2020

American Population Study of Pigmentation Based Genotype Interpretation for Phenotypic Determination of Hair and Eye Color using HIrisPlex

Emma Leigh Combs
West Virginia University, elcombs@mix.wvu.edu

2020

American Population Study of Pigmentation Based Genotype Interpretation for Phenotypic Determination of Hair and Eye Color using HIrisPlex

Emma Leigh Combs

American Population Study of Pigmentation Based Genotype Interpretation for Phenotypic Determination of Hair and Eye Color using HIrisPlex

Emma Leigh Combs, B.S.

Thesis submitted to the Eberly College of Arts and Sciences at West Virginia University

In partial fulfillment of the requirements for the degree of Masters' of Science in Forensic and Investigative Science

Tina Moroose, M.S., Chair
Casey Jelsema, Ph.D.
Craig Barette, Ph.D.

Department of Forensic and Investigative Science

Morgantown. West Virginia
2020

Keywords: Forensic DNA Analysis, DNA Phenotyping, HIrisPlex assay, DNA

Abstract

American Population Study of Pigmentation Based Genotype Interpretation for Phenotypic Determination of Hair and Eye Color using HIrisPlex

Emma Leigh Combs, B.S

Currently the largest limitation with DNA evidence is that a comparison to a known source sample is required for any interpretation with the current methods. Simply put, if an unknown sample from a crime scene is collected and results in a profile, but there is no suspect or match from CODIS to compare it to, the profile is essentially useless and no information can be gained from it. Research has been performed within the area of forensic DNA phenotyping as a potential tool to aid in taking steps forward to use genotypic information as an investigative tool. Populations studies have lead to the discovery of information used to develop an assay, HIrisPlex, for the purpose of understanding and predicting externally visible characteristics (EVCs) such as eye and hair color from DNA.

Development of these tools for forensic purposes could be utilized to begin establishing a physical description of a suspect to help further aid in an investigation. The current limitation with the HIrisPlex assay is in eye color based prediction related to the lack of understanding of the genetic basis of non-brown or non-blue eye colors which they refer to as intermediate eye colors. The goal of this study was to take the first step in using a higher level of diverse phenotypes present within an American population to evaluate the accuracy of the DNA variants selected in relation to the prediction of eye color phenotypes for the HIrisPlex assays.

Table of Contents

Table of Content (cont.)

1. Purpose, Goals, and Objectives

The information gained from biological evidence within the field of Forensic Biology is highly valued, as DNA evidence has the potential for a high power of discrimination. However, currently the largest limitation with DNA evidence is that a comparison to a known source sample is required for any type of analysis or interpretation of DNA evidence with the current methods used. Simply put, if an unknown sample from a crime scene is collected and results in a profile but there is no suspect or match from Combined DNA Index System (CODIS) to compare it to, the profile is essentially useless and no information can be gained from it to further the investigation. Currently there is research being performed within the area of forensic DNA phenotyping, as a potential tool to aid in taking steps forward to use genotypic information as an investigative tool.

Populations studies have lead to the discovery of information used to develop assays such as the IrisPlex and HIrisPlex for the purpose of understanding and predicting externally visible characteristics (EVCs) such as eye color and hair color from genotypes.[1] [2] [3] Development of these tools for forensic purposes could be utilized to begin establishing a physical description of a suspect to help further aid in an investigation when no suspect for Short Tandem Repeat (STR) profile comparison has been found [4]. This can also be a useful tool in aiding in identifying recovered remains, where visual identifications are not possible.[4]

Within the development and validation of these assays the researchers discuss the limitation of eye color based prediction related to the lack of understanding of the genetic basis of non-brown or non-blue eye colors which they refer to as intermediate eye colors.[2] They also discuss how worldwide population studies for use in relation to eye color prediction where known phenotypic information of the individual whose genotypic information is being used to evaluate the prediction model should be conducted. [2] This discussed area of interest for further research could be used as a foundation for a better understanding of the DNA variants in relation to intermediate eye color phenotypes.[2] The knowledge gained from this research could be utilized in forming an intermediate eye color prediction model that has high accuracy, similar to the results of the one developed with the IrisPlex assay in relation to predicting blue and brown eye color phenotypes.[2]

The goal of this study is to take the first step in using a higher level of diverse phenotypes present within an American population to evaluate the accuracy of the DNA variants

selected in relation to the prediction of eye color phenotypes for the IrisPlex and HIrisPlex assays. This aim aids in evaluating these developed assays in relation to the different variations of intermediate phenotypes using an American population study.

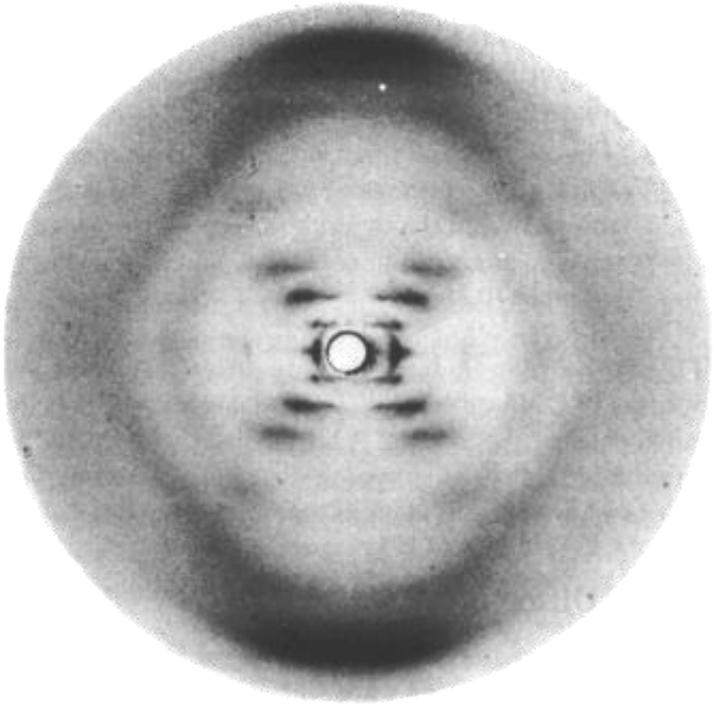
There are two main objectives of this study, the first will be to perform an internal validation of the HIrisPlex assay on the West Virginia University DNA lab Applied Biosystems® 3500 Genetic Analyzer for quality assurance purposes before being used for research samples. The second objective will be to run samples collected from individuals in the population who have different variations of intermediate eye color phenotypes. This will be used for evaluating the accuracy of the DNA variants used in this assay in relation to the eye color phenotype prediction. Specifically, this will evaluate the selected DNA variants on a wider range of intermediate eye color phenotype variations as well as examine the prediction probability values with recorded phenotypes to see if any significant trends can be observed.

2. Theoretical Background

2.1 DNA and inheritance

Deoxyribonucleic acid (DNA) is the genetic template for what constructs life.[5] In humans there are two types of DNA, nuclear DNA which is contained within the nucleus of cells, and mitochondrial DNA which is circular DNA that is contained within each mitochondria and is maternally inherited.[5] DNA has a very unique structure that was first discovered by James Watson and Francis Crick in 1954.[5] The signature double helix structure was pieced together by Dr. Watson and Dr. Crick after examining a crystallography image of a DNA molecule captured by Dr. Rosalind Franklin which can be seen in **Figure 1.**[6] [7]

The double helix structure is composed of two antiparallel helical strands which are held together by hydrogen bonds between the purine and pyrimidine bases.[5] Each of the strands is made up of a deoxyribose sugars and phosphate groups.[5] The phosphate groups form a phosphodiester bond between the 5’ carbon of the a deoxyribose sugar and the 3’ carbon of another sugar.[5] Branching off each of the sugars within the phosphate backbone are the purine



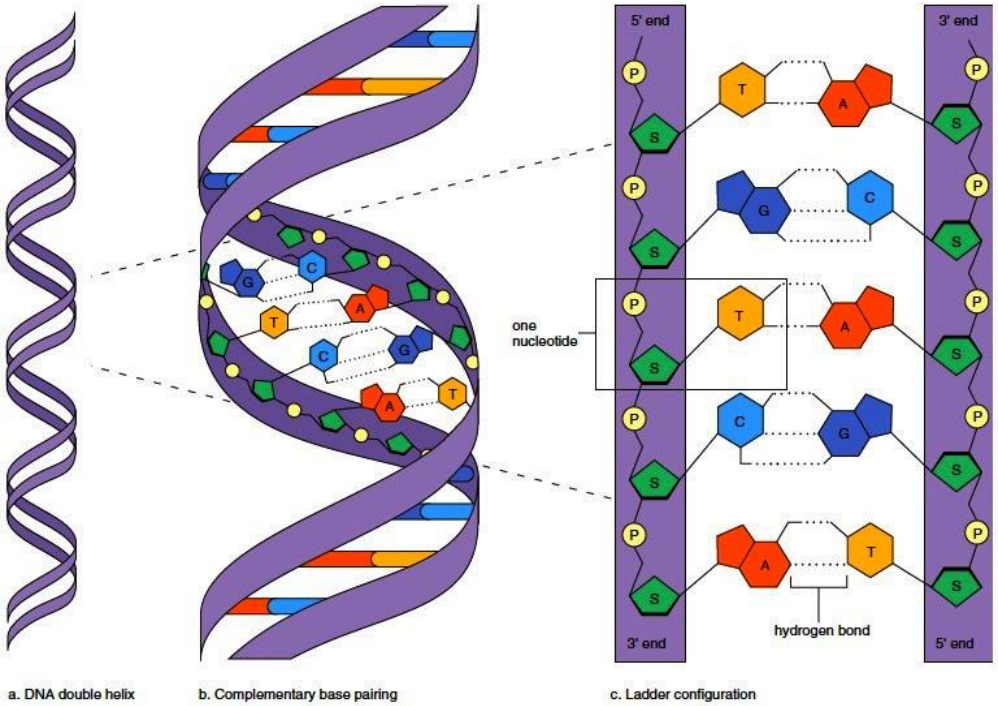
https://www.mun.ca/biology/scarr/Franklins_crystallograph.html

Figure 1. Crystallography image known a B-DNA captured by Dr. Rosalind Franklin in 1954. [7]

and pyrimidine nitrogenous bases.[5] The purine bases of one strand forms a hydrogen bonds with the a pyrimidine bases of the other.[5] With a specific complementary binding relationship, the purine base adenine (A) only binds with the pyrimidine base thymine (T), and the purine base guanine (G) only binds with the pyrimidine base cytosine (C).[5] These complimentary binding pairs pull the two antiparallel strands together in the signature double helix shape.[5] A visual diagram of the discussed structure can be found in **Figure 2**.

Nuclear DNA in humans consists of 23 pairs of chromosomes, 22 autosomal pairs and a single pair of sex chromosomes.[5] Parents of an offspring each contribute half of an individual's chromosomal information, meaning one chromosomes within each pair is inherited from the mother and the other half from the father.[5] Within the pair of sex chromosomes, an X chromosome is alway contributed by the mother, while a the father contributes the sex determining chromosome of either an X or Y chromosome.[5]

Chromosomes contain genetic information in the form of a gene, and all genes have a specific location on a specific chromosome.[5] The location is referred to as a locus.[5] At a

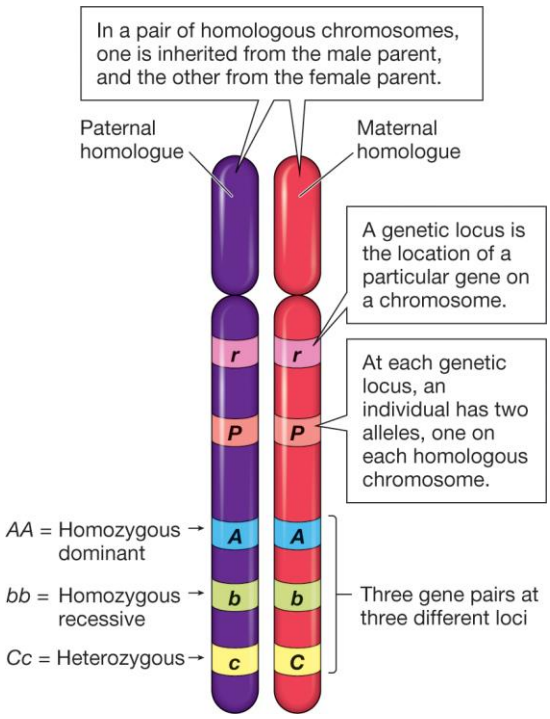


http://encyclopedia.lubopitko-bg.com/Nucleic_Acids.html

Figure 2. An illustration of the structure, a DNA double helix (a), complementary base pair binding (b), and the ladder configuration diagram to display the bonds between the sugar, represented by a green pentagon, and the phosphate represented by a yellow circle that makes up the phosphate backbone[8].

particular locus there can be varying forms of that gene which are referred to as an allele.[5] Each individual has two alleles per locus, one on each chromosome meaning one allele inherited from the father and one from the mother.[5] When an individual inherits the same allele for a gene from both parents they are homozygous at that locus, and if different alleles are inherited for the same gene from each parent, an individual is considered to be heterozygous at that locus.[5] A visual illustration of the discussed terminology is depicted in **Figure 3**.

The genetic information contained with an individual's DNA from all chromosomes is the same through all cells of a given human being, excluding gamete cells.[5] As well as being consistent throughout an individual, their genetic information is unique to that individual with the exception of monozygotic twins who share identical genetic information.[5]



https://digital.wwnorton.com/ebooks/epub/bionowcore/OPS/xhtml/Chapter07-3.xhtml

Figure 3. A visual diagram that explains the terminology in relation to alleles on a pair of homologous chromosome.[9]

2.2 Protein Synthesis

DNA functions as the template for the synthesis of all proteins that are produced and interact to construct an individual.[5] The sequence of nucleotides along a strand of DNA for a specific gene is used as the foundation for the transcription of ribonucleic acid (RNA).[5] The transcribed RNA is then free to move outside of the nucleus of the cell to a ribosome where it can be translated into a sequence of amino acids to synthesize the specific protein dictated by original DNA sequence.[5] The process of transcription occurs in the nucleus, and RNA polymerase II binds to the promoter region that sits up stream in the sequence of the DNA from the gene to be transcribed.[5] RNA polymerase II then builds a pre-messenger RNA sequence at the transcriptional start site.[5]

Like DNA, RNA is composed of four nucleotides; adenine (A), guanine (G), cytosine (C), and uracil (U).[5] Uracil forms a complementary binding pair with the adenines within the DNA sequence, while the adenines within the RNA form complementary binding with the thymines within the DNA sequence.[5] While the complementary binding relationship between cytosine and guanine remains the same.[5] The pre-mRNA undergoes a packaging process in

which a series of modifications to the strand specify it for a specific protein and increases the time a messenger RNA can be free flowing within the cytoplasm of the cell as it works its way to a ribosome to be translated before being broken down by the cell.[5] These modifications include the addition of a 5' cap and a poly-A-tail, as well as the removal of introns from the sequence.[5] Introns are the selection of the pre-mRNA that do not contain coding information, therefore modifications are made to cut out the introns and splice together the exons which are the regions of the pre-mRNA that contain the coding information for the protein.[5]

Once properly packaged and modified, the mRNA moves out of the nucleus and into the cytoplasm of the cell, where it facilitates protein synthesis.[5] This is done by translating the mRNA strand from non-overlapping three base pairs units, referred to as codons, into a chain of amino acids.[5] Each amino acid has its own chemistry and the charge of an amino acid, whether it be positive, negative or neutral, effects the way the sequences of amino acids will fold and dictates the structure of the protein.[5] When it comes to proteins, the structure of a protein decides its function, and ultimately the expression of the gene.[5] Therefore mutations that cause a single nucleotide difference within a coding region of a gene have the potential to alter the structure, and therefore the function of the protein.[5] These physical results of gene expression are referred to as phenotypes, and are decided by the genotype of the specific gene for an individual.[5]

2.3 Polymorphisms

Approximately 5% of the human genome codes for proteins, while the rest is composed of introns.[5] [10] The non-coding regions of the genome are often highly variable because the occurrence of mutations within this region of the DNA does not impact the phenotypes of an individual.[5] [10] This variability is known as a polymorphism, which can be in two main variations, sequences or length based polymorphisms.[10] Sequence based polymorphisms are when there is a change of a nucleotide within the DNA sequence, often a single nucleotide difference referred to as a single nucleotide polymorphism or a SNP.[10] An example of this can be seen in part A of **Figure 4**. There are also regions of DNA that contain repeating units of repetitive DNA that can be classified as micro or minisatellites based on the number of base pairs (bp) within the repeat unit.[10] The microsatellites contain repeat units between 2-8 base pairs in length and repeat a variable number of times. These repeat units are referred to as short tandem repeats (STRs).[10]

These are considered length polymorphism, and an example of this can be seen in part B of **Figure 4**.

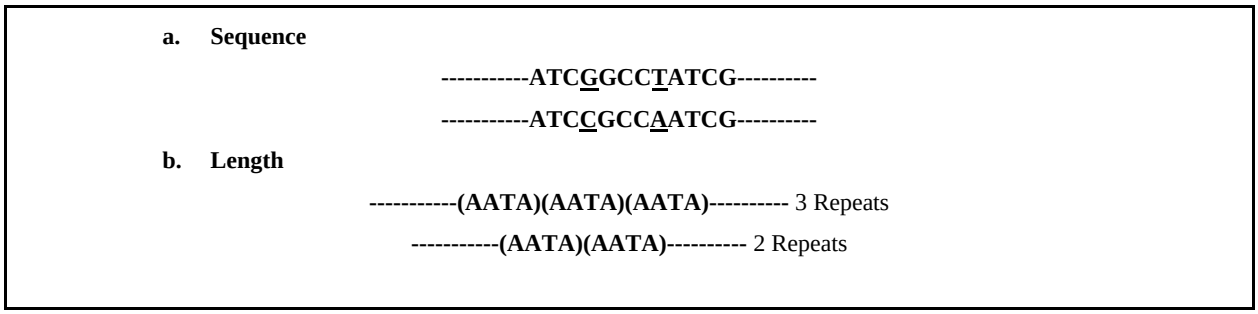


Figure 4. Visual illustration of the difference of what makes a sequence based polymorphism (a), and a length based polymorphism (b)[10].

2.4 STR vs. SNP profiling

The current method for DNA analysis within the field of forensics examines STRs by utilizing a PCR based technique, to analyze a minimum of 20 core loci set by the FBI standards.[10] [11] These core loci consist of units of four base pair length polymorphisms that vary in the number of times they repeat.[10] The process for obtaining a DNA profile using STR profiling involves extraction, quantitation, amplification, and capillary electrophoresis.[10]

SNPs detection methodology and instrumentation is traditionally different than STR profiling due to capillary electrophoresis being a size based separation technique.[12] SNPs are sequence polymorphisms therefore they are detected differently.[12] One of the most common and accurate methods utilized in the detections of SNPs is called pyrosequencing, which uses an enzyme cascade system that produces light whenever complementary base pairing is formed between nucleotides.[12] As the free floating nucleotides bind to the template DNA, the light signal is detected and translated into base pair information.[12] The next nucleotide is then added, generating nucleotide sequence data.[12] The SNaPshot™ kit from Applied Biosystems® can be used for SNPs detection using a capillary electrophoresis instrument.[12]

The SNaPshot™ kit strategically utilizes artificially introduced size based separation and the use of fluorescent dye labeled dideoxynucleotide triphosphates (ddNTPs) in place of fluorescently labeled primers.[12] These fluorescently labeled ddNTPs are incorporated into the sequence during the single base pair extension (SBE) portion of PCR that allow for specific nitrogenous bases to appear as peaks within specific dye channels of an electropherogram.[12] Artificial size base separation is created for each of the SNPs detected by adding different length

poly-tails to the 3’ end of the single base pair extension (SBE) primers.[12] This allows for SNP based profiles to be produced using a genetic analyzer, making it the easier methodology to implement within a forensic DNA laboratory.[12] However, the selection of this method has the limitation of only providing specific single nucleotides as opposed to the short target sequences that can be acquired using a sequence based method like pyrosequencing.[12]

2.5 DNA Phenotyping

Forensic DNA Phenotyping is used to predict possible traits related to an individual's appearance, or externally visible characteristics (EVCs) from specific regions of DNA.[4] The goal within this area of research is to be able to achieve the ability of interpreting an individual's physical description from viewing their genetic profile.[4] This information could then be utilized to aid in investigations where there are no suspects or a CODIS match to compare to a traditional STR profile produced from a crime scene sample.[4] Currently the most information is know in relation to EVCs that are dictated by pigmentation based genes, such as eye color, hair color, and skin tones.[4] Phenotype prediction models termed IrisPlex and HIrisPlex have been developed for use in predicting the eye color and hair color phenotypes of an individual based off of specific SNP genotype profiles.[4] With these predictions it is feasible to begin generating a possible physical description of the unknown individual that left a biological sample at the crime scene, resulting in information that could turn up possible leads to aid investigations.[4]

2.6 IrisPlex assay

2.6.1 Background

Human iris color is a highly polymorphic phenotype, and recent studies have aimed to increase genetic understanding of human eye color.[2] These studies that focused on genome-wide association and linkage analysis have resulted in understanding that the OCA2 gene located on chromosome 15, originally suspected to be the gene most informative in relation to human eye color, is less significant in its association with human eye color than the neighboring HERC2 gene.[2] It was found that exon 12 of the OCA2 gene functions as a modifier for the rs1291382 SNP on the HERC2 gene.[2] Though this HERC2/OCA2 region on chromosome 15 contributes most of the eye color information in relation to blue and brown, 5 other SNPs were found to have

a lesser degree of influence in human eye color variation.[2] This information along with that from a population study of 6168 Dutch European individuals looked at 15 SNPs originating from 8 different genes in relation to eye color, and was used in the selection of a subset of SNPs used to develop the single multiplex genotyping system they refer to as IrisPlex.[2]

The IrisPlex assay consists of six eye color informative SNPs that are currently the most accurate in predicting blue and brown eye color in humans based off the previously mentioned research and Dutch European population study. These are rs12913832 (HERC2), rs1800407 (OCA2), rs12896399 (SLC24A4), rs16891982 (SLC45A2 (MATP)), rs1393350 (TYR), and rs12203592 (IRF4).[2] This assay can generate DNA data to be used with the constructed prediction model to classify the eye color of an individual.[2]

2.6.2 Methods

DNA from buccal samples collected from their selected population was extracted using the QIAamp™ DNA Mini kit, following the manufacturer's protocol (Qiagen, Hagen, Germany).[2] The details of the six SNPs rs12913832, rs1800407, rs12896399, rs16891982, rs1393350 and rs12203592 and the primers used in the assay are summarized in **Table 1**.

The designed primer pairs were created using the Primer3Plus™ software, and each PCR fragment size was limited to no more than 150 bp to accommodate degraded DNA samples expected in relation to the intended implementation of IrisPlex for forensic purposes.[2] To ensure success of capillary separation between the single base extension (SBE) PCR products, poly-T tails of differing sizes were added to the SBE primers at the 5'end.[2] The SNaPshot™ kit by Applied Biosystems® was used for the amplification and detection of the multiplex SBE assay following the kit manufacture protocol.[2]

Table 1. Summary table information related to SNPs, and their PCR primer sequences and concentrations for IrisPlex assay.[2]

SNP-ID	Prediction rank	Chromosome Position	Gene	PCR products (bp)	Forward PCR Primers (5'-3') Reverse Primers (5'-3')	Each Primer Conc. (uM)	Single Base Extension primer (5'-3') with t-tail for length differentiation
rs12913832 (1F and 1R)	1	15 – 26039213	HERC2	87	TGGCTCTCTGTGTCTGATCC GGCCCCTGATGATGATAGC	0.416	ttttttttttttttttttttttGCGTGCAGAACTTGACA
rs1800407a(2F and 2R)	2	15 – 25903913	OCA2	127	TGAAAGGCTGCCTCTGTTCT CGATGAGACAGAGCATGATGA	0.416	ttttttGCATACCGGCTCTCCC
rs12896399 (3F and 3R)	3	14 – 91843416	SLC24 A4	104	CTGGCGATCCAATTCTTTGT CTTAGCCCTGGGTCTTGATG	0.416	ttttttttttttttttttttttttttttaTCTTTAGGTCAGTAT ATTTTGGG
rs16891982 (4F and 4R)	4	5 – 33987450	SLC45 A2(MA TP)	128	TCCAAGTTGTGCTAGACCAGA CGAAAGAGGAGTCGAGGTTG	0.416	ttttttttttAAACACGGAGTTGATGCA
rs1393350 (5F and 5R)	5	11 – 88650694	TYR	80	TTCCTCAGTCCCTTCTCTGC GGGAAGGTGAATGATAACACG	0.416	ttttttttttttttttttttttaTTTGTAAAAGACCACAC AGATTT
rs12203592a(6F and 6R)	6	6 – 341321	IRF4	115	ACAGGGCAGCTGATCTCTTC GCTAAACCTGGCACCAAAAG	0.416	ttttttttttttttaTTTGGTGGGTAAAAGAAGG

2.6.3 Results

The predicted phenotypes were determined from the results of the multinomial logistic regression model previously published in Liu et al. labeled as equations 1-3.[1 2]

$$\pi_{l} = \frac{exp(\alpha_{l} + \Sigma\beta(\pi_{l})_{k} x_{k})}{1 + exp(\alpha_{1} + \Sigma\beta(\pi_{1})_{k} x_{k}) + exp(\alpha_{2} + \Sigma\beta(\pi_{2})_{k} x_{k})} \quad (1)$$

$$\pi_2 = \frac{exp(\alpha_2 + \Sigma\beta(\pi_2)_k x_k)}{1 + exp(\alpha_1 + \Sigma\beta(\pi_1)_k x_k) + exp(\alpha_2 + \Sigma\beta(\pi_2)_k x_k} \quad (2)$$

$$\pi_3 = 1 - \pi_1 - \pi_2 \quad (3)$$

Equation 1-3 are were to calculate the probability of an individual having brown (π_1), blue (π_2), and intermediate (π_3) eye colors, where $x_k$represents the number of minor alleles of the *k*th SNP, and the alpha and beta values were obtained from data of the previously performed Dutch population study by Liu et al.[1 2] Prediction probability values for each category were calculated for each individual sample, which is used to determine the sample classification as brown, blue or intermediate based off of a 0.7 threshold for the probability value.[2] This threshold was determined using the receiver operating characteristic (ROC) curve from the Liu et al. Dutch population.[1 2]

The IrisPlex assay design kept the PCR fragment length between 80-128 bp so that the intended application of this assay within forensics could still be utilized on degraded DNA samples that are a commonly present sample type within forensic DNA analysis.[2] The PCR and SBE multiplex aimed to generate peak heights with approximately equal intensities, and overall allele balance.[2] Though peak height balance was almost achieved, there were two slight imbalances observed in relation to SNPs rs12896399 and rs16891982.[2] However, this slight imbalance did not display any issues with samples of DNA quantities above the determined sensitivity threshold.[2] Though the optimal DNA quantity in the range of 0.25-0.5 ng of template, full profiles were consistently obtained with as low as 31 pg of DNA, and only at 15 pg of DNA was allelic dropout first observed. This can be seen in **Figure 5.**[2]

Prevalence-adjusted prediction accuracy values obtained from the area under the receiver characteric operating curve (AUC) gave very high values for both brown and blue eyes.[2] The author's reported prediction accuracy values of the developed model as 0.93 for brown eyes, and a 0.91 for blue eyes, where a fully accurate prediction value should be equal to 1.[2]

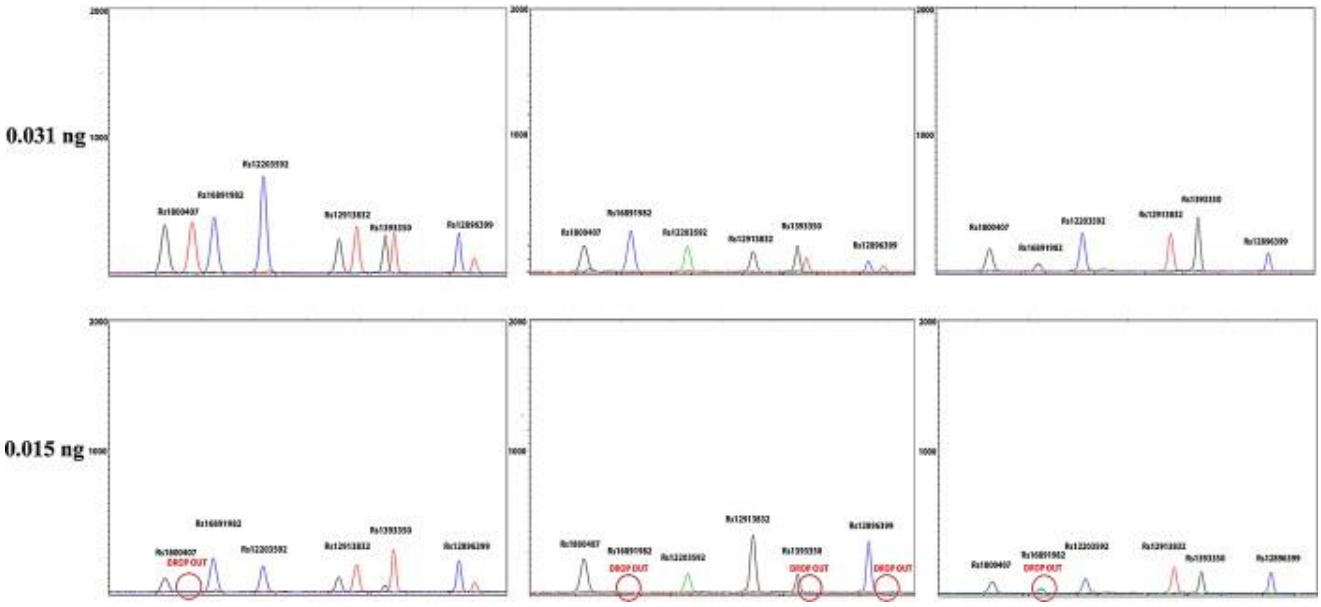


Figure 5. Images of a IrisPlex assay profile obtained from starting concentrations of 31pg and 15pg of DNA to show where allelic drop out occurred, which is circled in red. [2].

Once these prediction probability values were calculated for population used, they ordered the high resolution photos of the irises of the individuals sampled for the study in order based on the probability values.[2] Starting with the highest blue value in the top left corner and the highest brown value in the bottom right corner.[2] This is illustrated in **Figure 6**. The phenotypic eye color was not considered in the selected order of the images in **Figure 6** but it can be seen that there is a correlation between the visible color of the expressed phenotypes and the produced prediction values when viewed from top to bottom and left to right.[2] The authors took this as confirmation of the accuracy of their developed six SNP prediction model.[2]

For 37 out of the 40, or 92.5%, of the individuals sampled in the study the genetic eye color prediction agreed with the phenotypic color based on visual inspection.[2] It was observed that intermediate eye colors were more difficult to classify as they can appear a lighter or varying shades of blue or brown.[2] The study concluded that more work in relation to the genetic variants for the intermediate, non-brown and non-blue, eye colors needs to be done to obtain a higher predictive value.[2]

Each of the six SNPs contribute genetic information towards the prediction model, with the rs12913832 SNP within the HERC2 gene having been shown to have the most predictive

information in relation to eye color determination.[2] The other five SNPs rs1800407(OCA2 gene), rs12896399 (SLC24A4 gene), rs16891982 (SLC45A2 (MATP) gene), rs1393350 (TYR gene), and rs12203592 (IRF4 gene) within the IrisPlex assay a smaller degree of influence to the accuracy of the prediction model.[2] A visual representation of each SNPs informational contribution to eye color determination can be seen in **Figure 7**.[2]

2.6.4 Validation

A validation study of this assay was completed following SWGDAM guidelines.[13] Successful accuracy and reproducibility was found on a wide range of sample materials including blood, semen, saliva, hair, and touch DNA samples that included very low quantity samples.[13] Its sensitivity for obtaining a full profile from the six SNPs was shown at 31pg of template DNA.[13] Species testing revealed the complete assay is specific to human and primate DNA profiles.[13] The only noted potential issue with the IrisPlex assay is that it is unable to be applied to mixture samples.[13] However, as an intended investigation based tool to be used when no suspect or CODIS match can be found for comparison, it would fall within the DNA casework flow after an STR profile has been obtained, at which point the presence of a mixture should have been identified.[13]

Few alterations to the run parameters and the assay were made within the study to increase the sensitivity.[13] These modifications included an increase in the annealing temperature for the multiplex PCR, and the direction of the SNP primers for rs1800407, and rs12203592 in the SBE reaction.[13] This was done to increase the resulting peak heights at lower DNA concentrations.[13] The primer concentration for rs16891982 was increased to 0.5µM to increase the peak heights produced when the homozygous C/C alleles were present.[13] Lastly, the standard protocol for the ABI 3130xl Genetic Analyzer® was changed to have an increased injection voltage and time, as well as decreasing the processing time to a 500 second run time.[13] All of these minor changes resulted in an increase in the overall performance of the IrisPlex assay.[13]

2.7 HIrisPlex

2.7.1 Background

After recent studies demonstrated that hair color could be predicted based upon specifically selected DNA markers, the researchers who developed the IrisPlex assay worked to expand it.[3] The goal of the expanded assay was to exploit the strong genetic and phenotypic

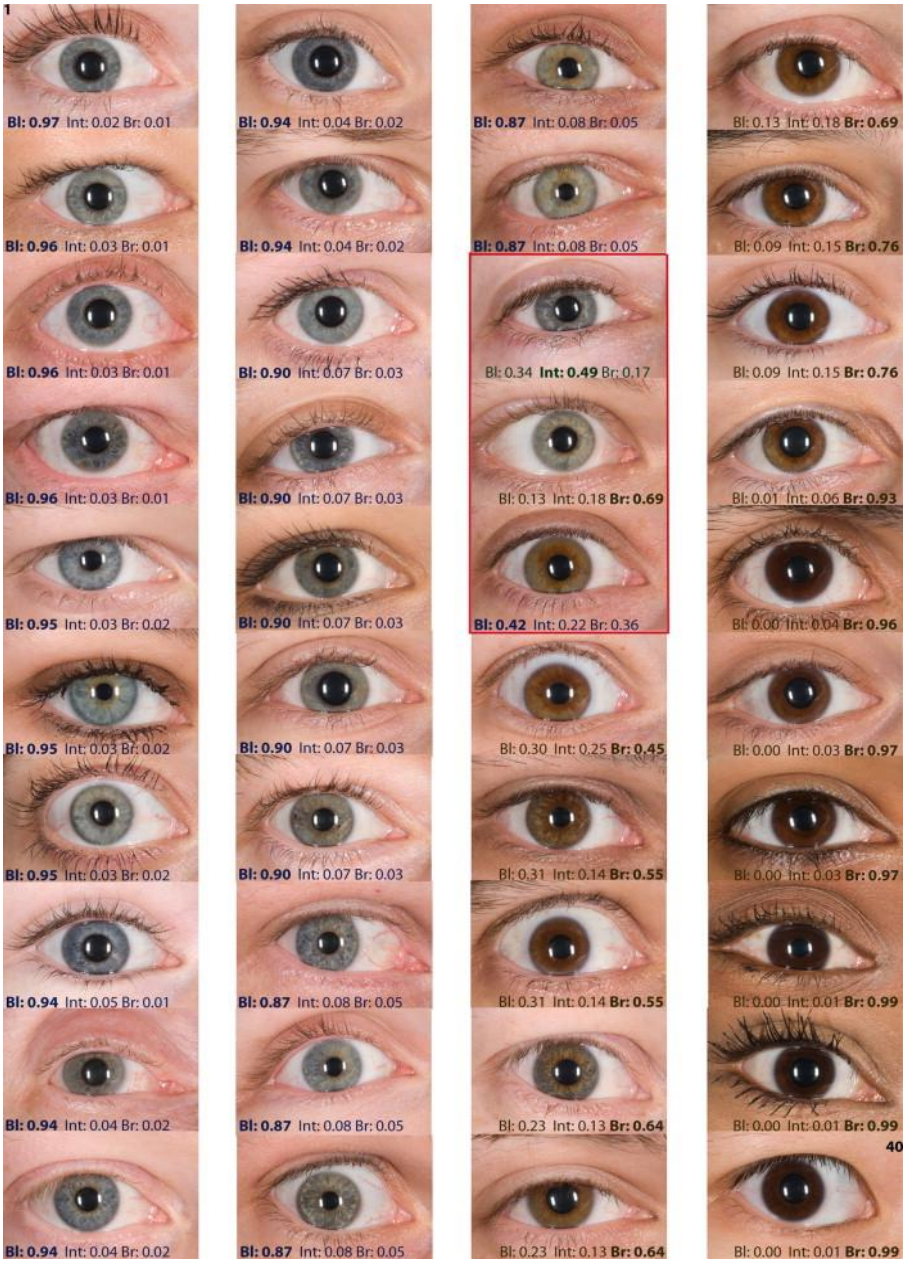


Figure 6. Illustrated visual example of how the IrisPlex analysis model worked in predicting eye color phenotypes of 40 individuals, with calculated model predicted probability values attach to corresponding phenotypes.[2]

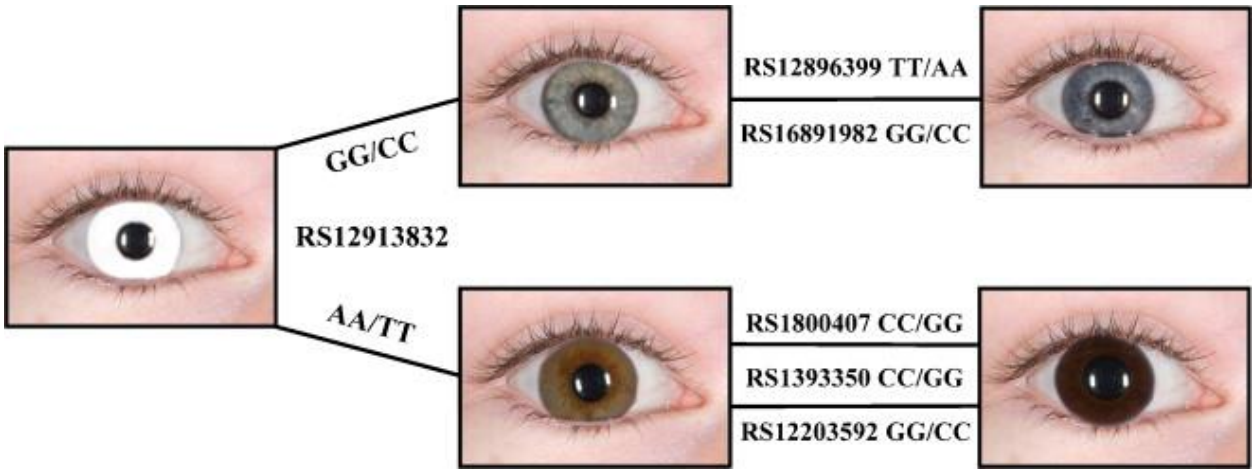


Figure 7. Hypothesized scenario for genetic determination of brown and blue eye colors showing the impact of the most influential SNP genotypes from the 6-SNP model.[2]

relationship observed between hair and eye color variation.[3] Their objectives were to increase the understanding of the genetic influence of hair color and to create a prediction model that works for hair color determination with an expanded assay.[3] Secondly, they aimed to create an assay that when combined in a single run gives a profile of the SNPs related to hair color determination as well as the previously studies six SNPs used in the IrisPlex assay.[3] This combined assay was termed HIrisPlex.[3]

Based on information gained from Valenzuela et al., three SNPs, rs12913832 (HERC2), rs16891982 (SLC45A2), and rs1426654 (SLC24A5), were found to provide the best prediction for light verse dark hair colors.[3] [14] 46 different SNPs from 13 genes were considered in relation to the variants within human hair color variation based of the authors previous study.[3] [15] The 46 DNA variants were evaluated for model base hair color prediction based off of population data in relation to determining hair color variants.[3] [15] Though a set of markers containing the most informative genes in relation to hair color prediction were identified in their previous study, they were unable to run genotyping simultaneously for all of the selected 22 DNA markers within a single reaction.[3] Therefore, this study focused on developing and evaluating a single-tube multiplex assay.[3] The assay included all 22 of the hair color predictive DNA variants from the previous study as well as the 6 SNPs from the IrisPlex assay.[3] This resulted in a total of 24 DNA markers as four were overlapping.[3] With this newly developed HIrisPlex assay, the author's also

made available a spreadsheet tool to be used in interpreting and performing the calculations for the prediction model classification for both the IrisPlex and HIrisPlex assay.[3] [16]

2.7.2 Methods

DNA samples from three separate European populations were collected from Poland, Greece, and Ireland with the intention of collecting samples that contained a range of all phenotype categories.[3] DNA samples were extracted using multiple different methods for each population, and all samples were genotyped with the HIrisPlex assay.[3] The assay includes 23 SNPs and 1 insertion or deletion (INDEL) polymorphism.[3] These DNA variants came from a total of 11 separate genes.[3] Sample collection included high resolution images of hair color and eye color of individuals under the same lighting and distance conditions.[3] All individuals sampled were given a questionnaire asking for basic information as well as questions related to self reported hair and eye color phenotypes.[3] Buccal swab samples of each individual were collected.[3] A summary of the PCR primers for the 24 DNA variants for the assay can be found in Table 2.[3] The authors followed the same parameters and methods for primer design used in the IrisPlex assay development, with the exception of changing the target fragment length from less than 150 bp to less than 160 bp.[2] [3]

The same reaction conditions from the IrisPlex PCR runs were used for the HIrisPlex with the adjusted PCR primer concentrations in Table 2, still utilizing the SNapShot™ kit from Applied Biosystems® for the multiplex SBE.[2] [3] The samples were run on a ABI 3130xl Genetic Analyzer® with POP-7 on a 36 cm capillary following the manufacturer's guidelines of the SNapShot™ kit, with the exception of the run parameter being set to 2.5 kV 10s inject voltage and a 500 second run time at 60°C.[3] These alterations in the run parameters were used to increase the sensitivity of the assay.[3] The analytical threshold used was set to 50-rfu. This was based on the production of a full profile in every replicate of the different input DNA levels from their conducted sensitivity study.[3]

The HIrisPlex assay consists of a total of 23 SNPs and 1 INDEL.[3] 6 of the SNPs are from the previously established IrisPlex assay used for eye color prediction.[3] The 22 DNA variants used in relation to hair color prediction are 11 SNPs, Y152OCH, N29insA, rs1805006, rs11547464, rs1805007, rs1805008,rs1805009, rs1805005, rs2228479, rs1110400, and rs885479, from the MC1R gene, as well as rs1042602 (TYR), rs4959270 (EXOC2), rs28777 (SLC45A2

(MATP)), rs683 (TYRP1), rs2402130 (SLC24A4), rs12821256 (KITLG), rs2378249 (PIGU/ASIP), rs12913832 (HERC2), rs1800407 (OCA2), rs16891982 (SLC45A2 (MATP)), and rs12203592 (IRF4).[3]

Though hair color can be displayed in many different variations, the authors categorized them into four main groups: red, blonde, brown and black.[3] For the four categories the highest calculated probability value was used for the category prediction.[3] A random subset of 80% of the samples from each selected population were used for training the prediction model for hair color to obtain the alpha and beta values to be used with the model.[3] The hair color prediction model was based on the same Multinomial Logistic Regression model (MLR) used for the IrisPlex assay published by Liu et al modified to the four discrete categories.[1 3] Individuals were divided into the previously discussed four categories based on their phenotypes, and the 22 DNA variants in relation to hair color prediction were used to test the prediction model.[3] Due to the division of the categories being based in color and not shade, a different approach was used to separate the individuals grouped into subsequent light verse dark categories for hair color.[3] This was performed utilizing a two-prong model approach which used only blonde and black category data to predict light or dark categories.[3] These shaded category predictions were based on the influence of genotypes associated with blonde and black phenotypes.[3] The remaining 20% of samples from each population were then used to evaluate the accuracy of the developed prediction models in terms of both the hair color category prediction as well as hair shade prediction.[3]

2.7.3 Results

Effort was put in to optimizing the single multiplex assay to have balanced peak heights, with the intent of minimizing allelic drop at lower DNA concentrations.[3] This was successful with the exception of the INDEL DNA variant N29insA, which is a difficulty known to occur with INDELs.[3] Though a lower peak height is observed at this DNA variant, it was not seen to affect the sensitivity of the HIrisPlex assay until DNA concentrations were lower than 63 pg[3]. However, it was noted that this technical issue related to N29insA DNA variant did not affect the overall practical use of the HIrisPlex assay.[3] It was observed that greater than 500pg of

Table 2. Summary table information related to DNA variants, and their PCR primer sequences and concentrations for HIrisPlex assay.[3]

Assay position	SNP	CHR	Position		Gene	Major Allele	Minor Allele	PCR primers		Concentration		Product size	SBE primers	Concentration
1	N29insA	16	89985753	Exonic	*MC1R*	C	insA	MC1Rset1F	Set1	0.55 μm	GCAGGGATCCCAGAGAAGAC	117bp	CCCCAGCTGGGGCTGGCTGCCAA	1.3 μm
2	rs11547464	16	89986091	Exonic	*MC1R*	G	A	MC1Rset1R		0.55 μm	TCAGAGATGGACACCTCCAG		ttttttttttttGCCATCGCCGTGGACC	0.1 μm
3	rs885479	16	89986154	Exonic	*MC1R*	C	T	MC1Rset2F	Set2	0.5 μm	CTGGTGAGCTTGGTGGAGA	158bp	ttttttttttttttttttGATGGCCGCAACGGCT	1.25 μm
4	rs1805008	16	89986144	Exonic	*MC1R*	C	T	MC1Rset2R		0.5 μm	TCCAGCAGGAGGATGACG		ttttttttttttttACAGCATCGTGACCCTGCCG	0.375 μm
5	rs1805005	16	89985844	Exonic	*MC1R*	G	T	MC1Rset3F	Set3	0.5 μm	GTCCAGCCTCTGCTTCCTG	147bp	ttttttttttttttttTGGTGGAGAACG-CGCTGGTG	0.75 μm
6	rs1805006	16	89985918	Exonic	*MC1R*	C	A	MC1Rset3R		0.5 μm	AGCGTGCTGAAGACGACAC		ttttttttttttttttttttCTGCCTGGC-CTTGTCGGA	0.75 μm
7	rs1805007	16	89986117	Exonic	*MC1R*	C	T	MC1Rset4F	Set4	0.4 μm	CAAGAACTTCAACCTCTTTCTCG	106bp	ttttttttttttttttttttttttttCTCCATCTTC-TACGCACTG	1 μm
8	rs1805009	16	89986546	Exonic	*MC1R*	G	C	MC1Rset4R		0.4 μm	CACCTCCTTGAGCGTCCTG		ttttttttttttttttttttttttttttttttttATCTGC-AATGCCATCATC	0.4 μm
9	Y152OCH	16	89986122	Exonic	*MC1R*	C	A						ttttttttttttttttttttttttttttttttCATCTT-CTACGCACTGCGCTA	0.6 μm
10	rs2228479	16	89985940	Exonic	*MC1R*	G	A						ttttttttttttttttttttttttttttttttttC-TGGTGAGCGGGAGCAAC	0.375 μm
11	rs1110400	16	89986130	Exonic	*MC1R*	T	C						ttttttttttttttttttttttttttttttttCTTCTA-CGCACTGCGCTACCACAGCA	0.3 μm
12	rs28777	5	33994716	Intronic	*SLC45A2*	A	C	rs28777_F	Set5	0.4 μm	TACTCGTGTGGGAGTTCCAT	150bp	ttttttttttttttttttttttttttttttttttttttCA-TGTGATCCTCACAGCAG	1.2 μm
								rs28777_R		0.4 μm	TCTTTGATGTCCCCTTCGAT			
13	rs16891982	5	33987450	Exonic	*SLC45A2*	G	C	Rs16891982_F	Set6	0.4 μm	TCCAAGTTGTGCTAGACCAGA	128bp	ttttttttttttttttttttttttttttttttttttttttt-AAACACGGAGTTGATGCA	1 μm
								Rs16891982_R		0.4 μm	CGAAAGAGGAGTCGAGGTTG			
14	rs12821256	12	87852466	Intergenic	*KITLG*	A	G	rs12821256_F	Set7	0.4 μm	ATGCCCAAAGGATAAGGAAT	118bp	ttttttttttttttttttttttttttttttttttttt-GGAGCCAAGGGCATGTTACTACGGCAC	0.1 μm
								rs12821256_R		0.4 μm	GGAGCCAAGGGCATGTTACT			
15	rs4959270	6	402748	Intergenic	*EXOC2*	C	A	Rs4959270_F	Set8	0.4 μm	TGAGAAATCTACCCCCACGA	140bp	ttttttttttttttttttttttttttttttttttttt-GGAACACATCCAAACTATGACACTATG	0.375 μm
								Rs4959270_R		0.4 μm	GTGTTCTTACCCCCTGTGGA			
16	rs12203592	6	341321	Intronic	*IRF4*	C	T	rs12203592_F	Set9	0.4 μm	AGGGCAGCTGATCTCTTCAG	126bp	ttttttttttttttttttttttttttttttttttttttt-tTCCACTTTGGTGGGTAAAAGAAGG	0.3 μm
								rs12203592_R		0.4 μm	GCTTCGTCATATGGCTAAACCT			
17	rs1042602	11	88551344	Exonic	*TYR*	G	T	rs1042602 _F	Set10	0.4 μm	CAACACCCATGTTTAACGACA	124bp	ttttttttttttttttttttttttttttttttttttttt-ttttttTCAATGTCTCTCCAGATTTCA	1.25 μm
								rs1042602 _R		0.4 μm	GCTTCATGGGCAAAATCAAT			
18	rs1800407	15	25903913	Exonic	*OCA2*	G	A	rs1800407_F	Set11	0.4 μm	AAGGCTGCCTCTGTTCTACG	124bp	ttttttttttttttttttttttttttttttttttttttt-ttttttttttttttGCATACCGGCTCTCCC	0.1 μm
								rs1800407_R		0.4 μm	CGATGAGACAGAGCATGATGA			
19	rs2402130	14	91870956	Intronic	*SLC24A4*	A	G	rs2402130_F	Set12	0.4 μm	ACCTGTCTCACAGTGCTGCT	150bp	ttttttttttttttttttttttttttttttttttttttttt-ttttttttttttTGAACCATACGGAGCCCGTG	0.75 μm
								rs2402130_R		0.4 μm	TTCACCTCGATGACGATGAT			
20	rs12913832	15	26039213	Intronic	*HERC2*	C	T	rs12913832_F	Set13	0.4 μm	TCAACATCAGGGTAAAA-ATCATGT	150bp	ttttttttttttttttttttttttttttttttttttttttt-ttttttttttttttttTAGCGTGCAGAACTTGACA	1.2 μm
								rs12913832_R		0.4 μm	GGCCCCTGATGATGATAGC			
21	rs2378249	20	32681751	Intronic	*ASIP/PIGU*	T	C	rs2378249_F	Set14	0.4 μm	CGCATAACCCATCCCTCTAA	136bp	ttttttttttttttttttttttttttttttttttttttttt-ttttttttttttttttttCCACACCTCTCCTCAGCCCA	0.18 μm
22	rs12896399	14	91843416	Intergenic	*SLC24A4*	T	G	Rs12896399_F	Set15	0.4 μm	CTGGCGATCCAATTCTTTGT	125bp	ttttttttttttttttttttttttttttttttttttttttt-ttttttttttttttttttTCTTTAGGT-CAGTATATTTTGGG	1.125 μm
								Rs12896399_R		0.4 μm	GACCCTGTGTGAGACCCAGT			
23	rs1393350	11	88650694	Intronic	*TYR*	C	T	Rs1393350_F	Set16	0.4 μm	TTTCTTTATCCCCCTGATGC	124bp	ttttttttttttttttttttttttttttttttttttttttt-ttttttttttttttttttttttCATTTGTA-AAAGACCACACAGATTT	1.1 μm
								Rs1393350_R		0.4 μm	GGGAAGGTGAATGATAACACG			
24	rs683	9	12699305	Exonic	*TYRP1*	T	G	rs683_F	Set17	0.4 μm	CACAAAACCACCTGGTTGAA	138bp	ttttttttttttttttttttttttttttttttttttttttt-ttttttttttttttttttttttGCTTTG-AAAAGTATGCCTAGAACTTTAAT	0.175 μm
								rs683_R		0.4 μm	TGAAAGGGTCTTCCCAGCTT			

DNA resulted in a balanced profile with high RFU levels, but full profiles, though imbalanced, were still obtained with as low as 63pg of DNA.[3] Allelic dropout did not occur until quantities under 63pg of DNA were run, at which drop out only occured at 5 instances.[3] Specifically at N29insA, rs1042602, rs4959270, rs1800407, and rs1393350.[3]

The hypothesized influence of each of the 22 DNA variants considered in the contribution to hair color is illustrated in **Figure 8.**[3] The provided results of the HIrisPlex prediction for a subset of 44 individuals' phenotype images was used to visually assess the performance of the prediction model.[3] The individuals hair color images were ordered according

to their phenotype predicted probability values obtained from the HIrisPlex analysis.[3] The actual hair color phenotype of the images was not considered in the ordering.[3] The results of this are depicted in **Figure 9.**[3]

Since overlap exists between hair colors and factors such as the presence of an A allele at N29insA or Y152OCH produces a red hair color probability value equal to one, the highest probability category approach was adjusted to consider these factors.[3] A prediction guide approach was developed for interpretation purposes, an example of this can be seen in **Figure 10.**[3]

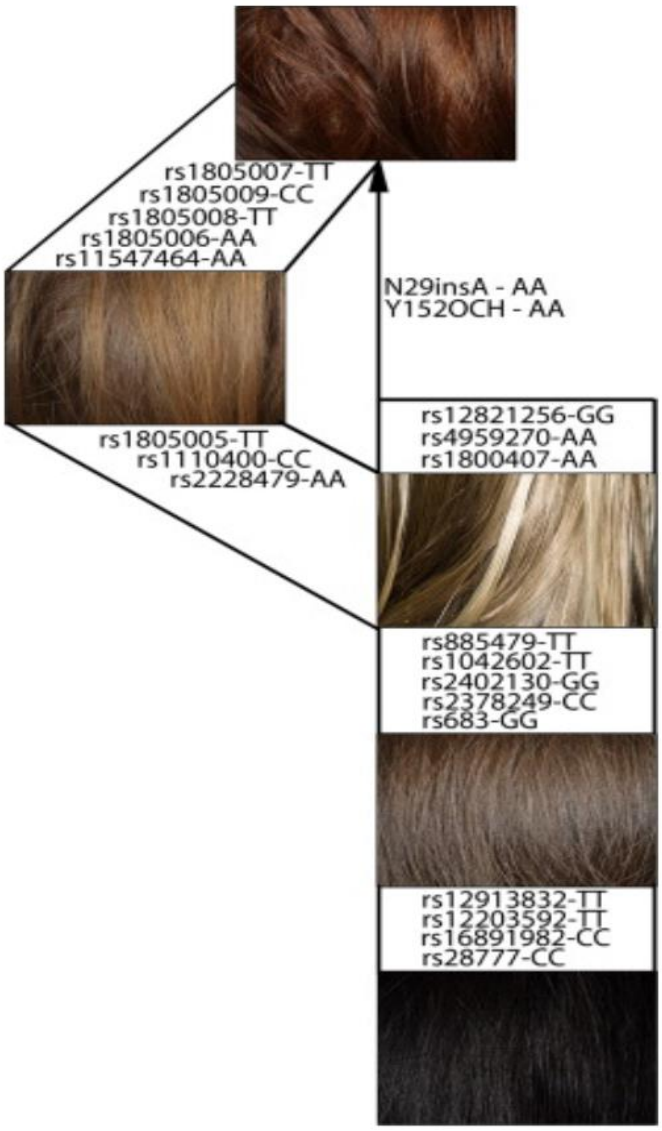


Figure 8. Hypothesized scenario for genetic determination of black, brown, red (two shades) and blond showing the impact of the most influential DNA variant genotypes from the 22 DNA variant model.[3]

Figure 9. Illustrated visual example of how the HIrisplex analysis model worked in predicting hair color phenotypes of 44 individuals, with calculated model predicted probability values attach to corresponding phenotypes.[3]

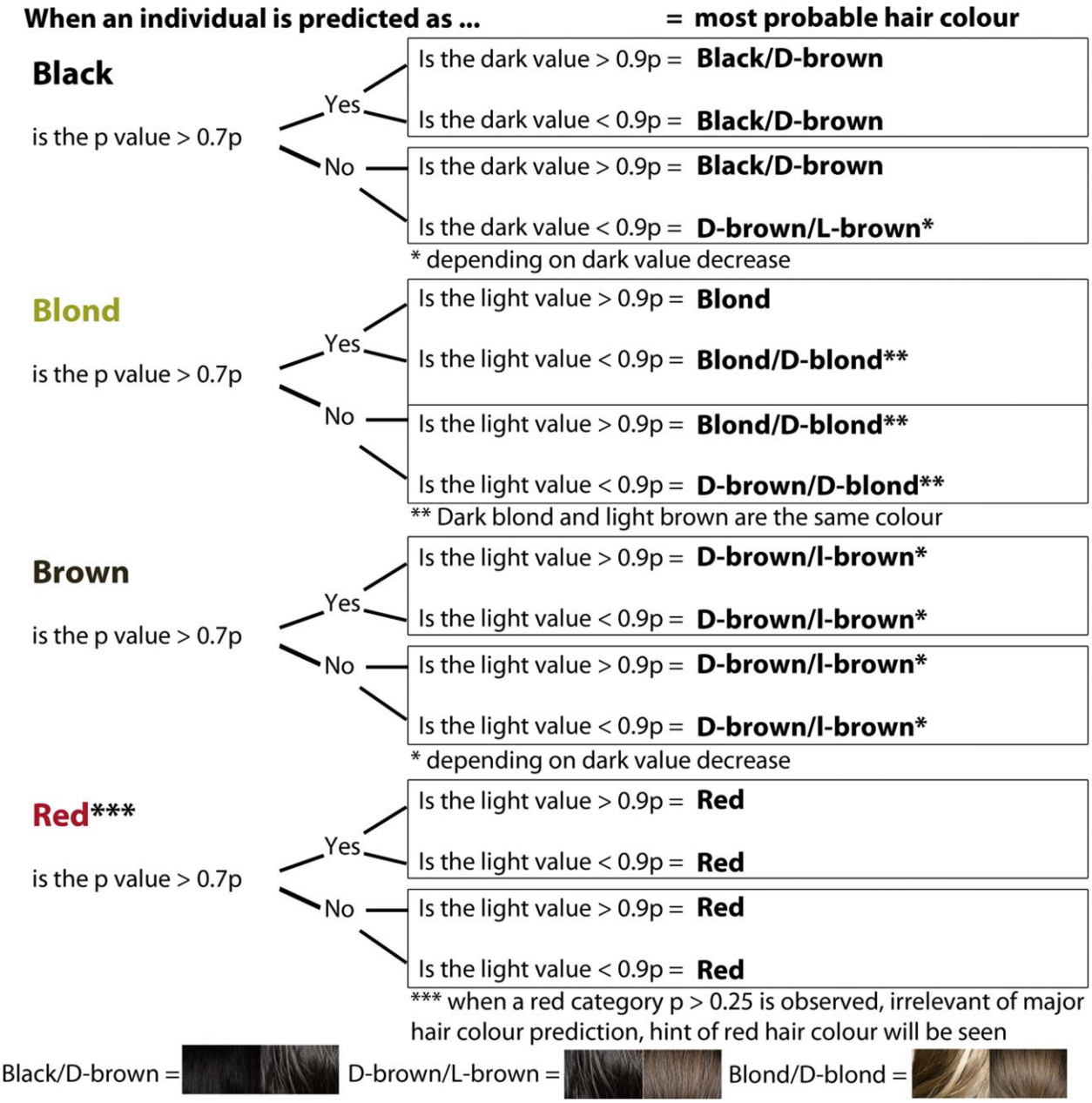


Figure 10. The HIrisPlex prediction guide for interpreting the individual hair color and shade prediction probability values.[3]

2.7.4 Validation

A validation study of this assay was completed following SWGDAM guidelines. Successful accuracy and reproducibility was found on a wide range of sample material, including blood, semen, saliva stains, hair, and touch DNA samples that included very low quantity samples.[17] Its sensitivity for obtaining a full profile with the HIrisPlex assay was shown at 63pg of DNA.[17] Species testing revealed the complete assay is specific to human DNA but a close

non-human partial profiles of 20 of the 24 DNA variants were obtained with primate DNA.[17] It was noted that due to the assay design which is designed to cater to degraded samples, the assay is successful at obtaining full profile from human remains that are up to several hundred years old.[17]

3. Methods and Results

3.1 Sample Collection

Due to the need for a collection gradient of a wide range of phenotypes, 70 buccal swab samples were collected in duplicate from willing participants.. The tips of the swabs were cut off into a clean 2ml microcentrifuge tube and labeled with the individual's sample identification number. The identification number corresponded with an extensive survey on genealogy and phenotypic history filled out by each participant. High resolution images of the iris and hair color phenotypes of each of the participants were taken from approximately a 7cm distance from the left eye and back of the head of the individual, using a Nikon camera with a 60mm macro lens set with an f/stop at 5.6, an ISO of 100 and under direct 5000K color light at 50% brightness.

Once all samples and phenotype images were collected, a population of 40 samples of the collected samples were selected to give a relatively equal distribution of all collected phenotypes to be used for internal validation of the HIrisPlex assay. A second population from the collected samples was selected independently of the previous population. This population contained all collected samples with individuals that self-identified their phenotype classification to be a variation of an intermediate eye color phenotype, not already included in the previous population.

3.1.1 Phenotype Image Analysis

Once captured, the iris phenotype images for each of the participants were analyzed in ImageJ, utilizing the RGB Measure Plugin. Due to the variation in iris size, pupil dilation, image impurities such as light reflections, the entirety of the Iris image could not be used to evaluate Red, Green, Blue color values. To minimize variability, five different 250-pixel diameter circles

were taken out of each captured iris phenotype image, and visual representation of this can be seen in **Figure 11**.

Figure 11. A visual representation of five different 250-pixel diameter circles taken out of each captured iris phenotype image.

Each 250-pixel diameter circle was analyzed using the ImageJ RGB Measure Plus Plugin, to give a red, green, or blue color threshold value, as well as a standard deviation value for each of the color thresholds. The color threshold values, as well as the standard deviation values for each of the color thresholds, were also averaged between all five 250-pixel diameter circles to get the overall color determination values for each participant image.

This process was then repeated in triplicate to eliminate any potential human error introduced from the manual selection of the 250-pixel diameter circles.The color threshold values for each replicate were then used to build a Random Forest Classification Model, as described in **Section 4.1.1.**

3.1.2 Phenotype History Survey

The phenotype history survey given to each participant collected information in relation to ancestral origin, the participants self identified hair and eye color phenotypes, as well as any changes to either of these phenotypes for any reason. (ie. Hair dye, gray hair, color changing intermediate eye color phenotype, etc.). This information was collected to be used within data and profile interpretation. A copy of the Phenotype History Survey can be found attached as **Supplement Document 1**.

3.2 Sample Preparation Troubleshooting

3.2.1 Following Published Protocol

Testing began with using my own sample, an intermediate eye color reference sample (ELC_R), and a blue and brown eye color phenotype reference sample (26908.L, and 32942.L) for initial testing following the methods presented in the Walsh et al, HIrisPlex developmental validation study.[17] Due to the selected base pair size of the target regions of the DNA variants for the assay, study samples were extracted using the QIAamp™ DNA Mini Kit from Qiagen following manufacturer's standard protocol. DNA concentrations of samples were quantified using the Quantifiler™ Trio DNA Quantification Kit from Thermo Fisher Scientific following the kit manufacturer protocols. The amplification of the samples was performed in two different amplification steps.

First the AmpliTaq Gold DNA Polymerase with Buffer II and $MgCl_2$ kit from Thermo Fisher Scientific were used for a single multiplex PCR step using 1 µl of extracted DNA between 3ng to 300pg in concentration. This was added to a 20 µl PCR reaction with 1X PCR reaction buffer, 2.5 mM$MgCl_2$, 1.75 U AmpliTaq Gold DNA polymerase and primer concentrations for each primer listed in **Table 2**. The thermal cycling conditions for the PCR reaction on the 9700 Thermal Cycler from Applied Biosystems® can be found in **Table 3**. The PCR products were then cleaned using the ExoSAP-IT PCR Product Cleanup Reagent from Thermo Fisher Scientific following the manufacturer's protocol.

Table 3. Thermal cycling run conditions for single multiplex PCR reaction on the 9700 Thermal Cycler from Applied Biosystems

Thermal Cycling Conditions			
95 °C for 10 minutes	33 cycles of		5 minutes at 60 °C
	95 °C for 30 seconds	60 °C for 30 seconds	

The second amplification step was completed using SBE multiplex PCR utilizing the Applied Biosystems® SNaPshot™ multiplex kit. 3µl of the cleaned PCR product was used with

5µl SNapShot™ reaction mix in a total reaction volume of 10µl, along with all listed SBE primer concentrations listed in **Table 2**. The thermal cycling conditions for the SBE PCR reaction on the 9700 Thermal Cycler from Applied Biosystems® can be found in **Table 4**. The SBE PCR products were cleaned using Shrimp Alkaline Phosphatase (SAP) from USB Corp following manufacturer's guidelines. Samples were prepared for capillary electrophoresis by taking 0.5µl of cleaned product to be run on the ABI 3500 Genetic Analyzer® with POP-7 polymer following the ABI SNaPshot™ Kit sample preparation guidelines, however the run parameters were altered to a 2.5 kV for a 10 second injection voltage and run times of 560 seconds at 60 °C for increased sensitivity.

Table 4. Thermal cycling run conditions for single multiplex PCR reaction on the 9700 Thermal Cycler from Applied Biosystems

Thermal Cycling Conditions			
96 °C for 2 minutes	25 cycles of		
	96 °C for 10 seconds	50 °C for 5 seconds	60 °C for 30 seconds

A singleplex run of each individual SBE primer for each eye color phenotype reference samples was run, in addition to a multiplex run with each of the three eye color phenotype reference samples. The singleplex samples were to aid in the determination of the SNP identification peaks bin set development, while the multiplex samples were to confirm that the multiplex run protocols were working.

Capillary electrophoresis runs were viewed within the GeneMapper-IDX software version 1.4, settings for profile viewing were modified for peak detection only and no bin set, or panel was set in the analysis parameters. Setting for peak detection only were used following the guideline within the ABI SNaPshot™ Kit Analysis Getting Started User Guide version 4.0.

3.2.2 Results for Following Published Protocol

The profiles resulting from the first run following the published protocol yielded electropherograms with no significant peaks in all samples.

3.2.3 Optimization Troubleshooting

3.2.3.1 DNA Concentration Testing, and Cleaning Reagent Adjustment

After researching possible areas within the amplification processes that could be adjusted based on the desired outcomes, two possible issues seemed most likely. Either too much of the cleaning reagents in either of the amplification steps were degrading the samples prior to capillary electrophoresis, or the initial concentration of DNA input into the first PCR step was too low.

Given these factors, a test run of the same three test samples was run with each sample being run at four different concentrations of DNA (1ng, 5ng, 10ng, and 25ng). In addition the concentration of ExoSAP-IT™ Cleanup reagent was decreased to 2 units per 15µL for the cleaning step following the first PCR step. The quantities of Shrimp Alkaline Phosphatase (SAP) reagents in the cleanup step for the second amplification were also adjusted to increase the volume of the second amplification product to 6µL, and adjusted the SAP reagent and buffer to 2 units from a 20µL total reaction volume, decreasing the amount of distilled water to 6µL.

The capillary electrophoresis run settings were also adjusted to increase the injection time to maximize the likelihood of peak visualization.

3.2.3.2 Resulting Electropherograms

The resulting electropherograms contained a range of strong to weak peaks in all samples of varying input DNA concentrations, and a visual of the resulting electropherogram for sample ELC_R at 1ng can be seen in **Figure 12**.

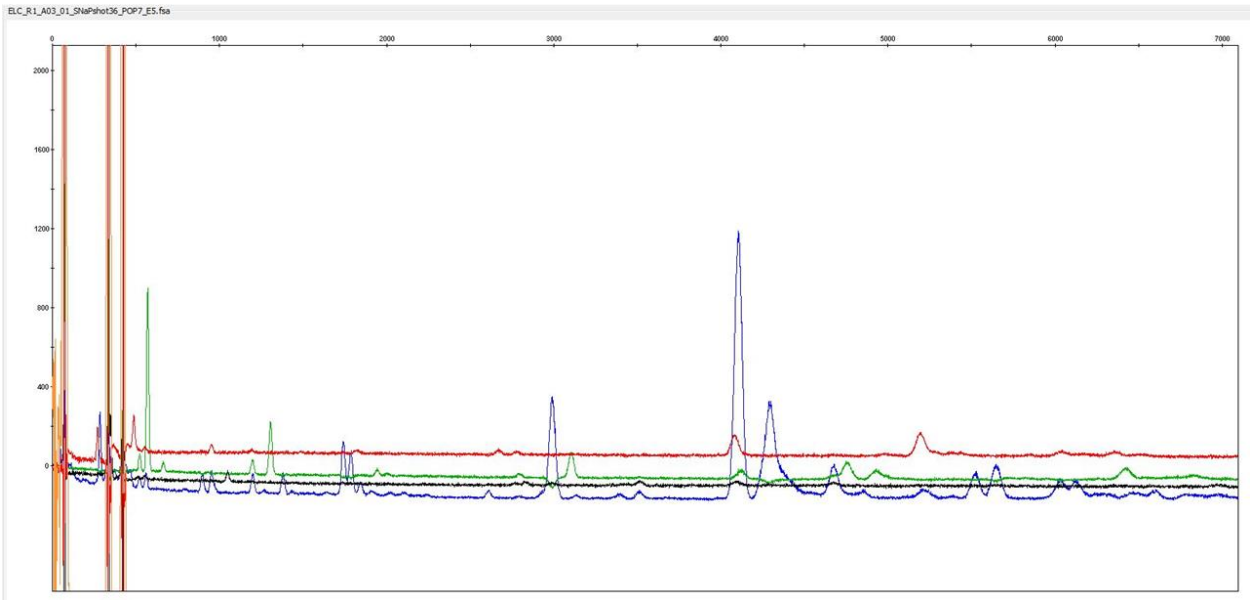


Figure 12. Electropherogram of sample ELC_R at 1ng from sample test run with a testing range of DNA concentrations, and cleaning reagent protocol adjustments.

From this it was concluded that targeting a concentration of DNA around 1ng should be sufficient for future optimization testing. However, the peak heights present within LIZ-120 size standard appeared to be significantly higher than all peaks presumptively being assumed to be data. This can be seen within **Figure 12**, where the LIZ-120 size standard is within the orange dye channel.

3.2.3.3 LIZ-Size Standard Testing

Due to the difficulty in visualization with the disproportionate RFUs values between the size standard peaks and the SNP data peaks, a run utilizing different size standards was done. The aim of the run being to assist with trying to solve a recurring sizing issue of the data within GeneMapper-IDX version 1.4.

A capillary electrophoresis run using the same adjusted run settings discussed in **Section 3.2.3.1** were used in a run with the 1 ng target samples of ELC_R and 32942.L from the previous DNA concentration testing, and cleaning reagent adjustment run. Both samples were run with replicates with LIZ-120 size standard, no size standard, a 1:10 dilution of the LIZ 120 size standard, and 1:10 dilution of the LIZ 600 size standard.

3.2.3.4 LIZ-Size Standard Testing Results

The electropherograms showed increased visualization of the data peak, that could best be viewed in the sample run with no size standard, this can best be seen in **Figure 13**. However, the 1:10 dilution of LIZ 600 size standard had the best coverage of the data peaks present; this can be seen in **Figure 14**.

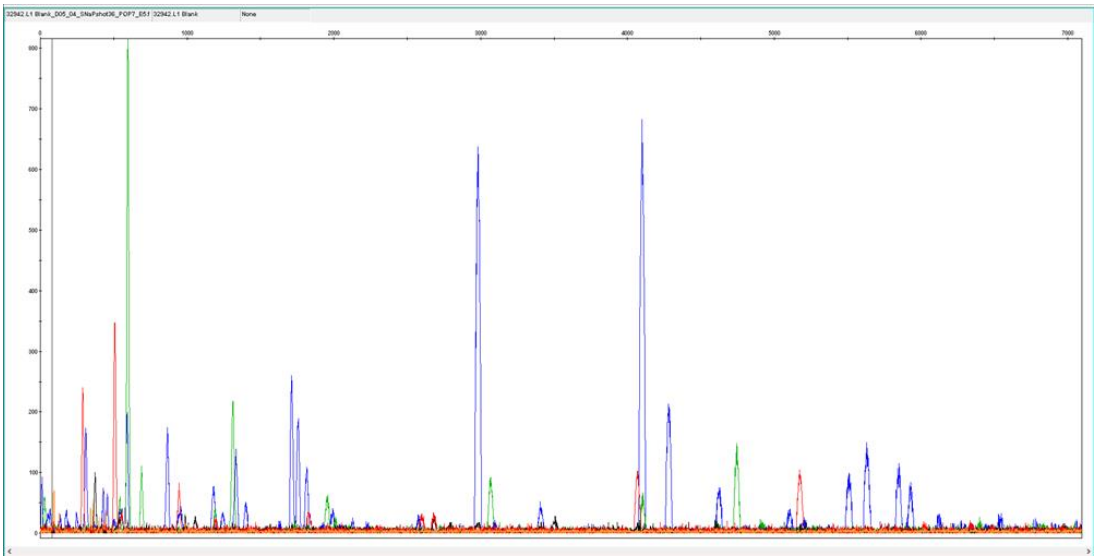


Figure 13. Electropherogram of sample 32942.L run with no size standard.

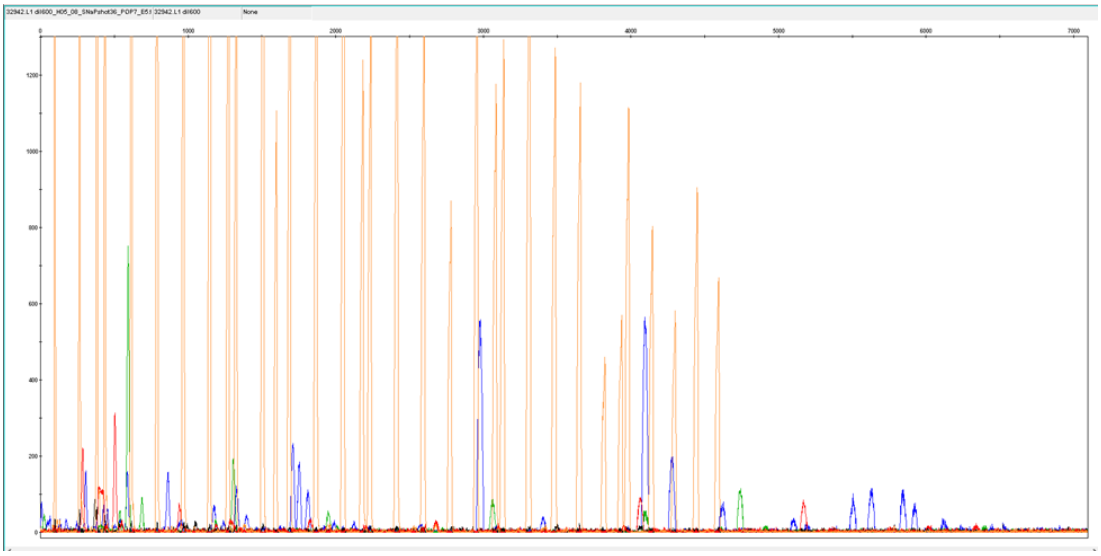


Figure 14. Electropherogram of sample 32942.L run with a 1:10 dilution of LIZ 600 size standard.

The WVU Genomics Core Director, Ryan Percifield, M.Sc. was consulted to view the electropherograms on their GeneMapper version 5 software with a pre-installed SNaPshot™ kit analysis protocol to confirm the peaks were consistent with those viewed using the custom setting within in GeneMapper-IDX version 1.4 in the WVU Forensic Biology Laboratory. After reviewing the profiles on the GeneMapper version 5 software with a pre-installed SNaPshot™ kit analysis protocol it was believed that the peaks present within the electropherogram could be true data.

3.2.3.5 Testing for Reproducibility of Methodology

Samples ELC_R, 26908.L, and 32942.L were rerun in duplicate through the modified amplification and capillary electrophoresis methodology described in **Section 3.2.3.1**, with the exception of using a 1:100 dilution of LIZ 600 size standard. Due to the height of the peaks in the samples run with a 1:10 dilution of LIZ 600 size standard described in **Section 3.2.3.4**, and seen in **Figure 14,** a further dilution of LIZ 600 size standard to a 1:100 dilution was used to test if it would result in better visualization.

3.2.3.6 Testing for Reproducibility of Methodology Results

The resulting electropherograms contained multiple peaks in a range of varying RFUs, that can be seen in **Figure 15**. Due to utilizing GeneMapper-IDX version 1.4 software for peak detection only at this phase in the project, the Base Pairs (bps) and data point values for the sizing standard peaks were collected. This was used to plot a sizing graph to determine the slope and y-intercept for each profile to calculate the value of each peak's bps size, using their reported data point value. The data point values and calculated bps peaks for sample 26908.L can be seen in **Table 5**, **Table 6**, and **Table 7**.

Table 5. LIZ™-600 Size Standard Data from sample 26908.L with a calculated slope of 0.1076721375 and y-intercept of 53.52320545.

BP (y)	Data Point (x)
80	272
100	444
114	566
120	620
140	805
160	984
180	1167
200	1350
214	1479
220	1547
240	1721
250	1824
260	1906
280	2094
288	2177
300	2283
314	2416
320	2470
340	2659
360	2846
380	3037
400	3221
414	3352
420	3406
440	3590
460	3796

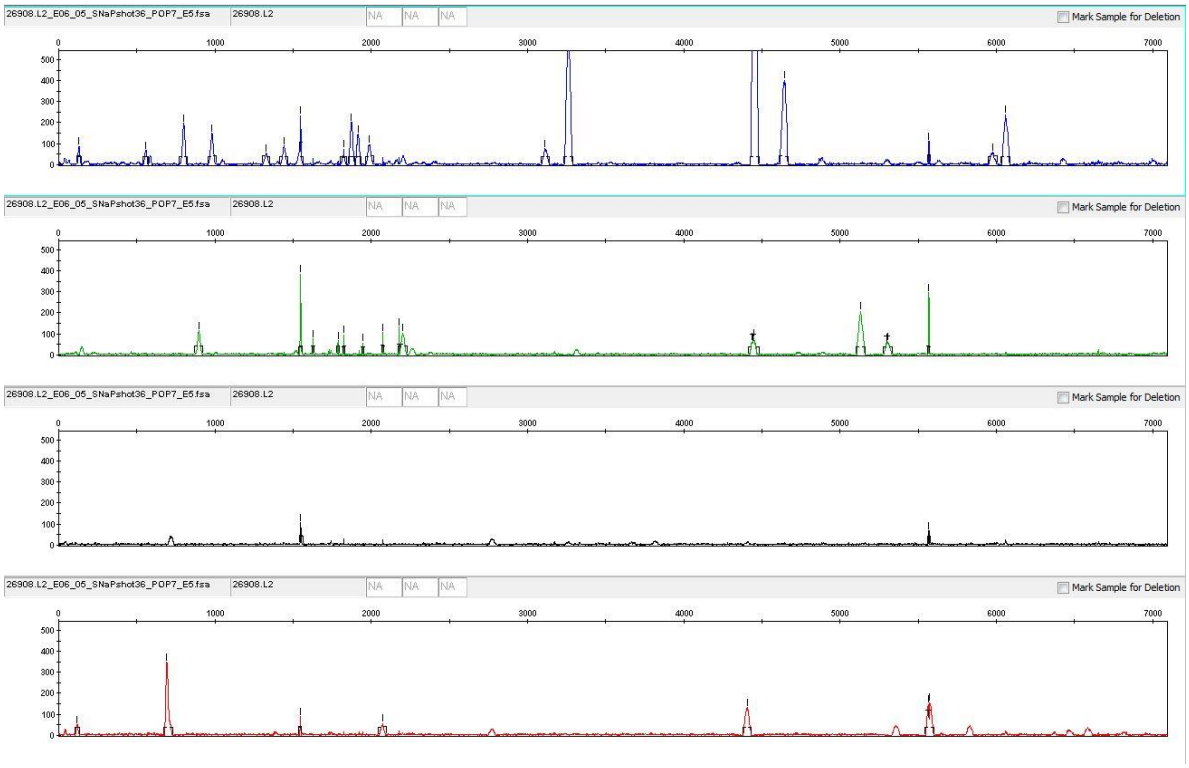


Figure 15. Electropherogram of sample 26908.L from the testing for reproducibility of methodology run described in **Section 3.2.3.5**.

Table 6. Data points from sample 26908.L electropherogram from run described in **Section 3.2.3.3**.

Blue	Green	Yellow	Red
129	896	1548	116
557	1547	5568	692
800	1627		1546
981	1787		2074
1328	1824		4405
1442	1945		5565
1548	2074		5573
1824	2177		
1873	2200		
1915	4438		
1988	4446		
3110	5131		
3262	5297		
4455	5307		
4646	5572		
5568			
5980			
6059			

Table 7. Calculated bps of peaks in sample 26908.L electropherogram using sizing data in Table 5 from the run described in **Section 3.2.3.3**.

Blue	Green	Yellow	Red
67.41291119	149.9974407	220.1996743	66.0131734
113.496586	220.0920022	653.041667	128.0323246
139.6609155	228.7057732		219.98433
159.1495723	245.9333152		276.8352186
196.511804	249.9171842		527.8189711
208.7864277	262.9455129		652.7186506
220.1996743	276.8352186		653.5800277
249.9171842	287.9254488		
255.193119	290.4019079		
259.7153488	531.3721517		
267.5754148	532.2335288		
388.3835531	605.9889429		
404.749718	623.8625178		
533.202578	624.9392391		
553.7679563	653.4723556		
653.041667			
697.4025877			
705.9086865			

3.2.3.7 Singleplex Reruns for Bin Determination

Due to obtaining profiles with distinct peaks using the same methodology for more than one run, I continued with the assumption that true profiles were being obtained. To be able to move forward more easily and no longer have to complete manual interpretation of profiles a singleplex run with each individual primer using sample ELC_R, and one sample with the full multiplex primer set was run following the protocol laid out in **Section 3.2.3.5**. The intention of the run being able to locate the expected peak positions for each individual SNP within the profiles to aid in the creation of a bin set determination protocol within GeneMapper-IDX version 1.4. The multiplex sample run with the singleplex samples was intended to be utilized as a reference to confirm the rerun results were consistent with the previous runs.

3.2.3.8 Singleplex Reruns for Bin Determination Results

Resulting profiles had inconsistent appearances of peaks within the majority of samples. A rerun of the samples was performed with the modification that the forward and reverse primers within the first amplification run both increased to 1μL per reaction, as well as increasing the SBE primer volume to 2μL per reaction within the second amplification step.

The resulting rerun produced sample profiles with peak issues similar to the first run, as well as issues with having no sizing data when viewed within GeneMapper -IDX version 1.4. From this, the amplified samples with the increased primer volumes were rerun through capillary electrophoresis with the sample volume increased to 1μL and utilizing non-diluted size standard. The resulting sample profiles appeared more consistent with the previously obtained profiles, however there was an issue of seeing multiple peaks, exceeding the two possible expected peaks for a heterozygous phenotype, in all singleplex samples which was also present in the multiplex samples. To gain better understanding of if this could be possible contamination or noise, focus was shifted into panel and bin set creation for easier data interpretation.

3.2.3.9 Panel and Bin Determination Protocol Creation in GeneMapper-IDX version 1.4

Following the panel and bin set creation instructions in the SNaPshot™ Kit Analysis Getting Started Guide version 4.0, a SNaPshot™ Panel was created within the software library. Within the Panel folder a bin set was created and the bin boundaries for each bin were estimated based on the bp number of the related SBE primer lengths, plus one for the addition of the

fluorescently labeled nucleotide with a plus or minus of 0.5 or 1 bps for error. The input values for the estimated bin set can be found in **Table 8**. A copy of the panel settings, and bin set file for the HIrisPlex assay SNP peaks, compatible with being uploaded into GeneMapper softwares can be found in **Supplementary Document 2** and **Supplementary Document 3** respectively.

Table 8. Estimate Bins for HIrisPlex assay SNP peaks

SBE Primer	Estimated Bin
N29insA	24 ±1
rs11547464	29±1
rs885479	35±0.5
rs1805008	34±0.5
rs1805005	36±0.5
rs1805006	39±1
rs1805007	45±1
rs1805009	49±1
Y152OCH	52±1
rs2228479	55±1
rs1110400	57±1
rs28777	59±1
rs16891982	63±1
rs12821256	67±1
rs4959270	69±0.5
rs12203592	70±0.5
rs1042602	74±1
rs1800407	78±1
rs2402130	81±1
rs12913832	84±1
rs2378249	87±1
rs12896399	89±1
rs1393350	93±1
rs683	96±1

Applying the SNaPshot™ Panel containing the estimated bin set to the GeneMapper-IDX version 1.4 analyze protocol on all of the resulting profiles from runs described in **Section 3.2.3.8,** revealed a few issues. With further examination it was determined that though the positive and negative control from the SNaPshot™ Multiplex kit worked, all of the peak data obtained from question samples were far out of the range to be consistent with true peak data for the HIrisPlex assay.

4. Data Analysis

4.1 Statistical Analysis of ImageJ Data

4.1.1 Random Forest Model for Classification

The set of color threshold values collected from each of the Iris images processed using the methods described in **Section 3.1.1** were used to create a Random Forest Classification Model, for the classification of Iris Phenotypes. The Random Forest methodology utilizes the development of multiple decision trees that are then used to aggregate the data into a classification model. These models work in 3 steps. In step one trees are drawn based on the value of ntree, or number of trees within the model, of the bootstrap samples. In the second step each of the bootstrap samples grow into unpruned trees that split based on sampling of predictors at each node. These predictors are referred to as mtry, which for classification based models is determined based off of the square root of p, where p is defined as the number of variables or features within the model. The final step is then to make a decision of the models classification based on the majority vote when taken into account the decisions of each individual tree within the Random Forest.

This was done using the 'randomForest' Package version 4.6 -14 within R. The R script used to create the Random Forests in the study can be found in **Supplemental Document 4 and 5** respectively.

4.1.2 Five Class Classification Random Forest Model

The first Random Forest created in this study was completed by using a model with five classifications, based on the eye color phenotype options contained within the phenotype history survey collected from each participant. The self-identified phenotype of the individuals that the images were collected from, was assumed to be their true classification of the images for the purpose of creating the model. The five classes were defined within the dataset using the numerical values contained within **Table 5**.

Table 5. Numerical key for iris phenotype classification within Random Forest Model One.

Numeric Value	Iris Phenotype Classification
1	Blue
2	Brown
3	Green
4	Blue/Green Hazel
5	Brown/Green Hazel

The data contained 135 observations and 37 variables, and was partitioned into a 70/30 split between a training dataset and a testing dataset for the construction and evaluation of the Random Forest. This created a training dataset which contained 96 observations and a testing set with 39 observations. The default ntree value of 500 was used for the creation of this model.

4.1.3 Five Class Classification Random Forest Model Results

The Random Forest Model constructed from the bootstrapped training data set with the five class classification categories produced from the iris phenotype image data was found to have an accuracy equal to 1, with a 95% CI [0.9623,1], and a p-value of <2.2e-16. However,

when the prediction error within the model was evaluated using the non-bootstrapped testing data an Out of the Bag (OOB) error rate of 9.38% was calculated. While the model had a zero percent classification error rate for both blue and brown eye color phenotype categories, the three intermediate eye color phenotype categories had classification error rates of 21.4%, 33.3%, and 4.0% for Green, Blue/Green Hazel and Brown/Green Hazel respectively. The Confusion Matrix for the Five Class Classification Random Forest can be found in **Figure 16**.

Confusion matrix:

	1	2	3	4	5	class.error
1	22	0	0	0	0	0.0000000
2	0	20	0	0	0	0.0000000
3	0	0	11	2	1	0.2142857
4	1	0	2	10	2	0.3333333
5	0	1	0	0	24	0.0400000

Figure 16. Random Forest Confusion Matrix for the Five Class Classification RScript found in Supplementary Document 4.

4.1.4 Three Class Classification Random Forest Model

Due to all misclassifications errors within the Five Class Classification Random Forest Model being contained within the classification of the three different types of intermediate eye color phenotypes, a second Random Forest Model was created utilizing only 3 classes for classifications as laid out in **Table 6**.

Table 6. Numerical key for iris phenotype classification within Random Forest Model Two.

Numeric Value	Iris Phenotype Classification
1	Blue
2	Brown
3	Intermediate (ie. Green, Blue/Green Hazel, or Brown/Green Hazel)

With keeping all other information the same within the model, but decreasing the possible classification categories to only three the resulting Random Forest Model produced for the iris phenotype image data was found to have the same accuracy equal to 1, with a 95% CI [0.9623,1], and a p-value equal <2.2e-16 as the five class classification model. However, the

OOB error rate of the second model was calculated to be 4.17%. While the second Random Forest model still had a zero percent classification error rate for the brown eye color phenotype category, it was found to have classification error rates for blue and intermediate eye color phenotype categories of 4.5%, and 5.5% respectively. The Confusion Matrix for the Three Class Classification Random Forest can be found in **Figure 17**.

Confusion matrix:

	1	2	3	class.error
1	21	0	1	0.04545455
2	0	20	0	0.00000000
3	2	1	51	0.05555556

Figure 17. Random Forest Confusion Matrix for the Three Class Classification RScript found in Supplementary Document 5.

5. Discussion

In conclusion, the SNP peak data collected in relation to the HIrisPlex assay was determined to not be true peak data. In addition, due to the results of the positive and negative controls within the sample runs using the SNaPshot™ Multiplex assay, it can be concluded that the kit itself was working properly. Ruling out kit component related issues, the difficulties with the HIrisPlex assay can be concluded to most likely be associated with an issue in the binding of the primers. The cause of the primer issues however, could be due to a multitude of factors. Such factors include undocumented storage temperature issues, improper manufacturing of primers, thermocycling condition issues, and so on.

On the other hand, the Random Forest Classification Models created from the collected ImageJ data has important applications within the advancement of Forensic DNA Phenotyping. One of the many topics of discussion within the field is how to objectively classify pigmentation based phenotypes, as color determination from individual to individual can be subjective. Utilization of the Iris Image Random Forest Classification Model built in this study removes the potential error of subjectivity in the classification of eye color phenotypes with the utilization of image color threshold data. Therefore, it could be used to standardize the determination of eye color phenotype classification for future Forensic DNA Phenotyping technologies as one of the first introduced objective classification methods for eye color phenotypes.

6. References

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

[2] 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, Forensic Sci. Int. Genet. 5 (2011) 170–180.

[3] S. Walsh, F. Liu, A.Wollstein, L. Kovatsi, A. Ralf, A. Kosiniak-Kamysz, W. Branicki, M. Kayser, The HIrisPlex system for simultaneous prediction of hair and eye colour from DNA, Forensic Sci. Int. Genet. 5 (2013) 98–115.

[4] M. Kayser, Forensic DNA Phenotyping: Predicting human apperance from crime scene material for investigative purposes, Forensic Sci. Int. Genet. 18 (2015) 33–48.

[5]Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P. Molecular biology of the cell. 4th ed. New York: Garland Science; 2002.

[6] Watson JD, Crick FHC. Molecular structure of nucleic acids. A structure for deoxyribose nucleic acid. 1953. Ann Intern Med 2003;138(7):581.

[7] https://www.mun.ca/biology/scarr/Franklins_crystallograph.html

[8] http://encyclopedia.lubopitko-bg.com/Nucleic_Acids.html

[9] https://digital.wwnorton.com/ebooks/epub/bionowcore/OPS/xhtml/Chapter07-3.xhtml

[10] Butler J. Basics of DNA, biology, and genetics. In: Fundamentals of forensic DNA typing. 3rd ed. Academic Press; 2009.

[11] “Quality Assurance Standards for Forensic DNA Testing Laboratories,” Federal Bureau of Investigation. https://www.fbi.gov/file-repository/quality-assurance-standards-for-forensic-dna-testing-laboratories.pdf

[12] B. Sobrino, M. Brion, A. Carracedo, SNPs in forensic genetics: a review on SNP typing methodologies, Forensic Sci. Int. 154 (2005) 181–194.

[13] S. Walsh, A. Lindenbergh, S.B. Zuniga, T. Sijen, P. Knijff, M. Kayser, K. N. Ballantyne, Developmental validation of the IrisPlex system: Determination of blue and brown iris colour for forensic intelligence, Forensic Sci. Int. Genet. 5 (2011) 464–471.

[14] R.K.Valenzuela, M.S. Henderson, H.M. Walsh, et al. Predicting phenotype from genotype: normal pigmentation, J. Forensic Sci. 55 (2010) 315-322.

[15] E. 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.

[16] https://hirisplex.erasmusmc.nl/, link to HIrisPlex spreadsheet and manual

[17] S. Walsh, L. Chaitanya, L. Clarisse, L. Wirken, J. Draus-Barini, L. Kovatsi, et al., Developmental validation of the HIrisPlex system: DNA-based eye and hair colour prediction for forensic and anthropological usage, Forensic Sci. Int. Genet. 9 (2014) 150–161.

Supplemental Document 1 - Phenotype History Survey

`

Sample Identification #:_____

Phenotype History Questionnaire for ***American Population Study of Pigmentation Based Genotype Interpretation for Phenotypic Determination of Hair and Eye Color using HIrisPlex***

Conducted by Emma L. Combs, B.S.

Name of Sample Donor:_____

Biological Sex: _____

Race:_____

Please list all known possible ancestral origins (be as specific as possible):

_____ _____

_____ _____

_____ _____

_____ _____

1. What would you classify your eye color as?

_____

2. Pick one of the categories below that best describes your eye color
   a. Blue
   b. Green
   c. Blue/Green Hazel
   d. Brown/Green Hazel
   e. Other Hazel
   f. Brown
   g. Heterochromia

3. Have your eyes ever changed color in your lifetime?

a. Yes
b. No
c. Unsure

4. If yes to question 3, please explain in detail below:

_____

_____

_____

_____

_____

5. What would you classify your natural hair color?

_____

6. Selected the hair color and hair shade that best represents your hair appearance:

a. Black
b. Brown
c. Red
d. Blonde

a. Dark
b. Light

7. Have you ever dyed your hair?
a. Yes
b. No

8. If yes to question 7, when was the last time you dyed your hair? What color was it dyed?

_____

9. Has your hair ever changed color naturally in your lifetime?
Example: Your hair was lighter as a child or your hair had begun to gray?

a. Yes
b. No
c. Unsure

10. If yes to question 9, please explained in detail below:

_____

_____

_____

_____

Please print your name, sign and date at the bottom of the page authorizing that all information given in the above questionnaire is the truth, to the best of your current knowledge.

_____
Name

_____ _____
Signature Date

Supplemental Document 2 - HIrisPlex Panel file for GeneMapper Software

`

Version	**GMID-X v 1.4**							
Kit type:	**MICROSATELLITE**							
Chemistry Kit		**SNaPshot**		**none**				
Panel	**HIrisPlex**	**test 5-31-2020**						
rs1805005_G	**blue**	**0.0**	**999.0**	**-**	**1**	**none**	**false**	
rs683_G	**blue**	**0.0**	**999.0**	**-**	**1**	**none**	**false**	
rs1393350_G	**blue**	**0.0**	**999.0**	**-**	**1**	**none**	**false**	
rs12896399_G		**blue**	**0.0**	**999.0**	**-**	**1**	**none**	**false**
rs2378249_G	**blue**	**0.0**	**999.0**	**-**	**1**	**none**	**false**	
rs12913832_G		**blue**	**0.0**	**999.0**	**-**	**1**	**none**	**false**
rs2402130_G	**blue**	**0.0**	**999.0**	**-**	**1**	**none**	**false**	
rs1800407_G	**blue**	**0.0**	**999.0**	**-**	**1**	**none**	**false**	
rs1042602_G	**blue**	**0.0**	**999.0**	**-**	**1**	**none**	**false**	
rs12203592_G		**blue**	**0.0**	**999.0**	**-**	**1**	**none**	**false**
rs1805008_G	**blue**	**0.0**	**999.0**	**-**	**1**	**none**	**false**	
rs885479_G	**blue**	**0.0**	**999.0**	**-**	**1**	**none**	**false**	
rs11547464_G		**blue**	**0.0**	**999.0**	**-**	**1**	**none**	**false**
N29insA_G	**Blue**	**0.0**	**999.0**	**-**	**1**	**none**	**false**	
rs4959270_G	**blue**	**0.0**	**999.0**	**-**	**1**	**none**	**false**	
rs12821256_G		**blue**	**0.0**	**999.0**	**-**	**1**	**none**	**false**
rs16891982_G		**blue**	**0.0**	**999.0**	**-**	**1**	**none**	**false**
rs28777_G	**blue**	**0.0**	**999.0**	**-**	**1**	**none**	**false**	
rs1110400_G	**blue**	**0.0**	**999.0**	**-**	**1**	**none**	**false**	
rs2228479_G	**blue**	**0.0**	**999.0**	**-**	**1**	**none**	**false**	
Y1520CH_G	**blue**	**0.0**	**999.0**	**-**	**1**	**none**	**true**	
rs1805009_G	**blue**	**0.0**	**999.0**	**-**	**1**	**none**	**false**	
rs1805007_G	**blue**	**0.0**	**999.0**	**-**	**1**	**none**	**false**	
rs1805006_G	**blue**	**0.0**	**999.0**	**-**	**1**	**none**	**false**	
rs1805008_A	**Green**	**0.0**	**999.0**	**-**	**1**	**none**	**false**	
rs683_A	**Green**	**0.0**	**999.0**	**-**	**1**	**none**	**false**	
rs1393350_A	**Green**	**0.0**	**999.0**	**-**	**1**	**none**	**false**	

rs12896399_A	Green	0.0	999.0	-	1	none	false
rs2378249_A	Green	0.0	999.0	-	1	none	false
rs12913832_A	Green	0.0	999.0	-	1	none	false
rs2402130_A	Green	0.0	999.0	-	1	none	false
rs1800407_A	Green	0.0	999.0	-	1	none	false
rs1042602_A	Green	0.0	999.0	-	1	none	false
rs12203592_A	Green	0.0	999.0	-	1	none	false
rs885479_A	Green	0.0	999.0	-	1	none	false
rs11547464_A	Green	0.0	999.0	-	1	none	false
N29insA_A	Green	0.0	999.0	-	1	none	false
rs4959270_A	Green	0.0	999.0	-	1	none	false
rs12821256_A	Green	0.0	999.0	-	1	none	false
rs16891982_A	Green	0.0	999.0	-	1	none	false
rs28777_A	Green	0.0	999.0	-	1	none	false
rs1110400_A	Green	0.0	999.0	-	1	none	false
rs2228479_A	Green	0.0	999.0	-	1	none	false
Y1520CH_A	Green	0.0	999.0	-	1	none	true
rs1805009_A	Green	0.0	999.0	-	1	none	false
rs1805007_A	Green	0.0	999.0	-	1	none	false
rs1805006_A	Green	0.0	999.0	-	1	none	false
rs1805005_A	Green	0.0	999.0	-	1	none	false
rs1805008_C	Yellow	0.0	999.0	-	1	none	false
rs683_T	Yellow	0.0	999.0	-	1	none	false
rs1393350_T	Yellow	0.0	999.0	-	1	none	false
rs12896399_T	Yellow	0.0	999.0	-	1	none	false
rs2378249_T	Yellow	0.0	999.0	-	1	none	false
rs12913832_T	Yellow	0.0	999.0	-	1	none	false
rs2402130_T	Yellow	0.0	999.0	-	1	none	false
rs1800407_T	Yellow	0.0	999.0	-	1	none	false
rs1042602_C	Yellow	0.0	999.0	-	1	none	false
rs12203592_C	Yellow	0.0	999.0	-	1	none	false

rs885479_C	Yellow	0.0	999.0	-	1	none	false
rs11547464_C	Yellow	0.0	999.0	-	1	none	false
N29insA_C	Yellow	0.0	999.0	-	1	none	false
rs4959270_C	Yellow	0.0	999.0	-	1	none	false
rs12821256_C	Yellow	0.0	999.0	-	1	none	false
rs16891982_C	Yellow	0.0	999.0	-	1	none	false
rs28777_C	Yellow	0.0	999.0	-	1	none	false
rs1110400_C	Yellow	0.0	999.0	-	1	none	false
rs2228479_C	Yellow	0.0	999.0	-	1	none	false
Y1520CH_C	Yellow	0.0	999.0	-	1	none	true
rs1805009_C	Yellow	0.0	999.0	-	1	none	false
rs1805007_C	Yellow	0.0	999.0	-	1	none	false
rs1805006_C	Yellow	0.0	999.0	-	1	none	false
rs1805005_C	Yellow	0.0	999.0	-	1	none	false
rs1805008_T	Red	0.0	999.0	-	1	none	false
rs1805005_T	Red	0.0	999.0	-	1	none	false
rs1805006_T	Red	0.0	999.0	-	1	none	false
rs1805007_T	Red	0.0	999.0	-	1	none	false
rs1805009_T	Red	0.0	999.0	-	1	none	false
Y1520CH_T	Red	0.0	999.0	-	1	none	true
rs2228479_T	Red	0.0	999.0	-	1	none	false
rs1110400_T	Red	0.0	999.0	-	1	none	false
rs28777_T	Red	0.0	999.0	-	1	none	false
rs16891982_T	Red	0.0	999.0	-	1	none	false
rs12821256_T	Red	0.0	999.0	-	1	none	false
rs4959270_T	Red	0.0	999.0	-	1	none	false
N29insA_T	Red	0.0	25.0561	-	1	none	false
rs11547464_T	Red	0.0	999.0	-	1	none	false
rs885479_T	Red	0.0	999.0	-	1	none	false
rs12203592_T	Red	0.0	999.0	-	1	none	false
rs1042602_T	Red	0.0	999.0	-	1	none	false

rs1800407_C	**Red**	**0.0**	**999.0**	**-**	**1**	**none**	**false**	
rs2402130_C	**Red**	**0.0**	**999.0**	**-**	**1**	**none**	**false**	
rs12913832_C	**Red**	**0.0**	**999.0**	**-**	**1**	**none**	**false**	
rs2378249_C	**Red**	**0.0**	**999.0**	**-**	**1**	**none**	**false**	
rs12896399_C	**Red**	**0.0**	**999.0**	**-**	**1**	**none**	**false**	
rs1393350_C	**Red**	**0.0**	**999.0**	**-**	**1**	**none**	**false**	
rs683_C	**Red**	**0.0**	**999.0**	**-**	**1**	**none**	**false**	

Supplemental Document 3 - HIrisPlex Bin set file for GeneMapper Software

`

Version GMID-X v 1.2

Chemistry Kit SNaPshot

BinSet Name HIrisPlex

Panel Name HIrisPlex

Marker Name rs1805005_G

A 36.0 0.5 0.5 virtual

Marker Name rs683_G

G 96.0 1.0 1.0 virtual

Marker Name rs1393350_G

G 93.0 1.0 1.0 virtual

Marker Name rs12896399_G

G 89.0 1.0 1.0 virtual

Marker Name rs2378249_G

G 87.0 1.0 1.0 virtual

Marker Name rs12913832_G

G 84.0 1.0 1.0 virtual

Marker Name rs2402130_G

G 81.0 1.0 1.0 virtual

Marker Name rs1800407_G

G 78.0 1.0 1.0 virtual

Marker Name rs1042602_G

G 74.0 1.0 1.0 virtual

Marker Name rs12203592_G

G 70.0 0.5 0.5 virtual

Marker Name rs1805008_G

G 34.0 0.5 0.5 virtual

Marker Name rs885479_G

G 35.0 0.5 0.5 virtual

Marker Name rs11547464_G

G 29.0 1.0 1.0 virtual

Marker Name N29insA_G

G 24.0 1.0 1.0 virtual

Marker Name rs4959270_G

G 69.0 0.5 0.5 virtual

Marker Name rs12821256_G

G 67.0 1.0 1.0 virtual

Marker Name rs16891982_G

G 63.0 1.0 1.0 virtual

Marker Name rs28777_G

G 59.0 1.0 1.0 virtual

Marker Name rs1110400_G

G 57.0 1.0 1.0 virtual

Marker Name rs2228479_G

G 55.0 1.0 1.0 virtual

Marker Name Y1520CH_G

G 52.0 1.0 1.0 virtual

Marker Name rs1805009_G

G 49.0 1.0 1.0 virtual

Marker Name rs1805007_G

G 45.0 1.0 1.0

Marker Name rs1805006_G

G 39.0 0.5 0.5 virtual

Marker Name rs1805008_A

A 34.0 0.5 0.5 virtual

Marker Name rs683_A

A 96.0 1.0 1.0 virtual

Marker Name rs1393350_A

A 93.0 1.0 1.0 virtual

Marker Name rs12896399_A

A 89.0 1.0 1.0 virtual

Marker Name rs2378249_A

A 87.0 1.0 1.0 virtual

Marker Name rs12913832_A

A 84.0 1.0 1.0 virtual

Marker Name rs2402130_A

A 81.0 1.0 1.0

Marker Name rs1800407_A

A 78.0 1.0 1.0 virtual

Marker Name rs1042602_A

A 74.0 1.0 1.0 virtual

Marker Name rs12203592_A

A 70.0 0.5 0.5 virtual

Marker Name rs885479_A

A 35.0 0.5 0.5 virtual

Marker Name rs11547464_A

A 29.0 1.0 1.0 virtual

Marker Name N29insA_A

A 24.0 1.0 1.0 virtual

Marker Name rs4959270_A

A 69.0 0.5 0.5 virtual

Marker Name rs12821256_A

A 67.0 1.0 1.0 virtual

Marker Name rs16891982_A

A 63.0 1.0 1.0 virtual

Marker Name rs28777_A

A 59.0 1.0 1.0 virtual

Marker Name rs1110400_A

A 57.0 1.0 1.0 virtual

Marker Name rs2228479_A

A 55.0 1.0 1.0 virtual

Marker Name Y1520CH_A

A 52.0 1.0 1.0 virtual

Marker Name rs1805009_A

A 49.0 1.0 1.0 virtual

Marker Name rs1805007_A

A 45.0 1.0 1.0 virtual

Marker Name rs1805006_A

A 39.0 0.5 0.5 virtual

Marker Name rs1805005_A

A 36.0 0.5 0.5 virtual

Marker Name rs1805008_C

C 34.0 0.5 0.5 virtual

Marker Name rs683_T

T 96.0 1.0 1.0 virtual

Marker Name rs1393350_T

T 93.0 1.0 1.0 virtual

Marker Name rs12896399_T

T 89.0 1.0 1.0 virtual

Marker Name rs2378249_T

T 87.0 1.0 1.0 virtual

Marker Name rs12913832_T

T 84.0 1.0 1.0 virtual

Marker Name rs2402130_T

T 81.0 1.0 1.0 virtual

Marker Name rs1800407_T

T 78.0 1.0 1.0 virtual

Marker Name rs1042602_C

C 74.0 1.0 1.0 virtual

Marker Name rs12203592_C

C 70.0 0.5 0.5 virtual

Marker Name rs885479_C

C 35.0 0.5 0.5 virtual

Marker Name rs11547464_C

C 29.0 1.0 1.0 virtual

Marker Name N29insA_C

C 24.0 1.0 1.0 virtual

Marker Name rs4959270_C

C 69.0 0.5 0.5 virtual

Marker Name rs12821256_C

C 67.0 1.0 1.0 virtual

Marker Name rs16891982_C

C 63.0 1.0 1.0 virtual

Marker Name rs28777_C

C 59.0 1.0 1.0 virtual

Marker Name rs1110400_C

C 57.0 1.0 1.0 virtual

Marker Name rs2228479_C

C 55.0 1.0 1.0 virtual

Marker Name Y1520CH_C

C 52.0 1.0 1.0 virtual

Marker Name rs1805009_C

C 49.0 1.0 1.0 virtual

Marker Name rs1805007_C

C 45.0 1.0 1.0 virtual

Marker Name rs1805006_C

C 39.0 0.5 0.5 virtual

Marker Name rs1805005_C

C 36.0 0.5 0.5 virtual

Marker Name rs1805008_T

T 34.0 0.5 0.5 virtual

Marker Name rs1805005_T

T 36.0 0.5 0.5 virtual

Marker Name rs1805006_T

T 39.0 0.5 0.5 virtual

Marker Name rs1805007_T

T 45.0 1.0 1.0 virtual

Marker Name rs1805009_T

T 49.0 1.0 1.0 virtual

Marker Name Y1520CH_T

T 52.0 1.0 1.0 virtual

Marker Name rs2228479_T

T 55.0 1.0 1.0 virtual

Marker Name rs1110400_T

T 57.0 1.0 1.0 virtual

Marker Name rs28777_T

T 59.0 1.0 1.0 virtual

Marker Name rs16891982_T

T 63.0 1.0 1.0 virtual

Marker Name rs12821256_T

T 67.0 1.0 1.0 virtual

Marker Name rs4959270_T

T 69.0 0.5 0.5 virtual

Marker Name N29insA_T

T 24.0 1.0 1.0 virtual

Marker Name rs11547464_T

T 29.0 1.0 1.0 virtual

Marker Name rs885479_T

T 35.0 0.5 0.5 virtual

Marker Name rs12203592_T

T 70.0 0.5 0.5 virtual

Marker Name rs1042602_T

T 74.0 1.0 1.0 virtual

Marker Name rs1800407_C

C 78.0 1.0 1.0 virtual

Marker Name rs2402130_C

C 81.0 1.0 1.0

Marker Name rs12913832_C

C 84.0 1.0 1.0 virtual

Marker Name rs2378249_C

C 87.0 1.0 1.0 virtual

Marker Name rs12896399_C

C 89.0 1.0 1.0 virtual

Marker Name rs1393350_C

C 93.0 1.0 1.0 virtual

Marker Name rs683_C

C 96.0 1.0 1.0 virtual

Supplemental Document 4 - R Script for Five Class Classification Random Forest Regression Model

`

```
#install.packages("tidyverse")
#install.packages("janitor")

library("tidyverse")
library("readxl")


data <- R_Code_Data_sheet

original_names <- colnames(data)

data <- data %>%
  janitor::clean_names()

new_names <- colnames(data)
names(new_names) <- original_names
names(original_names) <- new_names



str(data)
data$color <-as.factor(data$color)
table(data$color)

#Data Partition
set.seed(123)
#ind for Independent samples
ind <-sample(2, nrow(data), replace = TRUE, prob = c(0.7,0.3))
train <- data[ind==1,]
test  <- data[ind==2,]
```

```
#Random Forest
library("randomForest")
#rfNews()
set.seed(222)
rf <- randomForest( color ~ . , data=train,importance=TRUE, ntree = 500)
print(rf)

attributes(rf)
rf$confusion


#Dr. Jelsema plt code
rf_out01 <- importance( rf  )
rf_out02 <- data.frame(rf_out01) %>% rownames_to_column()
rf_out03 <- rf_out02 %>% arrange( -MeanDecreaseGini ) %>%
  mutate( Variable = recode_factor( rowname, !!!original_names )) %>%
  dplyr::select( Variable, MeanDecreaseAccuracy, MeanDecreaseGini )



ggplot( rf_out03 , aes(x=reorder(Variable, MeanDecreaseGini), y=MeanDecreaseGini ) ) +
  geom_bar( stat="identity") +
  coord_flip() +
  labs( x="", y="Mean decrease in Gini Index")
```

```
ggplot( rf_out03 , aes(x=reorder(Variable, MeanDecreaseAccuracy),
y=MeanDecreaseAccuracy ) ) +
  geom_bar( stat="identity") +
  coord_flip() +
  labs( x="", y="Mean decrease in OOB Error")

#prediction & Confusion Matrix - train
library(caret)
p1 <- predict(rf, train)
head(p1)
head(train$color)
confusionMatrix(p1, train$color)

#prediction & Confusion Matrix - test
p2 <- predict(rf, test)
confusionMatrix(p2, train$color)

#error rate of Random Forest
plot(rf)
```

Supplemental Document 5 - R Script for Three Class Classification Random Forest Regression Model

`

```
data <- R_Code_Data_sheet_all_intermeidate_as_3

#original_names <- colnames(data)

#data <- data %>%
#janitor::clean_names()

#new_names <- colnames(data)
#names(new_names) <- original_names
#names(original_names) <- new_names



str(data)
data$Color <-as.factor(data$Color)
table(data$Color)

#Data Partition
set.seed(123)
#ind for Independent samples
ind <-sample(2, nrow(data), replace = TRUE, prob = c(0.7,0.3))
train <- data[ind==1,]
test  <- data[ind==2,]



#Random Forest
library("randomForest")
set.seed(222)
rf <- randomForest( Color ~ . , data=train,importance=TRUE )
print(rf)
```

```
#round two

#install.packages("tidyverse")
#install.packages("janitor")

library("tidyverse")
library("readxl")


data <- R_Code_Data_sheet_all_intermeidate_as_3

original_names <- colnames(data)

data <- data %>%
  janitor::clean_names()

new_names <- colnames(data)
names(new_names) <- original_names
names(original_names) <- new_names



str(data)
data$color <-as.factor(data$color)
table(data$color)

#Data Partition
set.seed(123)
#ind for Independent samples
ind <-sample(2, nrow(data), replace = TRUE, prob = c(0.7,0.3))
```

```
train <- data[ind==1,]
test  <- data[ind==2,]




#Random Forest
library("randomForest")
#rfNews()
set.seed(222)
rf <- randomForest( color ~ . , data=train,importance=TRUE, ntree = 500)
print(rf)

attributes(rf)
rf$confusion


#Dr. Jelsema plt code
rf_out01 <- importance( rf  )
rf_out02 <- data.frame(rf_out01) %>% rownames_to_column()
rf_out03 <- rf_out02 %>% arrange( -MeanDecreaseGini ) %>%
  mutate( Variable = recode_factor( rowname, !!!original_names )) %>%
  dplyr::select( Variable, MeanDecreaseAccuracy, MeanDecreaseGini )



ggplot( rf_out03 , aes(x=reorder(Variable, MeanDecreaseGini), y=MeanDecreaseGini ) ) +
  geom_bar( stat="identity") +
  coord_flip() +
```

```
  labs( x="", y="Mean decrease in Gini Index")


ggplot( rf_out03 , aes(x=reorder(Variable, MeanDecreaseAccuracy),
y=MeanDecreaseAccuracy ) ) +
  geom_bar( stat="identity") +
  coord_flip() +
  labs( x="", y="Mean decrease in OOB Error")

#prediction & Confusion Matrix - train
library(caret)
p1 <- predict(rf, train)
head(p1)
head(train$color)
confusionMatrix(p1, train$color)

#prediction & Confusion Matrix - test
p2 <- predict(rf, test)
confusionMatrix(p2, train$color)

#error rate of Random Forest
plot(rf)
```