Journal Pre-proof

Estimating the number of contributors to a DNA profile using decision trees

Maarten Kruijver, Hannah Kelly, Kevin Cheng, Meng-Han Lin, Judi Morawitz, Laura Russell, John Buckleton, Jo-Anne Bright

PII: S1872-4973(20)30179-4

DOI: https://doi.org/10.1016/j.fsigen.2020.102407

Reference: FSIGEN 102407

To appear in: *Forensic Science International: Genetics*

Received Date: 21 June 2020

Revised Date: 30 September 2020

Accepted Date: 3 October 2020

Please cite this article as: { doi: https://doi.org/

Estimating the number of contributors to a DNA profile using decision trees

Maarten Kruijver[1], Hannah Kelly[1], Kevin Cheng[1], Meng-Han Lin[1], Judi Morawitz[1], Laura Russell[1], John Buckleton[1,2] , Jo-Anne Bright[1]

[1]Institute of Environmental Science and Research Limited, Private Bag 92021, Auckland, 1142 New Zealand
[2]University of Auckland, Department of Statistics, Auckland, New Zealand

Highlights

- We describe a decision tree method for assigning NoC
- The method is fast: less than one second for a plate of 90 samples
- The method is fully transparent to the analyst
- The decision tree method for NoC assignment has been shown to be over 77% accurate
- Machine learning approaches have better predictive performance

Abstract

The interpretation of DNA profiles typically starts with an assessment of the number of contributors. In the last two decades, several methods have been proposed to assist with this assessment. We describe a relatively simple method using decision trees, that is fast to run and fully transparent to a forensic analyst. We use mixtures from the publicly available PROVEDIt dataset to demonstrate the performance of the method. We show that the performance of the method crucially depends on the performance of filters for stutter and other artefacts. We compare the performance of the decision tree method with other published methods for the same dataset.

1 Introduction

The assignment of the number of contributors (NoC) to a forensic DNA profile is an important step in the interpretation process. Assessment of the NoC by an analyst is a time-consuming task and can be subjective. There have been numerous publications describing different methods for the assignment of NoC. The simplest of these is by counting the maximum number of alleles at a locus, dividing by two and rounding up (the maximum allele count or MAC method). An upward correction of the lower bound on NoC can often be obtained by applying heuristics. For instance, if peak height ratios are implausible under the lowest theoretical NoC, an upward correction is in order. However, for higher order mixtures (samples originating from four or more contributors) it becomes very hard to apply such heuristics. Additionally in higher order mixtures, estimating NoC using MAC alone is more difficult due to allele sharing amongst contributors [1, 2]. Allele counting has also been shown to be unreliable when alleles from one or more contributors have dropped out from the profile or when low level alleles from contributors are below the analytical threshold (AT) [3]. Uncertainty in NoC may also be an issue in the presence of a major contributor where trace peaks may be artefactual (for example, stutter) or alleles from a trace contributor [4], where it is not known which peaks in the profile are truly allelic peaks and which are artefacts such as stutters [5].

Another approach is to consider the Total Allele Count (sometimes called TAC) across all loci. The expected TAC under a range of NoCs can be computed [6] and compared to the observed count. Such an approach is complicated by alleles potentially dropping out and simulation methods have been proposed for the expected number of alleles, taking dropout into account [7]. Several likelihood-based approaches to assigning NoC have been described that mirror the models for DNA mixture interpretation: ranging from combinatorial (modelling allelic types only) to continuous models (modelling peak heights and artefacts). The combinatorial approach based on maximum likelihood [8,

9] takes into account the population frequencies of the observed alleles but does not take into account the peak heights. The combinatorial approach can be extended to allow for drop-in and dropout following a semi-continuous model (such as the model described in [10]). Penalised maximum likelihood approaches based on a semi-continuous or a continuous model have been proposed [11].

Similarly, continuous models can be used in a Bayesian context to assess NoC [12]. A drawback of using continuous methods for NoC assessment is that model parameters such as contributor template amounts have to be estimated using computationally intensive methods such as MCMC in the Bayesian approaches or by numerical optimisation. In recent years, approaches have been introduced that use statistical classification methods based on features (or covariates) such as MAC and TAC but also including engineered covariates such as the number of loci with 3 to 4 alleles or the standard deviation of the peak heights [13, 14]. These approaches allow for very fast classification based on the covariates without computationally intensive estimation of parameters. A drawback is that the classification methods need to be trained on a large set of ground truth known profiles (profiles for which the true NoC is given). Another issue is that it may be hard to discern how a system reaches a particular conclusion which is undesirable in a forensic application.

In our experience, a MAC method combined with an assessment of peak heights is currently the most commonly used method for NoC assignment by forensic laboratories. Despite the recent development of software solutions for NoC assignment, no single statistical method has become dominant and the relative performance of the different methods is not well understood. Comparison of different published approaches is confounded by the use of different multiplexes and sample types and also by the use of different metrics to score success. In casework there is no well-defined concept of the true number of contributors [15, 16]. When using ground truth known samples as in validation studies, we have the benefit of knowing experimental NoC; that is the number of individual contributor's DNA that was added to the PCR reaction. The experimental NoC may not be reflected in the electropherogram (epg) however. For example, if one or more contributors are present at sufficiently low level that their alleles have dropped out, the profile may appear as having originated from fewer contributors than the experimental NoC. Other methods for scoring the success of a NoC assignment method would be by comparison to values assigned manually by a trained analyst or values from another tool. These can also be inaccurate as manual methods implemented by analysts inherently have an amount of subjectivity. In addition, differences in NoC assignment are expected between software applying different models.

The performance of most methods for assessing NoC depends on data preprocessing including artefact removal and potential stutter filtering or identification. For instance, the MAC method requires all stutters to be removed and failing to remove a single stutter peak can easily lead to an overestimate of NoC. Models for stutter rates are well characterised and rates have been shown to be stable over a range of different platforms for the same kit and cycle number [5, 17, 18].

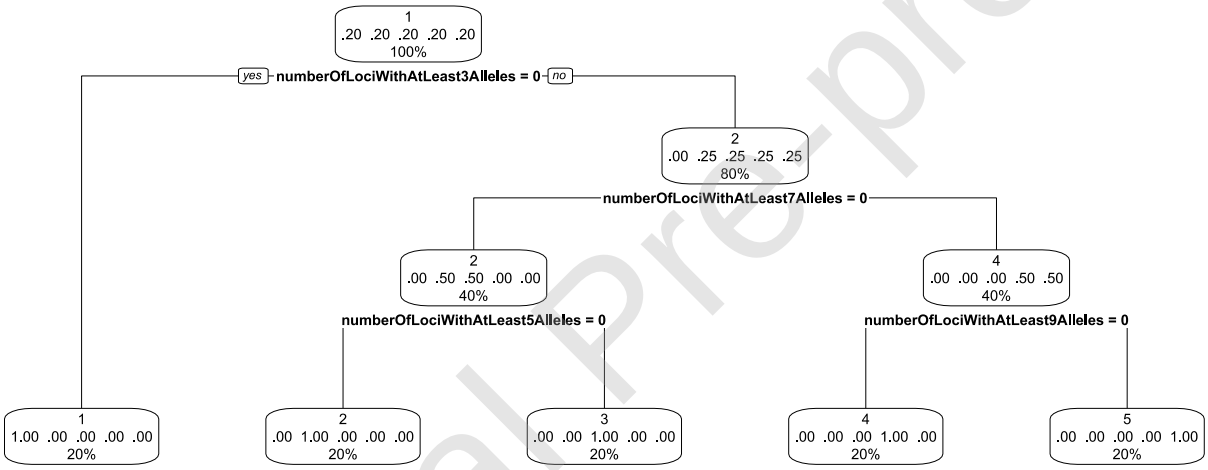
An assessment of NoC is important for mixture interpretation. If a known reference profile is available, it may be compared to the mixture after profile interpretation using a probabilistic genotyping (PG) system and a likelihood ratio (*LR*) assigned. PG software evaluates probabilistically which potential contributor genotypes are most supported by the observed profile given an assumed NoC. In most systems this number must be entered by the analyst, although some systems also allow the evaluation of a range of possible NoCs [13].

In this work we describe a simple method of assigning NoC using a decision tree. A decision tree is a flowchart where at every split one of two branches is taken depending on a test. For a NoC assignment decision tree, the ultimate outcomes are the possible NoC. Inputs of the tests are covariates relating to the DNA profile such as the number of alleles within the profile, minimum and maximum number of alleles at a locus, and allele heights. At each internal node a decision is made where the tree branches (edges). When the tree no longer splits, we reach the decision (a leaf node). A

node within a tree is simply an if/then type question. For these reasons, trees are very intuitive and easy to explain. Once a decision tree has been generated it takes a fraction of a second to assign NoC for a large dataset.

An example decision tree for NoC assignment following a MAC approach is given in Figure 1. The five numbers given at each node are the percentages of data that correspond to each contributor number (1 through 5 contributors). Following the tree in Figure 1, a DNA profile where the number of loci with at least three alleles is equal to zero would be assigned a NoC of one. A profile with one or more loci with at least three alleles would move right along the branch to the next node. If the same profile had at least one locus with at least seven alleles it would again move right to the next node. Finally, if that same profile had zero loci with at least nine alleles it would be assigned a NoC of four. The percentage given in each node is the proportion of data considered at each node, while the percentage given in the leaf node is the proportion of samples assigned to that NoC. In the data used to create Figure 1, there were equal numbers of one through five-person mixtures.

Figure 1: Decision tree representing a MAC rule. Each node corresponds to a subset of the data and starting from the whole dataset the decision rules consecutively refine the subsets. For each node, the bottom line of the label denotes the size of the subset relative to the full data, the middle line gives the fractions of one through five person profiles in this subset and the top line gives the number with the largest fraction (only meaningful in the bottom row because of ties elsewhere)



NOCIt is a Bayesian method for estimating the number of contributors, first described in [19]. NOCIt is described as a continuous method that utilises peak height information and parameters (such as the DNA mass of contributors) to infer the number of contributors to a profile. Peak heights are modelled using a Gaussian distribution. Using this model, NOCIt can compute the probability of observing the heights of peaks. This requires the modelling of a few variables (such as dropout rate, or the mean and standard deviation of true peaks using single-source profiles with specified DNA mass) to calibrate these parameters. Additionally, NOCIt utilises subpopulation allele frequency information and Amelogenin peak heights.

Grgicak et al. [12] demonstrated the capabilities of a new version of NOCIt in a large-scale validation study on 815 profiles of varying quality, number of contributors, and mixture proportions. Within this study, all profiles were analysed at 1 rfu in GeneMapper ID-X and artefacts were removed manually and using the CleanIt[1] module. The NOCIt parameters were calibrated using 100 single-source

[1] Refer to https://lftdi.camden.rutgers.edu/provedit/software/

profiles, and the number of contributors was estimated by assigning a posterior probability and comparing the results to the ground truth.

Benschop et al. [20] described a machine learning approach for NoC estimation. 590 Fusion 6C profiles from 1174 donors were used to test the approach. Ten different machine learning algorithms were trialled including random forest classification (RFC) and linear discriminant analysis (LDA), where 'features' or covariates were generated from allele counts, peak heights, and allele frequencies. The authors reported an accuracy of 83.3% using an RFC method with 19 covariates (RFC-19).

In this research we build decision trees for NoC assignment using the PROVEDIt dataset described in [12] and determine the effectiveness of this simple method. The method assumes that artefacts including stutter have been previously removed from the electropherograms. We compare the performance of the decision trees that were created using one of the three following methods for stutter and artefact filtering:

1. Using a combination of the Kalafut et al. [5] method to automatically filter allelic specific stutter from DNA profiles with a threshold of three standard deviations and manual DNA profile analysis. Referred to as '3SD'.
2. A decision tree method that was developed to filter stutter and artefacts, referred to as 'tree'.
3. An 'oracle' filter that serves as a best-case comparison.

Because the ground truth is known for the PROVEDIt data (the true contributors to each profile and their genotypes are given) we can construct an oracle filter (3) that leaves in any allelic peak and filters every artefact. The simple filter based on a decision tree (2) is created to mimic performance that should be easily achieved in practical settings, and a more advanced filter based on advanced statistical techniques that performs better than a simple method but not as well as the oracle.

The performance of the decision tree NoC method is compared to NOCIt [12], Benschop et al.'s. machine learning approach [20], and MAC.

2 Methods

2.1 Data

Publicly available DNA profiles were used in an attempt to enhance comparability with previous and future NoC assignment research. A previous publication evaluating the performance of NOCIt [12] used 815 samples that were described in the supplementary material. Most of those were downloaded (https://lftdi.camden.rutgers.edu/provedit/files/). These profiles are a subset of the freely available PROVEDIt dataset [21]. The samples consist of 100 single source and 666 two through five-person mixtures amplified using the GlobalFiler multiplex and electrophoresed on an Applied Biosystems 3500 Genetic Analyser machine using a 25 second injection protocol. A set of 49 mixed samples described in [12] as sonicated are not available for download. An overview of the samples is given in Table 1.

Table 1: Overview of the 766 samples that were analysed

NoC	**1**	**2**	**3**	**4**	**5**
Number of profiles	100	174	160	176	156
Total template (ng)	0.008-0.5	0.03-0.75	0.045-0.75	0.06-0.75	0.075-0.75
Contributor ratio	-	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

2.2 Analysis and filtering

Raw .hid files were analysed in FaSTR DNA (https://www.strmix.com/fastr) using an AT of 10 rfu. Pull-up and dye artefacts were removed at analysis. The data used for the tree filter and the oracle filter were exported with stutter peaks retained in the export. For the 3SD method, stutter was filtered in FaSTR DNA using per allele stutter models that were calibrated to a subset of the PROVEDIt data. Following Kalafut et al. [5] stutter was filtered with a threshold using the expected stutter ratio plus three standard deviations.

A filter for stutters and artefacts with more plausible performance was trialled. The tree filter uses a decision tree to assign peaks as either allelic or not allelic based on a number of covariates that were constructed from the data in the epg (allele call, fragment size, peak height). These covariates aid the identification of potential baseline noise, stutter and pull-up. More details on the tree filter are given in Appendix A.

Finally, in order to examine the effect of artefact filtering on NoC assignment performance, an oracle filter was applied to the unfiltered data. The oracle filter perfectly filters any peak that does not correspond to an allele of one of the contributors (including stutter and drop-in). The oracle filter can obviously only be applied in this experimental setting where the contributors are known.

2.3 Tree building

Decision trees were built in R using the *rpart* (Recursive Partitioning And Regression Trees) R package [22, 23] using the covariates described in Table 2 and experimental NoC for a portion of the analysed dataset. A training set of 300 profiles was randomly selected from the full dataset. The same training set was used throughout the experiments to keep results comparable. The remaining 466 profiles were used to compare the accuracy of the different approaches.

Table 2: Covariates used by Decision Trees for NoC classification

Covariate	Description
totalNumberOfAlleles	The total number of distinct alleles seen across all loci
numberOfAlleles*locus*	Number of alleles, one variable for every locus. E.g. numberOfAllelesTH01 denotes the number of alleles at TH01.
numberOfAllelesMin	Sample minimum of the numberOfAlleles*locus* variables
numberOfAllelesMedian	Sample median of the numberOfAlleles*locus* variables
numberOfAllelesMax	Sample maximum of the numberOfAlleles*locus* variables
numberOfAllelesSd	Sample standard deviation of the numberOfAlleles*locus* variables
numberOfLociWith1or2Alleles	The number of loci with either 1 or 2 alleles
numberOfLociWith*n*Alleles	Number of loci with exactly n alleles for $n = 0, 1, \ldots, 12$
numberOfLociWithAtLeast*n*Alleles	Number of loci with at least n alleles for $n = 3, 4, \ldots, 9$
minHeight	Smallest peak height across all alleles
maxHeight	Highest peak height across all alleles
minHeightToMaxHeight	minHeight divided by maxHeight
log10PRMNE	The probability that a Random Man can Not be Excluded at a locus is defined as $P(RMNE)=(\sum f_a)^2$,

	where f_a denotes the frequency of allele a and the sum is taken over all observed alleles. log10PRMNE is computed as the sum of $\log_{10}$(P(RMNE) of all loci

2.4 Comparison to other methods

For the purposes of the comparison between NoC assignment methods, we defined experimental NoC as the number of donors that was used to construct a profile. The number of contributors having alleles present in a profile was possibly lower. For example, sample E02_RD14-0003-40_41_42_43-1;1;1;1-M4e-0.06GF-Q5.6_05.25sec, was an experimentally designed four-person mixture. By visual inspection (MAC and peak height considerations), however, the mixture appeared to have originated from at least two, possibly three, contributors (see Supplementary Figure). There was apparent inhibition or degradation present in this profile. For all methods we define accuracy as the number of predictions equalling experimental NoC divided by the total number of predictions.

2.4.1 NOCIt

There are multiple ways to use the results from NOCIt. These include the incorporation of the APP (A Posteriori Probability) as a nuisance variable in an *LR* calculation; or simply assigning NoC as the value that gives the largest APP. This last approach is termed MAP (Maximum A Posteriori Probability). In this current work, NoC assignment using decision trees results in discrete values, therefore we will be comparing results with the MAP estimate. MAP was determined from the supplementary material from [12].

2.4.2 A machine learning approach

The source code was obtained (https://github.com/JenniferVdL/NOCmodel) and modified to enable the analysis of the PROVEDIt GlobalFiler profiles. The machine learning approach was trained and tested using the same PROVEDIt training (300 profiles) and test (466 profiles) datasets with stutter filtered using the 3SD method and the oracle filtering method. The NIST1036 Caucasian allele frequencies were used. As in Benschop et al. [20], each of the ten different classification methods was trialled and the performance of each was tested based on accuracy.

2.4.3 Maximum Allele Count

Although the theoretical performance of a decision rule based on MAC has been documented in the literature [1, 6], these publications only investigate the somewhat unrealistic case of the absence of stutter and allelic dropout. We evaluate the performance of a MAC rule on the datasets based on the three different stutter filters using the MAC decision tree shown in Figure 1.

2.5 Effective NoC

It is well known that the experimental NoC (the number of single source profiles that were combined to form the mixture) is not necessarily equal to the NoC that is observed in the epg, for instance due to dropout. An objective measure is needed for whether or not a contributor is present is the *LR*. We define an effective contributor to be a contributor for which the *LR* exceeds one when the profile is interpreted using experimental NoC.

Profiles were interpreted in STRmix [17, 23] and an *LR* assigned for the known contributors. The propositions considered were:

H_p: The DNA originated from the person of interest and N-1 unknown contributors,

H_d: The DNA originated from N unknown contributors,

where N was the experimentally designed number of contributors. *LR*s were assigned using allele frequencies from the NIST combined dataset with θ=0.01 [24]. To eliminate false exclusions a minimum *LR* per locus of 1/1000 was manually applied. This prevents a single locus exclusion (*LR*=0) caused by, for instance, a primer binding site mutation or sizing error to affect the results.

2.6 Sensitivity analysis

2.6.1 Training/test split

As for other classification methods, the decision tree approach is relatively sensitive to the training data that are chosen. A different training set may yield a different tree with different accuracy for the test set. The sensitivity of the method to a particular training/test split as well as the size of the training set was tested. The data was split into six different training data sets. The training sets contained 100, 200, 300, 400, 500, and 600 samples. The accuracy of the tree constructed using each training set was tested using the remaining profiles (of the 766 total). For example, a tree trained using 600 samples was tested on the remaining 166 samples. Each of the six training sets was generated 1000 times, each time selecting the 766 profiles differently.

2.6.2 Injection times (25 seconds versus 5 seconds)

In order to determine if decision trees generated using one injection protocol are transferable to profiles generated using a second injection protocol, the 766 profiles analysed using a five second injection protocol were also downloaded and analysed in FaSTR DNA. Stutter and artefacts (including drop-in) were filtered using an oracle filter.

Three experiments were undertaken:

1) The decision tree trained using the 25 second data was used to assign NoC for the five second data,
2) A tree was trained using a subset of the five second data (300 profiles) and tested on the remaining five second data (466 profiles), and
3) The tree trained using the five second data was used to assign NoC for the 25 second data.

The accuracy for all comparisons was determined.

3 Results

3.1 3SD filtered stutter

The decision tree constructed using the 3SD data is given in Figure 2. The confusion matrix for the assigned NoC using this 3SD tree on the test set is given in Table 3. The accuracy of this method was 77.9%.

Figure 2: Decision tree obtained for 3SD filtered data

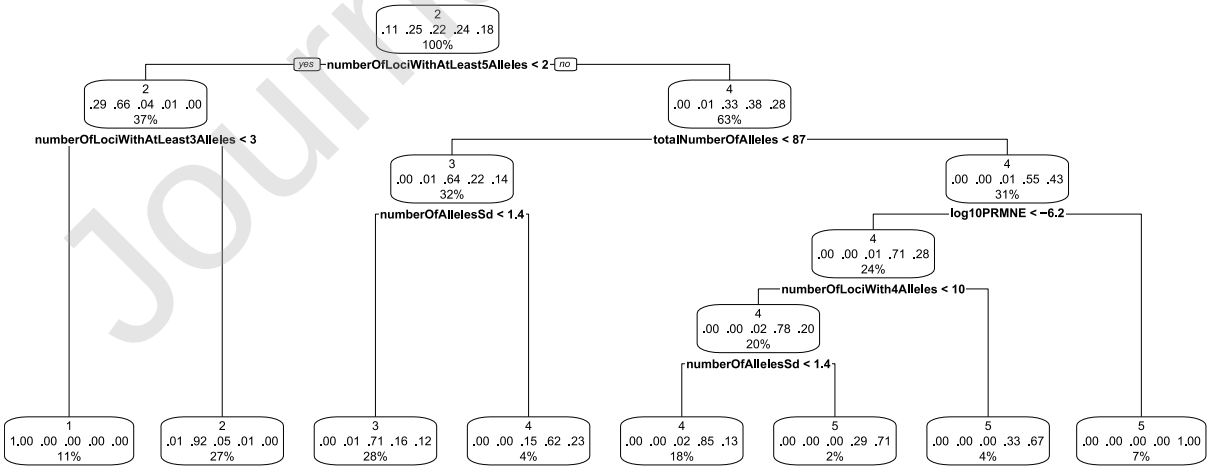


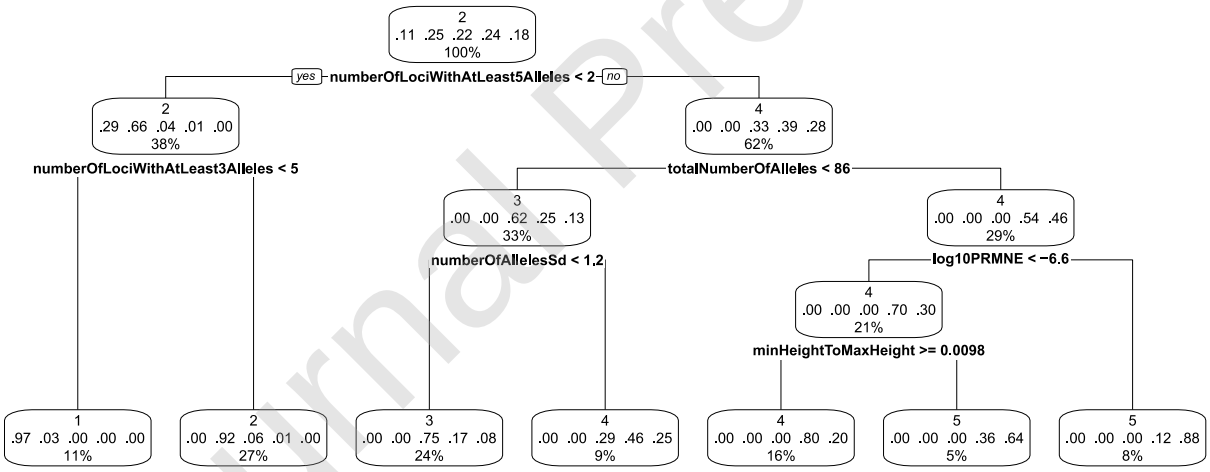
Table 3: Confusion matrix for NoC assigned for 466 profiles using 3SD filtered data, tree filtered data, and oracle filtered data

		3SD filtered (77.9% accuracy)					Tree filtered (78.8% accuracy)					Oracle filtered (85.2% accuracy)				
		1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
Prediction	1	61	0	0	0	0	66	1	0	0	0	67	0	0	0	0
	2	6	95	6	3	1	1	96	9	3	1	0	99	2	2	0
	3	0	4	77	20	7	0	2	72	11	7	0	0	83	12	6
	4	0	0	9	62	27	0	0	13	71	33	0	0	9	89	38
	5	0	0	2	18	68	0	0	0	18	62	0	0	0	0	59

3.2 Tree filtered stutter

After the data were filtered using the stutter filter based on a decision tree (as described in Appendix A) a decision tree was built using a training set (Figure 3). To enhance comparability, the training set consisted of the same samples for each of the trees. Using the tree, NoC was assigned to the 466 remaining profiles in the test set and the results are presented as a confusion matrix in Table 3. The performance of this tree was slightly better than the one operating on 3SD filtered data (Figure 2) with an accuracy of 78.8% (Table 3).

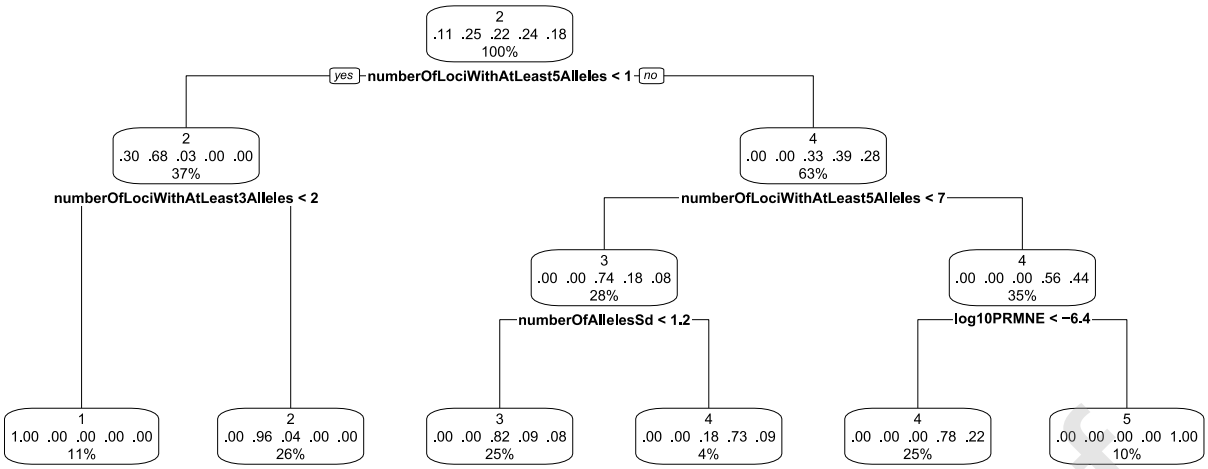
Figure 3: Decision tree estimated on data that were stutter filtered using a separate decision tree described in Appendix A



3.3 Oracle filtered stutter

Finally, the stutter was filtered from the dataset using an oracle filter where all but the known contributor alleles were removed from the dataset. The decision tree is given in Figure 4 and the confusion matrix with the results of the assigned NoC given in Table 3. The accuracy was 85.2%.

Figure 4: Decision tree estimated on data that was stutter filtered using an oracle filter



3.4 Effective Number of Contributors

We have defined the effective NoC as the number of contributors to a mixture for which an *LR* greater than one is obtained when compared to the mixture. All 766 profiles were interpreted in STRmix using experimental design NoC and *LR*s were assigned for known contributors. Where *LR*>1 for a contributor this was counted as an effective contributor. Table 4 compares the effective NoC to the experimental NoC for all 766 profiles (on the left) and also for the test set only (on the right). A large fraction of the higher order mixtures have an effective NoC that is smaller than the experimental NoC.

Table 4: Confusion matrix for effective NoC versus experimental design NoC for all data (left) and the test set (right)

		Experimental NoC									
		All data (93.7% accuracy)					Test set (93.6% accuracy)				
		1	2	3	4	5	1	2	3	4	5
Effective NoC	1	100	1	2	0	1	67	1	1	0	1
	2	0	173	5	0	0	0	98	2	0	0
	3	0	0	153	11	2	0	0	91	7	1
	4	0	0	0	165	26	0	0	0	96	17
	5	0	0	0	0	127	0	0	0	0	84

The effective NoC was used to revisit the performance of the tree that used 3SD filtered data (Section 3.1). The predictions are now scored against the effective NoC rather than experimental NoC. Table 5 shows that a comparable accuracy is obtained when scoring using effective NoC (77.7% versus 77.9% when using experimental NoC).

Table 5: Performance of a decision tree for 3SD filtered data predicting effective NoC (77.7% accuracy)

	Effective NoC (77.7% accuracy)

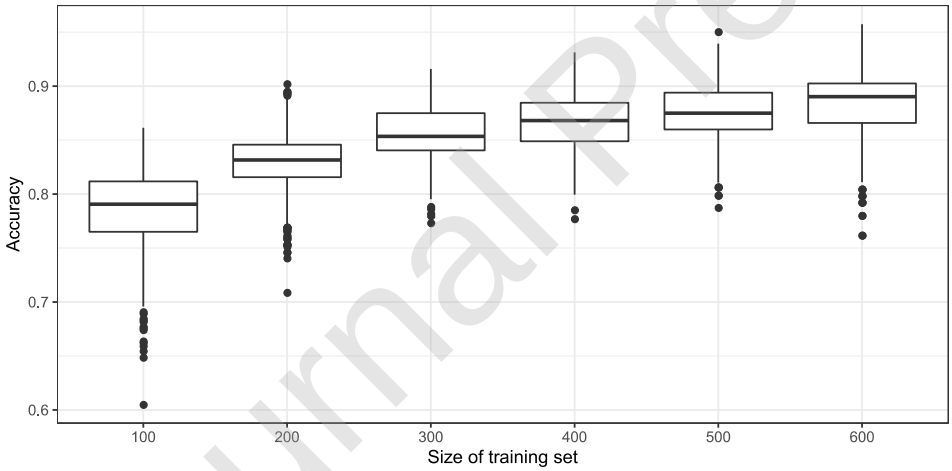
Prediction	1	2	3	4	5
1	61	0	0	0	0
2	8	94	6	3	0
3	1	5	79	20	3
4	0	1	12	66	19
5	0	0	2	24	62

3.5 Sensitivity analysis

3.5.1 Training data

Different decision trees are expected given different training datasets. In order to explore the effect of different training data, 1000 replicates each of a training/test split were generated for each sample size. After a tree was built, NoC was assigned for the remainder of the 766 profiles that was not contained in the training set and the accuracy was evaluated. The results (Figure 5) suggest that a higher accuracy is expected to be obtained when the training set size is increased past 300. When implementing a decision tree it is beneficial to use more samples but this comes at additional cost.

Figure 5: Relationship between accuracy of NoC classifications and size of the training set for oracle filtered data



3.5.2 Injection times (5 seconds versus 25 seconds)

We also explored the effect of different injection times. A new tree was built using 300 5 s injection profiles (oracle filtered data). The tree was used to predict NoC for both the 5 s and 25 s analysed data. In addition, the 25 s oracle filtered tree was used to predict NoC for the 5 s dataset. A summary of the accuracy for each comparison is given in Table 6. In general, the performance is better for 25 s data than the 5 s data, plausibly because more alleles are present in 25 s data.

Table 6: Accuracy for 25 s and 5 s datasets for different combinations of test and training datasets

		Test dataset	
		5 s	25 s
Training dataset	5 s	81.3%	88.0%
	25 s	80.0%	85.2%

3.6 Comparison to other methods

3.6.1 NOCIt

To facilitate comparisons with the results presented in the tables above, we present the NOCIt MAP results [12] as a confusion matrix in Table 7 for all 815 samples and the subset of 466 profiles using the results available in the supplementary materials in the large-scale validation study. The accuracy for the full dataset was 79.8%.

Table 7: Confusion matrix for NoC assigned using the NOCIt MAP estimate (all 815 profiles on left and test set of 466 profiles on right)

		Experimental NoC									
		All data (815 profiles) (79.8% accuracy)					Test set (466 profiles) (79.4% accuracy)				
		1	2	3	4	5	1	2	3	4	5
NOCIt MAP estimate	1	84	0	0	0	0	54	0	0	0	0
	2	16	174	1	2	0	13	93	1	2	0
	3	0	19	141	18	7	0	6	79	10	6
	4	0	0	27	152	49	0	0	13	85	30
	5	0	0	1	14	99	0	0	1	6	59
	6	0	0	0	0	11	0	0	0	0	8

Comparing this table with the confusion matrices from the FaSTR-filtered, tree-filtered, and oracle-filtered, we observe that the NOCIt assignment results in more over-assignments than the three other methods. NOCIt has an observed accuracy of 79.4% for the same dataset where the MAP for 370 of the 466 samples aligned with the experimental design.

3.6.2 Machine learning

Using the 3SD filtered training and testing dataset, we identified that the Random Forest Classifier with 45 features (RFC-45) and the Multilayer Perceptron classifier with 47 features (MLP-47) had the best test accuracy, where 413 of the 466 (88.6%) assigned NoC aligned with the experimental design NoC. However, the Linear Discriminant Analysis classifier with 26 features had a similar test accuracy (LDA-26), where 411 of the 466 (88.2%) assigned NoC aligned with the experimental design NoC (Table 8) with fewer features.

Using the oracle filtered training and testing dataset, we identified that four of the ten classifiers gave test accuracies of 96.4%. That is, 449 of the 466 assigned NoC aligned with the experimental design NoC (Table 8). The classifier that gave the best test accuracy with the fewest features was the Random Forest Classifier with 35 features (RFC-35).

Table 8: Confusion matrix for NoC assigned for the test set of 466 profiles using the Linear Discriminant Analysis classifier with 26 features on the 3SD filtered dataset (88.2% accuracy) and the Random Forest Classifier with 35 features on the Oracle filtered dataset (96.4%).

	Experimental NoC	
	3SD filtered (LDA-26) (88.2% accuracy)	Oracle filtered (RFC-35) (96.4% accuracy)

Prediction	1	2	3	4	5
1	62	2	0	0	0
2	4	96	3	2	0
3	1	1	85	6	2
4	0	0	6	85	18
5	0	0	0	10	83

	1	2	3	4	5
1	67	0	0	0	0
2	0	99	2	2	1
3	0	0	92	1	6
4	0	0	0	97	2
5	0	0	0	3	94

3.6.3 Maximum Allele Count

It is well known that a MAC rule is biased towards under assignment, especially for mixtures with more contributors [9]. This only applies, however, under the unrealistic assumption that there is no stutter or stutter can be perfectly filtered. The performance of a MAC rule was tested using more realistic data by using the three sets of profiles created using the three different stutter filters. Table 9 shows the results. For higher order mixtures there is still a strong tendency to under assign the number of contributors, but this effect is somewhat diminished when a non-perfect stutter filter is used (3SD and tree filtered data). On the other hand, for mixtures with one, two or three contributors there is a clear tendency to over assign the number of contributors. Using the 3SD or tree filter, there is a one in three and one in two chance respectively that a single source profile has at least one locus with three or more unfiltered peaks resulting in it being called a two-person mixture. These could be unfiltered stutters (those with heights above 3SD for example) or drop-in alleles.

Table 9: Performance of a MAC rule applied to the training set after removing stutters with three different filters

	Experimental NoC														
	3SD filtered (57.8% accuracy)					Tree filtered (56.4% accuracy)					Oracle filtered (67.0% accuracy)				
MAC assignment	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
1	42	0	0	0	0	37	0	0	0	0	67	0	0	0	0
2	25	79	2	1	1	30	83	3	2	1	0	99	2	2	0
3	0	19	79	41	24	0	16	87	46	28	0	0	92	47	24
4	0	0	13	60	69	0	0	4	55	73	0	0	0	54	79
5	0	0	0	1	9	0	0	0	0	1	0	0	0	0	0

4 Discussion

We have described an extremely fast (<1 second for a plate of 90 samples) method for assigning NoC to analysed data using a decision tree. An important benefit of the approach is that it is fully transparent to the analyst how the method operates. This enables the analyst to develop a good understanding of why a prediction is made for a particular sample. It is important that the method used to calibrate the tree aligns with the method used to filter the stutter from the casework data. From Table 6, we can see the accuracy of NoC assignment of the test data is weakly dependent on using the tree trained with the same injection parameters. For example, a tree that is estimated using perfectly stutter filtered data would not perform well on imperfectly filtered data. Trees that were calibrated

using the same injection parameters as the test samples also performed better than tress calibrated using samples injected using different parameters.

The benefit in interpretability comes at a cost of accuracy. As expected, the decision tree built on the oracle filtered data was the best performing decision tree method (85.2% accuracy). A more complex machine learning classification algorithm was shown to have better predictive performance with 96.4% accuracy.

The effective filtering of stutter peaks is an important factor for the accurate NoC estimation using the decision tree method described. For the decision trees, the oracle filtered data resulted in the highest accuracy with the manually analysed data using a 3SD filter the lowest (77.9%). Obviously, an oracle filter is not possible for casework profiles where the true contributor profiles are unknown, but it demonstrates firstly that highly accurate predictions are possible with a simple method and secondly that the accuracy of the decision tree approach hinges on the ability to filter artefacts effectively. In casework, an analyst who reviews the allele calls can make a judgment about artefacts and stutter informed by automated methods such as interpretation rules, threshold-based filters and neural networks. The performance of analysts is not known and there is likely to be variation between individuals. A reasonable lower bound on performance is the decision tree that removes artefacts as described in Appendix A. Using this automated filter, the decision tree approach predicted NoC with an accuracy of 78.8%, a surprising improvement over the more manual 3SD approach.

A large fraction of the observed differences between predicted and experimental NoC were under assignments likely due to alleles dropping out of the profile (60/466 test samples for the oracle filtered dataset). A notable example of an experimental five-person mixture being assigned as a two-person mixture is described in Appendix B. There were nine oracle-filtered profiles where NoC was over assigned; all were experimental three-person mixtures assigned as four (Table 3). As described in [21], the DNA for all of these three-person mixtures had been ‘compromised prior to amplification’ including treatment with UV light, DNAse, and sonication. The effects of the treatment were apparent in the electropherograms of these samples with alleles missing at the high molecular weight loci (as may be seen with degraded profiles), and/or a complete or partial loss of alleles at random loci across the molecular weight range (as may be seen with inhibited profiles). This fluctuation in the number of alleles observed across the profile led to the standard deviation of the number of alleles per locus (numberOfAllelesSd from Table 2) being in excess of the threshold set for this covariate based on the training data set. A higher standard deviation of the number of alleles is indicative of a greater number of contributors. As for any NoC assignment method, caution should be exercised for profiles that are highly degraded or inhibited.

MAC was the least accurate method (57.8% accuracy for 3SD-filtered data increasing to 67.0% for oracle filtered data). Using the oracle filtered data, NoC was under assigned for one third of the profiles. This is the expected result. We have demonstrated that the MAC rules can be formulated as a decision tree (Figure 1) suggesting that a decision tree approach can be used to refine a MAC rule towards an approach that is less biased towards under assignment for higher NoCs. An interesting observation is that MAC is biased upwards for lower order mixtures if stutter is not perfectly filtered. Although this is not surprising, it is worth mentioning because it is not well documented in the literature.

The decision tree method was compared to NOCIt. In Table 7, we represent the MAP results from NOCIt as a confusion table. The accuracy is 79.8% for all profiles and 79.4% for the 466 test samples that were used in the current study. From this subset of data, the accuracy of NOCIt is comparable to that of the 3SD-filtered decision tree (77.5%). The automatic assignment through the use of a decision tree, however, does not require the computationally expensive Bayesian modelling approach. A benefit of the Bayesian approach is that it yields posterior probabilities which can be used in weight of

evidence calculations to evaluate whether there is support for a person of interest being a donor to a sample when there is uncertainty about NoC.

Machine learning approaches have been demonstrated to achieve better predictive performance than a decision tree approach. Using the 3SD filtered training and test dataset, the Random Forest Classifier (45 features) and the Multilayer Perceptron classifier (47 features) gave test accuracies of 88.6%. However, the Linear Discriminant Analysis classifier had fewer features required in the model (26 features) and observed a similar test accuracy of 88.2%. Using the oracle filtered dataset, the Random Forest Classifier with 35 features had a test accuracy of 96.4%. This increase in test accuracy mirrors the increase in the test accuracy of the decision trees when stutter is perfectly filtered. This machine learning approach required some time to train, but once the classifiers were trained it could provide an almost instantaneous NoC prediction. Whilst the machine learning approach provides better predictive performance, the method may not be as simple to explain as a decision tree.

We remark that the overall accuracy across categories is a summary statistic that does not tell the whole story. In particular, it is emphasised that accuracy decreases for higher order mixtures.

Table 10 offers another perspective on the results by comparing the accuracy across the methods for different experimental NoC. In particular, it is emphasised that accuracy decreases for higher order mixtures.

Table 10: Accuracy of the compared methods for different experimental NoC on the test set of 466 profiles (3SD filtered data was used except for NOCIt which uses unfiltered data)

		Experimental NoC				
		1	2	3	4	5
Method	MAC	62.7%	80.6%	84.0%	58.3%	8.7%
	Decision tree	91.0%	96.0%	81.9%	60.2%	66.0%
	NOCIt	80.6%	93.9%	84.0%	82.5%	57.3%
	LDA-26	92.5%	97.0%	90.4%	82.5%	80.6%

An important difference between the current dataset and the dataset Benschop et al. used is the number of unique donors. Benschop et al. used 1174 unique donors to construct 590 profiles [20], whereas the PROVEDIt dataset only had 26 unique donors within the 766 profiles used. The PROVEDIt dataset contains the same combinations of donors multiple times. For instance, the 156 five-person mixtures that were used in this study were composed of just seven unique combinations of donors. Fewer unique contributors results in decreased diversity of alleles observed at each locus. There is a risk that classification methods identify patterns that are characteristic to particular combinations of contributors in the training set rather than the number of contributors. A simple thought experiment is to consider ten two-person mixtures constructed using DNA from only two donors, A and B. Assuming that seven of the ten profiles are used to train the models, we could use the absence and/or presence of donor A or donor B's alleles to predict the NoC perfectly in the other three profiles but not in unseen data. Out of the 26 and 35 selected features in LDA-26 and RFC-35 respectively, 13 and 14 of the features are related to the allele frequencies observed at specific loci. It is possible that these classifiers are reliant on the allele frequencies at specific loci which reflect particular combinations of donors, and the classifiers might not perform as adequately on profiles with other donors.

The value of the analytical threshold has an effect on the accuracy of automated methods for NoC estimation. The profiles used by Benschop et al. were analysed with dye specific thresholds with the minimum being 85 rfu. With this higher dye specific threshold, the authors obtained 83.3% test accuracy for Fusion profiles using the RFC-19 model. In this work we have analysed the data at 10 rfu. When samples were analysed using a higher AT of 50 rfu, the accuracy of the decision tree decreased (data not shown). An increased AT results in the filtering of additional stutter peaks, however it will also filter low-level allelic peaks. The filtering of low level allelic peaks can result in the under-assignment of NoC as one or more trace contributors may no longer be observed.

The decision tree approach requires a large set of training data. Figure 5 shows the relationship between accuracy and the size of the training dataset. The plot shows a general trend of increasing accuracy with increasing dataset size and suggests that larger training sets are beneficial. However, this increase does plateau to a level that depends on the information content of the profiles. Moreover, there are profiles where due to the amount of dropout the predicted NoC is never expected to equal experiment NoC irrespective of the NoC assignment method. For example, see samples E02_RD14-0003-40_41_42_43-1;1;1;1-M4e-0.06GF-Q5.6_05.25sec (supplementary material and described in Section 2.4). Limited improvements are expected when the size of the training dataset is further increased.

Some further limitations of the current study warrant discussion. The performance of the classification approaches discussed in this study, barring NOCIt, depends on the composition of the data set. The relative fractions of mixtures with each NoC influence the accuracy, because the methods tend to perform poorly on higher order mixtures. Moreover, real-world accuracy depends on the representativeness of the data that was used to train the model. Practitioners may be interested in tuning a model such that misclassifications in one direction are favoured over another direction. For instance, it may be preferable to misclassify a single source profile as a two-person mixture than the other way around, because a continuous genotyping system has – in most cases – little problem with identifying that a superfluous contributor would have contributed a negligible amount of DNA. On the other hand, a mixture interpretation may fail to run or will result in false exclusions if NoC is underestimated. It is an interesting area of future work to tune models to preferentially misclassify NoC.

In conclusion, the decision tree method for NoC assignment has been shown to be over 77% accurate, with increasing performance with improved stutter and artefact filters. This quick and simple method is close to the performance of more complicated methods but has the appeal of being computationally less intensive and easier to explain.

Acknowledgements

This work was supported in part by grant 2017-DN-BX-0136 from the United States National Institute of Justice.

Appendix A Tree filter for stutters and artefacts

A decision tree model was created for the removal of stutter and other artefacts. All peaks in the 766 profiles from the PROVEDIt dataset were annotated with an indicator for the class: either 'Allele' or 'Artefact', where the latter is assigned to any peak not consistent with the contribution of a known reference. The data consisted of 82,080 peaks of which 56,349 were allelic (69%) and 25,731 were artefactual (31%).

Several covariates were constructed for the dataset that were considered plausible predictors of a peak being allelic or artefactual. These included:

- for each type of stutter (back, forward, double back, half back): the parent height, expected stutter ratio and the expected stutter height,
- the total expected stutter height,
- the ratio of the observed height to the expected stutter height,
- the difference of the observed height and the expected stutter height,
- the height relative to the total fluorescence at the locus,
- the probability of observing a peak below the observed value conditional on it being a stutter peak, i.e. the quantile of the observed value in the stutter height distribution. Peak heights were modelled using a lognormal distribution.

After dividing the data randomly into a training set (20,000 peaks) and test set (62,080 peaks), a decision tree model was built. A very simple tree with only two decision nodes, shown in Figure 6, was found using the *rpart* package in R. The first branch acts on the quantile in the stutter distribution. If the peak is unusually high for stutter then it is labelled allelic. If not, then another division is made based on the height of the peak compared to the fluorescence at the locus: if the peak height is at least 1.4% of the fluorescence at the locus then it is assigned to be allelic.

The performance of the tree is demonstrated by the predictions shown in Table A.1. Despite the simplicity of the model, an accuracy of 97.4% was achieved, which makes it a useful model. The tree was applied to the full dataset (a fraction of which was earlier used as a training set) to obtain a filtered dataset that was used for NoC estimation.

For comparison, Table A.2 shows the accuracy of the 3SD filter that was applied in combination with manual removal of non-stutter artefacts. The accuracy is slightly lower than the decision tree filter (97.0% versus 97.4%). The number of artefacts that are labelled as allelic, however, is over twice as high when the 3SD filter is used (954 versus 474).

Figure 6: Decision tree for filtering stutter and other artefacts

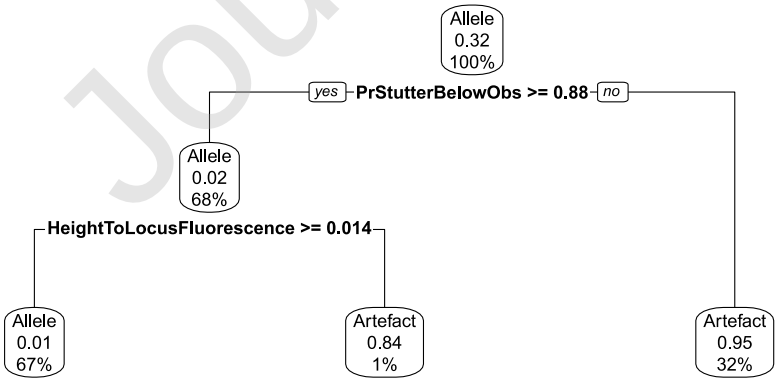


Table A.1: Performance of the decision tree filter for stutters and artefacts (97.4% accuracy)

		Reference	
		Allele	Artefact
Prediction	Allele	42,100	474
	Artefact	1153	19,217

Table A.2: Performance of the 3SD filter accompanied by manual profile analysis (97.0% accuracy)

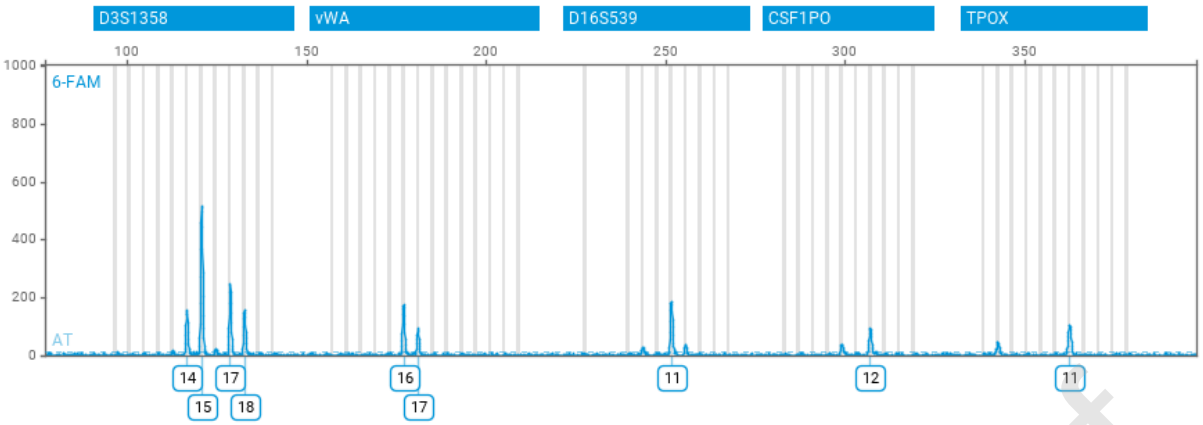
		Reference	
		Allele	Artefact
Prediction	Allele	42,296	954
	Artefact	957	18,737

Appendix B Noteworthy misclassification

B.1 B09_RD14-0003-33_34_35_36_37-1;1;4;1;1-M3e-0.12GF-Q2.2_02.25sec

The decision trees based on 3SD filtered and tree filtered data both assigned this experimental five-person mixture as a two-person mixture. Looking at the two respective trees (Figure 2 and Figure 3), there is only one possible path for a sample to be assigned a NoC = 2. A sample must have fewer than two loci that have at least five alleles, which is unusual for a five-person mixture. Additionally, the sample must have more than three (for 3SD filtered data) or five (tree filtered data) loci that have at least 3 alleles. Judging from the excerpt (blue channel only) of the epg shown in Figure B.1, it seems plausible that the sample is suitably low-level and degraded to be considered as having originated from fewer than five contributors. According to the sample naming conventional described for the PROVEDIt dataset [21], this sample was subjected to harshest DNase I degradation method (24 mU DNase I).

Figure B.1: Excerpt from the epg of an experimental five-person mixture that was misclassified as a two-person mixture



Appendix C The 26 and 35 selected features used within the machine learning Linear Discriminant Analysis and Random Forest Classifiers

LDA-26 Features	RFC-35 Features
• MAC • TAC • stdAllele • peaksBelowRFU • MinNOC_D21S11 • SumAF_D21S11 • AlleleCount_CSF1PO • HighAF_TPOX • Below/AboveRFU • PercAF_D1S1656 • LowAF_D18S51 • SumAF_D8S1179 • HighAF_D2S1338 • minHeight_D3S1358 • HighAF_D12S391 • PercAF_D2S441 • MinNOC_CSF1PO • SumAF_vWA • SumAF_TH01 • minHeight_TH01 • LowAF_D1S1656 • HighAF_D1S1656 • AlleleCount_D8S1179 • SumAF_D19S433 • medianHeight_D2S441 • AlleleCount_D5S818	• MAC • TAC • stdAllele • peaksAboveRFU • PercAF_D19S433 • AlleleCount_D19S433 • PercAF_D22S1045 • Below/AboveRFU • MAC3-4 • MinNOC_D22S1045 • PercAF_D1S1656 • AlleleCount_CSF1PO • SumAF_vWA • HighAF_D22S1045 • minHeight_D21S11 • AlleleCount_D10S1248 • HighAF_D21S11 • MAC0 • SumAF_TH01 • PercAF_D5S818 • MAC1-2 • MinNOC_D19S433 • HighAF_D16S539 • MinNOC_D1S1656 • SumAF_D22S1045 • MinNOC_D2S441 • SumAF_D3S1358 • minHeight_D3S1358 • maxHeight_D22S1045 • HighAF_D10S1248 • HighAF_D3S1358 • AlleleCount_D22S1045 • HighAF_D12S391

	• HighAF_D2S1338 • AlleleCount_D13S317

References

[1] 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 Science International: Genetics 19 (2015) 207-211.

[2] D.R. Paoletti, T.E. Doom, C.M. Krane, M.L. Raymer, D.E. Krane, Empirical analysis of the STR profiles resulting from conceptual mixtures, Journal of Forensic Sciences . 50 (2005) 1361-1366.

[3] 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, Legal Medicine 32 (2018) 1-8.

[4] 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 Science International: Genetics 12 (2014) 208-214.

[5] T. Kalafut, C. Schuerman, J. Sutton, T. Faris, L. Armogida, J.-A. Bright, J. Buckleton, D. Taylor, Implementation and validation of an improved allele specific stutter filtering method for electropherogram interpretation, Forensic Science International: Genetics 35 (2018) 50-56.

[6] T. Tvedebrink, On the exact distribution of the numbers of alleles in DNA mixtures, International Journal of Legal Medicine 128(3) (2014) 427-437.

[7] H. Haned, K. Slooten, P. Gill, Exploratory data analysis for the interpretation of low template DNA mixtures, Forensic Science International: Genetics 6(6) (2012) 762-774.

[8] T. Egeland, I. Dalen, P.F. Mostad, Estimating the number of contributors to a DNA profile, International Journal of Legal Medicine 117 (2003) 271-275.

[9] 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?, Journal of Forensic Sciences 56(1) (2011) 23-28.

[10] K. Slooten, Accurate assessment of the weight of evidence for DNA mixtures by integrating the likelihood ratio, Forensic Science International: Genetics 27 (2017) 1-16.

[11] Ø. Bleka, C.C.G. Benschop, G. Storvik, P. Gill, A comparative study of qualitative and quantitative models used to interpret complex STR DNA profiles, Forensic Science International: Genetics 25 (2016) 85-96.

[12] C.M. Grgicak, S. Karkar, X. Yearwood-Garcia, L.E. Alfonse, K.R. Duffy, D. Lun, A large-scale validation of NOCIt's A Posteriori Probability of the number of contributors and

its integration into forensic interpretation pipelines, Forensic Science International: Genetics (2020) 102296.
[13] 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.
[14] J.-A. Bright, J.M. Curran, Investigation into stutter ratio variability between different laboratories, Forensic Science International: Genetics 13(0) (2014) 79-81.
[15] K. Slooten, A top-down approach to DNA mixtures, Forensic Science International: Genetics 46 (2020).
[16] T. Bille, S. Weitz, J.S. Buckleton, J.-A. Bright, Interpreting a major component from a mixed DNA profile with an unknown number of minor contributors, Forensic Science International: Genetics 40 (2019) 150-159.
[17] J.-A. Bright, D. Taylor, J.M. Curran, J.S. Buckleton, Developing allelic and stutter peak height models for a continuous method of DNA interpretation, Forensic Science International: Genetics 7(2) (2013) 296-304.
[18] J.-A. Bright, J.M. Curran, J.S. Buckleton, Investigation into the performance of different models for predicting stutter, Forensic Science International: Genetics 7(4) (2013) 422-427.
[19] 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 Science International: Genetics 16 (2015) 172-180.
[20] C.C.G. Benschop, J. van der Linden, J. Hoogenboom, R. Ypma, H. Haned, Automated estimation of the number of contributors in autosomal short tandem repeat profiles using a machine learning approach, Forensic Science International: Genetics 43 (2019) 102150.
[21] 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 Science International: Genetics 32 (2018) 62-70.
[22] L. Breiman, J. Friedman, C.J. Stone, R.A. Olshen, Classification and Regression Trees, CRC Press, Boca Raton FL, 1984.
[23] D. Taylor, J.-A. Bright, J. Buckleton, The interpretation of single source and mixed DNA profiles, Forensic Science International: Genetics 7(5) (2013) 516-528.
[24] C.R. Hill, D.L. Duewer, M.C. Kline, M.D. Coble, J.M. Butler, U.S. population data for 29 autosomal STR loci, Forensic Science International: Genetics 7(3) (2013) e82-e83.