SELECTION OF A SUPPLEMENTARY ANCESTRY-INFORMATIVE MARKER (AIM) PANEL OF INDELs FOR DISTINGUISHING SOUTHWEST HISPANICS AND SOUTHWEST AISIANS

Bing Song[1], Lindsey M. Tompson[2], Xiangpei Zeng[1], Sarah Sturm[1], Kelly Sage[2], Frank R.Wendt[1], Jonathan King[1], Ranajit Chakraborty[1], Bruce Budowle[1,3] , Bobby LaRue[1,4]
[1]Institute of Applied Genetics, Department of Molecular and Medical Genetics, University of North Texas Health Science Center
[2]Sorenson Forensics
[3]Center of Excellence in Genomic Medicine Research (CEGMR), King Abdulaziz University
[4]Department of Forensic Science, Sam Houston State University

Compared to commonly used STRs, small insertion/deletion polymorphisms (INDELs) are better suited for analysis of degraded biological samples in which DNA can be fragmented < 180-200bps. In situations where no suspect has been identified, an ancestry-informative INDEL panel could provide investigative value. In a previous study, a 59 AIM-INDELs panel was designed using the data from the 1000 Genomes data to distinguish three major populations, i.e., Africans, East Asians and Caucasians. However, that panel failed to separate the southwestern Hispanics (SWH) and southwest Asians (SWA).

In this study, a supplementary panel of INDELs was identified to resolve SWH and SWA. From the 1000 Genomes Project Website, the genome data from more than 2000 individuals were downloaded. The populations selected included: CLM (Columbian in Medellin, Columbia), MXL (Mexican Ancestry in Los Angeles, California) and PEL (Peruvian in Lima, Peru) as representative SWH populations; and BEB (Bengali in Bangladesh), GIH (Gujarati Indian in Houston, TX), ITU (Indian Telugu in the UK), STU (Sri Lankan Tamil in the UK) as representative SWA populations.607 individuals are selected after training. Then using VCF tools and an Excel workbook, INDELs were selected with an -F_{ST} value greater than 0.35, a length of 3-6 base pairs, and high allele frequency divergence. After being tested for departures from Hardy-Weinberg Equilibrium (HWE) expectations and for linkage disequilibrium (LD) using the software, Genetic Data Analysis (GDA), a 19 marker panel was selected. Population structure was investigated by these markers using the program STRUCTURE v2.3.4 and principle component analysis was performed using R package “pca3d”. A set of primers was designed for future development of a PCR-based panel.