A SYSTEMS-BASED APPROACH TO VALIDATION IMPROVES MIXTURE INTERPRETATION OUTCOMES: HIGH-FIDELITY DATA FOR SINGLE-CELL AND BULK MIXTURE INTERPRETATION PIPELINES

Catherine M. Grgicak[1,2], Amanda J. Gonzalez[1,3], Harish Swaminathan[4], Kelsey C. Peters[4], Lauren Alfonse[4], Ken R. Duffy[5], Desmond S. Lun [2,6,7,8]
[1]Department of Chemistry, Rutgers University
[2]Center for Computational and Integrative Biology, Rutgers University
[3]Department of Biology, Rutgers University
[4]Biomedical Forensic Sciences Program, Boston University School of Medicine
[5]Hamilton Institute, Maynooth University
[6]Department of Computer Science, Rutgers University
[7]Department of Plant Biology and Pathology, Rutgers University
[8]School of Information Technology and Mathematical Sciences, University of South Australia

Much effort has focused on developing inference tools that determine the likelihood ratio (LR), which is the ratio of probabilities of the evidence given hypotheses provided by the prosecution and the defense. No matter how sound the reasoning behind these inference techniques, all must work with multi-cellular bulk-processed DNA signal containing peaks from an unknown number of, potentially, partial genomes, making mixture interpretation an arduous task. Mixture interpretation is further complicated by the fact that each laboratory works independently to validate their own process. Though the forensic DNA pipeline between laboratories consists of the same basic steps, the sensitivity of today's technology means that small modifications to laboratory protocols may have large impacts on signal detection in the low-template regimes, negatively impacting inference consistency between laboratories.

In response to these issues, we instituted a multi-institutional, inter-disciplinary effort with the aim of developing a validation tool designed to:

i) Cost-effectively optimize laboratory conditions by parameterizing an *in silico* model with the laboratories' own data to optimize resolution between noise and signal from a single copy of DNA.

ii) Demonstrate, using fully continuous probabilistic software designed to produce the distributions on the number of contributors (NOC) and the likelihood ratio (LR), respectively, that optimized laboratory conditions from (i) result in improved NOC and LR outcomes.

iii) Alleviate burdens associated with artifact filtering and reporting.

The full tool, named ValiDNA, carries out the simulation of a large number of artificial EPGs according to the model specified in[1]. Briefly, each simulation begins with random sampling of two distinct alleles at every locus. The objective is to investigate the ability to distinguish $\text{signal}_{1\text{copy}}$ from noise. As such, the number of copies of an allele that undergo the multitype Galton-Watson amplification is forced to be 1. After the last PCR cycle, we simulate the CE process, converting amplicons to fluorescence, and add noise signal in a random fashion. The sensitivity parameter, α, and the noise parameters are automatically estimated for each locus from the laboratory's own data from samples of known genotype. The sensitivity is a linear function of the nominal amplicon number versus peak height. Noise peaks are randomly simulated at every locus and their heights are sampled from a lognormal distribution[2]. The false positive detection rate (proportion of noise peaks above the signal threshold) and the false negative detection rate (proportion of allele peaks below the signal threshold) are reported for a range of thresholds. This allows for fast exploration of multifarious laboratory parameters, such as the number of PCR cycles, injection time and thresholds that enhance the discernment of signal from noise. The conditions that minimize detection error rates, while still maintaining a reasonable dynamic range, are chosen as the optimized post-PCR process. The simulated data is then tested using probabilistic inference software such as NOCIt and CEESIt, which are both fully continuous computational tools. NOCIt outputs the *a posteriori* probability distribution for n = 1 to 6 contributors, while CEESIt computes the LR and its distribution by comparing the profile to up to one billion randomly generated genotypes. We are, thus, able to confirm that data generated using optimized laboratory conditions result in inference outcomes that cannot be improved upon without material modification to the laboratory technology itself. This provides the laboratory a means to systematically evaluate the outcome of decisions on their entire laboratory process - from amplification to reporting - in a holistic manner.

Since ValiDNA reports optimized laboratory conditions for the single-copy regime, we used it to determine optimized conditions for single-cell pipelines using previously derived serial dilutions. Once completed, the LR obtained from single-cell data acquired by pico-picopipetting was compared with LRs obtained from equivalent bulk-mixture samples. As an example, a 1:1 v/v saliva mixture from two individuals, -A and –B, amplified at a target of 0.033 ng with the Globalfiler^{TM} kit showed that the LR obtained from the bulk sample (assigning a NOC of 2) resulted in a $\log_{10}$LR of -3.8 and 24.6 for Person -B and -A, respectively. When a portion of the same cellular admixture is processed through our single-cell pipeline, we obtain five electropherograms from five cells. Cell-03 resulted in few detected alleles while the other four cells resulted in high-quality profiles. No allele drop-in was detected. Using CEESIt to evaluate these high-fidelity profiles and assigning Person -B as the POI, we obtain $\log_{10}$LRs of 29.8 for Cells -01 and -05, respectively (Note: $\log_{10}$LRs were less than -48 for Cells – 02, -04 and -03). Though we do not advocate separate analysis of each EPG due to issues associated with family-wise error, we provide this as a demonstration that high-fidelity single-cell electropherograms generated from optimized laboratory conditions can provide information unavailable with traditional bulk-mixture pipelines.

[1] K.R. Duffy, N. Gurram, K.C. Peters, G. Wellner, C.M. Grgicak, Exploring STR signal in the single- and multicopy number regimes: Deductions from an in silico model of the entire DNA laboratory process, Electrophoresis (2016).

[2] J. Bregu, D. Conklin, E. Coronado, M. Terrill, R.W. Cotton, C.M. Grgicak, Analytical thresholds and sensitivity: establishing RFU thresholds for forensic DNA analysis, J Forensic Sci 58(1) (2013) 120-9.