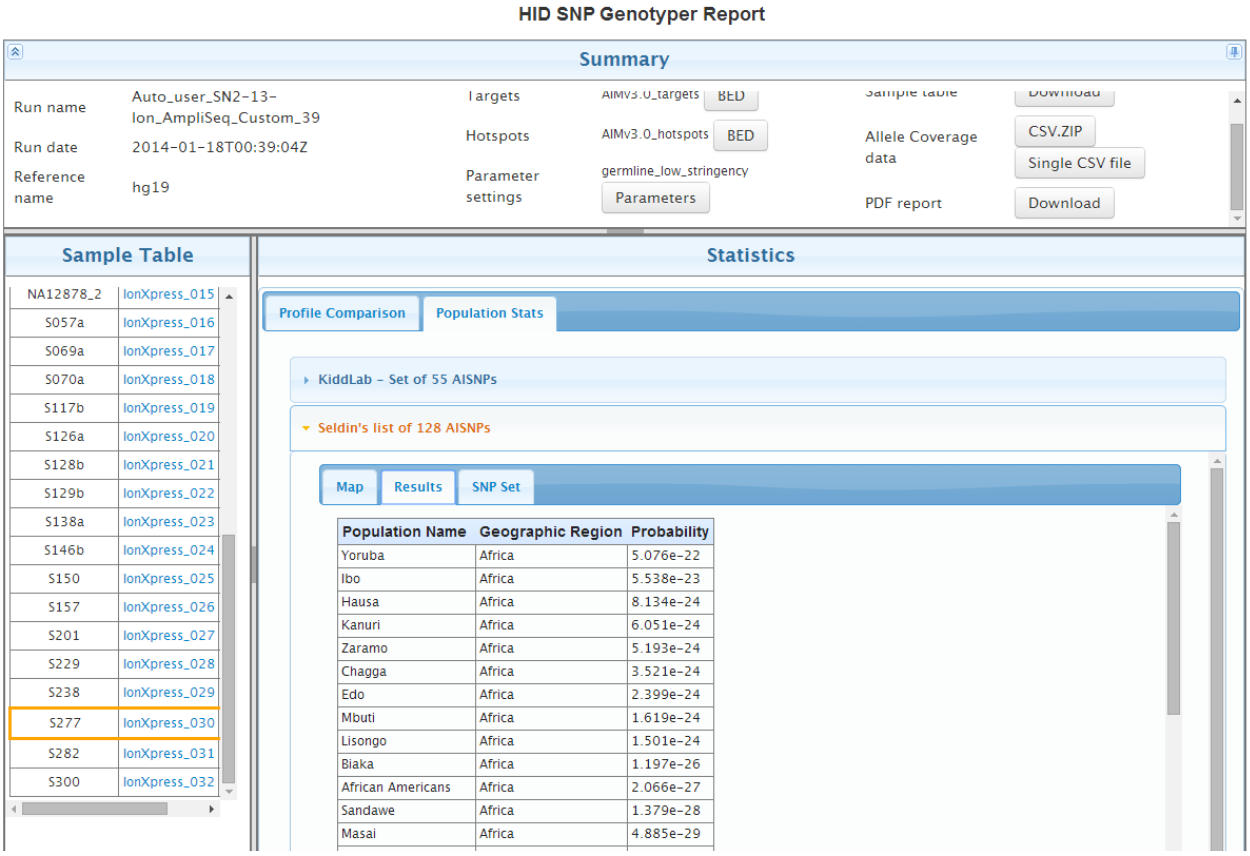


Figure 39. Screenshot of Genotyper v4.3 (beta) results output

In addition, the genotyper v4.3 (beta) also produces heat maps of the world that give "at a glance" information on where the sample most likely originated. For the West African individual, the heat maps using each of the two panels (40 SNPs of Kidd—55 and all of Seldin—128) can be seen figure 40, compared to heat maps generated from an East African individual (S150). It can be seen that, in addition to both individuals grouping within Africa,

one or both panels may be able to offer sub continental information as well. The 17 GWU samples grouped as expected, except for Hispanic individuals who tended to group most closely with western Asian populations. This is not surprising giving the admixed nature of the Hispanic population and the discrete nature of the populations in the ALFRED database. Using a model limited to samples more representative of the US population (as described elsewhere in this project) may improve this prediction; however, the sub continental information shown in the figure below would be lost.

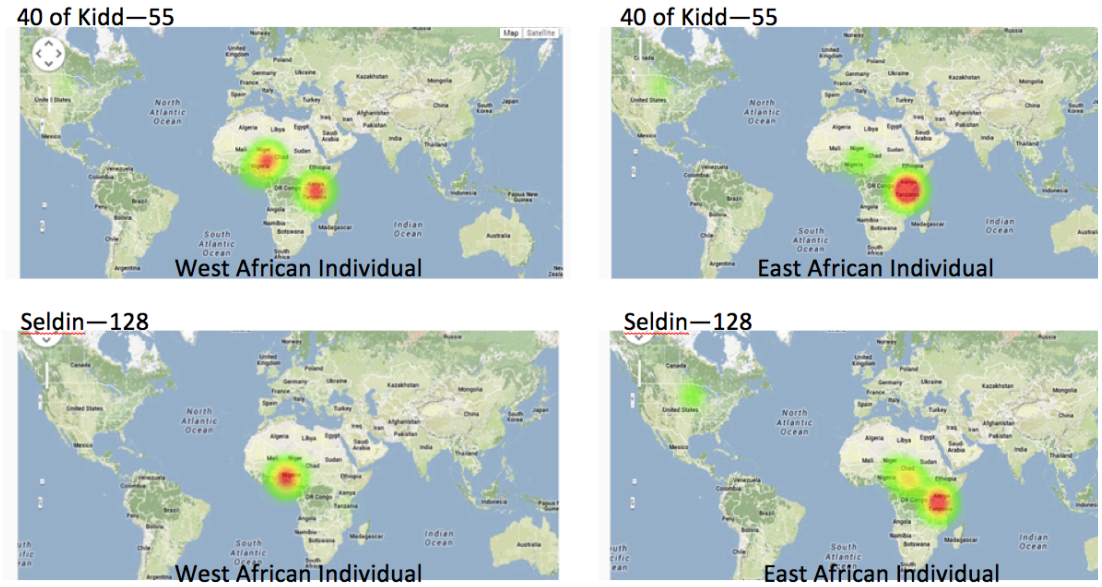


Figure 40. Examples of heat maps generated based on the predicted biogeographic ancestry of the tested sample.

Given the limited number of individuals tested, and the fact that some of them were purposely chosen because they were expected to be challenging, these results cannot be considered representative of the predictive power of the panel. Furthermore a comprehensive study, similar to the one described in the previous chapter, should be performed to define guidelines for interpretation of the output and the limitations of the assay. Nonetheless, given the number of markers tested in a single assay, these results support the idea that NGS platforms have the potential to significantly impact the forensic field in the near future as they have in other molecular biology-based disciplines.

7.5 Whole Genome Amplification as a potential mean for sample 'immortalization'

One of the most important outcomes of this project is the availability of a large number of US population DNA samples with detailed ancestry and phenotypic information of the donor of the sample. There is significant interest from fellow scientists in the field (personal communication) in testing these samples with multiple methods and SNP panels. Furthermore, as technology rapidly progresses, it is likely that in a few years new and comprehensive methods, even with greater capabilities than current NGS, will be developed. Additionally, as the understanding of the mechanics of expression of the human genome increase, it is reasonable to foresee an improved knowledge of how certain traits are obtained. Thus in the future the availability of these samples might be even more significant than today for the development of sound interpretational guidelines for the correct prediction of an individual's ancestry, pigmentation, hair type, height, etc. from evidence collected at a crime scene. Given that each volunteer donated three buccal swabs the amount of DNA available will soon become limited if several requests were to be satisfied.

A possible method to 'immortalize' these samples is Whole Genome Amplification (WGA). There are several WGA methods that have been developed (Park et al. 2005), some are PCR based like DOP-PCR, others, more suitable for this project, are strand displacement amplification based and use the ϕ29 enzyme. In this project we evaluated the strand displacement-based Repli-g Mini kit commercialized by Qiagen.

One important parameter to evaluate is the initial input DNA able to yield a balanced representation of the genome post WGA reaction. In fact one of the possible issues is that some regions amplify more efficiently than others resulting in more copies of one region versus another. If both these regions are targeted in a multiplex PCR assay performed after the WGA, one could end up being over represented, possibly even inhibiting the amplification of the other.

Following the manufacturer's recommendations multiple WGA amplification reactions were performed using as templates multiple dilutions of a single sample and different amounts of another six samples. Figure 41 shows a yield gel of the reaction. From this image there appears to be little difference from inputs down to 2.5 ng but below this amount yields decrease. Samples were quantified with QPCR with an in-house assay targeting a single copy autosomal marker located in the CSF region; the assay also contains an internal positive control (IPC) for monitoring inhibition. Results of the QPCR amplification are shown in table 13.

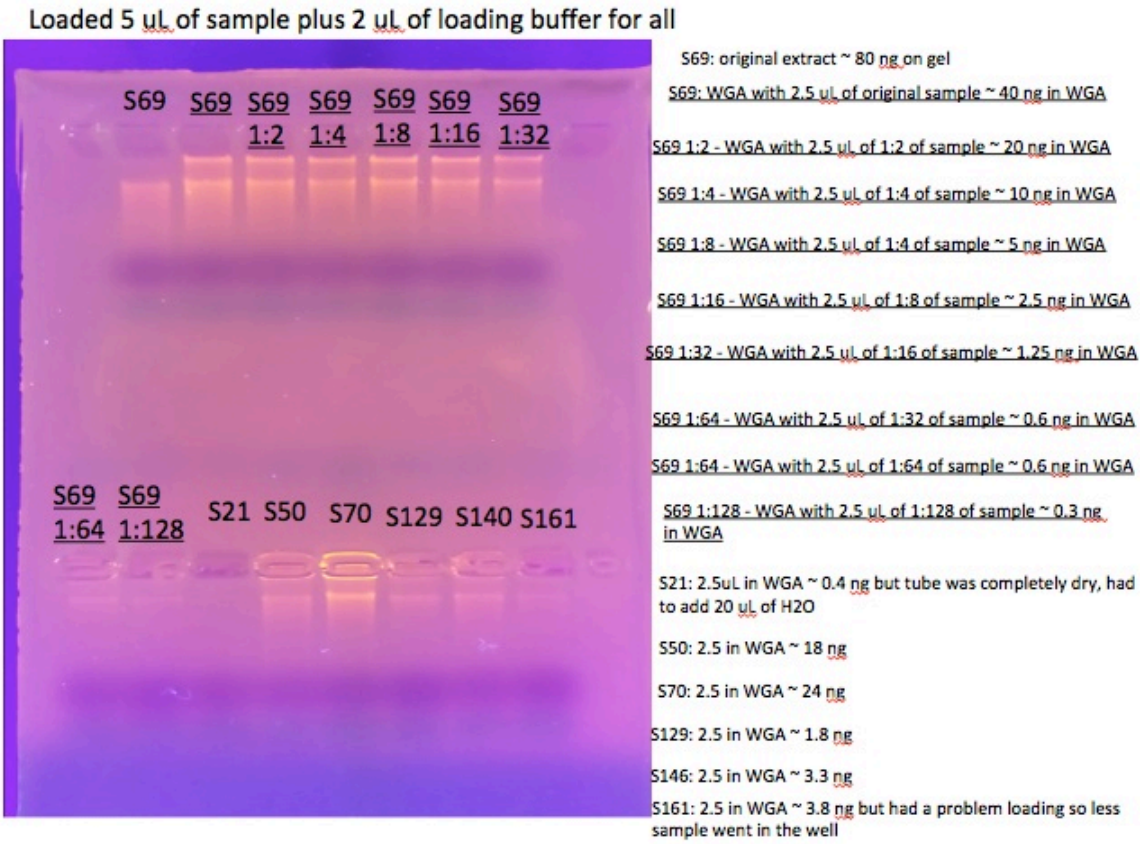


Figure 41. Image of a 1% agarose gel stained with EtBr. Sample S69, upper left, is the original extract in the other wells WGA product of the corresponding samples were loaded. The overall amount of DNA loaded in each WGA reaction is indicated in the legend to the right of the gel. The tube containing sample S21 was dry (possibly evaporated), 20 µL of water was added, the sample was then voretexed, spun down and 2.5 µL was loaded in the WGA reaction, the reaction did not yield amplification product. All wells were loaded with 5 µL of either original genomic DNA (only S69) or WGA product (all other wells) plus 2 µL of loading buffer. The yield of the reaction can be evaluated based on the intensity of the bands.

Sample	Loaded in WGA reaction ng	WGA yield ng/µL
S69-orig-extr		12.416
S69	40	7.15
S69-1:2	20	8.95
S69-1:4	10	8.85
S69-1:8	5	17
S69-1:16	2.5	15.65
S69-1:32	1.25	5.9
S69-1:64	0.6	5.321
S69-1:128	0.3	1.267
S21	?	ND
S50	18	6.4
S70	24	6.7
S129	1.8	5.94
S146	3.3	6.675
S161	3.8	2.2

Table 13. QPRC quantification results showing WGA yield based on input DNA.

A subset of samples was then amplified with all three SBE multiplexes and with AmpFLSTR® Identifiler® Plus. The protocol followed for the SBE amplification is the modified protocol described above using the new enzyme Kapa2G Fast, whereas AmpFLSTR® Identifiler® Plus amplification was performed with a modified 5 µL protocol developed for the purpose of saving reagents. Using this protocol, even with genomic samples, final adenilation issues with certain loci are enhanced (i.e. increased –A peaks), together with greater stochastic heterozygous peak height imbalance. PCR amplification for the three SNP assays was performed with 1 ng of template DNA whereas 300 pgrams were used for AmpFLSTR® Identifiler® Plus amplifications. In Figures 42 and 43 SNP assay and AmpFLSTR® Identifiler® Plus electropherograms respectively, obtained from samples that were whole genome amplified (WGAed), are shown. SNP allele dropouts appear only when the input of template DNA in the WGA is down to 0.3 ng. STR allele imbalance appears when the WGA input DNA is 2.5 ng or lower and full drop out becomes evident only when the WGA input is down to 0.3 ng. Results are consistent with what reported by the manufacturer, which recommends input amounts above 10 ng, although successful results can be obtained down to 1 ng with high quality template DNA.

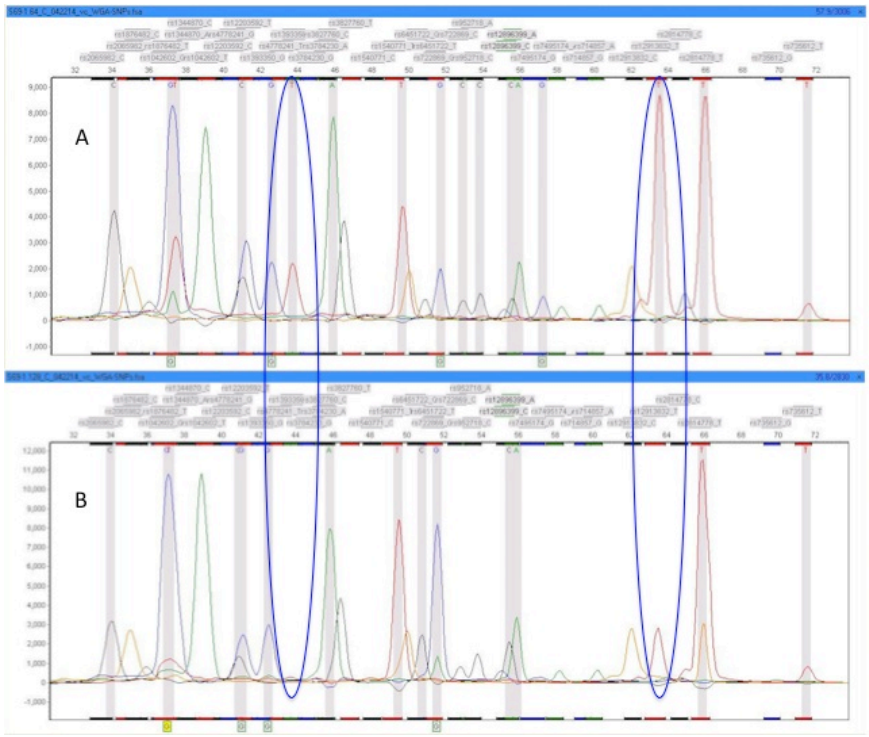


Figure 42. Electropherograms of WGA samples typed with multiplex C. **A)** Full profile from 1 ng of WGA product obtained from 0.6 ng of input DNA; **B)** Partial profile from 1 ng of DNA from WGA product obtained from 0.3 ng of input DNA.

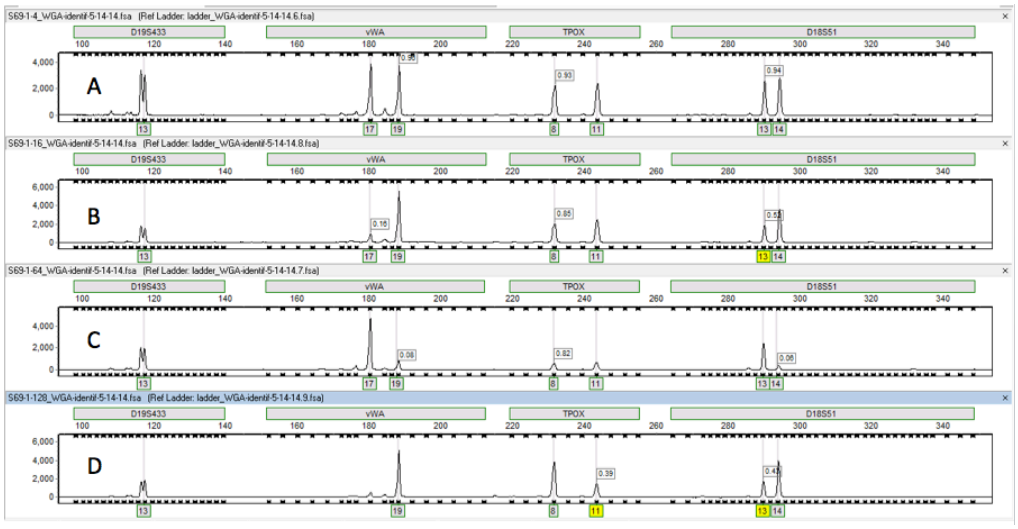


Figure 43. Elecropherograms of AmpFLSTR® Identifiler® Plus amplifications (yellow channel only) with 300 pg of WGA product obtained from 10 ng (A), 2.5 ng (B), 0.6 ng (C), and 0.3 ng (D) of input DNA all of the same sample. Allele preferential amplification appears when the amount of DNA used in WGA is down to 2.5ng while full drop out appeared only when the WGA input was down to 0.3 ng.

Although more tests are needed the results described in this section support the theory that the samples collected during this project can be 'immortalized'. Further funding will be sought to:

1) Evaluate other WGA kits. For example single-cell WGA kits are commercially available and could be more sensitive, reducing even further the amount of starting DNA necessary;
2) Perform WGA from WGA product to evaluate whether the intra and inter-locus balance is maintained over multiple re-amplification cycles;
3) Test WGA product on NGS platforms to determine compatibility and genotype consistency.

If the preliminary results described herein are confirmed the sample database could be immortalized and represent an important resource of US DNA samples with known ancestry and phenotype. Upon request it could be shared with other scientists in the US and abroad conducting research in the field.

7.6 References for Chapter 7

Bouakaze C, Keyser C, Crubezy E, Montagnon D, Ludes B. Pigment phenotype and biogeographical ancestry from ancient skeletal remains: inferences from multiplexed autosomal SNP analysis. Int. J. Legal Med. 2009; 123: 315-325.

Evett IW, Pinchin R, Buffery C. An investigation of the feasibility of inferring ethnic origin from DNA profiles. J. Foren. Sci. Soc. 1992; 4: 301-306.

Gettings KB. Forensic ancestry and phenotype SNP analysis and integration with established forensic markers [Doctoral dissertation]. Washington (DC): The George Washington University, 2013.

Gettings KB, Lai R, Johnson JL, Peck MA, Hart JA, Gordish-Dressman H, et al. A 50-SNP assay for biogeographic ancestry and phenotype prediction in the U.S. population. Foren. Sci. Int. Genet. 2014; 8: 101-108.

Halder I, Shriver M, Thomas M, Fernandez JR, Frudakis T. A panel of ancestry informative markers for estimating individual biogeographical ancestry and admixture from four continents: utility and applications. Hum. Mutat. 2008; 29: 648-658.

Kidd JR, Friedlaender FR, Speed WC, Pakstis AJ, Vega DL, Kidd KK. Analyses of a set of 128 ancestry informative single-nucleotide polymorphism in a global set of 119 population samples. Investig. Genet. 2011; 2: 1-13.

Kosoy R, Nassir R, Tian C, White PA, Butler LM, Silva G, et al. Ancestry informative marker sets for determining continental origin and admixture proportions in common populations in America. Hum. Mutat. 2009; 30: 69-78.

Lowe AL, Urquhart A, Foreman LA, Evett IW. Inferring ethnic origin by means of an STR profile. Foren. Sci. Int. 2001; 119: 17-22.

Nassir R, Kosoy R, Tian C, White PA, Butler LM, Silva G, et al. An ancestry informative marker set for determining continental origin: validation and extension using human genome diversity panels. BMC Genet. 2009; 10: 39-52.

Park JW, Beaty T, Boyce P, Scott A, McIntosh I, Comparing Whole Genome Amplification Methods and Sources of Biological Samples for Single Nucleotide Polymorphism. Clinical Chemistry. 2005; 51, 8: 1520-1523.

Phillips C, Salas A, Sanchez JJ, Fondevila M, Gomez-Tato A, Alvarez-Dios J, et al. Inferring ancestral origin using a single multiplex assay of ancestry-informative marker SNPs. Foren. Sci. Int. Genet. 2007; 1: 273-280.

Rajeevan H, Soundararajan U, Kidd JR, Pakstis AJ, Kidd KK. ALFRED: an allele frequency resource for research and teaching. Nucl. Acids Res. 2011; 1-6.

8 Appendix

Appendix Table 1a. Results of ancestry analyses for candidate SNP evaluation / reduction for the 50 selected SNPs.

SNP ID	Category	Chr	Gene/Region	X^2 (p-value) ancestry	X^2 rank ancestry	PCA Ancestry High Factor Loading	Snipper divergence ranking	Pairwise F_{ST}				
								AF-EU	AF-AS	AS-EU	AS-NA	NA-EU
rs10007810	AIM	4	*LIMGH1* intron	<0.0001	4	X	22	0.451	0.622	0.026	0.025	0.000
rs10108270	AIM	8	*CSMD1*	<0.0001	6	X	16	0.345	0.310	0.001	0.141	0.119
rs1042602	PIM	20	*TSIP*	<0.0001			62	0.179	0.006	0.210	0.000	0.207
rs10496971	AIM	2		<0.0001	7	X	18	0.002	0.484	0.448	0.088	0.172
rs1126809	PIM	11	*TYR*	<0.0001			77	0.090	0.007	0.115	0.054	0.019
rs11547464	PIM	16	*MC1R*	0.909			97	0.003	N/A	0.003	0.000	0.003
rs12203592	PIM	6	*IRF4*	<0.0001			82	0.069	0.005	0.085	0.012	0.053
rs12821256	PIM	12	*KITLG*	<0.0001			84	0.037	0.002	0.045	0.000	0.042
rs12896399	PIM	14	*SLC24A4*	<0.0001			80	0.183	0.127	0.007	0.003	0.020
rs12913832	PIM	15	*HERC2*	<0.0001			19	0.444	0.009	0.498	0.024	0.395
rs1344870	AIM	3		<0.0001	5	X	65	0.002	0.068	0.088	0.292	0.579
rs1375164	PIM	15	*OCA2* intron	<0.0001			21	0.486	0.004	0.432	0.191	0.065
rs1393350	PIM	11	*TYR*	<0.0001			75	0.107	0.006	0.129	0.006	0.107
rs1426654	PIM	15	*SLC24A5*	<0.0001			1	0.690	0.040	0.886	0.005	0.829
rs1540771	PIM	6	*IRF4*	<0.0001			83	0.165	0.054	0.035	0.028	0.000
rs1667394	PIM	15	*HERC2*	<0.0001			26	0.481	0.035	0.303	0.093	0.072
rs16891982	PIM	5	*SLC45A2*	<0.0001			3	0.370	0.001	0.400	0.012	0.495
rs1800407	PIM	15	*OCA2*	<0.0001			92	0.026	0.000	0.028	0.000	0.024
rs1800414	PIM	15	*OCA2*	<0.0001			14	0.001	0.393	0.388	0.393	0.001
rs1805007	PIM	16	*MC1R*	<0.0001			91	0.027	0.000	0.028	0.000	0.026
rs1805008	PIM	16	*MC1R*	<0.0001			95	0.020	0.002	0.029	0.010	0.008
rs1805009	PIM	16	*MC1R*	0.053			98	0.008	0.002	0.013	0.014	0.000
rs1834640	PIM	15	*SLC24A5*	<0.0001			5	0.678	0.004	0.745	0.086	0.420
rs1876482	AIM	2		<0.0001	11		10	0.036	0.565	0.431	0.075	0.179
rs2065982	AIM	13		<0.0001	11		36	0.000	0.354	0.352	0.021	0.503
rs260690	AIM	2	*EDAR*	<0.0001	1	X	15	0.274	0.151	0.689	0.003	0.639
rs26722	PIM	5	*SLC45A2*	<0.0001			70	0.020	0.138	0.213	0.030	0.105
rs2714758	AIM	15		<0.0001	19		9	0.600	0.604	0.000	0.003	0.002
rs2814778	AIM	1	*DARC*	<0.0001	1	X	4	0.815	0.841	0.005	0.002	0.001
rs3737576	AIM	1		<0.0001	9	X	56	0.025	0.023	0.000	0.352	0.346
rs3784230	AIM	14	*BRF1*	<0.0001	2	X	30	0.319	0.680	0.113	0.148	0.003
rs3827760	PIM	2	*EDAR*	<0.0001			2	0.008	0.713	0.663	0.007	0.748
rs4778138	PIM	15	*OCA2*	<0.0001			17	0.311	0.000	0.306	0.358	0.003
rs4778241	PIM	15	*OCA2*	<0.0001			55	0.126	0.035	0.275	0.059	0.090
rs4891825	AIM	18	*RAAN*	<0.0001	4	X	7	0.504	0.643	0.033	0.055	0.004
rs4911414	PIM	20	*ASIP*	<0.0001			90	0.054	0.008	0.022	0.048	0.005
rs4911442	PIM	20	*NC0A6*	<0.0001			88	0.033	0.002	0.041	0.001	0.035
rs4918842	AIM	10	*HABP2*	<0.0001	7	X	34	0.014	0.161	0.089	0.201	0.491
rs6451722	AIM	5		<0.0001	7	X	39	0.356	0.311	0.002	0.114	0.089
rs6548616	AIM	3	*ROBO1*	<0.0001	4	X	37	0.361	0.479	0.013	0.039	0.091
rs714857	AIM	11		<0.0001	13		32	0.432	0.015	0.312	0.100	0.075
rs7170852	PIM	15	*HERC2*	<0.0001			45	0.365	0.005	0.295	0.064	0.098
rs722869	AIM	14	*VRK1*	<0.0001	12		11	0.004	0.519	0.463	0.034	0.278
rs730570	AIM	14		<0.0001	9	X	43	0.287	0.006	0.362	0.048	0.579
rs735612	AIM	15	*RYR3*	<0.0001	14	X	51	0.032	0.291	0.151	0.517	0.155
rs7495174	PIM	15	*OCA2*	<0.0001			44	0.018	0.198	0.307	0.278	0.001
rs885479	PIM	16	*MC1R*	<0.0001			12	0.036	0.421	0.290	0.001	0.314
rs896788	PIM	2	*RNF144A*	<0.0001			67	0.022	0.074	0.167	0.002	0.134
rs916977	PIM	15	*HERC2*	<0.0001			24	0.464	0.025	0.312	0.095	0.075
rs952718	AIM	2	*ABCA12*	<0.0001	14		38	0.349	0.421	0.009	0.237	0.176

NOTE: For columns "X^2 with ethnicity", "X^2 rank for ethnicity", "Snipper divergence ranking" and "Pairwise F_{ST}", results are based on the four populations of primary interest in the U.S.: European (EU), East Asian (EA), African/African American (AA) and Native American (NA)

NOTE: For pairwise F_{ST}, X^2 testing shows values in gray are not significant at α=0.001.

Appendix Table 1b. Results of pigmentation analyses for candidate SNP evaluation / reduction for the 50 selected SNPs.

SNP ID	Category	Chr	Gene/Region	X^2 (p-value) Europeans			PCA European Pigmentation High Factor Loading (E-Eye, S-Skin, or H-Hair)
				eye	skin	hair	
rs10007810	AIM	4	*LIMGH1* intron	0.073	0.910	0.927	
rs10108270	AIM	8	*CSMD1*	0.875	0.384	0.049	
rs1042602	PIM	20	*TSIP*	0.558	0.071	0.030	E,S
rs10496971	AIM	2		0.422	0.937	0.588	
rs1126809	PIM	11	*TYR*	0.027	0.283	0.092	E,S
rs11547464	PIM	16	*MC1R*	0.618	0.205	0.768	E,H
rs12203592	PIM	6	*IRF4*	0.441	7.49E-05	0.001	E
rs12821256	PIM	12	*KITLG*	0.255	0.801	0.190	E,S
rs12896399	PIM	14	*SLC24A4*	0.607	0.167	0.219	E,H
rs12913832	PIM	15	*HERC2*	2.43E-15	0.002	0.016	H,S
rs1344870	AIM	3		0.456	0.130	0.660	
rs1375164	PIM	15	*OCA2* intron	0.002	0.492	0.274	E,H,S
rs1393350	PIM	11	*TYR*	0.018	0.891	0.200	E,S
rs1426654	PIM	15	*SLC24A5*	N/A	N/A	N/A	E,H,S
rs1540771	PIM	6	*IRF4*	0.054	0.295	0.001	
rs1667394	PIM	15	*HERC2*	1.15E-12	0.300	0.521	H,S
rs16891982	PIM	5	*SLC45A2*	1.44E-06	0.069	0.002	E,H,S
rs1800407	PIM	15	*OCA2*	0.051	0.917	0.846	E
rs1800414	PIM	15	*OCA2*	N/A	N/A	N/A	E,H
rs1805007	PIM	16	*MC1R*	0.053	0.214	1.68E-06	S
rs1805008	PIM	16	*MC1R*	0.167	0.072	4.61E-06	
rs1805009	PIM	16	*MC1R*	0.119	0.062	0.003	E,H
rs1834640	PIM	15	*SLC24A5*	0.266	0.713	0.290	E,H,S
rs1876482	AIM	2		0.051	0.861	0.517	
rs2065982	AIM	13		0.515	0.052	0.274	
rs260690	AIM	2	*EDAR*	0.658	0.790	0.814	
rs26722	PIM	5	*SLC45A2*	0.018	0.115	2.39E-05	E,H
rs2714758	AIM	15		0.924	0.137	0.622	
rs2814778	AIM	1	*DARC*	0.709	0.935	0.350	
rs3737576	AIM	1		0.772	0.868	0.745	
rs3784230	AIM	14	*BRF1*	0.092	0.608	0.644	
rs3827760	PIM	2	*EDAR*	0.266	0.713	0.290	
rs4778138	PIM	15	*OCA2*	2.24E-05	0.501	0.639	E,H,S
rs4778241	PIM	15	*OCA2*	2.05E-09	0.561	0.232	H,S
rs4891825	AIM	18	*RAAN*	0.734	0.072	0.457	
rs4911414	PIM	20	*ASIP*	0.154	0.144	0.470	E,H,S
rs4911442	PIM	20	*NC0A6*	0.279	0.063	0.205	E,H,S
rs4918842	AIM	10	*HABP2*	0.651	0.364	0.294	
rs6451722	AIM	5		0.694	0.923	0.049	
rs6548616	AIM	3	*ROBO1*	0.371	0.923	0.356	
rs714857	AIM	11		0.108	0.015	0.167	
rs7170852	PIM	15	*HERC2*	1.23E-10	0.155	0.630	E,H,S
rs722869	AIM	14	*VRK1*	0.146	0.638	0.105	
rs730570	AIM	14		0.021	0.070	0.557	
rs735612	AIM	15	*RYR3*	0.059	0.613	0.377	
rs7495174	PIM	15	*OCA2*	1.07E-04	0.021	0.592	
rs885479	PIM	16	*MC1R*	0.440	0.518	0.459	E,H,S
rs896788	PIM	2	*RNF144A*	0.436	0.771	0.007	E,H,S
rs916977	PIM	15	*HERC2*	1.65E-12	0.514	0.498	H,S
rs952718	AIM	2	*ABCA12*	0.558	0.925	0.002	

NOTE: For p-value in Europeans, after Bonferroni correction for multiple testing, values < 0.01 are significant.

Appendix Table 1c. Results of ancestry analyses for candidate SNP evaluation / reduction for the 49 eliminated SNPs.

SNP ID	Category	Chr	Gene/Region	X^2 (p-value) ancestry	X^2 rank ancestry	PCA Ancestry High Factor Loading	Snipper divergence ranking	Pairwise F_{ST}				
								AF-EU	AF-AS	AS-EU	AS-NA	NA-EU
rs1015362	PIM	20	*TSIP*	<0.0001			49	0.170	0.266	0.014	0.052	0.013
rs1041321	AIM	9	*ACO1*	<0.0001	24		66	0.071	0.011	0.135	0.210	0.010
rs10843344	AIM	12		<0.0001	26		69	0.152	0.021	0.079	0.100	0.001
rs10852218	PIM	15	*OCA2*	<0.0001			60	0.161	0.376	0.077	0.109	0.004
rs1110400	PIM	16	*MC1R*	0.231			99	0.005	N/A	0.005	0.000	0.005
rs1129038	PIM	15	*HERC2*	<0.0001			20	0.444	0.008	0.497	0.021	0.398
rs1160312	PIM	20		<0.0001			89	0.061	0.002	0.043	0.005	0.019
rs11636232	PIM	15	*HERC2*	<0.0001			78	0.106	0.012	0.156	0.040	0.056
rs11803731	PIM	1	*TCHH*	<0.0001			74	0.104	0.004	0.122	0.106	0.001
rs13400937	AIM	2	*CTNNA2*	<0.0001	14	X	61	0.295	0.007	0.221	0.203	0.000
rs1363448	AIM	5	*PCDHGA9*	<0.0001	17	X	71	0.148	0.194	0.004	0.179	0.134
rs1408799	PIM	9	*TYRP1*	<0.0001			41	0.137	0.109	0.408	0.016	0.315
rs1448484	PIM	15	*OCA2*	<0.0001			6	0.588	0.600	0.002	0.023	0.015
rs1454284	AIM	8		<0.0001	28		93	0.003	0.006	0.017	0.000	0.014
rs1470144	AIM	11		<0.0001	27		79	0.117	0.132	0.001	0.011	0.007
rs1513181	AIM	3	*LPP*	<0.0001	9	X	42	0.000	0.339	0.338	0.008	0.428
rs1545397	PIM	15	*OCA2*	<0.0001			8	0.001	0.613	0.591	0.144	0.206
rs1724630	PIM	15	*MYO5A*	<0.0001			63	0.007	0.047	0.088	0.015	0.031
rs1800401	PIM	15	*OCA2*	0.003			87	0.014	0.052	0.018	0.035	0.004
rs1800410	PIM	15	*OCA2*	<0.0001			25	0.035	0.332	0.523	0.178	0.126
rs1805005	PIM	16	*MC1R*	<0.0001			81	0.058	0.000	0.056	0.011	0.026
rs1805006	PIM	16	*MC1R*	0.789			96	0.001	N/A	0.001	N/A	0.001
rs1823718	AIM	15		<0.0001	16	X	64	0.097	0.041	0.231	0.223	0.000
rs1858465	AIM	17		<0.0001	12		53	0.417	0.283	0.020	0.073	0.019
rs2031526	PIM	13	*DCT*	<0.0001			29	0.031	0.426	0.266	0.000	0.255
rs2065160	AIM	1		<0.0001	14		40	0.110	0.114	0.400	0.008	0.491
rs2228478	PIM	16	*MC1R*	<0.0001			57	0.156	0.042	0.042	0.113	0.024
rs2228479	PIM	16	*MC1R*	<0.0001			76	0.026	0.103	0.037	0.120	0.041
rs2238289	PIM	15	*HERC2*	<0.0001			72	0.313	0.001	0.278	0.069	0.083
rs2304925	AIM	17		<0.0001	23		68	0.147	0.000	0.143	0.144	0.000
rs2352476	AIM	7		<0.0001	18		86	0.080	0.007	0.041	0.064	0.195
rs236336	AIM	1	*BCAR3*	<0.0001	20		13	0.446	0.518	0.005	0.018	0.040
rs2416791	AIM	12		<0.0001	5	X	33	0.601	0.393	0.040	0.071	0.200
rs2424984	PIM	20	*ASIP*	<0.0001			46	0.341	0.212	0.022	0.110	0.044
rs2733832	PIM	9	*TYRP1*	<0.0001			50	0.218	0.018	0.308	0.113	0.069
rs2946788	AIM	11		<0.0001	14	X	52	0.270	0.047	0.103	0.199	0.019
rs35264875	PIM	11	*TPCN2*	<0.0001			85	0.052	0.001	0.062	0.010	0.029
rs434504	AIM	1	*AJAP1*	<0.0001	22		31	0.000	0.435	0.427	0.257	0.030
rs4752566	PIM	10	*FGFR2*	<0.0001			27	0.121	0.538	0.193	0.169	0.001
rs4908343	AIM	1	*AHDC1*	<0.0001	8	X	73	0.464	0.246	0.049	0.078	0.004
rs559035	AIM	6	*CDC5L*	<0.0001	21		54	0.068	0.081	0.276	0.255	0.001
rs642742	PIM	12	*KITLG*	<0.0001			35	0.391	0.389	0.000	0.007	0.007
rs6950524 (me	PIM	7		0.425			94	0.007	0.000	0.004	0.007	0.001
rs697212	AIM	12	*STAB2*	<0.0001	15	X	59	0.165	0.344	0.042	0.267	0.108
rs741272	AIM	14	*FOXN3*	<0.0001	25		58	0.245	0.035	0.114	0.115	0.000
rs749846	PIM	15	*OCA2*	<0.0001			28	0.009	0.412	0.511	0.209	0.094
rs772262	AIM	12	*SARNP*	<0.0001	10	X	47	0.420	0.303	0.017	0.205	0.310
rs9522149	AIM	13	*ARHGEF7*	<0.0001	6	X	23	0.347	0.030	0.485	0.003	0.450
rs9530435	AIM	13	*TBC1D4*	<0.0001	3	X	48	0.398	0.604	0.053	0.001	0.040

NOTE: For columns "X^2 with ethnicity", "X^2 rank for ethnicity", "Snipper divergence ranking" and "Pairwise F_{ST}", results are based on the four populations of primary interest in the U.S.: European (EU), East Asian (EA), African/African American (AA) and Native American (NA)
NOTE: For pairwise F_{ST}, X^2 testing shows values in gray are not significant at α=0.001.

Four markers were eliminated prior to analysis:
rs6152 and rs6625163 are SNPs associated with baldness and the sample size was insufficient to assess correlation
rs3829241 and rs6119471 were eliminated due to genotyping issues / incompatiblily with SBE system

Appendix Table 1d. Results of pigmentation analyses for candidate SNP evaluation / reduction for the 49 eliminated SNPs.

SNP ID	Category	Chr	Gene/Region	X^2 (p-value) Europeans			PCA Europeans High Factor Loading (E-Eye, S-Skin, or H-Hair)
				eye	skin	hair	
rs1015362	PIM	20	*TSIP*	0.596	0.691	0.885	H
rs1041321	AIM	9	*ACO1*	0.677	0.820	0.561	
rs10843344	AIM	12		0.738	0.304	0.351	
rs10852218	PIM	15	*OCA2*	0.004	0.269	0.266	
rs1110400	PIM	16	*MC1R*	0.384	0.725	0.713	E,H,S
rs1129038	PIM	15	*HERC2*	0.027	0.283	0.092	H,S
rs1160312	PIM	20		0.074	0.703	0.932	
rs11636232	PIM	15	*HERC2*	N/A	N/A	N/A	E,S
rs11803731	PIM	1	*TCHH*	0.765	0.778	0.662	
rs13400937	AIM	2	*CTNNA2*	0.706	0.115	0.932	
rs1363448	AIM	5	*PCDHGA9*	0.905	0.218	0.473	
rs1408799	PIM	9	*TYRP1*	0.676	0.571	0.294	E,H
rs1448484	PIM	15	*OCA2*	0.794	0.659	0.409	H
rs1454284	AIM	8		0.637	0.038	0.473	
rs1470144	AIM	11		0.839	0.461	0.966	
rs1513181	AIM	3	*LPP*	0.270	0.420	0.558	
rs1545397	PIM	15	*OCA2*	0.645	0.841	0.147	E,S
rs1724630	PIM	15	*MYO5A*	0.175	0.374	0.875	H,S
rs1800401	PIM	15	*OCA2*	0.764	0.213	0.360	H
rs1800410	PIM	15	*OCA2*	0.551	0.861	0.138	E
rs1805005	PIM	16	*MC1R*	0.623	0.759	0.770	E,H,S
rs1805006	PIM	16	*MC1R*	0.257	0.285	0.904	E,H
rs1823718	AIM	15		0.640	0.606	0.323	
rs1858465	AIM	17		0.968	0.397	0.226	
rs2031526	PIM	13	*DCT*	0.298	0.216	0.507	E,H
rs2065160	AIM	1		0.501	0.456	0.906	
rs2228478	PIM	16	*MC1R*	0.840	0.845	0.158	E
rs2228479	PIM	16	*MC1R*	0.887	0.291	0.195	E,S
rs2238289	PIM	15	*HERC2*	1.44E-13	0.115	0.288	H
rs2304925	AIM	17		0.573	0.482	0.001	
rs2352476	AIM	7		0.026	0.039	0.004	
rs236336	AIM	1	*BCAR3*	0.414	0.677	0.571	
rs2416791	AIM	12		0.626	0.367	0.365	
rs2424984	PIM	20	*ASIP*	0.995	0.575	0.779	E,H
rs2733832	PIM	9	*TYRP1*	0.060	0.539	0.343	E,H,S
rs2946788	AIM	11		0.305	0.222	0.812	
rs35264875	PIM	11	*TPCN2*	0.729	0.953	0.076	
rs434504	AIM	1	*AJAP1*	0.242	0.523	0.597	
rs4752566	PIM	10	*FGFR2*	0.234	0.978	0.285	
rs4908343	AIM	1	*AHDC1*	0.353	0.287	0.294	
rs559035	AIM	6	*CDC5L*	0.540	0.228	0.964	
rs642742	PIM	12	*KITLG*	0.586	0.860	0.418	
rs6950524 (me	PIM	7		0.265	0.458	0.534	S
rs697212	AIM	12	*STAB2*	0.346	0.720	0.472	
rs741272	AIM	14	*FOXN3*	0.675	0.316	0.071	
rs749846	PIM	15	*OCA2*	0.168	0.956	0.400	H,S
rs772262	AIM	12	*SARNP*	0.554	0.194	0.176	
rs9522149	AIM	13	*ARHGEF7*	0.849	0.719	0.298	
rs9530435	AIM	13	*TBC1D4*	0.551	0.571	0.035	

NOTE: For p-value in Europeans, after Bonferroni correction for multiple testing, values < 0.01 are significant.

Four markers were eliminated prior to analysis:
rs6152 and rs6625163 are SNPs associated with baldness and the sample size was insufficient to assess correlation
rs3829241 and rs6119471 were eliminated due to genotyping issues / incompatiblily with SBE system

Appendix Table 2a. SNP markers contained in the 50 SNP assay, Multiplex A, with molecular and PCR primer information.

SNP ID	Gene/ Region	Chr	SNP Type P=phenotype A=ancestry	Base Change		PCR Primers	Concentration
Multiplex A							
rs885479	MC1R	16	P*	A / G	F	ATGCTGTCCAGCCTCTGCTT	0.6µm
					R	TAGTAGGCGATGAAGAGCGT	0.6µm
rs1834640	SLC24A5	15	P	A / G	F	CAACCGTTAGAGACCCATACTTG	0.04µm
					R	CCCTATACTTAGCAGCAGACAATCC	0.04µm
rs1805009	MC1R	16	P	C / G	F	CCTCATCATCTGCAATGCCATC	0.16µm
					R	GGTCCGCGCTTCAACACTTTCAGA	0.16µm
rs1805008	MC1R	16	P	C / T	F	CTGCAGCAGCTGGACAAT	0.06µm
					R	ATGAAGAGCGTGCTGAAGACGA	0.06µm
rs1126809	TYR	11	P	A / G	F	TCTTTCCATGTCTCCAGATT	0.3µm
					R	TGAAGAGGACGGTGCC	0.3µm
rs896788	RNF144A	2	P*	A / G	F	TCCTGCAGTGTAGATAAGGCCA	0.03µm
					R	TCACTGAGCATCTACAGTCACCAG	0.03µm
rs260690	EDAR	2	P	A / C	F	GAAACTCTGTGGCCAACGTA	0.16µm
					R	TGAAGGGCTCTTGAAAGCA	0.16µm
rs6548616	ROBO1	3	A*	C / T	F	CCTCACGCATTGCTAGTTGGATTG	0.08µm
					R	AGGAGTGGAATTCTCTTAGCTG	0.08µm
rs1667394	HERC2	15	P	A / G	F	CAGCTGTAGAGAGAGACTTTGAGG	0.24µm
					R	GGTCAATCCACCATTAAGACGCAG	0.24µm
rs26722	SLC45A2	5	P	C / T	F	CATTGCCAGCTCTGGATTTACG	0.16µm
					R	CACTTACAGAGGTTGCAAAGGG	0.16µm
rs10108270	CSMD1	8	A*	A / C	F	CTAGTGACCCTGGACACAATTC	0.5µm
					R	CCCTTTCTGTATCATCTCTCTCGG	0.5µm
rs1800414	OCA2	15	P	A / G	F	GTGCAGAGTAAATGAGCTGTGG	0.2µm
					R	GATCAAGATGAATGCCAGGGAC	0.2µm
rs4911442	NCOA6	20	P*	A / G	F	GGGAAGTACAGTAACTAGCTTGAGG	0.4µm
					R	TGGGCAACAGAGTGAGACT	0.4µm
rs4911414	ASIP	20	P*	G / T	F	TTGTTTGTAAGTCTTTGCTGAG	0.1µm
					R	CCATAGTCATCAGAGTATCCAGGG	0.1µm
rs11547464	MC1R	16	P	A / G	F	included in rs1805008	
					R	included in rs1805008	
rs12821256	KITLG	12	P*	C / T	F	GTGTGAAGTTGTGTGGCAGAAG	0.1µm
					R	AGTCATAAAGTTCCCTGGAGCC	0.1µm

* SNP used in ancestry prediction model

Appendix Table 2b. SNP markers contained in the 50 SNP assay, Multiplex B, with molecular and PCR primer information.

SNP ID	Gene/ Region	Chr	SNP Type P=phenotype A=ancestry	Base Change		PCR Primers	Concentration
Multiplex B							
rs3737576	none	1	A*	A / G	F	GTGTAGGGAACAAGAGATCGGATG	0.1µm
					R	GGAGAGATAGGAGGAAGAGCATAG	0.1µm
rs1375164	OCA2	15	P*	C / T	F	AGAAGTCCCTAGAGGTCATATCCC	0.06µm
					R	CATGATAGGTACCCTGTCCTGTTG	0.06µm
rs7170852	HERC2	15	P	A / T	F	CGATGATACACCAGGCCTTCTCTT	0.4µm
					R	GTTTCCTCAGTGTCTCTACAGTGC	0.4µm
rs4891825	RAAN	18	A*	A / G	F	GCCAGACCCTCAATCAAGACAAAC	0.08µm
					R	GGGAATCTCTAGGGTTGGTAAAGG	0.08µm
rs2714758	none	15	A*	A / G	F	TCTCCTGCACTGAGCTGT	0.2µm
					R	CACGCATGCATCTAGCAGGA	0.2µm
rs1426654	SLC24A5	15	P*	A / G	F	GATTGTCTCAGGATGTTGCAGG	0.1µm
					R	CTAATTCAGGAGCTGAACTGCC	0.1µm
rs16891982	SLC45A2	5	P*	C / G	F	CCAAGTTGTGCTAGACCAGAAAC	0.2µm
					R	CTCATCTACGAAAGAGGAGTCGAG	0.2µm
rs10496971	none	2	A*	G / T	F	GAGACAGTCAGAATGAGTCAGGAG	0.16µm
					R	CATCAAACCTACTCAGCAGCTC	0.16µm
rs916977	HERC2	15	P	A / G	F	GCCTTTCTGTTCTTCTTGACCC	0.22µm
					R	GAGAGACAGGGTGAACTGTTTG	0.22µm
rs1800407	OCA2	15	P	A / G	F	GCTTGTACTCTCTCTGTGTGTGTG	0.1µm
					R	GCGATGAGACAGAGCATGATGA	0.1µm
rs10007810	LIMGH1	4	A*	A / G	F	AACCGTCTTCTCTTGTAGACAGGG	0.1µm
					R	CTTCTGGAGTGTTCTTCCTCTCAG	0.1µm
rs4778138	OCA2	15	P	A / G	F	AGAAAGTCTCAAGGGAAATCAGA	0.24µm
					R	CCCATCGATTTAGCTGTGTTC	0.24µm
rs4918842	HTBP2	10	A*	C / T	F	GTTCTGCCTTACTGCACTTCTCTG	0.28µm
					R	GAATTAATCGGATGCTGAGCCTGG	0.28µm
rs730570	none	14	A*	A / G	F	ACTCACCTGCATCTCACACT	0.26µm
					R	TCCTTCCATATGGCTGAGCA	0.26µm
rs1805007	MC1R	16	P	**C** / G / **T**	F	CGCTACATCTCCATCTTCTACG	0.01µm
					R	ATGAAGAGCGTGCTGAAGACGA	0.01µm

* SNP used in ancestry prediction model

Appendix Table 2c. SNP markers contained in the 50 SNP assay, Multiplex C, with molecular and PCR primer information.

SNP ID	Gene/ Region	Chr	SNP Type P=phenotype A=ancestry	Base Change		PCR Primers	Concentration
Multiplex C							
rs2065982	none	13	A*	C / T	F	GTCCTTCAAGTTCTTCCCAAGG	0.1µm
					R	TAACTCACAGGAAGTGGTCAGTGC	0.1µm
rs1876482	LOC442008	2	A*	C / T	F	CACTTGGAGCATAGTGAGCTGTTG	0.1µm
					R	ATGGGCTGTACCCTCACTATTGG	0.1µm
rs1042602	TYR	11	P*	A / C	F	ATGACCTCTTTGTCTGGATG	1.6µm
					R	ACTCATCTGTGCAAATGTCA	1.6µm
rs1344870	none	3	A*	A / C	F	GAAGAAATATCACATTCGCTCTTAAGTATC	0.1µm
					R	AGGTAAGGTTGTCCCAGGATGT	0.1µm
rs12203592	IRF4	6	P	C / T	F	CAGCTGATCTCTTCAGGCTTTC	0.18µm
					R	CTTCGTCATATGGCTAAACCTGGC	0.18µm
rs4778241	OCA2	15	P	A / C	F	CCACTCTGGAAAGCAGTTTGAC	0.1µm
					R	CTCTGGGATTAATGTCCAGGAGTG	0.1µm
rs1393350	TYR	11	P	A / G	F	CTACTCTTCCTCAGTCCCTTCTCT	0.1µm
					R	CAGAGGCCATGTTAGGGAGATTTG	0.1µm
rs3784230	BRF1	14	A*	C / T	F	TGTGTCCGTGCTGGAGGTT	0.2µm
					R	CAAGTCTTCTTGGAGACTGCTG	0.2µm
rs3827760	EDAR	2	P*	C / T	F	TCCACGTACAACTCTGAGAAGG	0.1µm
					R	TCAAAGAGTTGCATGCCGTCTGTC	0.1µm
rs1540771	IRF4	6	P*	**A** / C / **G** / T	F	CACTGAAGACCACACTCAAGTC	0.2µm
					R	GTAGAAGAGAGAGGAGGGTTTCTG	0.2µm
rs6451722	none	5	A*	A / G	F	CTCTCTGTAAGCAGCTATTGCC	1.6µm
					R	CGGTACTGTCCTGGAAAGCAAA	1.6µm
rs722869	VRK1	14	A*	C / G	F	GCCTTCTGCACTTGGGCATATTCT	0.1µm
					R	GGTAGAGATCTAACAAACCACAGTCAG	0.1µm
rs952718	TBCT12	2	A*	A / C	F	TGAGCCTAGATCCTGACTTCCT	0.16µm
					R	CCAAAGGCCAGATATCTCACTGTC	0.16µm
rs12896399	SLC24A4	14	P*	G / T	F	CTGGCGATCCAATTCTTTGTTC	0.16µm
					R	CCTGTGTGAGACCCAGTACTTA	0.16µm
rs7495174	OCA2	15	P	A / G	F	TTTCCTGGGTCGCCTG	0.2µm
					R	CTTAGGAAGCAAGGCAAGTTCC	0.2µm
rs714857	none	11	A*	C / T	F	AATGGGTCTTGTGAACCTTGGC	0.1µm
					R	CAGAAGTTCTCCAAGGAAACACCC	0.1µm
rs12913832	HERC2	15	P*	A / G	F	CTTCATGGCTCTCTGTGTCTGA	0.1µm
					R	CCTGATGATGATAGCGTGCAGAAC	0.1µm
rs2814778	DARC	1	A*	A / G	F	ATACTCACCCTGTGCAGACAGTTC	0.1µm
					R	GCCCTCATTAGTCCTTGGCTCTTA	0.1µm
rs735612	RYR3	15	A*	G / T	F	CCTTGCAGGCATAACCCAATTCAC	0.1µm
					R	ACATTTCCAAAGATAAAGCAGAAGACTG	0.1µm

* SNP used in ancestry prediction model

Appendix Table 2d. SNP markers contained in the 50 SNP assay, Multiplexes A, B, and C, with SBE primer information.

SNP ID		Extension Primer with non-binding tail (as needed for differentiation)	Concentration
Multiplex A			
rs885479	R	$(t)^{3}$TGGCCGCAACGGCT	1.88µm
rs1834640	F	CATTATATCACAACCTCAGAAACCAC	0.5µm
rs1805009	F	$(t)^{2}$TCATCATCTGCAATGCCATCATC	0.5µm
rs1805008	F	tATCSTGACCCTGCCG	1.88µm
rs1126809	F	$(t)^{14}$GTATTTTTGAGCAGTGGCTCC	0.75µm
rs896788	R	$(t)^{15}$GCATCTACAGTCACCAGCCAC	0.5µm
rs260690	R	GCATGCATGCATGCCTCATAGTTGCTATGAACAGTTTAACAGT	0.38µm
rs6548616	R	$(t)^{11}$TTTCTCTTAGGAGTGGAATTCTCTTAGCTG	0.38µm
rs1667394	R	$(t)^{25}$CAGCAATTCAAAACGTGCATA	0.56µm
rs26722	F	$(t)^{12}$AGCTCTGGATTTACGTAACCATTTTTAACTTTCT	0.44µm
rs10108270	R	$(t)^{19}(ct)^{4}$CTTCTTTCAGGTGAGGACTTAGC	0.75µm
rs1800414	R	$(t)^{30}$GCAGAATCCCRTCAGATATCCTA	0.5µm
rs4911442	F	$(t)^{29}$GGTAACCTGTAAATGGTAGTACCAGAAT	0.75µm
rs4911414	F	$(t)^{26}$TTTTTGTTTGTAAGTCTTTGCTGAGAAATTCATT	0.25µm
rs11547464	R	$(t)^{36}$GTGCGTAGAAGATGGAGATGTAG	0.88µm
rs12821256	R	$(t)^{45}$AGGGCATGTTACTACGGCAC	0.5µm
Multiplex B			
rs3737576	R	TGAGGGGTTAGACCTGCATT	1.0µm
rs1375164	R	$(t)^{6}$TACCCTGTCCTGTTGTTGTCA	0.5µm
rs7170852	R	$(t)^{12}$GCTGTGCGTCTGTTTCC	1.25µm
rs4891825	R	$(t)^{4}(ct)^{4}$GATGGGTGTCTGAATGAAGC	0.5µm
rs2714758	R	$(t)^{17}$GCAGGACCTGGATATGGTCA	0.88µm
rs1426654	F	$(t)^{20}$TCTCAGGATGTTGCAGGC	0.63µm
rs16891982	R	$(t)^{20}$GGTTGGATGTTGGGGCTT	0.75µm
rs10496971	F	$(t)^{22}$CACCTTTAGGCAGAGGCATTT	0.5µm
rs916977	R	$(t)^{11}(ct)^{5}$cTGGGGATGCAGTTTGAGTAGA	0.63µm
rs1800407	F	$(t)^{30}$AGGCATACCGGCTCTCCC	0.38µm
rs10007810	R	$(t)^{15}(gcat)^{3}$gcGGAGATATAAAGGATGCACCACA	0.5µm
rs4778138	F	$(t)^{8}$AATTATATTGAACTGAATGAAAGTGAAAGTGAAAATATAACATATCAAAATTG	0.63µm
rs4918842	R	$(t)^{11}(ct)^{14}$CATCCCAAACTTGGTCCG	0.63µm
rs730570	R	$(t)^{35}$CCATTAATCACACAAATTTTGCAT	0.75µm
rs1805007	R	$(t)^{41}$GTCACGATGCTGTGGTAGC	0.63µm
Multiplex C			
rs2065982	F	tCTTCAAGTTCTTCCCAAGGAAA	0.31µm
rs1876482	F	$(t)^{6}$GCACATCAATTGCAGAGACAA	0.31µm
rs1042602	R	$(t)^{5}$CAAAATCAATGTCTCTCCAGATTTCA	0.63µm
rs1344870	F	TCGCTCTTAAGTATGTTTTCTTGGTC	0.25µm
rs12203592	F	$(t)^{8}$ACTTTGGTGGGTAAAAGAAGG	0.44µm
rs4778241	R	$(t)^{9}$TTGTTGGCTGGTAGTTGCAATT	0.31µm
rs1393350	F	$(t)^{16}$CTCAGTCCCTTCTCTGCAAC	0.31µm
rs3784230	R	$(t)^{11}(ct)^{5}$AGGACGCAGGCATTACCC	0.44µm
rs3827760	F	$(t)^{17}$CGTACAACTCTGAGAAGGCTG	0.31µm
rs1540771	R	$(t)^{17}$TGTTATGAACTGCACGAGTTGG	0.63µm
rs6451722	R	$(t)^{12}(ct)^{3}$cTTCTCAGGATACAGGATTTTGTG	0.63µm
rs722869	F	$(t)^{21}$GCATATTCTTAAATCCGTCTTGACT	0.31µm
rs952718	F	$(t)^{21}$ATTTGAATTTGATCATGAAAGTTGTA	0.44µm
rs12896399	R	$(t)^{24}$GGTTAATCTGCTGTGACAAAGAGA	0.44µm
rs7495174	F	$(t)^{35}$CACCCGTCTGTGCACACT	0.63µm
rs714857	R	$(t)^{29}$TTGTGTACAATTCTCTTAAATATGA	0.31µm
rs12913832	R	$(t)^{35}$TGATAGCGTGCAGAACTTGACA	0.44µm
rs2814778	R	$(t)^{8}(ct)^{15}$CCTCATTAGTCCTTGGCTCTTA	0.31µm
rs735612	F	$(t)^{38}$CCAATTCACTAAACATACATTTGTATTT	0.31µm

Appendix Table 3a. Binsets for the 50-SNP assay, Multiplex A

SNP	Locus Range		Allele	Start	End	Color	Allele	Start	End	Color
rs885479	26.17	30.13	C	26.17	27.17	Yellow	T	29.13	30.13	Red
rs1834640	30.44	33.50	G	30.44	31.44	Blue	A	32.50	33.50	Green
rs1805009	33.82	36.31	G	33.82	34.82	Blue	C	35.31	36.31	Yellow
rs1805008	37.44	40.88	C	37.44	38.44	Yellow	T	39.88	40.88	Red
rs1126809	40.10	42.65	G	40.10	41.10	Blue	A	41.65	42.65	Green
rs896788	42.35	45.40	C	42.35	43.35	Yellow	T	44.40	45.40	Red
rs260690	44.73	48.35	G	44.73	45.73	Blue	T	47.35	48.35	Red
rs6548616	48.94	51.26	G	48.94	49.94	Blue	A	50.26	51.26	Green
rs1667394	51.41	53.70	C	51.41	52.41	Yellow	T	52.70	53.70	Red
rs26722	53.80	55.52	C	53.80	54.80	Yellow	T	54.52	55.52	Red
rs10108270	55.59	58.02	G	55.59	56.59	Blue	T	57.02	58.02	Red
rs1800414	58.94	61.09	C	58.94	59.94	Yellow	T	60.09	61.09	Red
rs4911442	60.71	63.29	G	60.71	61.71	Blue	A	62.29	63.29	Green
rs4911414	63.92	67.00	G	63.92	64.92	Blue	T	66.00	67.00	Red
rs11547464	64.73	68.16	C	64.73	65.73	Yellow	T	67.16	68.16	Red
rs12821256	67.71	70.76	G	67.71	68.71	Blue	A	69.76	70.76	Green

Appendix Table 3b. Binsets for the 50-SNP assay, Multiplex B

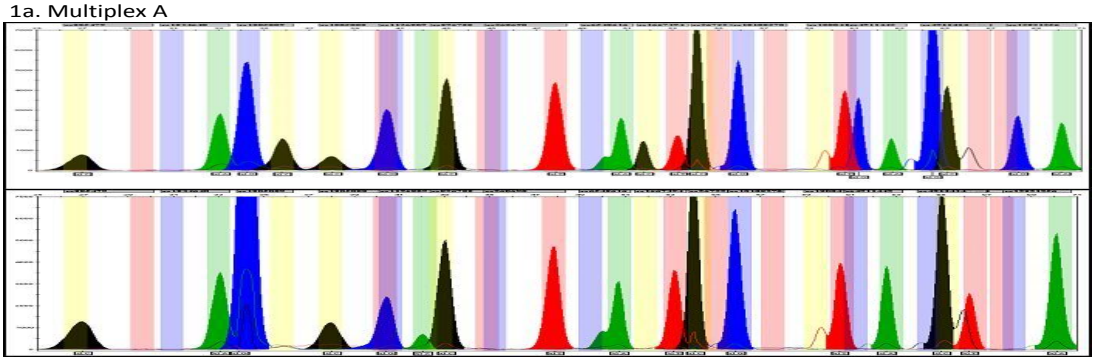
SNP	Locus Range		Allele	Start	End	Color	Allele	Start	End	Color
rs3737576	35.15	37.71	C	35.15	36.15	Yellow	T	36.71	37.71	Red
rs1375164	37.46	38.73	G	37.46	38.46	Blue	A	37.73	38.73	Green
rs7170852	39.26	42.40	A	39.26	40.26	Green	T	41.40	42.40	Red
rs4891825	42.28	45.03	C	42.28	43.28	Yellow	T	44.03	45.03	Red
rs2714758	44.97	47.52	C	44.97	45.97	Yellow	T	46.52	47.52	Red
rs1426654	45.22	47.16	G	45.22	46.22	Blue	A	46.16	47.16	Green
rs16891982	47.56	49.03	G	47.56	48.56	Blue	C	48.03	49.03	Yellow
rs10496971	48.91	51.48	G	48.91	49.91	Blue	T	50.48	51.48	Red
rs916977	51.13	53.27	C	51.13	52.13	Yellow	T	52.27	53.27	Red
rs1800407	53.09	55.04	G	53.09	54.09	Blue	A	54.04	55.04	Green
rs10007810	55.20	57.08	C	55.20	56.20	Yellow	T	56.08	57.08	Red
rs4778138	57.58	59.31	G	57.58	58.58	Blue	A	58.31	59.31	Green
rs4918842	60.06	61.99	G	60.06	61.06	Blue	A	60.99	61.99	Green
rs730570	62.37	64.44	C	62.37	63.37	Yellow	T	63.44	64.44	Red
rs1805007	64.18	65.90	G	64.18	65.18	Blue	A	64.90	65.90	Green

Appendix Table 3c. Binsets for the 50-SNP assay, Multiplex C

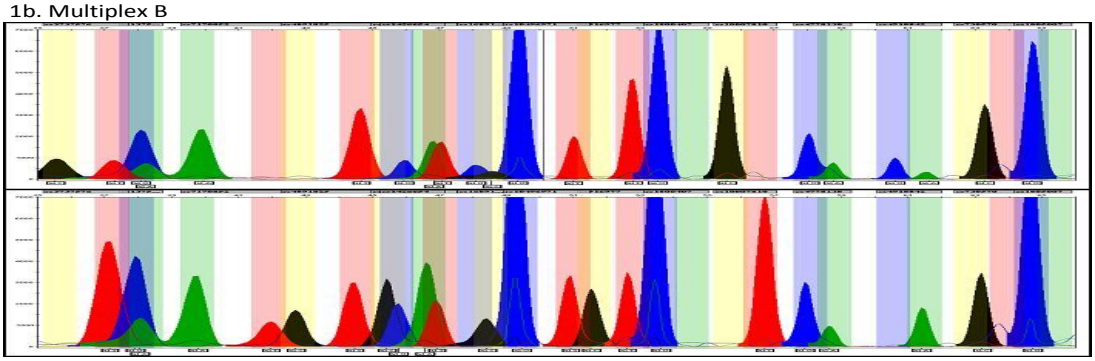
SNP	Locus Range		Allele	Start	End	Color	Allele	Start	End	Color
rs2065962	32.56	35.43	C	32.56	33.56	Yellow	T	34.43	35.43	Red
rs1876482	34.56	36.96	C	34.56	35.56	Yellow	T	35.96	36.96	Red
rs1046602	36.30	39.54	G	36.30	37.30	Blue	T	38.54	39.54	Red
rs1344870	37.37	38.71	A	37.37	38.37	Green	C	37.71	38.71	Yellow
rs12203592	39.62	41.77	C	39.62	40.62	Yellow	T	40.77	41.77	Red
rs4778241	39.71	43.47	G	39.71	40.71	Blue	T	42.47	43.47	Red
rs1393350	41.42	43.76	G	41.42	42.42	Blue	A	42.76	43.76	Green
rs3784230	43.87	45.71	G	43.87	44.87	Blue	A	44.71	45.71	Green
rs3827760	45.52	47.23	C	45.52	46.52	Yellow	T	46.23	47.23	Red
rs1540771	47.36	49.56	C	47.36	48.36	Yellow	T	48.56	49.56	Red
rs6451772	49.71	52.20	C	49.71	50.71	Yellow	T	51.20	52.20	Red
rs722869	50.85	52.96	G	50.85	51.85	Blue	C	52.06	52.96	Yellow
rs952718	52.96	54.40	A	53.40	54.40	Green	C	52.96	53.96	Yellow
rs12896399	54.61	56.15	A	55.15	56.15	Green	C	54.61	55.61	Yellow
rs7495174	56.50	58.35	C	56.50	57.50	Blue	A	57.35	58.35	Green
rs714857	58.86	60.47	G	58.86	59.86	Blue	A	59.47	60.47	Green
rs12913832	61.50	63.69	C	61.50	62.50	Yellow	T	62.69	63.69	Red
rs2814778	63.75	65.85	C	63.75	64.75	Yellow	T	64.85	65.85	Red
rs735612	69.00	71.50	G	69.00	70.00	Blue	T	70.50	71.50	Red

Appendix Figure 1: Examples of electropherograms of the three multiplexes incorporating the 50 slescted SNPs

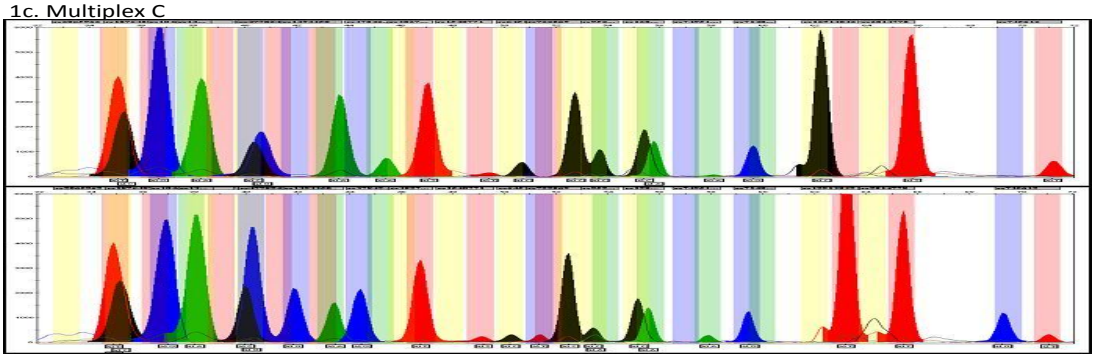
1a. Multiplex A



1b. Multiplex B



1c. Multiplex C



Data Collection Tools

Data Collection Tools: Adult Sample Collection Assent Form (given to volunteer, or parent/legal guardian of child volunteer <6 years old) .

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON DC

ADULT / PARENTAL INFORMATION SHEET

PROJECT TILTLE: **DNA based inference of ancestry and phenotypic traits for forensic applications**

GWU IRB # 060907 Expiration Date: 10/21/2010

Contact Information

Name: Daniele Podini and Katherine Butler
Department: Forensic Sciences
Email: forensicdnastudy@gmail.com (GWU students)
forensicdnastudy2@gmail.com (non GWU students)

Purpose of this Study: You are being asked to participate in a scientific research project funded by the National Institute of Justice that involves the study of your DNA. DNA is the substance that contains the information that makes us human and that makes us look different from each other. We are interested in studying (1) parts of DNA that are involved in determining eye, skin, and hair color, (2) parts of the DNA that affect specific traits like balding, freckles etc. and (3) parts of DNA that can help determine a person's ancestry, for example, whether your family originally came from Europe, Asia, African or a combination of these.

Procedures: This procedure can take place in a variety of locations, including your home, a classroom, or at the laboratory located at the Department of Forensic Sciences at The George Washington University. A member of the research team (Dr. Daniele Podini, Katherine Butler, Joni Johnson, or Ronald Lai) will always be present to help perform the procedure and to address your questions/concerns. The procedure will take around 15 minutes and is complete in one visit, no additional procedures or follow-up will be asked of you for this study. We intend to test approximately 200 individuals.

- You will collect your own DNA using a cotton swab that you will gently rub against the inside of your cheeks. This process is completely painless, and only takes a couple of seconds.
- Your hair and skin color will be measured using a small device known as a "spectrophotometer". This device is held up to the area of skin/hair, and automatically takes measurements of color. It is painless and completely safe, and only takes a couple of seconds. Multiple measurements may be taken at different sites of hair/skin.
- Your eye color will be compared to a color chart or known pictures of eye colors, and we will decide which one matches you the best.
- You will be asked to fill out a questionnaire. It will ask you questions such as where you think your family came from, what you think your hair/skin/eye color are, and information regarding certain traits like whether you are balding, your hair is curly, and if you have freckles. Again, your participation is voluntary, so you do not have to answer any particular question(s) that you do not want to.

NOTE: If you are a parent allowing this procedure to be performed on your child aged six (6) or under, the same basic procedure will be performed on your child. Differences from the above procedure will be (1) you or a member of the research team will collect the DNA sample from the child (method of collection is the same) and (2) you will be asked to fill out the questionnaire for your child.

Data Collection Tools: Adult Sample Collection Assent Form (given to volunteer, or parent/legal guardian of child volunteer <6 years old) (continued).

Voluntary Participation / Withdrawal: Your participation in this study is voluntary and you may decide not to participate or you may withdraw from the study at any time you wish. If you do choose to withdraw during the procedure, any DNA or data obtained from you will be immediately discarded. If you are a GWU student your academic standing will not, in any way, be affected should you choose not to participate or if you decide to withdraw from the study at any time and no member of GWU faculty will know whether you agree to participate to this study or not.

Confidentiality: We will not keep a list of names of people who participate to this study. All results will be anonymous, even to members of the research team. Once samples are collected it will not be possible to identify individuals in reports and/or publications at any point of this project. A number will be assigned to your sample and questionnaire but there will be no personal data associated with this number, its only purpose is to identify the sample and associate it to the data and questionnaire. Samples collected for this project may be used in the future for similar studies.

Risks: One risk is potential harm during the collections of the buccal swab which requires the insertion of a foreign object (long Q-tip) into the your mouth, a second risk is transferring potential microbes from one subject to the next when using the spectrophotometer to measure your skin color. To minimize risks we will use sterile cotton Q-tips and we will sterilize the spectrophotometer lens between each measurement. The level of risk to adults through the use of buccal swabs and spectrophotometer is to be considered very minimal and, when the collection is properly performed by an adult, the same low risk level exists for children.

Benefits: The ability to predict what someone looks like can greatly help in investigating a crime. For example, when there is a bloodstain found at a crime scene, if investigators have an idea of what the person who left the blood looks like, they can focus their search for the unknown victim/suspect better.

Questions: If you have questions, including questions about your rights, have concerns or complaints, or think you have been harmed. You can contact a member of the research team at forensicdnastudy@gmail.com (GWU students) or forensicdnastudy2@gmail.com (non GWU students). If you have questions on the rights of research subjects or simply want to talk to someone else, call the Office of Human Research at 202-994-2715.

DO NOT USE AFTER THE EXPIRATION DATE OF: 10/21/2010

A P P R O V E D
The George Washington University
Institutional Review Board
·FWA00005945·

Data Collection Tools: Child Sample Collection Assent Form (given to child volunteer >6 years old)

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON DC

Forensic DNA Research Study
CHILD ASSENT FORM
GWU IRB # 060907 Expiration Date: 10/21/2010

Contact Information

Name: Katherine Butler
Department: Forensic Sciences
Email: forensicdnastudy@gmail.com

We would like to take a sample from inside your mouth and look at your skin, hair, and eyes. We will also ask you or your parents what part of the world your family came from. This will help us learn why people look different. The National Institute of Justice, which is part of the US government is giving us the money for this study.

Knowing why people look different can help police solve crimes. It will help police if they know to look for someone with blue or brown eyes, or red or blonde hair.

If you let us, we can take these samples in our lab at the University, or in your home. There are four different people who can help you take samples and answer questions. Their names are Daniele, Katherine, Joni, and Ron. Your parents will also be present. It will take us around 15 minutes and you will not need to come back after we finish today. Here is what we would like to do:

- We will help you take a sample from inside your mouth. We will give you a long Q-tip for you to put in your mouth and rub inside your cheeks. It will not hurt and it is fast.
- We will look at the hair on your head and the skin on your arm with a special camera. It will not hurt and it is fast.
- We will look at your eyes and compare them to a color chart or to pictures of other people's eyes.
- You or your parents will be asked to fill out a questionnaire. It will ask you things like what part of the world you think your family came from and what you think your hair, skin, and eye colors are. You do not have to answer any questions that you do not want to.

You can stop at any time by saying "STOP". If you say "STOP" we will not take any more samples or ask any more questions. Also if you say "STOP" we will throw away any samples you gave us or questions you already answered. There is nothing wrong with saying "STOP" for any reason and you don't need to explain why you don't want to continue.

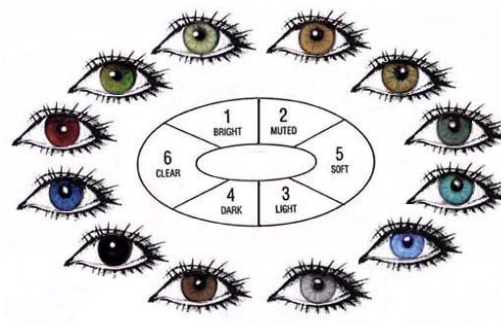
We will not put your name on your sample or on the question sheet and we will not keep track of who gave samples.

A P P R O V E D
The George Washington University
Institutional Review Board
·FWA00005945·

Data Collection Tools:. Sample Collection Checklist (completed by researcher)

CODE _____

Ancestry _____



Special Eye features _____

<u>Spectrophotometer Measurements checklist</u>

1. control (white calibration plate)
2. negative control (blank space)
3. wrist (right)
4. wrist 2
5. forearm
6. forearm 2
7. above elbow
8. above elbow 2
9. below armpit
10. below armpit 2
11. forehead
12. forehead 2
13. cheek
14. cheek 2
15. hair
16. hair 2
17. hair 3
18. control (white calibration plate)
19. negative control (blank space)

A P P R O V E D
The George Washington University
Institutional Review Board
·FWA00005945·

❑ Collect Swabs

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON DC

DEPARTMENT OF FORENSIC SCIENCES

QUESTIONNAIRE

Genetic Inference of Ancestry and Phenotypic Traits for Forensic Applications

November 2009

Department of Forensic Sciences
Forensic Molecular Biology Laboratory
2100 Foxhall Road, NW
Somers Hall – Bottom Level

CODE ID: _____

1. Sex: ❑ Female ❑ Male

2. Height: _____ Age: ❑ < 18 ❑ 18-39 ❑ 40-60 ❑ 60+

3. Body build: ❑ Light ❑ Medium ❑ Heavy

4. What is your Ethnic / Ancestral origin?

❑ European – N W S E
❑ Africa – N W S E
❑ Asia – N W S E
❑ African American
❑ Hispanic
❑ Middle Eastern
❑ Pacific Islander
❑ Native American – N W S E
❑ Other _____

Specify:_____
_____

5. What are the Ethnic / Ancestral origins of your **paternal** grandparents?

Grandfather (paternal):

❑ European – N W S E
❑ Africa – N W S E
❑ Asia – N W S E
❑ African American
❑ Hispanic
❑ Middle Eastern
❑ Pacific Islander
❑ Native American – N W S E
❑ Other _____

Specify:_____
_____

Grandmother (paternal):

❑ European – N W S E
❑ Africa – N W S E
❑ Asia – N W S E
❑ African American
❑ Hispanic
❑ Middle Eastern
❑ Pacific Islander
❑ Native American – N W S E
❑ Other _____

Specify:_____
_____

6. What are the Ethnic / Ancestral origins of your **maternal** grandparents?

Grandfather (maternal):

- ❑ European – N W S E
- ❑ Africa – N W S E
- ❑ Asia – N W S E
- ❑ African American
- ❑ Hispanic
- ❑ Middle Eastern
- ❑ Pacific Islander
- ❑ Native American – N W S E
- ❑ Other _____

Specify:_____
_____

Grandmother (maternal):

- ❑ European – N W S E
- ❑ Africa – N W S E
- ❑ Asia – N W S E
- ❑ African American
- ❑ Hispanic
- ❑ Middle Eastern
- ❑ Pacific Islander
- ❑ Native American – N W S E
- ❑ Other _____

Specify:_____
_____

7. (a) What is your natural head **hair** color?

- ❑ Light Blond
- ❑ Dark Blond
- ❑ Light Brown
- ❑ Dark Brown
- ❑ Black / Very Dark Brown
- ❑ Red
- ❑ Reddish Brown (Auburn)
- ❑ Other: _____

(b) From the chart, circle what you think best resembles your natural hair color.

Source: killerstrands.blogspot.com

8. (a) How would you classify the natural **type** of your head hair?

- ❑ Straight
- ❑ Wavy
- ❑ Curly
- ❑ Kinky / Coiled
- ❑ Other: _____

(b) Check the picture that best resembles the natural type of your hair.

❑ ❑ ❑ ❑ ❑

9. How would you classify the natural **texture** of your head hair?

- ❑ Fine
- ❑ Medium
- ❑ Coarse
- ❑ Other: _____

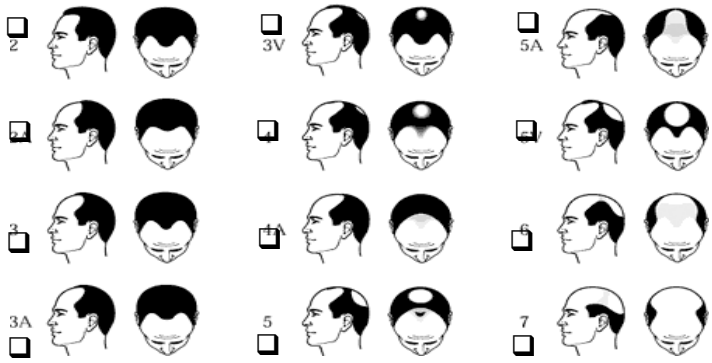
10. How would you classify the **thickness** of your head hair?

- ❑ Thin
- ❑ Medium
- ❑ Thick
- ❑ Other: _____

11. (a) Are you bald or in the process of **balding**? If so, when did this begin?

❑ Yes. I began balding at around ____ years old. ❑ No

(b) If yes, which of the following pictures best describes your current stage of balding?



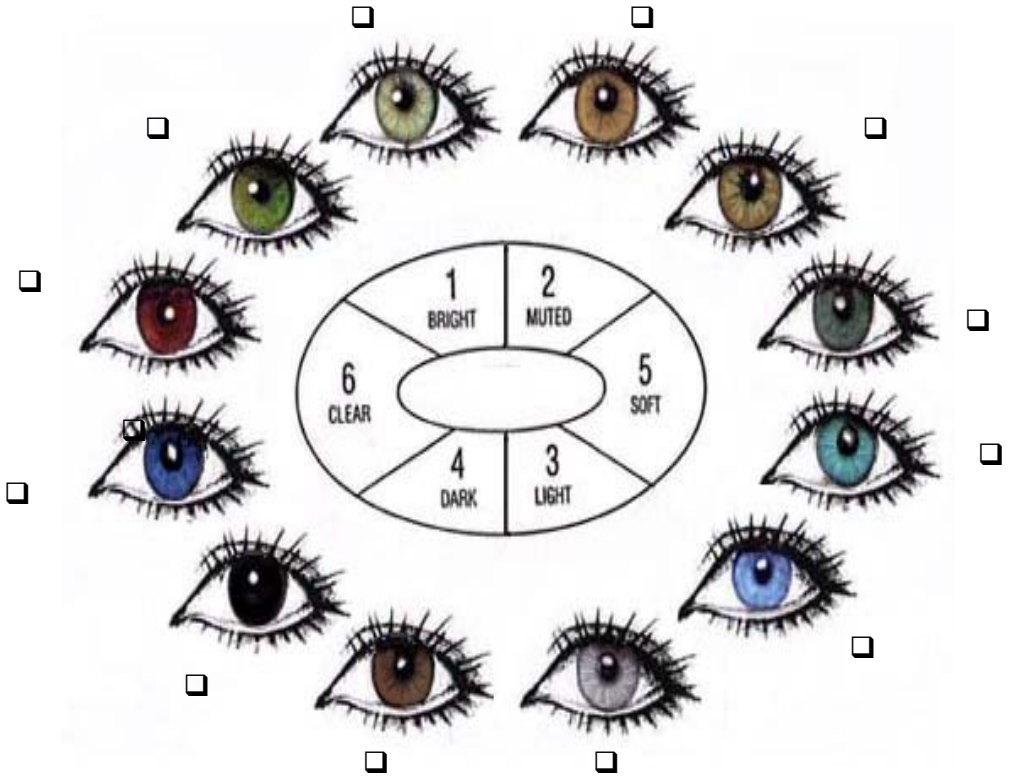
12. Is there a history of balding in your family? If so, please specify which side.

❑ Yes (circle one): Maternal / Paternal / Both ❑ No

13. (a) What is your natural **eye** color?

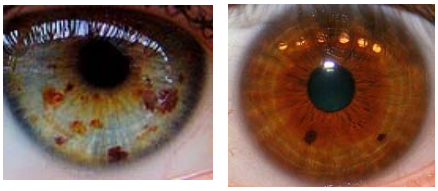
❑ Light Blue
❑ Dark Blue
❑ Grey
❑ Light Green
❑ Dark Green
❑ Light Brown
❑ Dark Brown
❑ Hazel
❑ Black / Very Dark Brown
❑ Other: _____

(b) Check the picture that best resembles your natural eye color.



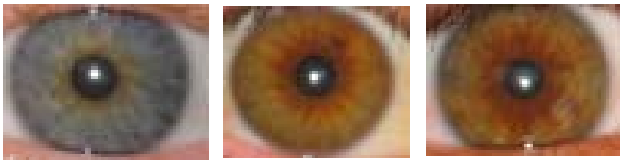
Source: www.color-chart.org

(c) Are there spots in your eyes similar to the pictures below? ❑ Yes ❑ No



If yes…
Which eyes? ❑ Both ❑ Right ❑ Left
How many? ❑ Less than 5 ❑ More than 5
What color are the spots? _____

(d) Are there visible rings around your pupils, similar to pictures below?



If yes…
Which eyes? ❑ Both ❑ Right ❑ Left
What size? ❑ Small ❑ Medium ❑ Big
What color? _____

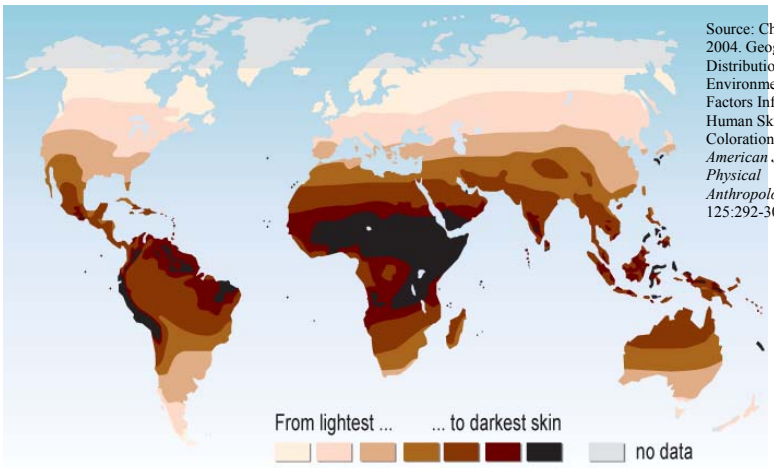
(e) Anything else special about your eyes? If so, please list and describe.
_____
_____
_____

14. (a) In your own words, describe your natural **skin** color: _____

(b) What would you classify your natural **skin** color as from the list below?

- ❑ Light – Pale white or freckled
- ❑ Fair – White
- ❑ Medium – White to light brown
- ❑ Olive – Moderate brown
- ❑ Brown – Dark brown
- ❑ Black – Very dark brown to black
- ❑ Other: _____

(c) Check the box that you think best resembles your natural skin color.



Source: Chaplin G. 2004. Geographic Distribution of Environmental Factors Influencing Human Skin Coloration. *American Journal of Physical Anthropology* 125:292-302.

❑ ❑ ❑ ❑ ❑ ❑ ❑

15. How abundant are **freckles** on your skin?

❑ None
❑ Scarce
❑ Moderate
❑ Abundant

16. Fill in the following information if known:

Father: eye color _____ skin color _____ hair color _____
Mother: eye color _____ skin color _____ hair color _____

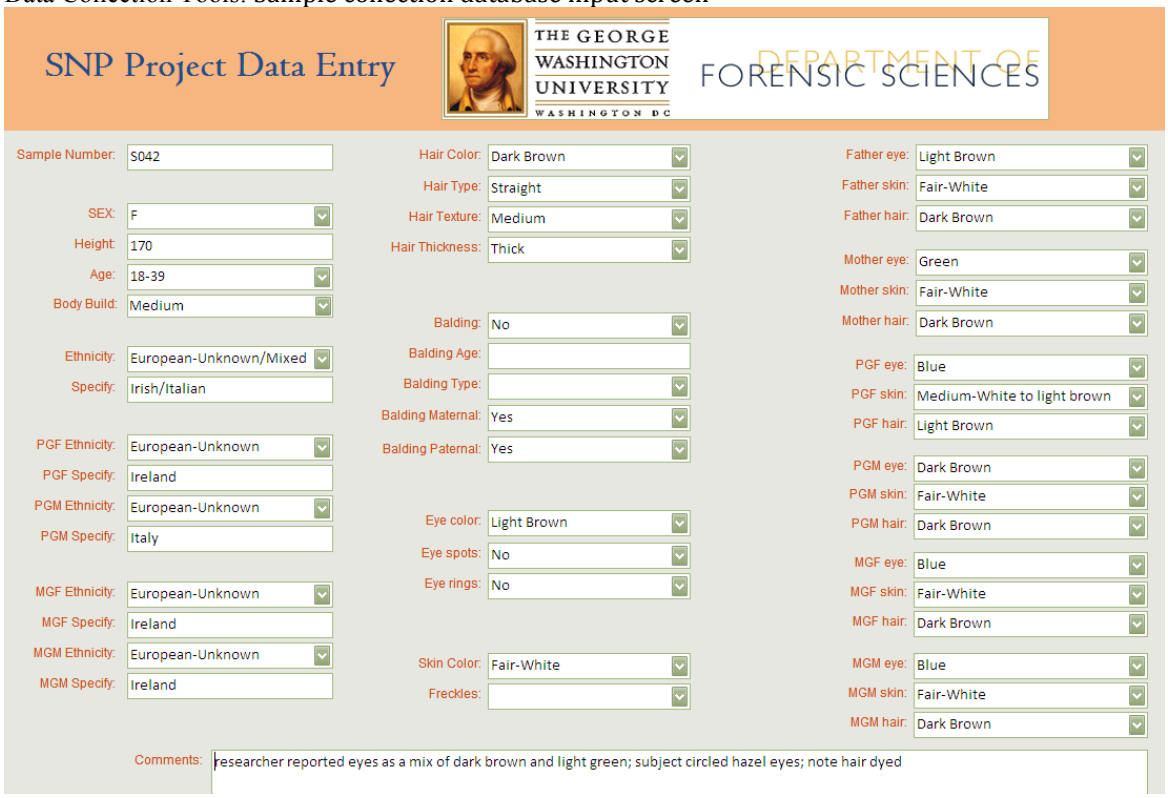
Paternal grandfather: eye color _____ skin color _____ hair color _____
Paternal grandmother: eye color _____ skin color _____ hair color _____

Maternal grandfather: eye color _____ skin color _____ hair color _____
Maternal grandmother: eye color _____ skin color _____ hair color _____

Thank you for your participation!

A P P R O V E D
The George Washington University
Institutional Review Board
·FWA00005945·

Data Collection Tools: Sample collection database input screen



SNP Project Data Entry

THE GEORGE WASHINGTON UNIVERSITY WASHINGTON DC

DEPARTMENT OF FORENSIC SCIENCES

Sample Number:	S042
SEX:	F
Height:	170
Age:	18-39
Body Build:	Medium
Ethnicity:	European-Unknown/Mixed
Specify:	Irish/Italian
PGF Ethnicity:	European-Unknown
PGF Specify:	Ireland
PGM Ethnicity:	European-Unknown
PGM Specify:	Italy
MGF Ethnicity:	European-Unknown
MGF Specify:	Ireland
MGM Ethnicity:	European-Unknown
MGM Specify:	Ireland
Hair Color:	Dark Brown
Hair Type:	Straight
Hair Texture:	Medium
Hair Thickness:	Thick
Balding:	No
Balding Age:	
Balding Type:	
Balding Maternal:	Yes
Balding Paternal:	Yes
Eye color:	Light Brown
Eye spots:	No
Eye rings:	No
Skin Color:	Fair-White
Freckles:	
Father eye:	Light Brown
Father skin:	Fair-White
Father hair:	Dark Brown
Mother eye:	Green
Mother skin:	Fair-White
Mother hair:	Dark Brown
PGF eye:	Blue
PGF skin:	Medium-White to light brown
PGF hair:	Light Brown
PGM eye:	Dark Brown
PGM skin:	Fair-White
PGM hair:	Dark Brown
MGF eye:	Blue
MGF skin:	Fair-White
MGF hair:	Dark Brown
MGM eye:	Blue
MGM skin:	Fair-White
MGM hair:	Dark Brown
Comments:	researcher reported eyes as a mix of dark brown and light green; subject circled hazel eyes; note hair dyed