A NEW PHYLOGENETIC APPROACH FOR DE NOVO DISCOVERY OF PUTATIVE MITOCHONDRIAL INSERTIONS INTO THE NUCLEAR GENOME (pNumts)

Utpal Smart[1], Bruce Budowle[1,2] , Angie Ambers [3], Rodrigo Soares Moura-Netoc [4], Rosane Silvad[5] and August E. Woerner[1,2]
[1]Center for Human Identification, University of North Texas Health Science Center
[2]Department of Microbiology, Immunology, and Genetics, University of North Texas Health Science Center
[3]Forensic Science Department,The Henry C. Lee College of Criminal Justice and Forensic Sciences, University of New Haven
[4]Laboratório de Biologia Molecular Forense, Instituto de Biologia, Universidade Federal do Rio de Janeiro
[5]Instituto de Biofisica Carlos Chagas Filho, Universidade Federal do Rio, de Janeiro, Rio de Janeiro

In forensic genetics, construing signal from noise usually involves estimating an arbitrary analytical threshold (AT) below which all sequence data are ignored. We present a novel method that leverages phylogenetic analyses, to deconstruct noise stemming from nuclear insertions of mitochondrial DNA (Numts) in massively parallel sequence read data. Our bioinformatic method is capable of discovering putative Numts (pNumts) in absence of a reference genome, using diagnostic statistics extracted from haplotype networks. We tested the new method on a whole mitochondrial genome dataset (n=41 individuals from an admixed population sample from Rio de Janeiro) and in the process identified 451 pNumts. Evaluation of these pNumts haplotypes against existing Numt databases showed 147 exact matches to previously discovered Numts, while 122 haplotypes differed only by a single base pair. None of the pNumt haplotypes matched exclusively to the mitochondrial genome. In general, these sequences were considerably more divergent from the mitochondrial genome than from those of the Numt database, indicating that these pNumts are probably variants that as yet remain uncatalogued. In contrast to previous approaches, our method is able to detect both polymorphic and fixed Numt sequences. The results also indicate an enrichment for pNumts in the D-Loop and associated Promoters, an area of the human mitogenome that carries markers of forensic importance. Our novel approach has the potential to be expanded to other scenarios which might require construing signal from noise, including the deconvolution of mixtures, and thus significantly refines how ATs are designated.