DNA PHENOTYPING: PREDICTING ANCESTRY AND PHYSICAL APPEARANCE FROM FORENSIC DNA

Ellen McRae Greytak, PhD* and Steven Armentrout, PhD
Parabon NanoLabs, Inc., 11260 Roger Bacon Dr., Suite 406, Reston, VA 20190
*Corresponding author: ellen@parabon.com, 703-689-9689

Introduction

In forensic investigations, there is a great need for techniques to derive information about an unidentified person directly from biological material. Traditional DNA forensics treats DNA like a fingerprint, using short tandem repeats (STRs) to match a DNA sample to a suspect or a database. However, when no suspect has been identified and there are no database hits, these markers cannot tell investigators anything new about the person who left a particular DNA sample. Other markers in the genome, known as single nucleotide polymorphisms (SNPs), are actual changes in the DNA sequence at a particular site. These types of sequence differences between individuals can affect the functioning or expression of proteins, thus forming the blueprint for changes in a person's traits, including physical appearance. Millions of SNPs can be genotyped in a single assay using genome-wide microarray genotyping. *DNA phenotyping* refers to the prediction of an unknown person's biogeographic ancestry and/or physical traits from SNP data. Such predicted phenotypic information can be used to generate investigative leads, narrow suspect lists, and aid in the identification of human remains. This paper discusses the development of the Parabon® Snapshot™ DNA Phenotyping System ("Snapshot"), which was built over the past four years with funding from the US Department of Defense and is now in active use by law enforcement.

Biogeographic Ancestry Inference

While human genetic variation is continuous across the world, there are genuine genetic differences between populations that can be detected using high-dimensional SNP data. Snapshot uses two distinct approaches for ancestry inference, principal component analysis and statistical clustering, both of which are performed at global and regional scales. Both require a database of reference DNA samples with well-defined ancestry, and thousands of subjects have been collected from populations around the world for this purpose.

Principal component analysis (PCA) combines correlated variables into a smaller set of uncorrelated variables that explain much of the variance present in the original data. Figure 1A shows the first two principal components (PCs) of global human genetic variation. Each point represents a single individual in the reference database, with location on the plot determined solely by their DNA, after which the points were colored according to the subject's known ancestry. Individuals with admixed backgrounds (African-American and Latino in this plot) show ancestry intermediate between the parent groups (African/European and Native American/European, respectively). In this way, it is possible to localize an unknown person to a broad population group by projecting his or her genotypes onto these PCs.

Principal component analysis can also be performed at a regional scale, as long as the populations are genetically distinct. Figure 1B shows a PC plot for only European individuals. While individuals from the various regions grade into one another, different geographic regions cluster in different parts of the plot. For Snapshot, regional PC analysis is performed within each continental group (or within several, if the subject is highly admixed), depending on the global ancestry determined in the first round of testing, further localizing an unknown subject within a region.

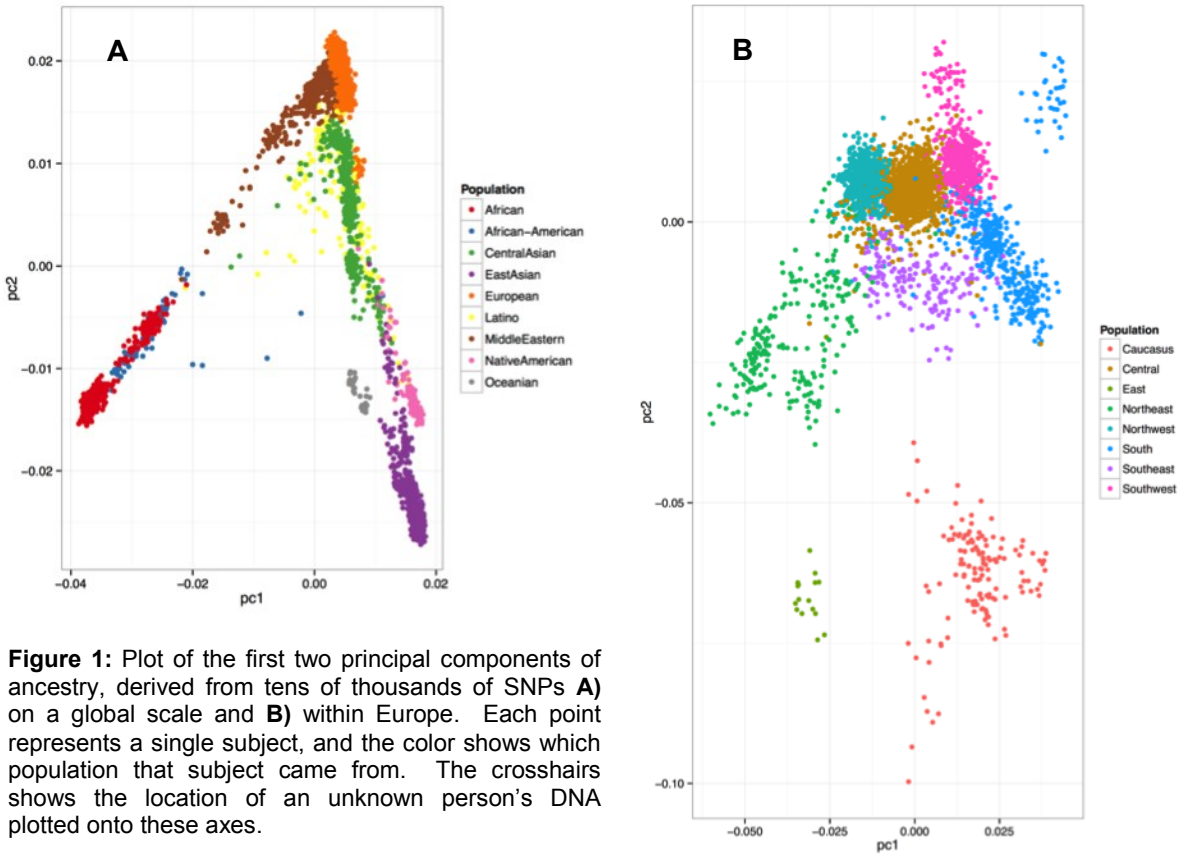

Figure 1: Plot of the first two principal components of ancestry, derived from tens of thousands of SNPs **A)** on a global scale and **B)** within Europe. Each point represents a single subject, and the color shows which population that subject came from. The crosshairs shows the location of an unknown person's DNA plotted onto these axes.

In addition to PCA, Snapshot also uses statistical clustering to assess ancestry. In this approach, a reference database of subjects is used to define a set of populations, against which an unknown subject's DNA is compared to determine its proportional membership in each. This explicitly allows for admixture, even between populations that have not previously been observed. In the first round of analysis, subjects from around the world are included (Figure 2A). In the second round, only subjects from the inferred continent(s) are included (Figure 2B). Significantly, this analysis can be performed even if the subject is admixed. For example, given a child of one East Asian parent and one European parent, this analysis can still determine which region of East Asia that ancestry comes from.

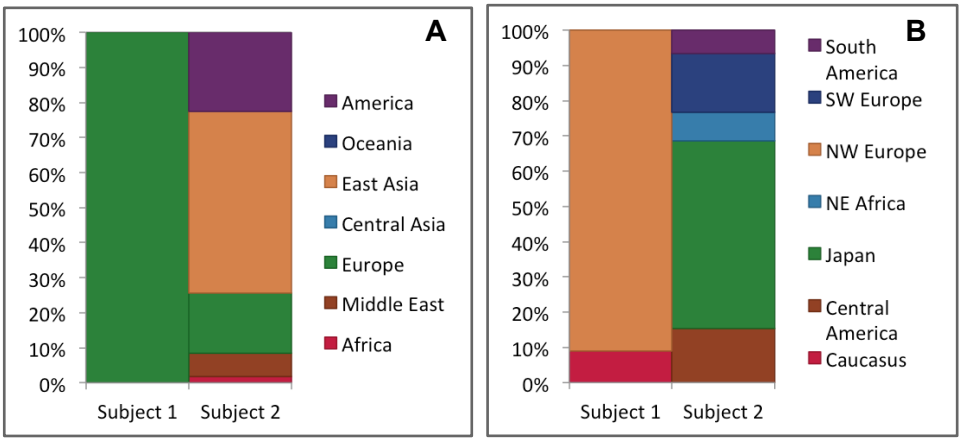

Figure 2: Snapshot ancestry predictions for two subjects at the **A)** global and **B)** regional scales. Only regional populations with >5% ancestry are included. Subject 1 is Northwest European, and Subject 2 is admixed, with one Japanese parent and one Latino parent.

Data Mining and Predictive Modeling

For each phenotype, a database was assembled that included genotypes from hundreds of thousands of SNPs and phenotype on each subject. Pigmentation phenotypes were scored from lightest to darkest according to the subject's self-described coloration (Table 1).

Table 1: Trait values for each pigmentation phenotype.

	1	2	3	4	5
Skin Color	Very Fair	Fair	Light Olive	Dark Olive	Dark
Eye Color	Blue	Green	Hazel	Brown	Black
Hair Color	Red	Blond	Brown	Black	
Freckling	None	Few	Some	Many	

Three-dimensional face morphology data was collected and converted to (x,y,z) coordinates for 7,150 quasi-landmarks on each face, for a total of 21,450 variables. PCA was performed on this data to construct a lower-dimensional "face space" that described the majority of variation among faces (Figure 3). Position along each PC was then used as a series of phenotypes. Subjects from all ethnic backgrounds were included, and phenotypes were corrected for sex and the principal components of ancestry.

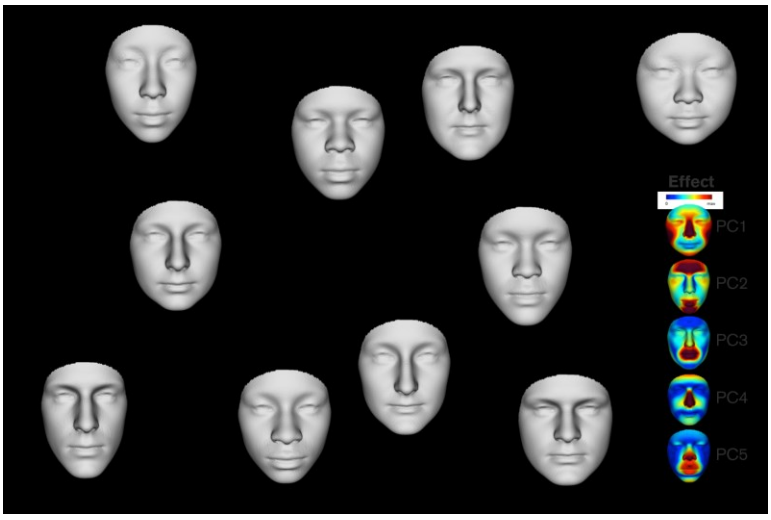

Figure 3: Schematic of the first 5 principal components of facial variation in subjects of all sexes and ancestries. Along each dimension, the two endpoints are shown. At right are heat maps showing the relative magnitude of the effect of each PC on each part of the face, where red is a large effect and blue is no effect.

To build predictive models for these forensic phenotypes, an enhanced genome-wide association study (GWAS) approach was used. Each SNP was assessed for its association with phenotype using linear regression and assigned a p-value based on the strength of association. In addition to this single-SNP association testing, each phenotype was also evaluated for non-additive interactions among SNPs, known as epistasis. Searches for high-order epistasis at a genome-wide scale are extraordinarily computationally intensive, involving a search space that simply cannot be exhausted. Parabon has developed Crush, a software application that uses a distributed evolutionary search algorithm [1] to efficiently search for epistasis. Each set of SNPs examined is evaluated for joint association with phenotype using multifactor dimensionality reduction (MDR) [2]. MDR is as an alternative to regression, which has low statistical power when some genotype combinations are not seen in the data, as is often the case in high-dimensional interactions.

The top SNPs from this data mining were carried forward to predictive modeling. Sex, ancestry PCs, and SNP genotypes were combined in a machine learning model for each phenotype. Model parameters were optimized across cross-validation (CV) folds using a custom evolutionary search algorithm. The best parameters were then used to build the final model. Within each CV fold, the machine learning

model was used to make predictions on the 10% test set that had been held out. These out-of-sample predictions were used to assess the accuracy of the model.

This entire data mining and predictive modeling process takes place within a ten-fold cross-validation (CV) framework. Thus, each of these steps was performed eleven times for each phenotype: once in each of the ten CV folds, using only the 90% training set in that fold, and once at the end, using the entire dataset. The model produced in each CV fold was used to make predictions on the remaining 10% of the data, which were then evaluated for accuracy. At the end of the process, a final model was built using the entire dataset. This model's prediction accuracy on new samples can be expected to be approximately the same as the accuracy across the ten CV folds [3].

Prediction on New Samples

Machine learning produces predictions that are a single value, which must then be interpreted statistically to convert them to phenotype predictions. The cross-validation results provide a framework for doing so and then placing confidence statements on each prediction. Snapshot compares a new, unknown person's prediction to those made on subjects with known phenotype. The prediction value is evaluated for its consistency with each possible phenotype category for that trait (e.g., blue, green, hazel, brown, and black for eye color) based on the distribution of prediction values observed for that category during CV. These consistency values can then be converted to confidence statements, and categories with <5% consistency can be excluded with very high confidence (Figure 4).

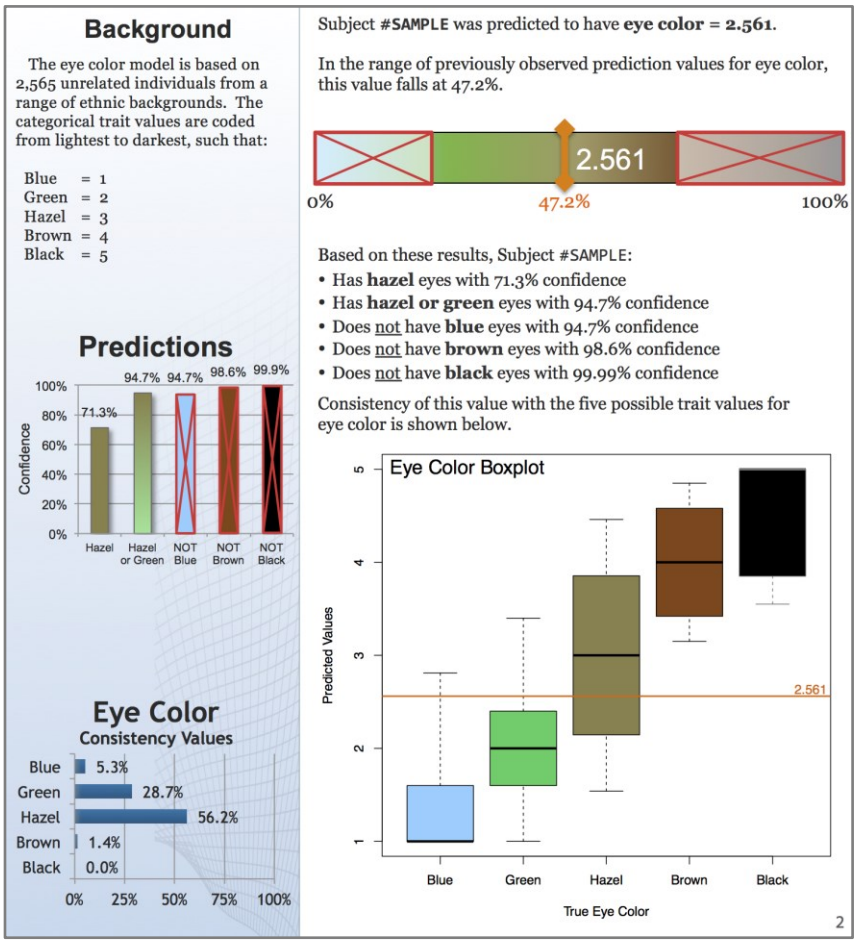

Figure 4: Example of a Snapshot eye color phenotype prediction, along with consistency values for each possible trait value (blue, green, hazel, brown, and black) and confidence statements for each prediction.

Age and BMI are not currently available from DNA, so predictions must be made in the absence of this information. Therefore, face morphology predictions are made at a standard age of 25 and a body mass index (BMI) of 22, which is the middle of the "normal" range. These predictions are solely of underlying facial structure, not of the texture of the skin or the appearance of facial features such as eyebrows or hairline. Thus, to emphasize the parts of the face that are distinctive, the prediction is compared to CV predictions made on subjects with the same sex and major continental ancestry. These comparisons are visualized as heat maps, which show the parts of the face that are changing in each spatial dimension (Figure 5).

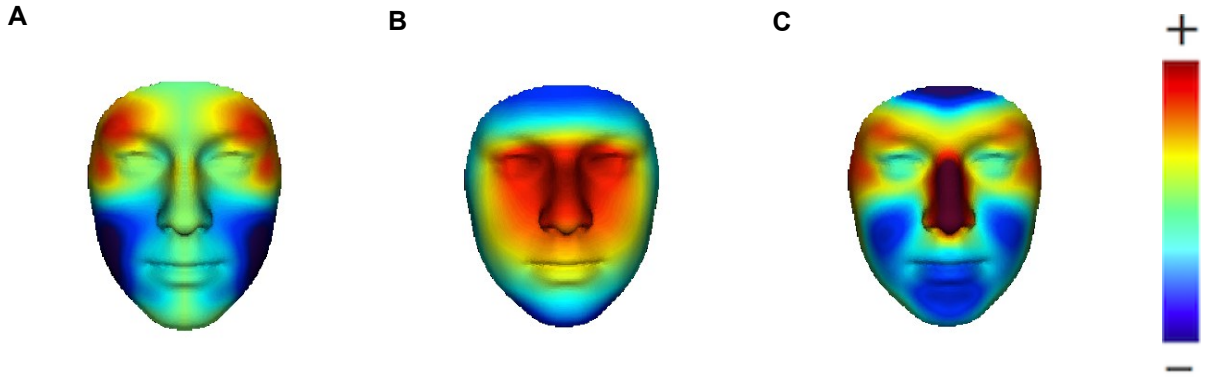

Figure 5: Heat maps comparing the face morphology prediction for an unknown individual to the average prediction for subjects with the same sex and ancestry in the **A)** X, **B)** Y, and **C)** Z dimensions. Intensities are relative within each heat map; red is an increase, and blue is a decrease.

Forensic Casework

Snapshot DNA phenotyping is now being actively used in forensic casework. Numerous additional challenges are encountered when analyzing forensic samples. Genotype data for Snapshot is generated on an Illumina® microarray scanner, which was designed for clinical use, and thus suggests input of at least 200 ng of high-quality DNA to ensure 100% call rate. Parabon and others have performed testing of this SNP technology using forensic quantities of DNA and have demonstrated that very high call rates (~98%) can be obtained from 2.5 ng of DNA, and even 1 ng of DNA can generate call rates near 95%. These results deliver sufficient SNPs for Snapshot to make predictions. However, the missing SNPs must be accounted for. Snapshot uses a machine learning method that allows for missing data, which many do not, so having no-calls at some SNPs still allows predictions to be made. For each case, the CV predictions are recalculated assuming that only that particular set of SNPs was available in the test sets. This ensures that phenotype predictions and confidence statements are based on relevant comparisons. In addition to issues with DNA quantity, forensic cases also often have low DNA quality due to degradation and/or mixing with another individual. These can both lead to decreased call rate and an inability to make accurate predictions. To accommodate such samples, Parabon is actively researching laboratory techniques to repair degraded DNA samples and is developing computational approaches for deconvoluting mixtures using microarray genotype data.

Distant Kinship

A novel kinship capability was developed that utilizes the massive amounts of data generated by SNP arrays. This method was built using a database of ~1,400 subjects from a range of populations with known relationships, out to 7th-degree relatives. New measures of pairwise similarity between a pair of genomes were calculated, and these were used as input variables to a machine learning model that predicts degree of relatedness. As with other Snapshot models, this model was built using cross-validation, and the results are shown in Figure 6. Critically, unrelated pairs could be distinguished from even distantly-related pairs with extremely high accuracy. This capability can be applied to remains

identification when immediate relatives are not available or to ascertain any relatedness between DNA samples recovered at a crime scene.

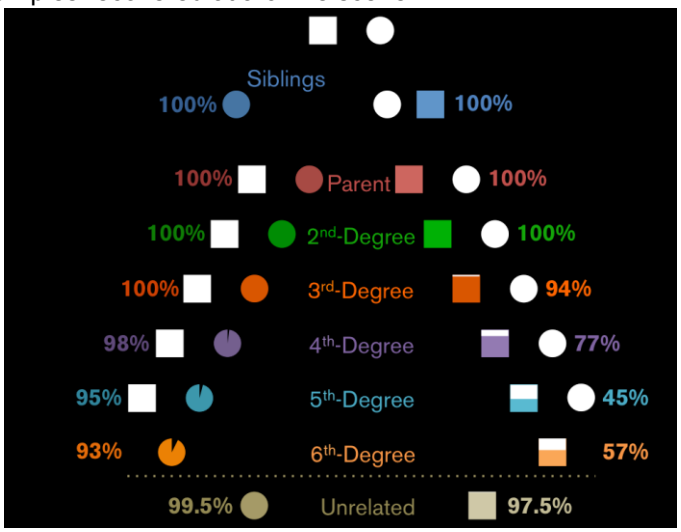

Figure 6: Cross-validation results for Snapshot's distant kinship prediction model. Absolute accuracy refers to correct prediction of the exact degree of relatedness, as opposed to accuracy within one degree.

Conclusion

DNA phenotyping represents a new way for investigators to use forensic DNA to generate investigative leads or learn additional information about unidentified remains. Parabon has developed the Snapshot DNA Phenotyping System, which makes highly accurate predictions about biogeographic ancestry, eye color, hair color, skin color, freckling, and face morphology. Snapshot is offered as a service to law enforcement and is actively being used in casework.

References

1. Sullivan, K., Luke, S., Larock, C., Cier, S., & Armentrout, S. (2008). Opportunistic evolution: efficient evolutionary computation on large-scale computational grids. *Proceedings of the 2008 Conference on Genetic and Evolutionary Computation*, pp. 2227-2232.
2. Ritchie, M. D., Hahn, L. W., Roodi, N., Bailey, L. R., Dupont, W. D., Parl, F. F., & Moore, J. H. (2001). Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *American Journal of Human Genetics*, *69*(1), 138–47.
3. Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *International Joint Conference on Artificial Intelligence*, *5*, 1137–1143.