Forensic Science International: Genetics 49 (2020) 102390

Contents lists available at ScienceDirect

Forensic Science International: Genetics

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

Research paper

Multi-laboratory validation of DNAxs including the statistical library DNAStatistX

Corina C.G. Benschop [a,*], Jerry Hoogenboom [a], Fiep Bargeman [a], Pauline Hovers [a], Martin Slagter [a], Jennifer van der Linden [a], Raymond Parag [a], Dennis Kruise [b], Katja Drobnic [c], Gregor Klucevsek [c], Walther Parson [d,e], Burkhard Berger [d], Francois Xavier Laurent [f], Magalie Faivre [f], Ayhan Ulus [f], Peter Schneider [g], Magdalena Bogus [g], Alexander L.J. Kneppers [a], Titia Sijen [a]

[a] Netherlands Forensic Institute, Division of Biological Traces, Laan van Ypenburg 6, 2497GB, The Hague, The Netherlands
[b] Netherlands Forensic Institute, Division of Digital and Biometric Traces, The Hague, The Netherlands
[c] National forensic laboratory, Police, Ministry of the Interior, Ljubljana, Slovenia
[d] Institute of Legal Medicine, Medical University of Innsbruck, Austria
[e] Forensic Science Program, The Pennsylvania State University, University Park, PA, USA
[f] Institut National de Police Scientifique, Ecully, France
[g] Institute of Legal Medicine, University Hospital of Cologne, Division of Forensic Molecular Genetics, Cologne, Germany

ARTICLE INFO

Keywords:
DNA profile interpretation
Likelihood ratio
Continuous model
Validation
DNAxs
DNAStatistX

ABSTRACT

This study describes a multi-laboratory validation of DNAxs, a DNA eXpert System for the data management and probabilistic interpretation of DNA profiles [1], and its statistical library DNAStatistX to which, besides the organising laboratory, four laboratories participated. The software was modified to read multiple data formats and the study was performed prior to the release of the software to the forensic community. The first exercise explored all main functionalities of DNAxs with feedback on user-friendliness, installation and general performance. Next, every laboratory performed likelihood ratio (LR) calculations using their own dataset and a dataset provided by the organising laboratory. The organising laboratory performed LR calculations using all datasets. The datasets were generated with different STR typing kits or analysis systems and consisted of samples varying in DNA amounts, mixture ratios, number of contributors and drop-out level. Hypothesis sets had the correct, under- and over-assigned number of contributors and true and false donors as person of interest. When comparing the results between laboratories, the LRs were foremost within one unit on log10 scale. The few LR results that deviated more had differences for the parameters estimated by the optimizer within DNAStatistX. Some of these were indicated by failed iteration results, others by a failed model validation, since unrealistic hypotheses were included. When these results that do not meet the quality criteria were excluded, as is in accordance with interpretation guidelines, none of the analyses in the different laboratories yielded a different statement in the casework report. Nonetheless, changes in software parameters were sought that minimized differences in outcomes, which made the DNAStatistX module more robust. Overall, the software was found intuitive, user-friendly and valid for use in multiple laboratories.

* Corresponding author.
E-mail addresses: c.benschop@nfi.nl (C.C.G. Benschop), j.hoogenboom@nfi.nl (J. Hoogenboom), f.bargeman@nfi.nl (F. Bargeman), p.hovers@nfi.nl (P. Hovers), m.slagter@nfi.nl (M. Slagter), j.van.der.linden@nfi.nl (J. van der Linden), r.parag@nfi.nl (R. Parag), d.kruise@nfi.nl (D. Kruise), katja.drobnic@policija.si (K. Drobnic), gregor.kljucevsek@policija.si (G. Klucevsek), Walther.parson@i-med.ac.at (W. Parson), Burkhard.berger@i-med.ac.at (B. Berger), fx.laurent@interpol.int (F.X. Laurent), magalie.faivre@interieur.gouv.fr (M. Faivre), ayhan.ulus@interieur.gouv.fr (A. Ulus), peter.schneider@uk-koeln.de (P. Schneider), magdalena.bogus@uk-koeln.de (M. Bogus), s.kneppers@nfi.nl (A.L.J. Kneppers), t.sijen@nfi.nl (T. Sijen).

https://doi.org/10.1016/j.fsigen.2020.102390
Received 16 June 2020; Received in revised form 30 July 2020; Accepted 26 August 2020
Available online 7 September 2020

1. Introduction

Software solutions facilitate the interpretation of DNA profiles in forensic casework. Such software undergo developmental validation by the developing institute followed by internal validation within the institute that aims to use the software in a casework setting according to established software validation guidelines [2–7]. As different laboratories use different short tandem repeat (STR) typing kits and different polymerase chain reaction (PCR), capillary electrophoresis (CE) and profile analysis settings, their DNA profiling data can have different characteristics as well as different formats. Because of these differences, each laboratory aiming to implement a particular DNA profile interpretation software, needs to perform its own internal validation prior to implementation in casework.

In this study, the DNA eXpert System 'DNAxs', which includes the statistical library DNAStatistX, was validated by multiple laboratories. Aside from DNAStatistX to compute likelihood ratios (LRs), DNAxs accommodates, among other things, features to import, view, infer and compare autosomal short tandem repeat (STR) profiles [8]. The DNAxs software was developed in 2017 and the developmental validation and internal validation at the developing institute were published in [1]. In that study, DNAStatistX was compared to EuroForMix [9] as both software tools are based on the model inferred from [10]. In this study, five laboratories participated; the National Forensic Laboratory in Slovenia, the Institute of Legal Medicine in Austria, Institut National de Police Scientifique in France, Institute of Legal Medicine in Germany, and the Netherlands Forensic Institute in the Netherlands. Prior to the multi-laboratory study, the DNAxs software was adapted so that genotyping data of different formats (different sample name patterns) can be loaded and more STR typing kits are accommodated. After an exercise to check the installation process and explore all main functionalities of DNAxs, LR calculations were performed using DNAStatistX. Both a dataset provided by the organising laboratory and laboratory-specific datasets were used to cover the variation in DNA profiles that can be encountered in forensic casework. This study describes the comparison of the results obtained in the various laboratories.

2. Materials and Methods

2.1. DNA profiles

Five datasets of DNA profiles were created, i.e. one per laboratory. These datasets were designed to include single donor samples and complex DNA mixtures with known content, and varying for at least the mixture proportion and level of drop-out. Most importantly, the sets were designed to cover the range of sample types encountered in casework at the particular laboratories. Table 1 presents an overview of the numbers of DNA samples used in this study and further details on the composition of the samples are provided in Supplementary Table 1. Laboratories 1 to 5 used 30, 17, 5, 11 or 8 different donors, respectively, to generate their dataset. The minimum amount of DNA per donor was 30 pg for laboratories 1, 3 and 5. Laboratory 2 and 4 used a minimum of 100 and 375 pg per donor, respectively (Supplementary Table 1). DNA profiles were generated using the STR kit and PCR, CE and analysis settings (e.g. stutter and detection thresholds) in use at each laboratory (Supplementary Table 2). The datasets were shared in the form of. csv files containing allele numbers and peak heights.

2.2. DNAxs and DNAStatistX

In this study, two DNAxs software versions were used. With DNAxs version 1.3.2.200b22a, LRs were calculated for the data from lab 1 in all laboratories. Updated DNAxs version 1.4.0.3366b12 accepts more data formats (sample name patterns) and was used for the datasets from laboratories 2 to 5. DNAxs imports DNA profiles from text (.txt,. csv and similar) containing the following columns: Sample Name, Marker, Allele 1, Allele 2, Allele n, Height 1, Height 2, Height n. The first versions of DNAxs required a sample name pattern that was specific to the developing laboratory. The updated DNAxs version reads the data if the sample name column contains a sample identifier (a sample name or number, uniquely identifying the trace or reference profile), possibly followed by a CE run (or replicate) identifier and/or followed by a case number. In both DNAxs versions, DNAStatistX (MLE, maximum likelihood estimate) v1.0.2 was used. DNAStatistX v1.0.2 uses the cumulative density of the Gamma distribution to perform model validation. In this model validation, probability-probability (PP) plots are generated, plotting the cumulative probabilities for the expected peak heights against the cumulative probabilities for the observed peak heights [9]. During model validation, a Bonferroni-corrected 'goodness of fit test' is performed where the default significance level of 0.01 was used. Along with the PP-plots, the DNAxs software returns the status of the model validation (passed, warning, failed, or unavailable), which was examined for the LR calculations performed in this study. Furthermore, DNAxs presents results obtained by the optimizer which were noted and examined in detail for analyses that yielded variation in log10 LRs between laboratories. Based on the results obtained in this study, DNAStatistX's optimizer settings were modified, yielding DNAStatistX v1.0.5. The performance of DNAStatistX v1.0.5 was examined with the data that yielded deviating LR results in the inter-laboratory test as well as with 318 propositions from a validation dataset described in [1].

In each of the participating laboratories, DNAxs was installed on computers with at least 8 GB of memory and at least 4 CPU cores. Processing time was monitored for LR calculations within lab 1. These calculations were run on Dell PowerEdge R6415 servers, each equipped with one AMD EPYC 7601 processor (32 cores, 64 threads) and 64 GB of memory. Analyses with three or four contributors were configured to use

Table 1
Overview of the number of DNA samples used in the multi-laboratory evaluation of DNAxs/DNAStatistX and the number of LR calculations performed using DNAStatistX.

	Data lab 1		Data lab 2		Data lab 3		Data lab 4		Data lab 5	
True number of contributors	# samples	# LRs	# samples	# LRs	# samples	# LRs	# samples	# LRs	# samples	# LRs
1	3	6	3	6	3	6	3	6	5	6
2	10	16	8	13	8	13	8	13	4	13
3	10	17	8	13	8	13	8	13	4	13
4	10	19	8	15	8	15	7	13	4	15
5	3	6	-	-	3	6	3	5	3	6
Total	36	64	27	47	30	53	29	50	20	53
Number of Hp-true tests		42		27		32		29		32
Number of Hd-true tests		15		13		14		14		14
Number of Hd-true tests using simulated brother		7		7		7		7		7
# labs that calculated LRs		5		2		2		2		2
Total # LRs per dataset		320		94		106		100		106
Total # LRs					726					

five threads, whereas analyses with less than three contributors were configured to use two threads.

2.3. Likelihood Ratio calculations

The organising laboratory (lab 1) computed LRs for each of the five datasets and labs 2-5 calculated LRs for the dataset of lab 1 and for their own dataset (Table 1). In addition, in lab 2, 15 results from lab 1 data and 15 results from lab 2 data were calculated twice by two different experts and results were compared within lab 2 prior to sending the results to lab 1 for further processing. This comparison showed no differences between the analyses of the two experts within lab 2. The propositions varied for: the number of contributors (NOC, one to six), the correctness of the NOC (true, under, or over-assigned number), the conditioning on known contributors (0, 1 or 2), the number of PCR replicates (one to three) and a true or a non-contributor (non-related or simulated brother) as the person of interest (POI) under the prosecution hypothesis (Hp, in the software this is denoted H1). Non-contributors were selected from existing genotypes, either from the same dataset or from a large dataset of 2085 Dutch donors [11]. The genotypes of the simulated brothers were manually created by modifying about 50% of the alleles of one of the true contributors. In addition, the propositions of lab 1 varied for F_{ST} correction (0, 0.01 or 0.03), whilst the propositions for the data of labs 2-5 used one fixed F_{ST} value chosen by the particular laboratory. All LR calculations were performed using the settings and thresholds that apply to the specific dataset. For the data of lab 1 population frequencies as published in [11] were used; for the data of labs 2-5, internal population data as provided by the participating lab were applied. Each set of propositions included two duplicate LR calculations (using the same data; denoted rerun) and two LR calculations using duplicated data (the LR calculation was performed using a data file that was copied and renamed; denoted duplicate) (Supplementary Tables 3–7). A rerun is exactly the same, though run at a different moment within the laboratory. The duplicate used the same data but the file name was different between the two comparing analyses. This was to confirm that the software would handle such analyses correctly and not, for example, presents, or mix up, stored data. Settings as used in the LR calculations are provided in Supplementary Table 2.

Propositions that led to LR differences of more than one unit on log10 scale when analysed by different laboratories were recalculated eight times by laboratory 1. Smaller differences were noted, but LR calculations were not repeated as such differences in LRs were regarded negligible, as they will not affect the reporting of these results in forensic casework.

2.4. Inference of a major contributor to DNA profiles

DNAxs includes the LoCIM method (Locus Classification & Inference of the Major) to infer the alleles of the most prominent component to a DNA profile, i.e. the major contributor [12]. This feature is available when the stochastic threshold that applies to the data, is provided to the DNAxs software. In this study, labs 3 and 4 provided their stochastic threshold (Supplementary Table 2) and the performance of the LoCIM method was examined using their laboratory- specific data. The numbers of Type I, II and III loci [12] were noted and the alleles marked as belonging to the major component (blue bars in DNAxs, see Supplementary Fig. 1 for an example) were compared to the true composition of the DNA profiles.

2.5. Estimation of the number of contributors

DNAxs provides DNA profile information that can be useful for estimating the number of contributor (NOC), such as the maximum allele count (MAC, maximum number of alleles at any of the loci) and the total allele count (TAC, the total number of alleles observed in the DNA profile). In addition, DNAxs includes NOC tools that automatically present the estimated NOC based on machine learning approaches. Two different NOC tools were implemented: the PPF6C RFC19 model that is specifically suited for PPF6C profiles as presented in [13] and the generic RFC11 model designed to be applied to data from any STR typing system. Both are based on a random forest (RFC) classifier, but use a different number of features (19 and 11, respectively) and are applicable to different data. The RFC19 model is described in Benschop et al. [13] and applies to PPF6C DNA profiles generated using the settings as used by default in lab 1 [13]. The RFC11 model is a generic model that is not only independent of the STR kit used but also of settings for PCR, CE and profile analysis. The development of the generic model followed the same approach as the RFC19 model regarding feature selection and model selection. This generic model only involves features of the 12 ESS (European Standard Set) and U.S. core loci available in the most common commercial STR typing kits (FGA, TH01, VWA, D1S1656, D2S441, D3S1358, D8S1179, D10S1248, D12S391, D18S51, D21S11, D22S1045). This model does not include features holding information on peak heights or fragment lengths. Further details on the development of the RFC11 model is described in Supplementary Material 1. Lab 1 applied a stand-alone version of the generic RFC11 model to the datasets of laboratories 1, 2, 3 and 4 as presented in Supplementary Table 1 plus some additional profiles that were supplied by the laboratories, but not used in the weight of evidence calculations as described previously. The dataset of lab 5 was not used as it was not available at the time of testing. The NOC as estimated by the generic RFC11 model was compared to the number of contributors according to the MAC and compared to the true (designed) NOC (the data for the RFC19 model are presented in [13]).

3. Results and discussion

3.1. Exercise and feedback on the software

After an instructive demonstration of the DNAxs software at the organising laboratory, labs 2-5 installed the software in their own laboratory and performed an exercise for which the software manual was used as a guide. This approach assessed whether the instructions in the manual were clear and sufficient to use the software independently. With some support, all participants succeeded in installing the software, which incited the programming of a software installer and an installation manual. The software received good feedback on its general functionalities and user-friendliness. Furthermore, it was suggested to expand the manual with further details on the DNAStatistX results, such as explanation on the model validation and advice on actions to take in case a failed model validation or failed iterations result is obtained. These details are now included in the DNAxs manual and/or the frequently asked questions section on the DNAxs website [8].

3.2. Comparison of log10 LRs

Each of the five laboratories calculated 64 LRs using the data from laboratory 1. Between 47 and 53 LRs were calculated per dataset of labs 2 to 5, which were performed by the respective lab and lab 1. Out of 726 LR calculations, 450 were Hp-true and 276 were Hd-true analyses. The LR outcomes were within the expected range: The Hp-true tests with a true donor as person of interest (POI) under Hp yielded LRs in favour of Hp, except for three instances in which the number of contributors was deliberately set too low to enable explaining the observed alleles (Supplementary Table 8A). In addition, the Hd-true tests yielded LRs in favour of Hd, except for a few non-donors for four- or five-person mixtures or simulated brothers of the true donors (Supplementary Table 8B). The maximum value obtained for an unrelated non-donor was LR = 36 and concerned a five-person mixture for which the POI had four unseen alleles and its mixture proportion was estimated by DNAStatistX at only 3.8%. The maximum value for a simulated brother as POI was log10 LR = 7.4. This related non-donor had very high resemblance with

the mixture profile; all, except one, of his alleles were observed in the mixture profile. These values are within the ranges as observed in the previously performed developmental and internal validations [1].

During developmental and internal validation of DNAStatistX, duplicate analyses yielded log10 LRs (overall and per locus) that were the same up to at least two decimal places [1]. However, small differences between repeated DNAStatistX calculations may be obtained as the MLE approach within DNAStatistX uses an optimizer that searches for the best fitting parameter values for mixture proportion, peak height expectation, peak height variance and degradation slope. As the true values of these parameters are unknown, the optimizer employs a trial and error approach to find the values for these parameters that best explain the DNA profiles obtained. Then, the optimizer tries to find a better fit on the data by changing the parameter values in small, randomly-chosen steps, with the ultimate goal of finding the largest (maximum) likelihood value. Obtaining slight differences in these parameter values can subsequently result in slight differences in the log10 LR. The maximum difference in [1] was log10 LR = 0.63 when comparing EuroForMix to DNAStatistX that use the same model but a different optimizer (EuroForMix uses the function nlm while DNAStatistX uses CMA-ES). Furthermore, during the developmental validation, up to 60 repeated LR calculations were performed for 180 sets of propositions. From these results, the probability of missing the optimal parameter values that best explain the data, was calculated as one in 16, 445 with three repeated optimizer runs, denoted iterations [1]. For some specific analyses the optimum log likelihood value was missed more often. The optimizer in DNAStatistX v1.0.0 (and the follow-up v1.0.2) was therefore programmed dynamically to have a minimum of three iterations. If with three iterations the likelihoods were not identical (with a maximum difference of two decimal places on log10 scale), the number of iterations is increased by one until the same largest likelihood is obtained three times. The maximum number of iterations was set to ten. Using this approach, the probability of missing the optimum was regarded almost negligible; the probability of missing the optimum after ten iterations was calculated as 6×10^{-12} [1]. DNAStatistX presents whether the optimizer iterations are scored as passed (green), warning (orange), or failed (red). See Supplementary Fig. 2C for an example. A passed result is presented if three iterations yield a sufficiently similar largest log likelihood. A warning is given if less than three iterations yield sufficiently similar largest log likelihood out of the maximum of (by default) ten iterations. A failed iteration result is presented if no sufficiently similar results were found after the maximum of ten iterations.

In the present study, all duplicate DNAStatistX calculations and analyses using duplicated experimental data yielded equal log10 LR results (up to at least two decimal places) within the laboratories, as was expected based on the developmental validation study [1]. When comparing the LR results between laboratories, 95.3% to 100% did not exceed one unit on log10 scale (Table 2). This means that the LRs obtained by a laboratory were for 95.3% up to 100% within one unit on log10 scale in comparison to the LRs that were obtained by the laboratory that created the particular dataset. Eight log10 LRs deviated more: five when labs 2-5 used data from lab 1, two when lab 1 used data from lab 4 and one when lab 1 used data from lab 5 (Supplementary Fig. 3). Data from labs 2 and 3 did not show such variations. Deviating log10 LR results were examined in more detail and showed that for each of these calculations, the optimum log likelihood was missed by at least one of the laboratories, under Hp, Hd or both. In these analyses, the largest log likelihood that was found by the optimizer was lower than the largest log likelihood found in the analyses performed by (an)other laboratory/laboratories. This indicates that these were local optima, though the aim is to obtain the global (maximum) optimum. The deviating results were reanalysed eight times by lab 1. For five of the eight sets of propositions, the variation in log10 LR was now observed as well within a laboratory (lab 1) (Fig. 1). Detailed analyses confirmed that in some of these repeated LR calculations the optimum log likelihood was missed by the optimizer under Hp, Hd or both. Evidently, the optimum log likelihood was missed by the optimizer more often than expected. A failed iteration result returned by the software may indicate that the optimum log likelihood is missed, though a passed iteration result is no guarantee that the global optimum was found. Therefore, further optimisation and testing of these optimisations was performed aiming to minimize the chance of missing the optimum log likelihood. These tests and results are described in Sections 3.4 and 3.5.

Differences in log10 LRs can be problematic in casework if they result in a different statement in the casework report. Reporting guidelines for LRs can differ between laboratories. In this study, the guidelines for PowerPlex Fusion 6C data generated in lab 1 were used [1]. In short: 1) LR calculations are performed if the person of interest (POI) has at most 15, 10 or 5 unseen alleles in the trace profile with an estimate of two, three or four contributors, respectively, 2) A lower reporting threshold of log10 LR = 4 (below this value no LR is reported) and an upper threshold of log10 LR = 9 is applied (above this value more than a billion is reported), and 3) LR results are not to be reported if the model validation and/or optimizer iterations fail and cannot be explained or solved by rerunning the calculation (possibly under different propositions) [1]. The DNAxs software presents the obtained LR results along with information on the model validation and iteration results (see Supplementary Fig. 2 for an example). These latter are presented as passed, failed, or warning (or unavailable for model validation in case a minus infinity result is obtained or the trace profile contains alleles below the predefined detection threshold). The model validation and iteration results were examined for the 13 repeated calculations (nine times by lab 1, once by labs 2-5) that were performed for the eight sets of propositions with deviating LRs. When following the guidelines by excluding the analyses with a failed model validation and/or optimizer iterations, the LRs are very similar and would not give a different statement in a casework report (Fig. 1, black bars).

Table 2
Percentages of log10 LRs that did not exceed one unit on log10 scale difference when compared to the results obtained by the laboratory that generated the dataset.

	Dataset lab 1 ($n = 64$)	Dataset lab 2 ($n = 47$)	Dataset lab 3 ($n = 53$)	Dataset lab 4 ($n = 49$)	Dataset lab 5 ($n = 53$)
Lab 1	-	100%	100.0%	95.9%	98.1%
Lab 2	98.4%	-	-	-	-
Lab 3	95.3%	-	-	-	-
Lab 4	96.9%	-	-	-	-
Lab 5	96.9%	-	-	-	-

3.3. Model validity

The DNAStatistX model includes, like EuroForMix, model validation plots. These so-called PP- plots are generated to check whether the Gamma distribution model and its settings in combination with the defined propositions can well explain the observed data. In the PP-plots, the cumulative probabilities for the expected peak heights are plotted against the cumulative probabilities for the observed peak heights [9]. The data is well explained if the expected peak heights are close to the observed peaks heights. The range of variation is set to a default significance level of 0.01 in DNAStatistX. This 0.01 line is based on a Bonferroni corrected significance level and is shown as an 'envelope' in the PP-plots (see Supplementary Fig. 2B for an example). When four or more values fall outside the 0.01 line, the model validation is scored as 'failed'. A 'warning' is presented if 2-3 data points fall outside the 0.01 line. A failed model validation indicates that the data and its propositions cannot be explained well by the applied model which is a

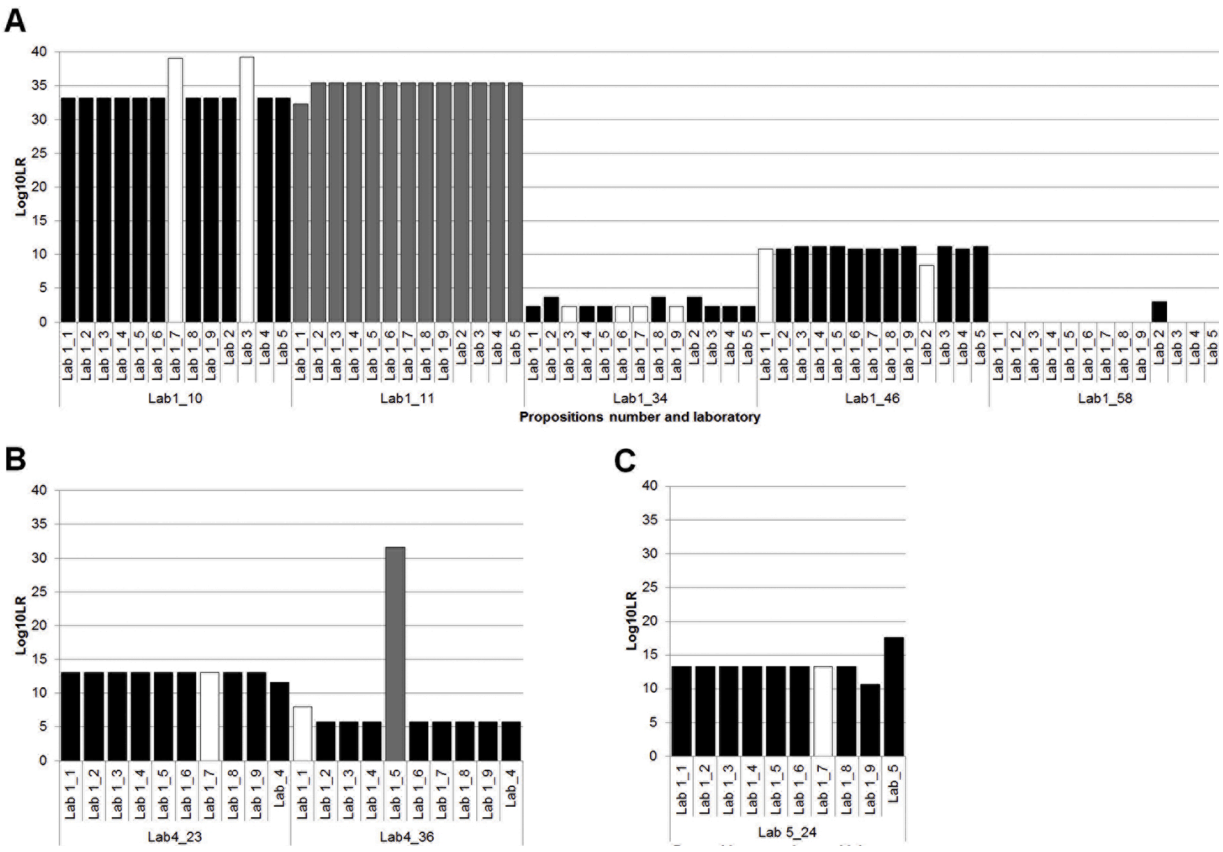

Fig. 1. Overview of the propositions for which the log10 LR could differ by more than one unit between laboratories. Variation in log10 LRs for analyses using the dataset of lab 1 (A), the dataset of lab 3 (B), or the dataset of lab 5 (C). Lab 1_58 shows one bar only, as the remainder of analyses yielded an LR of zero. Details on the propositions can be found in Supplementary Tables 3, 6 and 7. Black bars indicate that the LR was accompanied by a passed model validation and passed iterations result. Grey bars represent that the LR was accompanied by a failed model validation and white bars represent that it was accompanied by failed optimizer iterations. Following the suggested guidelines, only LR results from black bars meet the quality requirements and can be reported in forensic casework.

combination of the Gamma distribution model, drop-in model, and degradation model within DNAStatistX. Previous studies indicated the types of analyses that could result in failing model validations, e.g. incorrect parameter settings such as lack of degradation model application when data do show degradation, an under- assigned number of contributors and/or too low drop-in values so that the observed peaks cannot be well explained, incorrect kit settings, analyses with a non-contributor as POI under Hp (the model validation often, but not always, fails with a non-contributor), or analyses of replicates with (large) peak height variation.

In this study, the model validation plots were examined both under Hp and Hd and outcomes were compared between laboratories. Comparison of model validation results between laboratories showed high similarity (Table 3). Only two analyses deviated: one analysis showed a difference of one data point that was for lab 1 within but for labs 2-5 outside the 0.01 significance level; the other analysis yielded 2 data points outside the 0.01 significance level when analysed by lab 1, while it resulted in 13 data points outside the 0.01 significance level each time analysed by labs 2-5. In the latter case, the POI was a non-donor and all laboratories obtained log10 LR=-22.9 and thus the same statement in a casework report. As expected, most of the failed model validations were explained by the reasons mentioned above (Table 3). The model validations that failed due to an under-assigned number of contributors passed when using the true (designed) number of contributors under the propositions.

3.4. Adjusting optimizer settings

The LR results for seven sets of propositions (out of a total of 726 LR calculations, Table 1) showed some differences between laboratories (Fig. 1). These were caused by differences in parameter values estimated by the optimizer (see Section 3.2 for further information on the optimizer). Therefore, possible solutions to minimize the differences were examined. These included increasing the number of optimizer iterations, lowering the accepted log likelihood differences between iterations, and/or scoring optimizer iterations as equal if, in addition to the log likelihood, other model parameter values have predefined similarity. Options 1-6 shown in Table 4A were repeatedly applied to the eight sets of propositions: propositions with less than four unknowns under Hd were repeated 100 times, those with four unknowns 10 times (because of calculation time) and when these showed variable results, 25 times.

Using the original approach as programmed in DNAStatistX v1.0.2 (option 1 in Table 5A), the repeated analyses indeed showed that the optimum model parameters were missed with five of the eight sets of propositions (Table 4B). Best results were obtained with option 6 (Table 4AB), although for three sets of propositions the optimum values were still missed (Table 4B, lab1_analysis46 Hd and lab4_analysis23 Hd, lab5_analysis24 Hp). These were propositions with a larger number of contributors assigned than required to explain the observed alleles. When reanalysing these samples with a more realistic number of contributors (the true number of contributors), neither deviating LR results nor missed optimum model parameters were obtained. When examining these three sets of propositions in more detail, it was noted that unknowns received equal mixture proportions which means that a too large mixture proportion for the non-existing extra person was assigned (well above 1%) as the mixture proportion for this person was below 0.57% when the optimum log likelihood was achieved.

This knowledge prompted the testing of a 7th option (Table 4A),

Table 3
Number of analyses that resulted in a passed, warning or failed model validation, including reasons for failing. Analyses using data from lab 1 (A), 2 (B), 3 (C), 4 (D), or 5 (E).

	Model validation results						
	Total number	Number passed	Number warnings	Number failed			
A				Total	POI = non-donor[a]	Too many/large peaks for given drop-in and/or NOC[b]	Other
Lab 1	64	42[c]	2	20	16	4	0
Lab 2	64	40	3	21	17	4	0
Lab 3	64	40	3	21	17	4	0
Lab 4	64	40	3	21	17	4	0
Lab 5	64	40	3	21	17	4	0
B							
Lab 1	44[d]	34	7	3	3	0	0
Lab 2	44[d]	34	7	3	3	0	0
C							
Lab 1	53	42	3	8	7	0	1[e]
Lab 3	53	42	3	8	7	0	1[e]
D							
Lab 1	47[f]	34	0	13	11	1	1[g]
Lab 4	47[f]	34	0	13	11	1	1[g]
E							
Lab 1	53	35	4	14	12	2	0
Lab 5	53	35	4	14	12	2	0

[a] A non-donor (unrelated or brother of true donor) was the POI in 21/64 (lab 1), 20/47 (lab 2), 21/53 (lab 3), 21/50 (lab 4), and 16/47 (lab 5) of the analyses.
[b] Model validation did not fail when the number of contributors under the propositions was increased by 1.
[c] One sample gave a warning within labs 2-5 and was passed within lab 1. It showed a difference of one data point that was within (lab 1) or outside (labs 2-5) the envelope in the PP-plot. In addition, one sample that gave a warning within lab 1 and failed within labs 2-5 showed 2 data points (lab 1) or 13 data points (lab 2-5) under Hp that fell outside the envelope. The POI was a non-donor and all laboratories obtained log10 LR=-22.9. This would not affect the statement in a casework report.
[d] Only 44/47 results could be compared as three analyses yielded 'unavailable' model validation within lab 2. These data files included alleles below the detection threshold as specified in the particular kit settings. Model validation results were shown when the data files did not contain these below threshold alleles, but were not rerun by the particular lab.
[e] The number of contributors was over-assigned in this LR calculation. Rerunning the analyses with the true number of contributors yielded a passed model validation.
[f] Three of the 50 model validations were not retrieved due to an error which occurred during transit of the data for comparison within this study.
[g] No clear explanation. This model validation had 0 outliers under Hp and 4 outliers under Hd, just above the threshold.

which is explained below. As mentioned in Section 2.3, the parameters for mixture proportion, peak height expectation, peak height variance and degradation slope are estimated by the optimizer using a trial and error approach to find the values for these parameters that best explain the observed DNA profile. Except for the mixture proportions, which are chosen randomly, the initial parameter values are directly calculated from the peak heights in the profile. This is denoted as the 'initial guess'. Next, the optimizer tries to find a better fit on the data by changing the parameter values, with the ultimate goal of finding the largest (maximum) likelihood value. In option 7, in case the maximum likelihood value is obtained with equal mixture proportions for unknown contributors, the optimization is re-started with the mixture proportion for one of those unknown contributors initially set to 1% (Table 4A). In addition, in this option, optimizer settings were modified to reduce the range of parameter values sampled by the optimizer by staying closer to the 'initial guess' (Table 4A). This option 7 yielded the optimum parameter values for all of the eight sets of propositions (Table 4B) and was therefore programmed in DNAStatistX v1.0.5. When examining the LR calculation time, DNAStatistX v1.0.5 was in most cases faster than DNAStatistX v1.0.2 (Table 4C), which is an appreciated side effect.

3.5. Testing adjusted optimizer settings

To examine if the adjusted optimizer settings function robustly, a dataset of 318 sets of propositions was selected from [1] and subjected to the settings programmed in DNAStatistX v1.0.2 and v1.0.5 with 100 repetitions for propositions with up to three contributors and 10 repetitions for propositions with four or more contributors. With v1.0.2, 19 and with v1.0.5, two of the 318 analyses failed to yield the optimum log likelihood value in at least one of the repeated analyses (Table 5). Thus, improved optimizer estimates were obtained with v1.0.5. Details are presented in Table 5. For the two analyses for which the optimum was missed with v1.0.5, sometimes the iteration criteria were not met (Table 4A), but these were not the same analyses for which the optimum was missed. The effect of missing the optimum parameter values, however, hardly affected the LRs (Table 5B, -0.59 or +0.12 on log10 scale) and would not yield a difference for casework reporting.

In DNAStatistX, the time needed for LR calculation and for Hp and Hd model validation account for most of the processing time. The total processing time for DNAStatistX v1.0.2 and v.1.0.5 were compared using different numbers for the unknowns under Hd and the total number of contributors (Table 6). Processing times with DNAStatistX v1.0.5 were shorter when no extra optimisation steps were required and could be larger when extra optimisation steps were performed. Calculating the LR and performing model validation under both Hp and Hd took seconds for mixtures with up to two unknowns under Hd, minutes for mixtures with up to three unknowns and hours (less than a day) for mixtures with four unknown contributors under Hd (Table 6B).

DNAStatistX v1.0.5 is integrated within DNAxs v2.0, which is the version disseminated to the forensic community (see [12]). In this version, the maximum number of iterations can be adjusted (default is ten). In a future version, optimizer settings as described in Table 4A will be programmed into user settings enabling manual modification. Furthermore, opportunities to reduce the calculation time will be further examined and preliminary results are very promising (Supplementary Table 9).

Table 4
A) Different options tested for the optimizer. B) Number of missed optimum log likelihood values under Hp or Hd when analysing the eight sets of propositions that yielded variation on log10 LR between laboratories. C) Average calculation time per option and sample.

A	Feature	Option 1[a]	Option 2	Option 3	Option 4	Option 5	Option 6	Option 7
	Number of iterations required with sufficient similarity out of max. 10 iterations	3	4	3	4	3	4	3
	Maximum accepted variation for: Log likelihood	1%	1%	ln2	ln2	ln2	ln2	ln2
	Maximum accepted variation for: Mixture proportion	–	–	–	–	0.01	0.01	0.01
	Maximum accepted variation for: Degradation slope	–	–	–	–	0.01	0.01	0.01
	Maximum accepted variation for: Peak height expectation	–	–	–	–	1%	1%	1%
	Maximum accepted variation for: Peak height variation	–	–	–	–	5%	5%	5%
	Optimizer start points	Random	Random	Random	Random	Random	Random	Only Mx random.[b]
	If equal mixture proportions for unknowns, re-start with 1 set to 1%	–	–	–	–	–	–	√

B	Sample	Short description of the analysis	Number of repetitions	Number of times optimum log likelihood was missed under Hp or Hd													
				Option 1[a]		Option 2		Option 3		Option 4		Option 5		Option 6		Option 7	
				Hp	Hd	Hp	Hd	Hp	Hd	Hp	Hd	Hp	Hd	Hp	Hd	Hp	Hd
	Lab1_10	2p analysed as 3p, Hp-true	100	0	1	0	0	0	4	0	1	0	1	0	0	0	0
	Lab1_11	2p (3 replicates), Hp-true[c]	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Lab1_34	3p, Hp-true	100	0	4	0	0	0	0	0	0	0	3	0	0	0	0
	Lab1_46	4p analysed as 5p, Hp-true	25[d]	0	10	0	8	2	12	0	9	0	14	0	4	0	0
	Lab1_58	4p, Hd-true, POI=brother of true donor	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Lab4_23	3p analysed as 4p, Hp-true	25[d]	4	25	1	25	3	25	2	24	3	25	0	23	0	0
	Lab4_36	4p analysed as 5p, Hp-true	10	0	1	0	0	0	1	0	0	0	0	0	0	0	0
	Lab5_24	3p analysed as 4p, Hp-true	25[d]	3	11	0	9	5	14	1	8	2	15	2	8	0	0

C	Sample	Short description of the analysis	Number of repetitions	Average calculation time						
				Option 1[a]	Option 2	Option 3	Option 4	Option 5	Option 6	Option 7
	Lab1_10	2p analysed as 3p, Hp-true	100	0:00:55	0:01:11	0:00:55	0:01:09	0:00:55	0:01:08	0:00:29
	Lab1_11	2p (3 replicates), Hp-true[c]	100	0:00:18	0:00:22	0:00:31	0:00:23	0:00:19	0:00:23	0:00:07
	Lab1_34	3p, Hp-true	100	0:01:01	0:01:07	0:01:10	0:01:24	0:01:11	0:01:30	0:00:35
	Lab1_46	4p analysed as 5p, Hp-true	25[d]	4:55:49	3:46:46	4:18:32	6:26:29	4:37:00	5:40:59	4:17:50
	Lab1_58	4p, Hd-true, POI=brother of true donor	10	1:39:41	1:54:50	2:21:55	3:18:59	1:46:33	1:50:37	0:51:17
	Lab4_23	3p analysed as 4p, Hp-true	25[d]	1:17:36	2:20:03	1:21:47	1:40:04	1:22:31	1:42:51	1:51:58
	Lab4_36	4p analysed as 5p, Hp-true	10	5:51:17	6:38:01	10:35:29	7:44:09	9:09:00	7:24:54	3:08:08
	Lab5_24	3p analysed as 4p, Hp-true	25[d]	1:50:02	1:42:44	1:38:09	2:25:37	1:30:19	2:26:10	2:06:34
		Average calculation time for these complex analyses		1:55:07	2:03:08	2:32:19	2:42:17	2:18:29	2:22:19	1:32:07

[a] Option 1 represents the original setting incorporated in DNAStatistX v1.0.2.
[b] Parameters other than mixture proportion (Mx) randomly sampled from Gamma distribution with mean calculated from profile. Coefficient of variation: Expected peak height = 0.1, Peak height variance = 0.3, Degradation slope = 0.1.
[c] The 3 replicates had 1 (362 RFUs), 1 (425 RFUs) and 2 (608 and 138 RFUs) elevated stutters, respectively (these four peaks represent three different elevated stutters), which is too many and too high to be explained by the assigned number of contributors and used drop-in parameter values.
[d] For these analyses option 7 was repeated 10 times instead of 25 times.

3.6. Inference of the major contributor's genotype

With mixed profiles, it can be opportune to infer the genotype of a major contributor, for instance for DNA database storage or comparison to other profiles in the case. To this aim, DNAxs contains the LoCIM (Locus Classification & Inference of the Major) method. To enable LoCIM, the stochastic threshold relevant to datasets needs to be provided to DNAxs. This was done by labs 3 and 4 (Supplementary Table 2). LoCIM classifies loci as Type I, Type II or Type III based on criteria for stochastic threshold, mixture proportion and heterozygote balance [12]. Type I loci are regarded easiest to infer a major contributor's genotype from the (mixed) DNA profile and at these loci all criteria are met: peak heights of the major component are above the stochastic threshold, the alleles of a heterozygote major are sufficiently balanced and ratio between the largest peak of the major and the largest peak of minor contributor(s) is distinctive (>8:1 for homozygote and >4:1 for heterozygote major). Type II loci do not need to meet the stochastic threshold criterion and a lower mixture ratio (>4:1 for homozygote and >2:1 for heterozygote major) is utilised. Loci that do not meet the Type I or Type II criteria are classified as Type III. For Type I and II loci the largest allele plus all alleles within 50% of that allele's peak height are inferred as major component. For Type III loci, all alleles having at least 33% of the peak height of the largest allele are included which may evoke inclusion of more alleles than just those of the major component. In DNAxs, the LoCIM-inferred alleles are marked blue (see Supplementary Fig. 1 for an example). Even though LoCIM is not a probabilistic method, it is extremely fast and proved to be useful to casework examiners and outperformed manual inference [12,14]. The number of loci within the profile classified as I, II or III informs whether a major contributor's genotype can be inferred reliably.

The LoCIM method within DNAxs was applied to 30 DNA profiles of lab 3 and 23 DNA profiles of lab 4. These labs use different STR typing systems, PCR and CE settings (Supplementary Table 2), thus we regard the data separately. The alleles inferred by LoCIM were compared to the genotypes of the true major contributors. Results followed expected trends, i.e. most correct inferences were obtained for Type I loci, followed by Type II and Type III (Fig. 2). Although Type I loci are most often inferred correctly, it was noted that such loci might be incorrectly inferred in case the overall profile is of low quality or quantity, or has donors of about equal contribution. Such profiles are indicated by low numbers of Type I loci and many of Type III. In this study, an incorrect inference for a Type I locus was obtained only if the overall profile included four or less Type I loci out of a total of 21 markers. Type III loci often had extra alleles inferred due to the 33% inclusion rule. When regarding the results per DNA profile, samples with large differences in mixture proportion yielded many Type I loci and correct inferences, whilst samples with small differences in mixture proportion yielded mainly Type III loci and incorrect inferences (Fig. 3), which is according

Table 5
Overview and details for the analyses for which the optimum log likelihood was missed (out of 318). In A) original optimizer settings (DNAStatistX v1.0.2, option 1 in Table 4A), and in B) the adjusted optimizer settings (DNAStatistX v1.0.5, option 7 in Table 4A) were used.

A

Analysis number	Sample number	Short description of the analysis	Numb er of repeated analyses	% of analyses with optimum likelihood under Hp&Hd	Log10 LR obtained with optimal Hp & Hd likelihoods	Difference in log10 LR when optimum was missed	Hypothesis with missed optimum
214	1	Hd-true, 3p	100	0%[a]	−18.21	−0.59	Hp
81	2	Hp-true, 4p	10	30%	4.53	−0.41	Hp
85	2	Hp-true, 4p	10	60%	4.53	−0.41 (occurred 3x)	Hp
		duplicate of 81	10	60%	4.53	2.29 (occurred 1x)	Hd
283	3	Hp-true, 4p	10	60%	10.60	0.38	Hd
224	4	Hd-true, 4p	10	70%	−8.90	−10.02	Hp
263	3	Hp-true, 4p	10	70%	9.58	0.12	Hd
231	5	Hd-true, 4p	10	80%	−6.17	−10.23	Hp
243	3	Hp-true, 4p	10	80%	10.94	0.53	Hd
43	6	Hp-true, 3p	100	84%	14.68	0.07	Hd
44	6	Hp-true, 3p	10	85%	14.77	0.10	Hd
78	3	Hp-true, 4p	10	90%	10.74	0.48	Hd
79	3	Hp-true, 4p	10	90%	10.94	0.53	Hd
98	7	Hp-true, 4p 3 replicates	10	90%	0.00	5.71	Hd
185	3	Hp-true, 4p	10	90%	11.53	0.45	Hd
194	7	Hp-true, 4p, 3 replicates	10	90%	8.12	108.16	Hd
221	8	Hd-true, 3p as 4p	10	90%	−1.68	3.54	Hd
258	6	Hp-true, 3p	10	92%	14.77	0.10	Hd
318	6	Hp-true, 3p	10	96%	14.68	0.07	Hd
156	6	Hp-true, 3p	10	99%	14.76	0.08	Hd

B

Analysis number	Sample number	Short	Numb er of repeated analyses	% of analyses with optimum likelihood under Hp&Hd	Log10 LR obtained with optimal Hp & Hd likelihoods	Difference in log10 LR when optimum was missed	Hypothesis with missed optimum
214	1	Hd-true, 3p	100	24%	−18.21	−0.59	Hp
263	3	Hp-true, 4p	10	60%	9.58	0.12	Hd

[a] All analyses yielded the same optimum, but higher optimum with v1.0.5.

Table 6
Processing times for DNAStatistX v1.0.2 (A) and v1.0.5 (B). Results are based on 318 analyses selected from [1] that were performed repeatedly. '*n*' presents the total number of LR calculations performed per category.

A		**Calculation time (hr:min:sec)**		
Unknowns Hd/ total number of contributors	***n***	**Minimum**	**Average**	**Maximum**
1/1	400	0:00:01	0:00:01	0:00:03
1/2	1200	0:00:01	0:00:02	0:00:08
2/2	7900	0:00:02	0:00:13	0:00:36
2/3	2400	0:00:04	0:00:12	0:00:39
3/3	10300	0:00:31	0:05:05	0:33:57
3/4	150	0:03:31	0:06:25	0:17:40
4/4	780	0:03:47	6:27:23	24:44:42
B		**Calculation time (hr:min:sec)**		
Unknowns Hd/ total number of contributors	***n***	**Minimum**	**Average**	**Maximum**
1/1	400	0:00:00	0:00:00	0:00:01
1/2	1200	0:00:01	0:00:01	0:00:07
2/2	7900	0:00:02	0:00:07	0:00:20
2/3	2400	0:00:03	0:00:09	0:00:24
3/3	10300	0:00:14	0:02:19	0:08:16
3/4	150	0:03:01	0:07:07	0:12:35
4/4	780	0:02:41	4:59:07	20:35:32

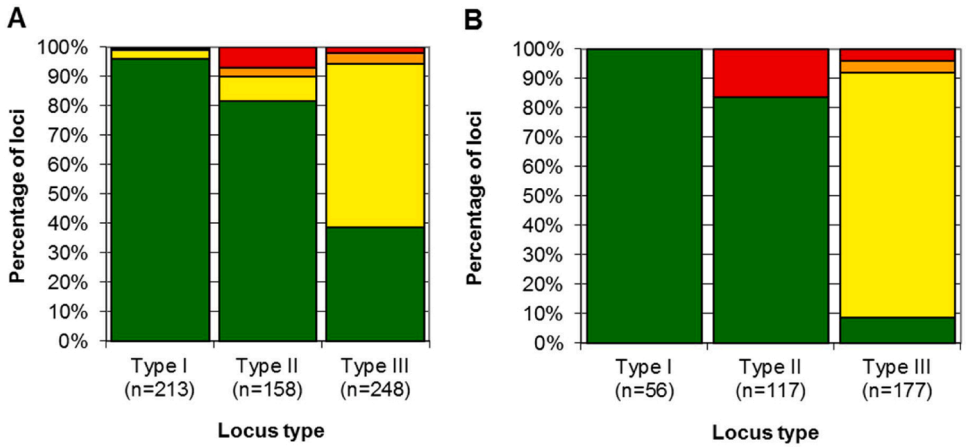

Fig. 2. LoCIM inference results for the data of lab 3 (A) and lab 4 (B). Overview of the number of Type I, Type II and Type III loci that yielded a correct major contributor inference (green), extra alleles (yellow), missing alleles (orange) or extra and missing alleles (red).

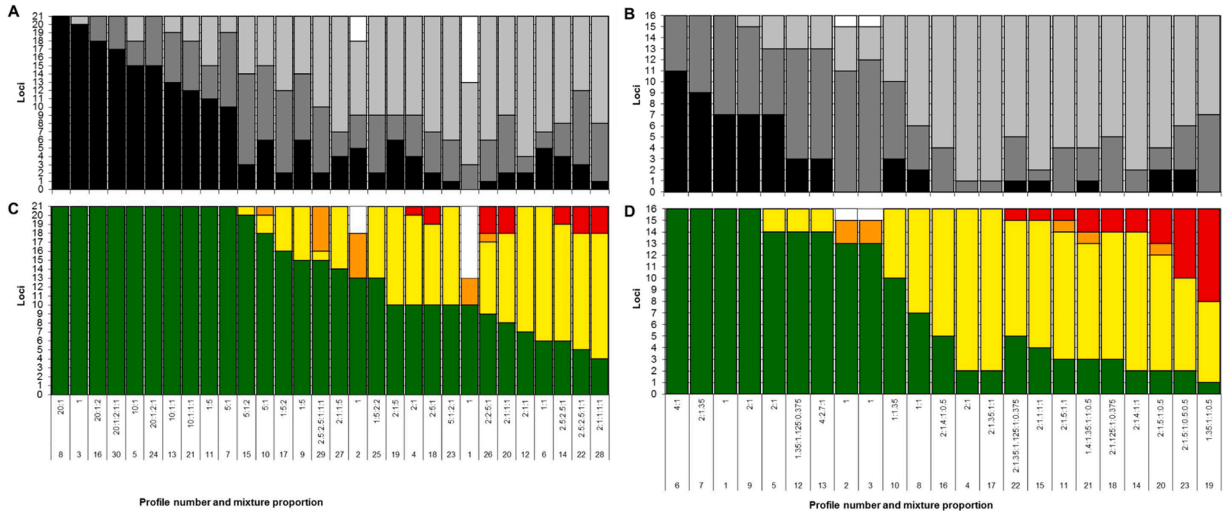

Fig. 3. LoCIM results per profile for of lab 3 (A & C) and lab 4 (B & D). A & B: Number of type I, II and III loci, shown as black, dark grey and light grey bars, respectively. White bars indicate locus drop-out. C & D: Inference results: green, correctly deduced; yellow, extra alleles; orange, alleles of the major are missing; red, both one/more extra allele(s) deduced and one/more allele(s) of the major is/are missing.

to expectations [12,14].

3.7. Estimation of the number of contributors using the generic RFC11 machine learning model

The generic RFC11 NOC model was applied to datasets of laboratories 1, 2, 3 and 4. Table 7 provides the percentage of correctly estimated NOCs when using this model and when using the MAC approach. The accuracies differ between laboratories as different datasets, with different complexities, were used. For few three person mixtures the MAC approach yielded a higher accuracy than the generic model. However, the overall results per dataset showed an increase in the percentage of correctly estimated NOC when using the generic RFC11 model in comparison to the conventional MAC approach (Table 7). A subset of PPF6C profiles from [15] were used in both the PPF6C RFC19 model and the generic RFC11 model. Results show, as expected, best

Table 7
Percentage of correctly estimated NOC for the datasets of lab 1, 2, 3 or 4 and applied to the MAC approach and generic RFC11 model.

	Percentage of correct predictions											
	Lab 1 PPF6C			Lab 2 GlobalFiler			Lab 3 GlobalFiler			Lab 4 NGMSE		
True NOC	*n*	MAC	Generic RFC11	*n*	MAC	Generic RFC11	*n*	MAC	Generic RFC11	*n*	MAC	Generic RFC11
1	3	100%	100%	3	66.7%	100%	8	100%	100%	3	100%	100%
2	14	71.4%	100%	12	91.7%	100%	12	33.3%	100%	12	75%	91.7%
3	14	100%	100%	12	83.3%	66.7%	12	83.3%	83.3%	12	58.3%	50%
4	14	21.4%	85.7%	12	58.3%	75%	12	16.7%	25%	12	58.3%	75%
5	3	33.3%	100%	3	0%	33.3%	3	0%	0%	3	0%	33.3%
Total	48	64.6%	95.8%	42	71.4%	78.6%	47	51.1%	70.2%	42	61.9%	71.4%
Increase in % correct predictions		31.2%			7.2%			19.1%			9.5%	

results using the RFC19 model (Supplementary Material 1) as this model is more specific to the data as it uses all loci, and more of the available information. These results show that it is preferable to have a NOC model that is developed for the specific data used in a laboratory. However, in absence of such model, or data to develop such model, the generic RFC11 model can be a useful alternative and serve as an addition to the reporting officer's toolbox to interpret mixed DNA profiles. Therefore, the Python code of the generic RFC11 model was translated to Java and implemented in DNAxs. The results that were obtained when the using the implementation in DNAxs were the same as those obtained by the Python code.

4. Concluding remarks

This study demonstrates a multi-laboratory application of DNAxs and its probabilistic genotyping module, DNAStatistX. The users' feedback and results obtained in this study yielded further insight in user-friendliness of the software, clarity of the manual and differences between software runs performed by different laboratories and using different datasets. These led to: 1) the development of an installer and installation manual to ease software installation, 2) a more detailed user manual, and 3) adjustments of the DNAStatistX optimizer settings which made the DNAStatistX module more robust and on average shortened the calculation times. Overall, the software performed as expected and DNAxs v2.0 (including DNAStatistX) was found valid for use in a multiple laboratories. Version 2.0 will be made available to the forensic community (see [12]). The dataset of laboratory 1 will be provided via https://www.forensicinstitute.nl/research-and-innovation/international-projects/dnaxs to aid the implementation of the DNAxs/DNAStatistX software within other laboratories.

Acknowledgements

This study was partly funded by the European Union's Internal Security Fund — Police (Proposal Number: 820838, Proposal Acronym: DNAxs2.0). The authors would like to thank Martin Pircher and Alexandra Kaindl-Lindinger for technical support and Martin Bodner (all Innsbruck) for helpful discussion.

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.2020.102390.

References

[1] C.C.G. Benschop, J. Hoogenboom, P. Hovers, M. Slagter, D. Kruise, R. Parag, et al., DNAxs/DNAStatistX: Development and validation of a software suite for the data management and probabilistic interpretation of DNA profiles, Forensic Sci. Int. Genet. 42 (2019) 81–89.

[2] European Network of Forensic Science Institutes (ENFSI), Best Practice Manual for the internal validation of probabilistic software to undertake DNA mixture interpretation ENFSI-BPM-DNA-01 issue 001, 17052017.

[3] Scientific working group on DNA analysis methods (SWGDAM), Guidelines for the Validation of Probabilistic Genotyping Systems, Accessed April 2019 (2015) http://media.wix.com/ugd/4344b0_22776006b67c4a32a5ffc04fe3b56515.pdf.

[4] Forensic Science Regulator, Software validation for DNA mixture interpretation, FSR-G-223 (1), 2018.

[5] M.D. Coble, J. Buckleton, J.M. Butler, T. Egeland, R. Fimmers, P. Gill, et al., DNA Commission of the International Society for Forensic Genetics: recommendations on the validation of software programs performing biostatistical calculations for forensic genetics applications, Forensic Sci. Int. Genet. 25 (2016) 191–197.

[6] J.A. Bright, I.W. Evett, D. Taylor, J.M. Curran, J. Buckleton, A series of recommended tests when validating probabilistic DNA profile interpretation software, Forensic Sci. Int. Genet. 14 (2015) 125–131.

[7] H. Haned, P. Gill, C. Lohmueller, C. Inman, N. Rudin, Validation of probabilistic genotyping software for use in forensic DNA casework: definitions and illustrations, Sci. Justice 56 (2016) 104–108.

[8] https://www.forensicinstitute.nl/research-and-innovation/international-projects/dnaxs.

[9] Ø. Bleka, G. Storvik, P. Gill, EuroForMix: an open source software based on a continuous model to evaluate STR DNA profiles from a mixture of contributors with artefacts, Forensic Sci. Int. Genet. 21 (2016) 35–44. http://euroformix.com.

[10] R.G. Cowell, T. Graversen, S. Lauritzen, J. Mortera, Analysis of forensic DNA mixtures with artefacts, J. R. Stat. Soc. Ser. C 64 (1) (2015) 1–48.

[11] A.A. Westen, T. Kraaijenbrink, E.A. Robles de Medina, J. Harteveld, P. Willemse, S. B. Zuniga, et al., Comparing six commercial autosomal STR kits in a large Dutch population sample, Forensic Sci. Int. Genet. 10 (2014) 55–63.

[12] C.C.G. Benschop, T. Sijen, LoCIM-tool: an expert's assistant for inferring the major contributor's alleles in mixed consensus DNA profiles, Forensic Sci. Int. Genet. 11 (2014) 154–165.

[13] 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 Sci. Int. Genet. 43 (2019), 102150.

[14] 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 Sci. Int. Genet. 25 (2016) 85–96.

[15] C.C.G. Benschop, A. Nijveld, F.E. Duijs, T. Sijen, An assessment of the performance of the probabilistic genotyping software EuroForMix: trends in likelihood ratios and analysis of Type I & II errors, Forensic Sci. Int. Genet. 42 (2019) 31–38.