IDENTIFYING STRS IN NEXT GENERATION SEQUENCING DATA

H.-H. Tam[1], S. Faith[2], D. Bornman[3], S. Nelson[2], P. Yan[1], B. Young[2], R. Bundschuh[4,5]
[1]The Ohio State University Medical Center
[2]Battelle Memorial Institute, Divisions of Applied Biology
[3]Battell Memorial Institute, Statistics and Information Analysis
[4]The Ohio State University Medical Center and Department of Physics
[5]The Ohio State University Department of Physics and Biochemistry

The production cost and analysis time for next generation sequencing (NGS) data has dramatically declined over the last few years and will soon make routine sequencing of individuals practical in medical as well as in forensic settings. In forensic settings it would be desirable to be able to link the results of NGS to traditional identification methods such as the CODIS panel of short tandem repeats (STRs). However, the standard approach to analyzing NGS data, namely mapping or aligning the reads to a reference genome, fails due to the variation in repeat length between the subject or analysis sample and standard haploid human reference sequences (e.g., hg19) that present only one STR allele for each locus . Thus, the common approach of analyzing NGS data using reference alignment is not a feasible method to type STR profiles from human samples. To circumvent this problem and make NGS data accessible to the forensics community, we have developed a Hidden Markov Model based computational approach that can identify raw NGS data reads that completely cover a given STR without the requirement of reference alignment. We have applied the algorithm to NGS data from an individual that has also been subjected to PowerPlex® 16 analysis and show that our computational approach reveals the defined allele calls for a majority of the STRs in the CODIS panel, as well as the pentameric STRs found in the PowerPlex 16® panel. We also discovered that the two most critical factors for applying this strategy were sufficient read length, encompassing the nucleotides of repeat region and the unique flanking sequences up- and down-stream of an STR locus, and coverage, the frequency of observing any given locus within the dataset. This method describes a possible path forward for the forensics community for examining human STRs with the vast array of NGS technologies presenting to the future market.