COMPUTATIONAL ALGORITHMS FOR DEVELOPMENT OF IDENTITY-LINKED SNP ISLANDS FOR ANALYSIS BY MASSIVELY PARALLEL SEQUENCING

M. Heath Farris[1], Andrew Scott[1], Ashley Williams[1], Marta Bartlett[1], Caroline Gary[1], Patricia Coleman[2], and David Masters[3]

[1] Homeland Security Systems Engineering & Development Institute (HSSEDI), The MITRE Corporation, McLean, Virginia
[2]Department of Homeland Security, Customs and Border Protection, Washington, D.C.
[3]Department of Homeland Security, Science and Technology Directorate, Washington, D.C.

Markers of identity located within the human genome have been shown to have utility in the differentiation of DNA from individual contributors. With the advancements of massively parallel DNA sequencing technologies and the development of human single nucleotide polymorphism (SNP) databases, the ability to identify islands within the human genome with identity-linked information allows for the design of suites of identity-linked target regions, amenable to sequencing in a multiplexed and massively parallel manner. Using SNP data from the *1000 Genomes Project*, regions of the human genome, containing identity-linked SNPs that are amenable to targeted resequencing on the Illumina DNA sequencing platform, were identified. Computational algorithm filters were used to exclude target regions that did not conform to restrictions in length and sequence variation and to further exclude target regions with non-optimal conservation within defined flanking region sequence lengths. These conserved flanking regions were used to design primers sets for amplification of the target regions. Algorithms were designed to find conserved regions within the flanking region sequence lengths using the *National Center for Biotechnology Information* (NCBI) *GenBank* database. SNP target regions and primer sites, identified in this manner, were amplified from contributor genomic DNA samples using the polymerase chain reaction (PCR). Amplicons were sequenced in a massively parallel manner using the Illumina MiSeq platform, and the resulting sequences were analyzed for SNP variations. Over 150 putative identity-linked SNPs were targeted in the genome regions that were amenable to PCR and targeted sequence analysis. DNA samples of 25 individuals were uniquely identified using the suite of identity-linked SNPs. The results of this study indicate that the custom computation algorithms allow the tunable identification of identity-linked target regions for use in uniquely identifying individuals using massive parallel DNA sequencing technologies.