B109 FONTANA: A FOrensic NexT-Generation ANAlysis Pipeline for High-Throughput Microhaplotype Data Analysis

Keylie M. Gibson, BS, The George Washington University, Ashburn, VA 20147; Fabio Oldoni, PhD, The George Washington University, Washington, DC 20007; Rebecca M. Hart, The George Washington University, Washington, DC 20007; Daniele S. Podini, PhD, Department of Forensic Science, Washington, DC 20007; Keith A. Crandall, PhD, The George Washington University, Washington, DC 20052*

Learning Overview: After attending this presentation, attendees will become familiar with a new computational software tool specifically designed for their community—FONTANA: the FOrensic NexT-generation ANALysis pipeline. Attendees will discover the strengths of this new approach to analyze next-generation sequencing data for microhaplotype discovery and analysis and experience applications of this new approach to improve human identification efforts.

Impact on the Forensic Science Community: This presentation will impact the forensic science community by introducing the community to a new computational tool for analyzing next-generation sequencing data focused on microhaplotypes. Additionally, this presentation will help bridge the gap between forensic science and computational biology.

Microhaplotypes (MHs), typed on Next-Generation Sequencing (NGS) platforms, can enhance mixture deconvolutions and provide increased discrimination power and ancestry predictions. MHs are loci of two or more single nucleotide polymorphisms (SNPs) within a short distance from each other (<300 nucleotides i.e. 'micro') with three or more allelic combinations ('haplotypes'). Unlike STRs, MHs have a low mutation rate and show no PCR stutter, and using NGS, phasing information between the SNPs can be attained. The amount of data generated by a single NGS run can be upwards of a million times greater (4 billion base pairs to 4 trillion base pairs or 1GB to 1TB of data) compared to about 4 million bases pairs (or 1MB of data) generated by a single Sanger sequencing run. Working with data on this scale requires new computational tools. A tool has been developed to assist the forensic science community with MH analyses. Introducing FONTANA: FOrensic NexT-generation ANALysis pipeline. FONTANA is a working pipeline, platform agnostic, created to analyze microhaplotypes from forensic samples. FONTANA currently consists of quality control, alignment to the human genome reference, variant calling, haplotype calling, and report generation. Quality control is executed by Flexbar, where low quality reads, low quality nucleotides, and adapter sequence contamination are removed. The sequencing reads are then aligned to the human genome reference, and ready for the next step: variant calling, completed with FreeBayes. Haplotypes, characterized by phasing the SNPs together, are called with the FreeBayes program for each sample from the previous step where all variants (at known and unknown SNP location) were identified. This entire pipeline is executed in the Snakemake workflow, which is a tool to create reproducible and scalable data analyses–ideal for a forensic application. FONTANA's environment has been configured in Bioconda. Bioconda is a package manager for bioinformatic software, and therefore, FONTANA can be accessible to a variety of operating systems (Linux and Mac OSX). An initial version of FONTANA was applied to three populations: Mexican Pima, European American, and Southwest Hispanic, each with 50 individuals, with ten microhaplotypes. The most unique alleles found at a single MH was 14, with a range between two to 14 alleles at a MH (median = 5 alleles). One of the microhaplotypes showed that admixed populations (Southwest Hispanic and European American) contain a greater diversity of alleles present than in a homogeneous population (Mexican Pima). Moreover, utilizing MHs facilitates biogeographic ancestry prediction in a sample. Over 70 MHs will be added to the pipeline; additionally, the goal of FONTANA is to be designed as a plastic pipeline, so that as new and more informative MHs are discovered, they can easily be added to FONTANA. FONTANA can be found at https://github.com/kmgibson/FONTANA, a GitHub page developed for tracking the progress and usage and providing guidelines for execution of the software program. Future directions include adding downstream applications for ancestry prediction, mixture deconvolution, and probabilistic genotyping. The development of this computational analysis tool and identifying additional SNPs within each MH strengthens the foundation of MHs for use in criminal casework and will help combat the problem of mixture deconvolution.

Microhaplotypes, Computational Software, Next Generation Sequencing