NOC*It* – a computational tool to infer the number of contributors to a forensic DNA sample

Harish Swaminathan[1], Catherine M. Grgicak[2], Muriel Medard[3], and Desmond S. Lun[1,4,*]

[1]Center for Computational and Integrative Biology, Rutgers University, Camden, NJ 08102, USA
[2]Biomedical Forensic Sciences Program, Boston University School of Medicine, Boston, MA 02118, USA
[3]Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[4]School of Mathematics and Statistics, University of South Australia, Mawson Lakes, SA 5095, Australia
*Corresponding author. Tel.: +1-856-225-6094. E-mail: dslun@rutgers.edu

Introduction

DNA profiling is used to identify the perpetrator(s) of crimes when biological evidence is available. If a DNA profile is obtained, an assumption about the number of contributors to a sample is needed to compare the crime scene profile with that of a known [1]. Usually, the number of contributors to a crime scene sample is unknown and is estimated by the analyst based on the electropherogram obtained.

There are a number of issues associated with the process of creating an STR DNA profile that hinder the interpretation of a DNA profile. Stochastic effects associated with DNA extraction, the PCR process and pipetting lead to non-detection of alleles (dropout). Further, allele overlap and PCR amplification artifacts like stutter occur frequently and make it difficult to interpret low-template, mixture profiles [2-3].

Though methods to infer the number of contributors to a forensic sample exist [4-9], there are a number of issues associated with them. One of the main issues is that these methods do not use the quantitative data obtained, i.e. the heights of the peaks in the signal. Also, they do not examine effects of stutter, baseline noise or drop-out. As a result, these methods are not suitable for low template mixture interpretation

In response to the aforementioned issues, NOC*It*, a computational tool that calculates the probability distribution for the number of contributors to a DNA sample was developed. In addition to using the allele frequencies, NOC*It* works upon the quantitative data in the signal. It accounts for dropout of alleles, the formation of stutter peaks and the propensity for the baseline noise to increase with target. It is a tool that provides statistical evaluations of the possible number of contributors by considering all known confounding factors related to PCR and instrument interferences.

This represents a study designed to investigate the ability to infer the number of contributors to complex mixtures. The results obtained from NOC*It* were compared to those obtained via traditional methods. Specifically, the most likely number of contributors obtained from NOC*It* was compared to the most likely number of contributors estimated via the allele count and Maximum Likelihood Estimator methods. The relationship between the ability to provide only one conclusion regarding the number of contributors from DNA samples containing varying masses is also discussed.

NOCIt will soon be available to download as a Java application at www.bu.edu/dnamixtures.

Materials and methods

We used 1555 single source samples from 58 donors with known genotypes to calibrate NOC*It*. These samples were generated using the AmpF*l*str® Identifiler® Plus kit developed by Applied Biosystems (Foster City, California). Samples were amplified from 7 low template DNA amounts (0.007 – 0.25ng) and injected using a 3 kV injection voltage and 3 times of injection (5, 10 and 20s). In the profiles obtained, the peaks were separated into 1 of 3 categories: True peaks (all peaks representing the alleles in the sample), Stutter peaks (all peaks in the *n-4* position of True peaks) and Noise peaks (all other peaks in the signal). The heights of the peaks were modeled using the Gaussian distribution. Calibration parameters (namely the mean and the standard deviation) for the 3 categories of peaks were computed for each DNA amount at the 3 injection times for every locus. Dropout rates and the rate of occurrence of stutter were also computed at each DNA amount at the 3 injection times for every locus.

A Monte Carlo approach is used by NOC*It* to compute the likelihood for the number of contributors. At each iteration of the Monte Carlo process, genotypes for the *n* contributors are picked based on the frequencies of the alleles in the frequency table. Allele frequencies from the Caucasian population specified

in the AmpF*l*str® Identifiler® Plus manual were used [10]. A mixture ratio is picked at random - all mixture ratios are assumed to occur with equal probability. Modeling of the dropout frequencies, the log of the means and standard deviations of stutter ratios and means and standard deviations of true peak heights was carried out using the following distributions: exponentially decreasing curve ($y = ae^{bx}$), exponentially decreasing curve ($\log(y) = ae^{bx}+c$) and a line with a positive slope ($y = mx$) respectively. Curve fitting was done using MATLAB® (2011b, The Mathworks, Natick, Massachusetts, USA). For every allele in the genotype of the contributors, dropout of the allele is simulated by a Bernoulli trial. Based on the evidence observed, the likelihood of observing the heights of the peaks is then computed using the calibration data. The average of the values computed is the likelihood of observing the evidence at a locus, given *n* contributors. The likelihood values at all the loci are multiplied with each other to give the overall likelihood for an *n*. The *n* that results in the highest likelihood is the number of contributors most supported by the evidence as calculated by NOC*It*.

To test the performance of the software, NOC*It* was run on 1, 2 and 3 person mixtures. The performance of NOC*It* was compared with the Maximum Allele Count (MAC) and the Maximum Likelihood Estimator (MLE) methods. MAC uses the number of peaks observed in the signal to determine the number of contributors while MLE uses the number of peaks as well as the allele calls of the peaks [7]. Both the methods depend upon the establishment of a threshold to determine the set of true peaks. NOC*It*, on the other hand, does not depend upon the setting of a threshold and works on the entire electropherogram obtained. Two types of thresholds were used for MAC and MLE for comparison purposes. The first threshold was a constant threshold of 50 RFU at all the loci. The second threshold is a variable threshold, set as the height of the highest noise peak observed in the calibration data at a particular DNA amount, dye color and time of injection. The stutter filter specified by Applied Biosystems in the AmpF*l*str® Identifiler® Plus manual [10] was used to filter out the stutter peaks at each locus while applying the MAC and MLE methods.

Results

NOC*It* resulted in a higher accuracy (i.e. (most likely NOC/actual NOC) × 100%) than the other methods at every target at all injections. As the signal-noise ratio increased with an increase in the injection time, so did the accuracy of MAC and MLE. The performance of NOC*It* was unaffected by changes in the time of injection. The accuracy of all the 3 methods increased with an increase in DNA mass. Both the MLE and MAC resulted in accuracies < 80 % when the template level was < 0.25 ng. In contrast, the accuracy obtained with NOC*It* was < 80 % at targets < 0.047 ng. The higher accuracy of NOC*It* compared to MAC and MLE suggests that using the entire signal obtained, instead of applying a threshold and subsequently losing information, gives a better estimate about the number of contributors. Even in the instances where NOC*It* failed to identify the correct number of contributors, *it identified the region in which the number is most likely to lie*, which can be useful in the case of complex and/or low template samples.

References

[1] John S Buckleton, James M Curran: A discussion of the merits of random man not excluded and likelihood ratios. *Forensic Science International Genetics*, 2 (2008) 343-348.

[2] John S Buckleton, James M Curran, Peter Gill: Towards understanding the effect of uncertainty in the number of contributors to DNA stains. *Forensic Science International Genetics*, 1(1):20-8, 2007.

[3] John M. Butler: *Fundamentals of Forensic DNA Typing*. Academic Press, 2009.

[4] David R. Paoletti, Travis E. Doom, Carissa M. Krane, Michael L. Raymer and Dan E. Krane: Empirical Analysis of the STR Profiles resulting from Conceptual Mixtures. *Journal of forensic sciences*, 50(6), 2005.

[5] Jaheida Perez, Adele A. Mitchell, Nubia Ducasse, Jeannie Tamariz, Theresa Caragine: Estimating the number of contributors to two-,three-, and four-person mixtures containing DNA in high template and low template amounts. *Croatian Medical Journal*, 52(3):314-326, 2011.

[6] A. Biedermann, S. Bozza, K. Konis, F. Taroni: Inference about the number of contributors to a DNA mixture: Comparative analyses of a Bayesian network approach and the maximum allele count method. *Forensic Science International: Genetics*,6 (2012) 689-696.

[7] Hinda Haned, Laurent Pene, Jean R. Lobry, Anne B. Dufour and Dominique Pontier: Estimating the Number of Contributors to Forensic DNA Mixtures: Does Maximum Likelihood perform better than Maximum Allele Count? *Journal of Forensic Sciences*, 56(1), 2011.

[8] Thore Egeland, Ingvild Dalen, Petter F. Mostad: Estimating the number of contributors to a DNA profile. *International Journal of Legal Medicine*, 117:271-275, 2003.

[9] David R. Paoletti, Dan E. Krane, Michael L. Raymer and Traves E. Doom: Inferring the Number of Contributors to Mixed DNA Profiles. *IEEE/ACM Transactions on Computational Biology and Bioinformatics,* 9(1), 2012.

[10] http://tools.lifetechnologies.com/content/sfs/manuals/cms_076395.pdf.