Forensic Science International: Genetics 43 (2019) 102140

Contents lists available at ScienceDirect

Forensic Science International: Genetics

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

Research paper

Developmental validation of PACE™: Automated artifact identification and contributor estimation for use with GlobalFiler™ and PowerPlex® fusion 6c generated data

Michael A. Marciano*[,1], Jonathan D. Adelman[1]

Forensic & National Security Sciences Institute, Syracuse University, 107 College Place 120 Life Science Building, Syracuse, New York, 13244 USA

ARTICLE INFO

Keywords:
DNA mixture
number of contributors
artifact identification
random forest
machine learning
complex interpretation

ABSTRACT

DNA mixture interpretation remains one of the major challenges in forensic DNA analysis. DNA mixture samples are inherently complex due to several factors including the variations in the quantity of DNA, the presence of non-allelic artifactual peaks and the presence of multiple contributors with variable levels of allele sharing. The Probabilistic Assessment for Contributor Estimation (PACE) is a fully continuous probabilistic machine learning-based method to predict the number of contributors (*n*) in a sample, and was previously developed for use with the Identifiler amplification kit. This system required manual preprocessing of data and was limited, exclusively, to samples amplified using said kit. This study introduces PACE™ v1.3.7 for use with both the GlobalFiler and PowerPlex Fusion 6c amplification kits. An automated artifact identification and management system has been added to accompany the rapid estimation of the number of donors in a given mixture. The artifact management module, when evaluated using previously unseen data, identified true allelic peaks and removed artifacts such as elevated baseline noise, stutter, and pull-up with accuracy over 93.5%. The systems yield the correct *n* classifications in over 90% of the samples, and demonstrate consistent accuracies as the number of donors and the overall mixture complexity increase. Misclassified samples generally exhibited high levels of allele sharing among donors, low DNA template amounts and high incidence of allelic dropout. This system offers a means for both artifact management and *n* estimation as well as a quantitative and reproducible method of assessing the quality of a profile.

1. Introduction

DNA mixture interpretation remains a primary focus of the forensic DNA analysis community. The advent and implementation of probabilistic genotyping has changed the landscape of how DNA mixtures can be interpreted [1,2]. These methods typically require an assessment of the number of contributors (*n*) in a DNA mixture, an interpretation that can be challenging due to stochastic effects, allele sharing, and variability in the level of contribution by the donors. This need helped spur the development of computationally intensive interpretational tools that can probabilistically assess the number of contributors in a DNA mixture. These include NOCit [3], which uses a Monte Carlo sampling algorithm to estimate the posterior probability on *n* given an electrophoresis profile, and PACE [4], which uses machine learning to estimate the optimal decision boundaries in feature space between the possible values of *n*, and subsequently constructs a predictive model that estimates *n* for a given profile.

The assessment of the number of contributors is widely considered one of the most critical elements leading to accurate DNA mixture interpretation. The assumption of *n* establishes a set of expectations for the number of alleles expected at any single locus and is typically required to formulate accurate conclusions when using either manual (i.e. binary) or probabilistic genotyping (PG) methods. Regardless of the method of deconvolution / interpretation, accurately predicting *n* remains of high importance. It is now well-known that incorrect assumptions of *n* are more probable in higher-order mixtures [5–8] and can negatively impact the resulting likelihood ratios (LR) [9,10] or potentially lead to false exclusions. Bright et al. 2018 provides one of the most comprehensive examinations of the effect an incorrect *n* assumption can have on resulting LRs. This study demonstrated that LRs decrease when *n* is overestimated (particularly when assessing minor contributors) and underestimations can likewise result in false exclusions [11]. More recently, Bille et al. 2019 demonstrated that, when using STRMix®, the LRs relating to minor contributors are reduced when the incorrect number of

* Corresponding author.
E-mail address: mamarcia@syr.edu (M.A. Marciano).
[1] Both authors contributed equally.

https://doi.org/10.1016/j.fsigen.2019.102140
Received 21 June 2019; Received in revised form 19 July 2019; Accepted 31 July 2019
Available online 08 August 2019

contributors is assumed [12]. Ultimately, determining *n* prior to mixture interpretation will lead to the most accurate conclusions.

We recently introduced PACE, probabilistic assessment for contributor estimation, a method that uses machine learning as a component of a fully continuous probabilistic system that predicts the number of contributors in a DNA mixture [4]. This study helped establish the utility of machine learning in the forensic DNA niche as well as introduce a rapid means of predicting the number of contributors in a DNA mixture. The original PACE system was developed for use with the Identifiler™ amplification kit (Thermo Fisher Scientific Inc.) and demonstrated accuracies of 100%, 98.1%, 95.9% and 100% for samples with 1-, 2-, 3-, and 4+ contributors, respectively. The rare misclassification was typically the result of a low-level minor contributor that displayed both high allele dropout and elevated levels of allele sharing. Although highly accurate, this system also required manual interpretation of the electropherogram to identify artifacts such as pull-up, electrical spikes, and dye blobs. However, this manual interpretation can be time-consuming and subjective.

In this study, we introduce PACE v1.3.7, which has been designed for use with samples amplified by the GlobalFiler™ (Thermo Fisher Scientific Inc.) or PowerPlex® Fusion 6c amplification kits (Promega Corporation). PACE now includes a fully automated means to perform artifact management, including identification and removal of pull-up, non-traditional stutter (a – 10 through a + 4) and noise above threshold. As with the preceding version of PACE, the new systems provide rapid, probabilistic assessments for contributor estimation.

2. Materials and methods

2.1. Sample sets and computational resources

This paper describes the development of PACE for use with samples amplified using the PowerPlex Fusion 6c and GlobalFiler Human DNA amplification kits. The PACE-GlobalFiler (PACE-GF) system was trained and tested using electronic data (.fsa/. hid files) obtained from 3921 non-simulated DNA samples of 1 to 5 contributors generated from a combination of 79 individuals and amplified using the GlobalFiler Human DNA amplification kit. Samples within this set included 99 different DNA template amounts ranging from 3.0 pg to 3.5 ng and 88 different ratios of contributors. Samples were run on three Applied Biosystems 3500 Genetic Analyzers (Thermo Fisher Scientific Inc.) using four different injection times (5,15,24,25 s) and two injection voltages, 1.2 and 1.5 kV (Table 1). This sample set included samples inhibited using humic acid, degraded using rDNase I, Fragmentase, sonication and damaged via UV-light [13].

The PACE-PowerPlex Fusion 6c (PACE-PPF6c) system was trained and tested using electronic data (.fsa/. hid files) obtained from 1969 non-simulated DNA samples of 1 to 5 contributors generated from a combination of 120 individuals and amplified using the PowerPlex Fusion 6c Human DNA amplification system. Samples within this set included 152 different DNA template amounts ranging from 3.0 pg to 5.1 ng and 49 different ratios of contributors. Samples were run on five Applied Biosystems 3500 Genetic Analyzers using 15, 20 and 24 s injections at 1 kV and three Applied Biosystems 3100 series Genetic Analyzers (Thermo Fisher Scientific Inc.) using 3, 5, 10, 11, 12, 15, 20 and 30 s injections times with a 3 kV injection voltage (Table 1).

Table 1
Summary of the PowerPlex Fusion 6c and GlobalFiler data sets used to train and test the PACE algorithms.

	PowerPlex Fusion 6c	GlobalFiler
Sample #	1969	3921
Individuals	120	79
Template range	3.0 pg-5.1 ng	3.0 pg-3.5 ng
Mixture ratios	49	88
Instruments	9 (5-3500 s, 4-31XX series)	3 (3500 s)
Injection time / voltage	9 times / 2 different kVs	4 times / 2 different kVs

Both data sets were compiled from publicly available data, e.g. PROVEDIt database [13,14] and validation data provided by public, CODIS-participating, crime laboratories. Genotypes for select samples in the data sets cannot be disclosed due to privacy restrictions. Further information regarding the samples can be found in the Supplementary Information, tables 1S and 2S.

Fragment analysis was performed using PACE embedded with Osiris 2.11 Beta 4 [15] and ArmedXpert (NicheVision Forensics) using a threshold of 10 relative fluorescent units (RFU) and without stutter filters. Manual preprocessing of the data was not performed.

All computations were performed on computers with either Intel Core i5–3230 M CPU @ 2.6 GHz, 8GB RAM or Intel Xeon E5-2699 v3 @ 2.30 GHz with 128MB of RAM and required less than 400MB of hard drive space. Models were generated using an Intel Xeon E5-2699 v3 @ 2.30 GHz with 128MB of RAM and took approximately 12 h. Note this is only experienced during the developmental phase and does not reflect the duration of analyses.

2.2. Allele and artifact identification

Because machine learning is ideally suited to mine information from large swaths of data, the choice was made to focus preprocessing on avoiding false negatives (i.e. avoiding the removal of true alleles) at the expense of having a larger number of false positives (i.e. retention of artifacts). Previous work demonstrated the efficacy of a locus- and sample-specific analytical threshold [16], which is calculated using the mean and standard deviations of the noise in regions flanking a locus within an individual sample. The same approach was used here as an initial preprocessing step. Traditional and non-traditional stutter (a-10 through a+4) was subsequently accounted for and removed throughout the profile; stutter rates for all peaks were determined using internally developed, empirically-derived stutter models [16] for both forward and reverse stutter. Similarly, full and partial pull-up were quantified and removed using empirically-derived models [17].

Two trimming algorithms described more fully in [16] were then used to remove remaining noise due to non-allelic, non-stutter-related peaks. The first algorithm finds the locus with the largest number of potential allelic peaks, stores the height of the smallest potential peak at that locus (or the mean value of all smallest peaks' heights, if multiple loci have the same maximum number of peaks) as a "global minimum peak height", and subsequently – for each of the remaining loci – removes peaks with a peak height of less than 5% of the global minimum peak height. The second algorithm removes potential allelic peaks from a locus if those peaks' heights are below a user-specific proportion of the highest peak height at that locus. This ratio was empirically optimized to 1.9% in order to balance the removal of noise and the retention of low-level allelic activity. A third trimming algorithm, not described in [16], removes many of the remaining false positive peaks by evaluating each potential allelic peak at each locus and removing the peak if the ratio of the peak area to the peak height is less than 1; as above, the threshold was empirically identified. Machine learning was used as a final pre-processing step to identify any additional artifacts not labeled as such by the previously described models and algorithms. Accuracy measures indicate the success of removing artifactual peaks.

2.3. Machine learning algorithms

PACE-PPF6c and PACE-GF were developed using a series of learning algorithms, each of which was responsible for learning a part of the overall feature vector used for *n* classification. Each of these individual classifiers was a form of supervised machine learning known as a random forest. Like other supervised learning algorithms, a random forest requires a data set with both a vector of pertinent features (i.e. characteristics that have predictive value when attempting to classify some other variable) and a label (i.e. the "correct answer" for that particular data instance). For example, an algorithm attempting to learn

the number of contributors based on information from a single locus might have features such as the number of observed peaks at that locus, or the ratio of the observed peak heights at that locus; the label would be the number of contributors for that particular DNA sample.

The random forest classification algorithm is an ensemble learning method, in which a large number of individual decision trees are used in concert. Each decision tree attempts to classify data using a subset of the original dataset as well as a subset of the initial feature vector. To continue the previous example, one decision tree might use the ratio of peak heights but not the total number of peaks, and it would be trained using only part of the initial data set used for training. The final random forest classification is derived from a majority vote of the individual decision trees [18], whereby each individual decision tree ($\mathscr{x}$, Θ_k) determines random forest classification (h) (Eq. 1) [19].

$$\{h(\mathscr{x}, \Theta_k),\ k= 1...\} \tag{1}$$

The individual decision trees, ($\mathscr{x}$, Θ_k), receive input x, where x is a sample with a feature vector. For example, given a hypothetical sample A, with a feature vector (x) that includes peak height, peak area, maximum peak height, and DNA template amount; Θ_k corresponds to the tree structure associated with the k^{th} decision tree, including the subset of features that will be used in the k^{th} decision tree. Recursive partitioning is used to build each decision tree where each branch point (or node) is a randomly selected variable from the full set of variables. The node with the lowest error is selected and becomes the parent node; two or more child nodes are then built in a similar manner to the parent. Branching continues until the terminal nodes have correctly identified the class or contain a specified number of correct classifications [18]. Each decision tree in the forest then "votes" on the appropriate classification of the sample(s); the class is determined by an unweighted majority vote. This yields comparatively high prediction accuracy at a lesser computational cost and prevents overfitting; models that are "over fit" result in poor generalizability, the inability to classify unknowns that are slightly disparate from the original data set. This is avoided through random selection of features used in each decision tree. In addition, random forests are able to perform with small sample sizes and comparably high numbers of features, thus partially mitigating "the curse of dimensionality" [20], which involves handling small sample sizes with comparably high numbers of features to ensure the trees are not correlated. To do so, cross-validation is used to predict expected classification error rate [21].

2.4. Feature selection, scaling, and importance

Feature selection was informed by our previous work that developed PACE for use with the Identifiler™ human DNA amplification kit [4]. Additional features were included due to the kit-specific reaction dynamics (e.g. stutter rates, expected heterozygote balance) and the presence of additional information such as additional autosomal loci and Y-specific loci. Precursor learning algorithms (such as locus-specific n classifiers) utilized similar, but unique, feature vectors; each classifier's selected feature vector is meant to provide features with the most predictive value for the specific classification being performed. Feature selection for the final learning algorithm that classifies n is described below.

Machine learning algorithms may not appropriately utilize the raw features in a feature vector because feature scales can be wildly different from one another. DNA template, for example, has mean and variance several orders of magnitude smaller than the sample-wide maximum observed peak height. The machine learning algorithm will spend a disproportionate amount of time minimizing the larger errors produced by features with larger variances, leading - in this example - to peak height's importance being artificially inflated. One solution is to standardize features (Eq. 2):

$$X_{std}^{(i)} = \frac{X^{(i)} - \mu_x}{\sigma_x} \tag{2}$$

where $X^{(i)}$ is a given feature, μ_x is the feature's mean, and σ_x is the corresponding standard deviation. Eq. 2 was used to scale all features in this study.

Feature importance metrics provide the relative value of individual features in aiding successful classification, *i.e.* the discriminatory power of a feature. One commonly used measure associated with random forests is Gini importance, which is defined as the total decrease in node impurity (approximately weighted by the proportion of samples that reach that node) averaged over all trees in the forest (Eq. 3):

$$i(\tau) = 1 - \rho_1^2 - \rho_0^2 \tag{3}$$

where ρ_k is the number of samples of class k divided by the total number of samples. The algorithm attempts to identify the features at a node that maximizes the change in i. The decrease in Gini impurity at node τ in individual decision trees are combined to provide a Gini importance for a particular feature in the random forest. [19,22].

For n classification, the features with the top fifteen Gini importance scores are included in Table 2. Some of these features (e.g. DNA template) are native to the data set, while others (e.g. Locus-derived Pr(n)1-4+, Pr(allele-sharing), and Pr(dropout)0-2+) are derived from precursor learning algorithms.

2.5. Data partitioning

All data sets were partitioned into stratified, non-overlapping subsets: the training sets, comprised of approximately 90% or less of all

Table 2
Top features used to construct the PACE machine learning algorithm for n classification.

Feature	Description
Pr(allele sharing)	Determined using a precursor learning algorithm
Sample wide peak count	Number of peaks observed across the sample
Template DNA amplified (ng)	Mass of DNA amplified
Locus-specific peak count	Number of peaks at each locus
Minimum number of contributors (MAC) – Pre-trimming	Allele counting prior to artifact removal
Estimated n	Sample-wide MAC
Locus-specific estimated n	MAC calculated per locus
Locus-derived Pr(n)1-4+	Probability of class 1 - 4+ contributors derived using data from only the given locus
# peaks at max locus	# of peaks at locus with the highest number of peaks
Peak height	Peak height in RFU
Peak area	Area of each peak
Min : Max ratio	Ratio of sample-wide peak with maximum height to peak with minimum height
Locus dropout metric	Average number of peaks across sample at loci with signal above threshold
Pr(dropout)0-2+	Quantitative, probabilistic assessment of the level of allele dropout
Estimated dropout peaks	Estimated number of dropout in a sample

samples from the original data set, were used to build the learning algorithms; the testing sets, comprised of approximately 10% of all data, were used to assess model performance on previously unseen data. All machine learning algorithms were exposed to the training data to learn patterns, identify decision boundaries between classes, and construct predictive models, and all models were evaluated solely using testing data. Most hyperparameters were kept at default values for Scikit-learn [23]; in cases where hyperparameters were modified, cross-validation was used as per [4].

2.6. Model evaluation

The ultimate success of machine learning is largely dependent on having a sufficient sample size from which to train the algorithm. A learning curve - a plot of training and cross-validation accuracy as a function of the number of training data - can be used to assess the accuracy of a machine learning algorithm as well as indicate classification accuracy when the algorithm is faced with previously unseen data. Simply, the machine learning algorithm is constructed by iteratively increasing the training data set size and the resulting classification accuracy of the training and testing set is plotted. The quality of a learning curve is demonstrated by the convergence of the training and testing accuracy (indicating the level of model bias) and the overall accuracy (indicating the overall model variance).

Accuracy is the ratio of true (correct) events to the total number of predicted events (Eq.4).

$$accuracy = \frac{number\ of\ correctly\ predicted\ events}{number\ of\ predicted\ events} \quad (4)$$

Precision (Eq. 5) is a measure of confidence, also known as the positive predictive value. In the context of predicting whether or not there were four donors in a DNA mixture, for example, precision represents the proportion of correctly identified 4-contributor mixtures to the total number of predicted 4-contributor mixtures. A high precision indicates few false positives.

$$precision = \frac{number\ of\ correctly\ predicted\ positive\ events}{number\ of\ predicted\ postive\ events} \quad (5)$$

Recall (Eq. 6), also known as the true positive rate or sensitivity, represents the predicted rate of positive identification for the specific class. Continuing the example above, recall represents the proportion of correctly predicted 4-contributor mixtures to the total number of 4-contributor mixtures. A high recall indicates few false negatives.

$$recall = \frac{number\ of\ correctly\ predicted\ positive\ events}{number\ of\ postive\ events} \quad (6)$$

The F1 score (Eq. 7) is the harmonic mean of precision and recall. Generally, predictive systems with F1 values approaching one will display higher accuracy.

$$F1 = 2 \times \left(\frac{precision \times recall}{precision + recall} \right) \quad (7)$$

Informedness (Eq. 8) represents the level of confidence that the system has in predicting the class [24]. In this case, informedness represents the relative level of confidence the system has in accurately predicting the presence of a 4-contributor mixture.

$$informedness = recall + (inverse\ recall) - 1 \quad (8)$$

Learning curves, which plot prediction accuracy as training size is varied, were generated for all machine learning-derived models using 10-fold cross-validation and were plotted with a band of +/- one standard deviation from each point.

A comparison of PACE results with human analyst interpretation was not possible due to the time constraints associated with manual analyses of hundreds of samples.

2.7. PACE input

PACE input consists of the raw. fsa or. hid file from the capillary electrophoresis instrument, trace data comprised of the *x,y* data from the electropherogram (automatically exported) and requires the analyst to input the DNA template amount for the particular sample of interest. Note, the DNA template amount is used as a general means of characterization of the sample, e.g. general expectations of peak height magnitude. Therefore, PACE remains robust given the variations in quantitation values caused by the inherent variability in DNA quantitation systems. Individual samples are routinely analyzed in seconds [4] and can do so on standard office computers.

3. Results

3.1. Peak identification and artifact removal

Allelic peaks are traditionally identified through a two-pronged process involving separating signal from noise and assessing the presence of process-related artifacts such as stutter and pull-up. PACE-GF and PACE-PPF6c perform these tasks in an automated fashion, allowing for hands-free use of the software. The locus- and sample-specific analytical threshold with noise reduction (LSST-NR) [16] correctly identified over 95% of true allelic peaks, 95.9% in PACE-GF training data and 96.8% in PACE-PPF6c training data (Table 3). Similarly, stutter was correctly identified and removed in nearly 95% of GF data and over 96% of PPF6c data, while pull-up peaks were correctly removed in 96.4% of PPF6c data and 97.0% of GF data. Excess noise above threshold was further filtered using (1) the area divided by peak height, (2) the low minimum height to maximum peak height algorithm, and (3) the machine learning algorithm designed to identify and remove remaining noise from signal, with accuracies in the PACE-GF set of 94.1%, 93.5% and 94.1%, respectively and accuracies in the PACE-PPF6c set of 96.7%, 96.7% and 95.7% respectively. In this context “incorrect” calls refer to instances where the LSST-NR system removed an allelic peak or retained artifactual signal.

3.2. Feature selection and importance

The influence of each feature on the resulting classification was calculated using the Gini importance. The locus-derived probabilities for each potential number of contributors displayed the highest importance in both PACE-6c and PACE-GF. Several of the most information-rich features were shared among the PACE versions, albeit with different overall rankings in each version respectively – sample-wide peak count, the number of peaks at the locus with the maximum number of peaks, maximum allele count estimation (pre- and post-artifact removal), peak heights, peak areas and probability of allele sharing (Tables 4 and 5). As expected, the locus-derived *n* classification (Locus-derived Pr(*n*)) was the most critical feature for prediction with Gini values of 0.221 and 0.150 for PACE-PPF6c and PACE-GF, respectively. The sample-wide peak count was also ranked in the top three features for the PACE-PPF6c data set and

Table 3
Results of automated peak identification and artifact removal for both the PACE-GF and PACE PPF6c testing data sets. The testing data set was not used to build the various peak/artifact assessment models.

	PACE-GF			PACE-PPF6c		
	Correct	Incorrect	Accuracy	Correct	Incorrect	Accuracy
LSST-NR	104551	4526	0.959	2709	89	0.968
Stutter	86398	4789	0.947	6264	234	0.964
Pull-up	18606	694	0.964	1617	50	0.970
Excess noise	3759	234	0.941	145	5	0.967
Area-div height	7670	532	0.935	470	16	0.967
Low-min max	14891	928	0.941	1323	60	0.957

Table 4
The ten most informative features for the PACE-PPF6c *n* classifier, with features ranked using Gini importance. Feature importance measures marked with an asterisk (*) indicate the maximum feature importance from a feature group; e.g. Locus-derived Pr(*n*) = max (vWA-derived Pr(*n*) feature importance, FGA-derived Pr(*n*) feature importance…) and Pr(dropout 0-2+)= max(Pr(dropout of 0 peaks) feature importance, Pr(dropout of 1 peak) feature importance, Pr (dropout of 2+ peaks) feature importance).

PACE PPF6c	
Feature	Importance
Locus-derived Pr(*n*)*	0.221
Locus dropout metric	0.090
Sample-wide peak count	0.074
# peaks at max locus	0.037
Min contributors (pre-trimming)	0.030
Pr(allele sharing)*	0.030
Estimated *n*	0.025
Estimated dropout peaks	0.012
Template DNA amplified (ng)	0.005
Pr(dropout 0 -2+)*	0.005

Table 5
The ten most informative features for the PACE-GF *n* classifier, with features ranked using Gini importance. Feature importance measures marked with an asterisk (*) indicate the maximum feature importance from a feature group; e.g. Locus-derived Pr(*n*) = max(vWA-derived Pr(*n*) feature importance, FGA-derived Pr(*n*) feature importance…) and Pr(dropout 0-2+)= max(Pr(dropout of 0 peaks) feature importance, Pr(dropout of 1 peak) feature importance, Pr(dropout of 2+ peaks) feature importance).

PACE-GF	
Feature	Importance
Locus Pr(*n*)*	0.150
Pr(allele sharing)*	0.083
Locus dropout metric	0.069
Sample-wide peak count	0.064
# peaks at max locus	0.041
Estimated *n*	0.034
Min contributors (pre-trimming)	0.027
Peak area*	0.015
Estimated dropout peaks	0.012
Peak height (RFU)*	0.008

top four in the PACE-GF data set. The locus-specific dropout metric, a sample-wide peak count that considers the level of dropout, was also critical to the success of the classifiers.

3.3. Sample size sufficiency – learning curves

Learning curves, which plot prediction accuracy as training size is varied, were generated for PPF6c and GF machine learning-derived models using 10-fold cross validation and were plotted with a band of +/- one standard deviation from each point. Both the PACE-PPF6c (Fig. 1) and the PACE-GF (Fig. 2) learning curves show low bias (i.e. high accuracy rates) and low variance (i.e. convergence of training and testing set accuracies), with overall accuracy for both training and testing data above 99.5%.

3.4. Model evaluation

Model performance is shown via summary statistics in Tables 6 and 9. Each confusion matrix portrays the performance of a model using only testing data, with each instance in the data set being a DNA sample with one or more donors. Summary metrics were calculated using Eqs.

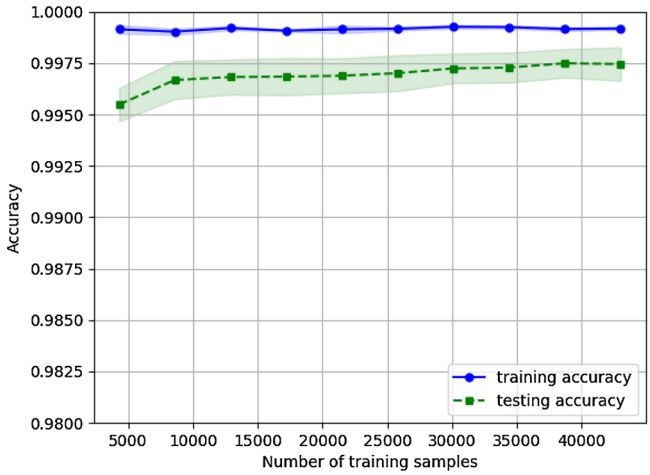

Fig. 1. Learning curve assessing the PACE-PPF6c random forest machine learning algorithm for n classification.

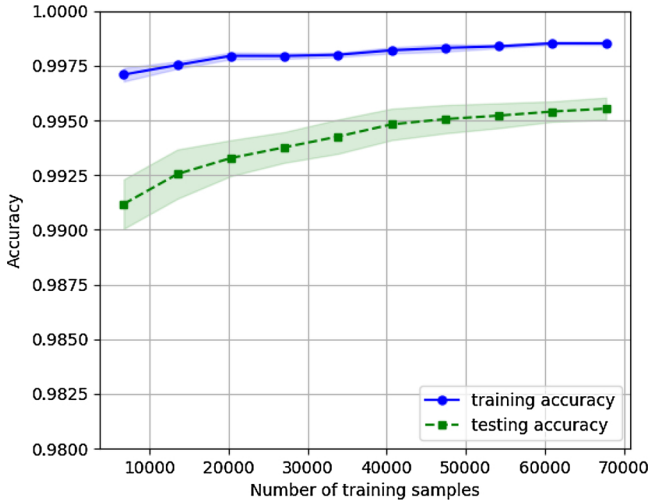

Fig. 2. Learning curve assessing the PACE-GF random forest machine learning algorithm for n classification.

4–7 respectively. Note that correct/incorrect calls were determined from the maximum probability of classes 1 through 4 + . If, for example, a sample's probabilities were estimated to be Pr(n = 1) = 0, Pr (n = 2) = 0.8, Pr(n = 3) = 0.2, and Pr(n = 4+) = 0, the class with the maximum probability would be n = 2, and the predictive model's "answer" would be 2 contributors. The unweighted mean F1 score for GlobalFiler data was 0.914; the same score for 6c data was 0.892. The minimum observed recall from both models was for the 3-contributor class for 6c (0.743); no other recall was below 0.85.

PACE-GF correctly classified *n* with an accuracy of 92.7% across classes. This includes correct classifications for many mixtures with high levels of allele dropout, allele sharing and high order mixtures (see Supplementary Information, tables 1S and 2S). The remaining 57 samples (7.3%) were misclassified based on the maximum probability returned by PACE-GF (Table 7). Approximately 77% (44/57) of the samples returned maximum probabilities that were within 0.43 of the second largest class probability (Fig. 3), with the correct class having a probability as high as 0.49, e.g. sample 1887. The average percent of allele dropout and allele sharing observed in the misclassified samples (24.2% and 60.2%, respectively) are elevated compared to the correctly classified samples (11.5% and 44.9%, respectively). Similarly, the distribution of the template amounts of the misclassified samples (ranging from 0.03 ng to 0.75 ng and a median of 0.165 ng) trend lower than the correctly classified samples (Fig. 4). Only two misclassified samples

Table 6
PACE-GF model summary statistics.

		Predicted *n*								
		1	2	3	4+	Accuracy	Precision	Recall	Informedness	F1
Actual *n*	1	271	0	0	0	0.985	0.958	1.000	0.977	0.978
	2	9	137	4	1	0.966	0.913	0.907	0.887	0.910
	3	3	10	119	8	0.944	0.838	0.850	0.814	0.844
	4+	0	3	19	201	0.961	0.957	0.901	0.885	0.928

Table 7
PACE-GF misclassified samples.

Sample ID	Actual *n*	Expected mixture ratio	Template DNA amplified (ng)	PACE predicted *n*	Class probability				% dropout alleles	Mean % allele sharing
					1	2	3	4+		
1263	2	1:2	0.045	1	0.70	0.30	0.00	0.00	35.3%	45.2%
1413	2	1:4	0.075	3	0.00	0.19	0.47	0.34	15.9%	41.7%
1426	2	1:2	0.375	4	0.00	0.25	0.30	0.45	5.9%	45.2%
1445	2	1:1	0.126	1	0.60	0.28	0.03	0.09	47.8%	47.6%
1452	2	1:2	0.045	1	0.68	0.31	0.00	0.00	35.3%	45.2%
1465	2	1:1	0.03	1	0.98	0.02	0.00	0.00	64.2%	47.6%
1472	2	1:2	0.189	3	0.00	0.35	0.62	0.03	1.5%	45.2%
1484	3	1:9:1	0.165	2	0.00	0.59	0.36	0.05	18.1%	62.7%
1510	3	1:2:2	0.155	4	0.00	0.00	0.27	0.73	9.5%	63.5%
1514	3	1:9:1	0.165	2	0.00	0.54	0.46	0.00	3.6%	62.7%
1570	3	1:2:1	0.06	2	0.00	0.63	0.18	0.20	31.8%	62.7%
1604	3	1:9:1	0.341	2	0.00	0.59	0.35	0.06	18.1%	62.7%
1620	3	1:9:1	0.693	2	0.00	0.55	0.45	0.00	1.2%	62.7%
1630	3	1:2:2	0.155	4	0.00	0.04	0.42	0.54	11.9%	63.5%
1634	3	1:2:1	0.06	1	0.70	0.30	0.00	0.00	58.8%	62.7%
1659	3	1:9:1	0.5	2	0.00	0.63	0.33	0.05	19.3%	62.7%
1734	5	1:4:4:4:1	0.21	3	0.00	0.06	0.75	0.18	16.2%	80.5%
1744	4	1:2:2:1	0.186	3	0.00	0.03	0.57	0.40	25.8%	72.6%
1765	2	1:4	0.075	1	0.46	0.44	0.00	0.10	52.2%	47.6%
1773	3	1:4:4	0.135	2	0.34	0.31	0.06	0.29	51.1%	57.9%
1790	2	1:1	0.03	1	0.90	0.10	0.00	0.00	64.7%	45.2%
1802	2	1:4	0.315	1	0.71	0.29	0.00	0.00	28.4%	47.6%
1875	3	1:4:1	0.09	2	0.00	0.77	0.17	0.06	43.3%	57.9%
1887	3	1:1:1	0.189	4	0.00	0.00	0.49	0.51	3.4%	55.6%
1900	3	1:1:1	0.045	2	0.05	0.66	0.24	0.05	37.5%	55.6%
1922	2	1:1	0.03	1	0.75	0.24	0.00	0.00	44.1%	45.2%
1939	3	1:4:1	0.09	4	0.00	0.00	0.39	0.61	8.9%	57.9%
2149	3	1:4:1	0.568	4	0.00	0.00	0.48	0.52	4.4%	57.9%
2175	3	1:4:1	0.09	4	0.14	0.11	0.16	0.59	53.3%	57.9%
2209	3	1:1:1	0.045	1	0.79	0.21	0.00	0.00	51.1%	55.6%
2229	3	1:4:1	0.378	4	0.00	0.00	0.44	0.56	2.2%	57.9%
2432	4	1:2:2:1	0.186	3	0.00	0.01	0.62	0.38	13.4%	72.6%
2439	4	1:1:4:1	0.105	3	0.00	0.00	0.62	0.37	21.4%	70.8%
2482	5	1:4:4:4:1	0.434	3	0.00	0.02	0.55	0.43	18.1%	80.5%
2492	5	1:4:4:4:1	0.21	3	0.00	0.03	0.49	0.47	22.9%	80.5%
2512	4	1:2:2:1	0.09	2	0.01	0.60	0.10	0.29	42.3%	72.6%
2536	4	1:1:1:1	0.06	3	0.00	0.17	0.42	0.41	26.3%	70.8%
2584	4	1:1:9:1	0.18	3	0.00	0.00	0.56	0.44	11.8%	67.9%
2597	4	1:4:4:1	0.15	3	0.00	0.00	0.70	0.30	20.2%	65.5%
2667	4	1:9:9:1	0.3	3	0.00	0.00	0.56	0.44	9.9%	67.9%
2677	4	1:9:9:1	0.75	3	0.00	0.02	0.67	0.32	7.9%	67.9%
2697	4	1:9:9:1	0.5	3	0.00	0.00	0.57	0.43	7.9%	67.9%
2757	4	1:9:9:1	0.75	3	0.00	0.06	0.56	0.39	20.8%	67.9%
2767	4	1:9:9:1	0.62	3	0.00	0.07	0.54	0.39	19.8%	67.9%
2887	3	1:1:1	0.045	2	0.00	0.72	0.24	0.04	19.3%	55.6%
2979	4	1:1:1:1	0.06	2	0.00	0.43	0.26	0.31	31.3%	70.8%
3025	4	1:1:1:1	0.124	3	0.00	0.03	0.52	0.45	8.1%	70.8%
3090	5	1:1:4:1:1	0.12	3	0.00	0.00	0.52	0.48	31.3%	75.7%
3123	5	1:1:1:1:1	0.075	3	0.00	0.05	0.57	0.38	36.8%	72.9%
3296	5	1:1:2:9:1	0.21	2	0.00	0.88	0.05	0.07	43.8%	76.2%
3372	2	1:25	0.3	1	0.60	0.40	0.00	0.00	30.4%	42.9%
3379	2	1:2	0.3	3	0.00	0.21	0.69	0.10	1.4%	34.5%
3439	2	1:3	0.6	3	0.00	0.11	0.89	0.00	0.0%	29.8%
3637	3	1:2:3	0.2	4	0.00	0.00	0.47	0.53	1.1%	54.8%
3702	3	3:1:1	0.2	2	0.00	0.91	0.09	0.00	29.2%	57.1%
3712	4	10:10:1:1	0.2	3	0.00	0.00	0.97	0.03	19.4%	66.7%
3755	4	20:10:1:1	0.1	3	0.00	0.09	0.65	0.26	19.4%	68.5%

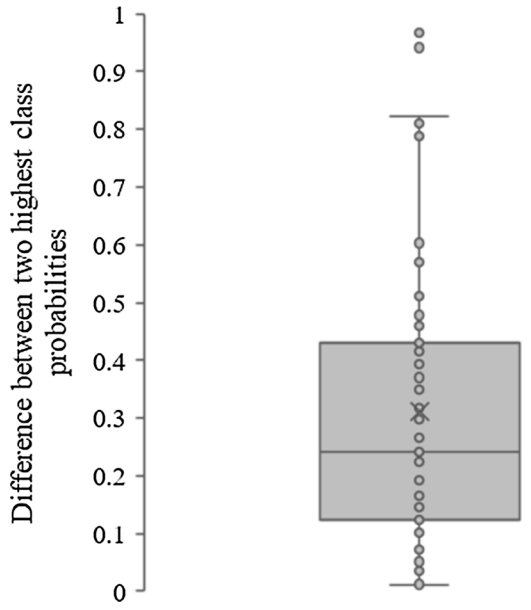

Fig. 3. PACE-GF - Distribution of the differences between the maximum class probability and the second largest class probability among the misclassified samples.

exhibited maximum probabilities of greater than 0.85 (1465 and 3712). Sample 1465 is a 2-contributor sample (1:1) amplified using 0.03 ng of DNA, which leads to approximately 50% of the expected alleles to dropout. Sample 3712, a 4-contributor sample, exhibits a high level of allele sharing (66.7%), a moderate level of dropout (19.4%) and disparate proportions of the two majors and two minors (10:10:1:1).

PACE-GF classification resulted in seven n calls that over- or underestimates by two contributors (13% of the incorrect samples) (Table 8), a high degree of allele sharing and/or allelic dropout accounted for nearly all of these instances. The misclassified samples exhibited high levels of allele sharing among contributors (45.2%–76.2%). The misclassified 3- and 4- contributor samples exhibited an average of 65% of the alleles being shared among the contributors and high levels of allele dropout, ranging from 29.9% to 58.8% of all expected alleles. Two of the 4+ contributor samples (2512 and 2979) and the 5- contributor sample (3296) had over 70.0% of alleles being shared (Table 8). Sample 1773, a 3 contributor sample classified as a single source, exhibited high levels of allele dropout (51.1% or 45 of 88 alleles) and approximately 58% of the alleles being shared. This, combined with a relatively low level of DNA template (0.135 ng), led to a probabilities of 0.34 for the 1-contributor class and 0.314 for the 2-contributor class, indicating there is underlying complexity where caution should be exercised during interpretation. Sample 1426, a 2-contributor sample classified as 4+-contributor (Supplementary information, Fig.1S), was a distinct misclassification; it was amplified using 0.375 ng of template DNA but was subjected to DNase I degradation (24mU enzyme) [3]. This led to low peak heights, with 57.8% of the peaks falling below 100RFU and peak height imbalance that typically indicates greater than two contributors. The LSST thresholding method did not remove any true alleles but, due to the noisy, low-level baseline, erroneously detected three non-allelic peaks at three loci. These peaks are not morphologically consistent with allelic peaks and thus an analyst, when interpreting the electropherogram, would not include them in the final interpretation. PACE output assigned probabilities of 0.25, 0.30, and 0.45 to the 2-, 3-, and 4+ contributor classes, respectively. Therefore, moderate weight was assigned to the correct class. The distribution of the probabilities across classes would indicate that the sample has underlying complexity, which, upon manual inspection, can be confirmed. Manual interpretation would further indicate that the low-level nature of the sample may not be appropriate for further analyses. Sample 1426 therefore represents a prime example of the value in using PACE as a tool to assist the analyst. Manual analysis would likely lead an analyst classifying this sample as a 2- or, potentially, a 3-contributor sample.

PACE-PPF6c class-specific accuracies are all over 90%, with 3- and 4+ contributor sample accuracy at 94.9% and 97.5%, respectively (Table 9). Nineteen samples were misclassified based on the maximum probability returned (Table 10); 58% (11/19) of the incorrect samples returned maximum probabilities within 0.48 of the second largest class probability (Fig. 5). The system overestimated *n* in five of 19 incorrectly classified samples, and only returned two samples that were over- and underestimated by more than 1 contributor. The misclassified samples exhibit higher average levels of allele sharing among contributors than the correctly called sample, 0.46 and 0.39, respectively (Fig. 6). Misclassified samples also displayed a 3-fold increase in the average proportion of allele dropout compared to the correctly classified counterparts, 0.22 and 0.075, respectively.

PACE-PPF6c resulted in correct calls (based on the maximum probability) in 90% of the testing set samples. Two samples exhibited a maximum probability for *n* that was ±1 contributor compared to the expected (Table 11).

Sample 951, a 2-contributor sample misclassified as a 4+ contributor was assigned class specific probabilities of 0.0, 0.403, 0.194

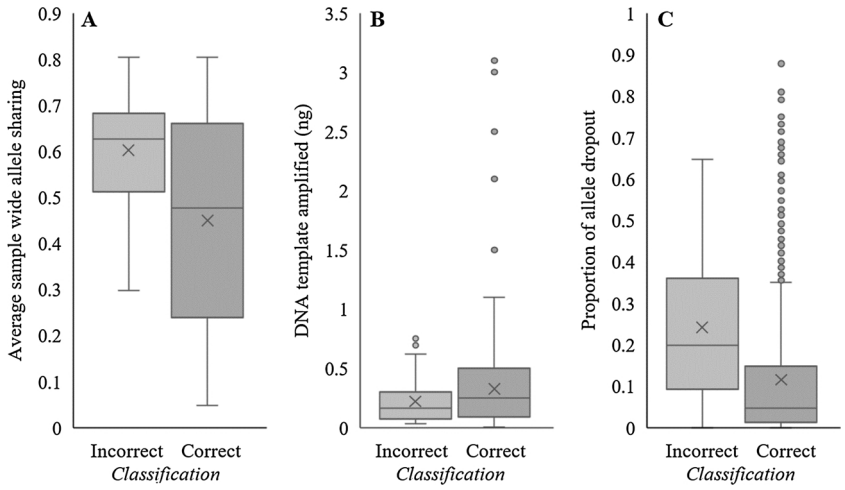

Fig. 4. PACE-GF - Distribution of the proportion of allele sharing (A), DNA template amount (B) and proportion of allele dropout (C) in the misclassified and correctly classified samples.

Table 8
A summary of PACE-GF incorrect calls that were over or underestimated by two contributors.

Sample identifier	Actual *n*	Expected mixture ratio	Template DNA amplified (ng)	Predicted *n*	Class probability				% dropout alleles	Sample-wide mean allele sharing
					1	2	3	4+		
1426	2	1:2	0.375	4	0.00	0.25	0.30	0.45	5.9%	45.2%
1634	3	1:2:1	0.06	1	0.70	0.30	0.00	0.00	58.8%	62.7%
1773	3	1:4:4	0.135	1	0.34	0.31	0.06	0.29	51.1%	57.9%
2209	3	1:1:1	0.045	1	0.75	0.25	0.00	0.00	51.1%	55.6%
2512	4	1:2:2:1	0.09	2	0.01	0.60	0.10	0.29	42.3%	72.6%
2979	4	1:1:1:1	0.06	2	0.00	0.43	0.26	0.31	31.3%	70.8%
3296	5	1:1:2:9:1	0.21	2	0.00	0.85	0.06	0.09	43.8%	76.2%

and 0.404, for 1-, 2-, 3- and 4+ contributor classes, respectively. Thus, output provided nearly equal weights to the correct and incorrectly identified classes. This sample exhibited high levels of baseline noise above threshold in all channels and lead to an increased number of erroneous allele calls (Supplementary Information Fig. 2S). In practice, this sample would be re-injected or re-amplified.

Sample 1198, a 4-contributor sample misclassified as a 2-contributor sample, exhibits high levels of allele-sharing (63.0%), with approximately 25% of all alleles dropping out. PACE artifact management and thresholding identified 92 of 93 alleles that were present, only removing a single allele at 22RFU. This misclassification was conservative, and appropriate, as the sample had a total of 0.0625 ng of DNA across 4-contributors at a 4:3:2:1 ratio. The two lowest minor contributors would be ideally expected to contribute a total of approximately 18 pg of DNA total, the same amount of DNA expected from approximately three diploid cells.

3.5. Probability thresholds

The robustness of each system was evaluated by assessing the accuracy of correct *n* calls as a function of the maximum probability. In this context, the maximum probability refers to the maximum probability returned by the system across all classes (1-, 2-, 3-, or 4+ contributors). At a maximum probability of 0.95 or greater, the PACE-PPF6c system is 98.3% accurate, while the PACE-GF system has an accuracy of 99.6% (Table 12). As the maximum probabilities decrease from 0.95 to 0.9 and to 0.8, we observe accuracies reach a minimum of approximately 96% for the PACE-PPF6c system and 99.0% for the PACE-GF system.

3.6. Assessing profile complexity

A detailed performance of PACE, both in predicting *n* and assessing profile complexity, is outlined in the following examples from the PACE-GF testing set. Sample 2624 (Fig. 7, 4-contributors, 9:1:1:1, 0.18 ng, degraded-12mU DNase I), has limited indications of a fourth contributor (only peak height imbalance). In this case, the sample experienced dropout of nine alleles and elevated levels of allele sharing (67.9%). PACE correctly classified this as a 4+ contributor sample (Pr(4+) = 0.82), while giving substantially less weight to a 3-contributor class (Pr(3) = 0.18). Sample 3201 (Fig. 8, 5-contributor, 9:2:1:1:1, 0.21 ng, untreated) indicates a potential fourth contributor at D19S433 (7 peaks) and SE33; however, many of these peaks fall in a-4 or a+4 position and could be interpreted as elevated stutter. The system correctly classified this sample as a 4+ contributor (Pr(4+) = 0.958). In this way, PACE can be used to assess the underlying complexity of a profile. This is exemplified by sample 2505 (Fig. 9, 5-contributor, 1:1:1:1:1, 0.075 ng, degraded- treated with 12mU DNase I), wherein PACE yields probabilities of Pr(1) = 0.11, Pr(2) = 0.38, Pr(3) = 0.07 and Pr(4+) = 0.43. This sample experienced widespread stochastic effects which led to the distribution of probabilities across the classes. Although PACE classifies the sample correctly, the class-based distribution clearly shows increased levels of uncertainty, and thus indicates the presence of underlying complexity. This interpretation would also likely be shared by the analyst and, when used in combination, would strengthen the analyst's conclusion. Fewer examples of PACE-PPF6c are available due to the confidentiality of the donor genotypes (based on the policy of the collaborating laboratory. Select PACE-PPF6c samples are included in the Supplementary Information (Fig. 3S and 4S). The overall PACE-PPF6c data set includes complex samples such as those in the PACE-GF sample set, where similar accuracies are achieved.

3.7. Repeatability

Because the eventual output from PACE's machine learning is a static, predictive model, results from PACE will always be similarly static. The testing data sets for both PACE-GF and PACE-PPF6c were run in duplicate to demonstrate the repeatability of results, i.e., the ability to obtain identical probabilities when two instances of the software are run on the same samples. The entirety of the output, for both PACE systems, including the class-specific *n* probabilities, were identical.

4. Discussion

PACE-GF and PACE-PPF6c both display high accuracy when faced with a multitude of sample types, including those that exhibit high levels of dropout and allele sharing. As expected, the overall accuracy of the system decreases as the number of contributors increase, reflecting the increase in the overall complexity of the sample. The current systems perform at this level with fully-automated artifact identification,

Table 9
PACE-PPF6c model summary statistics.

		Predicted *n*								
		1	2	3	4+	Accuracy	Precision	Recall	Informedness	F1
Actual *n*	1	58	2	0	0	0.965	0.921	0.967	0.930	0.943
	2	5	73	0	1	0.919	0.880	0.924	0.840	0.901
	3	0	7	26	2	0.949	0.963	0.743	0.737	0.839
	4+	0	1	1	22	0.975	0.880	0.917	0.899	0.898

Table 10
PACE-PPF6c misclassified samples. Note, the mixtures with a not applicable (NA) label in the expected mixture ratio column were differentially extracted samples where carryover occurred.

Sample ID	Actual *n*	Expected mixture ratio	Template DNA amplified (ng)	PACE predicted *n*	Class probability				% dropout alleles	Mean % allele sharing
					1	2	3	4+		
26	3	1:2:3	0.25	2	0.00	0.58	0.42	0.00	10.8%	63.0%
48	4	12:7:2:2	0.25	3	0.00	0.33	0.44	0.22	19.6%	68.5%
374	1	1:0	0.0528	2	0.23	0.75	0.02	0.00	2.4%	21.7%
411	2	NA	0.48	1	0.63	0.37	0.00	0.00	40.8%	43.5%
417	2	NA	1	1	0.99	0.01	0.00	0.00	36.0%	44.6%
419	2	NA	0.9975	1	1.00	0.00	0.00	0.00	32.9%	43.5%
433	3	NA	0.512	2	0.00	0.95	0.00	0.05	13.1%	58.0%
951	2	2:1	0.0375	4	0.00	0.40	0.19	0.40	27.7%	29.3%
1009	2	1:3	0.025	1	0.89	0.11	0.00	0.00	40.3%	56.5%
1076	3	1:10:1	0.075	2	0.04	0.83	0.13	0.00	36.4%	44.9%
1081	3	1:10:1	0.1	2	0.00	0.91	0.09	0.00	31.8%	44.9%
1086	3	1:2:3	0.0375	2	0.00	0.70	0.26	0.04	29.9%	44.9%
1099	3	1:6:1	0.2	2	0.00	0.70	0.22	0.08	13.1%	44.9%
1110	3	10:1:5	0.8	4	0.00	0.00	0.32	0.68	1.9%	44.9%
1122	3	2:3:1	0.6	4	0.00	0.00	0.29	0.71	0.9%	44.9%
1198	4	4:3:2:1	0.0625	2	0.01	0.68	0.15	0.17	25.2%	63.0%
1741	3	1:1:20	1	2	0.00	0.53	0.47	0.00	11.8%	55.1%
1808	2	NA	0.495	1	0.65	0.35	0.00	0.00	38.4%	48.9%
1829	1	NA	1	2	0.27	0.73	0.00	0.00	0.0%	13.04%

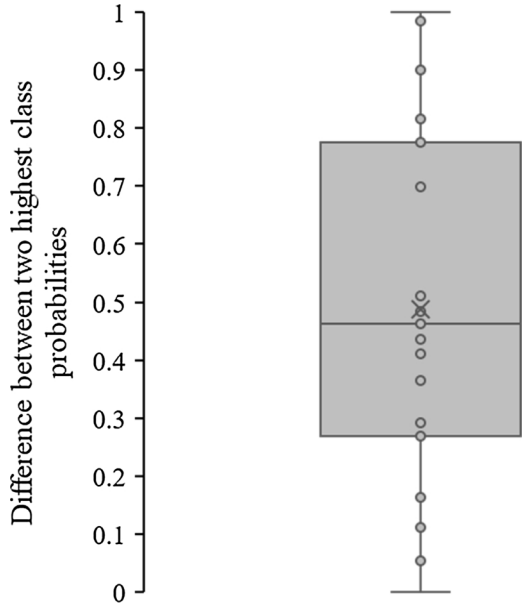

Fig. 5. PACE-PPF6c - Distribution of the differences between the maximum class probability and the second largest class probability among the misclassified samples.

in contrast to the PACE-Identifiler system described in [4], which required the analyst to manually remove all non-stutter artifacts. This manual preprocessing can lead to a substantial time commitment and is subjective, i.e. not reproducible from one analyst to the next. PACE artifact management and thresholding algorithms attempt to automate this process without a substantial reduction in the success of artifact removal or an increase in the erroneous removal of true alleles. Such automation, however, may result in infrequent over- or under-aggressive artifact removal. This is a result of the inherent process-based (stochastic) variation in sample processing. The *n* classification summary statistics were, expectedly, not as robust as were statistics associated with the previously designed system for use with samples amplified using the Identifiler amplification kit. However, the low incidence of erroneous classifications by the artifact management system can many times be recognized and corrected through manual interpretation. Ultimately, the inclusion of artifact management substantially improves reproducibility, decreases hands-on time and avoids the subjective nature of manual interpretation.

4.1. Classification results

The PACE-GF and PPF6c systems demonstrated correct n classifications in over 90% of the samples. This was achieved despite the presence of low quality/low quantity samples that exhibited high levels of allelic dropout, wide ranges of template amounts, and contributors with high levels of allele sharing, demonstrating the robust nature of the systems. Generally, the system experienced increased incorrect classifications as the contributor number increased, which is expected as the complexity increases accordingly. Overestimates resulted from inability of the internal artifact identification models to account for select noise and stochastic variation among stutter products. Underestimates of the n are primarily driven by the presence of elevated levels of allele sharing among contributors and stochastic effects, i.e. increased incidence of allele dropout. In many cases, locus-specific information such as the maximum allele count may indicate an additional contributor and such calls can be corrected through manual analyses. For example, a 3-contributor sample exhibiting high levels of dropout and overly aggressive artifact removal may be classified as a 2-contributor sample, despite observing five alleles at two loci. These peaks characterize the DNA profile after all identified artifacts have been removed, but because the accuracy of artifact removal is less than 100%, occasional non-allelic peaks may remain. The machine learning algorithm therefore "learns" that low-level peaks observed in conjunction with certain combinations of features are more likely to be artifacts; to continue the above example, for the loci with five peaks, the algorithm may determine that at least one of the peaks in each locus is likely to be an artifact and not a true allele. While most such cases are correctly evaluated by the learning algorithm's predictive model, occasionally the remaining low-level peaks are in fact alleles. These errors can be corrected through manual interpretation by an analyst, leading to a correct assessment of n.

Most samples that were erroneously classified were done so appropriately, where the presence of allelic dropout, low template levels, imbalanced mixture ratios and DNA degradation led to an under-representation of expected alleles in the electropherogram. However, in most cases PACE is able to detect low-level contributors, allowing the analyst to identify the true *n* even when stochastic effects are present. The analyst must weigh the level of sensitivity appropriately because it is possible that exceedingly low-level contributors do not influence the

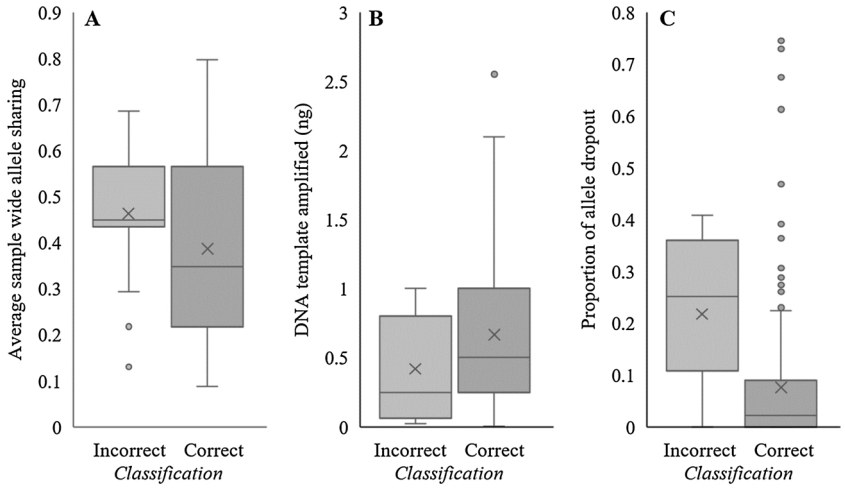

Fig. 6. *PACE-PPF6c* - Distribution of the proportion of allele sharing (A), DNA template amount (B) and proportion of allele dropout (C) in the misclassified and correctly classified samples.

overall interpretation of contributors above laboratory thresholds. Here we use the term effective number of contributors (n_e), which we define as the number of contributors that can be interpreted above laboratory thresholds and that contribute significantly to the overall profile complexity [25]. While n reflects that total number of donors in a sample regardless of the level of contribution, some contributors may be at significantly low levels and will not affect the overall interpretation as in mixture deconvolution or probabilistic genotyping. Under the current version of PACE, determining n_e requires analyst review in these situations; this function will be added to future versions of PACE without affecting the probabilistic output. Ultimately, regardless of whether n_e is used, the PACE systems' results are robust and reliable, mirroring the interpretation of an analyst thus demonstrating high accuracy in predicting n in complex samples.

4.2. Machine learning algorithm: random forest

While random forests were used for all machine learning for PACE, there are many other commonly used machine learning algorithms available. Random forests were specifically chosen for both theoretical and practical reasons. They often achieve high accuracy rates compared to other classifiers and tend to be resistant to overfitting when a large number of trees are used. These considerations are often trumped by those that are data-specific; often a researcher will not know which type of classifier works best until several are attempted. Random forests achieved high accuracy scores for PACE-6c and PACE-GF when compared to several other learning algorithms; only support vector machines were comparable, and these typically had worse accuracies as well as much longer learning periods and overall execution times, making them less practical for PACE model development.

Training and testing accuracies for both learning curves (Figs. 1 and 2) converged, with the difference in accuracies less than 0.002 for the PACE-GF model and less than 0.003 for the PACE-PPF6c model. The combination of high accuracies and low differences between training and testing accuracies are characteristic of low-bias, low-variance models, which in turn suggests that the models generalize well and do not exhibit overfitting. Converged training and testing accuracies do not guarantee a perfect coverage of a classification problem's feature space - for example, further reduction in model bias may be possible with a different or expanded set of features - but such an observation is a strong means of assessing comprehensive coverage of feature space in the absence of a drastic increase in sample size.

Summary metrics (Tables 6 and 8) were similar for both PACE-GF and PACE-PPF6c data. In both cases, model performance decreased as the number of contributors increased, with the exception of the final class (4+ contributors). This is not unexpected, as the last class does not represent a single number of contributors, i.e. represents any sample that has four or more contributors, and is therefore a less difficult group to classify. The dataset also does not contain mixtures with six or more donors; including these mixtures might well lead to decreases in model performance for the '4+' class due to the increased incidence of allele sharing and disparate mixture ratios in higher-order mixtures. This general trend of weaker predictions when faced with more complex mixtures has also been found previously [4–7]. The F1 score, which considers both false positives and false negatives and is often considered an appropriate balance between precision and recall, clearly demonstrates this pattern. Given the high accuracy in differentiating 3- and 4+ contributor samples, we believe that the system can be extended to 5/5+-contributor classification, given the appropriate training data is available.

While most of the precision and recall scores were very strong for both models, three results were less so: precision and recall for 3-contributor samples for PACE-GF data and recall for 3-contributor samples for PACE-PPF6c data. The first two results suggest a generally weaker model performance with 3-contributor samples. The third result, especially when paired with the high precision for 3-contributor samples, suggests that the 6c model disproportionately tends to under-classify 3-contributor samples when it predicts incorrectly. These misclassifications are often the result of extreme levels of allele-sharing, low DNA template levels and associated allele dropout. It is noteworthy that the 3-contributor class had slightly less favorable summary statistics than the 4+

Table 11
A summary of PACE-PPF6c incorrect calls that were over or underestimated by two contributors.

Sample identifier	Actual n	Expected mixture ratio	Template DNA amplified (ng)	Predicted n	Class probability				% dropout alleles	Sample-wide mean allele sharing
					1	2	3	4+		
951	2	2:1	0.0375	4	0.00	0.40	0.19	0.40	27.7%	29.3%
1198	4	4:3:2:1	0.0625	2	0.01	0.68	0.15	0.17	25.2%	63.0%

Table 12
The accuracy of correct n calls and the percentage of the total samples observed in each probability class as a function of the maximum probability output by the PACE-PPF6c and PACE-GF systems.

Maximum Class Probability	PACE-PPF6c		PACE-GF	
	% Correct	% of total samples	% Correct	% of total samples
0.95	98.3	60.1% (119/198)	99.6	58.9% (462/785)
0.9	97.1	68.1% (135/198)	99.3	67.7% (532/785)
0.8	96.1	74.2% (1547/198)	99.0	77.3% (607/785)

contributor class. This is due to the 4+ contributor class being a collection of both 4- and 5-contributor samples, and therefore requiring a less precise prediction to achieve an accurate result. An expanded set of 5-contributor samples would allow a model to be generated that explicitly permits differentiation between 4- and 5-contributor samples. A 5- contributor will be implemented in the future, at which time the summary statistics for 4-contributor samples would likely become equivalent to or slightly favorable than those for 3-contributor samples.

The PACE-PPF6c model was either similarly or slightly less predictive than the PACE-GF model across all classes. It is likely that the differences are the result of a combination of a smaller sample size (compared to the PACE-GF set) and differences between kit components leading to differences in, for example, heterozygous balance, stutter and baseline. The effect of a lower sample size, in this case, is likely stochastic in nature and does not indicate sample size insufficiency for downstream model generation. The learning curve for 6c data exhibited convergence, and thus indicates the PACE-PPF6c data set is of sufficient size for robust *n* classification.

4.3. Use and validation recommendations

The PACE system is an assist tool designed to help the analyst determine the *n* in a sample. In practice, the *n* of a sample is impossible to factually characterize, only evidence based assumptions can be made. PACE, when used to support the knowledge and experience of an analyst, will lead to more accurate assessments than either used in isolation. In practice, we suggest that the assessment of *n* begin with the analyst noting observations, as is performed routinely in forensic laboratories. Subsequently, the sample is run in PACE where the output is used to add weight to the decision-making process of the analyst. Although PACE is highly accurate, it is possible that PACE results contradict the analyst, e.g. an instance where the only indication of an additional contributor is the result of elevated stutter peaks. The analyst can then refer to the initial observations made by manual assessments and defend a decision that may contradict the PACE output.

This version of PACE can be used in three distinct, yet non-mutually exclusive manners: (1) a system to aid the analyst in identifying artifacts such as stutter, above-threshold noise and pull-up, (2) a means to determine profile complexity e.g. if a profile is suitable for interpretation and (3) a probabilistic assessment of the number of contributors in a DNA mixture. Our recommendations in this section will focus on the latter two.

Validation should focus on the probabilities that are the primary output of the PACE system. The output consists of four probabilities that correspond to the systems assessment for each of the four contributor classes 1 through 4 + . Thus, the sample set will require known samples containing 1–5 contributors. We suggest a minimum of 84–124 samples, approximately 20–30 per contributor class (1–4) and ten 5-contributor samples. The inclusion of 5-contributor samples is critical to ensure the laboratory is able to anticipate the performance of PACE when used to analyze samples containing greater than four contributors

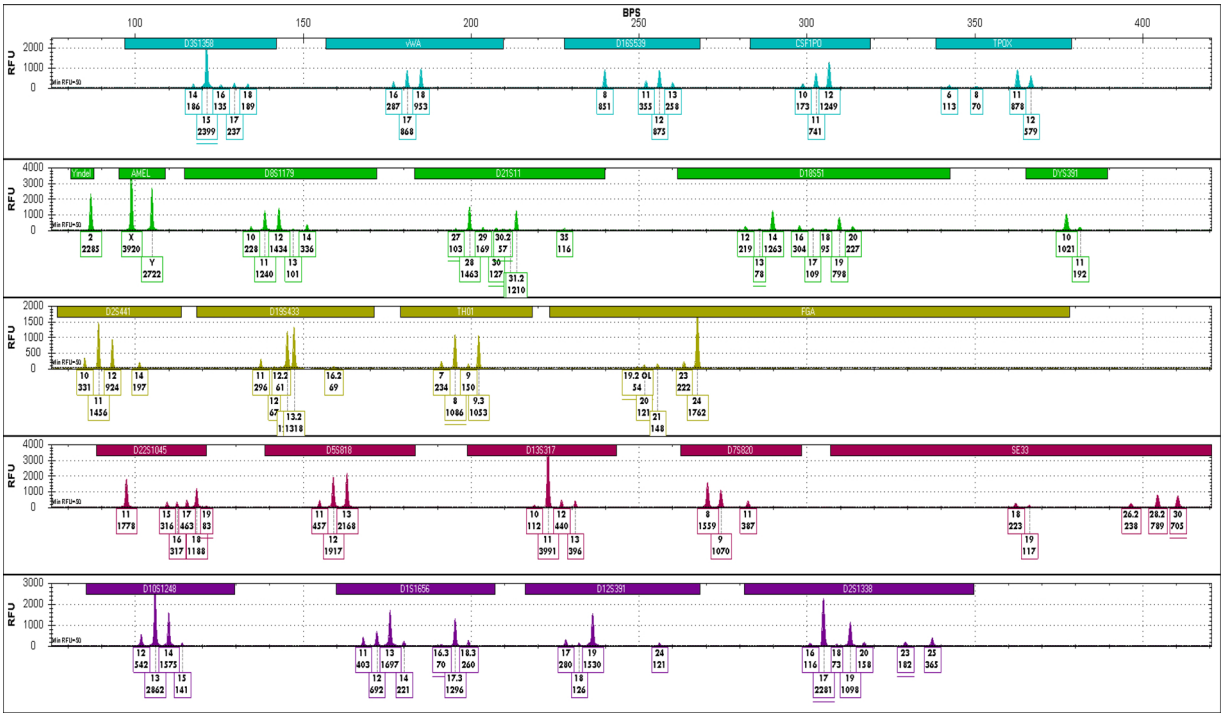

Fig. 7. PACE-GF sample 2624, (4-contributor, 9:1:1:1, 0.18 ng, degraded-12mU DNase I), analyzed at 25RFU for demonstrative purposes. PACE correctly classifies this sample with a Pr(4+) = 0.82 and Pr(3) = 0.18.

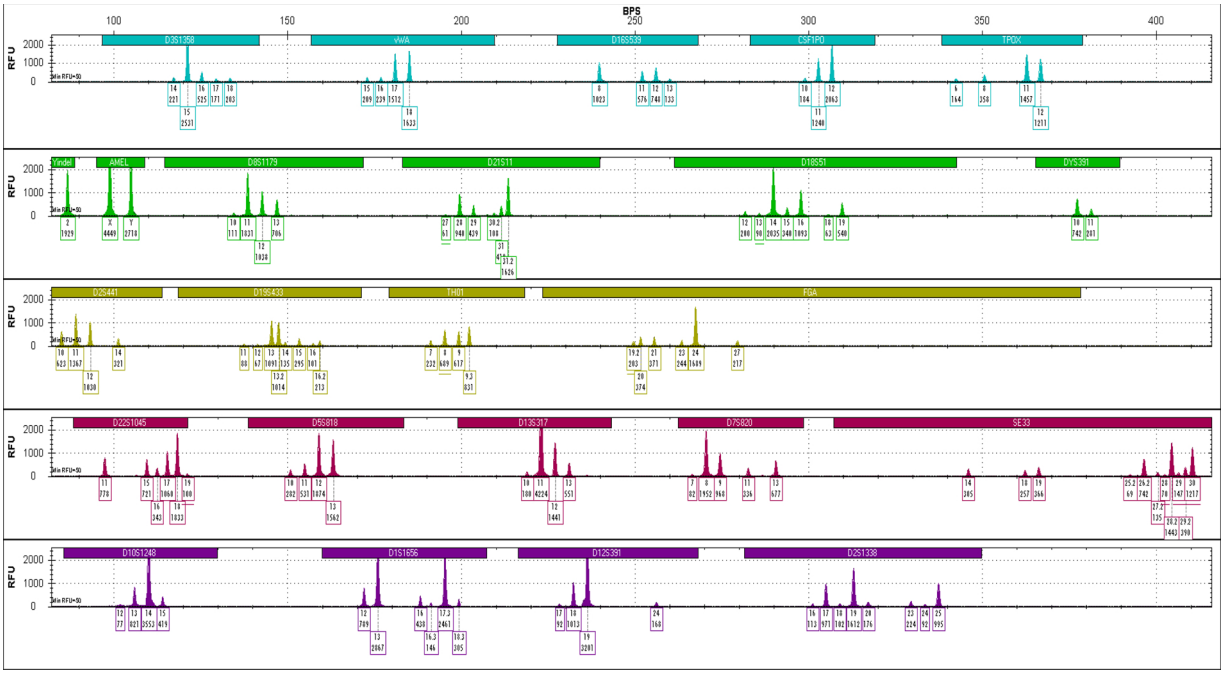

Fig. 8. PACE-GF sample 3201, (5-contributor, 9:2:1:1:1, 0.21 ng, untreated), analyzed at 25RFU for demonstrative purposes. PACE correctly classifies this sample with a Pr(4+) = 0.958. Note, most loci indicate three contributors with indications of a fourth contributor occur at SE33 and D19S433.

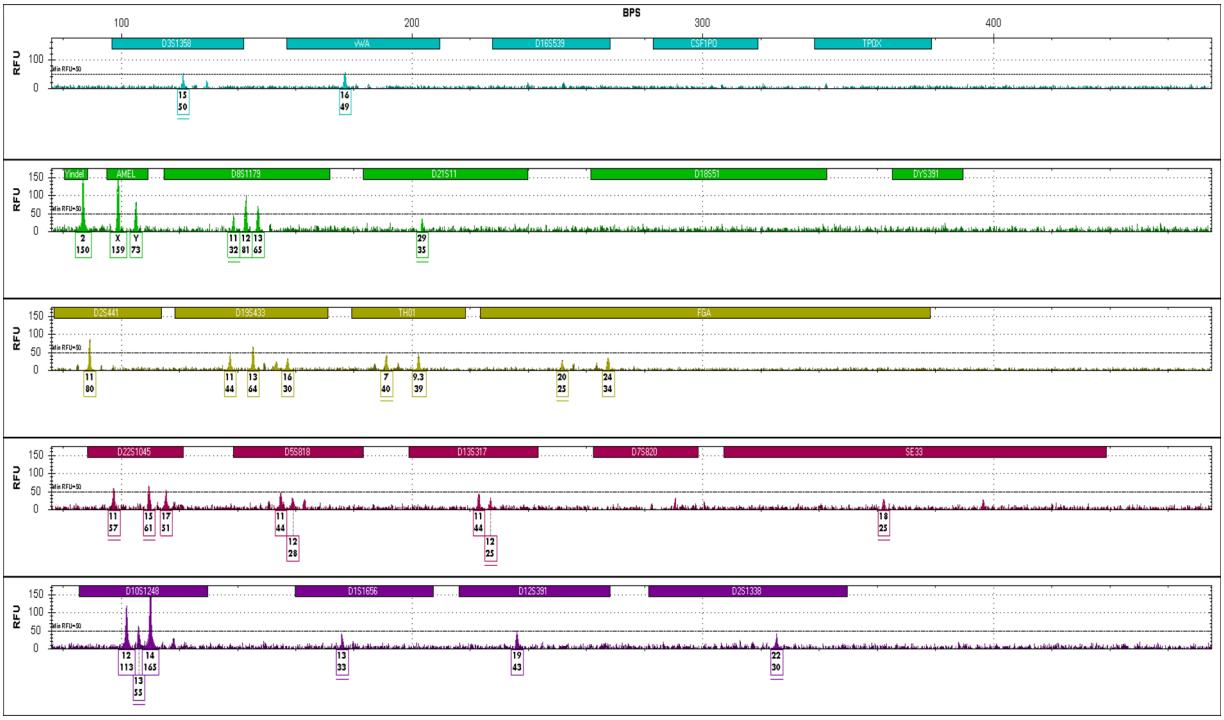

Fig. 9. PACE-GF sample 2505, (5-contributor, 1:1:1:1:1, 0.075 ng, degraded-12mU DNase I) analyzed at 25RFU for demonstrative purposes. PACE outputs probabilities of Pr(1) = 0.11, Pr(2) = 0.38, Pr(3) = 0.07 and Pr(4+) = 0.43. Despite, a correct classification, the distribution of the probabilities indicate a level of complexity that warrants caution when interpreting.

(Table 13). The samples should have varied template amounts (0.0156–1.0 ng, serially diluted) with varied ratios of contributors. The recommended sample set (Table 13) is provided as a guide; the laboratory should adjust the sample parameters based on various internal practices and protocols. This includes adjusting template amounts based on the laboratories' target amplification amount and mixture ratios to reflect those most commonly observed in casework.

Establishing an appropriate probability threshold for *n* or the assessment of sample interpretability should begin with the accuracy of correct calls given the PACE maximum probabilities in Table 12. We observed that when the class with the maximum probability is 0.8, correct calls were made in 96.1% and 99.0% of the instances in the PACE-PPF6c and PACE-GF data sets, respectively. Therefore, we recommend that a laboratory begin assessing the accuracy of the internal validation set in a similar manner. However, we also recommend that, when the maximum class probability is lower than the established

Table 13
Sample recommendations for internal validation.

n	Total number of samples	Final template (ng)	Ratios
1	14 - 21	0.0156 to 1	NA
2	20-30	0.0156 to 1	1:1 to 20:1
3	20-30	0.0156 to 1	1:1:1 to 20:1:1
4	20-30	0.0156 to 1	1:1:1:1 to 20:1:1:1
5	10	0.0156 to 1	1:1:1:1:1 to 20:1:1:1:1

threshold, it does not preclude the use of the result. In this case, agreement between PACE and the analyst is sufficient to allow a conclusion to be proffered. Ultimately, PACE output should be used to aid an analyst's judgement in determining *n*.

PACE functionality extends to the assessment of the interpretability/complexity of a sample, i.e. addressing questions such as: should a sample be analyzed and, if so, should it be analyzed in duplicate in PG systems using different values for *n*? The probability outputs may indicate that there is underlying complexity. For example, if there are 0.5 probabilities for two classes, $Pr(n=3) = 0.5$ and $Pr(n=4+) = 0.5$, and the sample may therefore be analyzed in PG software as a 3- and 4-contributor sample. Again, the lack of a high confidence PACE assessment does not preclude the sample's use for downstream analysis. Instances such as these may indicate that the sample should be analyzed in the PG software as a 3- and 4-contributor sample. Quality metrics of the PG system can then be used to further assess the quality of the interpretation.

5. Conclusion

This study presents a new version of the Probabilistic Assessment for Contributor Estimation (PACE), a fully continuous probabilistic method to identify artifacts and assess the number of contributors in DNA samples amplified using the GlobalFiler and the PowerPlex Fusion ® 6c human DNA amplification kits. This machine learning-based system is able to analyze samples in seconds using standard computing resources. The, now incorporated artifact management module is over 93.5% accurate in identifying true allelic peaks and removing artifacts such as high baseline noise, stutter, and pull-up. Similarly, the subsequent *n* prediction for both the GlobalFiler and PowerPlex Fusion 6c data sets yield correct classifications in over 90% of the samples, including those that exhibit low template DNA amounts and degraded or inhibited DNA. Generally, the system performs consistently as contributor number and complexity increases. The misclassified samples, the majority of which are underestimated by one contributor, are characterized by high levels of allele-sharing and low template amounts leading to high incidence of allelic dropout. This leads to less confidence in calls, just as it would during manual interpretation. However, in contrast to manual methods, PACE assigns weights to each probability class, allowing the analyst to assess the distribution of probabilities to aid in the decision-making process. The combination of PACE and manual interpretation is therefore more accurate and robust than either used in isolation.

Funding

This study was funded by NicheVision Forensics LLC.

Availability

The software is available commercially, inquires can be directed to Vic@NicheVision.com.

Declaration of Competing Interest

Syracuse University has licensed this intellectual property to NicheVision Forensics LLC.

Acknowledgements

The authors would like to extend special thanks to NicheVision Forensics LLC for supporting this project, including Luigi Armogida, Vic Meles, Brian Young and Tom Faris and the contributions of Ebrar Mohammed, Victoria Williamson, Angie Zhao and D. Spencer Eberst.

The authors would like to thank the following groups, without whom this would not be possible: Erie County Central Police Services, Office of the Chief Medical Examiner New York City (Department of Forensic Biology) the Connecticut Division of Forensic Services, Palm Beach County Sheriff's Office, Kansas City Police Crime Laboratory, Michigan State Police Forensic Science Division, Promega Corporation, Idaho State Police Forensic Services, Kentucky State Police Forensic Laboratory System, Washington DC Department of Forensic Sciences, Oakland Police Department, Indiana State Police Laboratory, San Diego County Sheriff's Department and Rutgers University, Onondaga County Center for Forensic Sciences.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.fsigen.2019.102140.

References

[1] J.-A. Bright, D. Taylor, C. McGovern, S. Cooper, L. Russell, D. Abarno, J. Buckleton, Developmental validation of STRmix™, expert software for the interpretation of forensic DNA profiles, Forensic Sci. Int. Genet. 23 (2016) 226–239, https://doi.org/10.1016/j.fsigen.2016.05.007.
[2] M.W. Perlin, M.M. Legler, C.E. Spencer, J.L. Smith, W.P. Allan, J.L. Belrose, B.W. Duceman, Validating TrueAllele® DNA mixture interpretation, J. Forensic Sci. 56 (2011) 1430–1447, https://doi.org/10.1111/j.1556-4029.2011.01859.x.
[3] 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 (2017) 308–316, https://doi.org/10.1111/1556-4029.13284.
[4] 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, https://doi.org/10.1016/j.fsigen.2016.11.006.
[5] 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, J. Forensic Sci. 50 (2005) 1361–1366, https://doi.org/10.1520/JFS2004475.
[6] J. Perez, A.A. Mitchell, N. Ducasse, J. Tamariz, T. Caragine, Estimating the number of contributors to two-, three-, and four-person mixtures containing DNA in high template and low template amounts, Croat. Med. J. 52 (2011) 314–326, https://doi.org/10.3325/cmj.2011.52.314.
[7] 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 (2007) 20–28, https://doi.org/10.1016/j.fsigen.2006.09.002.
[8] 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, https://doi.org/10.1016/J.LEGALMED.2018.02.001.
[9] C.C.G. Benschop, H. Haned, T.J.P. de Blaeij, A.J. Meulenbroek, T. Sijen, Assessment of mock cases involving complex low template DNA mixtures: a descriptive study, Forensic Sci, Int. Genet. 6 (2012) 697–707, https://doi.org/10.1016/J.FSIGEN.2012.04.007.
[10] 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, https://doi.org/10.1016/j.fsigen.2014.06.009.
[11] J.-A. Bright, R. Richards, M. Kruijver, H. Kelly, C. McGovern, A. Magee, A. McWhorter, A. Ciecko, B. Peck, C. Baumgartner, C. Buettner, S. McWilliams, C. McKenna, C. Gallacher, B. Mallinder, D. Wright, D. Johnson, D. Catella, E. Lien, C. O'Connor, G. Duncan, J. Bundy, J. Echard, J. Lowe, J. Stewart, K. Corrado, S. Gentile, M. Kaplan, M. Hassler, N. McDonald, P. Hulme, R.H. Oefelein, S. Montpetit, M. Strong, S. Noël, S. Malsom, S. Myers, S. Welti, T. Moretti, T. McMahon, T. Grill, T. Kalafut, M. Greer-Ritzheimer, V. Beamer, D.A. Taylor, J.S. Buckleton, Internal validation of STRmix™ – A multi laboratory response to PCAST, Forensic Sci. Int. Genet. 34 (2018) 11–24, https://doi.org/10.1016/J.FSIGEN.2018.01.003.
[12] 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 Sci. Int. Genet. 40 (2019) 150–159, https://doi.org/10.1016/j.fsigen.2019.02.017.
[13] 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, https://doi.org/10.1016/j.fsigen.2017.10.006.

[14] R.W. Cotton, C. Grgicak, M. Terrill, C. Word, S. Cortes, DNA - Additional information about the web application, Bost. Univ. Biomed. Forensic Sci. DNA Mix. (n.d.). http://www.bu.edu/dnamixtures/pages/help/introduction/. (accessed May 30 2019).

[15] R.M. Goor, L. Forman Neall, D. Hoffman, S.T. Sherry, A mathematical approach to the analysis of multiplex DNA profiles, Bull. Math. Biol. 73 (2011) 1909–1931, https://doi.org/10.1007/s11538-010-9598-0.

[16] 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, https://doi.org/10.1016/j.fsigen.2018.03.017.

[17] J.D. Adelman, A. Zhao, D.S. Eberst, M.A. Marciano, Automated detection and removal of capillary electrophoresis artifacts due to spectral overlap, Electrophoresis. (2019) elps.201900060, , https://doi.org/10.1002/elps.201900060.

[18] Y. Singh, P. Kumar Bhatia, O. Sangwan, A Review of studies on machine learning techniques, Omprakash Sangwan Int. J. Comput. Sci. Secur. (n.d.). https://www.cscjournals.org/manuscript/Journals/IJCSS/Volume1/Issue1/IJCSS-7.pdf (accessed June 30, 2018).

[19] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32, https://doi.org/10.1023/A:1010933404324.

[20] R.E. Bellman, Dynamic Programming, Princeton University Press, 1957 (accessed June 23, 2018), https://dl.acm.org/citation.cfm?id=862270.

[21] W.G. Touw, J.R. Bayjanov, L. Overmars, L. Backus, J. Boekhorst, M. Wels, A.F.T. Sacha van Hijum, Data mining in the life science swith random forest: A walk in the park or lost in the jungle? Brief. Bioinform. 14 (2013) 315–326, https://doi.org/10.1093/bib/bbs034.

[22] B.H. Menze, M. Kelm, R. Masuch, U. Himmelreich, P. Bachert, W. Petrich, F.A. Hamprecht, A Comparison of Random Forest and Its Gini Importance With Standard Chemometric Methods for the Feature Selection and Classification of Spectral Data, (2009), https://doi.org/10.1186/1471-2105-10-213.

[23] F. Pedregosa, G. Varoquaux, A. Gramfort, B. Thirion, O. Grisel, V. Dubourg, A. Passos, M. Brucher, É. Perrot, Matthieu Duchesnay, Scikit-learn: Machine Learning in Python, (2011) http://scikit-learn.sourceforge.net.

[24] D.M.W. POWERS, Evaluation: from precision, recall and F-Measure to roc, informedness, markedness & correlation, J. Mach. Learn. Technol. 2 (2011) 37–63 doi:10.1.1.214.9232.

[25] J.S. Buckleton, J.-A. Bright, K. Cheng, H. Kelly, D.A. Taylor, The effect of varying the number of contributors in the prosecution and alternate propositions, Forensic Sci. Int. Genet. 38 (2019) 225–231, https://doi.org/10.1016/J.FSIGEN.2018.11.011.