B140 Increased Ancestry Prediction of the United States Population Using Short Tandem Repeat (STR) Data in Addition to 32 Ancestry-Informative SNPs

Valerie Clermont Beaudoin, BS, 1930 37th Street, NW, Washington, DC 20007; Katherine B. Gettings, PhD, NIST, 100 Bureau Drive, Mail Stop 8314, Gaithersburg, MD 20899; Moses S. Schanfield, PhD, Dept of Forensic Sciences-GWU, 2100 Foxhall Road, Washington, DC 20007; and Daniele S. Podini, PhD, Department of Forensic Science, 2100 Foxhall Road, NW, Washington, DC 20007*

After attending this presentation, attendees will learn that the STRs used for Human Identification (HID) can support Single Nucleotide Polymorphisms (SNPs) -based ancestry prediction on United States-population individuals.

This presentation will impact the forensic science community by demonstrating that STRs can aid in increasing the accuracy of ancestry inference.

Often a very limited amount of information is available on the perpetrator of a crime, especially when there are no eyewitnesses, when the DNA left behind does not match any suspect, or the DNA isn't present in the databases; however, it is still possible to garner investigational information by analyzing Ancestry-Informative SNPs (AISNPs). An SNP assay was developed at George Washington University that predicts the most likely ancestry of an individual between the primary United States populations (grouped as African American, East Asian, European American, and Hispanic/Native American). The prediction is made by first determining the probability of the SNP profile within each population (i.e., the Random Match Probability (RMP)). The RMP is the probability of randomly finding an unknown individual with that specific profile in the given population. Then a Likelihood Ratio (LR) is calculated for each of the four populations. In this LR, the numerator is the RMP of the SNP profile from the population in which it is highest and the denominator is the RMP of the profile in the specified population. The LR expresses how much more likely it is to observe the profile if it originated from the population in the numerator versus if it originated from the population in the denominator. An empirical threshold of 1,000 was chosen, above which it is considered significant for a sample to be classified as belonging to one of the four populations. A sample with an LR lower than 1,000 was classified as inconclusive between the two populations with the highest RMPs, inferring that the sample most likely belongs to one or both of those populations. The samples are divided into four categories based on the results obtained: correct, incorrect, inconclusive correct, and inconclusive incorrect. Correct corresponds to the situation where the sample is classified as belonging to a single population, which is the same one as reported by the donor. Incorrect corresponds to the situation where the sample is classified as belonging to the wrong population. Inconclusive correct corresponds to the situation where the sample is classified as belonging to two populations and the donor reported one or both of these populations. Inconclusive incorrect corresponds to the situation where the sample is classified as belonging to two populations but these populations were not reported by the donor.

Thirty-two AISNPs were used to predict the ancestry of 134 samples (with self-reported ancestry information) using the method described above. Of these samples, 72% were correctly classified as belonging to one population. One sample was classified as inconclusive incorrect. The rest of the samples (27%) were classified as inconclusive correct. No samples were classified as incorrect.

To assess the impact of including HID STR allele frequencies into the ancestry prediction, the 134 samples were also genotyped with AmpflSTR® Identifiler® Plus. The random match probability was calculated with the STR data obtained and factored into the SNP RMPs. The LRs were then recalculated and the accuracy of the prediction was reevaluated. With the inclusion of the STR data in the analysis, the accuracy of the prediction increased to 80% correctly classified samples. This 8% increase is statistically significant at a 95% confidence (Z-score=-1.7252154; p=0.0422). This indicates that although STRs were not selected to provide ancestry information, they can improve the prediction of an SNP-based assay. Thus, available STR data should be included in ancestry prediction.

SNPs, Ancestry Prediction, STRs