Supplementary Material 1: Development of a generic NOC tool

A generic model to estimate the number of contributors was developed using a train, test and hold-out dataset as presented in Table 1. This is the same data as used in [13], except that for the generic model only the 12 ESS (European Standard Set) and U.S. core loci were used as these are available in the most common commercial STR typing kits (FGA, TH01, VWA, D1S1656, D2S441, D3S1358, D8S1179, D10S1248, D12S391, D18S51, D21S11, D22S1045).

A total of 15 profile features were examined (Table 2). These features are independent of the peak heights as these can be influenced by the number of PCR cycles, the CE settings or amplicon length in the STR typing kit.

Alike in Benschop et al. [13], ten different algorithms (classifiers) were examined (Table 3).

Supplementary material Table 1. Overview of the numbers of DNA profiles (12 ESS and US core loci from PPF6C profiles) in the train, test and hold-out dataset.

Number of contributors	Train	Test	Hold-out	Total
1	99	33	33	165
2	56	20	20	96
3	74	25	25	124
4	78	27	27	132
5	43	15	15	73
Total	350	120	120	590
Number of different donors	695	273	273	1174

Supplementary material Table 2. Overview of the 15 sample features that were examined. Sample features take account of all 12 ESS (European Standard Set) and U.S. core loci available in the DNA profiles.

Number	Feature	Details
1	MAC	Maximum allele count (MAC); Maximum number of alleles observed on a locus
2	TAC	Total allele count (TAC); Total number of alleles per profile
3	Mean Allele count	Mean, median or standard deviation of the number of alleles per locus
4	Median Allele Count	
5	Standard Deviation Allele Count	
6	Minimum Allele Count	Minimum number of alleles observed per locus
7	Minimum NOC	Minimum number of contributors (NOC); Minimum Allele Count / 2, rounded up to integer
8	MAC method	Maximum Allele Count / 2, rounded up to integer
9	AC 0	Number of loci with an allele count of 0 (*i.e.* empty loci/ locus drop-outs), 1 or 2, 3 or 4, 5 or 6, 7 or 8, or 9 alleles or more.
10	AC 1-2	
11	AC 3-4	
12	AC 5-6	
13	AC 7-8	
14	AC ≥9	
15	Match probability	The probability of a random, unrelated person matching to this DNA profile. The probability is calculated using the allele frequencies of 2085 male Dutch individuals database.

Supplementary material Table 3. Overview of the ten algorithms used in this study.

Number	Classifier	Number	Classifier
1	Decision Tree Classifier (DTC)	6	Linear Support Vector Classification (LSVC)
2	Gaussian Naive Bayes (GaussianNB)	7	Logistic Regression Classifier (LRC)
3	Gradient Boosting Classifier (GBC)	8	Multi-layer Perceptron Classifier (MLPC)
4	k-Nearest Neighbors Classifier (k-NN)	9	Random Forest Classifier (RFC)
5	Linear Discriminant Analysis (LDA)	10	Support Vector Classification (SVC)

Model selection

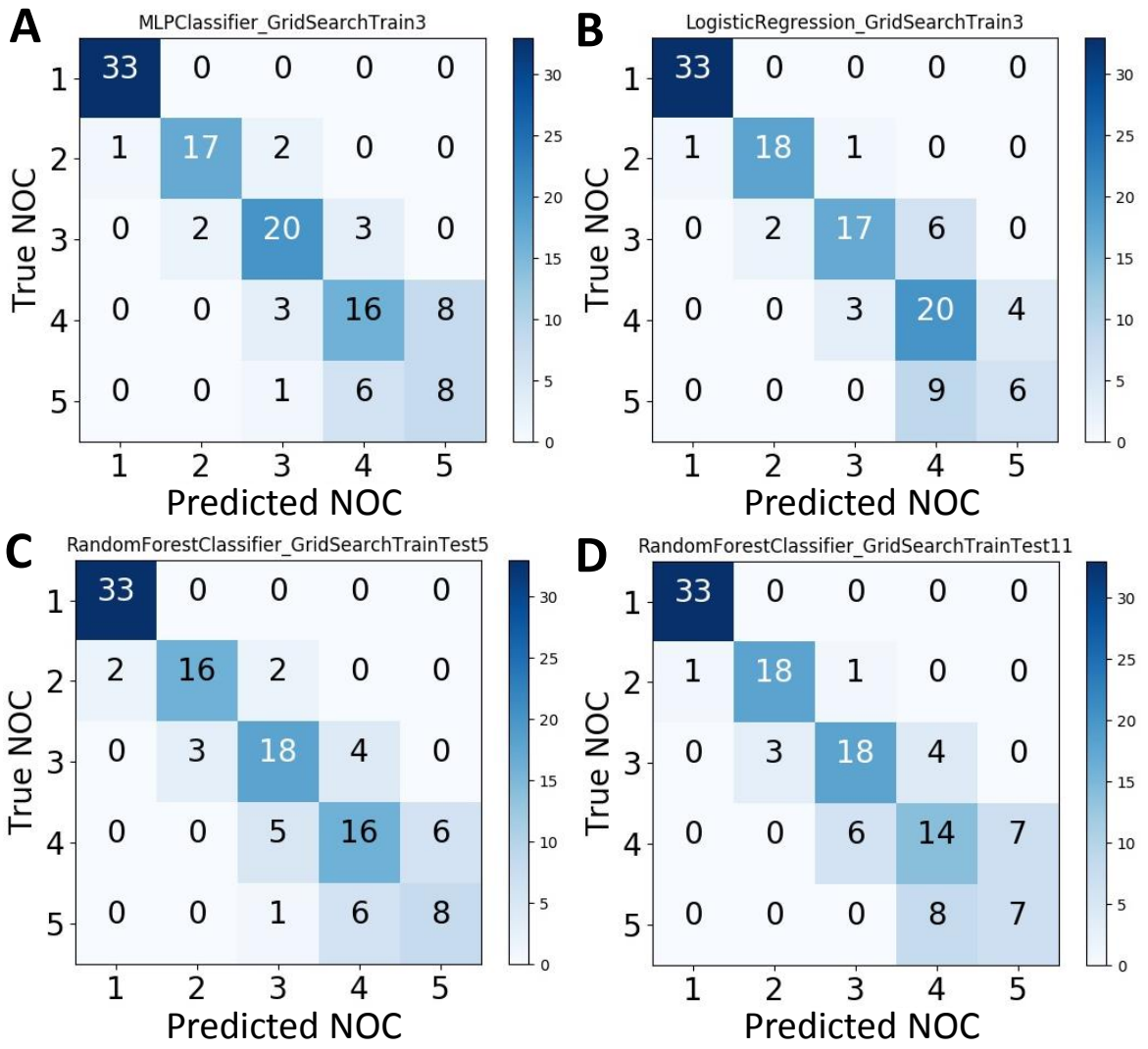
Every algorithm in combination with a varying number of features out of the total of 15 features were used to train and test model's performance. The highest percentage of correctly estimated NOC per algorithm is shown in Table 4 and the five best performing generic models are shown in Table 5. These five models show quite similar results, all yielded between 75% and 78% correctly estimated NOC when using the test set. The incorrect estimates deviated one NOC from the true NOC in all, but two occasions. That is, a five person mixture that was estimated as a three person mixture when using MLP3 or RFC5 (Fig. 1).

Supplementary material Table 4. Highest train and test accuracies per algorithm and number of features that were used to obtain these results.

Algorithm	Number of features	Train accuracy	Test accuracy
RFC	5	0.82	0.76
LDA	13	0.75	0.75
LRC	3	0.77	0.78
MLPC	3	0.81	0.78
GBC	3	0.80	0.78
SVC	9	0.80	0.78
LSVC	7	0.80	0.78
k-NN	7	0.78	0.77
DTC	13	0.73	0.71
GaussianNB	8	0.76	0.76

Supplementary material Table 5. Overview of the five generic models (algorithm + features) that yielded the highest accuracies on the training and test set.

	Train accuracy	Test accuracy
SVC 9	0.80	0.78
MLP 3	0.81	0.78
LR 3	0.77	0.78
RFC 5	0.82	0.76
RFC 11	0.82	0.75



Supplementary material Figure 1. Number of observations in categories for estimated (X-axes) and true (Y-axes) NOC in the test set using the generic models: A) MLP3, B) LR3, C) RFC5, and D) RFC11.

Alike with the PPF6C RFC19 model, every generic model presents a probability per estimated NOC. These probabilities appeared not informative when using the SVC9 model, as the NOC estimated by SVC9 was not always (40/590) the NOC that received the highest probability. Therefore, the SVC9 model was excluded from further examination.

The probabilities obtained using the four remaining models were examined in further detail using a selection of samples yielding an incorrect NOC estimate. To that end, samples yielding an incorrect estimate with all four models were excluded, as well as samples for which it is expected that the incorrect estimate will have no, or small, impact on the weight of evidence in forensic casework (i.e. two person mixtures estimated as three person mixtures, or three person mixtures estimated as four person mixtures).

Table 6 presents the number of samples that received an incorrect estimate for the NOC, that showed a probability difference of more than 0.1 when compared to the probability presented for the true NOC, and that could have a negative effect on computing the weight of evidence. A negative effect was defined as estimating a four person mixture as a five person mixture, since the maximum number of unknowns under Hd in DNAStatistX is four, which renders too many contributors to compute an LR. In addition, over- or underestimating the NOC by two was regarded more problematic than over- or underestimating the NOC by one.

Such negative effects were observed the least with the RFC11 model and this model had no incorrect estimates that differed more than one when compared to the true NOC (MLP3 and FRC5 did show such differences (Fig. 2)). Overall, RFC11 was preferred slightly over de other three models and was selected for further validation using the hold-out set. An overview of the 11 features as used in the RFC11 model is presented in Table 7.

Supplementary material Table 6. Overview of the number of incorrect predictions within the test set (A), those of which had a probability difference of >0.1 when compared to the probability presented for the true NOC (B) and those of which could possibly have a negative influence on computing a weight of evidence (WoE; C).

		Model			
		LR3	MLP3	RFC11	RFC5
A	Number of incorrect predictions	19	19	23	22
B	As A & probability difference to true NOC >0.1	11 (58%)	10 (53%)	7 (30%)	6 (27%)
C	As B & possibly problematic for WoE calculations	5 (26%)	7 (37%)	3 (13%)	5 (23%)

Supplementary material Table 7. Overview of the 11 sample features as used in the generic RFC11 model.

Nr.	Feature	Details
1	MAC	Maximum allele count (MAC); Maximum number of alleles observed on a locus
2	TAC	Total allele count (TAC); Total number of alleles per profile
3	Median Allele Count	Median of the number of alleles per locus
4	Standard Deviation Allele Count	Standard deviation of the number of alleles per locus
5	Minimum Allele Count	Minimum number of alleles observed per locus
6	Minimum NOC	Minimum number of contributors (NOC); Minimum Allele Count / 2, rounded up to 0 decimals
7	MAC method	Maximum Allele Count / 2, rounded up to 0 decimals
8	AC 0	Number of loci with an allele count of 0 (i.e. empty loci/ locus drop-outs)
9	AC 5-6	Number of loci with an allele count of 5 or 6 alleles.
10	AC ≥9	Number of loci with an allele count 9 alleles or more.
11	Match probability	The probability of a random, unrelated person matching to this DNA profile. The probability is calculated using the allele frequencies of 2085 male Dutch individuals database.

Performance of the generic RFC11 model

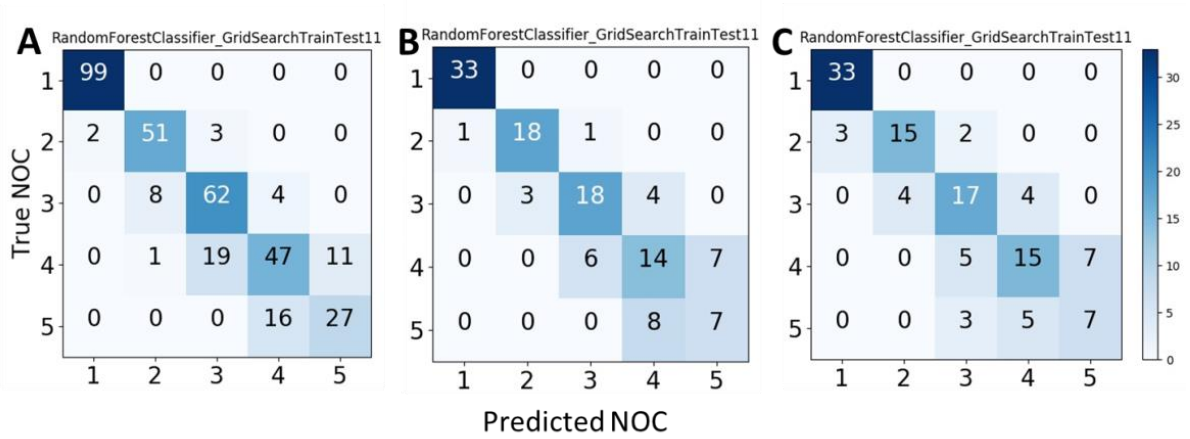
The generic RFC11 model was applied to the hold-out set and resulted in 33 samples with and incorrect estimate for the NOC (72.5% correct) in comparison to 30 when using the test set (75% correct) (Fig. 2). For 82% (27/33) of these incorrect predictions, the second highest probability was presented for the true NOC.

Next, the performance of the RFC11 model was compared to the PPF6C RFC19 model and to the MAC approach using a set of 120 PPF6C profiles from [15]. As expected, best results were obtained when using the RFC19 model (85% correct, Table 8) as this model is more specific to the data; it uses all loci, and more of the available information. The MAC approach yielded more correct predictions for three- and four-person mixtures when compared to RFC11. However, overall, the RFC11 model showed 10% higher correct estimates when compared to the MAC approach (Table 8), even though fewer markers were included in the prediction.

Finally, a challenging and complex dataset including six-person mixtures, low template DNA mixtures containing brothers and extremely degraded mixed DNA profiles were submitted to the RFC11 model to define the limitations of this NOC model. It was expected that the number of correct predictions would be lower for these complex data when compared to the train, test and hold-out set. The RFC11 model estimates a maximum of 5 contributors. A

prediction of 5 should thus be interpreted as ≥5 and this is regarded as the optimum (best) prediction that can be obtained for mixtures with six (or more) contributors. The RFC11 model is trained on single donor to five-person mixtures having no or moderate levels of drop-out and degradation. As expected, most six-person mixtures received an estimate of five and almost all samples including DNA of brothers or showing extreme levels of drop-out received an underestimated NOC.

Overall, in absence of a specific model, the generic RFC11 model can be a useful addition to the reporting officer's toolbox to interpret complex DNA profiles. In most instances it outperformed the MAC approach, though it uses fewer markers. As with every machine learning model, incorrect predictions are expected for complex data not used in training.



Supplementary material Figure 2. Estimated NOC (X-axes) in comparison to the true NOC (Y-axes) for DNA profiles in the train (A), test (B), or hold-out dataset (C) when using the generic RFC11 model.

Supplementary material Table 8. Percentage of correctly estimated NOC when using the MAC approach, the PPF6C RFC19 model or the generic RFC11 machine learning model for a selection of 120 two- to five-person mixtures presented in [15].

True NOC	*n*	Percentage of correct predictions		
		MAC	PPF6C RFC19	Generic RFC11
2	30	66.7%	100%	100%
3	30	96.7%	96.7%	80%
4	30	76.7%	83.3%	66.7%
5	30	36.7%	60.0%	70%
Total	120	69.2%	85.0%	79.2%

Supplementary material Table 9. Estimated NOC when applying the generic RFC11 machine learning model to extreme data types that were not used in training.

Type of extreme samples	*n*	True NOC	Predicted NOC				
			1	2	3	4	5
Six-person mixtures	16	6	-	-	-	3	13
DNA mixtures from brothers with allelic drop-out	5	2	3	2	-	-	-
	3	3	-	3	-	-	-
	2	4	-	2	-	-	-
Extremely degraded mixed DNA profiles	3	3	3	-	-	-	-
	3	4	1	-	2	-	-
	3	5	2	-	1	-	-