Forensic Science International: Genetics 16 (2015) 172–180

Contents lists available at ScienceDirect

Forensic Science International: Genetics

journal homepage: www.elsevier.com/locate/fsig

NOC*It*: A computational method to infer the number of contributors to DNA samples analyzed by STR genotyping

Harish Swaminathan [a], Catherine M. Grgicak [b], Muriel Medard [c], Desmond S. Lun [a,d,*]

[a] Center for Computational and Integrative Biology, Rutgers University, Camden, NJ 08102, USA
[b] Biomedical Forensic Sciences Program, Boston University School of Medicine, Boston, MA 02118, USA
[c] Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[d] School of Mathematics and Statistics, University of South Australia, Mawson Lakes, SA 5095, Australia

ARTICLE INFO

Article history:
Received 29 April 2014
Received in revised form 24 September 2014
Accepted 9 November 2014

Keywords:
Short tandem repeats
Number of contributors
DNA
Mixture interpretation

ABSTRACT

Repetitive sequences in the human genome called short tandem repeats (STRs) are used in human identification for forensic purposes. Interpretation of DNA profiles generated using STRs is often problematic because of uncertainty in the number of contributors to the sample. Existing methods to identify the number of contributors work on the number of peaks observed and/or allele frequencies. We have developed a computational method called NOC*It* that calculates the *a posteriori* probability (APP) on the number of contributors. NOC*It* works on single source calibration data consisting of known genotypes to compute the APP for an unknown sample. The method takes into account signal peak heights, population allele frequencies, allele dropout and stutter—a commonly occurring PCR artifact. We tested the performance of NOC*It* using 278 experimental and 40 simulated DNA mixtures consisting of one to five contributors with total DNA mass from 0.016 to 0.25 ng. NOC*It* correctly identified the number of contributors in 83% of the experimental samples and in 85% of the simulated mixtures, while the accuracy of the best pre-existing method to determine the number of contributors was 72% for the experimental samples and 73% for the simulated mixtures. Moreover, NOC*It* calculated the APP for the true number of contributors to be at least 1% in 95% of the experimental samples and in all the simulated mixtures.

1. Introduction

Short tandem repeats, or STRs, are repetitive sequences 1–7 base pairs in length that are scattered throughout the human genome. One of the commonly used applications of STRs is in the field of human identification for forensic purposes [1]. An STR DNA profile developed from a biological sample collected at a crime scene is compared with that of a person of interest or run against a database to check for a match.

The Scientific Working Group on DNA Analysis Methods (SWGDAM) recommends that forensic reports include a statement as to the assumption made about the number, or the minimum number of contributors, to the sample being investigated [2]. The number of contributors to a crime scene sample is generally unknown and must be estimated by the analyst based on the electropherogram obtained. The assumption on the number of contributors affects statistics used to assess the weight of DNA evidence (e.g., the likelihood ratio) [3]. Thus, it is useful to have a good estimate on the number of contributors to the sample.

There are issues associated with the process of generating a DNA profile that hinder the interpretation of a profile. Stochastic effects associated with DNA extraction, the PCR process and pipetting lead to non-detection of alleles (dropout). Further, allele sharing and PCR amplification artifacts like stutter occur frequently and make it difficult to interpret low-template, mixture profiles [4]. These make it difficult to accurately estimate the number of contributors to a sample.

Methods have previously been developed to infer the number of contributors to a DNA sample. The most widely used method is maximum allele count (MAC). This method seeks to identify the minimum number of individuals who could have contributed to a sample by counting the number of alleles observed at each locus, taking the maximum value over all the loci and dividing it by two. The MAC method may not work well with complex mixtures

* Corresponding author at: Department of Computer Science, Rutgers University, 227 Penn Street, Camden, NJ 08102, USA.
Tel.: +1 856 225 6094; fax: +1 856 225 6624.
E-mail address: dslun@rutgers.edu (D.S. Lun).

http://dx.doi.org/10.1016/j.fsigen.2014.11.010
1872-4973/

because of allele-sharing among the contributors [5]. Guidelines for estimating the number of contributors using the total number of alleles observed were established for high template and low template samples by Perez et al. [6]. Methods that do not solely rely upon the number of alleles observed but also use the frequencies of the alleles observed in the population, have been created. A probabilistic approach was developed by Biedermann et al. [7], employing a Bayesian network, to infer the number of contributors to forensic samples. This method was shown to work better than MAC with degraded DNA and with higher number of contributors. Haned et al. [8] extended the work of Egeland et al. [9] on diallelic markers to the multi-allelic markers that are commonly used in creating STR profiles to develop a maximum likelihood estimator (MLE) for the number of contributors, taking into account population substructure. This method was also shown to give more accurate results than MAC with higher number of contributors and degraded DNA. A probabilistic mixture model was used by Paoletti et al. [10] to infer the number of contributors to a sample based on the frequencies of the alleles observed. This method, like MLE, accounts for correction due to population substructure.

We have developed NOC*It* (*NOC*: number of contributors)—a computational tool that calculates the a posteriori probability (APP) on the number of contributors to a DNA sample. In addition to using the qualitative information contained in the signal, i.e. the allele frequencies, NOC*It* also makes use of the quantitative information present, i.e. the heights of the peaks. The heights of the peaks increase with an increase in the amount of input DNA and are an indicator of the mixture ratio and the number of copies of an allele that gave rise to a peak. This is information that could be used in estimating the number of individuals that gave rise to a sample. In addition, NOC*It* accounts for the dropout of alleles and the formation of stutter peaks. Out of the 278 experimental samples tested, NOC*It* correctly identified the number of contributors in 83% of the samples, while the accuracy of the best pre-existing method was 72%. NOC*It* also correctly identified the number of contributors in 85% of the 40 simulated mixtures used for testing.

2. Material and methods

2.1. Calibration of NOCIt

NOC*It* uses the quantitative information contained in the signal in the form of peak heights to calculate the probabilities for the number of contributors. This involves characterizing the dependence of variables such as probability of dropout, probability of stutter and true, stutter and baseline noise peak heights on the input DNA mass. This is done by using single source calibration samples with known genotypes obtained from samples amplified from a wide range of input DNA masses.

To generate the calibration samples (calibration set – Table 1), high molecular weight DNA was extracted from 35 single source samples using standard organic extraction procedures. The samples were whole blood, dried blood stains or saliva. The blood stains were either on Whatman® paper or cloth swatches. Saliva samples were either whole saliva or dried buccal swabs on cotton. Briefly, the organic extraction consisted of incubating the sample in 300 μg/mL of Proteinase K and 2% v/v SDS (sodium dodecyl sulfate) solution at 37 °C for 2 h to overnight. Purification was accomplished with phenol/chloroform and alcohol precipitation. The DNA was dissolved in 50 μL of TE buffer (10 mM Tris, 0.1 mM EDTA, pH 8.0) at 56 °C for 1 h. Absolute DNA quantification was performed using real-time PCR and the Quantifiler® Duo™ Quantification kit according to the manufacturer's recommended protocol and one external calibration curve [11,12]. A 7500 Sequence Detection System (Life Technologies, Inc.) was used for C_t (cycle threshold) detection. The extracted DNA was amplified using the manufacturer's recommended protocol (29 cycles) for AmpFℓSTR® Identifiler® Plus Amplification Kit (Life Technologies, Inc) [13]. Single source samples were amplified using 0.25, 0.125, 0.063, 0.047, 0.031, 0.016 and 0.008 ng of DNA. The PCR reaction consisted of 15 μL of master mix, the calculated volume of template DNA based on target mass required, and enough Tris-EDTA (TE) buffer (10 mM at pH 8.0) to bring the total reaction volume to 25 μL. Amplification was performed on Applied Biosystems' GeneAmp® PCR System 9700 using 9600 emulation mode. Positive and negative amplification controls were also run and showed expected results (data not shown). Fragment separation was accomplished by using a 3130 Genetic Analyzer (Life Technologies, Inc.) and a mixture containing appropriate amounts of HiDi (highly deionized) formamide (8.7 μL/sample) (Life Technologies, Inc.) and GeneScan™-600 LIZ™ Size Standard (0.3 μL/sample) (Life Technologies, Inc.). A volume of 9 μL of that mixture and 1 μL of sample, negative or ladder was added to the appropriate wells. The samples were incubated at 95 °C for 3 min and snap-cooled at −20 °C for 3 min. Five, ten, and twenty second injections at 3 kV were performed on each of the samples and run according to the manufacturer's recommended protocol [13]. Fragment analysis was performed using GeneMapper IDX v1.1.1 (Life Technologies, Inc.) using Local Southern sizing and an RFU threshold of 1. A threshold of 1 RFU was used in order to capture all peak height information, i.e. the allelic peaks, baseline noise and stutter peaks, in the signal. Known artifacts such as pull-up, spikes, −A, and artifacts due to dye dissociation were manually removed. A peak was considered pull-up if it was the same size (±0.3 bp) as a larger peak in another color and below 5% of the height of the larger peak. Peaks were determined to be 'spikes' if they were in greater than 2 colors and in the same position. Peaks were considered to be −A if they were one base pair smaller than an allele and peaks determined to originate from dye dissociation had to be in the same position, in the same color channel and be observed in multiple samples. The genotypes table, which included the file name, marker, dye, allele, size and height, was exported.

Table 1
Number of single source samples used for the calibration of NOC*It*.

Injection time (s)	DNA amount (ng)	Number of samples	Injection time (s)	DNA amount (ng)	Number of samples	Injection time (s)	DNA amount (ng)	Number of samples
5	0.008	35	10	0.008	56	20	0.008	35
	0.016	35		0.016	53		0.016	35
	0.031	36		0.031	54		0.031	36
	0.047	32		0.047	33		0.047	33
	0.063	34		0.063	49		0.063	35
	0.125	35		0.125	53		0.125	35
	0.25	35		0.25	59		0.25	33
	Total	242		Total	357		Total	242