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

Lauren E. Alfonse, M.S.[a], Amanda D. Garrett, M.S.[a], Harish Swaminathan Ph.D .[a], Kelsey C. Peters, B.S.[a], Genevieve Wellner, M.S.[a], Xia Yearwood-Garcia, B.S.[a], Lauren M. Taranow, M.S.[a], Jennifer Sheehan, B.S.[a], Sarah E. Norsworthy, M.S.[a], Ullrich Mönich, Ph.D.[f], Desmond S. Lun, Ph.D.[b,c,d,e], Ken R. Duffy, Ph.D.[g], Muriel Médard, Sc.D.[f], Robin W. Cotton, Ph.D.[a], & Catherine M. Grgicak, Ph.D.[a,*]

[a] *Biomedical Forensic Sciences Program, Boston University School of Medicine, United States*
[b] *Center for Computational and Integrative Biology, Rutgers University, United States*
[c] *Department of Computer Science, Rutgers University, Camden, United States*
[d] *Department of Plant Biology and Pathology, Rutgers University, New Brunswick, United States*
[e] *School of Information Technology and Mathematical Sciences, University of South Australia, Australia*
[f] *Research Laboratory of Electronics, Massachusetts Institute of Technology, United States*
[g] *Hamilton Institute, Maynooth University, Ireland*

* Corresponding and presenting author at: Biomedical Forensic Sciences, Boston University School of Medicine, 72 E. Concord Street Rm R806, Boston, MA 02118, USA
Tel.: +1 617 638 1968; fax: +1 617 638 1960.
E-mail address: cgrgicak@bu.edu (C.M. Grgicak).

Background

The amplification of STR fragments followed by capillary electrophoresis separation is the chief technique by which forensic laboratories conduct human identity testing. Since the publication of Mullis et al.'s [1] work describing PCR, there have been innumerable and substantive advances in the field of identity testing such as the introduction of real-time PCR and extremely sensitive PCR and electrophoresis systems into forensic laboratories. These advances regularly allow genotypic information obtained from only a few cells to be detected. The detection, evaluation, and interpretation of signal garnered from the amplification of only a few copies is difficult and becomes progressively more challenging as the sample becomes more complex.

Several interpretation tools [2-8], analysis techniques [9, 10], and interpretation standards/recommendations [11-13] have been developed and released. Some of these are in the form of published research, while others take the form of an open-source procedure, a freeware tool, or as a commercially available product. In the case of software solutions, these systems rely upon distinct assumptions and have computational nuances associated with their algorithms; thus, there is considerable interest in comparing their performance. Despite this interest, producing an empirical data set large enough to efficiently compare, contrast and validate these computational systems is costly, labor-intensive, and requires the amplification of many samples that may not be readily available. In response to these issues, Boston University released a set of 2,990, 1- to 4- person .fsa files on www.bu.edu/dnamixtures which have been available to the community since 2013.

In an effort to provide continued support to the forensic science community and to foster growth in both forensic research and operations, we announce the PROVEDIt initiative: Project Research Openness for Validation with Empirical Data.

PROVEDIt comprises 25,000 .fsa and .hid profiles, available at www.bu.edu/dnamixtures, as well as a suite of analysis, interpretation, and *in silico* software systems/procedures and models, at http://sites.bu.edu/grgicak/software, developed in a variety of software environments by a multi-disciplinary, inter- institutional team.

The collection of computational systems includes:

1. CEESIt[2]: Computational Evaluation of Evidentiary Signal. Provides the likelihood ratio, likelihood ratio distribution and *p*-value for an unknown.
2. NOCIt[3]: Number of Contributors. Provides the *a posteriori* probability distribution for the number of contributors from which the sample arose.
3. GGETIt: Genotype Generator & Evaluation Tool. A simulator that outputs the minimum number of contributors based on allele counts and compares it against the known number of contributors.
4. SEEIt: Simulating Evidentiary Eletropherograms. A dynamic model written in the Stella™ environment that simulates the entire forensic process and produces simulated, well-characterized electropherograms for up to six contributors.
5. CleanIt. An automated procedure for filtering bleed-through, complex bleed-through and minus A from an electropherogram.

Methods

The collection of 25,000 .fsa and .hid profiles was generated over a four year period and includes 1- to 5- person DNA samples, amplified with targets ranging from 1 to 0.0078 ng. In the case of multi-contributor samples, the contributor ratios ranged from equal parts of each contributor to mixtures containing 19 parts of one and 1 part of the other(s). These profiles were generated using a variety of laboratory conditions from samples containing pristine; damaged (i.e., UV-Vis); enzymatically/sonically degraded; and inhibited DNA. Table 1., provides a summary of the laboratory parameters used to generate the samples.

Table 1. A summary of the laboratory parameters used to produce the .fsa and .hid profiles. §QI (Quality Index) is a metric obtained from the Quantifiler® Trio kit, which provides a ratio of the concentration of small and large autosomal target.

Kit (PCR cycle no.)	**3130 (5, 10, 20 s)**	**3500 (5, 15, 25 s)**	**DNA Condition**				**§QI**
			Pristine	**UV**	**Degraded**	**Inhibited**	
IDPlus (28 cycles)	x		x	x	x	x	x
IDPlus (29 cycles)		x	x				
GlobalFiler (29 cycles)		x	x	x	x	x	x
PP16HS (32 cycles)	x		x				

We designed the names to contain as much information about the sample as possible. There are two sets of samples, and the naming conventions between them differ. The first sample set is designated with the project code RD12-0002, while the second contains project code RD14-0003. In general, single source sample names will follow the format below and are best explained through example:

RD14-0003-21d1x-0.5IP-Q0.8_002.20sec or RD12-0002-21d1-0.5IP-002.20sec

RD14-0003 and RD12-0002 are the project numbers, 21 is the sample identifier within that project and d_ is the dilution number which was used by laboratory personnel to distinguish between extracts. **NB:** Each sample is designated by the combination of project number and sample identifier. For example, RD14-0003-21 will have a different known genotype than RD12-0002-21. Within the RD14-0003 project sample names, the 'x' designator indicates DNA condition (see Table 2.), 0.5 represents the template mass in nanograms (typically ranges from 1-0.0078 ng), and IP is the amplification kit type (i.e. IP=Identifiler® Plus, GF=Globalfiler®, PP16=PowerPlex®16 HS). If the extracts were quantified with Quantifiler® Trio we provide the Quality Index in the form of a Q value (Q0.8 in this example). In some instances, the Q designator is followed by "LAND", which stands for "large autosomal not detected." This term is used for samples in which the large autosomal fragment was not detected during qPCR; thus, a numerical Q value was not obtained. The designator 002 is the capillary number (capillaries are numbered 001-004 on the 3130 CE and 01-08 on the 3500 CE), and 20 sec is the injection time (5, 10, and 20 second injection times are utilized for samples run on the 3130 CE, and 5, 15, and 25 second injection times are utilized for samples run on the 3500 CE).

Table 2. A summary demonstrating the type of damage induced for the RD14-0003 sample set, and a corresponding example using single source samples.

Damage Type	Single Source Sample Name	x	Description
DNase I Degradation	RD14-0003-21d2**a**-0.5IP-Q0.8_002.20sec	a to e	Letters a, b, c, d, and e indicate volume of DNase I added. a=not degraded (no enzyme) and e=most degraded (highest volume of enzyme).
Fragmentase® Degradation	RD14-0003-36d1**-15**-0.5IP-Q1.4_003.10sec	-15 to -45	-15 indicates enzyme digestion/incubation time in minutes. 15, 30, and 45 minute digestion times were utilized.
UV Damage	RD14-0003-03d2**U60**-0.5IP-Q8.8_003.10sec	U15 to U105	U60 indicates 60 minutes of UV exposure. Times range from 15-105 minutes.
Sonication	RD14-0003-12d3**S30**-0.0078IP-Q17.1_002.20sec	S2 to S30	S30 indicates DNA was damaged with 30 cycles of sonication. 2, 10, and 30 cycles were utilized.
Humic Acid Inhibition	RD14-0003-49d2**I22**-0.5IP-Q2.4_002.10sec	I15 to I35	I22 indicates volume of 2 mg/mL humic acid (in µL) added to whole blood lysate. Three volumes were utilized (15, 22, and 35 µL).

Mixtures are named similarly, with the addition of a mixture ratio. For example:

RD14-0003-31_32-1;2-M1x-0.062IP-Q14.4_003.10sec

This is a 2-person mixture of contributors RD14-0003-31 and RD14-0003-32. The ratio follows the sample names (i.e., 1;2); thus, person RD14-0003-31 is contributing one part to the mixture, and person RD14-0003-32 is contributing two parts. M1 is the mixture dilution number used by

laboratory personnel to distinguish between extracts. The rest of the sample name follows the convention outlined above for single-source samples.

Results

There are myriad ways a large collection of simulated or experimental samples may be useful; we show one example by presenting findings that suggest typical PCR and capillary electrophoresis procedures used during human identity testing are sensitive enough to detect signal from one allele copy. We do this by evaluating simulated data generated by SEEIt and comparing the synthesized data to a large empirical data set of single-source samples amplified using 0.0078 ng of DNA (see Figure 1.).

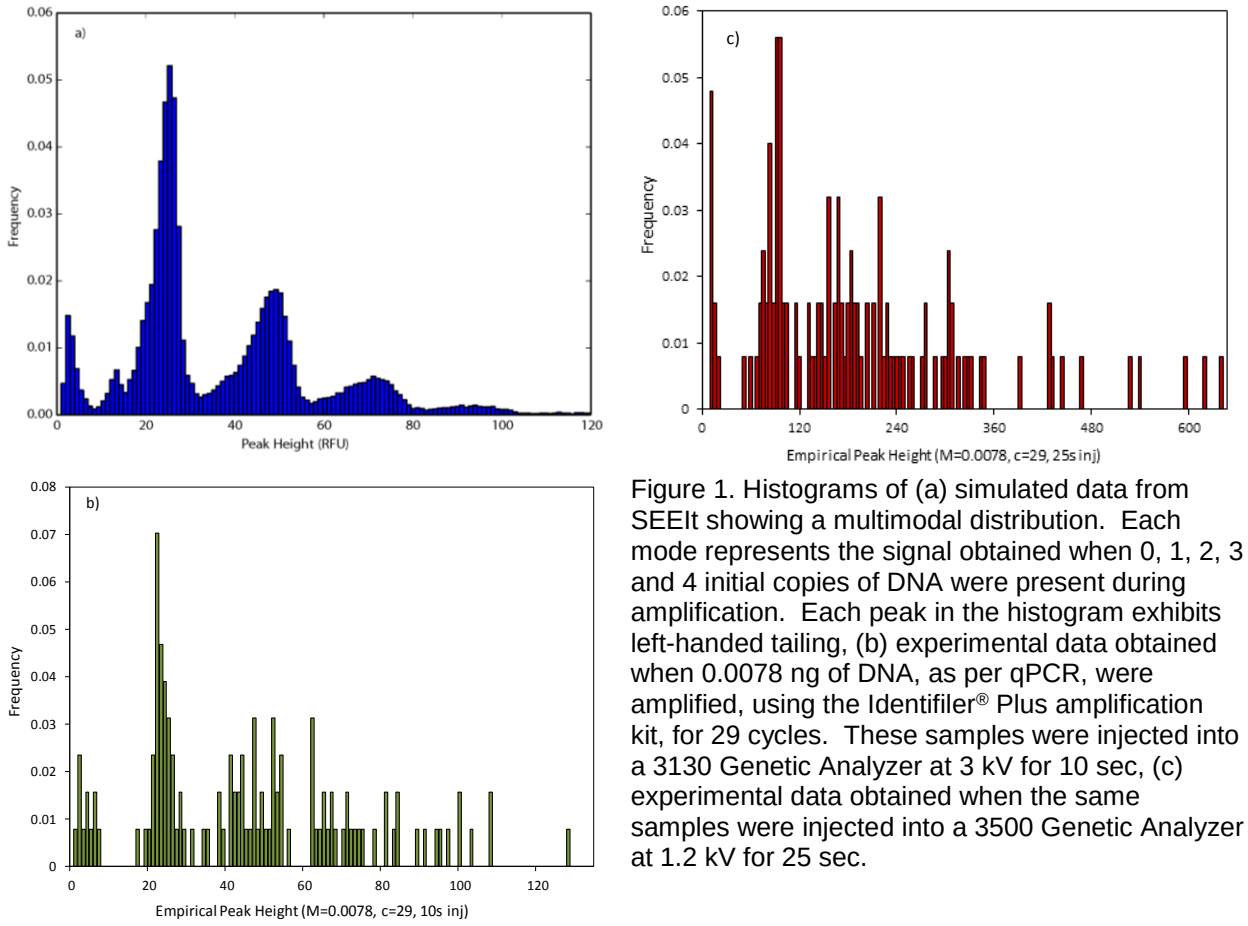

Figure 1. Histograms of (a) simulated data from SEEIt showing a multimodal distribution. Each mode represents the signal obtained when 0, 1, 2, 3 and 4 initial copies of DNA were present during amplification. Each peak in the histogram exhibits left-handed tailing, (b) experimental data obtained when 0.0078 ng of DNA, as per qPCR, were amplified, using the Identifiler® Plus amplification kit, for 29 cycles. These samples were injected into a 3130 Genetic Analyzer at 3 kV for 10 sec, (c) experimental data obtained when the same samples were injected into a 3500 Genetic Analyzer at 1.2 kV for 25 sec.

Histograms of the empirical and simulated signal in Figures 1a. and Figure 1b. demonstrate that the data are consistent; this confirms that simulated signal from SEEIt is a good representation of the signal obtained experimentally. The histogram derived from the experimental dataset (Figure 1b.) shows that low-level samples result in a multimodal pattern with, at least, three seemingly distinct peaks. For example, for the D8S1179 locus, the first, second and third signal groups are centered around 4, 24, and 47 RFU, respectively. Similarly, the simulated data (Figure 1a.) is multimodal. Through simulation we determine that the first mode consists largely of signal derived from instrumental noise; the second group is the signal obtained when one copy of DNA is amplified; and the third is the signal obtained when two copies of DNA are amplified.

Figure 1c. depicts the frequency of the peak height obtained when the same samples were injected for 25 sec on a 3500 Genetic Analyzer. A similar multimodal signal distribution is obtained; however, in this case, the first and second signal groups are centered around 16 and 86 RFU, respectively, suggesting that, regardless of platform, it is possible to implement a laboratory protocol and analytical threshold which together ensure that most amplicon-based signal is imported into the interpretation system(s) while retaining the ability to filter most of the noise.

This project was partially supported NIJ2011-DN-BX-K558 and NIJ2012-DN-BX-K050 and ARO RIF W911NF-14-C-0096 awarded by the National Institute of Justice, Office of Justice Programs, U.S. Department of Justice and the Department of Defense, Army Research Office, Rapid Innovation Fund, respectively. 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.

References

[1] Mullis, K., Faloona, F., Scharf, S., Saiki, R., Horn, G., Erlich, H., *Methods in Enzymology* 1986, *155*, 335-350.
[2] Swaminathan, H., Garg, A., Grgicak, C. M., Medard, M., Lun, D. S., *Forensic Sci. Int. Genetics*, *22*, 149-160.
[3] Swaminathan, H., Grgicak, C. M., Medard, M., Lun, D. S., *Forensic Sci. Int. Genetics*, *16*, 172-180.
[4] Perlin, M. W., Szabady, B., *Journal of Forensic Sciences* 2001, *46*, 1372-1378.
[5] Ballantyne, J., Hanson, E. K., Perlin, M. W., *Science and Justice* 2012, *53*, 103-114.
[6] Taylor, D., Bright, J.-A., Buckleton, J., *Forensic Science International: Genetics* 2013, *7*, 516-528.
[7] Cowell, R. G., Lauritzen, S. L., Mortera, J., *Forensic Sci. Int. Genetics* 2011, *5*, 202-209.
[8] Haned, H., Pene, L., Lobry, J. R., Dufour, A., Pontier, D., *Journal of Forensic Sciences* 2011, *56*, 23-28.
[9] Goor, R. M., Neall, L. F., Hoffman, D., Sherry, S. T., *Bulletin of mathematical biology* 2011, *73*, 1909-1931.
[10] Hansson, O., Gill, P., Egeland, T., *Forensic Sci. Int. Genetics* 2014, *13*, 154-166.
[11] Schneider, P. M., Fimmers, R., Keil, W., Molsberger, G., Patzelt, D., Pflug, W., Rothamel, T., Schmitter, H., Schneider, H., Brinkmann, B., *Int. J. Legal Med* 2009, *123*, 1-5.
[12] Gill, P., L. Gusmao, Haned, H., Mayr, W. R., Morling, N., Parson, W., Prieto, L., Prinz, M., Schneider, H., Schneider, P. M., Weir, B. S., *Forensic Science International - Genetics* 2012, *6*, 679-688.
[13] SWGDAM, 2010.