Hair Color Informative SNPs

Shehreen Hassan, Shatakshi Shewale, Allison Shannon

Background

Phenotyping from informative SNPs is becoming an important tool in forensics. Previously DNA profiling has been used in forensic analysis, through DNA database searches and matching. This has been beneficial but remained unhelpful in cases with unidentified suspects. As a result, the forensic community has begun to use forensic DNA phenotyping (FDP), which aims to construct an image of the suspect through phenotyping externally visible characteristics (EVCs). FDP is an important tool now being used to evaluate unidentified DNA from crime scenes to paint a picture and narrow the search of suspects (Söchtig et al., 2015).

In FDP pigmentation traits are some of the most variable of human phenotypes, which makes them informative characteristics. In terms of hair color, differentiation in phenotypes is mainly determined by the levels of melanin production. There are 2 types of melanin that influence hair color, pheomelanin, which ranges from yellow to reddish-brown, and eumelanin, which accounts for brown to black pigmentation (Söchtig et al., 2015).

Variation in hair pigmentation is generally confined to European populations, ranging from the lightest white-blonde to red to the deepest black. The eastern Baltic region is known to contain the largest variation of phenotypic range for hair color. Outside of Europe, hair color is mainly black with the only exception occurring in the Near East and Melanesian populations (Söchtig et al., 2015).

Data on hair color was obtained from 23andme files from openSNP. In these files hair color was self-reported. Only files that were self-reported to have specifically blonde, red, brown, and black hair were evaluated and counted. This was done because in these 23andme files people reported to have a range of hair colors like dirty blonde, strawberry blonde, auburn, light brown, dark brown, and many more. We acknowledge that this likely impacts results because hair color was self-reported and factors like culture and environment play a role in how people report their hair color. Upon evaluation of files, we found 19 files that reported red hair, 46 reported blonde hair, 65 reported black hair, and 153 reported brown hair, for a total of 283 files.

Through literature review it was discovered that HIrisPlex, HIrisPlexS, and Snipper were the most commonly used informative panels for hair color. These panels have been the basis of multiple human phenotypic pigmentation studies. Snipper contains 12 SNPs, HIrisPlex contains 24, and HIrisPlexS contains 41 (the 24 from HIrisPlex plus 18 more). HIrisPlex was created as a DNA Prediction tool originally used to predict eye color, mainly blue and brown, but was found to be useful in predicting hair color as well. It is particularly good at predicting red hair color when looking at European populations (Visser, Kayser, & Palstra, 2012). Snipper contains the 12 most strongly associated SNPs to hair pigmentation in 6 different genes (Söchtig et al., 2015). It should be noted that HIrisPlex and HIrisPlexS contain 2 more SNPs associated with MC1R than Snipper, rs2228479 and rs1805005.

Common genes associated with hair color and pigmentation in humans include MC1R, HERC2, OCA2, SLC24A5, SLC24A4, SLC45A2, and IRF4 (described in more depth later). Each of these genes contributes to the complex interplay of genetic factors determining hair color variation across different populations. Also, hair color is affected by complex relationships between a variety of genes and environment which can make it difficult to identify specific genes that contribute heavily to hair color.

Methods

The methods described below were done for each panel. The exact same code was used for all panels just with varying input files and pathways. For data loading, the paths to the folder containing the data files and the annotation file (list containing the SNPs in each panel) are defined. The annotation file is read in binary mode and reads its contents. Each line in the file is iterated over and decoded using the detected encoding, splits it into parts, and appends the first part to a list. This list is used to track the SNPs of interest. After collecting the SNPs from the annotation file, missing data for each SNP is tracked. Each data file is opened, split its name to extract additional information, and process its content. If the RSID matches one of the SNPs of interest, it updates the SNPs dictionary with the genotype information. Once all files have been processed, the script constructs a DataFrame using the collected data. It uses the SNPs and nucleotides to form column names and the file legends (extracted from file names) as row indices.

PCA:

After the data is loaded, PCA was conducted. This technique reduces the dimensionality of the data while preserving as much variance as possible. We computed the principal components of the standardized data to identify the directions of maximum variance. To select the appropriate number of principal components, the 95% cumulative explained variance ratio against the number of components was calculated. PCA loadings represent the coefficients of the original variables (SNPs) in the principal component space and are used to analyze which SNPs contribute the most to each principal component or check how much a specific SNP contributes to each principal component.

Clustering:

The K-means clustering algorithm was implemented to group the data into clusters, with the number of clusters (k=4) determined based on the knowledge that there were 4 different hair colors. The K-means algorithm was applied to the principal components to assign each data point to one of the four clusters. The first two principal components were plotted to visualize the clustering results, utilizing different colors to represent different clusters. Furthermore, the number of occurrences of each hair color were counted in each cluster, and the percentage distribution of each hair color within each cluster was calculated to understand the composition. To evaluate the clustering performance, metrics such as Adjusted Rand Index (ARI), Normalized Mutual Information (NMI), and V-measure were employed.

Predictive Models:

Four predictive models were used in total: baseline classifier (based off frequencies only), random forest, K-Nearest Neighbors (KNN), and support vector machine (SVM). Each model was run for 100 trials. Each trial encompassed splitting the dataset into training and testing sets with a training size of 90%. Stratification was done to ensure that all the classes are present in the Without class balancing, predictions were made based on a uniform distribution across all classes. Conversely, for balanced classes, class frequencies were derived from the training set to assign different weights to each class. The model's performance was evaluated by computing the overall accuracy and accuracy for each hair color.

For some of the models, parameter optimization was done through initial grid searches. Grid search with 5-fold cross-validation was used to optimize hyperparameters, including the number of trees, maximum tree depth, minimum samples required to split a node, and minimum samples required at each leaf node for the random forest model. Similarly, for the SVM model with a linear kernel, a grid search with 5-fold cross-validation was used to optimize the regularization parameter C within a specified range. For KNN a grid search with 5-fold cross-validation was used to optimize the number of neighbors (k) within a specified range. Parameter optimization is required since the training and test sets are split randomly.

Results & Discussion

PCA:

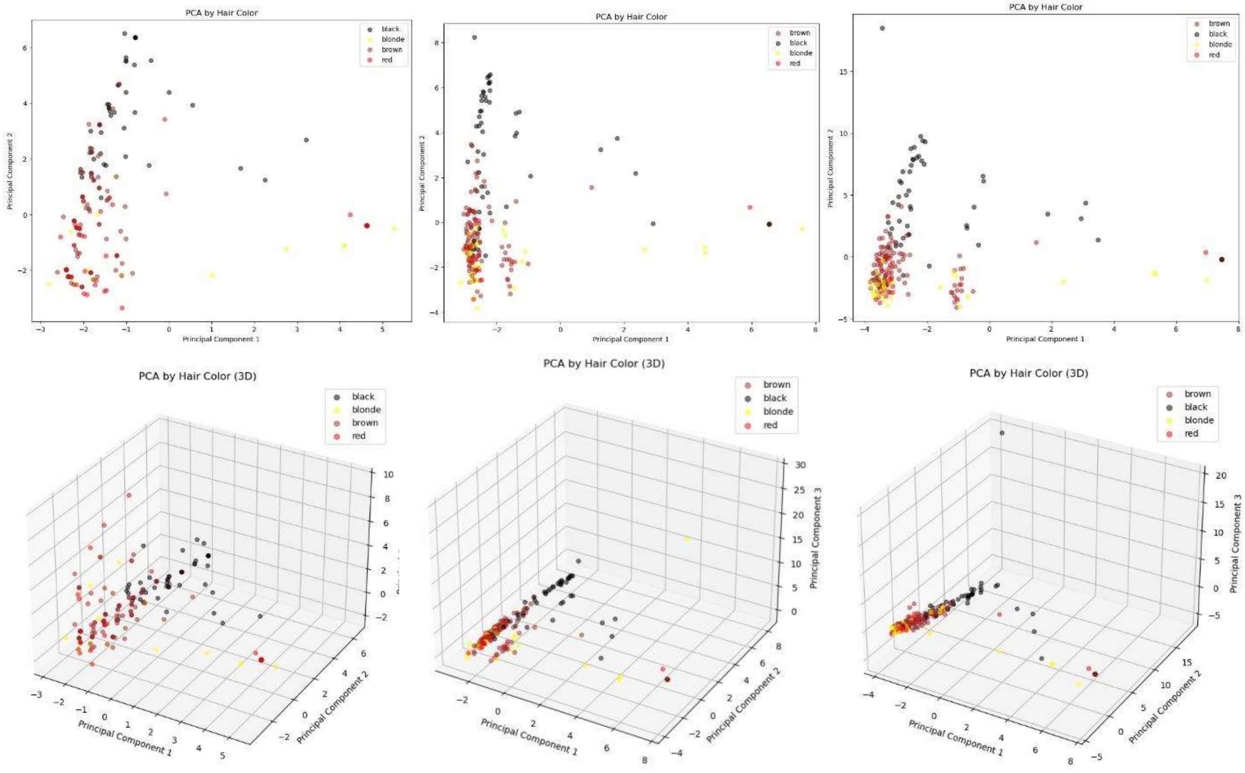

Figure 1: Principal components 1 and 2 plotted against each other (top) and components 1,2, and 3 plotted agaisnt each other (bottom) ordered: Snipper (left), HIrisPlex (center), and HIrisPlexS (right)

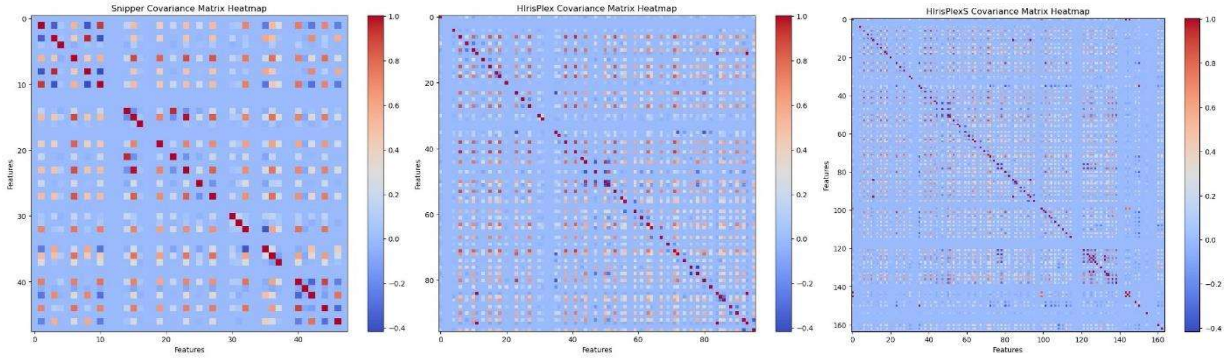

Figure 2: Covariance matrix ordered: Snipper (left), HIrisPlex (center), and HIrisPlexS (right)

Seen in Figure 1, the PCA shows that the principal component 2 axis is responsible for most of the spread. It is of note that for Snipper, there is more spread in the principal component

3 axis which is not present in HIrisPlex and HIrisPlexS. Additionally, in all three PCA's there is a slight sinusoidal pattern that can be seen. This is due to autocorrelation resulting from the hot encoding method used for loading the data and the orthogonal nature of the principal components. This can be seen in Figure 2, where there are square patterns in the covariance matrix.

Highest Rank SNPs (Snipper)			
PC	SNPs	Allele	Contribution
PC1	rs11547464	G	-0.328208489
PC2	rs12913832	A	0.4015372173
PC3	rs12931267	G	0.6317027430
PC4	rs1805008	T	0.4673126054
PC5	rs1805009	G	0.5735379982

Table 1: for the Snipper panel

Highest Rank SNPs (HIrisPlex)			
PC	SNPs	Allele	Contribution
PC1	rs1110400	T	-0.21122
PC2	rs1426654	C	0.275845
PC3	rs201326893	A	0.414619
PC4	rs312262906	C	0.353915
PC5	rs1667394	T	0.540973

Table 2: for the HIrisPlex panel

Highest Rank SNPs (HIrisPlexS)			
PC	SNPs	Allele	Contribution
PC1	rs1110400	T	-0.2442019
PC2	rs16891982	G	0.39764631
PC3	rs12896399	C	0.57153431
PC4	rs201326893	A	0.56665994
PC5	rs12203592	A	0.29448568

Table 3: for the HIrisPlexS panel

It was found that the SNPs rs12931267, rs1110400, rs16891982, rs12896399, rs201326893, rs12203592, rs1426654, rs312262906, and rs1667394 to be among the top 5 most significant results among the 3 panels. Notably, rs1110400, rs201326893, and rs312262906 are associated with MC1R gene, rs16891982 is associated with SLC45A2, rs12896399 is associated with SLC24A4, and rs1426654 is associated with SLC24A5. Additionally, rs1667394 is associated with HERC2, and rs1220355592 is associated with IRF4. After researching GWAS and ClinVar, it was found that rs16891982 is not associated with a known gene (National Human Genome Research Institute, n.d.).

The MC1R gene plays a crucial role in determining hair and skin color. It is linked with eumelanin production (brown/black pigment) and pheomelanin (red/yellow pigment). Variants in this gene, such as rs1805006, rs1805007, rs1805008, and rs1805009, are strongly linked to red hair color and fair skin (Söchtig et al., 2015). MC1R variants are among the first polymorphisms discovered to be associated with hair pigmentation (Rees, 2004). The SLC45A2 gene is involved in melanin synthesis, affecting the transport of tyrosine, an essential precursor for melanin production. Variants such as rs28777 and rs16891982 are associated with pigmentation traits, particularly influencing lighter skin and hair color (Söchtig et al., 2015). This gene is especially significant in determining pigmentation among Caucasian populations (Graf et al., 2005). HERC2 is on chromosome 15 and linked to hair, eye, and skin color. It can interact with OCA2 and SLC24A4 to regulate pigmentation. The SNP rs12913831 within HERC2 has a well-documented association with lighter hair and eye color (Sturm & Duffy, 2012). SLC24A4 is involved in the regulation of melanosome pH and the processing of melanin. The SNP rs12896399 within this gene is associated with pigmentation variation, particularly in determining lighter hair color (Sulem et al., 2008). Like SLC24A4, SLC24A5 plays a critical role in melanin production and pigmentation. The SNP rs1426654 is a key determinant of skin pigmentation differences between populations, significantly influencing hair color as well (Lamason et al., 2005). IRF4 is associated with hair, eye, and skin pigmentation. The SNP rs1220355592 has been linked to these traits through genome-wide association studies (GWAS). IRF4 regulates melanin synthesis and interacts with other pigmentation genes (Praetorius et al., 2013).

Common SNPs Contribution (Allele A)								
Panel	PC	rs1805007	rs1805009	rs28777	rs12913832	rs11547464	rs1805008	rs1805006
Snipper (Rank/48)	1	-1.58E-32 (33)	0 (43)	-0.28572 (8)	-0.12895 (13)	-0.0275 (23)	1.78E-32 (31)	-0.01561 (25)
	2	-1.04E-29 (36)	0 (40)	-0.15352 (9)	0.401537 (1)	-0.01108 (23)	1.16E-29 (34)	-0.00466 (25)
	3	1.02E-24 (38)	0 (44)	0.021286 (20)	0.027126 (16)	-0.07436 (10)	-1.15E-24 (36)	-0.02264 (18)
HIrisPlex (Rank/96)	1	0.0 (74)	0.0 (95)	-0.199 (12)	-0.111 (26)	-0.018 (45)	-2.52E29 (54)	-0.009 (48)
	2	0.0 (90)	0.0 (79)	-0.236 (7)	0.295 (3)	-0.021 (38)	0.0 (71)	-0.019 (41)
	3	-8.08E-28 (61)	-7.06E-29 (63)	0.002 (34)	0.010 (10)	-0.008 (14)	8.67E-19 (56)	0.004 (27)
HIrisPlexS (Rank/164)	1	0.0 (95)	0.0 (98)	-0.174 (14)	-0.092 (35)	-0.014 (66)	0.0 (101)	-0.008 (73)
	2	0.0 (97)	3.587 (87)	-0.154 (17)	0.244 (5)	-0.013 (67)	0.0 (92)	-0.003 (78)
	3	-2.52E-29 (87)	0.0 (95)	0.052 (38)	-0.038 (53)	0.048 (42)	0.0 (154)	-0.0 (75)

Table 4: Common 7 SNPs loading data for the first three principal components for allele A

Common SNPs Contribution (Allele T)								
Panel	PC	rs1805007	rs1805009	rs28777	rs12913832	rs11547464	rs1805008	rs1805006
Snipper (Rank/48)	1	-0.05269 (21)	0 (34)	0 (47)	0 (40)	0 (36)	-0.08111 (15)	2.66E-32 (29)
	2	-0.08316 (12)	0 (39)	0 (38)	0 (45)	0 (47)	-0.06922 (16)	1.74E-29 (32)
	3	0.062340 (30)	0 (43)	0 (40)	-2.17E-19 (30)	-1.11E-16 (26)	-0.18141 (6)	-1.71E-24 (34)
HIrisPlex (Rank/96)	1	-0.041735 (40)	0 (93)	0 (77)	0 (60)	0 (55)	-0.058964 (32)	0 (59)
	2	-0.088927 (22)	0 (74)	0 (80)	0 (63)	0 (61)	-0.10665 (16)	0 (60)
	3	-0.002728 (37)	3.48E-29 (67)	6.26E-30 (72)	0 (93)	-2.08E-17 (55)	0.01041 (11)	6.20E-25 (59)
HIrisPlexS (Rank/164)	1	0.01041 (56)	0 (148)	0 (143)	0 (136)	0.01041 (84)	-0.05261 (48)	0 (93)
	2	-0.05261 (38)	-1.12E-44 (88)	0 (126)	0 (129)	-6.94E-18 (83)	-6.94E-18 (34)	-3.85E-34 (85)
	3	0.022762 (56)	0 (94)	0 (127)	0 (125)	-3.33E-16 (81)	-0.03448 (55)	0 (113)

Table 5: Common 7 SNPs loading data for the first three principal components for allele T

Common SNPs Contribution (Allele G)								
Panel	PC	rs1805007	rs1805009	rs28777	rs12913832	rs11547464	rs1805008	rs1805006
Snipper (Rank/48)	1	-2.06E-32 (40)	-0.05947 (17)	0 (46)	-0.22746 (11)	-0.32821 (1)	0 (45)	1.62Ee-32 (32)
	2	-1.35E-29 (33)	0.03427 (21)	0 (37)	-0.3075 (6)	0.049374 (20)	0 (43)	1.06E-29 (35)
	3	1.32E-24 (35)	0.208399 (5)	0 (39)	-0.00348 (24)	0.040106 (14)	0 (45)	-1.06E-24 (37)
HIrisPlex (Rank/96)	1	0 (73)	-0.043044 (38)	0 (72)	-0.150874 (19)	-0.2375589 (2)	0 (67)	0 (57)
	2	0 (87)	0.062259 (28)	0 (81)	-0.2534682 (5)	0.007623 (43)	0.007623 (53)	6.16E-33 (54)
	3	1.26Ee-29 (70)	0.003823 (30)	1.58E-29 (69)	-0.007516 (17)	0.01136 (9)	5.42E-20 (57)	5.16E-26 (60)
HIrisPlexS (Rank/164)	1	0 (96)	-0.03688 (55)	0 (142)	-0.13334 (25)	-0.206268 (2)	0 (87)	0 (94)
	2	0 (118)	0.047561 (41)	0 (128)	0.047561 (8)	0.015527 (65)	0 (98)	-1.20E-35 (86)
	3	0 (90)	0.063891 (32)	0 (130)	0.086987 (18)	0.012873 (67)	0 (129)	-1.62E-27 (86)

Table 6: Common 7 SNPs loading data for the first three principal components for allele G

Common SNPs Contribution (Allele C)								
Panel	PC	rs1805007	rs1805009	rs28777	rs12913832	rs11547464	rs1805008	rs1805006
Snipper (Rank/48)	1	-0.31536 (3)	-0.02697 (24)	-0.05318 (20)	0 (35)	0 (38)	-0.30932 (4)	-0.2804 (9)
	2	0.08252 (13)	-0.03038 (22)	0.3219 (5)	-1.65E-24 (29)	-5.42E-20 (28)	0.074807 (15)	0.056095 (19)
	3	-0.21031 (4)	-0.03688 (15)	-0.00103 (25)	-1.36E-20 (31)	6.93E-18 (28)	0.093347 (8)	0.02430 (17)

HIrisPlex (Rank/96)	1	-0.22569 (6)	-0.01615 (46)	-0.05249 (35)	0 (62)	2.11E-22 (53)	-0.224362 (8)	-0.20779 (11)
	2	0.04451437 (33)	-0.035611 (36)	0.3945153941 (2)	0 (64)	0 (94)	0.394515 (31)	0.024815 (37)
	3	0.00558808 (23)	0.0027664 (36)	0.00022912 (49)	0 (92)	1.11E-16 (52)	-0.000526 (47)	0.0356889 (6)
HIrisPlexS (Rank/164)	1	-0.19462 (6)	-0.01422 (67)	-0.04117 (53)	0 (134)	1.69E-21 (85)	-0.19325 (8)	-0.18812 (9)
	2	0.033768 (48)	-0.02153 (58)	0.26745 (3)	0 (127)	0 (111)	0.034323 (47)	0.031965 (51)
	3	0.042675 (48)	-0.00798 (73)	-0.01118 (70)	0 (124)	-1.39E-17 (83)	0.074635 (27)	-0.17287 (10)

Table 7: Common 7 SNPs loading data for the first three principal components for allele C

To examine the theory that the SNPs common to all three panels would contribute significantly to the variance of the data, the contribution of the 7 SNPs was analyzed. Tables 4-7 showcase the contribution rank of each of 7 SNPs for the first three principal components of each panel; however, all principal components were used for the analysis (not displayed due to space).

SNPs rs28777 and rs12913832 for allele A both contribute to significant variance in the data especially in earlier principal components. Rs28777 is related to the OCA2 gene and rs12913832 is related to the HERC2 gene (regulates OCA2) which both involve the production of melanin in the iris. Although these two genes are mainly tied to affecting eye color, there has been some research done to indicate that they also play a role in hair color. Overall, it is important to note that the relationship between these SNPs and hair color is less established and understood compared to their association with eye color. Hair color is influenced by a complex interplay of multiple genes and environmental factors, making it challenging to pinpoint specific genetic contributions definitively. All 7 SNPs related to the T allele contributed the least overall to the variance of the data. This pattern held relatively true for all panels. For allele G, SNPs rs12913832 and rs11547464 play a relatively significant role throughout all the principal components. Rs 11547464 is related to the MC1R gene described in the background of the report. SNPs rs1805007 and rs1805006 related to the C allele both contribute significantly to all three panels and are related to the TNF gene which is primarily associated with immune responses and inflammation rather than directly influencing hair color. However, it's worth noting that some genes can have pleiotropic effects and there could be indirect relationships between TNF variants and hair color through complex biological pathways. In later PCs, all the 7 SNPs play a relatively less significant role for all the panels.

Once again, it is important to note that hair color is affected by complex relationships between a variety of genes and environment which can make it difficult to identify specific genes that contribute heavily to hair color. This can be seen by the fact that for all three of the panels, their highest contributing SNPs do not overlap with each other significantly and the associated genes vary in function from melanin production to immune regulation.

Clustering:

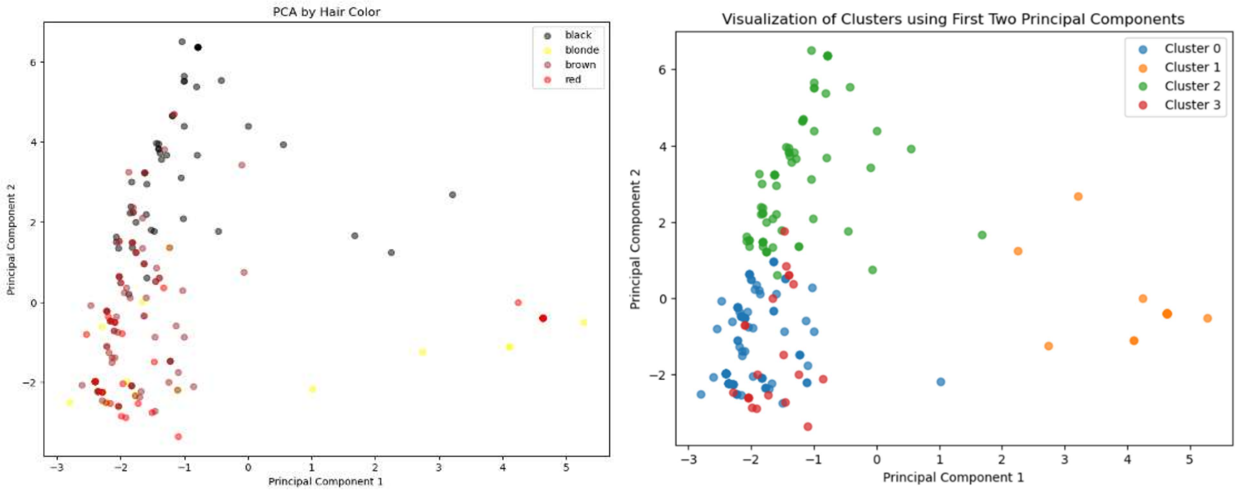

Figure 3: True Labels (left) vs Cluster Labels (right) for the Snipper panel.

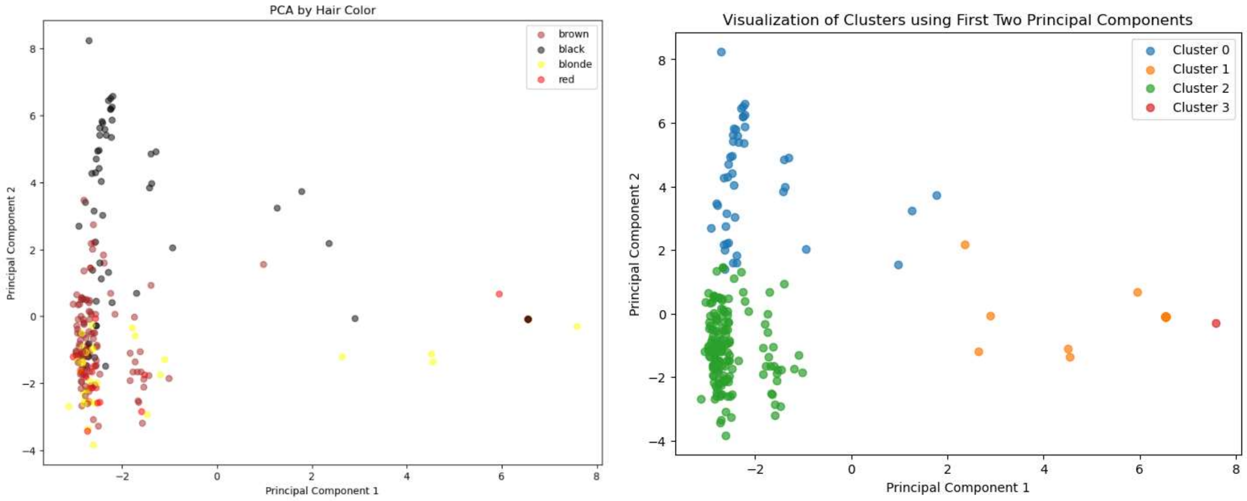

Figure 4: True Labels (left) vs Cluster Labels (right) for the HIrisPlex panel.

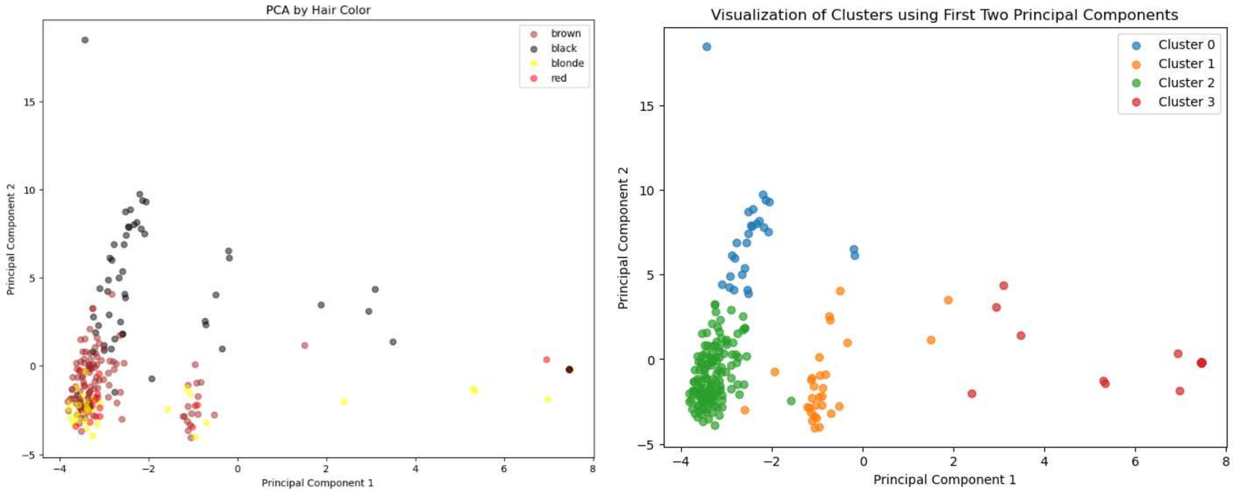

Figure 5: True Labels (left) vs Cluster Labels (right) for the HIrisPlexS panel.

Using K-means clustering with k=4, the dataset was divided into four distinct clusters. The first two principal components were plotted to visualize the clusters, revealing the separation among them. As shown in Figures 3, 4, and 5, each panel (Snipper, HIrisPlex, and HIrisPlexS) displays different cluster formations. The PCA plots anticipated a mixture of blonde and red clusters with distinct black and brown clusters. However, the actual clustering results showed a more mixed distribution of hair colors within each cluster.

The hair color distribution within each cluster was analyzed for the three panels. For the Snipper panel, the clusters were somewhat distinct but still exhibited some mixing of hair colors (Figure 3). The HIrisPlex and HIrisPlexS panels showed highly mixed distributions, with no clear clusters containing a single hair color (Figures 4 and 5). The distribution of hair colors within each cluster indicated that while some clusters had predominant hair colors, there was significant overlap, reflecting the complexity of genetic variation in hair color.

Panels	Adjusted Rand Index	Normalized Mutual Information	V-measure
Snipper	0.1271	0.1633	0.1633
HIrisPlex	0.1058	0.134	0.134
HIrisPlexS	0.1548	0.1755	0.1755

Table 8: Cluster evaluation metrics

The clustering performance was evaluated using Adjusted Rand Index (ARI), Normalized Mutual Information (NMI), and V-measure, as shown in the table. These metrics compared the cluster labels with the hair color labels, providing insights into the clustering performance against the ground truth. The ARI scores indicated a moderate level of agreement between the true labels and the cluster labels, with the HIrisPlexS panel having the highest ARI score (0.1548). The NMI and V-measure scores were also highest for the

HIrisPlexS panel (0.1755), suggesting a better alignment with the true hair color categories compared to the other panels. However, all scores were modest, indicating that the clusters were mixed and not perfectly aligned with the actual hair colors.

<u>Predictive Models:</u>

Average of 100 Baseline Accuracies (No Class Balancing)			
	Snipper	HIrisPlex	HIrisPlexS
Black	0.27	0.24	0.27
Blonde	0.28	0.24	0.23
Brown	0.25	0.25	0.25
Red	0.26	0.24	0.26
Overall	0.262	0.245	0.254

Average of 100 Baseline Accuracies (Class Balancing)			
	Snipper	HIrisPlex	HIrisPlexS
Black	0.26	0.25	0.24
Blonde	0.27	0.25	0.28
Brown	0.25	0.24	0.26
Red	0.24	0.26	0.24
Overall	0.256	0.252	0.254

Table 9: Class Accuracies for 100 trials using a baseline classifier model utilizing frequencies

Average of 100 SVM Accuracies (No Class Balancing)			
	Snipper	HIrisPlex	HIrisPlexS
Black	0.60	0.63	0.67
Blonde	0.03	0.33	0.31
Brown	0.95	0.85	0.88
Red	0.23	0.17	0.16
Overall	0.668	0.667	0.689

Average of 100 SVM Accuracies (Class Balancing)			
	Snipper	HIrisPlex	HIrisPlexS
Black	0.67	0.67	0.50
Blonde	0.60	1.00	0.80
Brown	0.19	0.38	0.50
Red	0.50	0.00	0.00
Overall	0.379	0.517	0.517

Table 10: Class Accuracies for 100 trials using SVM model

Average of 100 KNN Accuracies (No Class Balancing)			
	Snipper	HIrisPlex	HIrisPlexS
Black	0.58	0.49	0.47
Blonde	0.06	0.03	0.08
Brown	0.91	0.94	0.94
Red	0.01	0.01	0.00
Overall	0.634	0.624	0.627

Average of 100 KNN Accuracies (Class Balancing)			
	Snipper	HIrisPlex	HIrisPlexS
Black	0.67	0.51	0.51
Blonde	0.19	0.04	0.08
Brown	0.82	0.95	0.93
Red	0.09	0.00	0.01
Overall	0.631	0.633	0.630

Table 11: Class Accuracies for 100 trials using KNN model

Average of 100 Random Forest Accuracies (No Class Balancing)			
	Snipper	HIrisPlex	HIrisPlexS
Black	0.61	0.55	0.58

Average of 100 Random Forest Accuracies (Class Balancing)			
	Snipper	HIrisPlex	HIrisPlexS
Black	0.65	0.58	0.66

Blonde	0.11	0.03	0.03
Brown	0.93	0.94	0.95
Red	0.08	0.00	0.00
Overall	0.661	0.634	0.651

Blonde	0.74	0.44	0.45
Brown	0.46	0.61	0.65
Red	0.25	0.13	0.11
Overall	0.534	0.541	0.581

Table 12: Class Accuracies for 100 trials using random forest model

Compared to the baseline classifier where labels were randomly assigned based of class frequency, all the predictive models whether the classes were balanced or unbalanced performed better overall. However, for certain hair colors mainly red, the other predictive models tend to have lower accuracies when compared to the baseline. As expected, since there is no learning taking place in the baseline classifier all the class accuracies are around the same for each class. Additionally, whether the classes are balanced does not matter since the frequencies of each class are used to make predictions. For all the predictive models, Snipper was better at predicting red hair. This is likely due to the extra spread of principal components along the 3rd PC's axis containing mostly red hair. This spread is not present for the HIrisPlex and HIrisPlexS panels making the red hair color harder to predict. Overall, when the classes were balanced, the predictive models overall had a lower accuracy but the accuracies for blonde and red hair tended to increase. The reason for the overall lower accuracies is the drop in accuracies for brown hair. Black hair predictions tended to stay about the same. There is no clear classifier that is better than the other. The overall increase in accuracies compared to the baseline classifier indicated that the models do learn some information. Due to the overall low number of datapoints, the predictive models are not very reliable.

References

1. Söchtig, J., Phillips, C., Maroñas, O., Gómez-Tato, A., Cruz, R., Alvarez-Dios, J., Casares de Cal, M. A., Ruiz, Y., Reich, K., Fondevila, M., Carracedo, Á., & Lareu, M. V. (2015). Exploration of SNP variants affecting hair colour prediction in Europeans. *International Journal of Legal Medicine, 129*(5), 963-975. https://doi.org/10.1007/s00414-015-1226-y

2. Visser, M., Kayser, M., & Palstra, R. J. (2012). The common occurrence of epistasis in the determination of human pigmentation. *Genome Research, 22*(3), 446-455. https://doi.org/10.1101/gr.128652.111

3. Kastelic, V., & Drobnic, K. (2011). Single multiplex system of twelve SNPs: Validation and implementation for association of SNPs with human eye and hair color. *Forensic Science International: Genetics Supplement Series, 3*, e216-e217. https://doi.org/10.1016/j.fsigss.2011.08.108

4. National Human Genome Research Institute. (n.d.). GWAS Catalog. Retrieved May 14, 2024, from https://www.ebi.ac.uk/gwas/

5. National Center for Biotechnology Information. (n.d.). ClinVar. Retrieved May 14, 2024, from https://www.ncbi.nlm.nih.gov/clinvar/

6. Rees, J. L. (2004). The genetics of sun sensitivity in humans. The American Journal of Human Genetics, 75(5), 739-751.

7. Graf, J., Hodgson, R., & van Daal, A. (2005). Single nucleotide polymorphisms in the MATP gene are associated with normal human pigmentation variation. Human Mutation, 25(4), 278-284.

8. Sturm, R. A., & Duffy, D. L. (2012). Human pigmentation genes under environmental selection. Genome Biology, 13(9), 248.

9. Visser, M., Kayser, M., & Palstra, R. J. (2012). HERC2 rs12913832 modulates human pigmentation by attenuating chromatin-loop formation between a long-range enhancer and the OCA2 promoter. Genome Research, 22(3), 446-455.

10. Sulem, P., Gudbjartsson, D. F., Stacey, S. N., Helgason, A., Rafnar, T., Magnusson, K. P., ... & Stefansson, K. (2008). Two newly identified genetic determinants of pigmentation in Europeans. Nature Genetics, 40(7), 835-837.

11. Lamason, R. L., Mohideen, M. A. P. K., Mest, J. R., Wong, A. C., Norton, H. L., Aros, M. C., ... & Cheng, K. C. (2005). SLC24A5, a putative cation exchanger, affects pigmentation in zebrafish and humans. Science, 310(5755), 1782-1786.

12. Praetorius, C., Grill, C., Stacey, S. N., Metcalf, A. M., Gorkin, D. U., Robinson, K. C., … & Stefansson, K. (2013). A polymorphism in IRF4 affects human pigmentation through a tyrosinase-dependent MITF/TFAP2A pathway. Cell, 155(5), 1022-1033.