Contents lists available at ScienceDirect

Forensic Science International: Genetics

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

Research paper

A large-scale validation of NOCIt's *a posteriori* probability of the number of contributors and its integration into forensic interpretation pipelines

Catherine M. Grgicak[a,b,*], Slim Karkar[b], Xia Yearwood-Garcia[c], Lauren E. Alfonse[c], Ken R. Duffy[d], Desmond S. Lun[b,e,f]

[a] *Department of Chemistry, Rutgers University, Camden, NJ, 08102, USA*
[b] *Center for Computational and Integrative Biology, Rutgers University, Camden, NJ, 08102, USA*
[c] *Biomedical Forensic Sciences Program, Boston University School of Medicine, Boston, MA, 02118, USA*
[d] *Hamilton Institute, Maynooth University, Ireland*
[e] *Department of Computer Science, Rutgers University, Camden, NJ, 08102, USA*
[f] *Department of Plant Biology, Rutgers University, New Brunswick, NJ, 08901, USA*

ARTICLE INFO

Keywords:
Forensic DNA
DNA mixtures
Number of contributors
NOCIt
CEESIt
Likelihood ratios

ABSTRACT

Forensic DNA signal is notoriously challenging to interpret and requires the implementation of computational tools that support its interpretation. While data from high-copy, low-contributor samples result in electropherogram signal that is readily interpreted by probabilistic methods, electropherogram signal from forensic stains is often garnered from low-copy, high-contributor-number samples and is frequently obfuscated by allele sharing, allele drop-out, stutter and noise. Since forensic DNA profiles are too complicated to quantitatively assess by manual methods, continuous, probabilistic frameworks that draw inferences on the Number of Contributors (NOC) and compute the Likelihood Ratio (LR) given the prosecution's and defense's hypotheses have been developed.

In the current paper, we validate a new version of the NOCIt inference platform that determines an A Posteriori Probability (APP) distribution of the number of contributors given an electropherogram. NOCIt is a continuous inference system that incorporates models of peak height (including degradation and differential degradation), forward and reverse stutter, noise and allelic drop-out while taking into account allele frequencies in a reference population. We established the algorithm's performance by conducting tests on samples that were representative of types often encountered in practice. In total, we tested NOCIt's performance on 815 degraded, UV-damaged, inhibited, differentially degraded, or uncompromised DNA mixture samples containing up to 5 contributors. We found that the model makes accurate, repeatable and reliable inferences about the NOCs and significantly outperformed methods that rely on signal filtering.

By leveraging recent theoretical results of Slooten and Caliebe (FSI:G, 2018) that, under suitable assumptions, establish the NOC can be treated as a nuisance variable, we demonstrated that when NOCIt's APP is used in conjunction with a downstream likelihood ratio (LR) inference system that employs the same probabilistic model, a full evaluation across multiple contributor numbers is rendered. This work, therefore, illustrates the power of modern probabilistic systems to report holistic and interpretable weights-of-evidence to the trier-of-fact without assigning a specified number of contributors or filtering signal.

1. Introduction

Forensic DNA evidence is typically processed using the following steps: 1) the sample is collected and submitted for testing; 2) the DNA is extracted; 3) the concentration of DNA is approximated by qPCR; 4) a portion of the extract is amplified; 5) the amplified fragments are separated and detected using capillary electrophoresis and laser induced fluorescence, respectively; 6) the data are processed and the peaks are sized; and finally, 7) the peak information is interpreted to evaluate a NOC or LR. In Step 6, an analytical threshold (AT) and artifact filtering rules are often applied to the post-processed peaks, potentially significantly affecting assessments on the number of contributors and subsequent LR.

Since the NOC assumption can substantially affect downstream

[*] Corresponding author at: Department of Chemistry, Rutgers University, 315 Penn Street R306C, Camden, NJ, 08102, USA.
E-mail address: c.grgicak@rutgers.edu (C.M. Grgicak).

https://doi.org/10.1016/j.fsigen.2020.102296
Received 7 May 2019; Received in revised form 26 March 2020; Accepted 27 March 2020
Available online 11 April 2020

outcomes, particularly in forensic cases where the contributions from any one person within the DNA mixture are small or encumbered by signal from other contributors [1,2], there is interest in engineering methods that effectively and accurately provide inferences on the likely number of contributors that comprise the signal. Indeed, estimating the NOC to a sample has gained traction in other domains where the authors of [3] demonstrate that estimating the NOC in the gut contents of fish can inform predation rates.

The traditional method of inferring a sample's NOC is by a binary method termed 'Maximum Allele Count' (MAC), where the maximum number of peaks categorized as alleles per locus are counted, divided by two and rounded up. The MAC approach, therefore, reports the minimum number of contributors that explains the profile. Notably, MAC relies on the application of an AT and stutter filters. That is, peaks below the AT or stutter filter thresholds are not considered during interpretation, while peaks above those thresholds are used to assign the NOC, impacting downstream interpretation [2]. It is well established that reliance on categorical determinations of allele presence leads to underestimations of the NOC [4], which does not significantly improve by amplifying additional STRs [5,6] and is exacerbated in the presence of drop-out [7,8]. In addition, recently published results [9] demonstrate that effectively combining manual evaluation of peak height ratios with allele counting techniques is likely to remain unsubstantiated given the number of genotype combinations that explain a complex mixture. Despite numerous publications demonstrating counting methods provide an incomplete evaluation of the evidence, it is a method that continues to see widespread use in practice.

While evaluation of the NOC and determination of LRs have traditionally been distinct endeavors, a mathematical basis for computing an end-to-end, overall or full LR that treats the NOC as a nuisance parameter has recently been theoretically established by Slooten and Caliebe [10]. Briefly, for an electropherogram E, with N being the unknown number of contributors, H_p being the prosecution's hypothesis and H_d being the defense's hypothesis, if it is assumed that the *a priori* distribution of the number of contributors is the same under both hypotheses in the absence of any data, i.e., $P(N = n|H_p) = P(N = n|H_d)$, and that the posterior probability of the NOC given the defense's hypothesis is positive for all possible NOCs, $P(N = n|E, H_d) > 0$ for all n, then it is shown in [10] that the end-to-end LR satisfies

$$\begin{aligned} LR(E) = \frac{P(E|H_p)}{P(E|H_d)} &= \sum_n \left(\frac{P(E|H_p, N = n)}{P(E|H_d, N = n)} P(N = n|E, H_d)\right) \\ &= \sum_n \mathrm{LR}(E|N = n) P(N = n|E, H_d) \end{aligned} \tag{1}$$

That is, the likelihood ratio of the electropherogram given the two hypotheses is the sum of the APP that the NOC is n under the defense hypothesis given the data, $P(N = n|E, H_d)$, times the likelihood ratio conditioned on the number of contributors being n, $P(E \,|\, H_p, N = n)/P(E \,|\, H_d, N = n)$. For completeness, a derivation of Slooten and Caliebe's result, Eq. (1), is given in the appendix.

This result implies that if, in the absence of data, the prior prosecution and defense hypotheses regarding the NOC are equal, one can treat the NOC as a nuisance variable in the LR computation. This derivation assumes that the probabilistic model used for the computation of the APP and the conditional LRs is the same, and the data provided to each are coincident, which is important for its use in practice.

To evaluate the LR with Eq. (1) we require two distinct inferences: one that evaluates the APP of the NOC given the defense's hypothesis; and one that evaluates the LR for each possible NOC. While there are alternative expressions that can be used to compute the LR obtained by Eq. (1) [4], the expression given in Eq. (1) decomposes the LR into two parts that can be easily and independently interpreted, thus maintaining consistency with the prevailing approach among existing computational procedures for LR calculation that rely on an assignment for the NOC.

The current paper first focuses on validation of an updated version of NOCIt, which stands for Number of Contributors, a software system that computes an approximation to the APP, $P(N = n|E, H_d)$, for all values of n up to $n_{max} = 6$ by utilizing all post-processed information contained in the electropherogram. A prototype version of NOCIt, described in [11], utilized relationships between peak heights and the mass of DNA amplified as determined by qPCR. Despite its simplicity, the prototype outperformed the MAC and Haned et al.'s Maximum Likelihood Estimation (MLE) methods for samples containing up to five contributors [11,12] justifying further development of the biological models, which are fully described in [13,14]. The updated probabilistic models and computational environment underlying NOCIt have also been incorporated into CEESIt (for Computational Evaluation of Evidentiary Signal), a software that computes a Monte Carlo approximation to the conditional likelihood ratio $P(E \,|\, H_p, N = n)/P(E \,|\, H_d, N = n)$ [15].

NOCIt reports a probability distribution, the APP, which has a number of useful features. In general, it is not possible to determine the NOC without uncertainty and so, in reporting a probability distribution, a key feature of NOCIt is that it enables this uncertainty to be quantified and reported. In addition, it allows for the elimination of unlikely NOCs from consideration and, notably, provides the probability distribution on the NOC for full downstream interpretation.

Downstream interpretation in forensics could take a number of forms. The simplest method, conceptually, is to use the most likely NOC from the APP (the so-called Maximum A Posteriori Probability, or MAP, estimate) instead of one obtained by MAC or an alternative method [16–18] as the NOC assumption for LR computation. The advantage of the MAP estimate over alternative estimates is that, subject to the assumptions on the prior distribution of the NOC and the probabilistic model of the electropherogram, it theoretically minimizes the probability of error. Even with the MAP estimate, however, the probability of error may be significant. We have observed, for example, cases where the APP has significant, non-zero probabilities for two or more possible values of the NOC, implying that any estimator that produces only a single point estimate of the NOC may have significant probability of error. Indeed, while the MAP estimate (or alternative point estimates) are appealing because of their simplicity, previous studies have argued against their use [12]. An alternative approach is to use the APP to eliminate extremely unlikely NOCs from consideration and for LRs to be presented and interpreted only for the remaining NOCs. The APP, therefore, provides a quantitative basis by which the assumptions underlying the various LRs can be assessed. Yet another approach is to incorporate the APP into the LR calculation and treat the NOC as a nuisance variable as per Eq. (1), thus computing an overall LR across multiple n. In addition, non-forensic [3] or investigative uses, such as informing investigators of the number of individuals that explain the data in the absence of a suspect, may also gain traction. Whatever the method or application, we see the APP as being useful for determining the NOC and for downstream interpretation and [1–40] seek to report a large-scale experimental validation of NOCIt as a method for its computation.

A key design feature of NOCIt, and its sister program CEESIt, is that they work on all post-processed peaks, making the application of an AT a superfluous task. Since a decrease in the AT will necessarily increase the chance of detecting instrument and PCR artifacts – potentially increasing the burden associated with electropherogram review – we have implemented a module, named CleanIt, that automatically filters minus A (i.e., incomplete adenylation of newly polymerized DNA fragments), pull-up and raised baseline, using parameters provided by the laboratory. Though the application of an AT is unnecessary for the full interpretation pipeline, it remains an option as seemingly contradictory findings on the effects of ATs on LRs have been presented by the authors of [19], who demonstrated that for the 72 samples used in their study, a higher AT improved LR results for two probabilistic programs, and the authors of [20], who demonstrated that optimized ATs based on bioanalytical principles improve and stabilize inference outcomes across laboratory systems.

In what follows, we explore the performance of a continuous interpretation system that approximates the probability of n contributors given the evidence, $P(N = n|E, H_d)$, by evaluating the impact of various conditions and sample types on repeatability and true positive rates acquired by NOCIt. We supplement the literature by once again demonstrating that methods that rely on binary signal decisions do not provide correct NOC estimates in many cases. Motivated by the above and the description in [10], we show the way in which APPs calculated by NOCIt are used to calculate an overall LR for all probable values of n. The performance of NOCIt was tested on a dataset of 815, 1- to 5-contributor DNA profiles garnered from pristine, damaged, inhibited and differentially degraded samples [8]. This dataset is publicly available, making our study a potential benchmark against which newly developed NOC methods can be compared. Finally, we demonstrate that through the use of NOC inferences and LR computations, conditional on the use of the same probabilistic model, an end-to-end inference of the LR can be made.

2. Description of NOCIt

NOCIt computes the APP distribution $\mathrm{P}(N = n \mid E, H_d)$ for $n = 0, \ldots, n = n_{\max}$ as $\mathrm{P}(N = n \mid E, H_d) = (\mathrm{P}(E \mid N = n, H_d) \times \mathrm{P}(N = n|H_d))/\mathrm{Pr}(E|H_d)$, where E is the evidence (the electropherogram), N is the number of contributors, H_d is the defense hypothesis (i.e., that, given that $N = n$, the electropherogram arises from n random unrelated contributors from a population with known DNA profile frequencies), and $n_{\max}$ is a maximum number of possible contributors in a sample that is assumed *a priori*. We assume a uniform prior for N over $0, 1, \ldots, n_{\max}$ (that is, $\mathrm{P}(N = n|H_d) = 1/(n_{\max} + 1)$ for $n = 0, \ldots, n = n_{\max}$, and $\mathrm{Pr}(N = n|H_d) = 0$ otherwise). Since we assume a uniform prior, we have $\mathrm{P}(N = n \mid E, H_d) \propto \mathrm{P}(E \mid N = n, H_d)$.

In the absence of evidence external to the electropherogram to inform prior probabilities on the NOC, the choice of an uninformative uniform prior is reasonable. It is interpreted as saying that, prior to the observation of the electropherogram, it is assumed that all numbers of contributors between 0 and $n_{\max}$ are equally likely. If evidence external to the electropherogram suggests distinct a priori probabilities, that information can readily be incorporated by taking the APP obtained under the assumption of an uninformative prior and applying a simple re-weighting. All results presented in this paper are obtained using a uniform prior.

For a given sample and a given NOC, **n**, we denote by Θ the n-dimensional vector of DNA mixture proportions with components Θ_i. As $\sum_{i=1}^{n} \Theta_i = 1$, Θ takes values in the $(n-1)$-simplex $\Delta^{n-1} = \{(\theta_1, \ldots, \theta_n) \in \mathbb{R}^n \mid \sum_{i=1}^{n} \theta_i = 1,\ \theta_i > 0\ \forall\ i\}$. Because the degree of DNA degradation may differ among the contributors, the DNA mixture proportions may be different at distinct DNA fragment lengths. Let Θ_i^l be the mixture proportions for a length l bp fragment. We set $\Theta_i = \Theta_i^0$, the DNA mixture proportions at a putative 0 bp fragment length. Furthermore, we define $\Lambda = (1, \Lambda_2, \ldots, \Lambda_n)$ to be a n-dimensional vector with components $\Lambda_i = (\Theta_i^{200}/\Theta_1^{200})/(\Theta_i/\Theta_1) > 0$ that record the change in mixture proportion for contributor i, relative to the mixture proportion of contributor 1, at a reference length of 200 bp from that at 0 bp. For example, suppose that we have a given NOC of $n = 4$, that $\Theta = (3/8,\ 2/8,\ 2/8,\ 1/8)$, and that $\Lambda = (1,\ 1/2,\ 1/2,\ 1)$. This would imply that, at 0 bp, the mixture proportions are $\Theta^0 = \Theta = (3/8,\ 2/8,\ 2/8,\ 1/8)$ and that, at 200 bp, the mixture proportions are $\Theta^{200} = (3/6,\ 1/6,\ 1/6,\ 1/6)$. (The vector Θ^{200} can be computed by taking the Hadamard, or entrywise, product $\Theta \circ \Lambda = (3/8,\ 2/8,\ 2/8,\ 1/8) \circ (1,\ 1/2,\ 1/2,\ 1) = (3/8,\ 1/8,\ 1/8,\ 1/8)$ and normalizing by $\sum_{i=1}^{n} (\Theta \circ \Lambda)_i = 6/8$.) We see that, relative to the proportion contributed by contributor 1, the proportions contributed by contributors 2 and 3 have decreased (implying a greater differential rate of degradation) because $\Lambda_2 = \Lambda_3 = 1/2$, while the relative proportion contributed by contributor 4 stays the same because $\Lambda_4 = 1$. Note that, while the absolute mixture proportions for contributors 1 and 4 have increased from 0 bp to 200 bp, this example is valid only if, owing to overall degradation of the sample, the absolute contribution levels of contributors 1 and 4 have decreased from 0 bp to 200 bp, since each contributor's absolute contribution level can only decrease with increasing fragment size, not increase.

To quantify the overall signal amplitude of the sample, we fit an estimate of decayed amplitude for each dye color as previously described [13,14]. In brief, at a locus l, for n observed peaks of height $h_1, \ldots h_n$ at alleles of size $s_1, \ldots s_n$ bps, we define the amplitude of the signal at the locus as $H_l = \sum_{i=1}^{n} h_i$ and the weighted average size $\bar{s}_l$ of the alleles at the locus by $\bar{s}_l = (\sum_{i=1}^{n} h_i s_i)/H_l$. For each dye color c, we have a set of m loci $l_1, \ldots, l_m$ of that color, each with their corresponding weighted average sizes $\bar{s}_{l_1}, \ldots, \bar{s}_{l_m}$ and amplitudes $H_{l_1}, \ldots, H_{l_m}$. To these weighted average sizes and amplitudes, we fit an exponential regression curve of the form $f_c(s) = A_c e^{B_c s}$. Thus, for each dye color c, we have parameters A_c and B_c, which we call the quantification parameters for c.

Thus, NOCIt computes:

$$\mathrm{P}(E|N = n, H_d)$$
$$= \iint_{(\theta,\lambda)\in(\Delta^{n-1}\times\Omega^{n-1})} \mathrm{P}(E|\ \Theta = \theta,\ \Lambda = \lambda,\ N = n,\ H_d) f_{\Theta}(\theta) f_{\Lambda|\Theta}(\lambda|\theta)\mathrm{d}\theta\mathrm{d}\lambda$$

Where f_Θ is the probability density function of Θ and $f_{\Lambda|\Theta}$ is the conditional probability density function of Λ given Θ. In practice, we approximate the integral over θ as a discrete sum, where each θ_i takes values in $\left\{\frac{1}{d}, \ldots, \frac{d-1}{d}, 1\right\}$, $d \in \mathbb{N}$, and d is a parameter that represents the *Discretization Level*, as input by the user, while Θ is assumed to be uniform over the discrete set of points in Δ^{n-1} satisfying these discretization levels. For Λ, let $\Omega^{n-1} = \{(\lambda_2, \ldots, \lambda_n) | \lambda_i \in \{0.5, 1, 2\}\}$. For a given $\Theta = \theta$, we assume that Λ is uniform over Ω^{n-1}, excluding those points in Ω^{n-1} that would result in the absolute contribution level of one or more contributors increasing from 0 bp to 200 bp. More specifically, for each dye color we have an exponential curve $A_c e^{B_c s}$ that defines the expected signal amplitude as a function of fragment size s. At 0 bp, the absolute contribution level of contributor i is $A_c \theta_i$, while at 200 bp, their absolute contribution level is $A_c \theta_i^{200} e^{200 B_c}$. Thus, we require $A_c \theta_i \geq A_c \theta_i^{200} e^{200 B_c}$ for all contributors i and remove any points from Ω^{n-1} where this condition is not satisfied. Note that, even though Λ is defined relative to contributor 1, this distribution for Λ does not result in bias because of symmetry of the contributors.

Given Θ and Λ, the electropherogram for each locus l, E_l, is assumed independent of the electropherogram at every other locus. Therefore,

$$\mathrm{P}(E|\ \Theta = \theta,\ \Lambda = \lambda,\ N = n\ ,\ H_d) = \prod_{l \in L} \mathrm{P}(E_l|\ \Theta = \theta,\ \Lambda = \lambda,\ N = n,\ H_d),$$

where L is the set of loci.

Let G be the genotypes of the contributors. Let $D \in \{0,1\}^{2\times|L|\times n}$ be a $2 \times |L| \times n$ matrix that represents the drop-out configuration for the alleles of these contributors. If D_{ijk}, the i, j, k th entry of D, is 0, then the ith allele at the j th locus of the k th contributor has dropped out, while if D_{ijk} is 1, it has not dropped out. We have

$$\mathrm{P}(E_l|\ \Theta = \theta,\ \Lambda = \lambda,\ N = n,\ H_d)$$
$$= \sum_{g,d \in \Gamma^{\mathrm{n}} \times \{0,1\}^{2\times|L|\times n}} \left\{ \begin{matrix} P(E_l|G = g,\ D = d,\ \Theta = \theta,\ \Lambda = \lambda,\ N = n,\ H_d) \\ \bullet P(D = d|\ \Theta = \theta,\ \Lambda = \lambda)\mathrm{P}(G = g) \end{matrix} \right\}$$

Where Γ^n is the space of all possible genotypes for n contributors. The electropherogram E_l is a vector of peak heights. Given G and D, $\mathrm{P}(E_l|G = g,\ D = d,\ \Theta = \theta,\ \Lambda = \lambda,\ N = n,\ H_d)$ is calculated by classifying each peak as an allele peak, a forward stutter peak, a reverse stutter peak, noise, or the additive combination of allele and stutter, and assessing the probability of achieving the vector of peak heights E_l given their classification.

In prior work [13,14], we assessed and evaluated a range of probabilistic models for each one of these classifications. In the language of that paper, the final recommended models were TP4, SP1, NP2, TDO1, SDO1 and NDO2, which were used in the present paper. A complete description of those can be found in that published article, but, in summary, peak heights were modeled using Gaussian random variables, and drop-out events were modeled as Bernoulli random variables. The means and standard deviations of the Gaussians are functions of the decayed amplitude $A_c e^{B_c s}$, where s is the size of the allele in bp, for true and noise peak heights, but stutter peak height values were instead a function of parent peak heights. True peak drop-out probabilities were modeled as exponentially decaying in decayed amplitude, stutter peak drop-out probabilities as exponentially decaying in the parent peak height, and the probability of noise drop-out was estimated by determining the relative frequency of possible allele positions (i.e., bins) with no RFU signal. The parameters for these models were estimated using a set of calibration data (cf. Section 3.2), and these parameters were used for all the validation tests. The terms $\mathrm{P}(E_l|G = g,\ D = d,\ \Theta = \theta,\ \Lambda = \lambda,\ N = n,\ H_d)$ and $\mathrm{P}(D = d|\ \Theta = \theta,\ \Lambda = \lambda)$ are calculated from these models in the same manner as CEESIt, as described in [15]. The distribution $\mathrm{P}(G = g)$ is calculated from allele frequency distributions at each locus, which are derived from population-specific data tables, assuming a value of 0.01 for F_{ST} (referred to as 'theta' in in NOCIt's GUI). Specifically, at a given locus, the probability that the n + 1 st allele is of type a given that n alleles have been sampled, of which n_a were type a is calculated as $\frac{n_a F_{ST} + (1 - F_{ST})p_a}{1 + (n - 1)F_{ST}}$ [21], where p_a is the frequency of allele a. The allele frequencies we used are those found in the GlobalFiler™ Amplification Kit Manual following the method described in [22].

Because the set $\Gamma^{\mathrm{n}} \times \{0{,}1\}^{2n}$ is too large to be enumerated in full for larger values of n, NOCIt estimates $\mathrm{P}(E_l|\ \Theta = \theta,\ \Lambda = \lambda,\ N = n,\ H_d)$ using importance sampling. Specifically, rather than summing over all $(g,\ d) \in \Gamma^{\mathrm{n}} \times \{0{,}1\}^{2n}$, NOCIt generates $\lceil Z_n/\mathrm{Card}(\Delta^{n-1} \times \Omega^{n-1})\rceil$, where $\mathrm{Card}(\Delta^{n-1} \times \Omega^{n-1})$ is the cardinality of $\Delta^{n-1} \times \Omega^{n-1}$ after discretization, random samples of $\mathrm{g} \in \Gamma^{\mathrm{n}}$ and $d \in \{0{,}1\}^{2n}$. The random samples of g and d come from sampling distributions selected to achieve efficient importance sampling, which, for g, is a distribution whose probabilities are proportional to the allele heights in the electropherogram E_l and, for d, is a uniform distribution over its possibilities. The parameter Z_n, called the batch size, is calculated as $Z_n = \min(z_1 \times m^n,\ z_{\max})$ where $(z_1,\ m,\ z_{\max})$ are the user-defined parameters *Sample Batch Size*, *Multiplicative Factor* and *Maximum Samples in a Batch*, respectively. Thus, by generating $\lceil Z_n/\mathrm{Card}(\Delta^{n-1} \times \Omega^{n-1})\rceil$ samples of g and d, the number of samples grows exponentially with the number of contributors n, and the total number of samples used for a given n is kept constant regardless of the size of Δ^{n-1} and Ω^{n-1} (i.e. regardless of how finely the contribution and relatively degradation levels of the contributors are discretized).

Finally, NOCIt computes an estimate of

$$\mathrm{P}(N = n|E,\ H_d) = \frac{\mathrm{P}(E|N = n,\ H_d)}{\sum_{n\in\{0,\ldots,\ n_{\max}\}}\mathrm{P}(E|N = n,\ H_d)}.$$

In addition, NOCIt computes the standard error s_n of the estimate of $\mathrm{P}(N = n|E,\ H_d)$. An ad-hoc heuristic adds batches of $\lceil Z_n/\mathrm{Card}(\Delta^{n-1} \times \Omega^{n-1})\rceil$ samples to specific estimates of $\mathrm{P}(E_l|\ \Theta = \theta,\ \Lambda = \lambda,\ N = n,\ H_d)$ until the standard error $s_n \le \mathrm{s}_{\max}\ \forall\ n$ where $s_{\max}$ is a user-defined precision criterion (with default value set to 0.05). Alternatively, the user can define a computation time limit: in that case, sample batches are added until either the time or precision criteria is first met.

The computation performed by NOCIt bears significant similarity to that performed by CEESIt [15], which computes the conditional likelihood ratio $\mathrm{LR}(E|N = n) = \Pr(E|N = n,\ H_p)/\Pr(E|N = n,\ H_d)$. Both methods use the same signal model, i.e., both methods calculate $\mathrm{P}(E_l|G = g,\ D = d,\ \Theta = \theta,\ \Lambda = \lambda,\ N = n,\ H_d)$ in the same way. Where the two methods differ is in the way in which computation is allocated to calculate accurate approximations of their respective key statistics, as exact computation is prohibitively expensive in terms of computational running time, and sampling methods are employed to obtain those approximations. For example, in NOCIt, we compute $P(E|H_d,\ N = n)$ for all $n = 0{,}1,\ \ldots, n_{\max}$, but a crude estimate of $P(E|H_d,\ N = n_{max})$ often suffices, since $P(E|H_d,\ N = n)$ for some $n < n_{\max}$ is often orders of magnitude larger. Since the ultimate goal of NOCIt is to compute the APP, large uncertainty in $P(E|H_d,\ N = n_{max})$ is insignificant if it is dwarfed by $P(E|H_d,\ N = n)$ for some $n < n_{\max}$. In contrast, in CEESIt, if the given assumption for n is large, there is no option other than to utilize many samples to obtain a good approximation of $P(E|H_d,\ N = n)$ to compute $\mathrm{LR}(E|N = n)$. Thus, provided that NOCIt and CEESIt each calculate accurate approximations of the APP and the conditional likelihood ratios, respectively, the two methods are compatible in the sense that they can be separately used to calculate the terms of the end-to-end LR described in Eq. (1).

3. Materials and methods

3.1. Software and GUI verification

GUI and software verification was conducted in accordance with the General Principles of Software Validation, Version 2.0 by the Center for Devices and Radiological Health [23]. As suggested in [24], we utilized two software teams: a software development team and a test team. The test team performed all software testing while the development team implemented modifications to the code, if necessary. First, the test team categorized NOCIt as software that is both *critical* and *complex*: critical because of its ability to substantially influence forensic DNA interpretation, statistical conclusions and the accuracy of the results, and complex because it contains many lines of code, complex algorithms and interconnected modules; thus, NOCIt software testing included: 1) functional; 2) reliability; and 3) regression testing. Functional tests are those engineered to verify that each function of the software application operates as expected and are further subdivided into positive, negative, boundary and fuzz tests. Positive tests confirm that the natural inputs yield the expected outputs; negative tests confirm that incorrect inputs yield the expected outputs; boundary tests verify that the software renders expected outcomes when inputs are at the limits; and fuzz testing checks that expected outputs are obtained when nonsensical inputs are used. Reliability tests are designed to test the software in the laboratory environment and confirm the site and resources can handle the application's need, while regression testing confirms that an already verified function is not modified or terminated because of seemingly unrelated changes to the software.

A total of 273 software tests were completed on six 64-bit PC or MAC computers using Java Version 8. For each test case, we randomly selected samples from the PROVEDIt database [8]. Therefore, one test case may have utilized samples generated using one kit, CE platform, injection time, etc., while another test case utilized samples generated using different laboratory conditions. In total, the software functions and modules were tested using three kits (i.e., GlobalFiler®, Identifiler® Plus, and PowerPlex® 16HS), three different cycle numbers (i.e., 28, 29, and 32 cycles), four injection times (i.e., 5, 10, 15 or 25 s) on two different capillary electrophoresis platforms (3130 and 3500 Genetic Analyzers), nine computers, four distinct operating systems, and four testers at two sites using samples containing anywhere from 1- to 5-contributors [8].

If the software output or behavior satisfied pre-recorded expectations, the software functionality was categorized as "By Design," and the test was passed. If NOCIt failed to meet the acceptance criteria for a test case, remediation was required and the outcome of the software test was categorized as either a "Minor," "Major" or "Critical" failure. The level depended on the functionality tested and how disparate the outcome was compared to the pre-defined requirement. All discrepancies between expectation and output were uploaded into an

Table 1
Summary of the 815 GlobalFiler™ sample set used to validate NOCIt.

NOC	1	2	3	4	5
Number	100	193	170	186	166
Total Template Mass (ng)	0.5- 0.008	0.75- 0.03	0.75- 0.045	0.75- 0.06	0.75- 0.075
Contributor Ratio	N/A	1:1 – 1:9	1:1:1 – 1:9:9	1:1:1:1 – 1:9:9:1	1:1:1:1:1 – 1:9:9:9:1

internal database accessible to both software teams. Once appropriate software modifications were made, a new distribution of NOCIt was released for developmental testing.

3.2. NOCIt performance and validation

We calibrated NOCIt with all but 100 of the 1-Person PROVEDIt samples amplified with the GlobalFiler™ Kit and injected on the 3500 Genetic Analyzer for 25 s (i.e., 2611 single source samples). Though the entire PROVEDIt 25 s GlobalFiler™ dataset was used, smaller sets of calibration data containing 50 serially diluted samples have been shown to adequately calibrate NOCIt models [12].

The 100 single-source electropherograms excluded from the calibration data were used as the 1-person test samples. In addition, all 666, 2- to 5- person GlobalFiler™/25 s PROVEDIt mixtures were used for performance evaluation, as were 49 differentially degraded mixture samples. Table 1 summarizes the samples used to test NOCIt's performance, and the full list of mixture samples and their NOCIt results is available in Supplement 1. All pertinent information related to a sample is contained in its name as detailed in [8]. Briefly, the DNA target masses of the test samples ranged from 0.75 to 0.0078 (designated as 0.008 ng in this work) ng, and the mixture ratios ranged from equal parts from all contributors to a 1:9 ratio between any two contributors. Samples were analyzed with GeneMapper® *ID-X* at 1 RFU. Spikes and dissociated dye artifacts were manually removed during analysis, while minus A, pull-up and raised baseline (or complex pull-up) were removed using the CleanIt module available in the software; details regarding artifact removal are available in [8]. The data were imported into NOCIt as a CSV.

Within the context of this study, the nominal NOC is taken to be the TrueNOC as it is reasonable to assume that each contributor's DNA is represented in the signal since it has previously been demonstrated that for this dataset: 1) the RFU signal from a single amplifiable molecule of DNA is fully resolved from noise at these laboratory and AT conditions [20]; 2) the smallest minor contributor of any mixture sample did not fall below 0.016 ng, which is approximately two copies of DNA (NB: some single-source samples were amplified at 0.008 ng); and 3) all 25 s GlobalFiler™ single-source PROVEDIt samples amplified at 0.016 ng or higher rendered RFU signal at a minimum of 12 known allele locations, regardless of the laboratory or sample condition [8]. Though it is likely that some signal at these locations may be attributed to noise, it is unlikely that all 12 can. As such, RFU signal is expected to be present even for the most degraded, lowest quantity and quality minor contributor for the samples used in this study.

NOCIt's functionality was assessed by evaluating its performance pursuant to the recommendations set forth by the Scientific Working Group on DNA Analysis Methods (SWGDAM) in their Guidelines for the Validation of Probabilistic Genotyping Systems [25]. Although that document does not specifically address probabilistic systems that compute the APP of the NOC, we use it as the foundation for performance assessment.

3.2.1. Precision

Precision was evaluated for three NOCIt run conditions (Table 2).

Table 2
NOCIt settings for *Condition 1, 2* and *3* on a 12-core system with 3.4 GHz of processor speed.

NOCIt Parameters		Parameter Value	
	Condition 1	*Condition 2*	*Condition 3*
Discretization Level	12	8	12
Standard Error Tolerance	0.05	0.05	0.05
Refinement Time Limit (s)	14400	14400	14400
Sample Batch Size	4000	4000	1000
Multiplicative Factor	3.0	3.0	2.0
Maximum Samples in a Batch	175,000	175,000	175,000
Average Run Time per sample(min)	37	20	19

The NOCIt settings that are likely to impact precision or accuracy are the *Discretization Level,* batch size (which is the *Sample Batch Size* multiplied by the *Multiplicative Factor* as described in Section 2) and the *Maximum Samples in a Batch.* We ran the 815 test samples three times for each of the NOCIt conditions and recorded the *n* at which we obtained the Maximum A Posteriori Probability (MAP) for repeated NOCIt run 1 (R1). The APP obtained from run 2 (R2) and 3 (R3) at that *n* were also recorded, and the largest absolute difference in probabilities across runs at that *n* was termed the APP Range. For example, if $\text{APP}(5)_{\text{R1}} = 0.896$, $\text{APP}(5)_{\text{R2}} = 0.928$, and $\text{APP}(5)_{\text{R3}} = 0.893$, then APP Range = max(0.032, 0.003, 0.035) = 0.035. However, if $\text{APP}(5)_{\text{R1}} = 0.892$, $\text{APP}(5)_{\text{R2}} = 0.002$, and $\text{APP}(5)_{\text{R3}} = 0.925$, then APP Range = 0.923. Run conditions resulting in many samples exhibiting small APP Ranges are preferred over larger ranges, since it represents higher run-to-run precision. The F_{ST} value was set to 0.01 for all runs.

3.2.2. Correctly including the TrueNOC into the interpretation pipeline

Recall that the APP distribution may be used in a variety of ways during interpretation. Whether it is used to eliminate unlikely values of *n* from consideration or as a nuisance variable, reporting system performance is a necessity. One way to test performance is to assess the proportion of times the TrueNOC was deemed probable. Using the results of the first NOCIt run, we counted the number of samples, out of 815, rendering $P(TrueNOC|E) \geq \alpha$. Because α is essentially an arbitrary threshold, we determine these proportions for α values of 0.001, 0.01, 0.1, 0.2 and 0.5. Note that if the APP for an *n* is greater than 0.5, that *n* is necessarily NOCIt's MAP estimate. We emphasize that these thresholds do not represent recommendations; rather, they are used to allow for comparisons between NOC inference systems, different conditions, and to provide an overall assessment of model performance.

3.2.3. Comparison to alternative methods

NOCIt's performance was compared to MAC and to Haned et al.'s MLE method [18], whose method improves on the MAC approach by accounting for population frequencies, but does not explicitly account for drop-in or drop-out of alleles or degradation effects, instead taking as input the alleles deemed present at each locus. Because NOCIt assumes a uniform prior on the NOC (cf. Section 2), a MAP estimate derived from NOCIt's APP is also a maximum likelihood estimate, albeit one based on different data or "evidence" (i.e., the full sequence of electropherogram peak heights, as opposed to the alleles present at each locus) and on models of the evidence distinct from that used in Haned et al.'s MLE method. Nevertheless, for the purposes of this work, whenever we refer to the MLE method, we refer specifically to Haned et al's MLE method and, in particular, not to NOCIt's MAP estimate.

To compare performance between the MAC, Haned et al.'s MLE and

NOCIt's APP methods, we determined the proportion of samples where NOCIt's APP exceeded α and plotted these alongside the number of times the MAC or MLE method resulted in the correct n. For the MAC method, the minimum NOC was calculated by applying an AT of 100 RFU, and filtering potential stutter peaks that met the following two conditions: 1) the peak-in-question was one STR repeat shorter than a higher RFU peak to the right of it; and 2) the RFU intensity of said peak was ≤ the manufacturer's stutter filter thresholds [26]. For MLE's n, we used the same filtered data as was used for the MAC method as well as the forensim package available at http://forensim.r-forge.r-project.org/ [18,27]. We chose to compare the APP to MAC since filtering signal and counting alleles is often considered an attractive option, particularly if one seeks to provisionally inform their beliefs. Though MLE still requires the application of an AT and stutter filters, it estimates the NOC based on allele frequencies and was, therefore, taken to represent the category of software that do not deal with quantitative profile data, but are more sophisticated than counting methods. We note that exhaustive comparisons between MAC, MLE and NOCIt to other existing or future methods is afforded by the use of the publicly available PROVEDIt data [8] and by the data available in Supplement 1.

3.2.4. Robustness of NOCIt across sample qualities

As per PCAST recommendations [28], we report whether NOCIt was robust across the range of sample types typically encountered in casework. To do so, we used multiple logistic regression and determined the probability that $APP(TrueNOC) \geq 0.001$ and the probability that $APP(TrueNOC) \geq 0.5$ for: μ, the degree of electropherogram sloping (continuous); the mass of the minor contributor supplying the smallest quantity of DNA to the mixture [ng] (nominal); and the TrueNOC (nominal). Note that we drop the '|E' in the notation for ease of exposition. Multiple logistic regression was performed using JMP® Pro 14.2.0. If NOCIt's APP is robust, $\mathrm{P}(APP(TrueNOC) \geq \alpha)$ should remain relatively stable across large changes in all three of μ, the mass of the minor and the TrueNOC.

The degree of sloping, μ, was determined for each sample using the contour of the STR signal, which we modeled as exponentially decaying in fluorescence with molecular weight as per,

$$F_l = \varphi e^{\mu \bar{w}_l}$$

Where F_l is the sum of the peak heights associated with the known genotypes at locus l, $\bar{w}_l$ is the average base pair size of the known STR alleles at locus l, and φ and μ are the parameters obtained for each sample using least squares regression; thus, for the 815 samples used in this study, we obtain 815 sloping parameters μ. In extreme cases of decay, the highest molecular weight peaks may not reach detectable levels. If high molecular weight markers exhibit low peak heights due to degradation or inhibition of the PCR reaction, μ takes a large negative value. In contrast, if there is good signal balance across all loci, indicating efficient PCR and high quality template DNA, μ will be near zero.

The mass of the minor contributor was grouped into bins of sizes: $0.008-0.016$; $0.017-0.035$; $0.036-0.066$; and $0.067-0.5$ ng. The first bin, therefore, represents the amplification of ca. 1–2 copies of DNA, the second represents ca. 3–4 copies, while the third and fourth bins represent 5–10 and > 10 copies of the minor's DNA, respectively. Given that the limit of detection for our post-PCR method was 1 copy (i.e., all intact alleles that survived the pre-PCR steps are detected), allele drop-out is solely the result of the ability to sample intact alleles which can be described by binomial probabilities with binomial parameters T, the total copy number of allele and $\frac{V_{aliquot}}{V_{tot}}$, representing the volume fraction of extract transferred to the PCR container [29]. For the 815 samples used in this study, $V_{aliquot}$ ranged between 1 and 10 μL, V_{tot} = 48 μL, and the probability of drop-out is $\mathrm{P}(Binomial(T, \frac{V_{aliquot}}{V_{tot}}) = 0)$. In the absence of degradation, when the smallest minor contributor is 0.008 ng, T ranged from 60 $\left[\frac{0.008ng/_{1\mu L} \bullet 48\mu L}{0.0063^{ng}/_{cell}}\right]$ to 6 $\left[\frac{0.008ng/_{10\mu L} \bullet 48\mu L}{0.0063^{ng}/_{cell}}\right]$ copies. When the smallest contributor contributes 0.016 ng to the mixture, the value of T ranged from 121 to 12 copies. Thus, the dropout rate for the smallest minor in the first bin is expected to range between 28 and 6%. The second bin contains samples with minor contributors that have expected drop-out rates between 0.3 and 6%, whereas the remaining bins all have drop-out rates less than 0.3 % [20] when neither the DNA is degraded or PCR is inhibited. Thus, the bin ranges represent minor contributor quantities exhibiting considerable, some, low and extremely low allelic drop-out due to sampling.

3.2.5. Performance at the signal boundary

To test performance at the signal boundary, we evaluated the APPs obtained when an amplification negative, containing no known fragments of DNA or obvious spurious signal, was run without an AT (i.e., AT = 1 RFU) and with an AT of 150 RFU.

3.3. Integrating the APP into the LR calculation

To illustrate and further explore the impact of reporting the full LR, we ran four of the 815 GlobalFiler samples with CEESIt and combined NOCIt's APP and CEESIt's LRs as per Eq. (1). We compared the end-to-end LR and the LRs acquired when using a single MAC-based n assignment. Specifically, we evaluated LRs from the following samples:

1 A single-source sample (RD14 − 0003-15d2U60 − 0.25GF-Q4.5_01.25 s) where NOCIt's APP indicated that n = 2 was highly probable and the MAC assignment incorrectly assigned a value of 2;
2 A single-source sample (RD14 − 0003-17d2U60 − 0.25GF-Q13.3_01.25 s) where both the APP and the MAC assignment correctly indicated the NOC was likely one;
3 A two-person mixture (RD14 − 0003-39_40 − 1S2;2a-0.5GF-Q1.1_0.1.25 s) where NOCIt's APP suggested that n = 2 was highly probable, but MAC incorrectly assigned a value of 3; and
4 A four-person mixture (RD14-0003–40_41_42_43 − 1;1;1;1-M4d-0.06GF-Q1.7_01.25 s) where NOCIt's APP suggested that n = 3 and 4 were highly probable, but the MAC-based assignment was underestimated by one.

4. Results

4.1. Software and GUI verification

Over the course of development, the development team provided the test team with 11 software distributions. At the end of testing each distribution, the test team reported software failures or inadequacies to the development team who would modify the software, based on those reports, to meet expectations. The subsequent distribution was transferred to the test team for further testing and the reporting steps were repeated. Fig. 1A illustrates that over the course of GUI development, there was an increase in the proportion of test cases that passed, demonstrating that good software development practices, which included a standardized method of software validation and verification, resulted in marked improvements to NOCIt's GUI and functionality. The introduction of new modules in Distributions 7 and 9, which were the result of modifications to the user requirements, resulted in a decrease in pass rates, demonstrating the importance of rigorous software and GUI testing when introducing significant changes to software functionality.

Fig. 1B depicts the percentage of each test type used to verify the GUI was performing as expected. Most software tests were of the positive variety, though regression testing became prevalent at the end stages. The final distribution was verified solely by regression tests. In addition, different users at two sites using multiple processors/computers and operating systems confirmed GUI performance. We note that

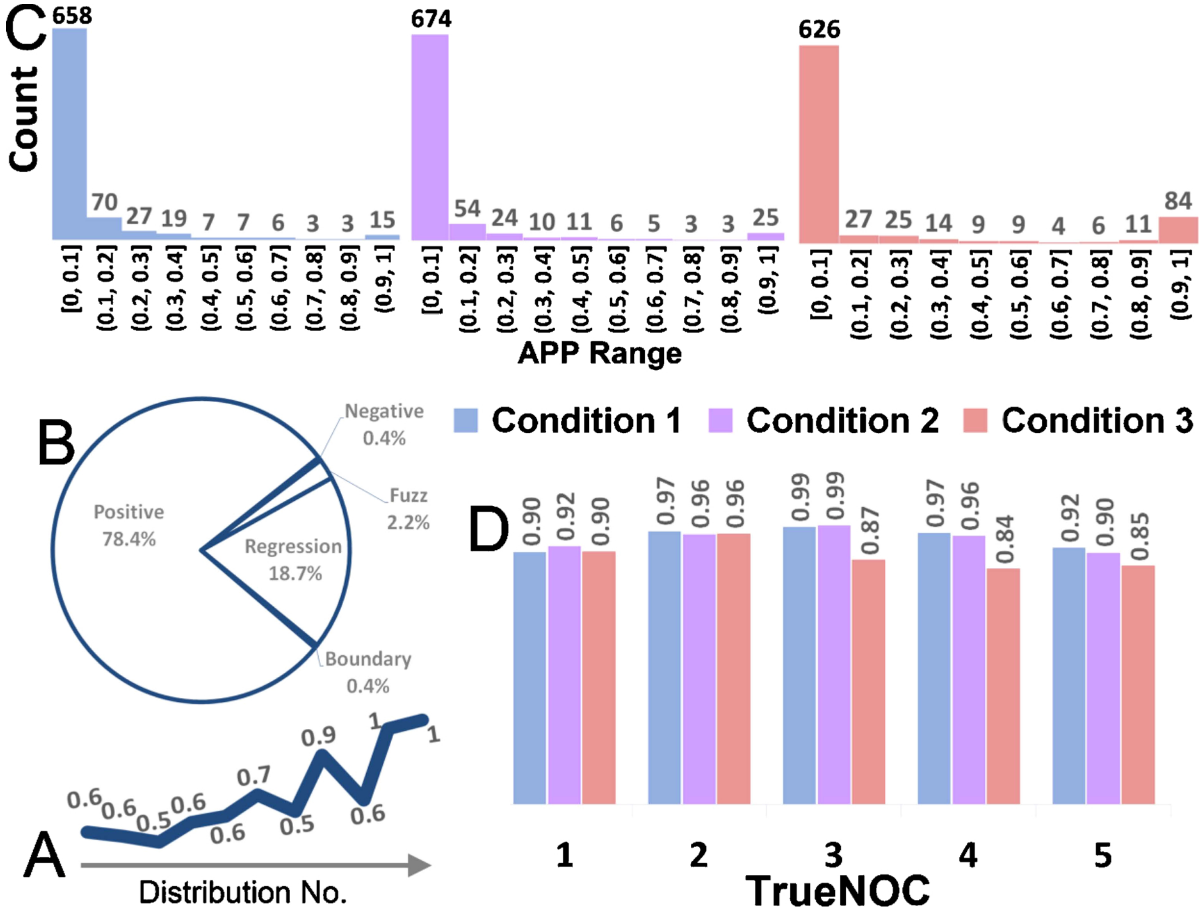

Fig. 1. (A) Proportion of GUI test cases that passed across 11 NOCIt distributions. (B) Percentage of positive, negative, fuzz, boundary and regression software tests across 11 NOCIt distributions. A distribution represents an internal release of NOCIt to the software test team from the development team. (C) Histograms of the APP Range, (D) Proportion of instances where NOCIt determined an APP ≥ 0.001 for the TrueNOC using (■) Condition 1, (■) Condition 2, and (■) Condition 3 for run 1 (R1).

although we obtained a test pass rate of 100 % with the final distribution, the failure of any software to meet pre-determined expectations for a given test does not necessarily invalidate it. Many minor failures or a few critical failures, however, may lead the user or team to conclude that the software is not fit for its intended use. Ultimately, the decision to implement new software is dependent upon pre-defined acceptance criteria and a review of the entirety of the results.

4.2. NOCIt performance and validation

Once GUI verification was completed the final distribution was used to test the biological models. Model validation was completed with 815 samples consisting of degraded, differentially degraded, inhibited, low-template and high-template samples containing anywhere from 1- to 5-contributors in any proportion. The mixtures were generated under controlled laboratory conditions and so the actual number of contributors, i.e., TrueNOC, for each sample is known. These 815 samples were specifically selected to test the biological models because they represent mixtures that: 1) contain information from a kit containing the extended CODIS loci; 2) consist of 1- to 5- contributors, some of which were subjected to conditions that compromise the integrity of the PCR; and 3) were run using an optimized forensic DNA pipeline engineered to produce allele signal that is well resolved from noise when in the single-copy regime [20].

NOCIt outputs the APP distribution for $n=$ 0 to 6 contributors, providing the user with a means by which to assess the probability that one, two, three or more contributors comprise the evidence. For example, a sample rendering APPs of 0.8 and 0.2 for $n=$ 2 and 3, respectively suggests the mixture may be well explained by either number, though $n=$ 2 is more likely given the evidence. Traditionally, the value 2 or 3 would be assigned to the mixture, and the LR would be calculated using those assignments. As suggested by the authors of [10], however, an alternative approach is to use the APP values in the determination of an overall LR, making the NOC assignment an unnecessary task. Since the APP calculation is pertinent to computing the overall LR or deciding upon probable values for n, we focus on assessing the performance of NOCIt to output reproducible, robust and accurate APP results.

4.2.1. Precision

As shown previously [30,31], some probabilistic genotyping approaches may not produce the same outcome from repeat analyses due to their Monte Carlo-based algorithms. Therefore, where applicable, validation studies should report the range of values from multiple analyses of the same data, which is then used as the basis for establishing acceptable or expected degrees of variation in output [25]. As a result, parameter settings that can reduce variability were evaluated.

Increasing the *Sample Batch Size* or the *Discretization Level*, as described in Section 2, associated with the Monte Carlo settings may influence the variation in the results. To evaluate effects of these parameters on repeatability, we ran all 815 test samples using three conditions (Table 2). For each sample, every condition was run in triplicate.

Each condition employed the same values for *Refinement Time Limit, and Maximum Samples in a Batch*. The *Refinement Time Limit* is a user input that pertains to the time (in minutes) that NOCIt is allowed to run.

Once this time limit is reached, NOCIt automatically stops and the user can restart it, if desired. Since NOCIt uses a Monte Carlo-based approach, this feature was added in cases where the user deems it advantageous to re-start a run rather than wait for an extensive period for convergence. Similarly, *Maximum Samples in a Batch* is the maximum number of samples NOCIt randomly selects during Monte Carlo sampling. Preliminary results demonstrated little change in the outcomes when these parameters were modified; therefore, these settings were not modified for this work. The parameters in *Condition 1* were chosen because they represent relatively stringent settings, resulting in the longest average run time of 37 min per sample. *Condition 2* employed a lower *Discretization Level* than *Condition 1* but the same *Batch Size*. The third condition was equivalent to the first, except the *Batch Sizes* decreased, significantly decreasing the average run time per sample between *Conditions 1* and *3*.

Fig. 1C are histograms of the APP Ranges across three NOCIt settings. Differences between the histograms of *Condition 1* and *2* versus *Condition 3* are evident. For example, we observe that the last condition, which employed smaller batch sizes, resulted in more samples with APP Ranges exceeding 0.5. Specifically, the less stringent *Condition 3* resulted in 14.0 % of samples exhibiting APP Ranges in excess of 0.5, while *Conditions 1* and *2* resulted in 4.2% and 5.2% of the samples with this outcome, respectively. Focusing on the APP Range (0.9,1], signifying that the most likely *n* for one run was nearly improbable for at least one of the other repeated runs, we observe that *Condition 3* has a substantial number of samples in this range, indicating that the *Batch Size* influences repeatability more than *Discretization Level*does. Fig. 1D, demonstrates that the proportion of samples for which $APP(TrueNOC) \geq 0.001$ for *Conditions 1* and *2* are not substantially different, though the higher *Discretization Level* seems to exhibit good accuracies across all TrueNOC. Accordingly, we utilized *Condition 1* as the standard NOCIt run parameters for all other tests, as it resulted in high accuracies, good repeatability and tractable run times.

4.2.2. Correctly including the TrueNOC into the interpretation pipeline

Fig. 2 are stacked plots presenting the APPs for all test samples obtained using *Condition 1* separated by the nominal number of contributors. Fig. 2A demonstrates that for most samples, NOCIt outputs an APP distribution suggesting only one or two values of *n* explain the evidence well and, typically, the *n* corresponding to the MAP was the TrueNOC. Even in the presence of significant allele-sharing, most 5-contributor samples resulted in only one or two probable values of *n*, and more than half of the samples resulted in a MAP at *n*= 5, demonstrating that even in the presence of abundant allele sharing and severe allele drop-out, a fully quantitative evaluation of the data without analytical thresholds or stutter filters produced APPs indicating the *n*= TrueNOC explained the data well. This is notable given that, based on the known genotypes, the 5-person samples do not have any locus with greater than eight alleles, and as many as 75 allelic peaks per profile (65 % of the profile) were not detected [8].

By further exploring Fig. 2A, we observed that the APP distribution was always unimodal, and it is possible that the mode was not necessarily equal to TrueNOC. That is, we did not observe any instance where the APP was high for small values of the NOC, *n*, low for medium *n* values and then increased again for large values of *n*. Fig. 2B and Table 3 summarize the number of samples that resulted in one, two or three *APP(n)* ≥ 0.001.

No sample had greater than three APPs exceeding 0.001.

4.2.3. Comparison to alternative methods

Following the testing principles described in the SWGDAM Guidelines [25], we compared the results of NOCIt to those acquired by MAC and the allele frequency based MLE method described in [18]. Prototype implementations of NOCIt were previously compared to allele counting and MLE methods [11,12] and showed promising results. To ensure that NOCIt continued to outperform common methodologies, we plot the proportion of samples that resulted in $\text{APP}(TrueNOC) \geq \alpha$ and the proportion of times MAC and MLE provided the correct estimate for the 815 samples used in this study.

Interestingly, we observed that the proportion of times TrueNOC was correctly included when using the MAC method is lower than might be expected (i.e., 66 %) for single-source samples. We note, however, that the 100 single source samples used as the test samples included many compromised, low-template samples. In particular, a great majority of the 1-person test samples were subjected to UV-damage or sonication and were at the extreme low template regime (Supplement 1). Closer inspection established that many of them exhibited excessive stutter products due to their low-template nature, which is not unexpected as previous work has demonstrated that amplifying samples in the single-copy regime results in nearly 8% of all alleles exhibiting stutter ratios > 15 % at the end-point of PCR [29]. Therefore, it is not surprising that low-template profiles containing 21 autosomal STR loci would regularly exhibit indications of an ‘additional contributor’ within the MAC paradigm. Since the MLE method also depends on the same noise and stutter filters as MAC, the proportion of samples supplying correct estimates did not much improve for the MLE method, though there were some improvements for the 4- and 5- person mixture samples when compared to MAC estimations. Also plotted in Fig. 3A are the proportion of samples where APP*(TrueNOC)* was at least 0.001, 0.01, 0.1, 0.2 and 0.5. Note that an APP threshold of 0.5 means that TrueNOC was the MAP estimate (i.e., most probable *n*). Fig. 3A demonstrates that in most cases, NOCIt's MAP estimate was equal to TrueNOC. It also demonstrates, however, that rather than choosing a single *n* for downstream interpretation, increasing the range of *n* is warranted. For example, at $\alpha = 0.001$, correctly including TrueNOC into the interpretation pipeline did not fall below 90 % regardless of the TrueNOC that comprised the sample.

In Fig. 3B is plotted MAC's NOC Assignment versus the APP *(TrueNOC)*. Notably, in many cases, NOCIt's MAP correctly indicated the TrueNOC when MAC did not, while the reverse (i.e., MAC correctly assigned the TrueNOC, while NOCIt's MAP did not) was rarely observed. Of note was the relatively large set of three-person samples where MAC and MLE estimates did not equal the MAP estimate. Closer inspection, however, of Fig. 3B for TrueNOC= 3 shows that many of the APP(3) ranged between 0.1 and 0.5, suggesting NOCIt still assigned a significant probability to *n*= 3. The MLE results were similar to those of the MAC method.

As suggested by Fig. 3, the models of NOCIt seem less sensitive to elevated stutter effects than methods that rely on signal filtering because they determine the relationship between the stutter peak heights and the parent peak height [13,14]. The poor response of MAC and MLE, even with seemingly simple samples, demonstrates that the NOC estimations are unlikely to represent the TrueNOC if assigned by binary or counting methods alone.

4.2.4. Robustness of NOCIt across sample qualities

To explore the impact of sample quality on the APP, we evaluated effects of electropherogram sloping, TrueNOC and the mass of the smallest contributor on the ability of NOCIt to output an APP greater than or equal to 0.001 or 0.5 for *n*= TrueNOC. That is, we evaluated the degree to which degradation, sample complexity and template mass of the smallest contributor impacted NOCIt's ability to determine whether the TrueNOC would be included in downstream interpretation by performing multiple logistic regression where the dependent variable is assigned a status of one if the APP*(TrueNOC)* > *α* and 0 otherwise. The resultant AIC and BIC values were 227 and 268, respectively, for α= 0.001 while the AIC and BIC values for α= 0.5 were 658 and 699, respectively. These represented the lowest AIC and BIC values obtained from several modeling options (data not shown). Full summaries of the logistic fit details for each *α* are provided in Supplement 2 and 3. The results are depicted in Fig. 4 where we plot the logistic curves for *α*= 0.001 and 0.5 against the degree of sloping separated by

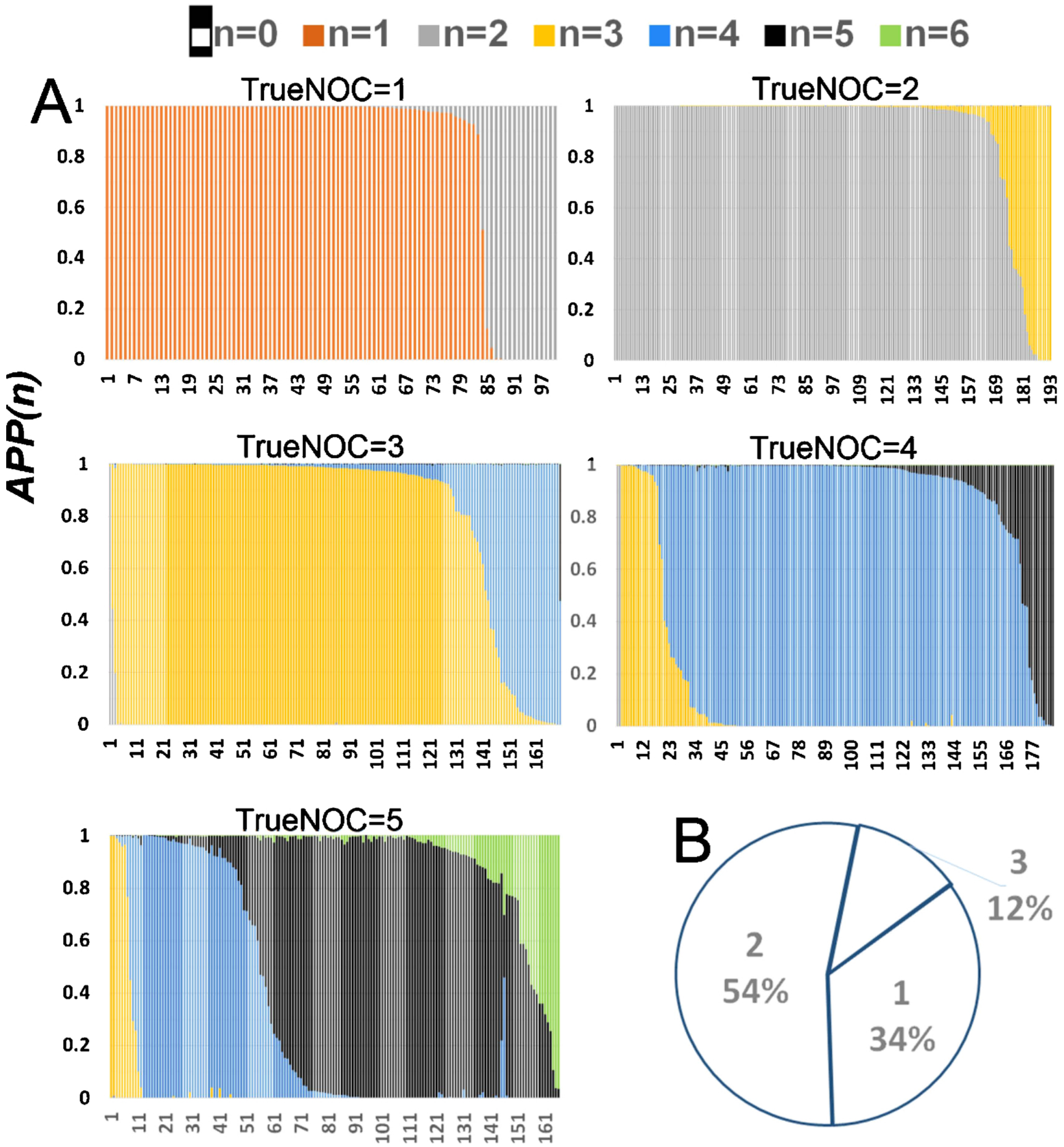

Fig. 2. (A) Stacked Plots of APP(*n*) obtained using *Condition* 1 NOCIt settings, where TrueNOC is the true number of contributors that comprise the mixture and the APP for *n* = 0 to 6 is depicted for each sample for *n*= (white bar)0; (■)1; (■)2; (■)3; (■)4; and (■) 5; and (■) 6. (B) Pie Chart depicting the percentage of samples resulting in one, two or three APP*(n)* ≥ 0.001. No sample exhibited greater than three APPs exceeding 0.001.

Table 3
The number of samples rendering one, two or three APP(*n*) ≥ 0.001.

	TrueNOC				
Number of APP(*n*) ≥ *0.001*	1	2	3	4	5
1	67	126	43	36	9
2	33	67	115	127	96
3	0	0	12	22	61

the TrueNOC and the mass of the smallest contributor.

Fig. 4 demonstrates that samples used in this work represent the wide variety of complexity expected in forensic samples with μ values as low as -0.025 and minor contributor's values as low as 0.016 ng (single-source masses went to 0.008 ng). Overall, we see that the chance of including the TrueNOC in downstream interpretation is, in general, higher for less stringent APP thresholds demonstrating the value of considering multiple NOCs during downstream interpretation. Although a less stringent APP threshold necessarily improves the chance of correctly incorporating TrueNOC in downstream LR interpretation, it also increases the range of *n.* In common parlance, this means that if the APP threshold were lowered there would be a relatively good chance the expert would be required to consider more than one *n* assignment to the possible exclusion of only improbable *n*s. The degree of electropherogram sloping, which is indicative of DNA damage or PCR inhibition, engendered a significant impact for α= 0.001, wherein the probabilities of correctly including TrueNOC into downstream interpretation decreased as the severity of electropherogram sloping increased. Interestingly, the greatest impact of μ was seen in the α= 0.001 case, since the chance of correctly including TrueNOC within the probable range of *n* is substantially larger at low degradation levels than for α= 0.5. Moreover, Fig. 4 reveals that P(APP*(TrueNOC)* > 0.001) for non-degraded samples is relatively constant until the mass of the minor contributor reaches ca. 2 cells' worth of DNA. This was in contrast to the α= 0.5 case where the probability of correctly incorporating the TrueNOC in downstream interpretation decreased as

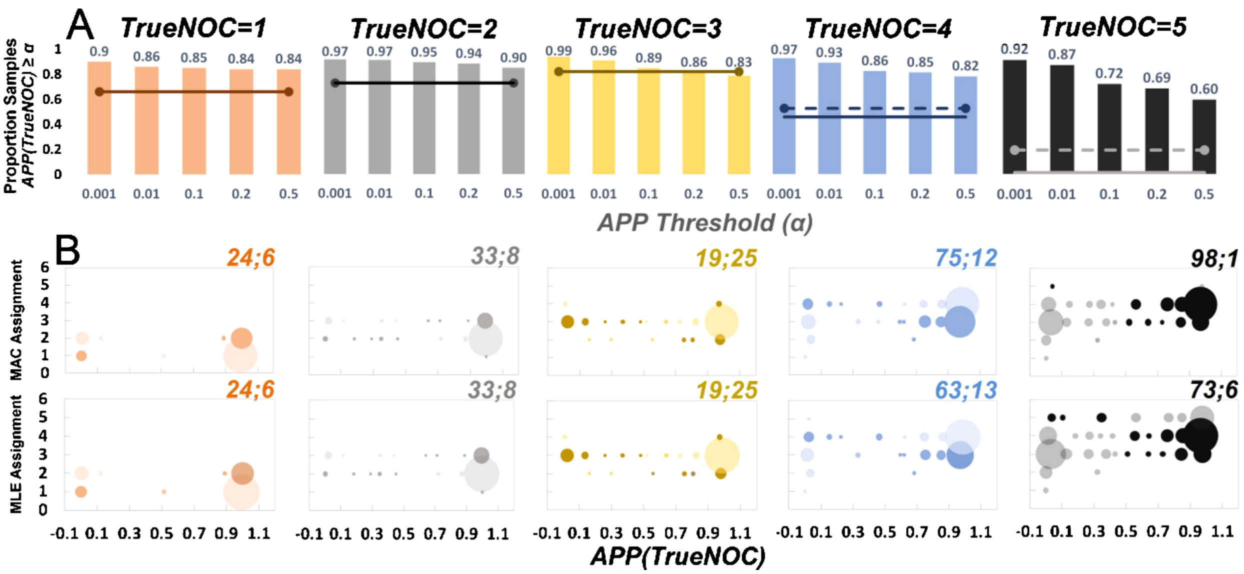

Fig. 3. (A) The proportion of samples resulting in APP*(TrueNOC)* ≥ *α*, where *α* was assigned values of 0.001, 0.01, 0.1, 0.2 and 0.5, separated by TrueNOC (■) 1, (■) 2, (■) 3, (■) 4, and (■) 5. Also included are the proportion of samples where (—) MAC and (- -) MLE estimates equaled the TrueNOC. Since these methods provide a single *n* value rather than a probability distribution, there is no *α*-value; therefore, the proportion is unchanging across the bar graphs. (B) Bubble charts, separated by TrueNOC, of the MLE or MAC estimates plotted against NOCIt's APP*(TrueNOC)*. The larger the diameter of the disk, the larger the proportion of samples represented. The darker shaded disks highlight instances where the MAC/MLE estimates were inconsistent with NOCIt's MAP estimate. On the top right of the plots are the number of samples that rendered APP*(TrueNOC)* ≥ *0.5* and MAC or MLE estimates ≠ TrueNOC; and the number of samples that rendered APP*(TrueNOC)* ≤ *0.5* and MAC or MLE estimates = TrueNOC.

the mass decreased across the entire range. In general, the TrueNOC consistently influenced the probability that the actual contributor number will be within the range of *n* considered during downstream interpretation, though not in a unidirectional manner. For example, Fig. 4 once again expresses the results already depicted in Figs. 2 and 3 which shows that the 1-person samples resulted in lower than anticipated probabilities of correctly incorporating the TrueNOC into downstream interpretation, indicating that DNA profile 'complexity' is present within all mass and NOC regimes.

4.2.5. *Performance at the signal boundary*

There are one of two ways a laboratory may choose to import

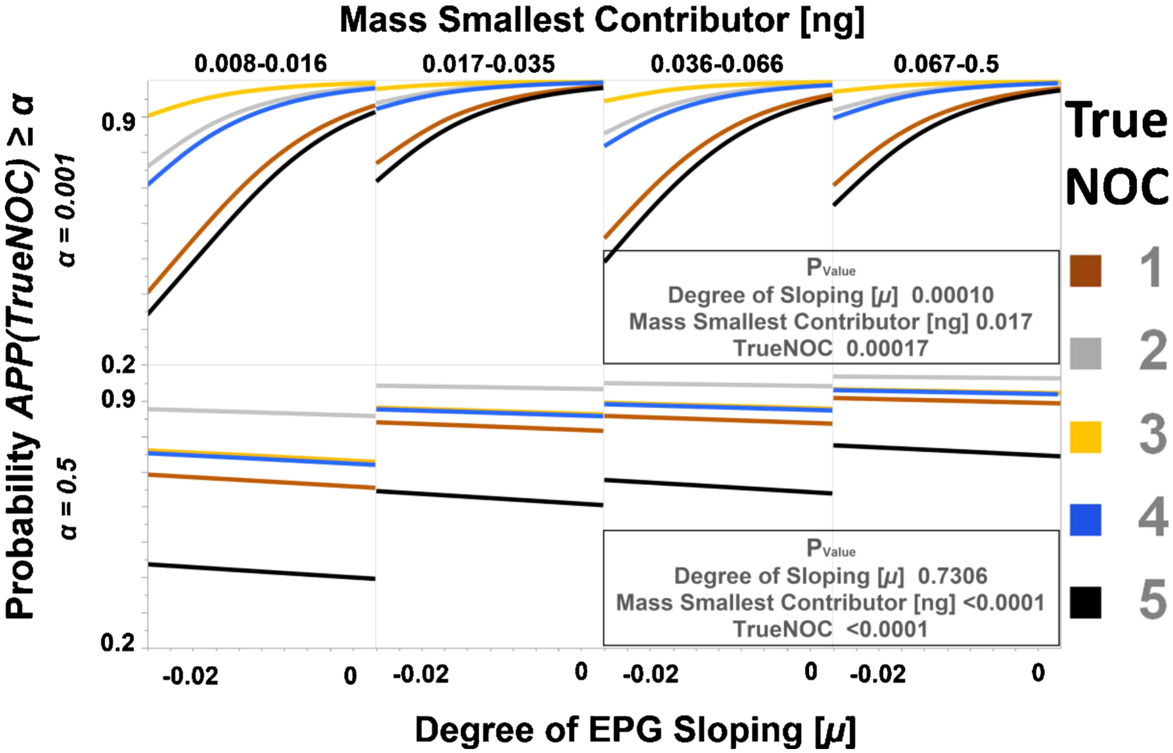

Fig. 4. Plotted are the probabilities that APP*(TrueNOC)* ≥ 0.001 and APP*(TrueNOC)* ≥ 0.5, determined with multiple logistic regression, against the degree of electropherogram sloping, μ, delineated by the mass of the contributor with the lowest target mass in the mixture [ng]. Also presented are the P_{Values} demonstrating the significance of each impact variable on correctly classifying TrueNOC as within the range of *n* for *α* = 0.001 and 0.5.

evidentiary or unknown data: The first option is to import all of the peak information (without AT applied), while the other is to import the peak information after the application of an AT. To examine the impact of an AT at the signal boundary, we analyzed a negative amplification control using both approaches. In the first approach, all of the data, including baseline, are imported into NOCIt (i.e., no AT is applied and all low-level, post-processed peaks are interpreted as part of the signal). We also applied an AT to the signal thereby importing no information. As expected, the results between these two scenarios differ: in the first case, the AAP was almost 1 at $n = 0$, and the next largest APP was 5.05×10^{-6} at $n = 1$ (see Supplement 4 for the full NOCIt report). In the second case, using the same negative sample with an AT of 150 RFU applied, the probabilities for all n were 1/7, which is equal to their prior probability (see Supplement 5 for the full NOCIt report). This example not only demonstrates that NOCIt works well at the signal boundaries, but that the application of a high-pass threshold designed to filter data is unnecessary when evaluating evidence using probabilistic systems and can, in some cases, provide the end-user with less information than is actually available. This suggests that further studies examining the impact of ATs on probabilistic pipelines are warranted.

4.3. Integrating the APP into the LR calculation

Fig. 6 provides an illustration of a sample progressing through the interpretation pipeline depicted in Fig. 5. We see in Fig. 6A that for this true 2-person example, the MAC method would suggest two persons explain the evidence while peaks near the baseline would indicate the presence of three. Consistent with that observation, Fig. 6B depicts the output acquired when the sample of Fig. 6A is analyzed by NOCIt and shows that the probable NOCs are 2 or 3 with APPs of 0.736 and 0.264, respectively.

The MAP of 0.736 suggests that 2 persons explain the mixture well, though an APP of 0.264 suggests that it is far from implausible that 3 persons have generated this mixture. Together, these APPs constitute the range of likely NOCs since the other n values had APPs of 0. Fig. 6C portrays the summary statistics obtained using CEESIt and four input conditions: (1) suspect 1 (s_1) and n = 2; (2) s_1 and n = 3; (3) suspect 2 (s_2) and n = 2; and (4) s_2 and n = 3, while Table 4 shows the result obtained for both suspects when Eq. (1) is used to compute the overall LR using NOCIt to obtain the $APP(n) = P(N = n|E, H_d)$ and CEESIt to approximate $\mathrm{LR}(E|N = n)$ [10] for each n. The question of how to assign the NOC [9], or whether to use different NOCs in the numerator and denominator [1,32–34] has been debated in the literature with some suggesting the LRs from all reasonable NOC assignments be determined and reported and others advocating maximizing the H_p and H_d likelihoods separately [34]. If one were to take the former approach, in this example the LR slightly decreases when the NOC assignment shifts from 2 to 3. In addition, the LR distribution under H_d shifts closer to 1 as the NOC assignment increases in number. Because protocols may vary between laboratories, one laboratory may compute the LR for n = 2 and 3 and report the lower LR, while another laboratory might only assign n = 2 (based on MAC) and report the higher LR value. Without standardization or consensus, reporting LRs at distinct n values is unlikely to aid in normalizing LR results across laboratories. Rather than computing and reporting LRs from different n separately, an alternative is to use the APPs from a system such as NOCIt, the LRs by a coincident system such as CEESIt and Eq. (1) [10] to compute an overall LR. Since NOCIt and CEESIt systems incorporate the same models for drop-out, stutter, peak heights and noise, combining the APP with $\mathrm{LR}(E|N = n)$ (Table 4) represents the full evaluation of evidentiary DNA signal without analytical thresholds applied.

Continuing, we compared the end-to-end LR to the MAC-based LR for the sample depicted in Fig. 6 and three additional samples and summarize the MAC, APP and LR results in Table 5. For the NOCIt results, we note any APP less than 1×10^{-6} was assigned a value of 0. For Sample 3, a true 2-person mixture, the LRs obtained using MAC and APP do not differ since the one 'extra' peak present in the profile that forced the MAC assignment to 3 was of a comparatively low peak height and in stutter position; thus, the LR was negligibly affected when an 'additional' contributor was added. Likewise, the LRs for Sample 2 did not significantly change since the MAC assignment and the APP result were consistent – i.e., both suggesting the NOC to this sample was one. Despite instances where the single LR and full LR agree, there were other instances when they did not. Notably, Sample 1 demonstrates that

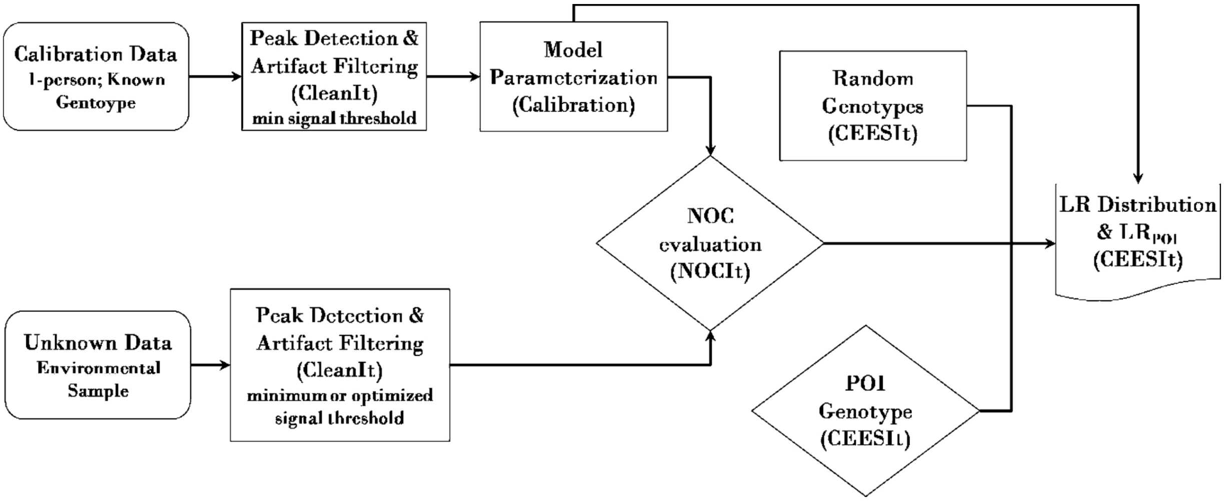
Calibration Data
1-person; Known Gentoype
Peak Detection & Artifact Filtering (CleanIt)
min signal threshold
Model Parameterization (Calibration)
Random Genotypes (CEESIt)
NOC evaluation (NOCIt)
LR Distribution & $\mathrm{LR}_{\mathrm{POI}}$ (CEESIt)
Unknown Data
Environmental Sample
Peak Detection & Artifact Filtering (CleanIt)
minimum or optimized signal threshold
POI Genotype (CEESIt)

Fig. 5. A schematic of the interpretation pipeline that includes APP*(n)*. All of the data are analyzed using a peak detection software of choice. The calibration data are garnered from single-source profiles of known genotype analyzed using the lowest possible signal threshold setting, and well-characterized artifacts, such as pull-up and minus A, are filtered with the CleanIt module and user-defined criteria. These calibration data are used to parameterize the models utilized by NOCIt. NOCIt determines the APP distribution on the NOC from data acquired from an unknown sample containing any number of contributors in any proportion. As with the calibration data, the STR data acquired from the environmental sample will undergo pre-processing steps wherein peak detection and artifact filtering are completed. Unlike the calibration data, however, an analytical threshold, may be applied to the unknown, if desired. The data, containing information on the peak height, peak position and allele call, are imported into NOCIt for evaluation. NOCIt outputs $P(N = n|E, H_d)$ for all n. Using the same models, CEESIt is then utilized to compute the LR for the person-of-interest (POI), likelihood ratio distributions for randomly generated genotypes, or other reporting statistics. For this pipeline, data from the calibration and unknown samples are expected to be acquired using the same DNA laboratory processing protocols (i.e., same STR assays, PCR cycle numbers, and electrophoresis settings).

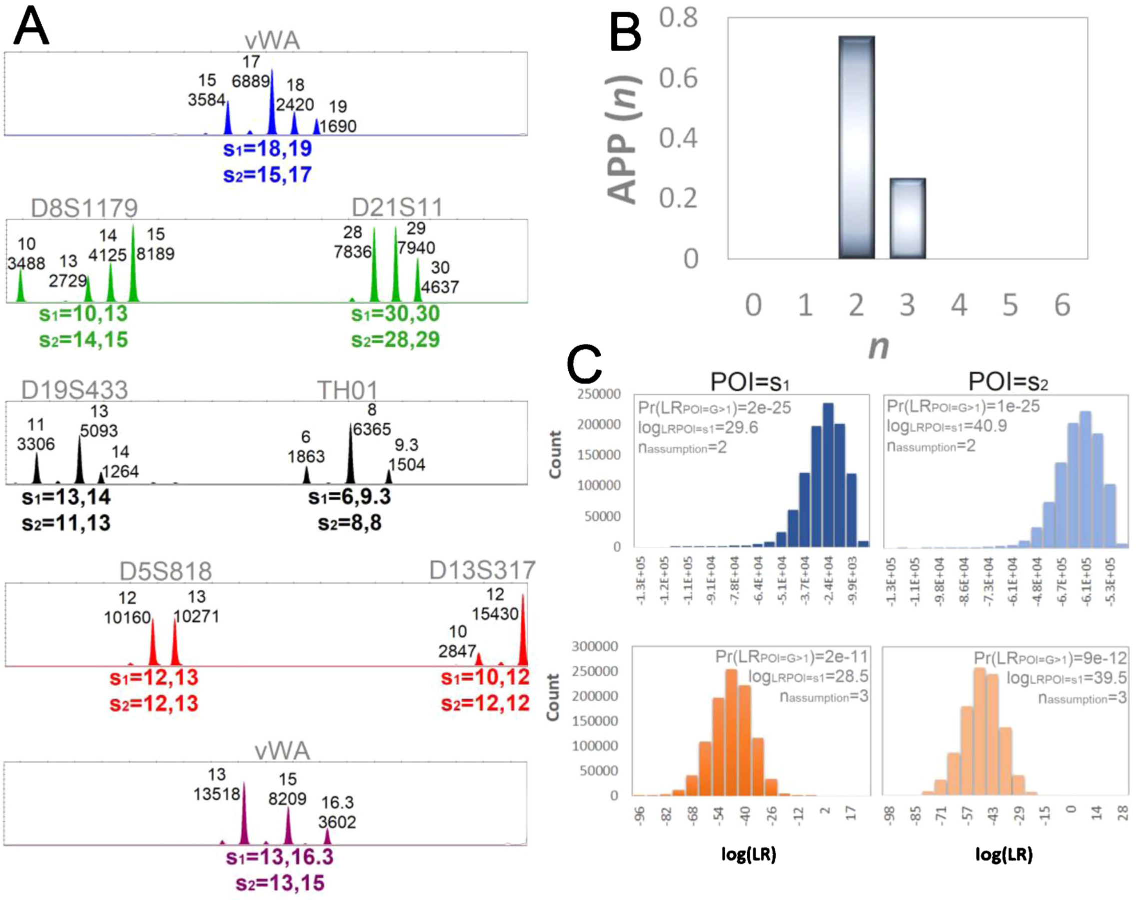

Fig. 6. Example of a sample progressing through the full interpretation pipeline depicted in Fig. 5. (A) The GeneMapper® *ID-X* electropherogram of eight representative STR loci of a 2-person mixture containing one part of a degraded contributor (s_1 = PROVEDIt ID #39) to two parts of an un-degraded contributor (s2 = PROVEDIt ID #40). The total target mass was 0.5 ng. Each known allele peak is labelled with the STR number and peak height above the peaks, while the known genotypes for both contributors, s_1 and $s_{2,}$ are below the peaks. The data were exported from the peak detection software GeneMapper IDX and filtered with CleanIt. (B) The filtered data were imported into NOCIt, and the *aposteriori* distribution on the NOC is shown. NOCIt results suggest that this sample probably arose from 2 or 3 contributors. (C) Both assumptions are used to compute the evidentiary summary statistics in CEESIt for contributors s_1 and s_2 as potential donors to the mixture. The distribution of $\text{LR}_{\text{POI}=\text{G}}$, the likelihood ratio for a potential donor with genotype G drawn uniformly from the background population, is shown under both assumptions. When the smaller, and in this case correct, NOC is assumed (i.e., $n = 2$, top two panels), the LR distribution shifts left, and the LR_{POI} increases. When a larger number is assumed (i.e., $n = 3$, bottom two panels), the LR_{POI} decreases for both s_1 and s_2 while the probability that a random person would result in LR > 1 increases. We note that though the $\Pr(\text{LR} > 1|\text{H}_d)$ can be useful when evaluating matters such as the benefits of including more information into the interpretation pipeline, it is not presented as an alternative to the LR_{POI}. The overall LR for this sample is $10^{29.5}$ and $10^{40.8}$ for s_1 and s_2 respectively, as detailed in Table 4.

Table 4
Computing the overall LR using Eq. (1) [10], the APP from NOCIt, and the $\text{LR}(E|N = n)$ from CEESIt for the sample in Fig. 6.

n	$P(N = n\|E)$	$\sum_{n=1}^{6} \text{LR}(E\|N = n)\text{P}(N = n\|E)$	
		s_1	s_2
0	0	$(0.736{*}10^{29.6}) + (0.264{*}10^{28.5}) = \mathbf{10^{29.5}}$	$(0.736{*}10^{40.9}) + (0.264{*}10^{39.5}) = \mathbf{10^{40.8}}$
1	0		
2	0.736		
3	0.264		
4	0		
5	0		
6	0		

for this single source sample, the MAC assignment of two is consistent with NOCIt's MAP, suggesting the plausibility of 2-persons contributing to this sample; however, the APP also suggested that one contributor might explain the data, leading to a difference in the final LR outcomes. Sample 4, a low-level complex mixture, rendered NOCIt APPs of 0.695 and 0.305 for $n = 3$ and 4, respectively. Given the complexity of the signal, the MAC-based NOC assignment may be in doubt resulting in two or more reported LRs; one which may be categorized as limited in support, while the other as relatively strong. The full LR, however, provides a means to report a single LR across all prominent *n*. Since it is unlikely that manual examination of peak heights is beneficial when the sample is composed of more than two contributors [9] or exhibits allele drop-out [7], integrating a validated APP distribution into the interpretation pipeline as depicted in Fig. 6 may be an important addition for contentious samples. As such, additional large-scale studies that explore, authenticate and report the use of Eq. (1) in forensic casework are warranted.

5. Conclusion

Probabilistic-based interpretation pipelines typically require an assumption on the number of contributors to a sample [35,36]. As a result, work on NOC estimations and studies that examine the effects of the NOC assumptions [1,6,7] on statements of evidential strength have catalyzed the development of methods that address this limitation [11,16,37]. Moreover, interpretation schemes that do not rely upon signal thresholds [15,38,39] or seek to minimize their effects [20] [40] demonstrate that continued development in this field is ongoing. Here

Table 5
Computing the overall LR using Eq. (1) [10], the APP from NOCIt, and the LR($EN = n$) from CEESIt for three PROVEDIt samples. Also shown are the LR values obtained when using n = TrueNOC and the MAC Assignment only.

Sample	MAC Assign.	P($N = n\|E$) n = 0;1;2;3;4;5;6	LR($E\|n$ = *TrueNOC*)	LR($E\|n$ = *MAC*)	Overall LR
1	2	0; 0.161; 0.839; 0; 0; 0; 0	$10^{36.4}$	$10^{30.9}$	$10^{35.5}$
2	1	0; 1; 4.14 × 10^{-10}; 0;0;0;0	$10^{34.4}$	$10^{34.4}$	$10^{34.4}$
3 (Minor)	3	0; 0; 0.736; 0.264; 0; 0; 0	$10^{29.6}$	$10^{28.5}$	$10^{29.5}$
3 (Major)	3	0; 0; 0.736; 0.264; 0; 0; 0	$10^{40.9}$	$10^{39.5}$	$10^{40.8}$
4 (PROVEDIt ID 42)	3	0; 0; 0; 0.695; 0.305; 0; 0	$10^{1.7}$	$10^{3.2}$	$10^{3.0}$

Sample Names as described in [8]: Sample 1. RD14 − 0003-15d2U60 − 0.25GF-Q4.5_01.25 s is a single-source sample amplified with target mass of 0.25 ng and UV-bombarded for 60 min; Sample 2. RD14 − 0003-17d2U60 − 0.25GF-Q13.3_01.25 s is a single source sample amplified at target mass of 0.25 ng, and UV-bombarded for 60 min; Sample 3. RD14 − 0003-39_40 − 1S2;2a-0.5GF-Q1.1_0.1.25 s is a 2-person differentially degraded sample (the minor contributor was subjected to 2 s of sonication) with a target mass of 0.5 ng; and Sample 4. RD14-0003–40_41_42_43 − 1;1;1;1-M4d-0.06GF-Q1.7_01.25 s is a 4-person degraded sample with a target mass of 0.06 ng using PROVEDIt Sample ID 42 as the POI.

we focus on reporting the performance of a fully continuous system, NOCIt, which calculates the P($N = n|E,H_d$).

For most samples, NOCIt's APP suggested that one or two values of n often described the electropherogram well, and the APP distribution was unimodal for all samples tested. The average run time of NOCIt for $n_{max} = 6$ using the most stringent settings tested was 37 min per sample. A comparison between NOCIt, allele counting and allele frequency-based methods indicated that NOCIt outperformed procedures that rely on signal filtering at all levels of complexity, including single-contributor samples. Tests with samples for which TrueNOC= 0 demonstrated that one of the most important features of probabilistic systems is the ability to model noise directly, while still filtering artifacts such as pull-up and minus A, constituting a complete evaluation of the evidentiary electropherogram; thus, we demonstrate that the application of an AT is unnecessary for the interpretation of evidence. We showed that accurately including the TrueNOC into downstream interpretation is affected by the stringency of the APP threshold, the presence of severe electropherogram sloping and, to a lesser extent, the TrueNOC. Finally, using experimental data, two fully continuous systems using the same underlying model (i.e., NOCIt and CEESIt) and the principles articulated in [10] we are the first to demonstrate the way in which NOC APPs may be coupled with LRs to obtain an 'overall LR' using the entire signal (i.e., no analytical threshold). We emphasize that all samples used in this work were obtained from PROVEDIt – a free, open, and forensically relevant database, facilitating direct comparisons between emerging or analogous systems.

Funding disclaimer

This work was partially supported by NIJ2011-DN-BX-K558 and NIJ2014-DN-BX-K026 awarded by the National Institute of Justice, Office of Justice Programs, U.S. Department of Justice as well as W911-NF-14-C-0096 awarded by the Army Research Office, U.S. Department of Defense. The opinions, findings, and conclusions or recommendations expressed in this publication are those of the author(s) and do not reflect those of the Departments of Justice or Defense. Desmond S. Lun and Catherine M. Grgicak are authors of Patent No: PCT/US14/59503.

Acknowledgements

The authors thank Amanda D. Garrett and Lauren Taranow for help with sample preparation, James Kelley with GUI development and Harish Swaminathan for helpful discussions. We also extend our sincere gratitude to two anonymous reviewers whose invaluable input improved the manuscript.

Appendix A

Derivation of Eq. (1), a result from [10]. Assuming that the *a priori* distribution of the number of contributors is the same under both hypotheses in the absence of any data, i.e., $P(N = n|H_p) = P(N = n|H_d)$, and that the posterior probability of the NOC given the defense's hypothesis is positive for all possible NOCs, $P(N = n|E, H_d) > 0$ for all n, the following equality holds due to repeated application of Bayes' Theorem:

$$LR(D) = \frac{P(E|H_p)}{P(E|H_d)} = \sum_n \frac{P(E|H_p, N = n)P(N = n|H_p)}{P(E|H_d)}$$

$$= \sum_n \left(\frac{P(E|H_p, N = n)P(N = n|H_p)P(E, N = n|H_d)}{P(E|H_d)P(E, N = n|H_d)}\right)$$

$$= \sum_n \left(\frac{P(E|H_p, N = n)P(N = n|H_p)P(E, N = n|H_d)}{P(E|H_d)P(E|H_d, N = n)P(N = n|H_d)}\right)$$

$$= \sum_n \left(\frac{P(E|H_p, N = n)}{P(E|H_d, N = n)}\frac{P(N = n|H_p)}{P(N = n|H_d)}\frac{P(E, N = n|H_d)}{P(E|H_d)}\right)$$

$$= \sum_n \left(\frac{P(E|H_p, N = n)}{P(E|H_d, N = n)}\frac{P(N = n|H_p)}{P(N = n|H_d)}P(N = n|E, H_d)\right)$$

$$= \sum_n \left(\frac{P(E|H_p, N = n)}{P(E|H_d, N = n)}P(N = n|E, H_d)\right)$$

$$= \sum_n LR^{(n)}(E)\, P(N = n|E, H_d)$$

We note that, by considering reciprocals and following the same derivation that leads to (1), one can also show that under the same assumptions, but with the posterior constraint being conditioned on the prosecution's hypothesis,

$$LR(E) = \frac{P(E|H_p)}{P(E|H_d)} = \frac{1}{\sum_n \left(\frac{P(E\mid H_d, N=n)}{P(E\mid H_p, N=n)} P(N=n|E, H_p)\right)}$$

$$= \frac{1}{\sum_n \left(\frac{1}{LR^{(n)}(E)} P(N=n|E, H_p)\right)}$$

Thus, it is also possible to evaluate the APP under the prosecution's hypothesis, and still treat the NOC as a nuisance variable in the computation of the likelihood ratio.

Appendix B. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.fsigen.2020.102296.

References

[1] C.C. Benschop, H. Haned, L. Jeurissen, P.D. Gill, T. Sijen, The effect of varying the number of contributors on likelihood ratios for complex DNA mixtures, Forensic Sci. Int. Genet. 19 (2015) 92–99.

[2] J.A. Bright, J.M. Curran, J.S. Buckleton, The effect of the uncertainty in the number of contributors to mixed DNA profiles on profile interpretation, Forensic Sci. Int. Genet. 12 (2014) 208–214.

[3] S.A. Sethi, W. Larson, K. Turnquist, D. Isermann, Estimating the number of contributors to DNA mixtures provides a novel tool for ecology, Methods Ecol. Evol. 10 (1) (2019) 109–119.

[4] J.S. Buckleton, J.M. Curran, P. Gill, Towards understanding the effect of uncertainty in the number of contributors to DNA stains, Forensic Sci. Int. Genet. 1 (1) (2007) 20–28.

[5] M.D. Coble, J.-A. Bright, J.S. Buckleton, J.M. Curran, Uncertainty in the number of contributors in the proposed new CODIS set, Forensic Sci. Int. Genet. 19 (2015) 207–211.

[6] G.M. Dembinski, C. Sobieralski, C.J. Picard, Estimation of the number of contributors of theoretical mixture profiles based on allele counting: Does increasing the number of loci increase success rate of estimates? Forensic Sci. Int. Genet. 33 (2018) 24–32.

[7] S. Norsworthy, D.S. Lun, C.M. Grgicak, Determining the number of contributors to DNA mixtures in the low-template regime: exploring the impacts of sampling and detection effects, Leg. Med. 32 (2018) 1–8.

[8] L.E. Alfonse, A.D. Garrett, D.S. Lun, K.R. Duffy, C.M. Grgicak, A large-scale dataset of single and mixed-source short tandem repeat profiles to inform human identification strategies: PROVEDIt, Forensic Sci. Int. Genet. 32 (2018) 62–70.

[9] P.C. Lynch, R.W. Cotton, Determination of the possible number of genotypes which can contribute to DNA mixtures: non-computer assisted deconvolution should not be attempted for greater than two person mixtures, Forensic Sci. Int. Genet. 37 (2018) 235–240.

[10] K. Slooten, A. Caliebe, Contributors are a nuisance (parameter) for DNA mixture evidence evaluation, Forensic Sci. Int. Genet. 37 (2018) 116–125.

[11] H. Swaminathan, C.M. Grgicak, M. Medard, D.S. Lun, NOCIt: a computational method to infer the number of contributors to DNA samples analyzed by STR genotyping, Forensic Sci. Int. Genet. 16 (2015) 172–180.

[12] L.E. Alfonse, G. Tejada, H. Swaminathan, D.S. Lun, C.M. Grgicak, Inferring the number of contributors to complex DNA mixtures using three methods: exploring the limits of low-template DNA interpretation, J. Forensic Sci. 62 (2) (2017) 308–316.

[13] S. Karkar, L.E. Alfonse, C.M. Grgicak, D.S. Lun, Statistical modeling of short-tandem repeat capillary electrophoresis profiles, 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (2018) 869–876.

[14] S. Karkar, L.E. Alfonse, C.M. Grgicak, D.S. Lun, Statistical modeling of STR capillary electrophoresis signal, BMC Bioinformatics 20 (16) (2019) 584.

[15] H. Swaminathan, A. Garg, C.M. Grgicak, M. Medard, D.S. Lun, CEESIt: A computational tool for the interpretation of STR mixtures, Forensic Sci. Int. Genet. 22 (2016) 149–160.

[16] M.A. Marciano, J.D. Adelman, PACE: Probabilistic Assessment for Contributor Estimation- A machine learning-based assessment of the number of contributors in DNA mixtures, Forensic Sci. Int. Genet. 27 (2017) 82–91.

[17] 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 Sci. Int. Genet. 6 (6) (2012) 689–696.

[18] H. Haned, L. Pene, J.R. Lobry, A.B. Dufour, D. Pontier, Estimating the number of contributors to forensic DNA mixtures: does maximum likelihood perform better than maximum allele count? J. Forensic Sci. 56 (1) (2011) 23–28.

[19] Y. You, D. Balding, A comparison of software for the evaluation of complex DNA profiles, Forensic Sci. Int. Genet. 40 (2019) 114–119.

[20] K.C. Peters, H. Swaminathan, J. Sheehan, K.R. Duffy, D.S. Lun, C.M. Grgicak, Production of high-fidelity electropherograms results in improved and consistent DNA interpretation: standardizing the forensic validation process, Forensic Sci. Int. Genet. 31 (2017) 160–170.

[21] J.M. Curran, C.M. Triggs, J. Buckleton, B.S. Weir, Interpreting DNA mixtures in structured populations, J. Forensic Sci. 44 (5) (1999) 987–995.

[22] J.M. Curran, J.S. Buckleton, An investigation into the performance of methods for adjusting for sampling uncertainty in DNA likelihood ratio calculations, Forensic Sci. Int. Genet. 5 (5) (2011) 512–516.

[23] C.F.D.A.R. Health, General Principles of Software Validation; Final Guidance for Industry and FDA Staff, (2002).

[24] N. Adams, R. Koppl, D. Krane, W. Thompson, S. Zabell, Letter to the editor-appropriate standards for verification and validation of probabilistic genotyping systems, J. Forensic Sci. 63 (1) (2018) 339–340.

[25] SWGDAM, Guidelines for Validation of Probabilistic Genotyping Systems, (2015).

[26] ThermoFisher, GlobalFiler™ PCR Amplification Kit User Guide, (2016).

[27] H. Haned, P. Gill, Analysis of complex DNA mixtures using the Forensim package, Forensic Sci. Int. Genet. Supplement Series 3 (1) (2011) e79–e80.

[28] Report to the President: Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods, (2016).

[29] 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).

[30] T.W. Bille, S.M. Weitz, M.D. Coble, J. Buckleton, J.A. Bright, Comparison of the performance of different models for the interpretation of low level mixed DNA profiles, Electrophoresis 35 (21-22) (2014) 3125–3133.

[31] J.A. Bright, K.E. Stevenson, J.M. Curran, J.S. Buckleton, The variability in likelihood ratios due to different mechanisms, Forensic Sci. Int. Genet. 14 (2015) 187–190.

[32] S. Presciuttini, T. Egeland, About the number of contributors to a forensic sample, Forensic Sci. Int. Genet. 25 (2016) e18–e19.

[33] P. Gill, A response to "about the number of Contributors to a forensic sample", Forensic Sci. Int. Genet. 26 (2017) e9–e13.

[34] C.H. Brenner, Fairness in evaluating DNA mixtures, Forensic Sci. Int. Genet. 27 (2017) 186.

[35] M.W. Perlin, J.M. Hornyak, G. Sugimoto, K.W.P. Miller, TrueAllele® genotype identification on DNA mixtures containing up to five unknown contributors, J. Forensic Sci. 60 (4) (2015) 857–868.

[36] D. Taylor, J.-A. Bright, J. Buckleton, The interpretation of single source and mixed DNA profiles, Forensic Sci. Int. Genet. 7 (2013).

[37] D. Taylor, J.A. Bright, J. Buckleton, Interpreting forensic DNA profiling evidence without specifying the number of contributors, Forensic Sci. Int. Genet. 13 (2014) 269–280.

[38] M.W. Perlin, A. Sinelnikov, An information gap in DNA evidence interpretation, PLoS One 4 (2009).

[39] M.W. Perlin, B. Szabady, Linear mixture analysis: a mathematical approach to resolving mixed DNA samples, J. Forensic Sci. 46 (6) (2001) 1372–1378.

[40] M.A. Marciano, V.R. Williamson, J.D. Adelman, A hybrid approach to increase the informedness of CE-based data using locus-specific thresholding and machine learning, Forensic Sci. Int. Genet. 35 (2018) 26–37.