B214 The PROVEDIt Initiative: The Development and Release of a Collection of Computational Tools and a Large-Scale Empirical Data Set for Forensic Research and Validation

Lauren Elizabeth Alfonse, MS, Boston University School of Medicine, Biomedical Forensic Sciences, 72 E Concord Street, Rm L805, Boston, MA 02118; Amanda D. Garrett, MS, Boston University School of Medicine, 72 E Concord Street, Rm R806, Boston, MA 02118; Harish Swaminathan, PhD, Boston University School of Medicine, 72 E Concord Street, Boston, MA 02118; Kelsey C. Peters, BS, Boston University Biomedical Forensic Sciences, 72 E Concord Street, Rm R806, Boston, MA 02118; Genevieve Wellner, MS, Illumina Madison, 5602 Research Park Boulevard, Ste 200, Madison, WI 53719; Lauren M. Taranow, BA, 7 Price Road, #1, Allston, MA 02134; Jennifer L. Sheehan, BS, Boston University School of Medicine, 72 E Concord Street, Rm R806B, Boston, MA 02118; Sarah E. Norsworthy, BA, RTI International, Center for Forensic Sciences, 3040 E Cornwallis Road, Bldg 7, Rm 222, Research Triangle Park, NC 27709; Desmond S. Lun, PhD, Rutgers University, Center for Computational & Integrative Biology, Camden, NJ 08102; Ken Duffy, PhD, Hamilton Institute, Maynooth, IRELAND; Muriel Medard, ScD, Massachusetts Institute of Technology, Dept of Elec Engineering & Computer Science, 77 Massachusetts Avenue, Cambridge, MA 02139; Robin W. Cotton, PhD, Boston University School of Medicine, Biomedical Forensic Sciences, 72 E Concord Street, R 806, Boston, MA 02118; and Catherine M. Grgicak, PhD, Boston University, School of Medicine, Biomedical Forensic Sciences, 72 E Concord Street, Rm R806D, Boston, MA 02118*

After attending this presentation, attendees will be aware of the release of a large-scale database of DNA samples containing up to five contributors. Attendees will also have knowledge of five software-based applications that can be utilized during forensic research, validation, or pedagogical pursuits.

This presentation will impact the forensic science community by providing an easily accessible large-scale database of compromised, mixed DNA samples of varying templates. This will allow all members of the forensic science community to test software implementations and various hypotheses from a single, well-characterized data set.

The interpretation of forensic mixtures is difficult and becomes increasingly more difficult as: (1) the number of contributors increases; (2) the number of copies decreases; and; (3) Polymerase Chain Reaction (PCR) becomes less efficient because of PCR inhibitors or DNA damage. Several interpretation tools, analysis techniques, and interpretation standards/recommendations have been developed and released. In the case of software solutions, all of these systems rely upon assumptions and have computational nuances associated with their algorithms; thus, there is considerable interest in comparing their performances.

In an effort to provide support to the community and to foster growth in both forensic research and operations, the Project Research Openness for Validation with Empirical Data Initiative (PROVEDIt) is announced.

PROVEDIt comprises 25,000 .fsa and .hid profiles as well as a suite of analysis, interpretation, and *in silico* software systems/procedures and models developed in a variety of environments. The profiles and tools are available to the community on Boston University's DNA Mixture Website (www.bu.edu/dnamixtures).

The collection of computational systems includes: (1) Computational Evaluation of Evidentiary Signal (CEESIt) — outputs the likelihood ratio, likelihood ratio distribution, and p-value for an unknown; (2) Number

of Contributors (NOCIt) — outputs the a posteriori probability distribution for the number of contributors; (3) Genotype Generator & Evaluation Tool (GGETIt) — a simulator that outputs the minimum number of contributors based on allele counts; (4) Simulating Evidentiary Eletropherograms (SEEIt) — a dynamic model that simulates the entire forensic process and produces well-characterized electropherograms for up to six contributors; and, (5) CleanIt — an automated procedure for filtering bleed-through, complex bleed-through, and minus A from an electropherogram.

The collection of .fsa and .hid profiles includes one- to five-person DNA samples, amplified with targets ranging from 1ng to 0.007ng, with varying levels of damage and contributor ratios.

The usefulness of large data sets is demonstrated by plotting a histogram of the empirical signal for each locus and confirming that signal from one copy is detected. These histograms exhibit at least three seemingly distinct peaks. For example, for the D8S1179 locus, the first signal group (median 4 Relative Fluorescence Units (RFU)) consists largely of instrumental noise; the second group (median 24 RFU) is the signal obtained when one copy of DNA is amplified; and the third (median 47 RFU) is the signal obtained when two copies of DNA are amplified. Next, the signal obtained when these same samples were injected for twice as long is presented. The same multi-modal pattern is observed but, with doubled injection time, the first and second signal groups appear at 4-11 RFU and 36-65 RFU, respectively. These data suggest that single-copy DNA signal is regularly detected using modern forensic laboratory implementations.

This project was partially supported by the National Institute of Justice, Office of Justice Programs, United States Department of Justice, and the Department of Defense, Army Research Office, Rapid Innovation Fund. The opinions, findings, and conclusions or recommendations expressed in this publication are those of the author(s) and do not reflect those of the Department of Justice or Department of Defense.

PROVEDIt, DNA Mixtures, Low-Template DNA