Forensic Ancestry and Phenotype SNP Analysis and Integration with Established Forensic Markers

by Katherine Butler Gettings

B.S. in Biology, December 1997, Virginia Polytechnic Institute & State University
M.S. in Criminal Justice, May 2001, Virginia Commonwealth University

A Dissertation submitted to

The Faculty of
the Columbian College of Arts and Sciences
of The George Washington University
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy

August 31, 2013

Dissertation directed by

Daniele S. Podini
Assistant Professor of Forensic Molecular Biology and of Biological Sciences

The Columbian College of Arts and Sciences of The George Washington University certifies that Katherine Butler Gettings has passed the Final Examination for the degree of Doctor of Philosophy as of July 9, 2013. This is the final and approved form of the dissertation.

Forensic Ancestry and Phenotype SNP Analysis
and Integration with Established Forensic Markers

Katherine Butler Gettings

Dissertation Research Committee:

Daniele S. Podini, Assistant Professor of Forensic Molecular Biology and of Biological Sciences, Dissertation Director

Ioannis Eleftherianos, Assistant Professor of Molecular Biology, Committee Member

Moses S. Schanfield, Professor of Forensic Sciences and Anthropology, Committee Member

Dedication

"I learned this, at least, by my experiment: that if one advances confidently in the direction of his dreams, and endeavors to live the life which he has imagined, he will meet with a success unexpected in common hours." "If you have built castles in the air, your work need not be lost; that is where they should be. Now put the foundations under them." *Walden*, Henry David Thoreau

This work is dedicated to my husband, Rob. From its inception, you believed in my dream of returning to school for this degree, and you happily made sacrifices so I could live the life I had imagined. And to our own genetics experiment, Owen: may you be inspired to follow your dreams.

Acknowledgments

The author wishes to acknowledge the invaluable contributions of several individuals. First and foremost, as an advisor, Dr. Daniele Podini has provided constant guidance, patience, and encouragement over the past four years. Through his leadership and perseverance, many obstacles have been overcome.

In addition, Dr. Moses Schanfield (GWU) donated samples from his collection, provided statistical support by performing and explaining PCA for SNP selection and CHAID decision tree analysis for ancestry and eye color determination, and assisted with diplotype evaluation. Dr. Heather Gordish-Dressman (CNMC) also provided statistical support by performing chi-squared analysis for SNP selection and performing and explaining MLR for ancestry determination. Dr. Joseph Devaney (CNMC) provided support and advice, specifically facilitating the attempt to obtain NGS data from forensic STR loci. Drs. John Butler and Peter Vallone (NIST) provided support and advice, and donated samples that had been well characterized, greatly facilitating portions of this project. Becky Hill and Erica Butts (NIST) prepared the test set samples and provided support with interpretation of results.

Many GWU students have contributed to this project by collecting volunteer samples, helping in SNP assay design and with SNP genotyping. Specifically, Ron Lai, Joni Johnson, Lorena Lara, Resham Uttamchandani, Michelle Peck, and Jessica Hart all made significant contributions.

The National Institute of Justice provided funding for this project in the form of a Forensic DNA Research and Development Grant 2009-DN-BX-K178 and a PhD Fellowship Grant 2011-CD-BX-0123.

Abstract of Dissertation

Forensic Ancestry and Phenotype SNP Analysis
and Integration with Established Forensic Markers

When an evidential DNA profile 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 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. With this research, single base primer extension (SBE) technology was used 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. The assay has been optimized to a sensitivity level comparable to current forensic DNA analyses, and has shown robust performance on forensic-type samples. In addition, three prediction models were developed and evaluated for ancestry in the U.S. population, and two models were compared for eye color prediction, with the best models and interpretation guidelines yielding correct information for 98% and 100% of samples, respectively. Also, because data from additional DNA markers (STR, mitochondrial and/or Y chromosome DNA) may be available for a forensic evidence sample, the possibility of including this data in the ancestry prediction was evaluated, resulting in an improved prediction with the inclusion of STR data and decreased performance when including mitochondrial or Y chromosome data. Lastly, the possibility of using next-generation sequencing (NGS) to genotype forensic STRs (and thus, the possibility of a multimarker multiplex

incorporating all forensic markers) was evaluated on a new platform, with results showing the technology incapable of meeting the needs of the forensic community at this time.

Table of Contents

List of Figures

List of Figures (continued)

List of Tables

List of Symbols / Nomenclature

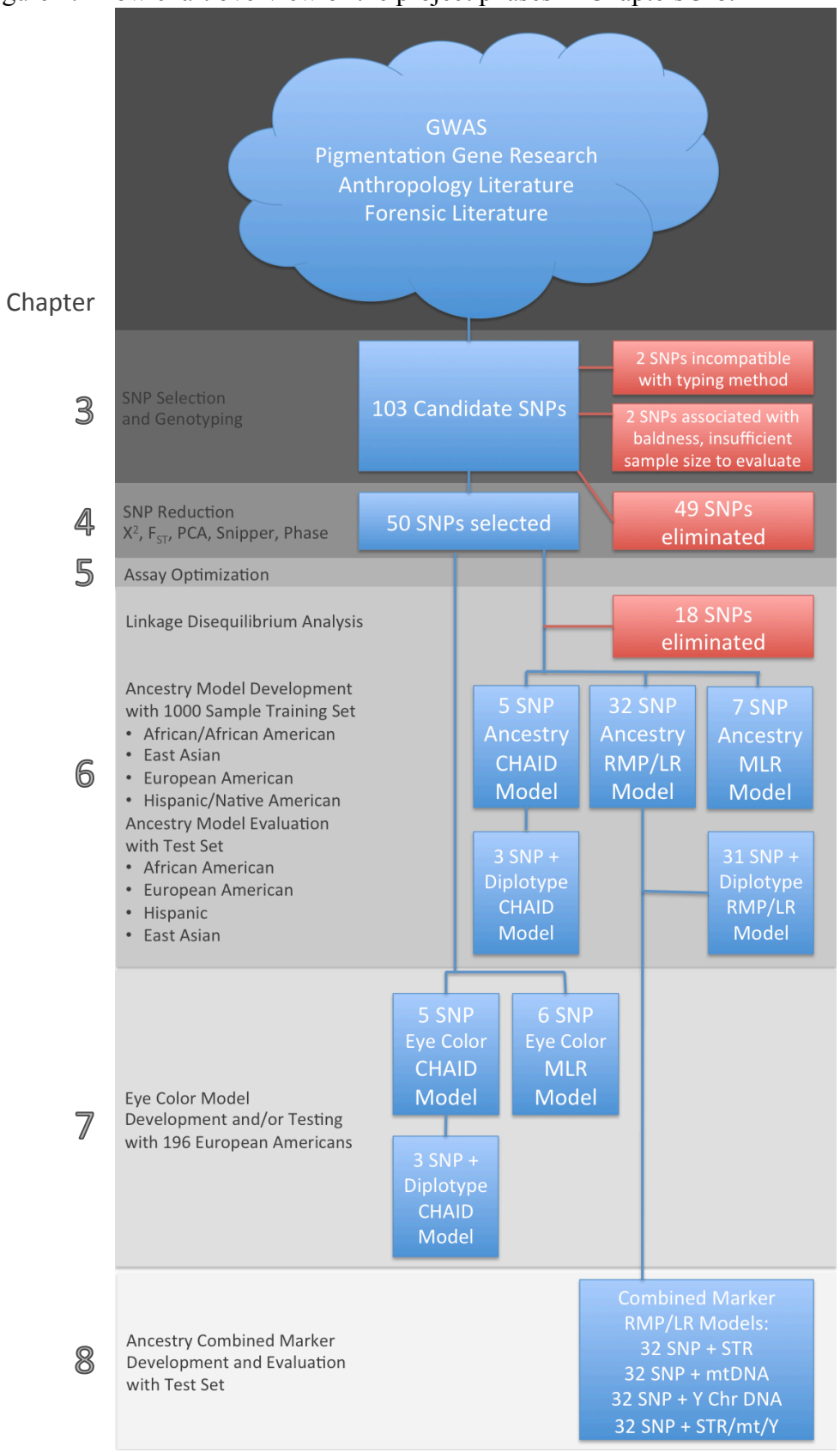
CHAID:	Chi-squared Automatic Interaction Detector
LD:	Linkage Disequilibrium
LR:	Likelihood Ratio
MLR:	Multinomial Logistic Regression
mtDNA:	Mitochondrial DNA
NGS:	Next Generation Sequencing
PCA:	Principle Component Analysis
RMP:	Random Match Probability
SBE:	Single Base Extension
SNP:	Single Nucleotide Polymorphism
STR:	Short Tandem Repeat
Y-STR:	Y Chromosome Short Tandem Repeat

Chapter 1: Overview

Current Forensic DNA casework typically employs Short Tandem Repeat (STR) analysis of crime scene evidence and comparison of the resulting profile to known profiles or databases. However, cases often go unsolved when an evidence DNA profile does not match any of the suspects, or any of the profiles in the available databases. An investigative tool that could provide more information regarding the donor of the unmatched profile would be extremely useful in these cases. Advances in genetic knowledge and technologies present new possibilities for maximizing the information content obtained from DNA samples, but adapting these technologies to the nuances of forensic samples is challenging.

The research described herein focuses on the development of a tool that can aid investigators by providing ancestry and phenotypic information on an unmatched profile. The project can be divided into seven distinct phases, presented as separate chapters. An overview of these phases can be seen in Figure 1. First, as described in Chapter 3, 103 candidate SNPs were chosen from the relevant literature, then a sample set was genotyped at these candidate SNPs. This sample set was composed of volunteer samples with both phenotype and ancestry data, and laboratory samples with only ancestry data. Additional sample genotype data was added from available databases (with only ancestry data). This overall data set was evaluated for candidate SNP reduction (Chapter 4) using several statistical approaches for both ancestry and pigmentation prediction. Ancestry prediction targeted the root populations forming the primary U.S. populations: African, East Asian, European, and Native American. Fifty SNPs were selected for a final assay to be used in forensic casework, and this assay was optimized (Chapter 5). In Chapter 6,

Figure 1. Flow chart overview of the project phases in Chapters 3-8.



ancestry prediction models were developed based on a training set composed of African/African American, East Asian, European/European American, and Hispanic/Native American. These models were evaluated with a separate test set of African American, East Asian, European American, and Hispanic American individuals. Chapter 7 contains results of the eye color model development and evaluation based on the volunteer samples with phenotype data. In Chapter 8, the possibility of combining traditional forensic markers with the SNP data for ancestry prediction is explored. Lastly, Chapter 9 contains an initial attempt at adapting a next-generation sequencing method to forensic STRs, which is the first step in designing a multi-marker multiplex for forensic purposes.

Candidate SNP Selection, Sample Collection, and SNP Genotyping

In this first phase of research, a list of candidate SNPs was culled from the literature on ancestry and phenotype markers. One hundred and three SNPs that were compatible with the genotyping system were selected: 43 ancestry markers, 53 phenotype markers associated with pigmentation, and seven markers associated with other physical characteristics such as hair form or baldness. Eleven assays were developed to genotype this set of SNPs using the single base extension (SBE) methodology. Concurrently, volunteer DNA samples were collected over a two-year period, with corresponding ancestry and phenotype data. These volunteer samples, along with a selection of samples already available in the laboratory (with ancestry data only), were genotyped at the candidate SNPs. Genotypes for additional samples with ancestry data only were gathered from publicly available resources to complete the candidate SNP dataset.

Candidate SNP Evaluation / Reduction

During this phase, a number of statistical approaches were employed to evaluate the ancestry and phenotype information content of the candidate SNPs. For ancestry association, these included chi-squared analysis, principle component analysis (PCA), pairwise F_{ST}, and Snipper (an online tool that ranks SNPs based on their ability to diverge predefined groups. Pigmentation phenotype associations were evaluated among European Americans using chi-squared analyses and PCA for hair, skin, and eye colors; and haplotype association to skin color in gene regions where many candidate SNPs were located. By cross-referencing the results of all these analyses, a subset of 50 SNPs were selected for inclusion in a final assay.

Development / Optimization of 50-SNP Assay

Once the panel of SNPs most predictive of ancestry and phenotype were chosen, the next phase was to develop an assay for genotyping these SNPs. This assay was built with the forensic practitioner in mind: achieving the sensitivity of currently used forensic methodologies, showing robust results on mock forensic evidence samples, and using the same equipment found in forensic DNA casework laboratories. The assay consists of three reactions to reduce primer interactions and improve balance among the SNPs. Once optimized, the method was used to genotype a set of samples that would become the test set for prediction model evaluation.

Development / Evaluation of Ancestry Models

While the preceding phases are important steps toward the final goal, in a practitioner's hands, the SNP genotype data is useless without a prediction model. The ideal model would incorporate all of the ancestry information content from the 50 SNPs and consistently predict the correct ancestry for a test set of samples, composed of the

populations of interest. Several statistical frameworks were evaluated within this project: a random match probability/likelihood ratio (RMP/LR) model that incorporates all SNPs which do not show evidence of linkage disequilibrium (LD); a multinomial logistic regression (MLR) model and a chi-squared automatic interaction detector (CHAID) decision tree model, both using a small subset of highly informative SNPs. Further, toward incorporating all informative SNP data, a haplotype approach was evaluated for the RMP/LR and CHAID models.

Development / Evaluation of Pigmentation Models

Determining an appropriate statistical framework and the limitations thereof is key to providing investigative information on phenotype as well. Due to a disproportionately European-centered body of research on pigmentation, and the fact that our sample set with corresponding phenotype data is also disproportionately European American in origin, pigmentation prediction models were only evaluated among this population. The relative complexity of hair and skin pigmentation prevented model development in our limited sample set. Eye color models evaluated include a published model based on MLR and a CHAID decision tree model, both using a small subset of highly informative SNPs. A haplotype approach was evaluated for the CHAID model.

Integration of Established Forensic Markers

The gold-standard forensic DNA analysis of individually-identifying STR loci would currently always precede any SNP analysis, so being able to harness and incorporate any ancestry information present in the STR profile could improve ancestry prediction. This possibility, as well as the possibility of incorporating lesser-used mitochondrial DNA (mtDNA) and/or YSTR data, was evaluated in this phase of the project. Because the

current forensic STR interpretation is based on a RMP calculation, these data could be incorporated directly into the SNP-based RMP/LR framework. The mtDNA and Y haplotype information was also incorporated based on the haplotype frequencies in the different populations, and the impact of integration was evaluated for each marker type.

Evaluation of Next-Generation Sequencing for Forensics

In this final phase of the project, a preliminary analysis of an emergent next-generation sequencing (NGS) technology was evaluated for use on forensic samples. The starting point for such an analysis is the ability of a new technology to genotype the forensic STR loci (because most NGS methods are designed for SNP typing, and some methods are not amenable to genotyping repeat-motifs, and because no new technology could replace current forensic methods without the ability to genotype STRs). Five different forensic STR loci were selected for their significant sequence variation, which would add to the discriminating ability compared to the current repeat-unit counting method. Working with collaborators at the Children's National Medical Center (CNMC), these loci were evaluated on the PacBio RS instrument.

Overall, a DNA based assay that can provide ancestry and phenotypic information of an individual complements a criminal investigation when no STR match is found. Also in missing person cases when a corpse is found and physical traits are unidentifiable due to the conditions of the remains, a genetic prediction of ancestry and phenotype could add to the anthropological data for a description of the individual and aid the identification process.

Along with providing this information, it is imperative to give a statistical weight to the ancestry estimation/phenotype prediction. The creation of a statistical framework

allows for a confidence level to be assigned to this information, and gives investigators perspective when incorporating this information into their investigation.

Lastly, while next-generation sequencing technologies are currently out of reach for most forensic laboratories, advances in medical genetics are leading to rapid decreases in expense and increases in efficiency in these technologies. Exploring these technologies, and determining which ones are amenable the nuances of forensic evidence samples, provides a foundation for other researchers.

Chapter 2: Literature Review

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 these SNPs can also be valuable to the field of forensic science (Butler 2005, Brookes 1999). A composite profile from a battery of ancestry and phenotype informative SNPs can provide an estimate of ancestry and physical morphology, with a significant advantage over eyewitness testimony in that these data can be statistically supported (Kayser 2011). Such a tool could help prioritize suspect processing, corroborate witness testimony, and determine the relevance of a piece of evidence to a crime (Butler 2007, Butler 2008). Additionally, adding existing information such as mtDNA or Y chromosome haplotype, and/or autosomal STR genotypes could boost ancestry prediction (Nelson 2007, Brion 2005, Phillips 2012), and combining all forensically-relevant loci into one multimarker multiplex would dramatically improve casework processing efficiency.

Ancestry Informative SNPs

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 have been used to infer the ancestral origin of a sample (Nelson 2007, Brion 2005); however, as both genomes are inherited without recombination, each can only provide information about either maternal or paternal

lineages. Particularly for admixed individuals (*e.g.* from more than one distinct ancestral population, such as African American or Mexican American), this method would provide an incomplete picture of ancestry. Although autosomal ancestry informative SNPs are subject to greater variation due to recombination, there are several autosomal SNPs where markedly different population frequencies occur due to an adaptation to a particular environment or other evolutionary forces.

An example of a powerful ancestry informative SNP is rs2814778, which is present in the *DARC* gene, part of the Duffy blood group. One Duffy phenotype (A-B-, homozygous C) lacks the receptor for *P. vivax*, which leads to a reduced susceptibility to malaria (Miller 1976). This represents an adaptation to presence of malaria and it occurs predominantly in Sub-Saharan African populations, as seen in Figure 2. Using markers such as rs2814778, two recent studies have shown high ancestral group classification probabilities with panels of only 10 or 34 autosomal SNPs (Phillips 2007, Lao 2006). With NIJ support, another laboratory has recently presented significant data (4,781 individuals) genotyped with a 128 SNP panel, which is particularly useful in estimating admixture (Kidd 2011).

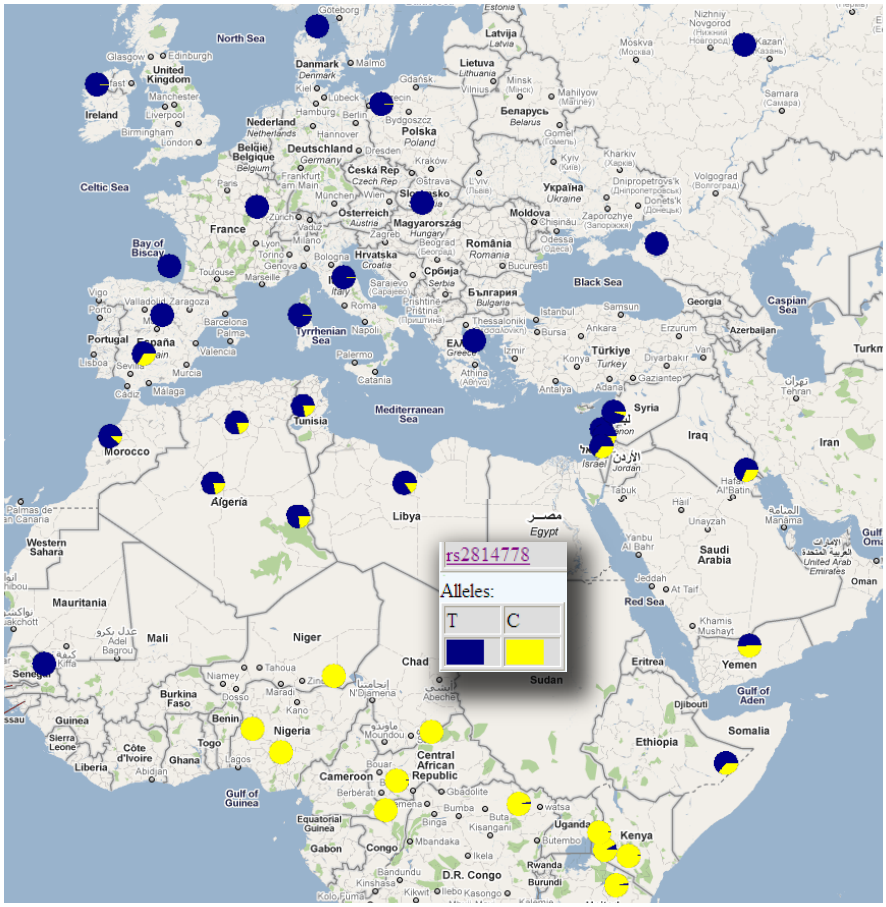


Figure 2. Data from www.alfred.med.yale.edu showing distribution of alleles at rs2814778. The C allele, an adaptation to malaria, is nearly monomorphic in sub-Saharan Africa, while the T allele is dominant to the north in the absence of selective pressure.

Phenotype Informative SNPs

The primary phenotype informative SNPs with forensic predictive value are those associated with pigmentation. Just as selective pressures create the ideal ancestry informative SNPs (as seen in the preceding example), these forces are also responsible for the variation in pigmentation found among humans. Significant hypotheses regarding the advantages of dark pigmentation near the equator and lighter pigmentation away from the equator have been proposed (Jablonski 2004).

A longstanding popular theory on the evolution of skin pigmentation has been the vitamin D hypothesis (Loomis 1967). Holick (1995) postulated that early tetrapods

required this vitamin to maximize calcium use in maintaining a rigid skeleton, and that the vitamin D had to either be synthesized by the organism or ingested in a sufficiently vitamin D-rich diet. Vitamin D_3 is needed for proper bone formation, and precursors of this vitamin are formed in the body upon exposure to UV radiation from the sun (Wharton 2003). In equatorial regions, there is sufficient UV radiation throughout the year to allow adequate synthesis of Vitamin D_3 precursors, even in darkly pigmented individuals (Jablonski 2004). Zones farther from the equator experience a corresponding increase in time of the year (depending on the rotation of the earth), where less than adequate UV radiation exists to produce sufficient levels of vitamin D_3; however, lighter skin allows more UV penetration which works to overcome the decrease in UV radiation (Jablonski 2004). A deficiency in Vitamin D_3 leads to a bone disease known as rickets, which is characterized by the failure of developing bones to mineralize, due to poor absorption of calcium and phosphate (Wharton 2003). This deficiency manifests as bowing of the legs, delay in fontanel closure, and female narrowing of pelvic bones, the last of which leads to high levels of death in childbirth (Wharton 2003). The significant impact of this disease on development and reproduction make it an ideal candidate for selection.

A relatively newer hypothesis in skin pigmentation research points to dark pigmentation protecting against the photodegradation of folate in regions of high UV radiation (Jablonski 2000). In recent decades, folate (folic acid, a B vitamin) has been shown to significantly impact cell division during pregnancy, where a lack of folate is associated with early pregnancy loss (Suh 2001). Darker skin pigmentation would be advantageous because it would prevent UV radiation from penetrating to the highly

vascularized dermis, where folate is present in the bloodstream (Jablonski 2004). Lighter skin pigmentation would be problematic, particularly with increasing proximity to the equator. Interestingly, it is noted that in areas where there is significant seasonal change in UV radiation (on the latitude of the Mediterranean Sea) populations are most able to develop facultative pigmentation (suntan), which provides some protection (Jablonski 2004).

These two hypotheses can be viewed as a merged model of ideal pigmentation balance: depending on the level of UV exposure, a population will evolve a skin pigmentation that allows sufficient vitamin D_3 synthesis while protecting against folate degradation, to maximize overall fitness.

Phenotype informative SNPs having predictive value for hair, eye, and/or skin pigmentation fundamentally depend on the amount, type, and distribution of melanin in these tissues. Both the amount and type of melanin and the shape and distribution of melanosomes contribute to overall pigmentation (Parra 2004). A recent NIJ funded project on polymorphisms associated with 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 variation across populations (Brilliant 2008). Two other studies point to one particular SNP (rs12913832) in the *HERC2* gene which is predictive of light eye color: 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, Sturm 2008).

This is consistent with a recent study that showed that the *HERC2* region encompassing rs12913832 functions as an enhancer, regulating transcription of *OCA2*,

which encodes for the trans-melanosomal membrane protein "P" (Tully 2007). In darkly pigmented human melanocytes, transcription factors HLTF, LEF1, and MITF were found 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. In lightly pigmented melanocytes carrying the rs12913832 C allele, chromatin-loop formation, transcription factor recruitment, and *OCA2* expression were all reduced (Visser 2012). Figure 3 shows the distribution of rs12913832 alleles, with the light eye color C allele (represented in the figure as "G" due to opposing strand being genotyped) absent in sub-Saharan Africa and becoming increasingly frequent to the north.

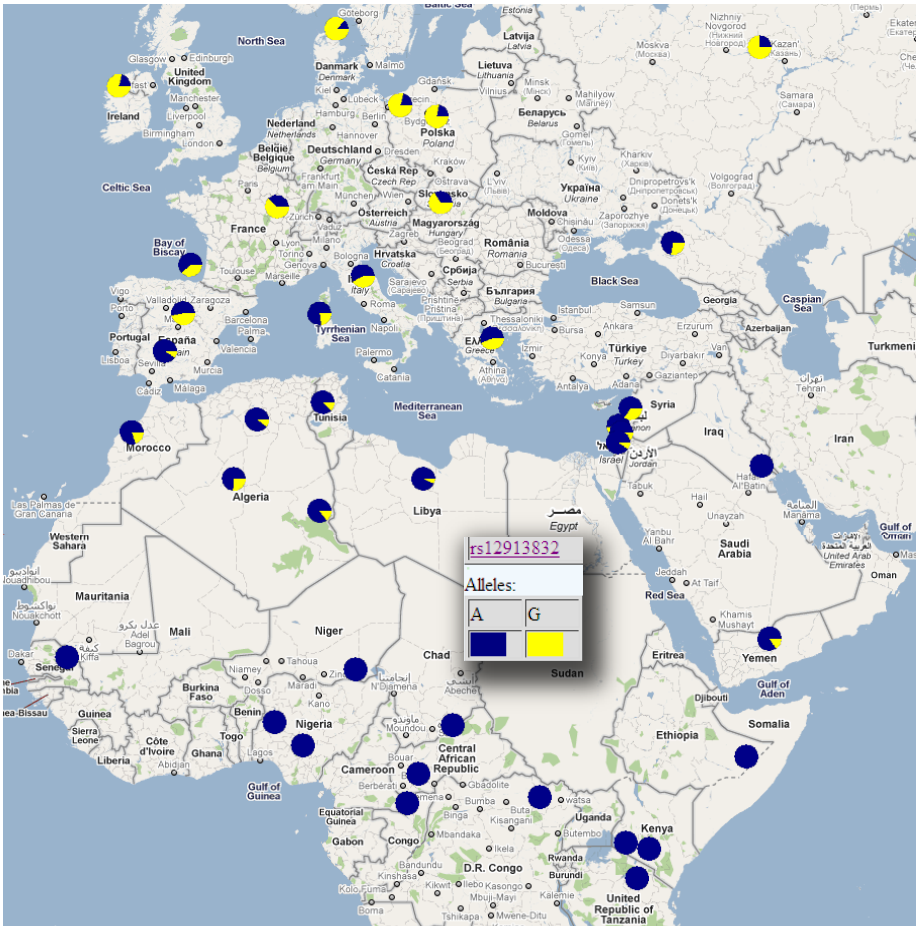


Figure 3. Data from www.alfred.med.yale.edu showing distribution of alleles at rs12913832. The A allele (associated with darker pigmentation) is monomorphic in sub-Saharan Africa, while the G allele (associated with lighter pigmentation) becomes increasingly prevalent in central to northern Europe.

Other Forensic Markers

The field of forensic science relies upon mtDNA to identify missing persons, locate maternal relatives, identify victims in mass disasters, and exclude individuals as contributors of forensic samples. Two hypervariable regions of the mtDNA genome are sequenced for forensic analysis (total of ~600bp), often being amplified with overlapping small primer sets (approximately 100-200 bases in each amplicon) to maximize degraded evidence, creating a laborious process with the currently used Sanger sequencing methodology. Methods have been described that allow for the genotyping of haplotype-defining SNPs within these hypervariable regions and inferring ancestry by way of maternal lineage (Nelson 2007).

Y chromosome STR loci are very useful in forensic kinship analyses involving the paternal lineage or when a forensic sample contains a male-female mixture where the female component outweighs the male (such as body swabs from sexual assault cases). YSTR data can also be used to inform an ancestry prediction by way of paternal lineage with the use of a haplotype predictor (Athey 2005). Several male specific SNPs have also been well-characterized on the Y chromosome and can also aid in ancestry determination (Vallone 2004).

Lastly, although forensic autosomal STR loci were specifically chosen for their ability to discriminate among individuals across populations, these STR allele frequencies do vary among populations. Because of this, forensic casework requires reporting the frequency of an STR profile in multiple populations, and frequency data sets exist for all major populations. A comparison of the 13 FBI CODIS core STR loci to a panel of 39 ancestry informative SNPs shows the STR loci are useful for admixture

analysis but less precise than a combination of STR and SNP loci (Barnholtz-Sloan 2005). A recent publication not only suggests inclusion of STR data into a SNP-based ancestry determination, but also provides a web-based statistical calculation tool allowing for ancestry prediction using both marker types (Phillips 2012).

Next-Generation Sequencing in Forensics

Next-generation sequencing methods have been rapidly evolving over the past five years as the medical genetics community concurrently moves away from Sanger-based sequencing (Metzker 2010) (the technology currently used in forensic mtDNA testing). The future of this technology holds the ability of a doctor to provide instant genotyping data that can aid in disease diagnosis and personalized treatment options.

Forensic science can greatly benefit from these advances in technology by generating more information from a smaller amount of sample as compared to the assays in place today. Recent studies show the benefits of combining different marker systems to maximize information, for example combining STR and SNP data to aid in identification of skeletal remains (Fondevila 2008) and using SNP data to determine if an individual is present in a complex DNA mixture (Homer 2008). Additionally, sequencing STRs and YSTRs could provide valuable information on sequence differences between individuals, which could help when mutations are suspected, or when a common YSTR profile is obtained. The exponential increase in information that could be obtained from a forensic sample, along with studies such as those that link mental disorders to DNA polymorphisms, raise significant ethical concerns that have yet to be addressed (Asplen 2013, Kayser 2009, Karayiorgou 2010).

One next-generation technology that holds particular promise for forensics is the

Pacific Biosciences real-time sequencer. While many next generation methods are limited to a small size amplicon (*i.e.* <100bp), which would not be able to encompass an STR repeat region, Pacific Biosciences has the flexibility to sequence fragments anywhere from 40 to 25,000bp (Travers 2010). Additionally this technology does not require amplification. In the case of a multiplex that includes nuclear and mtDNA markers, this could help overcome the copy number difference. However, including an amplification step would enrich the sample and increase sensitivity.

The true benefit in designing NGS methods for forensics would be the possibility of a multimarker multiplex, wherein STR, YSTR, mtDNA and the various types of SNPs could be analyzed concurrently to maximize the information obtained from one sample in one assay.

Chapter 3: Candidate SNP Selection, Sample Collection, and SNP Genotyping

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 (Bouakaze 2009, Branicki 2009, Brilliant 2008, Duffy 2007, Halder 2008, Han 2008, Iida 2009, Kidd 2008, Kosoy 2009, Lao 2006, Mengel-From 2010, Shekar 2008, Stokowski 2007, Sturm 2009, Sulem 2007). 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.

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 Figures 1 and 2), volunteers completed a comprehensive questionnaire (Appendix Figure 3) 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). The spectrophotometer software (Konica Minolta CM-SA) automatically calculates a melanin index, which is an integral of measurements across the wavelengths of normal human pigmentation (Stamatas 2004). See Figure 4 for relative melanin index measurements obtained; generally the face measurements were significantly darker than the arm due to increased UV exposure. Due to the desire to avoid facultative pigmentation (suntan), sample collection was suspended during the summer months.

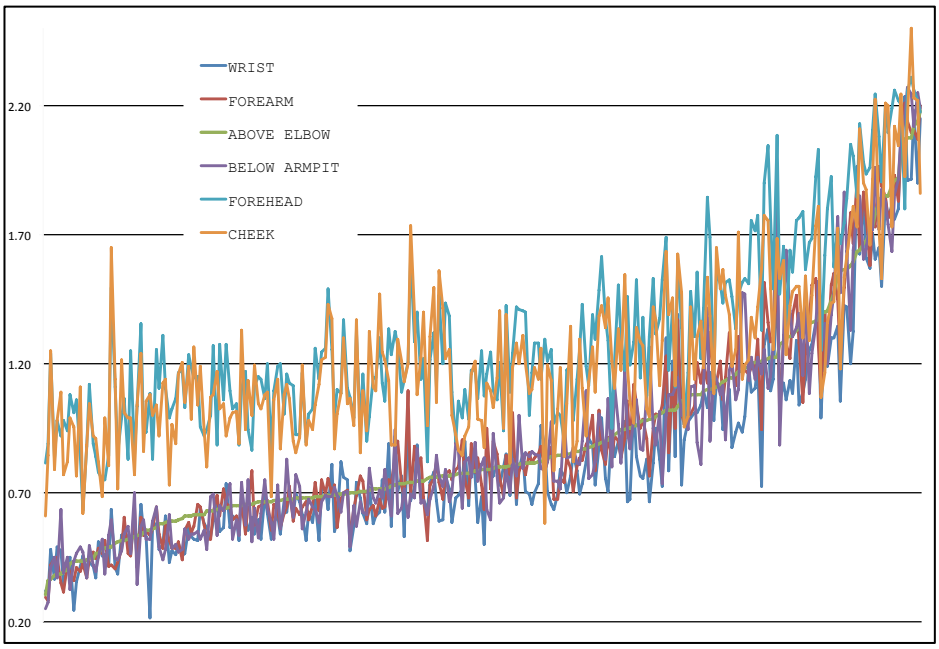


Figure 4. Skin melanin index measurements collected from volunteers, sorted from low to high based on "above elbow" values. The face measurements are consistently higher than arm measurements due to increased UV exposure over time.

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 checklist (Appendix Figure 4) 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 (Appendix Figure 5). 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 American 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, and 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 5 for complete breakdown of samples/sources/information.

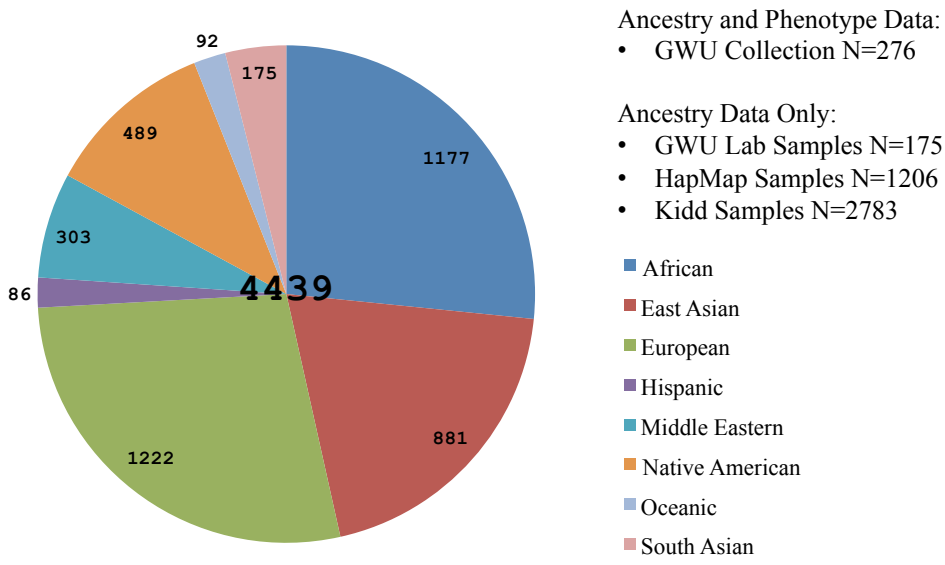


Figure 5. Sample breakdown by ethnicity, and sample sources

SNP Genotyping

The Single Base Primer Extension (SBE) technique (Sokolov 1990, Pastinen 1997, Syvänen 1999) allows for the simultaneous typing from 1 to over 30 SNPs (Phillips 2007). Once the assay is optimized, it allows one to obtain robust results over a broad range of both quantity and quality of genomic DNA template, utilizing equipment already available in Forensic DNA laboratories.

The SBE 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. After the multiplex PCR amplification is performed the reaction product is purified to eliminate unincorporated PCR primers and dNTPs. Using the purified PCR product as template, a complimentary 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, then the appropriate ddNTP is incorporated at the SNP site (Figure 6). Following the SBE

reaction, samples are loaded onto a capillary electrophoresis instrument.

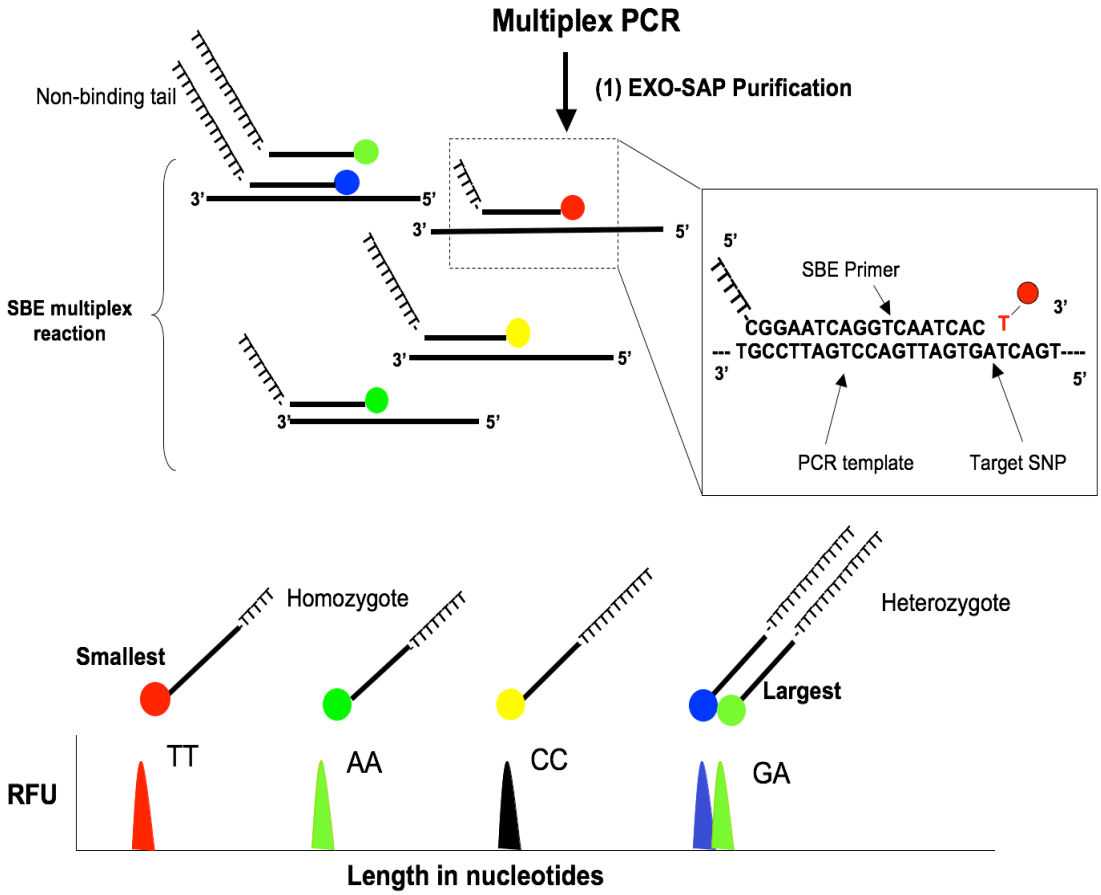


Figure 6. Schematic representation of the SBE assay. In this example the targets are 4 diploid loci of which the first three (left to right) and homozygous and the forth is a heterozygote. Multiplexing of a SBE assay is accomplished by adding a non-binding tail sequence to the 5' end of the SBE primer. 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. Requirements of the assay are that the amplicons must flank the desired SNP site and retain the SBE primer annealing site.

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 after troubleshooting (including different primer pairings, increased primer concentrations, and different multiplex combinations) and were eliminated during this phase. Figure 7 shows

electropherograms of a sample analyzed with five of the multiplexes developed to genotype the candidate SNPs.

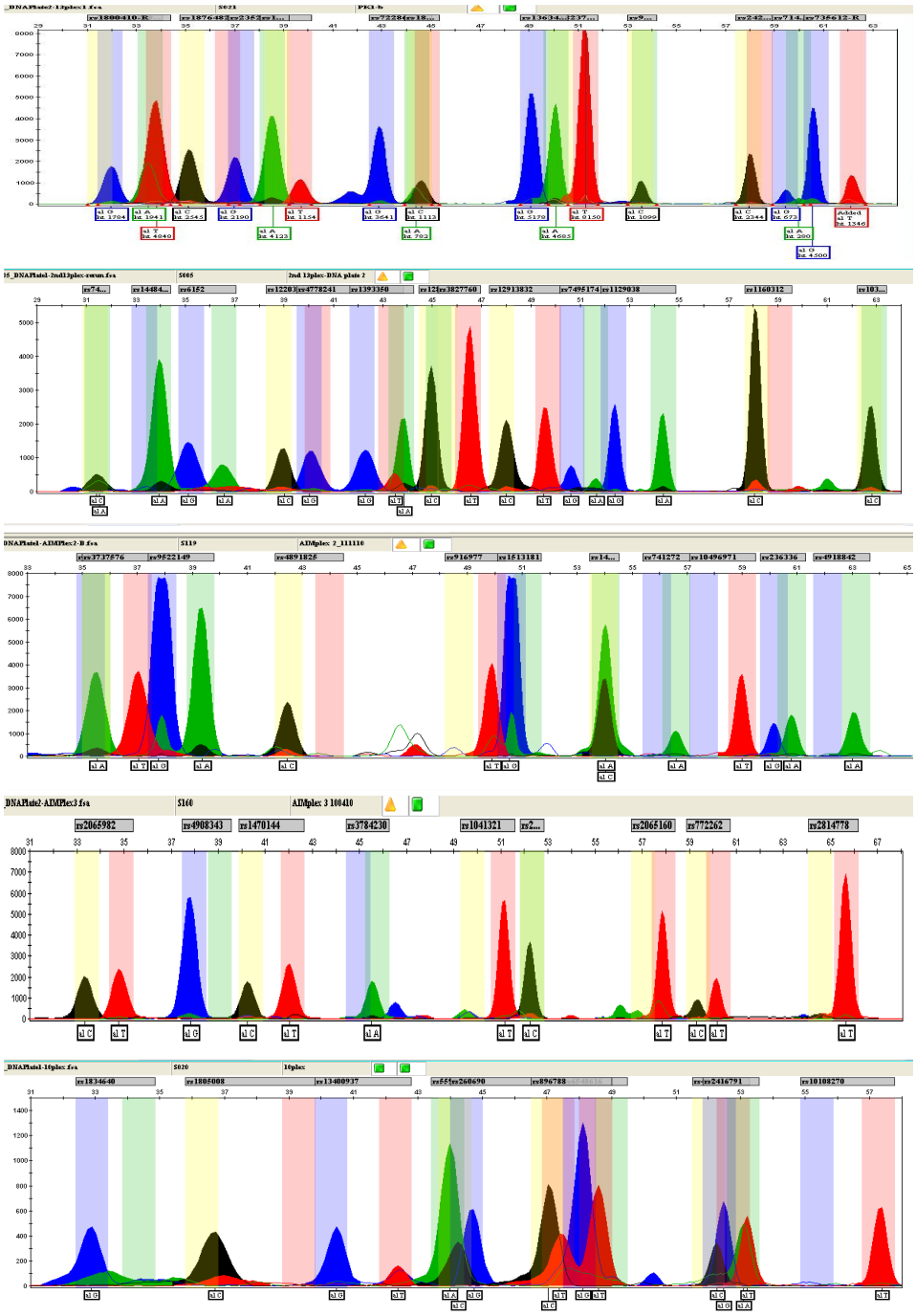


Figure 7. Examples of five SNP multiplexes that were used to screen volunteer samples.

Chapter 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 8). 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.

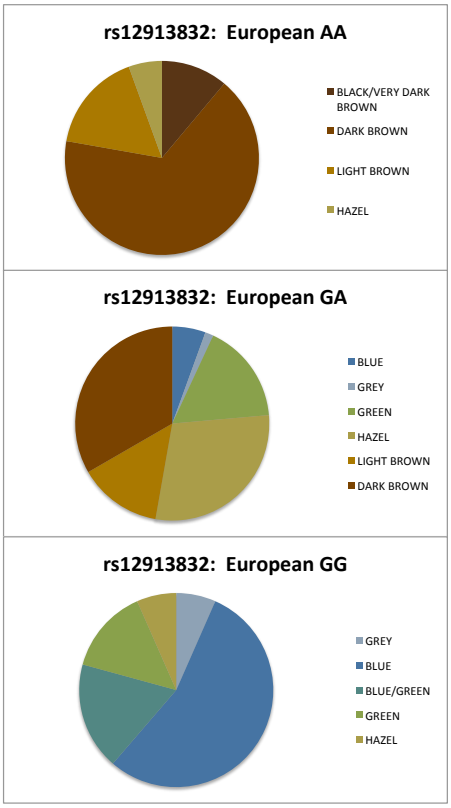


Figure 8a. rs12913832: Of 196 European Americans with phenotype data available, homozygous A individuals (9%) have brown eyes; whereas homozygous G individuals (54%) have light colored eyes. The remaining individuals (37%) are heterozygous and present both phenotypes.

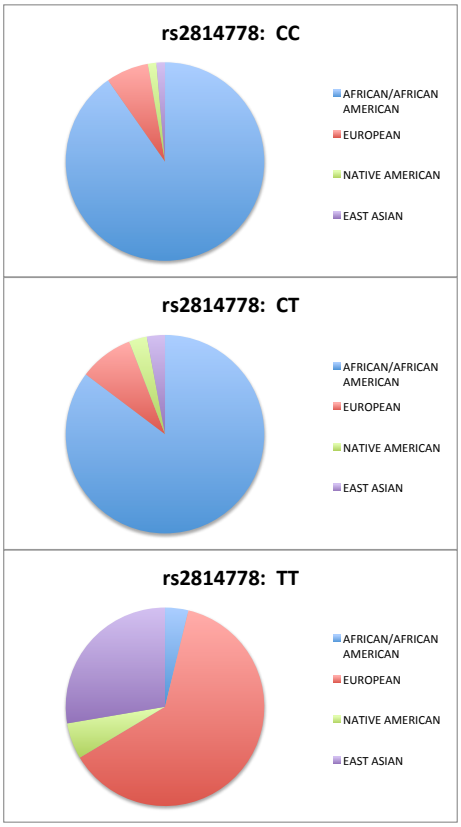


Figure 8b. rs2814778: the C allele represents an adaptation to malaria, thus the presence of at least one C allele is indicative of sub-Saharan African ancestry or admixture. Out of 395 individuals tested in the four populations of interest, 90% of homozygous C individuals were African American or African, 85% of heterozygous C/T individuals were African American, and only 4% of homozygous T individuals were African American.

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 European/European Americans. 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).

Materials and Methods - Ancestry

X^2 Analysis: This analysis evaluated the 99 remaining SNPs in relation to ancestry for the four populations of interest using a chi-squared test. This calculation compares the

observed allele frequencies to those expected under Hardy-Weinberg Equilibrium within each category, in this case the four populations. The resulting p-value represents the probability that deviation of the observed frequencies from those expected is due to chance alone. Using a p-value of 0.01 means deviation of observed from expected by chance could happen 1% of the time; therefore, the lower the p-value, the greater the significance. 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.

All available data for the 43 ancestry SNPs were divided into eight categories for PCA: African, East Asian, European, Hispanic, Middle Eastern, Native American, Oceanic, and South Asian. 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, pairwise F_{ST} analysis was performed (pairs included African/African American—European/European American, African/African American—East Asian, East Asian—European/European American, East Asian—Native American, and Native American—European/European American), which compares allele frequencies and levels of heterozygosity in the subpopulation to the total of the two populations. The resulting number is the difference in levels of heterozygosity, where a higher number indicating greater diverging power of the SNP. Performing this analysis in a pairwise fashion allows for determining the SNPs that best differentiate any two populations, or one population from multiple other 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.

Materials and Methods – Pigmentation in European Americans

X^2 Analysis: This analysis evaluated the 99 SNPs in relation to the specific phenotypes of eye, skin and hair color in samples of European descent (N=196) using a X^2 analysis evaluated with a conservative p-value. Table 1 shows the categorization of phenotypes for this analysis.

Table 1. Phenotype categorization in Europeans. Melanin index was measured on inner arm, above elbow.

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

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 the European American population only. The latter analysis was 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 American ancestry for which phenotype data was available (N=196) were categorized as before for hair and eye color, and with an adjusted scale for melanin indices (light = minimum-0.59, medium = 0.6-0.89, and dark = 0.9-maximum). PCA was performed on the data 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 v. 2.1 was used to generate the statistically most likely haplotypes from the genotype data (Stephens 2003) and to evaluate the likelihood of recombination (Li 2003 and Crawford 2004). All samples with genotype data in these gene regions were divided by ethnicity: European/European American, 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 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 was unlikely were then evaluated to determine which SNPs are definitive of the haplotype and/or appear to be associated with pigmentation.

Results and Discussion - Ancestry

X^2: As seen in Appendix Table 1, 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: SNPs with the highest factor loading for ancestry are indicated in Appendix Table 1. A subset of 25 SNPs with high factor loading was selected from the 43 ancestry SNPs. The ability of this subset to diverge the populations of interest was evaluated with *STRUCTURE 2.3.1* software analysis, a population genetics and anthropology software package based on Bayesian statistics, developed to analyze the genetic composition of individuals and populations. Figure 9 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, African American, Asian and Hispanic) can still be well-differentiated with the subset of 25 AIMs.

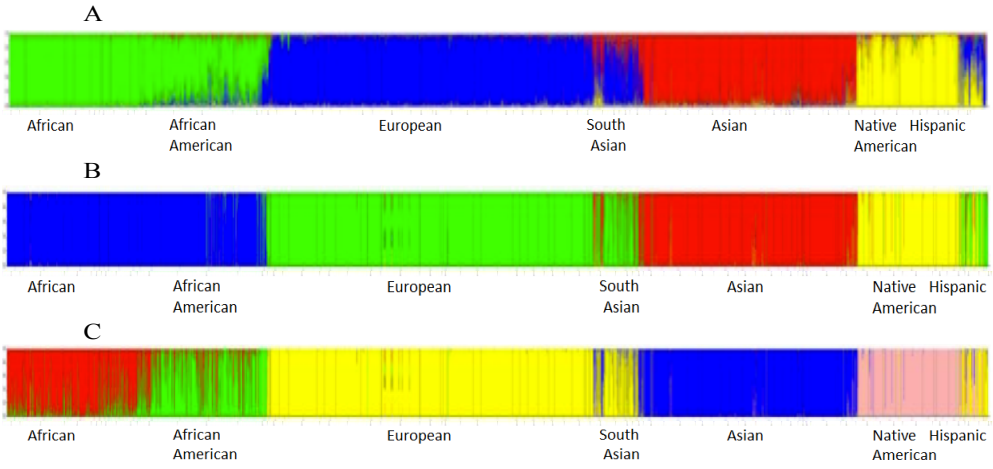


Figure 9. Structure plots (A) 43 ancestry informative markers, K=4, (B) 25 ancestry informative markers, K=4, and (C) 25 ancestry informative markers, K=5 analyzed on 4440 individuals from multiple populations. The 25 A ancestry informative markers were selected from the 43 with Principal Component Analysis (PCA *STATISTICA Data Miner* software).

Snipper: This analysis produced a divergence ranking value for each SNP (1 being the most divergent SNP and 99 being the least divergent), seen in Appendix Table 1. 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/European American were all over 90%; however, the rate was lower for Native Americans (81%). There were three misclassified Native Americans, all classified as European. 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}: In Appendix Table 1, 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 European/European Americans 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.

Results and Discussion - Pigmentation

X^2: This analysis showed a significant relationship for European American eye color with 11 SNPs, European American skin color with 2 SNPs and European American hair color with 11 SNP at the α=0.01 significance level. Many additional SNPs showed weaker evidence of a relationship with a p-value between 0.01 and 0.1. While many SNPs appeared associated with only one of the phenotype, several showed significance for multiple phenotypes. Specifically, rs12913832 in the *HERC2* gene region (previously described) and rs1129038 in the *SLC45A2* gene region were significant for all three phenotypes at α=0.1.

PCA: 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 10), the black hair color is separated the farthest from all other hair colors but when analyzing European Americans only (Figure 11), 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 European Americans, the SNPs that account for the most variability in the data were selected, and the analysis was repeated with a subset of 20 such SNPs (Figure 11). These results indicate this subset of 20 SNPs is similarly effective at differentiating the groups as 53 SNPs.

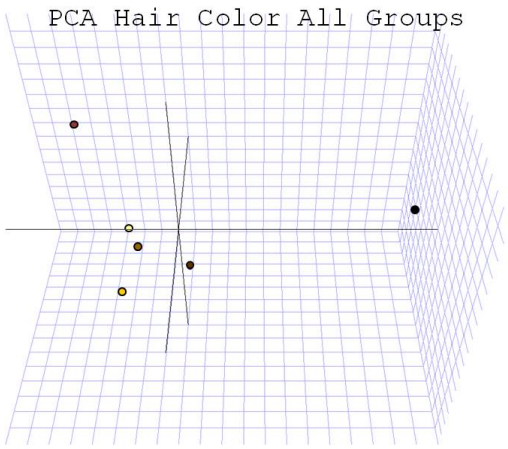
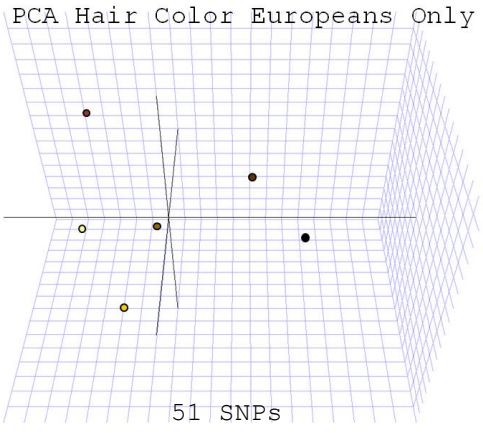
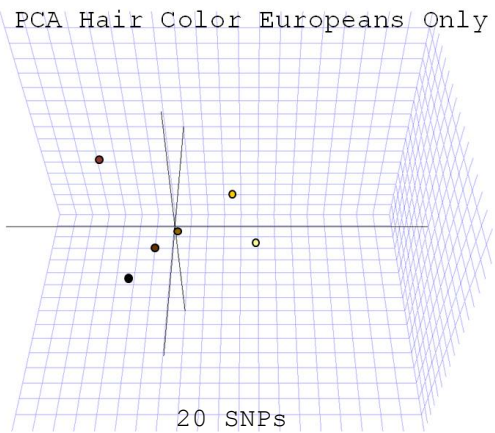


Figure 10. Tridimensional PCA plot of the 53 phenotype SNPs analyzed on all individuals with known phenotype. Individuals were divided in 6 groups based on their hair color represented by the color of the dot: black, dark brown, light brown, dark blonde, light blonde, red/auburn.

Figure 11. (below left) Tridimensional PCA plot of the phenotype SNPs 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 pigmentation SNPs analyzed only on individuals with known phenotype and of European descent.





PHASE - 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 literature (Li 2003) 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 2.

Table 2. Recombination analysis of the 19 SNPs genotyped in the *OCA2/HERC2* region, values in bold indicate recombination is likely.

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

Based on these analyses of the *OCA2/HERC2* SNPs, recombination is likely between SNPs 5 and 6 in both the European/European American and East 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 - 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 12, where 12 haplotypes are found among the European American individuals, three are present in the African/African American individuals, and four are found in the East Asian individuals (analysis performed only on 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 European Americans (Figure 13). Therefore, the three SNPs that define these three haplotypes (rs1805009, rs1805008, and rs1805007, respectively) are good candidates for the final assay.

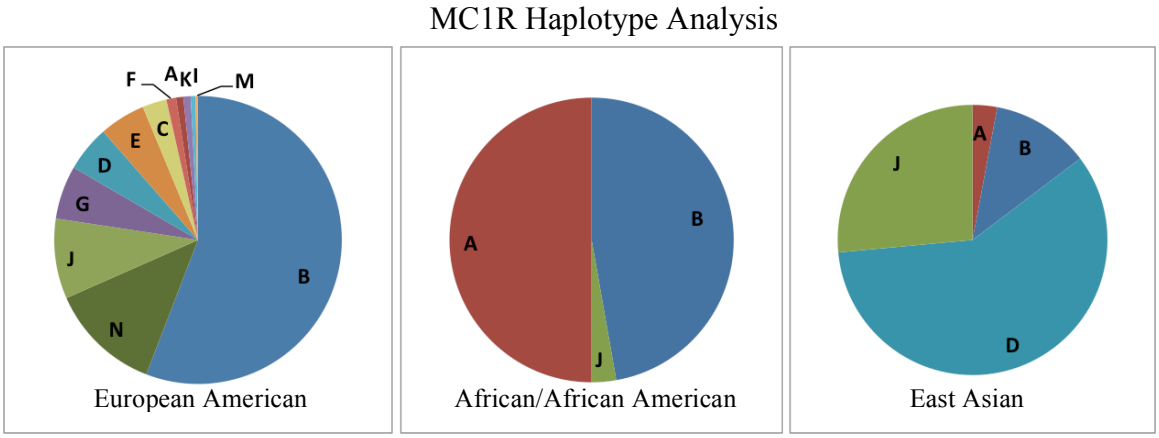


Figure 12. *MC1R* Haplotype distribution in the different populations that were tested, each chart contains only the samples for which phenotype information was available. The number of each haplotype found was: European American- B(206), N(50), J(34), G(23), D(19), E(19), C(10), F(4), A(3), K(3), I(2), and M(1); African/African American- A(18), B(17), and J(1); and East Asian- A(1), B(4), D(20), and J(9).

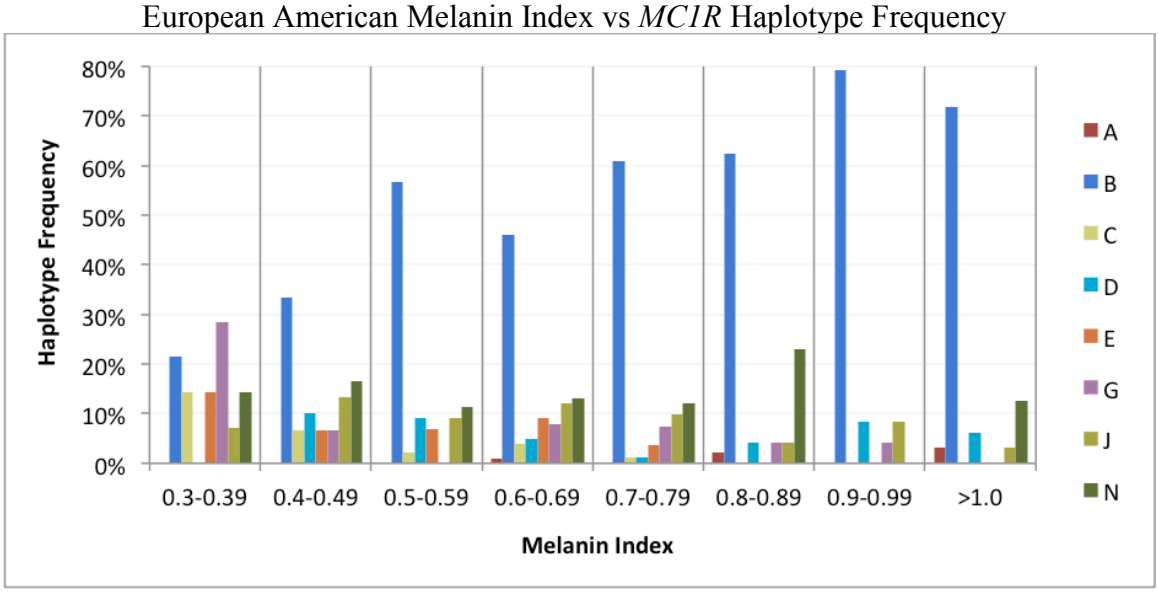


Figure 13. The graph compares melanin index (measured above elbow) on x-axis to frequency of haplotype on y-axis, among European Americans. Haplotype B increases in frequency and diversity decreases as melanin index increases. The frequency of haplotypes C, E, and G decrease steadily as melanin index increases.

The results for *OCA2*/*HERC2* haplotype distribution in linked regions were not nearly as informative. Comparing results for European/European American, 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 14). 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/European American 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 15). This is not surprising, since literature associates this gene region more strongly with eye color than with skin pigmentation (Tully 2007, Kayser 2008, Sturm 2008. Visser 2012).

OCA2/HERC2 Haplotype Analysis

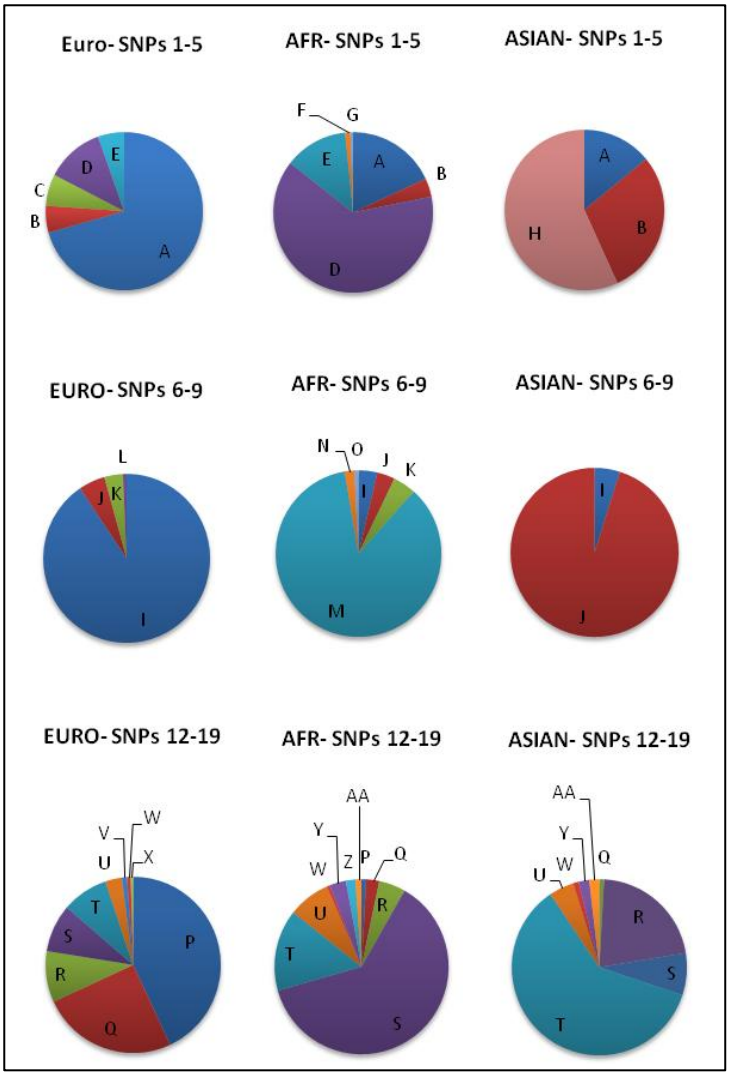


Figure 14. OCA2/HERC2 haplotypes in the three determined linked regions for European/European Americans (EURO), African/African Americans (AFR) and East Asians (ASIAN). Compared to the haplotype distribution for the MC1R SNPs, these gene regions show more similar number and distribution of haplotypes between the three populations. The large size of this gene region makes chance mutation more likely.

The number of each haplotype found was: EURO SNPs 1-5- A(1912), B(225), C(146), D(95), E(17); AFR SNPs 1-5- A(416), B(85), D(1462), E(295), F(24), G(13); ASIAN SNPs 1-5- A(242), B(502), H(978); EURO SNPs 6-9- I(2170), J(122), K(86), L(17); AFR SNPs 6-9- I(84), J(80), K(103), M(1966), N(43), O(19); ASIAN SNPs 6-9- I(84), J(1638); EURO SNPs 12-19- P(1026), Q(597), R(224), S(209), T(203), U(73), V(23), W(16), X(11); AFR SNPs 12-19- P(20), Q(56), R(115), S(1442), T(348), U(181), W(13), Y(71), Z(41), AA(26); ASIAN SNPs 12-19- Q(15), R(370), S(133), T(1040), U(79), W(18), Y(33), AA(32).

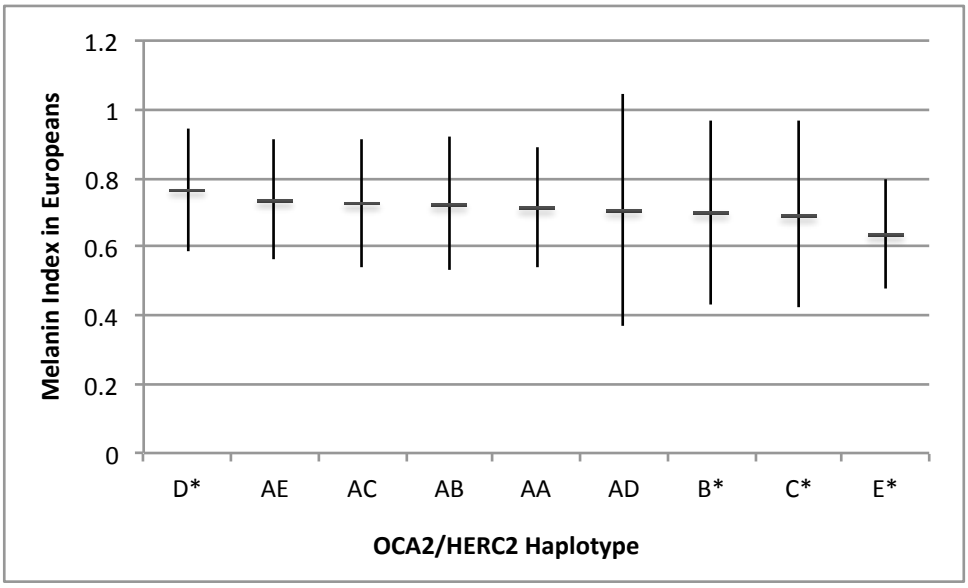


Figure 15. Example of OCA2/HERC2 haplotypes (x-axis) compared to skin melanin content above elbow (y-axis) in European Americans (SNPs 1-5) with phenotype data available. Two letters indicate an individual's two predicted haplotypes, whereas one letter and "*" indicates one predicted haplotype combined with any other haplotype. Based on the average (horizontal line) +/- one standard deviation (vertical line), no haplotype shows a clear relationship to skin melanin content. The number of each haplotype combination was D*(11), AE(16), AC(16), AB(14), AA(89), AD(31), B*(4), C*(8) E*(4).

Final Selection

By cross-referencing each of these analyses and paying particular attention to SNPs for which published prediction models already exist (Branicki 2011, Walsh 2011a), 50 SNPs were selected from the initial 103 that are expected to be most predictive of ancestry, specific phenotype traits, or both (see Appendix Tables 1 and 2 for the list of 50 SNPs and the results of each statistical approach). The set of 50 SNPs includes 19 AIMs and 31 pigmentation PIMs, 13 of which also have a strong association to ancestry.

Chapter 5: Development / Optimization of 50-SNP Assay

The objective of the optimization phase was to type the 50 SNPs 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 goal of this phase was to be able to type all the selected SNPs with no more that 1 ng of DNA.

Materials and Methods

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 3 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), cycling parameters, and reaction volumes (10 µl vs. 25 µl). The optimized reaction was compared to the 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. 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 2.5 pg to 10 ng of input DNA, using a sample quantified via UV-Vis spectrophotometry (NanoDrop 2000, Thermo Scientific).

Additional testing was performed on seven highly heterozygous samples, also quantified with UV-Vis spectrophotometry, at 100 pg, 150 pg and 200 pg of input DNA. The multiplexes were evaluated for robustness with various types of mock forensic samples, all of which had previously yielded STR profiles with 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 4); however, these will require adjustment based on the capillary electrophoresis polymer used and other laboratory-specific conditions.

Results and Discussion

The best peak balance with the least background was found in a 25 µL amplification reaction volume. Evaluation of PCR reaction mixture components showed that increasing DNA polymerase and dNTP input improved results, while the Identifiler Plus® reaction mix (proprietary concentrations of dNTPs, DNA polymerase and buffer) performed poorly in comparison. The multiplexes performed best with a high PCR cycle number, 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 8 ul reaction best balanced sensitivity and background. The optimal SBE parameters were 28 cycles with a 55°C annealing temperature. See Appendix Figure 6 for example electropherograms.

Recommended Protocol: PCR reaction components in a 25 µL reaction include: 1X PCR Buffer Gold® (Applied Biosystems), 2.5 mM $MgCl_2$ (Applied Biosystems), 0.22 mM

dNTPs (Roche Diagnostics, Indianapolis, IN), 0.0568 mg/mL BSA (Fisher Scientific, Waltham, MA), 4.375 U AmpliTaq Gold DNA Polymerase® (Applied Biosystems), 2 µL multiplex-specific PCR primer mix (Integrated DNA Technologies, Coralville, IA; see Appendix Table 3 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 2 µL 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 0.25 µL H20, in a final volume of 3 µL. 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 1 µL SNaPshot Reaction Mix® (Applied Biosystems), 1 µL multiplex-specific SBE primer mix (Integrated DNA Technologies, see Appendix Table 3 for primer sequences and reaction concentration), 3 µL H_20, and 3 µL 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, 10 µl of LIZ 120 size standard (Applied Biosystems) was added to 400 µl of Hi-Di formamide (Applied Biosystems), and 1 µl of sample was added to 10 µl of the Formamide/ILS mixture. Samples were electrophoresed on the 3130 Genetic Analyzer (Applied Biosystems), using a 36 cm capillary (Applied Biosystems, refurbished from Gel Company Inc.) and POP-7 polymer (Applied Biosystems), with injection parameters of 1.2kV for 16 seconds.

Sensitivity: Initial sensitivity testing detected all 50 SNPs at 100 pg 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 mistyped as heterozygotes at or below 200 pg of input DNA (see Table 3 for evaluation of nine SNPs with relatively low peak heights; SNPs not included in this table were correctly typed to 100 pg as heterozygotes). Careful evaluation of results and controls is required at or below this level. To minimize stochastic effects, recommended input range is 0.5-2 ng DNA per multiplex; however, the goal of genotyping all 50 SNPs with 1 ng of DNA was met, as concordant results would generally be expected with inputs totaling 1 ng.

Table 3. 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

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

Table 3 (continued).

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 16 for electropherograms showing loci and multiplex performance on a forensic sample.

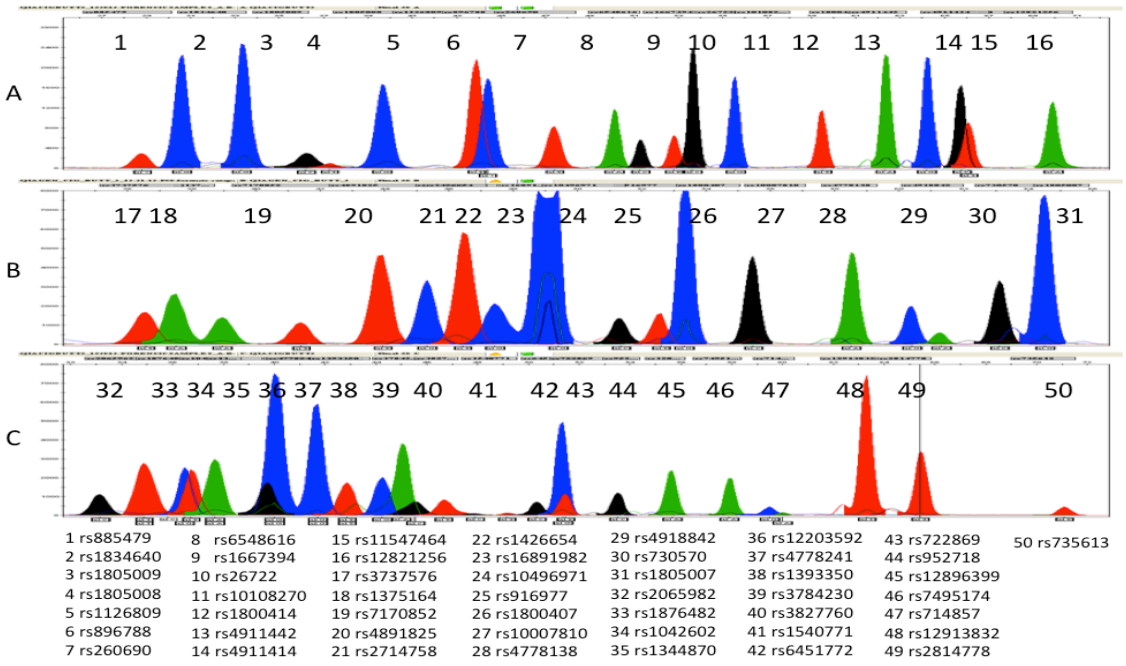


Figure 16. Electropherograms results of the 50 SNP assay (three multiplexes, loci distributed as shown); profile obtained from a cigarette butt.

Chapter 6: Development / Evaluation of Ancestry Models

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.

Materials and Methods

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 as traditional statistical approaches assume loci are unlinked. LD 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 LD (performed for *MC1R* and *OCA2/HERC2* SNPs) addressed in Chapter 4.

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 LD. Thirty-six of the remaining 44 SNPs were included in the linkage disequilibrium analysis (the remaining eight were not present in the HapMap data set). The LD analysis was reviewed conservatively, such that only one SNP from each gene region was selected (except for OCA2/HERC2, where multiple tests showed recombination between the two SNPs selected). This reduced the number of SNPs to be included in the biogeographic ancestry prediction to 32 (see Appendix Table 5 for this subset of SNPs and Appendix Table 6 for complete results of

the linkage disequilibrium evaluation).

Excluded SNPs which are clearly linked to included SNPs could be combined with the included SNPs to form haplotypes, then the haplotype frequencies could be used in ancestry and/or pigmentation models. A small-scale version of this approach has been recently published (Ruiz 2013), where two *HERC2* SNPs are combined (rs12913832 and rs1129038) and the diplotype frequency is used in place of allele frequencies. These two SNPs show nearly identical allele frequencies in the training set used herein, development of which is described below (a slight difference in allele frequencies is seen due to missing data for one allele in one sample); therefore, additional SNP pairs were evaluated to determine the effects of this approach on the ancestry and pigmentation models.

By evaluating the SNPs in the same gene regions based on their ability to diverge populations (see Chapter 4 and Appendix Table 1), two SNPs were selected for diplotype analysis and inclusion in the models. One SNP that would be included in all models, rs12913832, was chosen, along with rs916977. The pairwise F_{ST} shows differences between these two SNPs (which would yield varied haplotypes), and despite their distance on the chromosome (>140,000 bases), both PHASE and WGA viewer analysis show them to be linked (Appendix Table 6).

Further, using the training set (described on the following page) and the European American samples with corresponding eye color information (N=190), these two SNPs were evaluated for linkage disequilibrium with r2 and D' using the allocation method (Andersson 1985), and LD is still indicated, see Table 4 for results.

Table 4. Results of linkage disequilibrium testing for rs12913832 and rs916977 using the training set and European American samples with eye color information. Calculations could not be performed for the European/European American and East Asian training set samples due to one SNP being monomorphic in each population. The composition of the training set samples (selected for maximal divergence) is not ideal for LD analysis.

	European American Eye Color Samples	European/European American	African American	Hispanic/Native American	East Asian
D'	1	n/a	1	1	n/a
r^2	0.488	n/a	0.069	0.128	n/a

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 2010)), 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. This approach was used in order to create the most divergence between population groups, which would result in optimally performing models.

The training set was composed of 266 European/European Americans, 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 (See Appendix Table 7 for training set samples and allele frequencies).

Test Set: The samples tested under each ancestry model were composed of 31 European Americans, 32 African Americans, 32 Hispanic Americans and 32 East Asians. The majority of these test samples (European American, African American and Hispanic American) 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 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 prediction probabilities for the 32 East Asian samples, because these individuals helped inform SNP selection and 2) deflation of prediction probabilities 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.

32 SNP RMP/LR Ancestry Prediction Model: The allele frequencies were used to calculate the random match probability (RMP) in all four populations. See Appendix Table 7 for the training set allele frequencies.

As described by Evett (1992) and Brenner (1997) using forensic STR profiles, one RMP value can be divided by another, yielding the likelihood of the profile if it originated in the population of the numerator compared to that of the denominator. An LR1 value was calculated for each sample by dividing the highest RMP by the second

highest RMP obtained among the four populations.

$$LR1 = \text{highest RMP} / \text{second highest RMP}$$

LR1 thresholds were evaluated, above which a sample would be classified as belonging to a specific population (the population of the numerator RMP), and below which samples were defined as inconclusive. The latter designation is still informative, as the inconclusive result is 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 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. A practitioner would be expected 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 this 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.

31 SNP + Diplotype Ancestry Prediction Model: The above approach was repeated with the use of diplotype frequencies for rs12913832 and rs916977 (the latter being previously excluded from analysis). This approach could also be performed in Snipper by using the "Classification with a custom Excel file of frequencies" method, and uploading a file with frequency data for each allele or diplotype, rather than genotype data.

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 regression model. Any SNPs showing evidence of a significant association with ancestry (via the pseudo r^2 values provided 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 model, which was iteratively adjusted for inclusion/exclusion of SNPs until the final model was chosen. Because the ancestry of the 1000 subjects used in building dataset were well defined, the final model was simplified 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 (Chi-squared Automatic Interaction Detector) is one of the oldest methods of the latter form of data mining, originally proposed by Kass (1980). This method builds non-binary decision trees, based on a relatively simple algorithm, using the Chi-square test to determine the next split at each step in the decision tree.

In this case, the predictors were the genotypes of the SNPs used and the items being classified were ancestry groups. The categorical predictors are discontinuous so they are easily divided. Bi-allelic SNPs have 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

coded as 1, 2, or 3, respectively. The goal of the algorithm is 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 Component Analysis).

Using an "Exhaustive CHAID" algorithm, which performs a more thorough merging and testing of predictor values, all decisions were reduced until only two categories remained for each predictor. To carry out this analysis, the training set spreadsheet was loaded into Statistica (12^{th} edition, 64 bit) (StatSoft, Tulsa, OK), the dependent and categorical variables were chosen (ancestry groups and SNPs, respectively) and the algorithm was run to create a classification tree. The classification tree is represented as a graph, allowing the user to envision the process by showing the SNP genotype involved in each split. The resulting tree was then used to predict ancestry in the test set.

CHAID based 4 SNP + Diplotype Decision Tree Ancestry Prediction Model: The above approach was repeated with the use of diplotype frequencies for rs12913832 and rs916977 (the latter previously being excluded from analysis).

Results and Discussion

SNPs included in each model are listed in Appendix Table 5.

32 SNP RMP/LR Ancestry Model Performance: Various thresholds were considered prior to interpretation of this data. With no threshold, 7.1% (N=9) of samples were predicted erroneously (i.e. the highest RMP value was from a population other than the known population of the sample). The highest LR1 value that resulted in an incorrect prediction was on the level of 10^5; therefore, a threshold of 10^6 (below which results would be considered inconclusive between the highest and second highest RMP population) was

considered. Under such a threshold, 57.5% (N=73) of samples were inconclusive, limiting the usefulness of the test. Based on these results, a threshold of 1000 was chosen, showing the best balance of sensitivity and accuracy. Above this level, results were considered significant (predictive of a single population), and below this level, results were considered inconclusive between the highest and second highest RMP.

See Figure 17a 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 a haplogroup E Y chromosome, which is found at high frequencies in African populations but is also noticeably present in southern European and Middle Eastern populations (Semino 2004). 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 17b 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 (described above). 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.

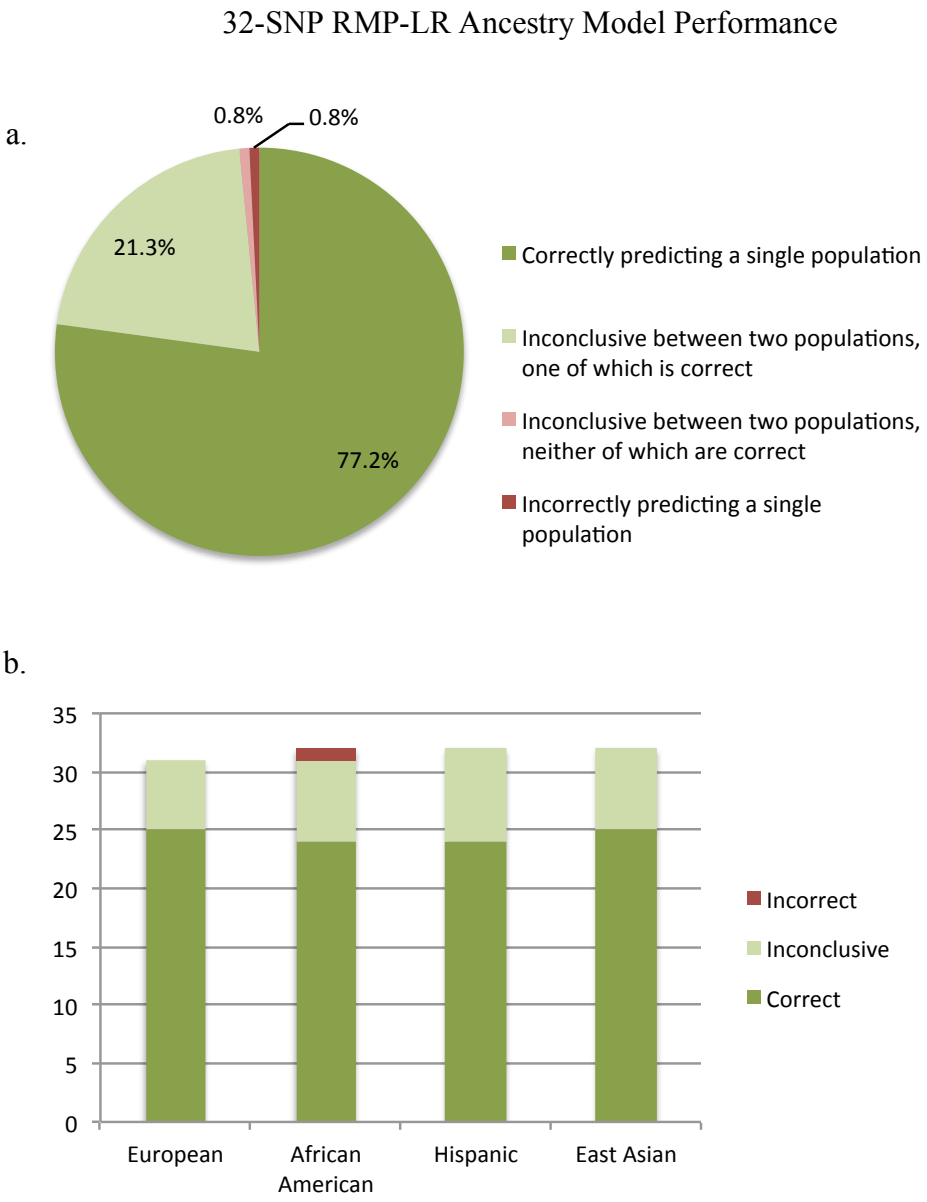


Figure 17. (a) Summary of overall 32 SNP RMP/LR ancestry model performance. Number of individuals in each category: Correct (107), Inconclusive including correct population (30), Inconclusive not including correct population (1), Incorrect (1). (b) 32 SNP RMP/LR ancestry model performance by population; similar distribution of inconclusive samples seen in each group, incorrect prediction only seen in African American population. Number of individuals in each category: European Correct (25), Inconclusive (6); African American Correct (24), Inconclusive (6), Incorrect (2); Hispanic Correct (24), Inconclusive (8); East Asian Correct (25), Inconclusive (7).

31 SNP + Diplotype RMP/LR Ancestry Model Performance: Figure 18a shows the overall performance of including the rs12913832 + rs916977 diplotype in place of the

single rs12913832 SNP. A total of five individuals are misclassified under this model, consisting of four African American and one Hispanic American sample. One of these African American samples was previously incorrect, and one was previously inconclusive between two populations, neither of which was correct, so essentially three new individuals are misclassified under this model. In addition, one East Asian individual who was previously correctly classified has become inconclusive between East Asian and Hispanic/Native American.

The breakdown of results by population seen in Figure 18b still shows a balance of correctly predicted individuals across the populations. The increase in incorrectly predicted individuals in the African American and Hispanic American populations only is likely due to European admixture, because the SNPs in the diplotype both differentiate European individuals. It is surprising that inclusion of the diplotype does not improve the prediction ability within the European population; on the contrary, one individual who was previously correctly predicted becomes inconclusive under this model.

Overall, performance of the model with the inclusion of this diplotype is poorer than performance using the single rs12913832 SNP. This is likely due to the dilution of predictive power for rs12913832. When used singly, there are only three possible genotypes for a biallelic SNP; whereas, in diplotype, there are a total of nine genotype combinations. Therefore, it becomes possible to dilute the power of a highly predictive SNP if it is combined with a less predictive SNP. The only instance in which the combination would improve prediction is if both SNPs are extremely and distinctly informative.

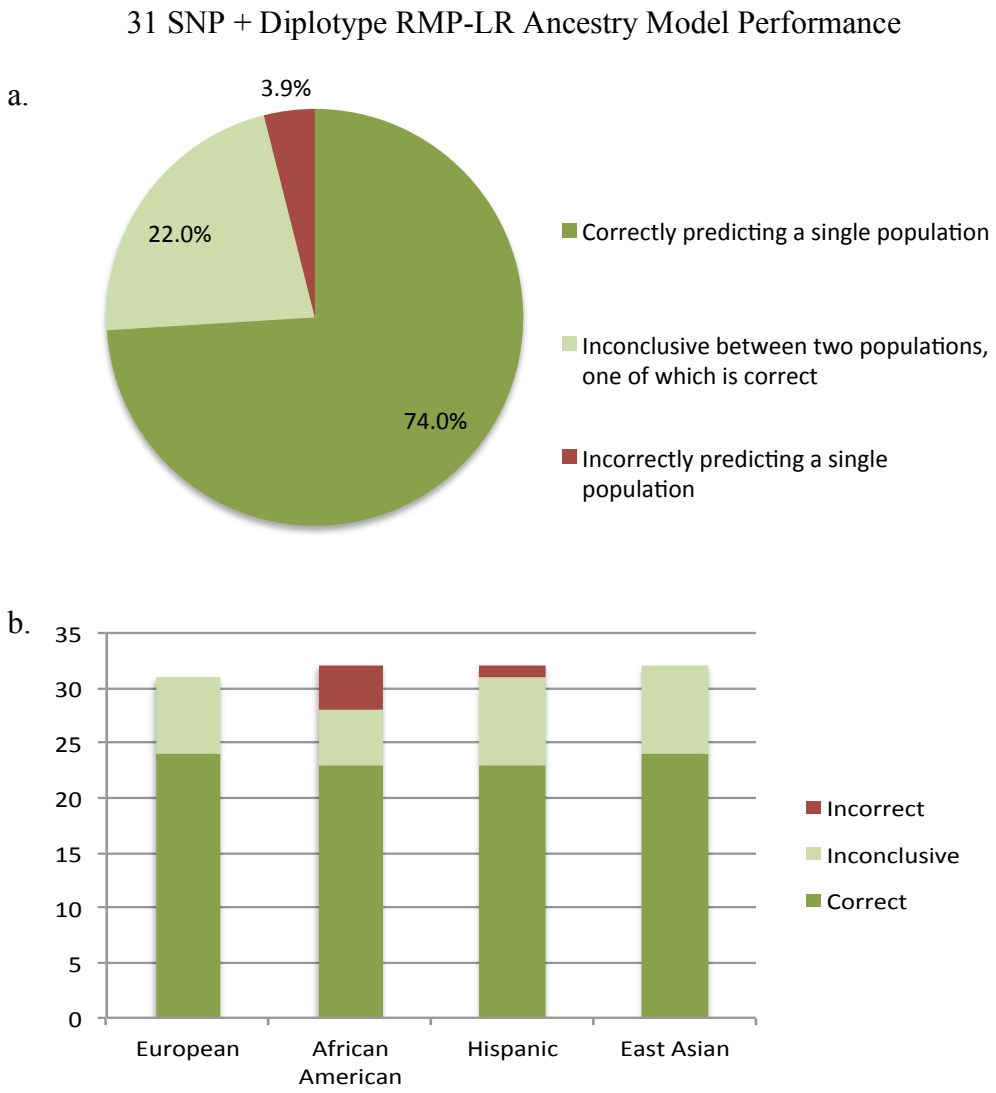


Figure 18. (a) Summary of overall 31 SNP + diplotype RMP/LR ancestry model performance, showing a worse outcome than the 32 SNP model. Number of individuals in each category: Correct (94), Inconclusive including correct population (28), Incorrect (5). (b) 31 SNP + diplotype RMP/LR ancestry model performance by population; similar distribution of correct samples seen in each group, incorrect prediction increases in African American and Hispanic population as compared to the 32 SNP model. Number of individuals in each category: European Correct (24), Inconclusive (7); African American Correct (23), Inconclusive (5), Incorrect (4); Hispanic Correct (23), Inconclusive (8), Incorrect (1); East Asian Correct (24), Inconclusive (8).

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 19a) show a significantly higher proportion of individuals (13.4%) would be incorrectly classified when compared to the previous model. As seen in Figure 19b, 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/Native American or East Asian individual). The proportion of inconclusive results varies widely among the populations (Figure 18b): at the low end, no inconclusive results were seen for the European American 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.

It should be noted that the benefits of using a smaller panel of SNPs could outweigh the lower performing ancestry prediction compared to the RMP/LR model. By using less SNPs, a single reaction could easily genotype all the loci, and such an assay might be more sensitive than the current reactions which contain at least twice as many loci. Additionally, by using less ancestry SNPs, other types of markers (such as individually identifying SNPs) could be added to increase the versatility of the assay.

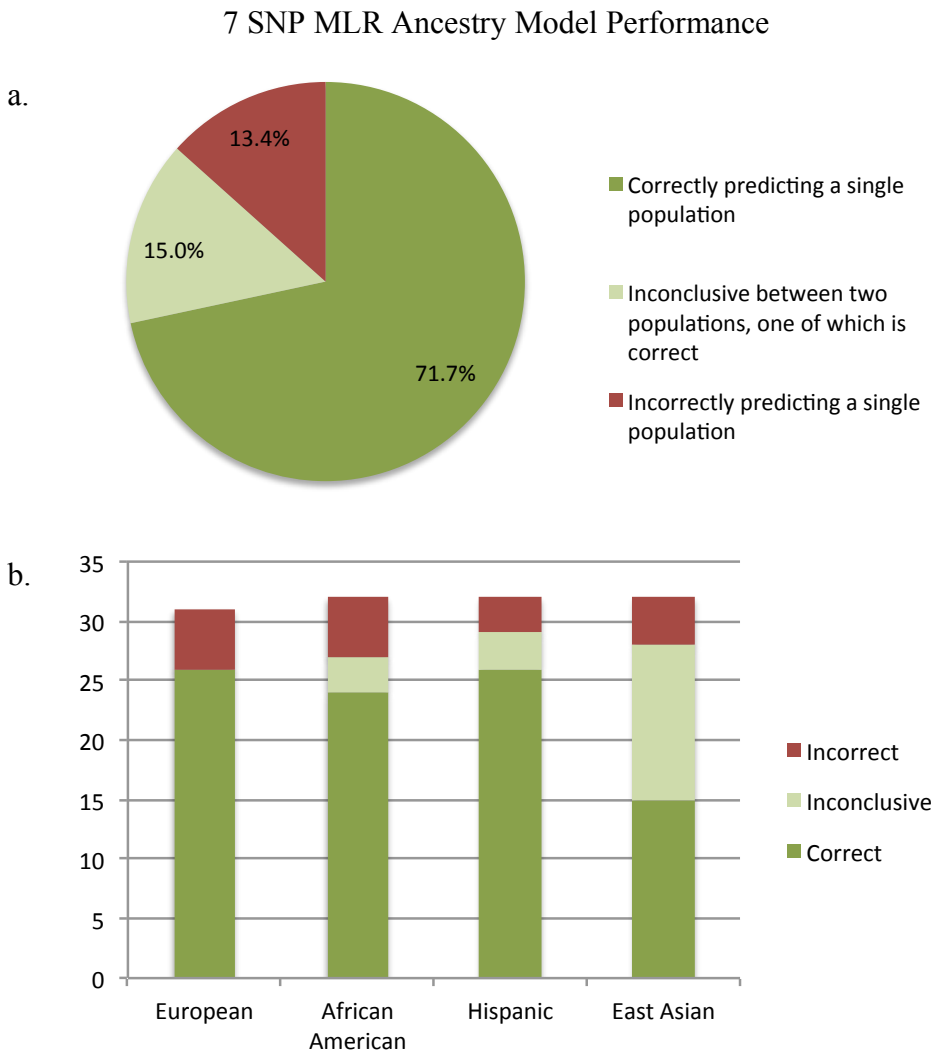


Figure 19. (a) Summary of overall 7-SNP MLR ancestry model performance. Number of individuals in each category: Correct (91), Inconclusive including correct population (19), Incorrect (17). (b) 7 SNP MLR ancestry model performance by population; similar distribution of incorrectly predicted samples seen in each group, varying distribution of inconclusive samples among the groups, indicating an imbalance of predictive ability for different populations. Number of individuals in each category: European Correct (26), Incorrect (5); African American Correct (24), Inconclusive (3), Incorrect (5); Hispanic Correct (26), Inconclusive (3), Incorrect (3); East Asian Correct (15), Inconclusive (13), Incorrect (4).

CHAID 5 SNP Decision Tree Ancestry 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 20a) show very similar overall results when compared to the 7-SNP MLR model, and a significant increase in incorrect predictions compared to the 32-SNP model. As seen in Figure 20b, 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 American and East Asian individuals, with European American individuals falling somewhere in between.

a.

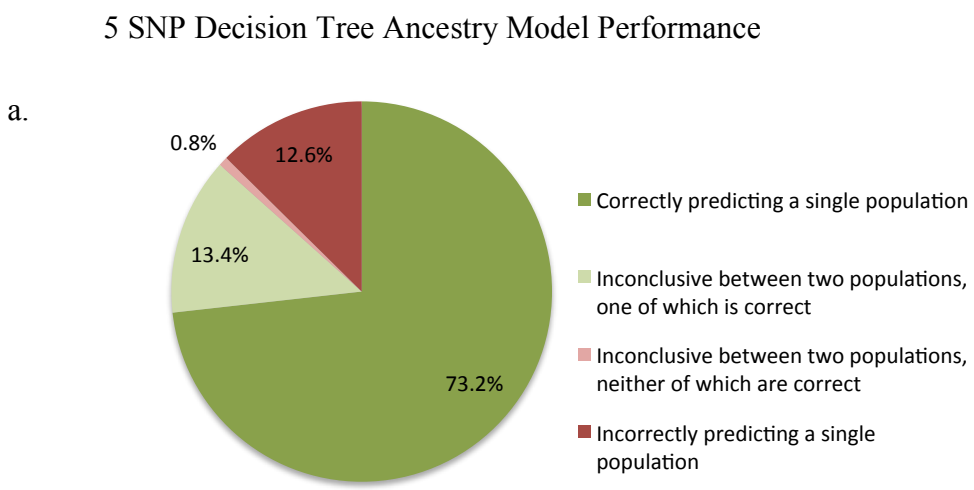


Figure 20. (a) Summary of overall 5 SNP decision tree ancestry model performance. Number of individuals in each category: Correct (93), Inconclusive including correct population (17), Inconclusive including incorrect populations (1), Incorrect (16).

5 SNP Decision Tree Ancestry Model Performance (cont.)

b.

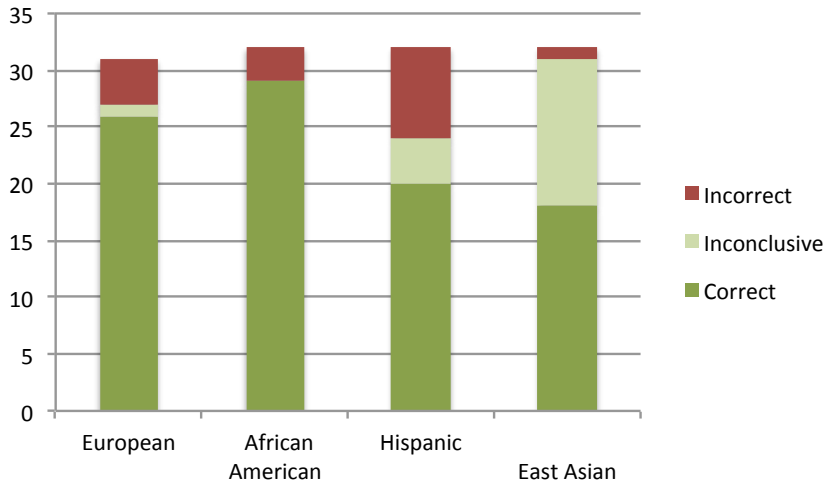


Figure 20. (b) 5 SNP decision tree ancestry model performance by population; imbalanced distribution of incorrectly predicted and inconclusive samples seen in each group, indicating an imbalance of predictive ability for different populations. Number of individuals in each category: European Correct (26), Inconclusive (1), Incorrect (4); African American Correct (29), Inconclusive (0), Incorrect (3); Hispanic Correct (20), Inconclusive (4), Incorrect (8); East Asian Correct (18), Inconclusive (13), Incorrect (1).

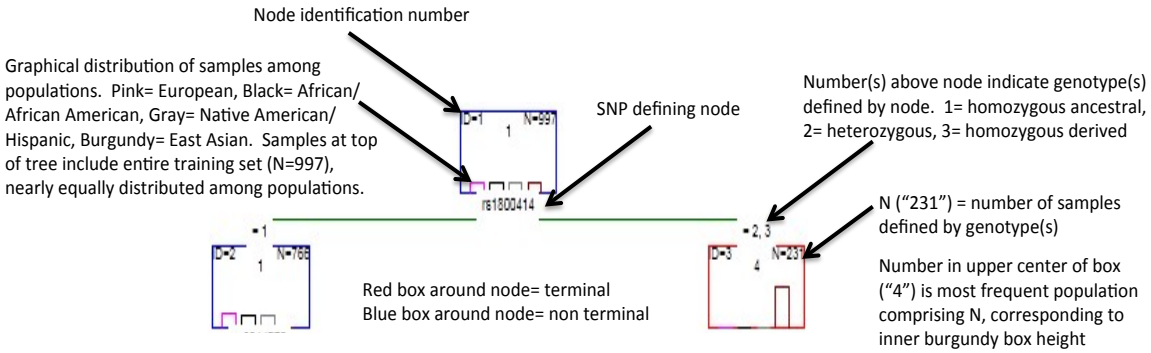
All but one of the inconclusive results (where the highest prediction probability is less than 0.8) are such that the correct population is one of the highest two predicted; therefore, correct information could still be given to the investigator for all but one of these individuals. Overall, correct information would be given for 86.7% of samples in the test set under this model.

An interesting aspect of this model is the generation of a decision tree, seen in Figure 21. This diagram shows how the model was built, using the five SNPs to best discriminate among the training set samples. Once the model is build upon the training set samples, it is used to evaluate each test set sample. A decision tree model would have the advantage of being straightforward to implement in a casework laboratory and easy to explain in court. Once the genotypes are determined for an unknown sample, the

practitioner would simply follow the tree and use the prediction and associated probability defined by the terminal node reached on the tree.

Decision Tree for 5 SNP Ancestry Model

a.



b.

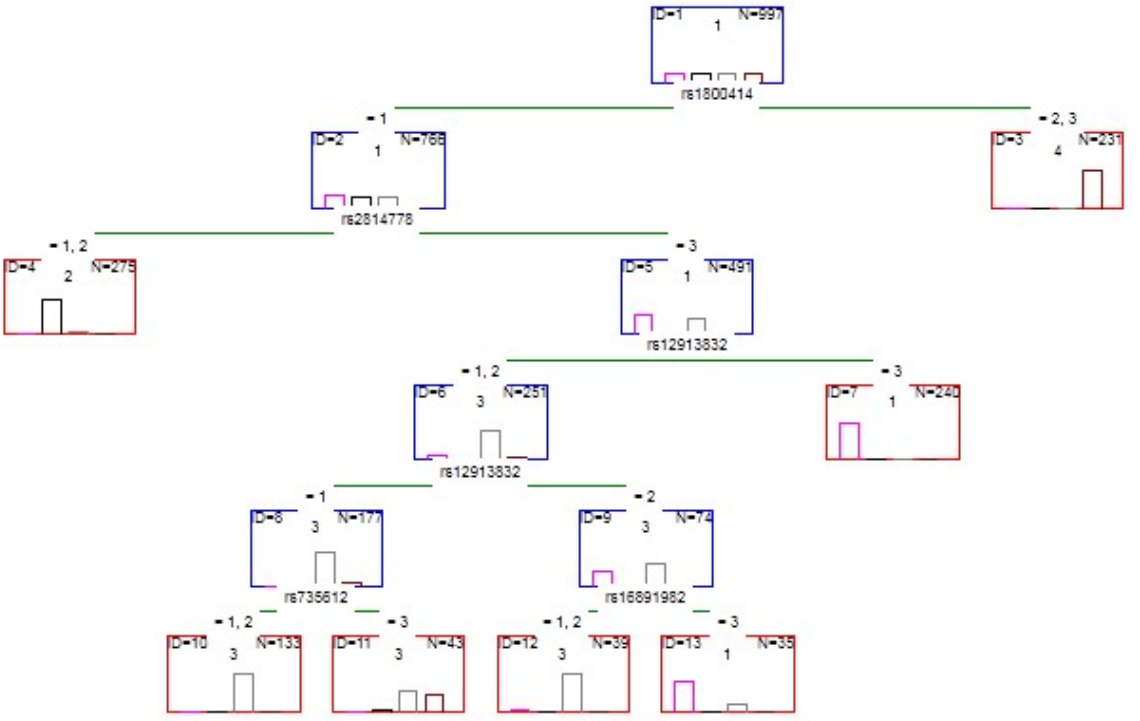


Figure 21. (a) Guide to reading decision tree. (b) 5 SNP decision tree created with training set samples. A single SNP may appear twice if each of the three possible genotypes is used to differentiate the samples.

CHAID based 4 SNP + Diplotype Decision Tree Ancestry Prediction Model: These results were also evaluated with prediction probability thresholds of 0.8 and 0.9. Seventy-one percent of individuals were inconclusive under a 0.9 threshold; therefore, despite having a lower error rate (4% at 0.9 compared to 11% at 0.8), the 0.8 threshold was used to further evaluate the results as described below.

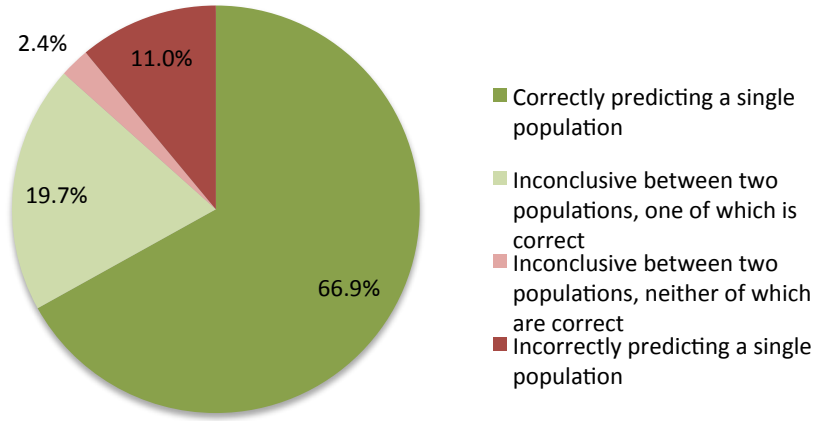
The results summary (Figure 22a) shows a slightly lower error rate for the 4 SNP + diplotype CHAID results compared to the 5 SNP results; however, when combining the misclassified samples with those that are inconclusive between two populations, neither of which are correct, the percentage of "misinformed" samples is the same between the two models (13.4%). In addition, the diplotype model results in a 6.3% increase in samples that are inconclusive between two populations, one of which is correct (and a corresponding decrease in samples correctly classified).

The results by population in Figure 22b show an improvement in African American and East Asian predictions, with poorer performance among European American and Hispanic American samples. This may be due in part to the diplotype and/or in part to this model's use of rs1375164 as opposed to the 5 SNP CHAID model's use of rs1800414 (SNPs are chosen by the algorithm as described in Materials and Methods). The remaining three SNPs in each model are identical.

Overall the CHAID model performs better with the single rs12913832 SNP as opposed to the diplotype. The decision tree for this model can be seen in Figure 23.

4 SNP + Diplotype Decision Tree Ancestry Model Performance

a.



b.

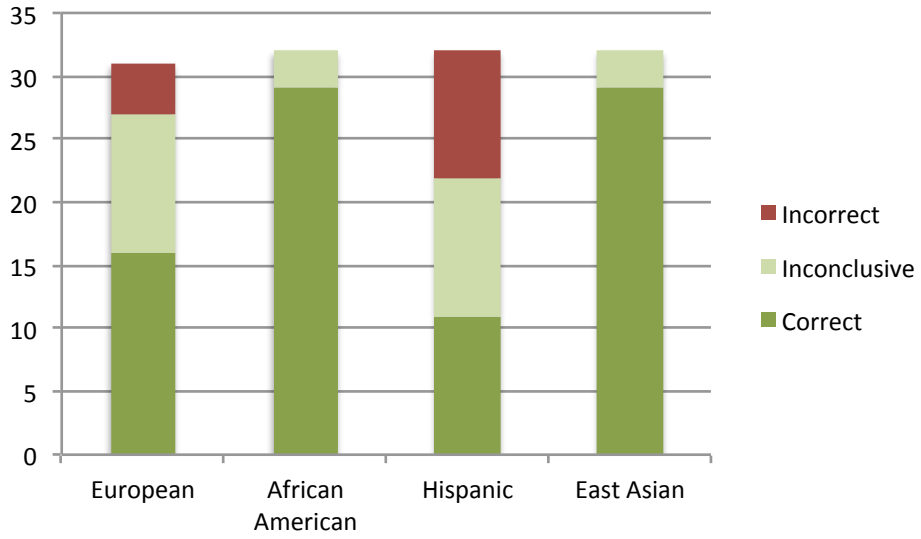


Figure 22. (a) Summary of overall 4 SNP + diplotype decision tree ancestry model performance. Number of individuals in each category: Correct (85), Inconclusive including correct population (25), Inconclusive including incorrect populations (3), Incorrect (14). (b) 4 SNP + diplotype decision tree ancestry 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. Number of individuals in each category: European Correct (16), Inconclusive (11), Incorrect (4); African American Correct (29), Inconclusive (3), Incorrect (0); Hispanic Correct (11), Inconclusive (11), Incorrect (10); East Asian Correct (29), Inconclusive (3), Incorrect (0).

4 SNP + Diplotype Decision Tree

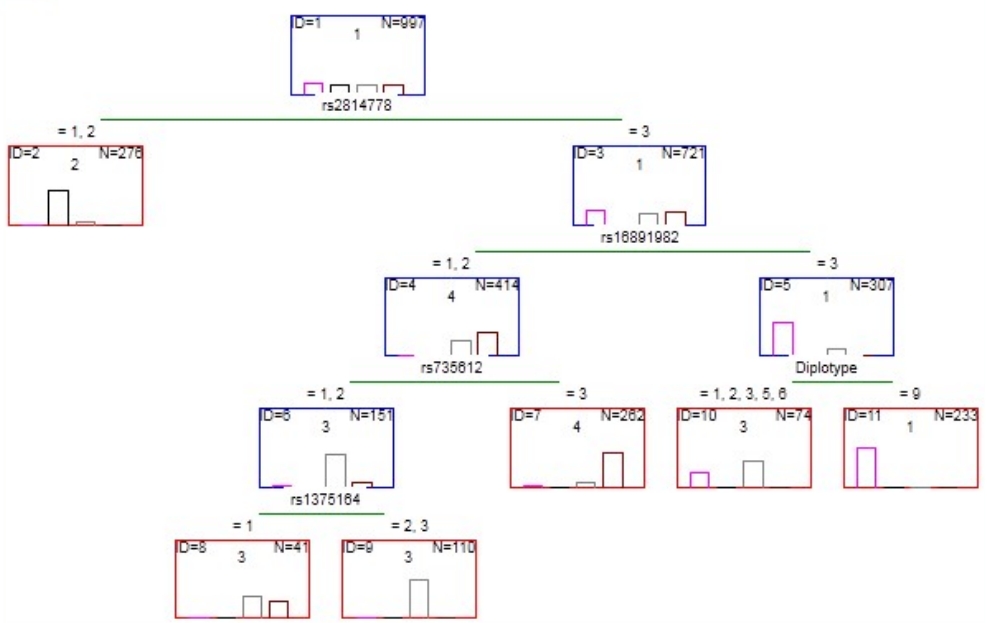


Figure 23. Decision tree model created with training set samples and diplotype for rs12913832/rs916977. See Figure 21 (a) for guide to reading decision tree.

A comparison of all ancestry models, overall and by population, can be seen in Table 5. Because a significant proportion of the U.S. population is variably admixed, it should be noted for all models that any prediction method which defines categories for ancestry determination will always produce errors from this spectrum of admixture.

Table 5. Comparison of all ancestry models. Whole numbers are number of individuals in each category. "Inconclusive (Correct)" category is inconclusive between two populations, one of which is correct. "Inconclusive (Incorrect)" category is inconclusive between two populations, neither of which is correct.

Overall (N=127)	**32 SNP RMP/LR**		**31 SNP + diplotype RMP/LR**		**7 SNP MLR**		**5 SNP CHAID**		**4 SNP + diplotype CHAID**	
Correct	98	77.2%	94	74.0%	91	71.7%	93	73.2%	85	66.9%
Inconclusive (Correct)	27	21.3%	28	22.0%	19	15.0%	17	13.4%	25	19.7%
Inconclusive (Incorrect)	1	0.8%	0	0.0%	0	0.0%	1	0.8%	3	2.4%
Incorrect	1	0.8%	5	3.9%	17	13.4%	16	12.6%	14	11.0%
European (N=31)										
Correct	25	80.6%	24	77.4%	26	83.9%	26	83.9%	16	51.6%
Inconclusive (Correct)	6	19.4%	7	22.6%	0	0.0%	0	0.0%	11	35.5%
Inconclusive (Incorrect)	0	0.0%	0	0.0%	0	0.0%	1	3.2%	0	0.0%
Incorrect	0	0.0%	0	0.0%	5	16.1%	4	12.9%	4	12.9%
African American (N=32)										
Correct	24	75.0%	23	71.9%	24	75.0%	29	90.6%	29	90.6%
Inconclusive (Correct)	6	18.8%	5	15.6%	3	9.4%	0	0.0%	0	0.0%
Inconclusive (Incorrect)	1	3.1%	0	0.0%	0	0.0%	0	0.0%	3	9.4%
Incorrect	1	3.1%	4	12.5%	5	15.6%	3	9.4%	0	0.0%
Hispanic (N=32)										
Correct	24	75.0%	23	71.9%	26	81.3%	18	56.3%	11	34.4%
Inconclusive (Correct)	8	25.0%	8	25.0%	3	9.4%	13	40.6%	11	34.4%
Inconclusive (Incorrect)	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%
Incorrect	0	0.0%	1	3.1%	3	9.4%	1	3.1%	10	31.3%
East Asian (N=32)										
Correct	25	78.1%	24	75.0%	15	46.9%	18	56.3%	29	90.6%
Inconclusive (Correct)	7	21.9%	8	25.0%	13	40.6%	13	40.6%	3	9.4%
Inconclusive (Incorrect)	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%
Incorrect	0	0.0%	0	0.0%	4	12.5%	1	3.1%	0	0.0%

Chapter 7: Development / Evaluation of Pigmentation Models

Once ancestry prediction has been established for a sample, phenotype predictions can provide additional investigative information. The most straightforward pigmentation prediction is that of eye color among Europeans, and models for this prediction were evaluated herein. Development of robust hair and skin pigmentation prediction models would require a much larger sample set than was available.

Materials and Methods

IrisPlex: The six SNPs comprising this MLR-based, published eye color model (Walsh 2011a) are included in the 50-SNP assay; therefore, the supplementary excel-based calculator was used to evaluate this model on the European American samples (N=196) and non-European American samples (N=77: African=9, African American=10, East Asian=17, Hispanic=12, Middle Eastern=1, Oceanic=2, South Asian=26,) for which eye color information was available. 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 5 SNP Eye Color Decision Tree: The methodology used in this analysis mimics that used in the CHAID ancestry model described in the preceding chapter. One hundred and eighty-seven European American individuals for whom eye color was known (blue,

brown or intermediate) and complete genotype information was available were evaluated with this method, using bootstrapping (*e.g.* all samples were used to build the model then each sample was removed and evaluated using the model) based on five SNPs chosen by the algorithm to best predict the trait (see Appendix Table 5 for full list of SNPs included under each model). The same method of evaluating results used for the IrisPlex model was used for the 5 SNP decision tree model.

CHAID 3 SNP + Diplotype Eye Color Decision Tree: This method used the diplotype described in Chapter 6 (rs12913832 + rs916977) in place of the single SNP rs12913832. The algorithm also selected one less SNP compared to the CHAID 5 SNP eye color model. The samples, classification, and evaluation method remain the same as the above 5 SNP model.

Results and Discussion

IrisPlex: As seen in Figure 24, results from testing 196 European American individuals for whom eye color information was available in the IrisPlex model show an expected

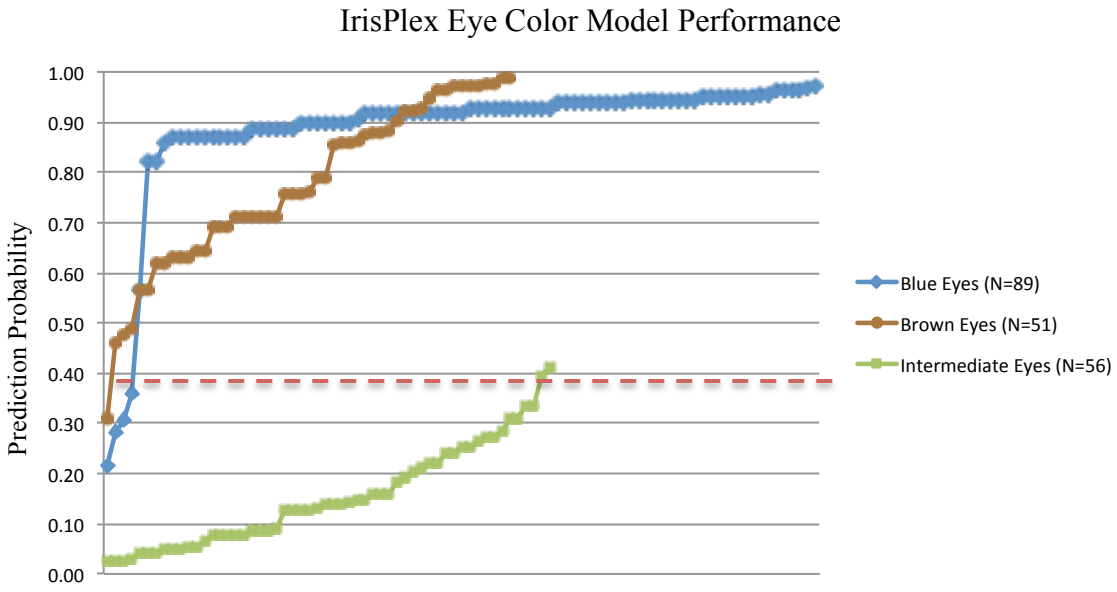


Figure 24. Results from IrisPlex model showing the prediction probability (y-axis) for the known eye color of each sample. Red dashed line indicates the level below which the known eye color is not the predicted eye color (with no threshold).

trade-off between accuracy and sensitivity, and an overall issue with predicting intermediate eye color. All 77 non-European American individuals were correctly predicted to have brown eyes.

Establishing a threshold below which a prediction probability is inconclusive will aide a practitioner in interpreting and delivering the results of this model. Figure 25 shows the results of applying 0.5, 0.7 and 0.9 thresholds. Using a 0.5 threshold, >90% of European American samples are classified but the error rate is high at 23%. At a 0.7 threshold, 75% of European American samples are classified with a 14% error rate, and lastly, at a 0.9 threshold, only 48% of European American samples are classified with a 7% error rate. At each threshold, the error rate is largely comprised of individuals with intermediate eye color who are predicted to have blue or brown eyes.

IrisPlex Eye Color Model Performance with Thresholds

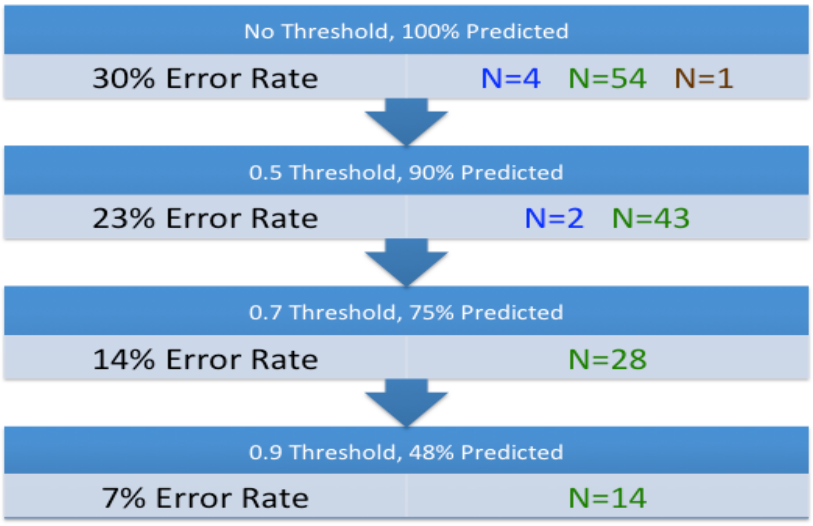


Figure 25. Results for the IrisPlex model at various thresholds. The "N" values correspond to individuals with the color-coded known eye color who are erroneously predicted to have a different eye color.

Based on this data set, the use of a 0.7 threshold allows for eye color prediction in three-fourths of European American 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. Among non-European Americans, the 0.7 threshold allows for prediction of all samples (and all are correct).

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, 100% of individuals tested 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. As the authors of the model hypothesize, this could be due to inconsistencies in phenotype categorization and/or the existence of unidentified variants that could better predict this phenotype.

5 SNP CHAID Eye Color Decision Tree: As was seen with the IrisPlex model, there is again a trade-off between accuracy and sensitivity, and an overall issue with predicting intermediate eye color, although this determination is slightly improved (Figure 26).

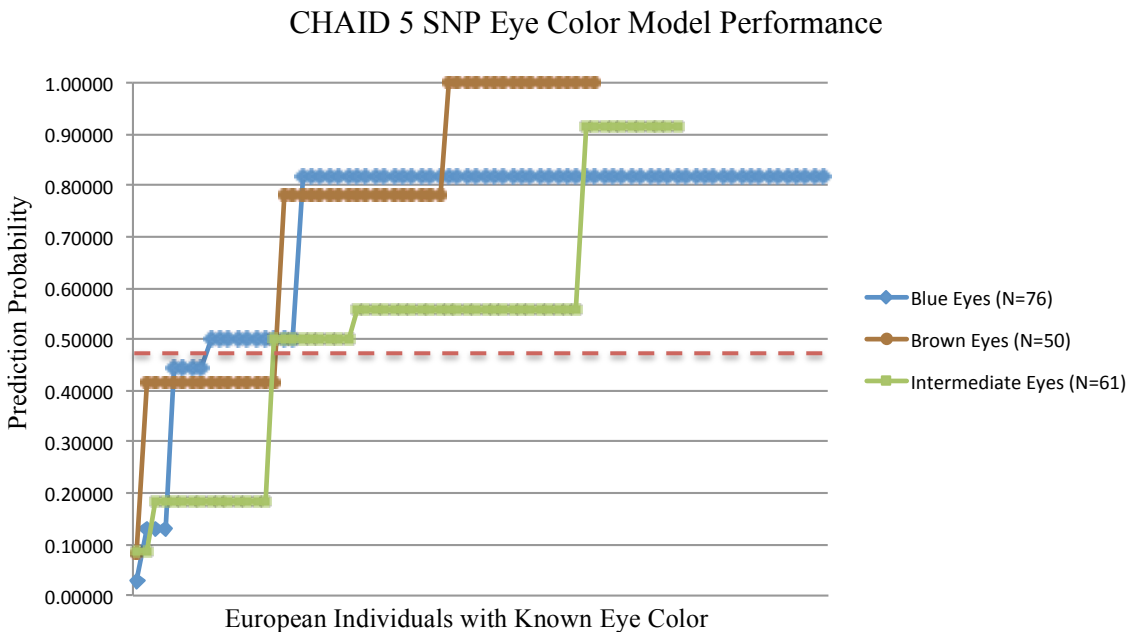


Figure 26. Results from 5 SNP decision tree model showing the prediction probability (y-axis) for the known eye color of each sample. Red dashed line indicates the level below which the known eye color is not the predicted eye color (with no threshold).

Nineteen samples have equal probabilities for two eye colors (ten known blue and nine known intermediate have 0.5 probability blue and 0.5 probability intermediate); therefore, these 19 samples are inconclusive with or without a threshold. Figure 27 summarizes the results at the different thresholds. The same results were obtained using no threshold or a 0.5 threshold: 90% of samples were predicted, with a 23% error rate. At a 0.7 threshold, 79% of samples are predicted, with a 13% error rate and the 0.9 threshold reduces the number of predicted individuals to an unacceptably low 16%. In comparison to the IrisPlex model, the error rates are similar and the best threshold under both models (0.7) has slightly higher percent predicted and slightly lower error rate under the 5 SNP model. However, under this model at 0.7 threshold, four individuals of known blue or brown eye color (three blue and one brown) are incorrectly predicted. Therefore, the previously described option of delivering the prediction as "not blue" or "not brown" would be incorrect for 2% of samples in this set under the 5 SNP model. The 5 SNP

decision tree generated from this sample set is shown in Figure 28.

CHAID 5 SNP Eye Color Model Performance with Thresholds

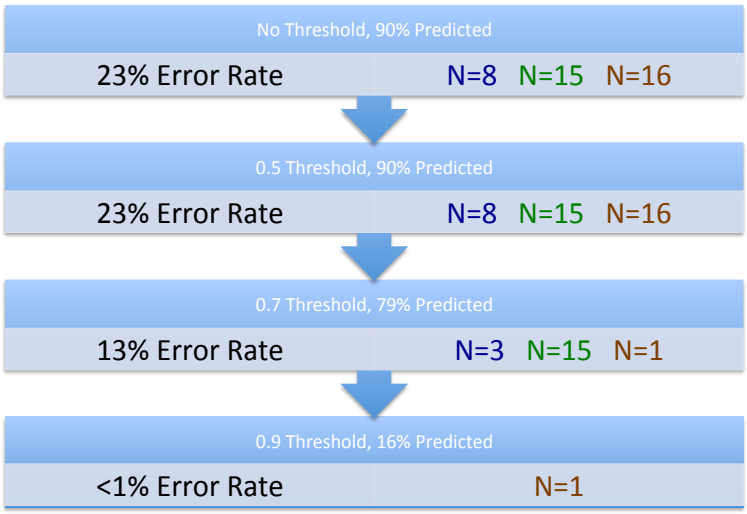


Figure 27. Results for the 5 SNP decision tree model at various thresholds. The "N" values correspond to individuals with the color-coded known eye color who are erroneously predicted to have a different eye color.

CHAID 5 SNP Eye Color Model Decision Tree

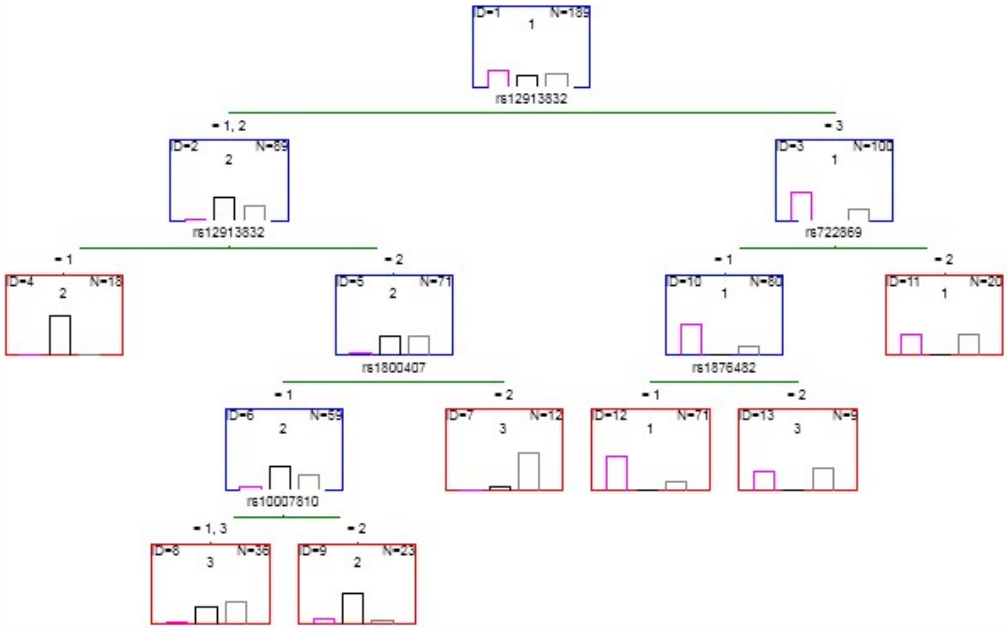


Figure 28. 5 SNP CHAID decision tree. The number in the center of each box represents the most frequent known eye color present in that node (1=blue, 2=brown, 3=intermediate), and the bars inside each box also represent the proportion of known eye colors present (pink=blue, black=brown, gray=intermediate). See guide to decision tree, Chapter 6 Figure 21a, for more information.

3 SNP + Diplotype CHAID Eye Color Decision Tree: Inclusion of the diplotype performs very similarly to the 5 SNP model, when comparing Figure 29 (below) to Figure 26.

However, differences can be seen by evaluating the performance at the three thresholds (Figure 30 below compared to Figure 27). With no threshold or 0.5 threshold, the percent predicted and error rate are the same under the 5 SNP or diplotype model, but the error rate has a different composition of samples (more known intermediate and less known blue in the diplotype model). At the 0.7 threshold, which still best balances sensitivity and accuracy, the error rate is 3% higher under the diplotype model while the percent predicted remains the same (79%). The 0.9 threshold applied to the diplotype model yields identical results to the 5 SNP model.

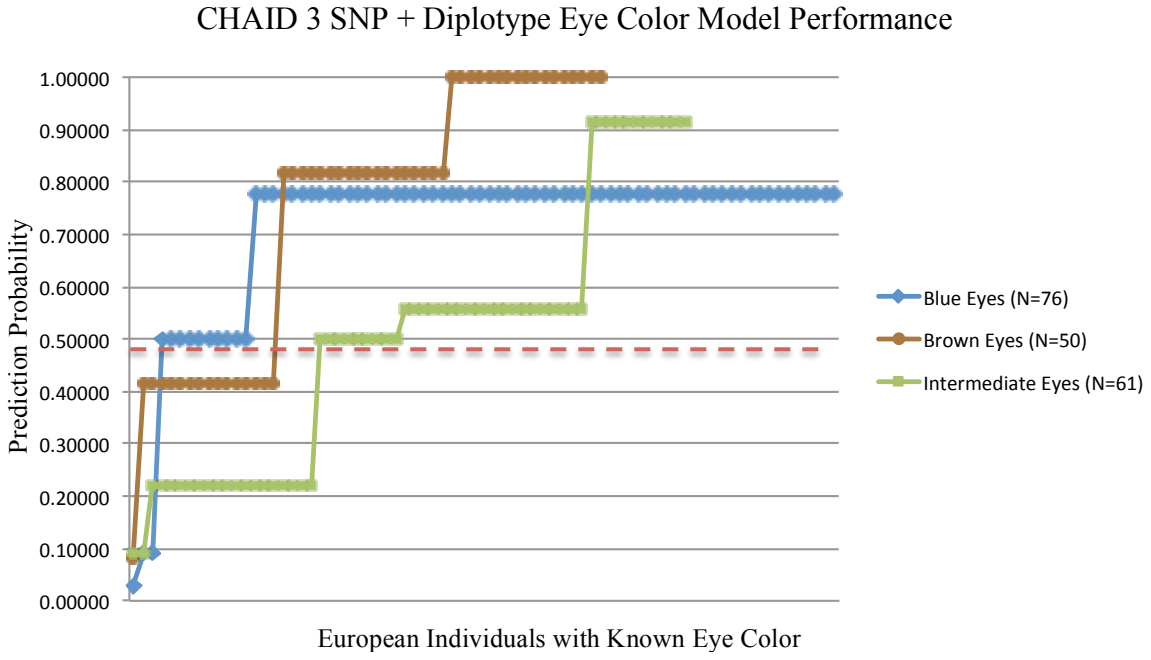


Figure 29. Results from 3 SNP + diplotype decision tree model showing the prediction probability (y-axis) for the known eye color of each sample. Red dashed line indicates the level below which the known eye color is not the predicted eye color (with no threshold).

As seen with the 5 SNP model, the option of delivering the prediction as "not blue" or "not brown" would likewise be incorrect for 2% of samples in this set under the

diplotype model. Overall, there appears to be no benefit in using the diplotype model. The 3 SNP + diplotype decision tree generated from this sample set is seen in Figure 31.

CHAID 3 SNP + Diplotype Eye Color Model Performance with Thresholds

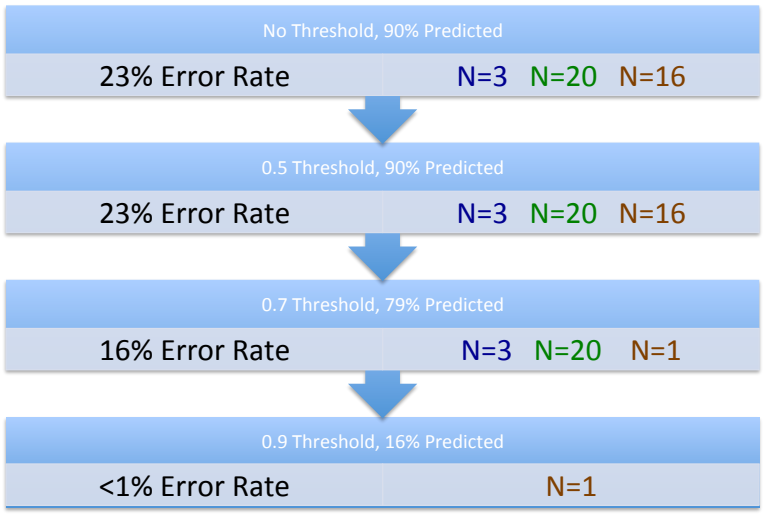


Figure 30. Results for the 3 SNP + diplotype decision tree model at various thresholds. The "N" values correspond to individuals with the color-coded known eye color who are erroneously predicted to have a different eye color.

CHAID 3 SNP + Diplotype Eye Color Model Decision Tree

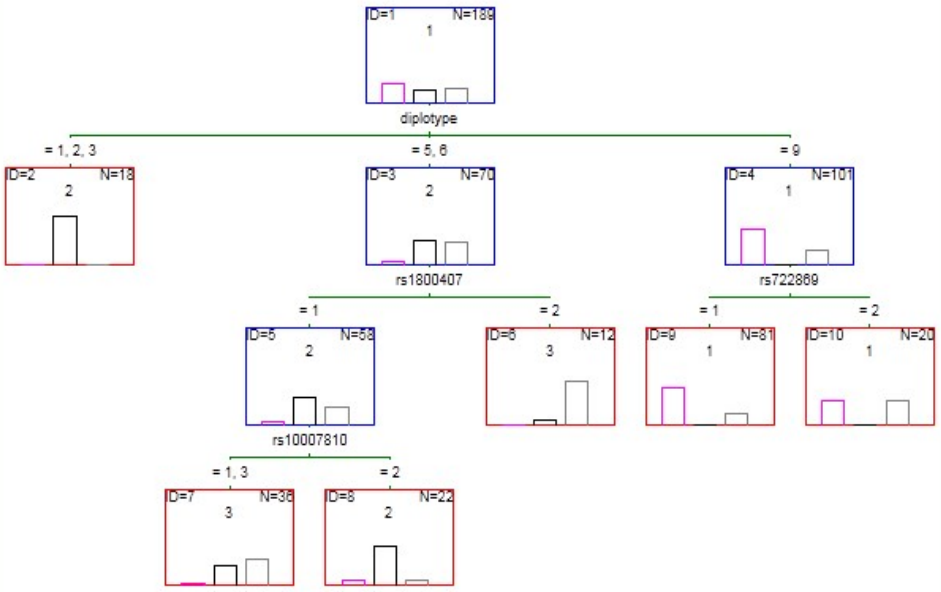


Figure 31. 3 SNP + Diplotype CHAID decision tree. The number in the center of each box represents the most frequent known eye color present in that node (1=blue, 2=brown, 3=intermediate), and the bars inside each box also represent the proportion of known eye colors present (pink=blue, black=brown, gray=intermediate). See guide to decision tree, Chapter 6 Figure 21a, for more information.

Chapter 8: Integration of Established Forensic Markers

From the literature it is clear and expected that far less ancestry information is contained in the forensic STR loci compared to ancestry SNPs, 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. In addition, other forensic marker information may be available, namely the mitochondrial DNA (mtDNA) and/or Y-chromosome haplotype. This chapter examines benefits and considerations in incorporating other forensic markers into the SNP ancestry prediction model.

Materials and Methods

LD Analysis: The forensic STR loci and the 32 SNPs in the previously described ancestry model are all present on autosomes; therefore, the possibility of linkage disequilibrium between the STR loci and the 32 SNP loci was evaluated. This statistical analysis was performed in a similar fashion to that described in Chapter 6 using WGAviewer; however, it was necessary to identify tag SNPs that flanked the STR regions, and then to evaluate LD between the tag SNP and the ancestry SNP. These tag SNPs varied in distance from the STR repeat regions. An increasing distance between the tag SNP and STR would result in a decreasing distance between the tag SNP and the ancestry SNP of interest, and thus an overestimation of LD.

See Appendix Table 10 for results of this analysis; no LD was found between the 15 STR loci and the 32 ancestry SNPs. WGAviewer does not calculate a corresponding

statistical significance; however, these results combined with the large distance between SNP and STR loci (the closest SNP to STR locus, rs952718 and D2S1338, are nearly three million bases apart) make LD highly unlikely.

STR analysis – RMP/LR: Forensic STR genotyping results for the previously described test set samples from the European American, African American, and Hispanic American populations were provided by NIST, and a subset of the East Asian samples (those collected or available at GWU, N=10) were genotyped in the 15 forensic STR loci of the Applied Biosystems Identifiler kit.

Frequency data were gathered from the four populations at the 15 STR loci, combining different datasets in an attempt to mimic the subpopulation demographics found in the SNP training set (e.g. using a combination/average of Hispanic American and Native American populations for Hispanic/Native American frequencies), see Table 6 for composition of the data set, and Appendix Table 8 for the STR allele frequencies. For rare alleles, a minimum allele frequency of 5/2n (where n is the number of individuals in the database) was assigned.

This frequency data was used to calculate the STR-based RMP for each sample in the test set, using p^2 for homozygous loci and 2pq for heterozygous loci, then calculating the product across all loci (these forensic STR loci are unlinked). To evaluate the STR data alone, the LR_1 was calculated for the STR data in the same way it was calculated for the SNPs, as described in Chapter 6. Then, STR RMPs were multiplied by the 32 SNP RMP result for each test set sample, and the combined results were evaluated to determine if the addition of STR data improves or confounds the ancestry determination.

Table 6. Sources of population frequency data for STR allele frequencies. Hardy-Weinburg calculations are included in the publications.

Reference	European	African American	Hispanic/Native American	East Asian
Life Technologies Corporation 2012	U.S. Caucasian N=349	African American N=357	U.S. Hispanic N=290 Native American N=191	
Butler 2003	U.S. Caucasian N=302	African American N=258	U.S. Hispanic N=140	
Hashiyada 2003				Japanese N=526
Shengjie 2008				Chinese Yunnan Han N=497

Mitochondrial DNA and Y chromosome analysis – RMP/LR: The test set samples from the European American, African American, and Hispanic American 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 American sample, and Y data was missing for one African American sample. No mtDNA or Y chromosome data were available for the East Asian samples; therefore, these were excluded from this analysis.

To evaluate the haplogroup frequencies and whether including the haplogroups would improve the overall analysis, population haplogroup frequency data were gathered for the four populations in the mitochondrial genome and the Y chromosome, attempting to mimic the subpopulation demographics found in the SNP training set, see Table 7. For rare alleles, a minimum allele frequency of 5/n (where n is the number of individuals in the database) was assigned. Appendix Table 9 contains the haplogroup frequencies.

Table 7. Sources of population frequency data for mtDNA and Y chromosome haplogroup frequencies.

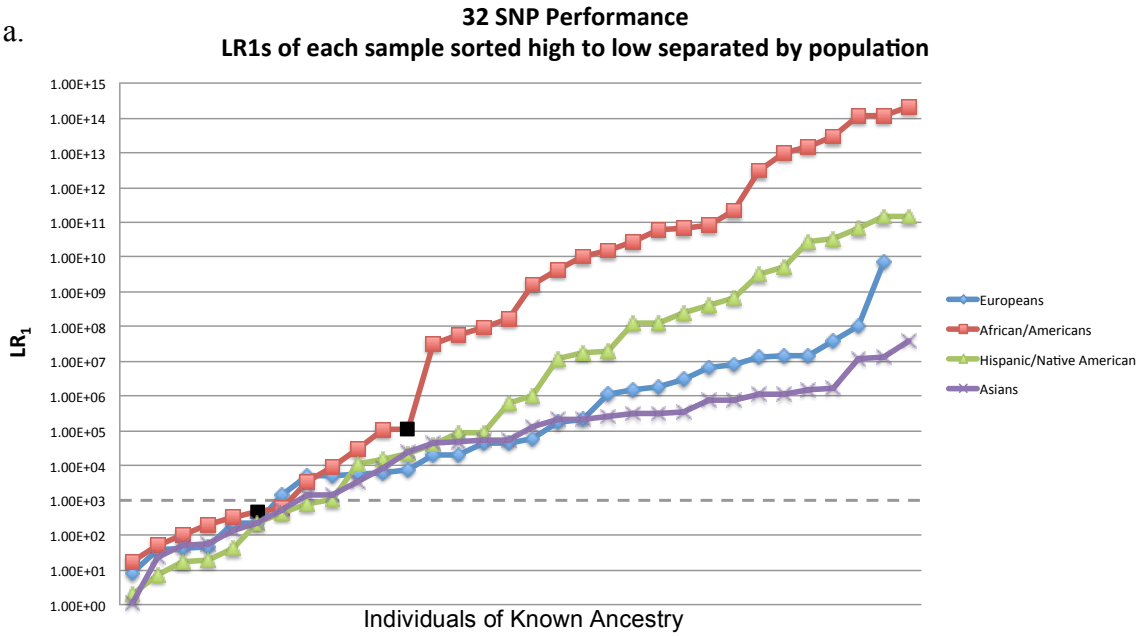
Reference	European	African American	Hispanic/Native American	East Asian
Mitochondrial DNA Haplogroups				
Allard 2002	U.S. Caucasian N=1771			
Allard 2005		African American N=1128		
Budowle 2002			Apaches and Navajos N=326	
Allard 2006			U.S. Hispanic N=686	
Lee 2006				Korean N=694
Irwin 2009				Chinese (Hong Kong) N=369
Irwin 2008				Vietnamese N=185
Y Chromosome Haplogroups				
Willuweit 2007	European Metapopulation 48 populations: • 31 Western • 13 Eastern • 4 South-Eastern N=4234	Sub-Saharan African Metapopulation 7 populations N=317	Native American Metapopulation 47 populations N=1325	Korean and Japanese Metapopulation 30 populations: • 9 Korean • 21 Japanese N=3015

Test set samples with predetermined mtDNA and/or Y chromosome haplotype data were assigned haplogroup frequencies, and the frequencies were evaluated to determine the ancestry prediction success rate based solely on mtDNA or Y data within each population. Then, the frequencies were multiplied by the 32-SNP RMP result for each test set sample (separate calculations for mtDNA and Y chromosome), and the LR_1 calculated. Lastly, all markers (32 SNPs, 15 STRs, mtDNA and Y) were combined into a single LR_1.

Results and Discussion

STR analysis – RMP/LR: As shown in Figure 32, the overall lower LR_1 values seen in the 15 forensic STR loci as compared to the 32 SNP ancestry panel are indicative of the considerably lower ancestry information content in the STR loci. Also the number of individuals that would be erroneously predicted is far higher, if this determination were based solely on the STR data (this is clearly not recommended).

a.



b.

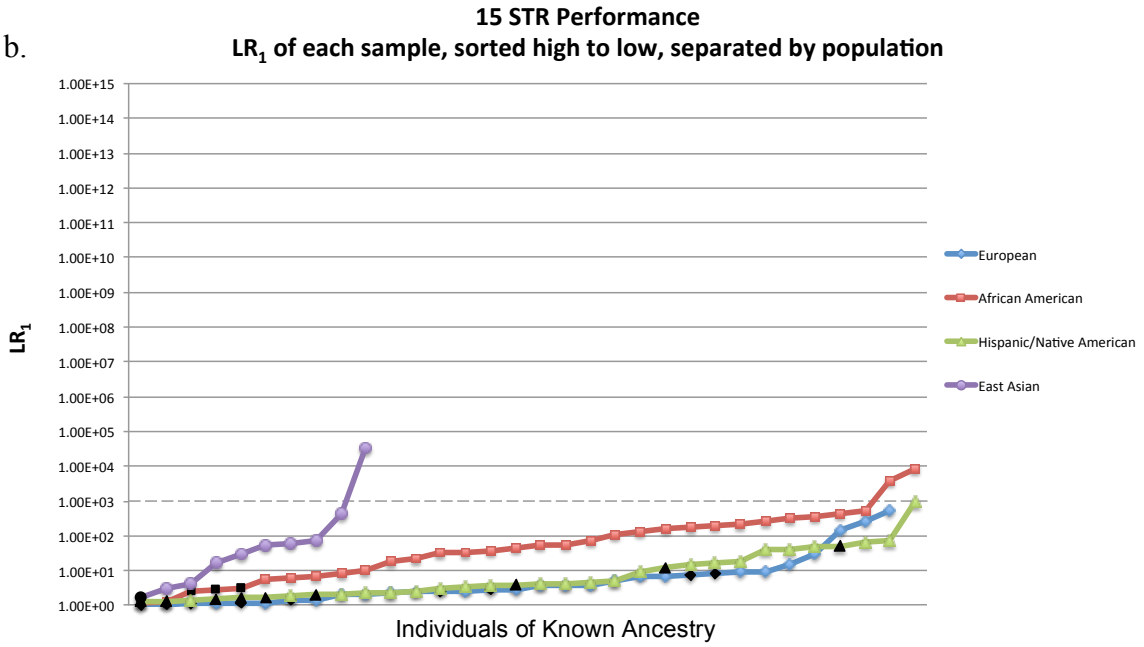


Figure 32. (a) Comparison of LR_1 values (y-axis, logarithmic scale) obtained based on the 32 SNP data alone. The dotted line is the likelihood ratio threshold of 1000, above which samples are predicted for a single population and below which samples are inconclusive between two populations. Black markers indicate samples that are incorrectly predicted. (b) Comparison of LR_1 values (y-axis, logarithmic scale) obtained based the STR data alone. The STR data figure shows fewer East Asian samples, reflecting the lack of STR data for 22 East Asian test set samples.

When combining the STR data into the 32 SNP ancestry model, however, seven samples, which were previously considered inconclusive between two populations

(because they failed to reach the likelihood ratio threshold of 1000) were conclusively and correctly predicted, seen in Figure 33. The error rate remains the same as it was under the SNP model, after adjusting for East Asian samples lacking STR data. Specifically, under the combined model, two samples are now misclassified above the 1000 threshold and would be incorrectly predicted for a single population; whereas, under the SNP model, one sample was misclassified above and one sample below the threshold.

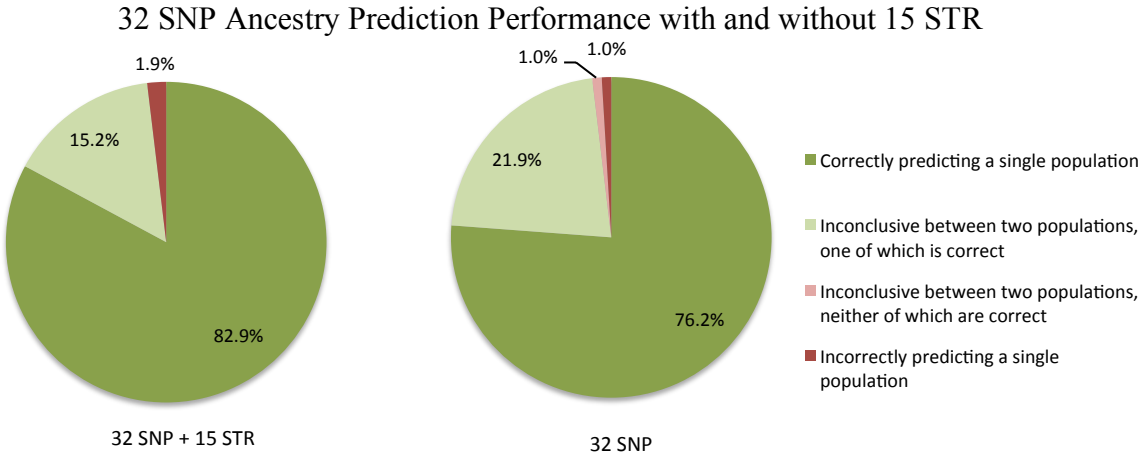


Figure 33. Performance comparison of the 32 SNP + 15 STR model (left) and the 32 SNP model (right). The former shows improvement in the number of samples correctly predicted, while the error rate remains the same. Note: East Asians for whom STR data was unavailable (N=22) were excluded from the 32 SNP model performance for the purpose of this comparison, resulting in slightly different percentages than those shown in Chapter 6, Figure 16. Number of individuals in each category are: 32 SNP + 15 STR- Correct (87), Inconclusive including correct population (16), Incorrect (2); 32 SNP- Correct (80), Inconclusive including correct population (23), Inconclusive including incorrect population (1), Incorrect (1).

Because of the relatively small change in percentages and low number of individuals tested, these results are not statistically significant; additional samples exhibiting the same trend would be needed to achieve significance. However, as the STR data already exists and this preliminary analysis shows improvement, combining the STR data into the 32 SNP ancestry model is recommended. For the purpose of forensic practitioners, the Snipper RMP-LR calculator can incorporate both the SNP and STR data if the "Classification with a custom Excel file of frequencies" option is used (wherein the

training set is replaced by a file containing *training set based* frequency data for the 32 SNPs and *literature based* frequency data for the 15 STRs), as opposed to the "Classification with a custom Excel file of populations" option described in Chapter 6.

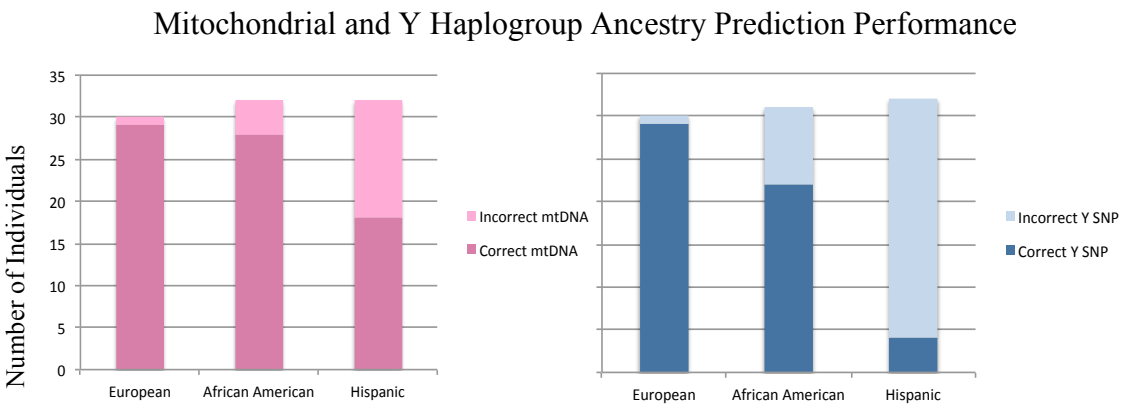


Figure 34. mtDNA and Y haplogroup ancestry prediction results by known population of test samples. (left) Dark pink portion of column indicates how often the mtDNA haplogroup is consistent with the known ancestry of the individual, light pink indicates the haplogroup is not consistent with the known ancestry. Number of individuals in each category: European- Correct (29), Incorrect (1); African American- Correct (27), Incorrect (4); Hispanic- Correct (18), Incorrect (4). (right) 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. Number of individuals in each category: European- Correct (29), Incorrect (1); African American- Correct (22), Incorrect (9); Hispanic- Correct (4), Incorrect (28).

Mitochondrial DNA and Y chromosome analysis – RMP/LR: Figure 34 shows wide population variation in the ability of mtDNA and Y chromosome DNA to predict overall ancestry.

For European American individuals, including the mitochondrial and Y chromosome information would improve the ancestry prediction for the majority of samples. This information will worsen the ancestry prediction 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 will be improved by including the mitochondrial haplogroup (87.5% of samples) and the Y

chromosome haplogroup (71% of samples) in the evaluation. The samples that would be negatively impacted by including these results include four samples with an "incorrect" mitochondrial haplogroup (two samples where the haplogroup is most common in Europeans and two samples where the haplogroup is most common in East Asians) and nine samples where the Y chromosome haplogroup is most frequent in Europeans.

The Hispanic American 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 known Hispanic American individuals that would be incorrectly classified based on both the mtDNA and Y chromosome haplogroups would be predicted European.

Figure 35 shows the effect on performance across populations when combining SNP and mtDNA, SNP and Y chromosome DNA, and all markers (SNP, STR, mtDNA and Y chromosome DNA). Table 8 shows the number of individuals in each category under each model, and the statistical significance of the differences between models.

The highest number of correctly predicted samples (nearly 85%) results from the combination of all markers; however, under this model, four samples are misclassified as opposed to two under either the SNP only or SNP+STR model. 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. Regarding the fact that East Asian individuals were not included in this analysis, it is expected that the addition of this population would improve the results when combining markers, due to the relative lack of admixture in this population (similar to the results for European American samples).

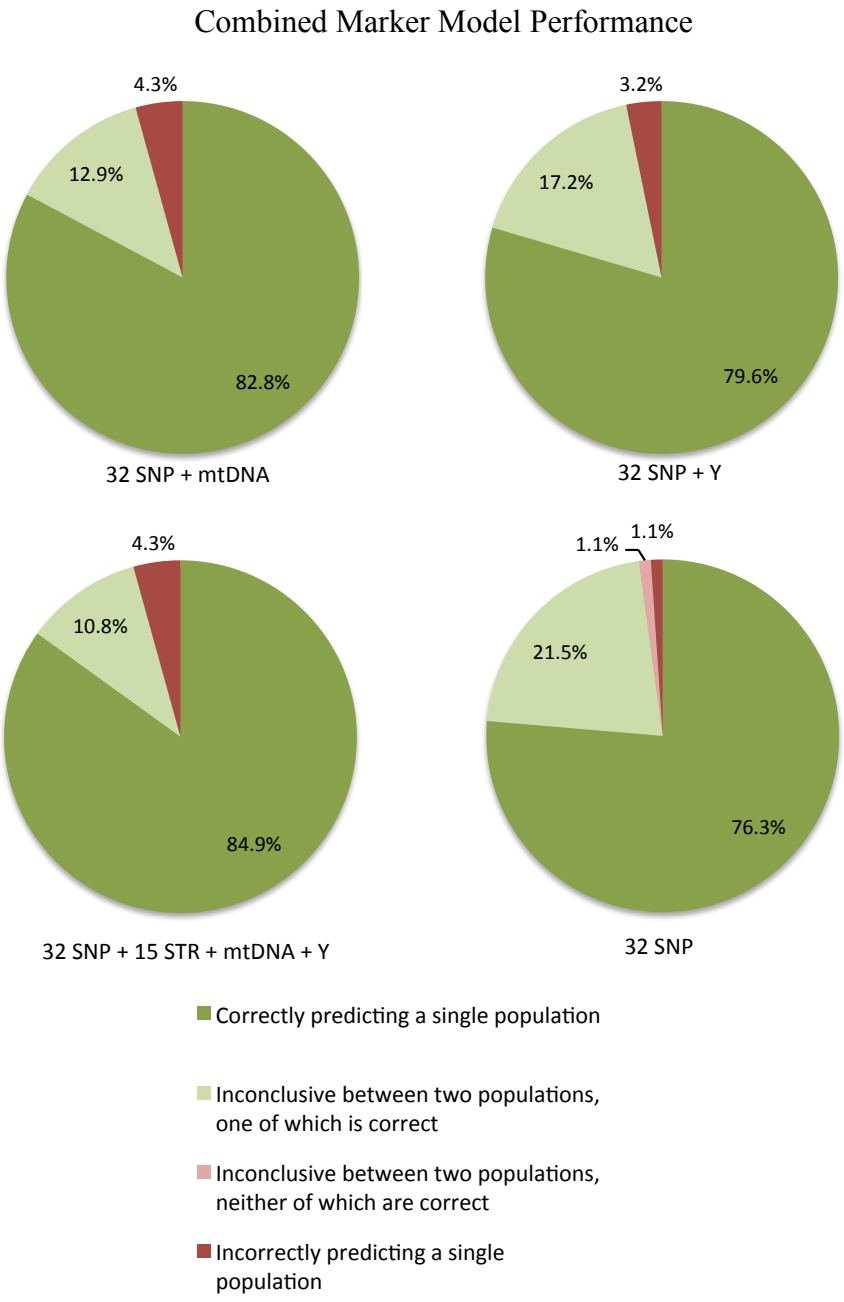


Figure 35. Performance of combined marker models. The highest percentage of correctly predicted samples results from inclusion of all markers; however this model also has the highest error rate. Note: All East Asian samples, as well as one European American and one African American sample were excluded from the SNP model performance for the purpose of this comparison because no corresponding mtDNA and/or Y chromosome DNA data was available, resulting in slightly different percentages from the SNP only results shown in Chapter 6, Figure 16.

Table 8. Comparison of all combination ancestry models. Whole numbers are number of individuals in each category. "Inconclusive (Correct)" category is inconclusive between two populations, one of which is correct. "Inconclusive (Incorrect)" category is inconclusive between two populations, neither of which is correct.

Overall (N=93)	**32 SNP RMP/LR**		**32 SNP + 15 STR**		**32 SNP + mtDNA**		**32 SNP + Y Chr**		**All Combined**	
Correct	71	76.3%	75	80.6%	77	82.8%	74	79.6%	79	84.9%
Inconclusive (Correct)	20	21.5%	16	17.2%	12	12.9%	16	17.2%	10	10.8%
Inconclusive (Incorrect)	1	1.1%	0	0.0%	0	0.0%	0	0.0%	0	0.0%
Incorrect	1	1.1%	2	2.2%	4	4.3%	3	3.2%	4	4.3%
European (N=30)										
Correct	24	80.0%	25	83.3%	26	86.7%	26	86.7%	29	96.7%
Inconclusive (Correct)	6	20.0%	5	16.7%	4	13.3%	4	13.3%	1	3.3%
Inconclusive (Incorrect)	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%
Incorrect	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%
African American (N=31)										
Correct	23	74.2%	26	83.9%	24	77.4%	25	80.6%	26	83.9%
Inconclusive (Correct)	6	19.4%	4	12.9%	5	16.1%	5	16.1%	4	12.9%
Inconclusive (Incorrect)	1	3.2%	0	0.0%	0	0.0%	0	0.0%	0	0.0%
Incorrect	1	3.2%	1	3.2%	2	6.5%	1	3.2%	1	3.2%
Hispanic (N=32)										
Correct	24	75.0%	24	75.0%	27	84.4%	23	71.9%	24	75.0%
Inconclusive (Correct)	8	25.0%	7	21.9%	3	9.4%	7	21.9%	5	15.6%
Inconclusive (Incorrect)	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%
Incorrect	0	0.0%	1	3.1%	2	6.3%	2	6.3%	3	9.4%

As was found with adding STR data, the relatively small change in percentages and low number of individuals tested render these results statistically insignificant, and more samples exhibiting a similar trend would be needed to definitively evaluate; however, these results discourage the incorporation of mtDNA or Y chromosomal DNA data. Practically speaking, it would be rare to have mtDNA results in a forensic case as most laboratories are not equipped to perform this analysis and the cost-benefit of outsourcing mtDNA sequencing is minimal in most cases. It is far more likely that a casework forensic laboratory would generate Y chromosome DNA results, in the form of a Y-STR profile; however, in order to determine the Y haplogroup, an additional prediction step based on the linkage disequilibrium between Y-STR loci and Y SNPs

defining Y haplogroups would be required (such as the "Y-Haplogroup Predictor" (Athey 2005)).

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.

Chapter 9: Evaluation of Next-Generation Sequencing for Forensics

The goal of this portion of the project was to perform a preliminary evaluation of the Pacific Biosciences RS next-generation DNA sequencer with forensic STR loci. Theoretically, the Pacific Biosciences RS has no limitations that would prevent it from being able to sequence all the currently established STR and YSTR loci, the mtDNA hypervariable region (or the entire mtDNA genome), and hundreds of SNPs, all in one assay.

Materials and Methods

First, five STR loci that are known to contain significant levels of sequence variation were chosen (D3S1358, D13S317, D21S11, FGA, and SE33). The gender typing locus Amelogenin was also included. Unlabeled forward and reverse primers were ordered from Integrated DNA Technologies (Coralville, IA), using sequences from commercial forensic kits (Krenke 2002, McLaren 2012). Seven known samples were amplified at these six loci individually according to a forensic STR kit manufacturer's monoplex protocol (Promega 2011). The amplicons were run on a 0.5% agarose product gel and normalized to approximately 1 ng/µl. The amplicons were pooled so that each sample contained approximately equal proportions of each locus. These samples (singleplexes and pooled samples) were run on the PacBio RS instrument at CNMC.

Results and Discussion

First runs of these samples (also some of the first instrument runs at CNMC) showed read lengths too short for STRs and no identifiable repeat sequences. Improvements in the chemistry are ongoing; but at this time, the results obtained indicate the Pacific Biosciences RS may not be a viable option for forensic samples.

Chapter 10: Discussion / Conclusions

The overall goal of this research was to provide a DNA based assay and corresponding statistical model that can provide ancestry and phenotypic information of an individual to complement a criminal investigation when no STR match is found. The project was divided into six phases.

In the first phase of this project, a large sample collection effort was designed and executed for candidate SNP evaluation. Beyond its use in the research presented here, this collection of volunteer samples (over 300 samples to-date and collection by laboratory staff continues) including a questionnaire of ancestry and phenotype data, DNA sample, and spectrophotometric skin measurements, represents a valuable repository for future research and a resource for other forensic researchers.

Also in this phase, SBE assays were built for genotyping the candidate SNPs. The SBE methodology is advantageous for researchers because assays can be custom designed in the research laboratory, and for forensic practitioners as the equipment needed to perform the assay is already present in the forensic casework laboratory, making such an assay easy to implement. Compared to the technologies used in genetics laboratories, however, this methodology is labor intensive and low throughput. Ideally it is a jumping-off point for forensic DNA laboratories, so that SNP data can be used in casework immediately, and will be replaced by higher throughput technologies once these become less expensive and more amenable to forensic samples.

The second phase of this project employed many statistical approaches in order to reduce the number of candidate SNPs to include only the most informative markers. Aside from some obvious choices, SNPs that have shown very strong associations to

ancestry and/or pigmentation phenotypes, a multifactorial approach to SNP selection was used.

SNPs were evaluated for their ancestry content using chi-squared, PCA, $F_{ST,}$ and web-based Snipper analyses for the most common populations in the United States: European American, African American, East Asian, and Hispanic/Native American. Because the sample set with corresponding phenotype information was largely European American in origin, and because the existing pigmentation research is skewed heavily toward European-specific changes in melanogenesis, phenotype analyses were performed among European Americans only. Phenotype SNPs were chosen based on chi-squared, PCA, and haplotype analyses; the ability of markers to serve dual-roles (providing both ancestry and phenotype information); and the presence of the markers in established pigmentation models. By evaluating the results of both the ancestry and phenotype analyses, a subset of 50 SNPs was selected.

Several SNPs were included that have been associated with premature balding; however, these could not be evaluated due to a limited number of balding males in the sample set. This phenotype would be useful in specific circumstances; it would only apply to male perpetrators in the age range affected by premature balding.

In the third phase of the project, the 50 SNPs were incorporated into an SBE assay, made up of three multiplexes, for use in forensic casework. In order to be suitable for forensic use, the assay must be inexpensive, easy to implement, sensitive, and robust. This assay was optimized to 100 pg of input DNA per multiplex; however, several SNPs must be interpreted with care at this low input level, and at least 0.5 ng of input DNA per multiplex is ideal. Robust results were obtained with mock forensic samples and

different forensic extraction methods.

Along with providing the genotype information, it is important to give a statistical weight to the ancestry and/or phenotype prediction, and this was addressed in the next phase of the research. Most statistical methods assume independence among included loci; therefore, the first step was to evaluate linkage disequilibrium and choose the best markers in linked regions. One approach to overcoming this issue is to include linked loci as a haplotype. This was evaluated on a small-scale by incorporating a diplotype frequency in place of a single SNP into several ancestry and phenotype models; however, no improvement in model performance was seen. The possibility remains that a different set of SNPs or incorporation of a larger haplotype could improve performance.

An ancestry training set was established so that all models could be evaluated using the same set of highly divergent samples. Selecting samples that are ideal examples of the four populations of interest was expected to produce optimally performing models when similar samples are input as unknowns, and a clear deterioration in performance when mixed samples are input. A test set numbering approximately ten percent of the training set, was also created. The ancestry models tested included one based on RMP-LR statistics, a MLR method, and a CHAID decision tree method. The RMP-LR method outperformed the other two, giving correct information in 98.4% of test set samples (e.g. samples are either correctly predicted for a single population or samples are determined inconclusive between two populations, one of which is the correct population); however, this method includes the most SNPs at 32. The other methods include far less SNPs, and this may be desirable so that a smaller, and more sensitive assay could be developed, or so that phenotype or individually identifying SNPs could be

incorporated into the same assay.

The phenotype model analysis was limited to eye color among European Americans, because this is the only population with sufficient sample size for testing, and because eye color is more straightforward and well-researched than hair or skin color (the latter would require a much larger sample set than that available). One published MLR model was evaluated, in addition to a CHAID decision tree model, with similar overall performance found between the two. The latter model performed better on intermediate eye color (green/hazel) while the former had higher performance for the blue/brown phenotypes. An approach of predicting a sample as "not blue" or "not brown" showed 100% correct results with the published model, and a 2% error rate with the CHAID model. The previously described diplotype approach was attempted with the CHAID eye color model, and again it did not improve results.

Determining the effect of incorporating other forensic markers into the ancestry prediction was the next phase of the research. In a forensic case, an STR profile would already exist prior to any SNP typing, so any level of ancestry information present should be added to the ancestry statistical model. After determining STR genotype frequencies for the four populations of interest, these were applied to the STR profiles of the test set and the STR RMP was multiplied into the SNP RMP, to produce an overall likelihood ratio. This was found to improve the SNP ancestry determination, by causing samples that had previously been inconclusive between two populations to become correctly predicted for a single population.

Mitochondrial and Y chromosome haplotype data for the test set was used similarly, by determining haplogroup frequencies in the populations of interest and

evaluating if inclusion of either of these markers would improve the ancestry determination. Both mtDNA and Y chromosome data inclusion caused a marked increase in the error rate, particularly for African American and Hispanic American individuals. This is not surprising as both of these markers are lineage specific, and these populations are admixed.

In the final phase of this research, a preliminary evaluation of an NGS method that theoretically is well-suited to forensic samples was performed. At this time, the technology is not sensitive enough, nor is the read length sufficient to provide a viable alternative to current forensic typing methods. However, exploring this high throughput genotyping method, and determining what would make it amenable the nuances of forensic evidence samples, is an important step toward a future forensic capability of generating vastly more information about an unknown sample that current forensic analyses allow.

The overall goal of this project was achieved. 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 American, 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 in a report format. Additionally, using a published model and calculator, eye color information can be provided.

Future projects and studies that could stem from this work include:

Explaining the results to practitioners and developing reporting guidelines

The first part of this project is to provide training to practitioners (forensic scientists) in this methodology to encourage implementation. The second part is to work with the practitioners to develop guidelines for relaying results of these analyses to investigators. It is imperative that forensic scientists properly relay the limitations of these tests, to prevent the results from misguiding an investigation.

Evaluating the limiting factors to implementing this assay in a forensic laboratory

The assay described herein has been optimized for forensic use, but there are a number of factors that may limit its use in a forensic casework laboratory. Even though the assay doesn't require any new equipment beyond what is typically already present in a forensic casework lab, reagents such as primers and the SNaPshot kit would need to be purchased. Current forensic DNA methods include primers in kit form, so practitioners are not accustomed to ordering primers. A different polymer may need to be purchased, depending on the polymer currently in use, and changing the polymer on the capillary electrophoresis instrument is an additional step that may be required. Evaluating all of these factors and providing information to the practitioners would facilitate implementation.

Evaluating mixture interpretation

As previously mentioned, this SNP assay would only be used after a forensic laboratory had already developed an STR profile. Since the sensitivity of STR analysis is in the same range as this SNP assay, the user would already know if a sample was a mixture of multiple individuals prior to SNP typing, and may know the approximate contributions of each individual in a two-person mixture. However, the SNP assay presents additional challenges in mixed samples: because the SNPs are biallelic, overlapping alleles are far

more prevalent than with an STR assay, and the SNaPshot fluorophores show inherently imbalanced peak heights, although predictably so. The previously discussed IrisPlex eye color model has been evaluated on mixed samples (Walsh 2011b), and it was found that while a mixture may be detectable with this six SNP assay, it was not possible to differentiate SNP profiles in mixed samples. Following a similar analysis for this 50 SNP assay would be an aid to laboratories developing interpretation guidelines.

Collecting additional samples with phenotype data and evaluating additional pigmentation models.

As indicated previously, an insufficient number of samples with corresponding phenotype data were collected to develop robust models for hair and skin color prediction in all populations, and eye color prediction in populations other than European American. Collecting additional samples in each population, and evaluating the robustness of the pigmentation models periodically (for example, after the collection of every 100 samples in each population) would greatly enhance this existing research.

Evaluating additional diplotype/haplotype combinations

Herein, one diplotype combination was evaluated and found to not improve ancestry or eye color prediction. Other groups of linked SNPs could be evaluated in an attempt to identify combinations that improve predictive power.

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, K. Miller, M. Wilson, K. Monson, B. Budowle. Characterization of the Caucasian haplogroups present in the SWGDAM forensic mtDNA dataset for 1771 human control region sequences. Scientific Working Group on DNA Analysis Methods. J. Foren. Sci. 47 (2002) 1215-1223.

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.

L. Andersson. The estimation of blood group gene frequencies: a note on the allocation method. Animal Blood Group and Biochem. Genet. 16 (1985) 1-7.

J. Asplen, "Next generation sequencing for forensics." *Forensic Magazine* April/May 2013. URL: http://www.forensicmag.com/article/next-generation-sequencing-forensics, Accessed May 6, 2013.

T.W. Athey. Haplogroup prediction from Y-STR values using an allele-frequency approach. J. Genet. Genealogy 1 (2005) 1-7.

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

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.

C.H. Brenner. Probable race of a stain donor. Proc. Seventh Int. Sym. Hum Id. (1997) 48-52.

M. H. Brilliant. Gene polymorphisms and human pigmentation. NIJ Grant# 2002-IJ-CX-K010 final report (2008). URL: https://www.ncjrs.gov/pdffiles1/nij/grants/223980.pdf.

M. Brión, J.J. Sanchez, K. Balogh, C. Thacker, A. Blanco-Verea, C. Børsting, *et al.* Introduction of an single nucleotide 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, R. Schoske, P.M. Vallone, J.W. Redman, M.C. Kline. Allele frequencies for the 15 autosomal STR loci on U.S. Caucasian, African American, and Hispanic populations. J. Foren. Sci. 48 (2003) 908-911.

J.M. Butler. Forensic DNA typing : biology, technology, and genetics of STR markers. 2nd ed., Elsevier Academic Press, Burlington, MA, 2005.

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.

D.C. Crawford, T. Bhangale, N. Li, G. Hellenthal, M.J. Rieder, D.A. Nickerson, M. Stephens. Evidence for substantial fine-scale variation in recombination rates across the human genome. Nat. Gen. 36 (2004) 700-706.

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.

I.W. Evett, R. Pinchin, C. Buffery. An investigation of the feasibility of inferring ethnic origin from DNA profiles. J. Foren. Sci. Soc. 4 (1992) 301-306.

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. Foren. Sci. Int. Genet. 2 (2008) 212-218.

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

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.

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.

M. Hashiyada, Y. Itakura, T. Nagashima, M. Nata, M. Funayama. Polymorphism of 17 STRs by multiplex analysis in Japanese population. Foren. Sci. Int. 133 (2003) 250-253.

M.F. Holick. Environmental factors that influence the cutaneous production of vitamin D^{1-3}. Am. J. Clin. Nutr. 61 (1995) 638S-645S.

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.

N. Jablonski. The evolution of human skin color. Annu. Rev. Anthropol. 33 (2004) 585-623.

N. Jablonski, G. Chaplin. The evolution of skin coloration. J. Hum. Evol. 39 (2000) 57-106.

G.V. Kass. An exploratory technique for investigating large quantities of categorical data. App. Stat. 29 (1980) 119-127.

M. Karayiorgou, T.J. Simon, J.A. Gogos. 22q11.2 microdeletions: linking DNA structural variation to brain dysfunction and schizophrenia. Nat. Rev. Neuro. 11 (2010) 402-416.

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.M. Schneider. DNA-based prediction of human externally visible characteristics in forensics: Motivations, scientific challenges, and ethical considerations. Foren. Sci. Int. Genet. 3 (2009) 154-161.

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. Population genetics of SNPs for forensic purposes. NIJ Grant # 2004-DN-BX-K025 final report (2008). URL: https://www.ncjrs.gov/pdffiles1/nij/grants/223982.pdf.

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) 1-13.

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.

B.E. Krenke, A. Tereba, S.J. Anderson, E. Buel, S. Culhane, C.J. Finis, *et al.* Validation of a 16-Locus Fluorescent Multiplex System. J. Foren. Sci. 47 (2002) 773-785.

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.

N. Li, M. Stephens. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. Genetics 165 (2003) 2213-2233.

Life Technologies Corporation. AmpF*l*STR® Identifiler® Plus PCR Amplification Kit User's Guide. Foster City, CA (2012) 102-113.

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.

W.F. Loomis. Skin-pigment regulation of vitamin-D biosynthesis in man. Science 157 (1967) 501-506.

R.S. McLaren, J. Patel, D.R. Storts, C.R. Hill, M.C. Kline, J.M. Butler. Improved primer pair for the SE33 locus in the PowerPlex ESI 17 Pro system. Promega Corporation Web site. http://www.promega.com/resources/articles/profiles-in-dna/2012/improved-primer-pair-for-the-se33-locus-in-the-powerplex-esi-17-pro-system/ Updated 2012. Accessed April 9, 2013.

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.

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

L.H. Miller, S.J. Mason, D.F. Clyde, M.H. McGinniss. The resistance factor to Plasmodium vivax in blacks. The Duffy-blood-group genotype, FyFy, New Eng. J. Med. 295 (1976) 302-304.

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, R.A. Kittles, M.D. Shriver. Implications of correlations between skin color and genetic ancestry for biomedical research. Nat. Gen. 36 (2004) S54-S60.

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

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/

Promega Corporation. PowerPlex® 16 and PowerPlex® ES Monoplex Systems Technical Bulletin. Madison, WI (2011) 3.

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

Y. Ruiz, C. Phillips, A. Gomez-Tato, J. Alvarez-Dios, M. Casares de Cal, R. Cruz, *et al*. Further development of forensic eye color predictive tests. Foren. Sci. Int. Genet. 7 (2013) 28-40.

O. Semino, C. Magri, G. Benuzzi, A. A. Lin, N. Al-Zahery, V. Battaglia, L. Maccioni, C. Triantaphyllidis, P. Shen, P. J. Oefner, L. A. Zhivotovsky, R King, A. Torroni, L. L. Cavalli-Sforza, P.A. Underhill, A. S. Santachiara-Benerecetti. Origin, Diffusion, and Differentiation of Y-Chromosome Haplogroups E and J: Inferences on the Neolithization of Europe and Later Migratory Events in the Mediterranean Area. Am. J. Hum. Genet. 74 (2004) 1023–1034.

N. Shengjie, Y. Jinyong, Y. Hongtao, Y. Yanmei, G. Tao, T. Wenru, *et al*. Genetic data of 15 loci in Chinese Yunnan Han population. Foren. Sci. Int. Genet. 3 (2008) e1-e3.

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. URL: http://www.snl.salk.edu/~shlens/pca.pdf.

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

G.N. Stamatas, B.Z. Zmudzka, N. Kollias, J.Z. Beer. Non-invasive measurements of skin pigmentation in situ. Pigment Cell Res. 17 (2004) 618-626.

M. Stephens, P. Donnelly. A comparison of Bayesian methods for haplotype reconstruction from population genotype data. Am. J. Hum. Genet. 73 (2003) 1162-1169.

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.

J.R. Suh, A.K. Herbig, P.J. Stover. New perspectives on folate catabolism. Annu. Rev. Nutr. 21 (2001) 255-282.

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.

A.C. Syvänen. 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.

G. Tully. Genotype versus phenotype: Human pigmentation. Foren. Sci. Int. Genet. 2 (2007) 105-110.

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. Foren. 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 Res. 3 (2012) 446-455.

(a) 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.

(b) S. Walsh, A. Lindenbergh, S.B. Zuniga, T. Sijen, P. de Knijff, M. Kayser, K.N. Ballantyne. Developmental validation of the IrisPlex system: determination of blue and brown iris colour for forensic intelligence. Foren. Sci. Int. Genet. 5 (2011) 464-471.

B. Wharton, N. Bishop. Rickets. Lancet 362 (2003) 1389-1400.

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

Appendices

Appendix Figure 1. Adult Sample Collection Assent Form (given to volunteer, or parent/legal guardian of child volunteer <6 years old)

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.

Appendix Figure 1 (continued). Adult Sample Collection Assent Form (given to volunteer, or parent/legal guardian of child volunteer <6 years old)

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·

Appendix Figure 2. 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.

APPROVED
The George Washington University
Institutional Review Board
·FWA00005945·

Appendix Figure 3. Sample Collection Questionnaire (completed by volunteer)

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

Appendix Figure 3 (continued). Sample Collection Questionnaire (completed by volunteer)

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:_____
_____

Appendix Figure 3 (continued). Sample Collection Questionnaire (completed by volunteer)

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 <u>natural</u> 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 <u>natural</u> hair color.

Source: killerstrands.blogspot.com

Appendix Figure 3 (continued). Sample Collection Questionnaire (completed by volunteer)

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: _____

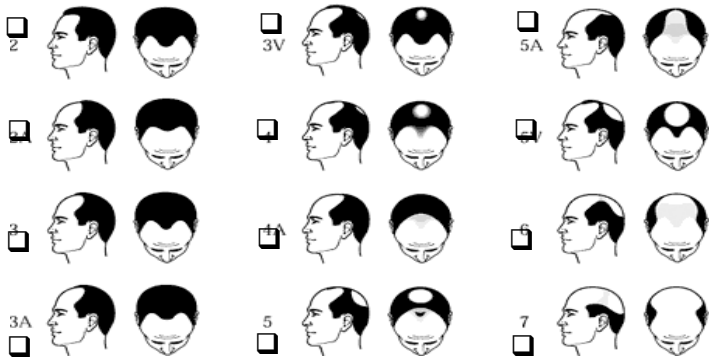
10. How would you classify the **thickness** of your head hair?

❑ 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?



Appendix Figure 3 (continued). Sample Collection Questionnaire (completed by volunteer)

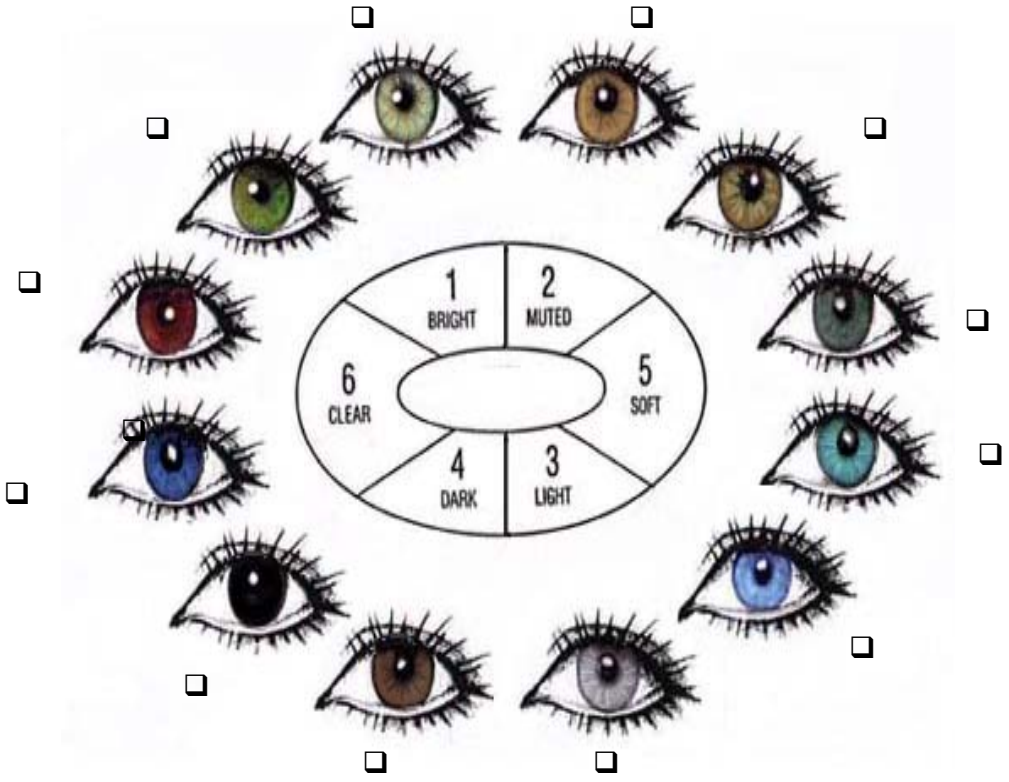
12. Is there a history of balding in your family? If so, please specify which side.

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

13. (a) What is your <u>natural</u> **eye** color?

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

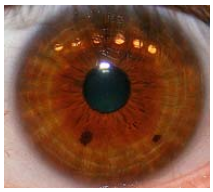
(b) Check the picture that best resembles your <u>natural</u> eye color.



Source: www.color-chart.org

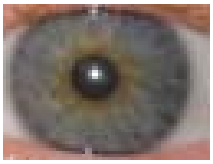
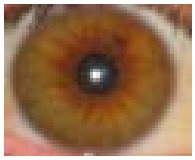
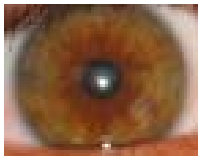
Appendix Figure 3 (continued). Sample Collection Questionnaire (completed by volunteer)

(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.
_____
_____
_____

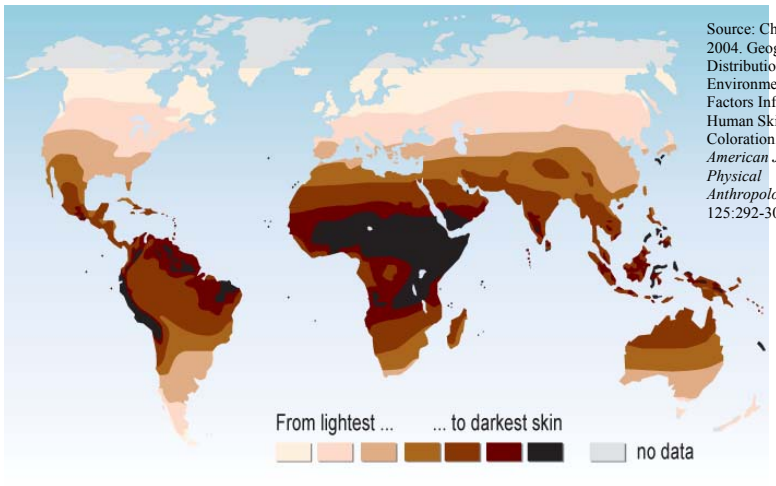
14. (a) In your own words, describe your <u>natural</u> **skin** color: _____

(b) What would you classify your <u>natural</u> **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: _____

Appendix Figure 3 (continued). Sample Collection Questionnaire (completed by volunteer)

(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 _____

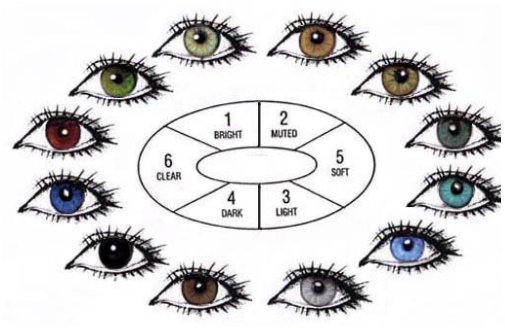
Maternal grandfather: eye color _____ skin color _____ hair color _____
Maternal grandmother: eye color _____ skin color _____ hair color _____

Thank you for your participation!

Appendix Figure 4. Sample Collection Checklist (completed by researcher)

CODE _____

Ancestry _____



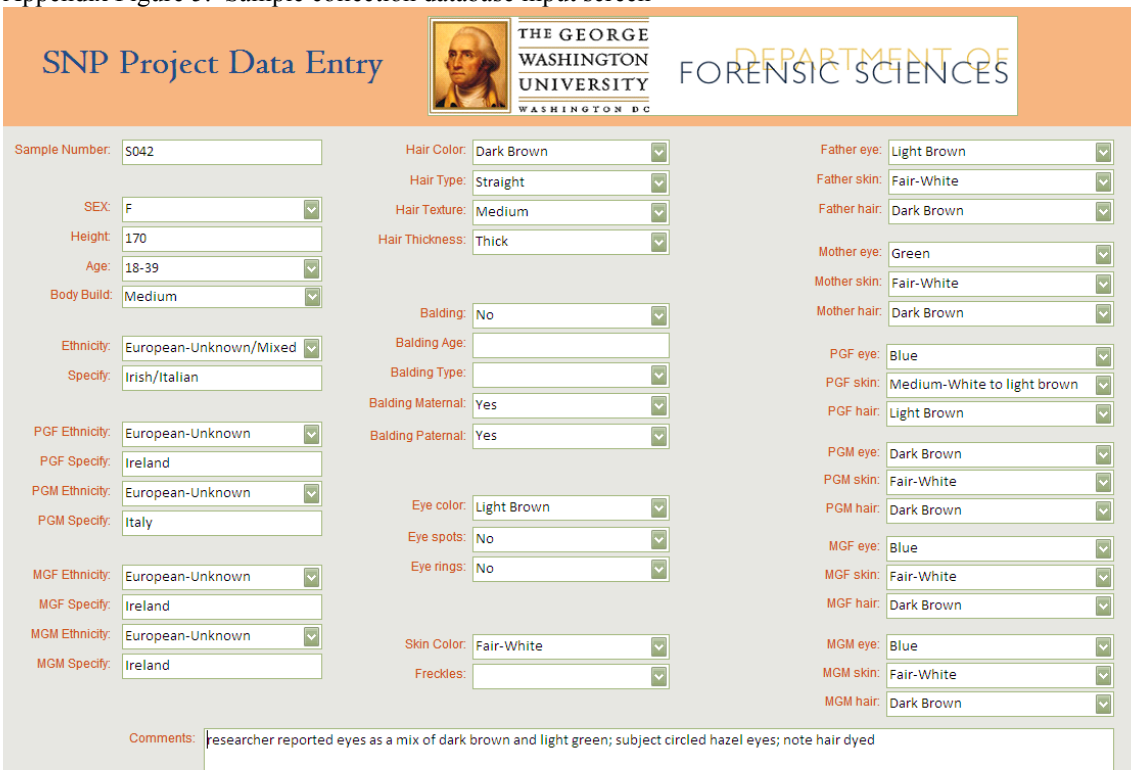
Special Eye features _____

Spectrophotometer Measurements checklist

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)

❑ Collect Swabs

Appendix Figure 5. Sample collection database input screen



Appendix Table 1a. Results of ancestry analyses for candidate SNP evaluation / reduction for the 50 selected SNPs. First column numbering represents order in final assay.

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

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 ancestry analyses for candidate SNP evaluation / reduction for the 49 eliminated SNPs, listed numerically by rs number.

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 2a. Results of pigmentation analyses for candidate SNP evaluation / reduction for the 50 selected SNPs. First column numbering represents order in final assay.

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

Appendix Table 2b. Results of pigmentation analyses for candidate SNP evaluation / reduction for the 49 eliminated SNPs, listed numerically by rs number.

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	

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 3a. SNP markers contained in the 50 SNP assay, Multiplex A, with molecular and PCR primer information.

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

Appendix Table 3b. SNP markers contained in the 50 SNP assay, Multiplex B, with molecular and PCR primer information.

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

Appendix Table 3c. SNP markers contained in the 50 SNP assay, Multiplex C, with molecular and PCR primer information.

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

Appendix Table 3d. SNP markers contained in the 50 SNP assay, Multiplexes A, B, and C, with SBE primer information.

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

NOTE: lower case “t” represents non-binding tail.

Appendix Table 4.

4a. Multiplex A Bin Set in GeneMapper

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

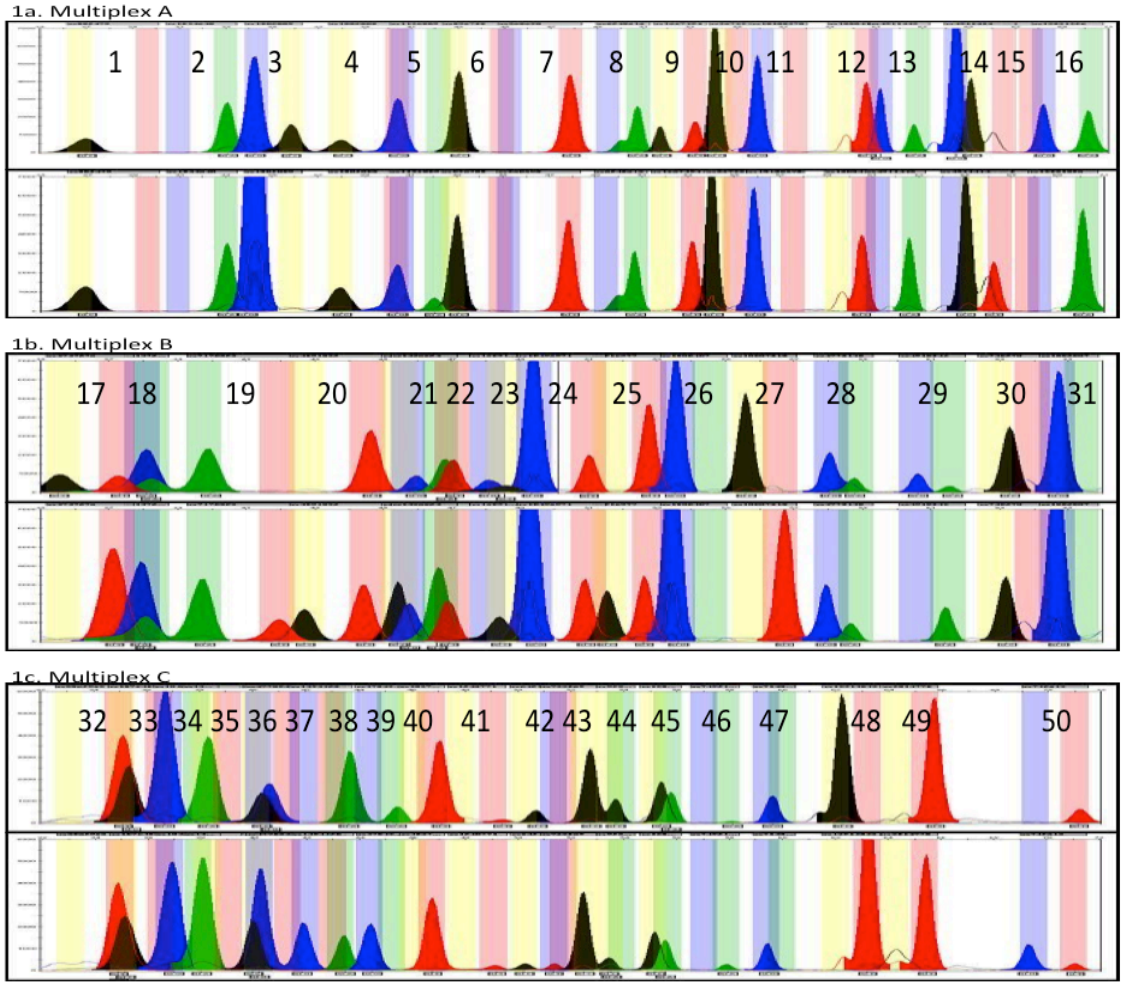
4b. Multiplex B Bin Set in GeneMapper

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

4c. Multiplex C Bin Set in GeneMapper

Position	SNP	Locus Range		Allele	Start	End	Color	Allele	Start	End	Color
32	rs2065982	32.56	35.43	C	32.56	33.56	Yellow	T	34.43	35.43	Red
33	rs1876482	34.56	36.96	C	34.56	35.56	Yellow	T	35.96	36.96	Red
34	rs1042602	36.30	39.54	G	36.30	37.30	Blue	T	38.54	39.54	Red
35	rs1344870	37.37	38.71	A	37.37	38.37	Green	C	37.71	38.71	Yellow
36	rs12203592	39.62	41.77	C	39.62	40.62	Yellow	T	40.77	41.77	Red
37	rs4778241	39.71	43.47	G	39.71	40.71	Blue	T	42.47	43.47	Red
38	rs1393350	41.42	43.76	G	41.42	42.42	Blue	A	42.76	43.76	Green
39	rs3784230	43.87	45.71	G	43.87	44.87	Blue	A	44.71	45.71	Green
40	rs3827760	45.52	47.23	C	45.52	46.52	Yellow	T	46.23	47.23	Red
41	rs1540771	47.36	49.56	C	47.36	48.36	Yellow	T	48.56	49.56	Red
42	rs6451772	49.71	52.20	C	49.71	50.71	Yellow	T	51.20	52.20	Red
43	rs722869	50.85	52.96	G	50.85	51.85	Blue	C	52.06	52.96	Yellow
44	rs952718	52.96	54.40	A	53.40	54.40	Green	C	52.96	53.96	Yellow
45	rs12896399	54.61	56.15	A	55.15	56.15	Green	C	54.61	55.61	Yellow
46	rs7495174	56.50	58.35	G	56.50	57.50	Blue	A	57.35	58.35	Green
47	rs714857	58.86	60.47	G	58.86	59.86	Blue	A	59.47	60.47	Green
48	rs12913832	61.50	63.69	C	61.50	62.50	Yellow	T	62.69	63.69	Red
49	rs2814778	63.75	65.85	C	63.75	64.75	Yellow	T	64.85	65.85	Red
50	rs735612	69.00	71.50	G	69.00	70.00	Blue	T	70.50	71.50	Red

Appendix Figure 6. Example electropherograms for the 50 SNP Assay.



1 rs885479	8 rs6548616	15 rs11547464	22 rs1426654	29 rs4918842	36 rs12203592	43 rs722869	50 rs735613
2 rs1834640	9 rs1667394	16 rs12821256	23 rs16891982	30 rs730570	37 rs4778241	44 rs952718	
3 rs1805009	10 rs26722	17 rs3737576	24 rs10496971	31 rs1805007	38 rs1393350	45 rs12896399	
4 rs1805008	11 rs10108270	18 rs1375164	25 rs916977	32 rs2065982	39 rs3784230	46 rs7495174	
5 rs1126809	12 rs1800414	19 rs7170852	26 rs1800407	33 rs1876482	40 rs3827760	47 rs714857	
6 rs896788	13 rs4911442	20 rs4891825	27 rs10007810	34 rs1042602	41 rs1540771	48 rs12913832	
7 rs260690	14 rs4911414	21 rs2714758	28 rs4778138	35 rs1344870	42 rs6451772	49 rs2814778	

Appendix Table 5. SNP loci used in the different ancestry and eye color models. First column numbering represents order in final assay.

		32 SNP RMP-LR Ancestry	31 SNP+Diplotype RMP-LR Ancestry	7 SNP MLR Ancestry	5 SNP CHAID Ancestry	4 SNP+Diplotype CHAID Ancestry	Irisplex	5 SNP CHAID Eye Color	3 SNP+Diplotype CHAID Eye Color
1	rs885479	✓	✓						
2	rs1834640			✓					
3	rs1805009								
4	rs1805008								
5	rs1126809								
6	rs896788	✓	✓						
7	rs260690								
8	rs6548616	✓	✓						
9	rs1667394								
10	rs26722								
11	rs10108270	✓	✓						
12	rs1800414				✓				
13	rs4911442	✓	✓						
14	rs4911414	✓	✓						
15	rs11547464								
16	rs12821256	✓	✓						
17	rs3737576	✓	✓						
18	rs1375164	✓	✓	✓		✓			
19	rs7170852								
20	rs4891825	✓	✓						
21	rs2714758	✓	✓	✓					
22	rs1426654	✓	✓						
23	rs16891982	✓	✓	✓	✓	✓	✓		
24	rs10496971	✓	✓						
25	rs916977		✓*			✓*			✓*
26	rs1800407						✓	✓	✓
27	rs10007810	✓	✓					✓	✓
28	rs4778138			✓					
29	rs4918842	✓	✓						
30	rs730570	✓	✓						
31	rs1805007								
32	rs2065982	✓	✓						
33	rs1876482	✓	✓					✓	
34	rs1042602	✓	✓						
35	rs1344870	✓	✓						
36	rs12203592						✓		
37	rs4778241								
38	rs1393350						✓		
39	rs3784230	✓	✓						
40	rs3827760	✓	✓	✓					
41	rs1540771	✓	✓						
42	rs6451722	✓	✓						
43	rs722869	✓	✓					✓	✓
44	rs952718	✓	✓						
45	rs12896399	✓	✓				✓		
46	rs7495174								
47	rs714857	✓	✓						
48	rs12913832	✓	✓*	✓	✓	✓*	✓	✓	✓*
49	rs2814778	✓	✓		✓	✓			
50	rs735612	✓	✓		✓	✓			

* = used in diplotype

Appendix Table 6. Linkage disequilibrium evaluation for ancestry model. Calculations performed using WGAviewer (r^2 and D') and Phase (results previously indicated in Table). Highlighted SNPs were eliminated for possible linkage.

Chr	Gene Region	SNP	distance	pop	r^2	D'
1		rs3737576	57,465,120	not included in LD analysis		
		rs2814778				
		rs896788	10,213,413	not included in LD analysis		
		rs1876482				
			92,151,033	CHB	0.03	0.526
				JPT	0.004	0.265
		rs3827760				
	EDAR		66,137	CHB	0.001	1
				JPT	0.002	0.076
		rs260690				
2				CEU	0.008	1
			36,190,205	CHB	0.004	1
				JPT	0.024	1
				YRI	0.01	0.279
		rs10496971				
				CEU	0	0.019
			70,118,681	CHB	0.001	0.221
				JPT	0.025	0.29
				YRI	0.001	0.14
		rs952718				
		rs1344870				
				CHB	0	0.022
3			58,092,174	JPT	0.004	0.094
				YRI	0.004	1
		rs6548616				
4		rs10007810				
		rs16891982				
	SLC45A2		12,177	CHB	0.008	1
		rs26722				
5				CHB	0.011	0.143
			9,747,508	JPT	0.006	0.161
				YRI	0.006	1
		rs6451722				
		rs12203592				
6	IRF4		69,712	CEU	0.049	0.578
		rs1540771				
8		rs10108270				
10		rs4918842				
		rs714857				
			72,937,307	CEU	0.008	0.383
		rs1042602				
11			99,350	CEU	0.169	1
	TYR	rs1393350				
			6,915	CEU	0.857	1
		rs1126809				
12		rs12821256				
13		rs2065982				
		rs12896399				
				CEU	0.016	0.286
			4,503,342	CHB	0.007	0.39
				JPT	0.013	0.265
				YRI	0.001	1
		rs722869				
				CEU	0.007	0.521
14			3,865,885	CHB	0.038	1
				JPT	0.002	0.233
				YRI	0.013	1
		rs730570				
				CEU	0.029	0.333
			4,536,165	CHB	0.026	1
				JPT	0.072	0.308
		rs3784230				

Appendix Table 6 (continued). Linkage disequilibrium evaluation for ancestry model.

Chr	Gene Region	SNP	distance	pop	r^2	D'		pop	r^2	D'
		rs2714758								
			2,717,500	CHB	0.02	1				
				JPT	0.025	1				
		rs1800414								
			33,281	not included in LD analysis, Phase shows linkage in EU, AF, AS						
		rs1800407						CHB	0.041	1
			61,494	not included in LD analysis, Phase shows recombination in EU, AS				JPT	0.069	0.742
		rs1375164								
				CEU	0.206	0.658				
			44,008	CHB	0.196	1	Phase shows linkage in EU, AF, AS			
				JPT	0.322	0.852				
				YRI	0.003	1				
		rs4778138								
				CEU	0.147	0.456				
			2,893	CHB	0.424	0.759				
				JPT	0.32	0.849	Phase shows recombination in EU			
				YRI	0.001	0.081		CEU	0.214	0.746
	OCA2 / HERC2	rs4778241								
			5,525	not included in LD analysis, Phase shows recombination in AF, AS						
		rs7495174								
			21,380	not included in LD analysis, Phase shows recombination in EU, AF, AS						
		rs12913832								
			62,368	CEU	0.452	0.912	Phase shows linkage in EU, AF, AS			
15		rs7170852								
				CEU	0.79	0.923				
			85,378	CHB	0.811	0.931	Phase shows linkage in EU, AF, AS			
				JPT	0.916	1				
				YRI	0.473	1				
		rs916977								
				CEU	1	1				
			16,818	CHB	0.934	1				
				JPT	1	1	Phase shows linkage in EU, AF, AS			
				YRI	1	1				
		rs1667394								
				CEU	0.003	0.11				
			5,546,460	CHB	0.029	0.458				
				JPT	0.013	1				
				YRI	0	0.024				
		rs735612								
				CHB	0.014	0.196				
			14,315,523	JPT	0.003	1				
				YRI	0.001	0.216				
		rs1834640								
				CHB	0.115	1				
	SLC24A5		34,319	JPT	0.241	1				
				YRI	1	1				
		rs1426654								
		rs11547464								
			26							
		rs1805007								
			27	No neighboring SNPs included in LD analysis						
16	MC1R	rs1805008		Phase shows all linked in EU, AF, AS				CEU	0.014	1
			10							
		rs885479								
			392							
		rs1805009								
18		rs4891825								
		rs4911414								
20			625,602	CEU	0.07	0.585				
		rs4911442								

Appendix Table 7. Training set allele frequencies by population. First column numbering represents order in final assay. No more than two alleles at each locus were found in the data comprising training set.

		European (N=266)				East Asian (N=250)				African American (N=250)				Hispanic (N=234)			
		A	C	G	T	A	C	G	T	A	C	G	T	A	C	G	T
1	rs885479		0.92		0.08		0.35		0.65		0.99		0.01		0.63		0.37
2	rs1834640	1.00		0.00		0.06		0.94		0.05		0.95		0.65		0.35	
3	rs1805009		0.02	0.98			0.01	1.00			0.01	1.00			0.01	0.99	
4	rs1805008		0.90		0.10		1.00		0.00		1.00		0.00		0.99		0.01
5	rs1126809	0.23		0.77		0.01		1.00		0.01		0.99		0.11		0.89	
6	rs896788		0.82		0.18		0.32		0.68		0.67		0.33		0.66		0.34
7	rs260690			0.08	0.92			0.96	0.04			0.66	0.34			0.50	0.50
8	rs6548616	0.80		0.20		0.87		0.13		0.93		0.07		0.82		0.18	
9	rs1667394		0.01		1.00		0.86		0.14		0.97		0.03		0.43		0.57
10	rs26722		1.00		0.00		0.60		0.40		0.95		0.05		0.67		0.33
11	rs10108270			0.70	0.30			0.69	0.31			0.07	0.93			0.81	0.19
12	rs1800414		0.01		1.00		0.68		0.32		0.01		1.00		0.01		1.00
13	rs4911442	0.89		0.11		1.00		0.00		0.99		0.01		0.97		0.03	
14	rs4911414			0.64	0.36			0.80	0.20			0.86	0.14			0.73	0.27
15	rs11547464		1.00		0.00		1.00		0.01		1.00		0.01		0.98		0.02
16	rs12821256	0.84		0.16		1.00		0.01		1.00		0.00		0.97		0.03	
17	rs3737576		0.05		0.95		0.13		0.87		0.06		0.94		0.19		0.81
18	rs1375164	0.13		0.87		0.94		0.06		0.99		0.01		0.44		0.56	
19	rs7170852	0.01			0.99	0.84			0.16	0.86			0.14	0.44			0.56
20	rs4891825		0.11		0.89		0.01		1.00		0.85		0.15		0.14		0.86
21	rs2714758		0.05		0.95		0.02		0.98		0.85		0.15		0.08		0.92
22	rs1426654	1.00		0.00		0.01		0.99		0.06		0.94		0.54		0.46	
23	rs16891982		0.99	0.01			0.00	1.00			0.04	0.96			0.43	0.57	
24	rs10496971			0.05	0.95			0.80	0.20			0.05	0.95			0.28	0.72
25	rs916977		1.00		0.01		0.14		0.86		0.04		0.96		0.57		0.43
26	rs1800407	0.04		0.96		0.01		1.00		0.00		1.00		0.03		0.97	
27	rs10007810		0.80		0.20		0.91		0.09		0.06		0.94		0.76		0.24
28	rs4778138	0.97		0.03		0.16		0.84		0.22		0.78		0.75		0.25	
29	rs4918842	0.91		0.09		0.57		0.43		0.93		0.07		0.54		0.46	
30	rs730570		0.14		0.86		0.80		0.20		0.76		0.24		0.59		0.41
31	rs1805007	0.09		0.91		0.01		1.00		0.00		1.00		0.03		0.97	
32	rs2065982		0.05		0.95		0.74		0.26		0.07		0.93		0.43		0.57
33	rs1876482		0.94		0.06		0.25		0.75		1.00		0.01		0.78		0.22
34	rs1042602			0.67	0.33			1.00	0.00			0.98	0.02			0.74	0.26
35	rs1344870	0.98	0.02			0.73	0.27			0.94	0.06			0.54	0.46		
36	rs12203592		0.86		0.14		1.00		0.01		0.99		0.01		0.93		0.07
37	rs4778241			0.97	0.03			0.12	0.88			0.37	0.63			0.57	0.43
38	rs1393350	0.24		0.76		0.01		1.00		0.01		0.99		0.11		0.89	
39	rs3784230	0.62		0.38		0.93		0.07		0.04		0.96		0.54		0.46	
40	rs3827760		0.01		0.99		0.89		0.11		0.00		1.00		0.44		0.56
41	rs1540771		0.49		0.51		0.73		0.27		0.91		0.09		0.47		0.53
42	rs6451722		0.81		0.19		0.81		0.19		0.14		0.86		0.82		0.18
43	rs722869		0.91	0.09			0.12	0.88			0.89	0.11			0.61	0.39	
44	rs952718	0.06	0.94			0.06	0.94			0.69	0.31			0.31	0.69		
45	rs12896399	0.46	0.54			0.34	0.66			0.03	0.97			0.25	0.75		
46	rs7495174	1.00		0.00		0.22		0.78		0.81		0.19		0.86		0.14	
47	rs714857	0.04		0.96		0.63		0.37		0.79		0.21		0.24		0.76	
48	rs12913832		0.95		0.05		0.01		1.00		0.01		0.99		0.15		0.85
49	rs2814778		0.01		0.99		0.00		1.00		0.96		0.04		0.06		0.94
50	rs735612			0.35	0.65			0.04	0.96			0.57	0.43			0.56	0.44

* = used in diplotype

Appendix Table 8. STR allele frequency data used for RMP/LR calculation. Minimum allele frequency = 5/2N. See main text for origin of frequencies.

D3S1358	AA	EU	HI-NA	EA
<12	0.0085	0.0075	0.0133	0.0050
12	0.0085	0.0075	0.0133	0.0050
13	0.0085	0.0075	0.0133	0.0050
14	0.1047	0.1303	0.0737	0.0291
15	0.3037	0.2578	0.3643	0.3716
15.2	0.0085	0.0075	0.0133	0.1695
16	0.3104	0.2404	0.2943	0.2667
17	0.1999	0.1985	0.1546	0.1217
17.1	0.0085	0.0075	0.0133	0.0065
18	0.0636	0.1584	0.0995	0.0340
19	0.0092	0.0110	0.0138	0.0050
>19	0.0085	0.0075	0.0133	0.0035
VWA				
11	0.0085	0.0075	0.0133	0.0050
12	0.0085	0.0075	0.0133	0.0050
13	0.0113	0.0075	0.0133	0.0015
14	0.0747	0.0886	0.0665	0.2218
15	0.1932	0.1121	0.1129	0.0296
16	0.2578	0.2168	0.3123	0.2011
17	0.2240	0.2630	0.2562	0.2567
18	0.1511	0.2125	0.1700	0.1981
19	0.0646	0.0936	0.0656	0.0743
20	0.0178	0.0098	0.0149	0.0166
21	0.0085	0.0075	0.0133	0.0050
>21	0.0085	0.0075	0.0133	0.0050
D8S1179				
<9	0.0085	0.0175	0.0133	0.0050
9	0.0085	0.0098	0.0133	0.0035
10	0.0264	0.0992	0.0772	0.1269
11	0.0421	0.0716	0.0499	0.1003
12	0.1371	0.1627	0.1263	0.1119
13	0.2248	0.3151	0.3239	0.2231
14	0.3006	0.1898	0.2728	0.2041
15	0.1929	0.1065	0.1106	0.1490
16	0.0581	0.0291	0.0216	0.0748
17	0.0106	0.0075	0.0133	0.0091
>17	0.0085	0.0075	0.0133	0.0050

FGA	AA	EU	HI-NA	EA
<18	0.0085	0.0075	0.0133	0.0050
18	0.0085	0.0266	0.0134	0.0286
18.2	0.0130	0.0075	0.0133	0.0050
19	0.0646	0.0573	0.0789	0.0757
19.2	0.0085	0.0075	0.0133	0.0050
19.3	0.0085	0.0075	0.0133	0.0050
20	0.0630	0.1330	0.0954	0.0732
20.2	0.0085	0.0075	0.0133	0.0050
21	0.1225	0.1771	0.1476	0.1268
21.2	0.0085	0.0075	0.0133	0.0045
22	0.2059	0.1941	0.1424	0.1791
22.2	0.0085	0.0125	0.0133	0.0035
22.3	0.0085	0.0075	0.0133	0.0050
23	0.1605	0.1430	0.1382	0.2021
23.1	0.0085	0.0075	0.0133	0.0050
23.2	0.0085	0.0075	0.0133	0.0035
23.3	0.0085	0.0075	0.0133	0.0050
24	0.1486	0.1368	0.1535	0.1620
24.2	0.0085	0.0075	0.0133	0.0050
24.3	0.0085	0.0075	0.0133	0.0050
25	0.1019	0.0785	0.1346	0.0788
25.1	0.0085	0.0075	0.0133	0.0050
25.2	0.0085	0.0075	0.0133	0.0045
25.3	0.0085	0.0075	0.0133	0.0050
26	0.0580	0.0251	0.0558	0.0377
26.2	0.0085	0.0075	0.0133	0.0035
27	0.0206	0.0076	0.0267	0.0070
27.2	0.0085	0.0075	0.0133	0.0030
28	0.0130	0.0075	0.0133	0.0015
29	0.0085	0.0075	0.0133	0.0035
29.2	0.0085	0.0075	0.0133	0.0030
30	0.0085	0.0075	0.0133	0.0050
>30	0.0092	0.0075	0.0133	0.0035

Appendix Table 8 (continued). STR allele frequency data used for RMP/LR calculation. Minimum allele frequency = 5/2N. See main text for origin of frequencies.

D21S11	AA	EU	HI-NA	EA
<24.2	0.0085	0.0075	0.0133	0.0030
24.2	0.0085	0.0075	0.0133	0.0050
24.3	0.0085	0.0075	0.0133	0.0050
25	0.0085	0.0075	0.0133	0.0050
25.2	0.0085	0.0075	0.0133	0.0050
26	0.0085	0.0075	0.0133	0.0050
26.2	0.0085	0.0075	0.0133	0.0050
27	0.0642	0.0359	0.0204	0.0040
28	0.2439	0.1633	0.0834	0.0376
28.2	0.0085	0.0075	0.0133	0.0065
29	0.1957	0.2000	0.1932	0.2623
29.2	0.0085	0.0075	0.0133	0.0030
29.3	0.0085	0.0075	0.0133	0.0050
30	0.1732	0.2651	0.2990	0.3119
30.2	0.0120	0.0305	0.0289	0.0091
30.3	0.0085	0.0075	0.0133	0.0050
31	0.0804	0.0773	0.0689	0.1043
31.1	0.0085	0.0075	0.0133	0.0050
31.2	0.0634	0.0968	0.1286	0.0632
32	0.0106	0.0112	0.0155	0.0201
32.1	0.0085	0.0075	0.0133	0.0050
32.2	0.0584	0.0778	0.1184	0.1284
32.3	0.0085	0.0075	0.0133	0.0050
33	0.0085	0.0075	0.0133	0.0066
33.1	0.0085	0.0075	0.0133	0.0050
33.2	0.0364	0.0295	0.0330	0.0391
33.3	0.0085	0.0075	0.0133	0.0050
34	0.0113	0.0075	0.0133	0.0035
34.1	0.0085	0.0075	0.0133	0.0050
34.2	0.0085	0.0075	0.0133	0.0045
35	0.0262	0.0075	0.0133	0.0050
35.2	0.0085	0.0075	0.0133	0.0030
36	0.0092	0.0075	0.0133	0.0050
>36	0.0085	0.0075	0.0133	0.0050

D5S818				
<7	0.0085	0.0075	0.0133	0.0050
7	0.0085	0.0075	0.0891	0.0111
8	0.0513	0.0075	0.0133	0.0075
9	0.0279	0.0458	0.0516	0.0887
9.2	0.0085	0.0075	0.0133	0.0050
10	0.0686	0.0527	0.0515	0.2006
11	0.2440	0.3768	0.3841	0.3016
12	0.3586	0.3682	0.2920	0.2257
13	0.2269	0.1479	0.1150	0.1489
14	0.0199	0.0075	0.0133	0.0141
15	0.0085	0.0075	0.0133	0.0035
>15	0.0085	0.0075	0.0133	0.0035

D18S51	AA	EU	HI-NA	EA
<11	0.0085	0.0083	0.0133	0.0015
11	0.0085	0.0143	0.0144	0.0035
12	0.0740	0.1330	0.1235	0.0492
13	0.0482	0.1269	0.1158	0.1781
13.2	0.0085	0.0075	0.0133	0.0050
14	0.0703	0.1523	0.1879	0.2092
14.2	0.0085	0.0075	0.0133	0.0050
15	0.1779	0.1476	0.1583	0.1795
15.2	0.0085	0.0075	0.0133	0.0050
16	0.1617	0.1376	0.1202	0.1304
16.2	0.0085	0.0075	0.0133	0.0050
17	0.1671	0.1246	0.1390	0.0803
18	0.1210	0.0767	0.0538	0.0471
19	0.0796	0.0412	0.0399	0.0467
20	0.0565	0.0196	0.0242	0.0276
20.2	0.0085	0.0075	0.0133	0.0050
21	0.0155	0.0090	0.0138	0.0216
21.2	0.0085	0.0075	0.0133	0.0050
22	0.0085	0.0075	0.0133	0.0151
>22	0.0085	0.0075	0.0133	0.0105

D13S317				
<8	0.0085	0.0075	0.0133	0.0030
8	0.0319	0.1174	0.0891	0.2769
8.1	0.0085	0.0075	0.0133	0.0050
9	0.0291	0.0762	0.1831	0.1525
10	0.0304	0.0477	0.1092	0.1189
11	0.2756	0.3185	0.2368	0.2333
12	0.4431	0.2780	0.2200	0.1664
13	0.1496	0.1179	0.0994	0.0381
13.3	0.0085	0.0075	0.0133	0.0050
14	0.0392	0.0426	0.0602	0.0106
15	0.0085	0.0075	0.0133	0.0030
>15	0.0085	0.0075	0.0133	0.0050

D7S820				
6	0.0085	0.0075	0.0133	0.0050
6.3	0.0085	0.0075	0.0133	0.0050
7	0.0115	0.0155	0.0161	0.0050
8	0.2119	0.1579	0.1230	0.1415
8.1	0.0085	0.0075	0.0133	0.0050
8.2	0.0085	0.0075	0.0133	0.0050
9	0.1232	0.1766	0.0848	0.0437
9.1	0.0085	0.0075	0.0133	0.0050
9.3	0.0085	0.0075	0.0133	0.0040
10	0.3378	0.2576	0.2623	0.1890
10.1	0.0085	0.0075	0.0133	0.0050
10.3	0.0085	0.0075	0.0133	0.0050
11	0.2010	0.1938	0.2776	0.3631
11.3	0.0085	0.0075	0.0133	0.0050
12	0.0974	0.1568	0.2022	0.2177
13	0.0147	0.0361	0.0348	0.0411
14	0.0085	0.0076	0.0133	0.0025
>14	0.0085	0.0075	0.0133	0.0050

Appendix Table 8 (continued). STR allele frequency data used for RMP/LR calculation. Minimum allele frequency = 5/2N. See main text for origin of frequencies.

D16S539	AA	EU	HI-NA	EA
<8	0.0085	0.0075	0.0133	0.0050
8	0.0356	0.0176	0.0184	0.0040
9	0.1933	0.1088	0.1184	0.3063
9.3	0.0085	0.0075	0.0133	0.0030
10	0.1126	0.0560	0.1431	0.1584
11	0.3166	0.3203	0.2905	0.2334
11.3	0.0085	0.0075	0.0133	0.0050
12	0.1919	0.3142	0.2754	0.2067
13	0.1333	0.1568	0.1363	0.0803
14	0.0162	0.0251	0.0199	0.0096
15	0.0085	0.0075	0.0133	0.0010
THO1				
<5	0.0085	0.0075	0.0133	0.0050
5	0.0085	0.0075	0.0133	0.0050
6	0.1173	0.2185	0.2161	0.1713
7	0.4248	0.2039	0.3517	0.2739
8	0.2007	0.0993	0.0776	0.0637
8.3	0.0085	0.0075	0.0133	0.0050
9	0.1371	0.1380	0.1181	0.4390
9.3	0.1106	0.3294	0.2283	0.0331
10	0.0099	0.0075	0.0133	0.0181
>10	0.0085	0.0075	0.0133	0.0035
TPOX				
<6	0.0085	0.0075	0.0133	0.0035
6	0.0841	0.0075	0.0133	0.0050
7	0.0197	0.0075	0.0133	0.0050
8	0.3667	0.5340	0.4491	0.4736
9	0.1948	0.1175	0.0728	0.1279
10	0.0907	0.0495	0.0375	0.0271
11	0.2167	0.2512	0.3134	0.3334
12	0.0259	0.0442	0.1205	0.0336
13	0.0085	0.0075	0.0133	0.0020
>13	0.0085	0.0075	0.0133	0.0015
CSF1PO				
<6	0.0085	0.0075	0.0133	0.0050
6	0.0085	0.0075	0.0133	0.0050
7	0.0496	0.0075	0.0143	0.0065
8	0.0678	0.0075	0.0133	0.0050
9	0.0374	0.0146	0.0379	0.0431
10	0.2679	0.2296	0.2573	0.2418
10.3	0.0085	0.0075	0.0133	0.0050
11	0.2275	0.3096	0.2652	0.2157
11.1	0.0085	0.0075	0.0133	0.0050
12	0.2947	0.3446	0.3603	0.4087
12.1	0.0085	0.0075	0.0133	0.0050
13	0.0451	0.0846	0.0573	0.0657
14	0.0099	0.0112	0.0133	0.0171
15	0.0085	0.0075	0.0133	0.0015

D2S1338	AA	EU	HI-NA	EA
15	0.0085	0.0076	0.0132	0.0030
16	0.0556	0.0402	0.0288	0.0106
17	0.1034	0.1777	0.1692	0.0872
18	0.0475	0.0710	0.0707	0.1258
19	0.1448	0.1258	0.2341	0.1981
20	0.0816	0.1461	0.1236	0.1014
21	0.1421	0.0334	0.0285	0.0266
22	0.1309	0.0391	0.0956	0.0642
23	0.1094	0.1163	0.1092	0.1676
24	0.0890	0.1203	0.0780	0.1274
25	0.0766	0.0995	0.0527	0.0647
26	0.0158	0.0286	0.0132	0.0211
27	0.0085	0.0076	0.0132	0.0040
28	0.0085	0.0076	0.0132	0.0035
D19S433				
9	0.0085	0.0076	0.0132	0.0050
10	0.0127	0.0076	0.0132	0.0035
11	0.0667	0.0076	0.0132	0.0040
11.2	0.0085	0.0076	0.0132	0.0050
12	0.1109	0.0792	0.0525	0.0497
12.2	0.0490	0.0076	0.0167	0.0070
13	0.2722	0.2712	0.1961	0.2984
13.2	0.0547	0.0126	0.0909	0.0397
14	0.2166	0.3550	0.3136	0.3043
14.2	0.0605	0.0133	0.0432	0.0888
15	0.0628	0.1548	0.1297	0.0532
15.2	0.0468	0.0311	0.0771	0.1189
16	0.0169	0.0458	0.0345	0.0076
16.2	0.0254	0.0161	0.0242	0.0251
17	0.0085	0.0076	0.0132	0.0030
17.2	0.0085	0.0076	0.0185	0.0030
18	0.0085	0.0076	0.0132	0.0030
18.2	0.0085	0.0076	0.0132	0.0035

Appendix Table 9. mtDNA and Y chromosome haplogroup frequency data used for combined marker statistical analysis. See main text for origin of frequencies.

mtDNA Haplogroup Frequencies

	AA	EU	HI/NA	AS
A	0.0106	0.0028	0.3893	0.0618
B	0.0062	0.0028	0.2036	0.1653
C	0.0044	0.0028	0.1660	0.0289
D	0.0044	0.0028	0.0346	0.2648
F	0.0044	0.0028	0.0049	0.1380
G	0.0044	0.0028	0.0049	0.0570
H	0.0284	0.4570	0.0455	0.0040
I	0.0044	0.0190	0.0049	0.0040
J	0.0044	0.1000	0.0198	0.0040
K	0.0044	0.0890	0.0217	0.0040
L	0.9140	0.0028	0.0791	0.0040
M	0.0142	0.0190	0.0049	0.1774
N	0.0044	0.0028	0.0049	0.0546
R	0.0044	0.0028	0.0049	0.0257
T	0.0044	0.1050	0.0109	0.0040
U	0.0133	0.1560	0.0188	0.0040
V	0.0044	0.0190	0.0049	0.0040
W	0.0044	0.0190	0.0049	0.0040
X	0.0044	0.0190	0.0049	0.0040
Y	0.0044	0.0028	0.0049	0.0088
Z	0.0044	0.0028	0.0049	0.0112

Minimum allele frequency = 5/N

Y Chromosome Haplogroup Frequencies

	AF	EU	NA	EA
A	0.0158	0.0010	0.0038	0.0017
B	0.0284	0.0012	0.0038	0.0017
E	0.9117	0.0723	0.0075	0.0017
G	0.0158	0.0378	0.0053	0.0017
H	0.0158	0.0035	0.0038	0.0017
I	0.0158	0.2008	0.0060	0.0017
J	0.0158	0.0661	0.0136	0.0017
K	0.0442	0.5973	0.9426	0.6769
Q	0.0158	0.0012	0.8415	0.0083
R	0.0379	0.5491	0.0974	0.0017

Minimum allele frequency = 5/N

A not be used because all populations are 5/N

H not be used because EU frequency is less than the minimum frequency for AF/NA

Appendix Table 10. Linkage disequilibrium analysis for integrating STR data into SNP ancestry model. Calculations performed with WGAviewer.

Chr	Locus	position/range	pop	r2	D'
2	**TPOX**	1,483,854			
		1,484,085			
	rs7591940	1,546,857			
			CEU	0.001	0.099
			CHB	0.058	0.412
			JPT	0.115	0.411
	rs771313	7,011,166			
	rs896788	7,149,155			
	rs952718	215,888,624			
			JPT	0.008	0.195
			YRI	0.013	0.138
	rs3731861	218,899,500			
	D2S1338	218,879,515			
		218,879,706			
3	rs1344870	21,307,401			
			JPT	0.013	0.224
	rs11130040	45,024,190			
	D3S1358	45,582,205			
		45,582,335			
	rs2245705	45,724,726			
			CEU	0.001	0.066
			CHB	0.02	0.718
			JPT	0.001	0.033
	rs6548616	79,399,575			
4	rs10007810	41,554,364			
			CEU	0.051	0.409
			CHB	0.025	0.298
			JPT	0.072	1
			YRI	0	1
	rs951728	155,504,502			
	FGA	155,508,848			
		155,509,043			
5	rs16891982	33,951,693			
	rs6451722	43,711,378			
			CEU	0.011	0.414
			CHB	0.001	0.05
			JPT	0.01	0.732
			YRI	0.006	0.093
	rs25759	123,107,433			
	D5S818	123,111,185			
		123,111,333			
	CSF1PO	149,455,735			
		149,456,053			

Chr	Locus	position/range	pop	r2	D'
7	**D7S820**				
8	rs10108270	4,190,793			
			CEU	0.022	0.248
			CHB	0.03	0.526
			JPT	0.008	1
			YRI	0.005	1
	rs9649956	125,906,763			
	D8S1179	125,907,080			
		125,907,260			
11	**TH01**	2,192,277			
		2,192,522			
	rs11021705	2,193,265			
			JPT	0	0.053
			YRI	0.047	0.33
	rs714857	15,974,389			
12	**vWA**	6,093,104			
		6,093,254			
	rs11064114	6,236,300			
			JPT	0.014	1
	rs10858978	89,327,984			
	rs12821256	89,328,335			
13	rs2065982	34,864,240			
	rs7320715	34,865,449			
			CHB	0	0.026
			JPT	0.001	0.052
	rs1119122	82,721,127			
	D13S317	82,722,059			
		82,722,243			
16	**D16S539**	86,386,257			
		86,386,413			
	rs16943289	86,388,211			
			CHB	0.08	1
			JPT	0.013	0.349
	rs885479	89,986,154			
18	**D18S51**	60,948,844			
		60,949,149			
	rs9946533	60,949,983			
			YRI	0.001	0.221
	rs4891825	67,867,633			
19	**D19S433**				
21	**D21S11**				

KEY:
Blue Text = tag SNP
Bold Text = STR locus

STR ranges from *GRCh37.p10 Primary Assembly*, www.ncbi.nlm.nih.gov