B63 A Multivariate Statistical Approach for Enhancing Ancestry Prediction From Microhaplotype Data

Chiara Della Rocca, Rome 00185, ITALY; Fabio Oldoni, PhD, Arcadia University, Glenside, PA 19038; Eugenio Alladio, PhD, Department of Chemistry, University of Torino, Torino, Piedmont 10125, ITALY; Andrea Berti, PhD, Carabinieri R.I.S. - Scientific Investigation Dept, Rome 00191, ITALY; Filippo Barni, PhD, Carabinieri R.I.S. - Scientific Investigation Dept, Rome 00191, ITALY; Fulvio Cruciani, PhD, Sapienza University, Rome 00185, ITALY; Daniele S. Podini, PhD, Department of Forensic Science, Washington, DC 20007*

Learning Overview: After attending this presentation, attendees will be able to appreciate the advantages of using Multivariate Data Analysis (MDA) techniques on Microhaplotype (MH) data.

Impact on the Forensic Science Community: This presentation will impact the forensic science community by providing an alternative statistical approach to infer Biogeographical Ancestry (BGA) at higher resolution compared to conventional Principal Component Analysis (PCA) methods that are commonly used in forensic routine analysis.

MHs are newly developed multi-allelic markers of at least two Single Nucleotide Polymorphisms (SNPs) within < 300bp. Due to the presence of small amplicons and a low recombination rate, absence of stutter and preferential amplification, they are promising candidates for human identification, mixture deconvolution, and BGA prediction. The global interest in BGA inference aims to maximize the amount of forensically relevant information retrievable from DNA evidence. Current protocols for ethnic origin estimation are generally based on PCA and Bayesian approaches. Early studies demonstrated the capability of multivariate statistical tools to predict the ancestry affiliation from autosomal Short Tandem Repeat (STR) data. This study aimed to extend this alternative and dynamic statistical approach to the analysis of MH data.

Multivariate techniques such as Partial Least Squares Discriminant Analysis (PLSDA) and Support Vector Machines (SVM) were used for the analysis of MH profiles from an initial set of 347 individuals from four United States population groups (88 African American [AA], 114 European-American [EA], 102 Southwest Hispanic [His], and 43 East-Asian American [EAA]). The DNA samples were typed using a bioassay of 74 MHs on the Ion S5™ sequencing platform. Overall both PLSDA and SVM approaches significantly improved ancestry inference by enhancing the separation of the four population clusters. Furthermore, PLSDA results were used to assess the accuracy of ethnic affiliation for four test individuals, one per each reference population group, to test their ancestry affiliation in terms of Likelihood Ratio (LR). The resulting LR values provide an indication of how much more likely it is to observe the MH profile of interest if it originated from the test population at the numerator than if it originated from the other three populations at the denominator. LR values for each of the four test individuals confirmed the correct ancestry affiliation and, in particular, $LR = 10^{99}$, $LR = 10^{62}$, $LR = 10^{4}$, and $LR = 10^{1}$ were obtained for AA, EAA, EA, and His, respectively. Overall, high LR values were observed for well-genetically defined populations while lower values from admixed ones, as expected.

The multivariate statistical approach will be shortly implemented in an open-source and user-friendly R Shiny app to assist law enforcement agencies and forensic DNA analysts with the interpretation of BGA data.

Microhaplotypes, Biogeographical Ancestry, Multivariate Data Analyses

*Presenting Author