B186 Development of a Novel DNA Phenotyping System Using Genome-Wide SNP Data

Ellen McRae Greytak, PhD, Parabon NanoLabs, Inc, 11260 Roger Bacon Drive, Ste 406, Reston, VA 20190*

After attending this presentation, attendees will: (1) learn the differences between traditional DNA analysis using Short Tandem Repeats (STRs) and modern DNA genotyping using Single Nucleotide Polymorphisms (SNPs); (2) understand how the genetic content of DNA encodes the observed variations in human appearance; (3) gain an appreciation for how these genetic variants can be used to predict traits (phenotypes); and, (4) learn about a new methodology for producing predictive models for forensic traits from genome-wide SNP data.

This presentation will impact the forensic science community by presenting a novel methodology for the discovery of significant SNPs for forensic traits and the development of those SNPs into predictive models that can be used for DNA phenotyping or kinship analysis.

Traditional forensic DNA analysis uses STRs to match an individual to a DNA sample by testing crime scene DNA against known suspects or a DNA database, such as the Combined DNA Index System. When this fails to yield a match, alternate approaches are required to generate leads from a DNA sample. DNA phenotyping refers to the use of SNPs, DNA variants that code for differences between individuals, to predict an individual's appearance based only on his or her DNA. Prior work on DNA phenotyping has been limited to using only a few SNPs which have been identified in the literature as being individually significant for prediction of eye and hair color and ~100 SNPs to calculate large-scale ancestry (i.e., African, European, or East Asian). Presented here are the results of work to build a novel DNA phenotyping system, developed under funding from the United States Department of Defense, that employs thousands of SNPs for the prediction of complex traits, resulting in improved prediction accuracy over a broader phenotypic range. This system can be applied to individuals from any ethnic background, even admixed individuals.

The predictive models were built using genome-wide genotype data from more than 2,500 subjects for eye color and hair color and from more than 500 subjects for skin color and freckling. Each subject's proportional ancestry in seven global populations (Africa, Middle East, Europe, Central Asia, East Asia, Oceania, and America) was calculated by comparing more than 20,000 SNPs, carefully selected from across the genome, against a set of more than 2,200 background subjects from known populations. This approach can detect even low levels of admixture. Within-continent ancestry (e.g., Northeast vs. Northwest Europe) and the principal components of ancestry were also inferred. These values were then used as covariates for the discovery of significant SNPs associated with each forensically relevant pigmentation phenotype. SNP discovery was performed using advanced data mining techniques that search not only for SNPs that individually contribute to phenotype, but also those that interact in a non-additive (epistatic) fashion. This was achieved using a custom distributed implementation of the Multifactor Dimensionality Reduction (MDR) algorithm.

Using the SNPs discovered during data mining as well as ancestry SNPs, predictive models were constructed using advanced machine-learning algorithms. These techniques allow non-linear variable combinations and are not negatively impacted by the inclusion of extraneous variables. The entire mining and modeling process was performed within a ten-fold cross-validation framework to allow all of the data to be used to build a model while still allowing for accuracy evaluation using out-of-sample predictions. A statistical procedure for evaluating confidence has been developed, which calculates a consistency value for each new prediction for each possible category (e.g., red, blond, brown, and black hair color). Each prediction is presented with a measure of confidence as well as a list of trait categories that can be excluded with very high confidence. DNA from an unknown subject can be run through these predictive models to produce a physical profile. Blind validation testing has been performed for ancestry, eye color, hair color, and skin color on 24 subjects (Table 1) from a range of ethnic backgrounds (European, African-American, Central Asian, and Middle Eastern). The final system has been successfully tested using as little as two ng of extracted DNA.

Table 1: Results of blind validation testing. For each trait, the average consistency for the absolute correct category and the frequency at which the absolute correct category was found to have the highest consistency (two very conservative estimates of accuracy) are reported.

Trait	Average Consistency for Absolute Correct Category	Frequency of Highest Consistency for Absolute Correct Category
Eye Color	76.1%	84.8%
Hair Color	67.7%	95.7%
Skin Color	56.0%	82.6%

A method to determine the degree of relatedness between two subjects using genome-wide SNP data was also developed. This model uses hundreds of thousands of SNPs to ascertain the precise level of similarity between two individuals' genomes. This method has >90% accuracy up to third-degree relatives and can distinguish up to sixth-degree relatives from unrelated pairs with >95% accuracy. Validation results for both the DNA phenotype and kinship inference systems will be presented, along with plans for future enhancement of the underlying processes and software.

SNP, DNA Phenotyping, Kinship Inference