A comparative study of qualitative and quantitative models used to interpret complex STR DNA profiles

Øyvind Bleka*[1,2], Corina C.G. Benschop[3], Geir Storvik[2], Peter Gill[1,4]

Abstract

The investigation of the performance of models to interpret complex DNA profiles is best undertaken using real DNA profiles. Here we used a data set to reflect the variety typically encountered in real casework. The "crime-stains" were constructed from known individuals and comprised a total of 59 diverse samples: pristine DNA/DNA extracted from blood, 2-3 person mixtures, degradation/no-degradation, differences in allele sharing, dropout/no dropout etc. Two siblings were also included in the test-set in order to challenge the systems. Two kinds of analyses were performed, namely tests on whether a person of interest is a contributor based on weight-of-evidence (likelihood ratio) calculations, and deconvolution test to estimate the profile of unknown constituent parts. The weight-of-evidence analyses compared *LRmix Studio* with *EuroForMix* including exploration of the effect of applying an *ad hoc* stutter-filter. For the deconvolution analysis we compared *EuroForMix* with *LoCIM-tool.* When we classified persons of interests into being true contributors or non-contributors, we found that *EuroForMix*, overall, returned a higher true positive rate for the same false positive levels compared to *LRmix.* In particular, in cases with an unknown major component, *EuroForMix* was more discriminating for mixtures where the person of interest was a minor contributor. Comparing deconvolution of major contributors we found that *EuroForMix* overall performed better than *LoCIM-tool.*

Keywords: NGM STR DNA, comparison study, weight-of-evidence, deconvolution

1 Introduction

Interpretation of short tandem repeat (STR) deoxyribonucleic acid (DNA) typing data is challenging when more than one individual contributes their biological material, especially when this is low quantity [1]. The likelihood ratio formula has become an established method in routine casework to report the weight-of-evidence of whether an individual is a contributor or not. In order to adapt this method to complex data (mixtures from several contributors), a number of advanced statistical models have been developed and implemented as software (e.g. *LoComatioN* [2], *LRmix* [3], *FST* [4], *LikeLTD* [5, 6], *Lab Retriever* [7], *STRmix*^{TM} [8], *EuroForMix* [9], *DNAmixtures* [10], *TrueAllele*® [11], *LiRaHt* [12] and *NOCIt*[13]). Some of these software are based on models that can take into account the variability of the quantitative information from the polymerase chain (PCR) products (i.e. *LikeLTD*, *STRmix*^{TM}, *DNAmixtures*, *EuroForMix*, *TrueAllele*®, *LiRaHt*, *NOCIt*), exploiting more of the data, while others (*LRmix*, *FST*, *LoComatioN*, *Lab Retriever*) only utilize presence/absence of alleles. Only a few studies have incorporated results from large numbers of complex data generated from biological material based on such models (e.g. [8, 13, 14, 15, 16, 17]). Here we use the material described in [18], and data described in [17] using open source software along with a new set of an accessible data (available at `www.euroformix.com/data`), to enable others to re-evaluate our analyses.

*Corresponding author at: Department of Forensic Biology, Norwegian Institute of Public Health, Rikshospitalet, Sognsvannsveien 20, 0372 Oslo, Norway. Tel.: +4721077643, E-mail address: Oyvind.Bleka@fhi.no

[1] Department of Forensic Biology, Norwegian Institute of Public Health, Oslo, Norway

[2] Department of Mathematics, University of Oslo, Oslo, Norway

[3] Division of biological traces, Netherlands Forensic Institute, The Hague, The Netherlands

[4] Department of Forensic Medicine, University of Oslo, Oslo, Norway

Haned et al. [17] and Benschop et al. [19] previously carried out a large data study with *LRmix/ LRmix Studio* (available through the R-package **forensim** [20] at `lrmixstudio.org`). With this method, hereafter described as the 'qualitative model' peak heights are not taken into account. A set of rules were defined in order to predict the number of contributors based on maximum allele count, distribution of allele counts over markers, the amount of allele drop-out and the total allele count. Using these rule-sets, the two articles assessed the sensitivity of the weight-of-evidence analysis using these rules relative to the underlying 'truth' (i.e. assuming the correct number of contributors and amount of drop-out). *EuroForMix* (available through the R-package **euroformix** at `www.euroformix.com`) is described as the 'quantitative model', utilising quantitative information (e.g. peak height, stutter) in a parameterized model. In the work presented here, a comparative study of the two models (*LRmix* and *EuroForMix*) was carried out to discover the differences and similarities for hypothesis testing (based on weight of evidence). In addition, the decision making process to predict the number of contributors was examined.

A complex profile typically includes two or more unknown contributors. Whenever possible, performing deconvolution, where an unknown component is extracted, can be valuable for different purposes, e.g. for searching a national DNA database without retrieving a large number of adventitious matches. However, an extraction can sometimes be very difficult as several allele combinations can be candidates for the unknown component source. By utilizing a statistical model for the quantitative information, *EuroForMix* infers the probability of different allele combinations for the unknown components, hence the uncertainty of assignments is quantifiable (see [9]). An alternative software to extract potential contributors, utilizing the peak heights, is *LoCIM-tool*, which is able to extract the major component and to categorize markers into levels of assignment difficulty [18]. We present a comparison of *EuroForMix* and *LoCIM-tool* with respect to the deconvolution problem.

In the following sections, data and methods used in the comparison study are presented. In the results section the important differences and similarities of the methods are presented for both the weight-of-evidence and for the deconvolution comparison. The supplementary material is copious, containing studies and detailed information which are important for the article.

2 Data

2.1 STR profiling

DNA profiles were generated using the NGM kit (Life Technologies) with 29 cycles and a 9700 thermal cycler (Life Technologies). Amplification products were separated by capillary electrophoresis (CE) on a 3130xl Genetic Analyzer (Life Technologies) at 3 kV for 5 s. The results were analyzed with GeneMapper® ID-X Software v.1.1.1 (Applied Biosystems) using a marker specific stutter filter as described in Westen et al. [21]. Alleles with peak heights below 50 relative fluorescence units (RFU) were removed in order to avoid baseline signal noise (i.e. the detection threshold was chosen as 50 RFU).

2.2 DNA profiles

A total of four two-person mixtures and 55 three-person mixtures were generated using known reference profiles of 33 individuals (subset described by Benschop and Sijen [18] and Haned et al. [17])[1]. Two siblings were included in the study, references 9A and 10B. Table 1 gives a summary of all the 59 samples where we let ":" separate the amount of DNA (pg) between the contributors. Table S1 in the supplementary material section B: "Details about data" shows a more detailed overview of the samples that were used in the comparison study.

Contributors typically consisted of a moderate-template component (i.e. a component with at least 100 picogram (pg) amount of DNA) together with one or more low-template component(s) i.e.

[1] All data can be found in the zip-file "NFIdata" at `www.euroformix.com/data`

Sample(s)	#contr.	DNA (pg)	Degraded
0.5.(1-4), 0.24.(1-4)	2	150:30	No
0.9.(1-4), 0.28.(1-4)	2	300:30	No
0.6.(1-4)	3	150:30:6	No
0.7.(1-4)	3	150:30:30	No
0.10.(1-4)	3	300:30:6	No
0.11.(1-4)	3	300:30:30	No
8.7d.(2-4)	3	500:250:250	Yes
9.6d.(2-4)	3	500:250:50	Yes
1.1, 2.1, 3.1, 6.1, 8.1, 9.1, 10.1, 11.1, 12.1, 14.1	3	100:50:50	Yes
1.2, 2.2, 3.2, 6.2, 8.2, 9.2, 10.2, 11.2, 12.2, 14.2	3	250:50:50	Yes
2.3, 3.3, 6.3, 8.3, 9.3, 10.3, 11.3, 12.3, 14.3	3	250:250:50	Yes
1.5, 2.5, 3.5, 6.5, 8.5, 9.5, 10.5, 11.5, 12.5, 14.5	3	500:50:50	Yes
1.6, 2.6, 3.6, 6.6, 8.6, 9.6, 10.6, 11.6, 12.6, 14.6	3	500:250:50	Yes

Table 1: The table gives a summary over all samples considered, with corresponding amounts of DNA (quantified in picograms (pg)) for the contributors. “#contr.” is the number of contributors and “DNA (pg)” denotes the amount of DNA for each contributors (separated by “:”). The bracketed information in the ’Sample(s)’ column denotes the replicate number, e.g. (2-4) means the replicates ’2’, ’3’ and ’4’. The first eight samples include components that are low-template (i.e. less than 50 pg). The next two samples, ’8.7d’, and ’9.6d’ have components with more than 50 pg but are greatly degraded. The rest of the samples consist of one replicate, but with different amounts of DNA. “Degraded” indicates whether the samples are degraded or not.

components with 30-50 pg of DNA. Ten of the samples were replicated, originating from separate amplifications of the same DNA extract. Eight of these samples had four non-degraded replicates with little DNA. The last two of these sample (sample ’8.7d’ and ’9.6d’) contained three very degraded replicates (i.e. the peak heights decreased as the fragment lengths increased). All the other samples were non-replicates and were degraded by varying degrees.

The number of allele dropout events are determined by counting the number of alleles in the reference that has corresponding peak height below 50 RFU (homozygotes were counted twice). For replicates, this number was summed up across all samples. All the samples had low-template components with drop-outs (see Table S1 in section B in supplementary material). The moderate-template components had mostly no drop-out (but sometimes one or two and even up to six), except for the very degraded samples ’8.7d’ and ’9.6d’ who had 42 and 38 drop-outs respectively. The number of drop-outs for the low-template components varied from 0 to 14 for contributors in the non-replicated samples and from 0 to 62 for replicated samples.

Replicates

Replicates are defined as DNA profiles obtained from independent PCR amplifications from the same DNA extract. All replicates within a sample were amplified simultaneously using the same PCR plate and PCR machine. For low template, stochastic effects cause much variation in peak height, heterozygote balance and drop-out [22].

Stutter filter

Stutters from an allele a occur due to strand slippage during PCR ([23]), typically resulting in ’back stutter’ of $a-1$ STR repeat unit (-4bp for a tetrameric repeat). Other stutter-artefacts can also be observed at allele $a-2$ repeat units (i.e. double back stutter) and $a+1$ repeat units (i.e. forward stutter [24]), however these occur less often, and are usually much smaller in peak height[2]. A filter can be optionally applied in GeneMapper to remove alleles that are coincident with stutters in both $a-1$ and $a+1$ positions [25]. In practice, stutter filters are calibrated based on the average stutter peak height ±3 standard deviations (SD) per marker. It does not consider stutter peak variation on an allelic basis. A problematic situation occurs when contributors with large amounts of DNA (major contributors)

[2] typically falling below the detection threshold of 50 RFU

produce stutters that are similar in peak height to the alleles from minor contributors. Therefore there is no guarantee that alleles from true contributors will not be removed as well. Application of the stutter model has the effect of slowing the speed of calculation which may be problematic if there are a large number of contributors. If a major contributor is the POI, then pretreatment with the GeneMapper stutter filter is an acceptable way forward. If minor contributors are evidential and their alleles are at the same peak height as backward or forward stutter, then we naturally approach the limits of interpretation. However, results from section E.1: "Comparison of the stutter model in *EuroForMix* versus GeneMapper stutter filter" in supplementary material show that the GeneMapper stutter filter is sometimes useful for evidential minors as well.

2.3 Allele frequency database

A total of 2085 Dutch male donors were typed with the NGM kit in order to create a representative population database for the allele frequencies [21]. With this number of samples, we found that the uncertainty of allele sampling has some effect on likelihood ratio (LR) calculations, but not large: from a consideration of ten samples with six references each, we found that the width of the 90% LR coverage interval is typically up to $10^{0.25}$ for *LRmix* and up to $10^{0.55}$ for *EuroForMix* (see details in supplementary material section C: "The sampling effect of allele frequencies"). All of the references used in this study are a subset from this typed population except for the ones contributing to the samples of type "0.x" (i.e. the samples having four replicates).

2.4 Design of experiment

For weight-of-evidence calculations, a person of interest (POI) is compared with a given mixture sample (see Table 1).

1) For a given mixture sample (out of the 59 samples), the POI is considered as each of the 33 reference samples in turn, giving 33 comparisons per mixture sample. Only two or three are actual contributors, the rest are non-contributors.

2) For 29 of the mixture samples, one of the contributors may be conditioned as *a priori* 'known' beforehand (listed under "Above Ts" in Table S1 in section B in supplementary material). These contributors had most of their peak heights above a stochastic threshold of T_s=175 RFU. This gives an additional 32 comparisons for each of the 29 mixture samples.

The stochastic threshold is an estimated RFU where one of the alleles in a heterozygote pair drops out with a defined probability. With the method tested here the probability of allele dropout is less than 0.01 when the remaining allele has peak height equal or greater than 175 RFU. See Gill et al. [26] for a method of determination. By repeating 1) for all 59 mixture samples and 2) for the 29 "conditioning" mixture samples, we end up with 228 comparisons where the POI is a true contributor, and 2646 comparisons where the POI is a non-contributor. Notice that the latter number reduces to 2634 when comparisons involving siblings were omitted.

3 Method

In this comparison study we compared the 'qualitative model' *LRmix* versus the 'quantitative model' *EuroForMix* for weight-of-evidence and hypothesis testing, and *EuroForMix* versus *LoCIM-tool* to estimate the most likely profile of the unknown major component. Note that the statistical models assumed in *LRmix* and *EuroForMix* require that the number of contributors is specified, whereas this is not the case for *LoCIM-tool*. In this section we introduce the different models and define how the number of contributors are estimated, along with the other unknown parameters used in the two models.

3.1 The likelihood ratio formula

To report a weight-of-evidence quantity to determine if a person of interest (POI) is a contributor to the sample E or not, the likelihood ratio (LR) formula is used. This is given as $LR = \frac{P(E|H_p)}{P(E|H_d)}$ where the hypotheses H_p: "POI contributes to the sample" and H_d: "POI does not contribute to the sample" are compared.

With l as a specific marker and allele a as one of the possible alleles in the population, the observed sample E is given as a set of peak heights $\{y_{l,a}\}$. If the peak height is below the detection threshold T, only this binary information is recorded. We use $T = 50$ RFU in this work. The LR method presented here requires that the number of contributors to the sample, K, are specified.

For the NGM kit, the evidence E consists of a total of $L = 15$ markers (excluding the amelogenin marker), such that $E = (E_1, ..., E_L)$ and $E_l = (y_{l,1}, ..., y_{l,A_l})$ where A_l is the number of alleles at marker l. By assuming that the observations at the different markers are independent for a given hypothesis $H = H_p$ or $H = H_d$, the LR is given by

$$LR = \frac{P(E|H_p)}{P(E|H_d)} = \frac{\prod_{l=1}^{L} P(E_l|H_p)}{\prod_{l=1}^{L} P(E_l|H_d)}. \tag{1}$$

The quantities $P(E_l|H_p)$ and $P(E_l|H_d)$ will depend upon the genotype(s) from the contributors. If there is only one contributor and $S_{l,1}$ is the (known) locus genotype of POI at marker l, $P(E_l|H_p) = P(E_l|S_{l,1})$. For K contributors with only POI (here the first contributor) known,

$$P(E_l[H_p) = \sum_{S_{l,2},...,S_{l,K}} P(E_l|S_{l,1}, S_{l,2}..., S_{l,K})P(S_{l,2}..., S_{l,K}|H_p). \tag{2}$$

Under H_d, assuming all K contributors are unknown,

$$P(E_l|H_d) = \sum_{S_{l,1},S_{l,2},...,S_{l,K}} P(E_l|S_{l,1}, S_{l,2}..., S_{l,K})P(S_{l,1}, S_{l,2}..., S_{l,K}|H_d). \tag{3}$$

The probabilities $P(E_l|\mathbf{S}_l)$ where $\mathbf{S}_l = (S_{l,1}, S_{l,2}..., S_{l,K})$, are defined through statistical models (including model parameters that need to be specified) and depend upon the number of alleles of type a in the locus genotype of contributor k (i.e. $S_{l,k}$). The probabilities $P(\mathbf{S}_l|H)$ uses the allele frequencies described in section 2.3.

When a reference V is known to be a contributor, the alternative hypothesis set becomes H_p: "POI and V contribute to the sample E" and H_d: "V contributes to the sample E, whereas POI does not". The equation (1) still holds, assuming that the reference V corresponds to contributor 2, the sums in equations (2) and (3) fix $S_{l,2}$ to the locus genotype of the reference, and $P(S_{l,2}|H_p) = P(S_{l,2}|H_d) = 1$.

LRmix only utilizes the binary information $y_{l,a} \geq T$. The statistical model behind *LRmix* introduces for each contributor a parameter d_k which is defined to be the probability of drop-out of an allele for contributor k. We follow the methodology in Haned et al. [17] and assume that the drop-out parameters are the same for all contributors, $d_1 = ... = d_K$ (i.e. the *BasicDrop* model), except for the situation when a reference V is known to be a contributor. For this situation, we fix the drop-out parameter of V to zero (i.e. the *SplitDrop* model). See technical model specification of *LRmix* in Appendix section A.1.

EuroForMix assumes the peak heights $y_{l,a}$ to be gamma distributed, where peak heights below T are truncated to zero. The parameters contained in the statistical model are the expectation and coefficient of variation of a heterozygote peak heights, mixture proportion for each contributor and an exponential decaying degradation slope parameter. *EuroForMix* also incorporates a model for back-stutters (-4bp) by including an expected stutter proportion parameter. In the supplementary material section D: "Validation data" we carried out a study based on 30 sample replicates for three different dilutions (20 pg, 25 pg and 30 pg amount of DNA). We found that the statistical model for the peak heights was adequate when compared with the empirical peak height variability and drop-out distribution. See technical model specification of *EuroForMix* in Appendix section A.2.

Replicates

More generally we can have R number of replicates of the sample information E_l, given by $E_l^{(1)}, ..., E_l^{(R)}$. We assumed these replicates to be independent and to include the same contributors, such that $P(E_l|\mathbf{S}_l) = \prod_{r=1}^{R} P(E_l^{(r)}|\mathbf{S}_l)$. In this work we assumed that the model parameters for *LRmix* and *EuroForMix* were constant across all markers and the same for all replicates.

Sub-population structuring

Both *LRmix* and *EuroForMix* include a model for $P(\mathbf{S}_l|H)$ to adjust for sub-population relatedness using the coancestry coefficient F_{st} [27]. For all analyses we follow Haned et al. [17] by applying $F_{st} = 0.01$ in order to accommodate the possibility that contributors belong to a sub-population of the population database.

Drop-in

A set of $N = 14757$ negative control samples[3] were generated with the same settings as described in section 2.1. From this data, a total of $x = 80$ false positive alleles were found (excluding the amelogenin marker), so that the relative frequency of drop-in per STR marker (out of total $L = 15$ markers) was estimated as $\frac{x}{N \times L} = 0.00036$. By assuming a shifted exponential distribution starting from 50 RFU as a model for all allele drop-in peak heights (similar to Taylor et al. [8], but different from Puch-Solis [28]), the maximum likelihood estimate for the rate parameter is $\lambda = 0.02$, and this was used as a plug-in value to model the drop-in peak height in *EuroForMix* (see section D.3: "Drop-in data" in the supplementary material for other model suggestions). From section E.2: "Application of the drop-in model" in the supplementary material we describe how the drop-in model was implemented in *LRmix* and *EuroForMix.* In section E.3: "The effect of applying the drop-in model to accommodate an extra allele" we demonstrate the effect of a spurious allele drop-in. Here we found that *EuroForMix* was relatively insensitive to allele drop-in provided that the event was a small peak height. If there is no drop-in then the model makes no difference to the LR. Drop-in should not be used to explain more than one mismatching allele per profile [1].

3.2 Model inference

Maximum likelihood estimation

Appendices A.1 and A.2 describe how a set of locus genotypes $\mathbf{S}_l$ is related to the observed sample E_l by assuming statistical models for $P(E_l|\mathbf{S}_l)$. Within these models, a set of unknown parameters, θ_p under hypothesis H_p and θ_d under hypothesis H_d are involved. The probabilities of the evidence $P(E|H)$ in equation (1) (where H is either H_p or H_d) are not completely defined without specification of the parameters involved. By following a maximum likelihood estimation approach we infer $P(E|H)$ with $P(E|H, \hat{\theta})$ where $\hat{\theta} = \arg\max_\theta P(E|H, \theta)$ is the maximum likelihood estimate (MLE) for the model parameters θ. Doing so we construct the maximum likelihood based LR quantity (the MLE method) as

$$LR = \frac{P(E|H_p, \hat{\theta}_p)}{P(E|H_d, \hat{\theta}_d)}. \tag{4}$$

The MLE method is one of the outputs for *EuroForMix* (in addition to the Bayesian approach). This method was also applied to *LRmix* in order to compare results, in addition to the standard conservative method (see paragraph "Bayesian and conservative LR quantities" later in this section).

[3] such samples are not expected to contain any DNA

Model selection

Appendices A.1 and A.2 describe the parametric models for *LRmix* and *EuroForMix* where the number of contributors are specified. However this number is typically unknown in real casework. One possible framework to take care of this is to predict it by establishing a criterion. For instance, the criterion could be based on the maximum number of alleles found at any markers in the profile (MAC), or the total number of alleles in the sample (TAC) (see other criteria in supplementary material 1 in Benschop et al. [19] which were based on samples with no or little dropout). However, these criteria are typically based on samples without drop-out. In this work we predicted the number of contributors used for the *LRmix* and *EuroForMix* results based on their corresponding parametric model themselves.

In the supplementary material section F.1: "Estimating number of contributors and drop-out parameter in *LRmix*" we performed a simulation study to show how the maximum likelihood for the *LRmix* model can be used as a criterion for estimating the number of contributors for different degree of allele drop-out. Here we found that penalizing the logarithm of the maximum likelihood value with the assumed number of contributors was necessary to avoid overestimation (but with the cost of being more likely to underestimate). We applied this criterion to predict the number of contributors for the *LRmix* model.

For *EuroForMix*, in addition to the number of contributors, other model alternatives include optional use of the stutter model and the degradation model. To select the optimum model $\widehat{M}$, the framework described by Bleka et al. [9] is followed, using the Akaike information criterion (AIC). For a specific model M with the inferred model $P_M(E|H,\hat{\theta})$ from section 3.2, the criterion is defined as $AIC_M = -2\log P_M(E|H,\hat{\theta}) + 2|\theta|$ where $|\theta|$ is number of parameters in the model. The optimum model out of a model set $\mathbb{M}$ is selected as the one with smallest AIC_M (i.e. $\widehat{M} = \arg\min_{M\in\mathbb{M}} AIC_M$).

The profile genotype probability for the unknown contributors is part of both models (*LRmix* and *EuroForMix*) and is influenced by the value of F_{st}. An increase of F_{st} also increases the likelihood of allele sharing so that a model with more contributors is more likely. We considered two options, $F_{st} = 0$ or $F_{st} = 0.01$, and selected the one with largest maximum likelihood.

Bayesian and conservative LR quantities

The likelihood ratio (LR) calculations in section 3.2 are based on the maximum likelihood estimated parameters (i.e. single points in the parameter space), which are most likely to explain the data. Such inference does not take into account the uncertainty of the estimators which again leads to an uncertainty in the LR quantity.

The "full" Bayesian approach takes into account the uncertainty of the model parameters by calculating the integral $\int_\theta p(E|H,\theta)p(\theta|H)d\theta$, where an *a priori* distribution on the model parameters, $p(\theta|H)$, has been assumed. For low-dimensional parameter sets, such integrals can be calculated by standard integration techniques, while for high-dimensional cases, Monte Carlo approaches can be applied.

An alternative approach [29] is to consider $LR = LR(\theta_p, \theta_d)$ as a function of the parameters involved and derive the posterior distribution of LR through the posterior distribution of the parameters given the data E. Such a distribution can be approximated by Monte Carlo simulations. A conservative approach is to use a lower quantile of the LR distribution as a measure of evidence.

We assumed that all parameters involved are *a priori* independent and uniformly distributed. We generated 1000 samples from the posterior distribution of the model parameters under each hypothesis to get the two sample vectors, $\tilde{\theta}_p = \{P(E|H_p,\theta_p^{(1)}),...,P(E|H_p,\theta_p^{(1000)})\}$ under H_p and $\tilde{\theta}_d = \{P(E|H_d,\theta_d^{(1)}),...,P(E|H_d,\theta_d^{(1000)})\}$ under H_d. The main differences between the method used for *EuroForMix* and *LRmix* is that for *LRmix* we followed the strategy in [29] where only the total number of alleles in the sample (TAC) was used as data, whereas for *EuroForMix* the full data E was used. For *EuroForMix* we constructed the random variable $LR = LR(\tilde{\theta}_p, \tilde{\theta}_d)$, whereas for *LRmix* we followed [29] and created the two separate random variables $LR_p = LR(\tilde{\theta}_p, \tilde{\theta}_p)$ and $LR_d = LR(\tilde{\theta}_d, \tilde{\theta}_d)$ where the lower 0.05-quantile of each was calculated. The smallest of these two values were then used as the "conservative" LR quantity. For *EuroForMix* we simply used the lower 0.05-quantile of LR as the "conservative" LR quantity.

Deconvolution

Both *EuroForMix* and *LoCIM-tool* are able to perform deconvolution by utilizing the peak height information, meaning that they are capable of inferring the most likely profile genotypes (i.e. DNA profiles) for the unknown contributor(s) in a sample. However, the two software differ.

LoCIM-tool requires that the stochastic threshold, the heterozygote imbalance threshold and major-to-minor(s) proportions are informed beforehand from validation data as described by Benschop and Sijen [18]. If replicates are analysed, a consensus profile is created by keeping the alleles which are presented in at least half of the replicates which are summed across all replicates. The summation, plus the parameters of the stochastic threshold, heterozygote balance and major to minor(s) ratio, are used to classify every marker as a type '1', type '2' or type '3', representing classes of increasing complexity. Based on the type of marker, *LoCIM-tool* applies an inclusion percentage to deduce the alleles for the major contributor. Type '3' markers are most complex and its inclusion percentage is lower compared to type '1' and '2' markers which are aimed at inferring the major contributor's alleles. Note that the method does not require that the number of contributors is specified; it is only suitable for extracting the major contributor, and it is not possible to condition on known profile(s).

EuroForMix applies a statistical model which consists of a set of parameters which are inferred by maximizing the likelihood function. Given the estimated model parameters $\hat{\theta}$, each marker can be handled independently and the probability of a specific locus genotype combination $\mathbf{S}_l$ for the unknown contributors for marker l is calculated (using Bayes' theorem) as

$$p(\mathbf{S}_l|E_l, H, \hat{\theta}) = \frac{p(E_l|\mathbf{S}_l, \hat{\theta})p(\mathbf{S}_l|H)}{p(E_l|H, \hat{\theta})}.$$

From this, the marginal probability of a locus genotype $g_{l,k}$ for each contributor $k = 1, .., K$ is calculated as

$$p_{l,k}(g_{l,k}) = \sum_{\mathbf{S}_l : S_{l,k} = g_{l,k}} p(\mathbf{S}_l|E_l, H, \hat{\theta}).$$

EuroForMix calculates the marginal probabilities $p_{l,k}$ for all possible locus genotypes $g_{l,k}$ for each unknown contributor k and marker l and ranks them. These probabilities can then be used to provide information to determine whether the locus genotype $g_{l,k}$ for an unknown contributor k is likely or not. The most likely genotype is used as the predicted genotype. A predicted genotype for a given marker is flagged as 'certain' if its probability is at least twice as large as the second likeliest genotype possibility.

4 Results

4.1 Inferring the number of contributors

Three methods were used to predict the number of contributors: The AIC based on the quantitative model (*EuroForMix*); the penalized maximum likelihood value based on the qualitative model (*LRmix*) and by manual inspection (MI) (i.e. checking maximum allele count, peak height variability etc. with visual inspection) carried out by a forensic scientist. Table S4 and Table S5 in supplementary material section F.2: "Inferred number of contributors for all samples" show an overview of the inferred number of contributors for the qualitative and quantitative models, and the criterion difference to the second most likely alternative number of contributors.

The numbers of contributors were incorrectly estimated on several occasions: out of 59 samples there were eight occurrences for the qualitative model, seven for the quantitative model and three using manual inspection:

Two contributors instead of three:

- qualitative model, samples '0.6', '3.5', '6.6'
- quantitative model, samples '0.6', '0.7', '9.2'
- manual inspection, samples '6.1', '6.2', '9.2'

Four contributors instead of three:

- qualitative model, samples '8.1', '8.5', '9.5', '10.6', '11.2'
- quantitative model, samples '2.5', '8.2', '8.5', '9.6d'
- manual inspection, none

For situations where the models underestimated the number of contributors, there was a lot of drop-out and/or a high amount of allele sharing between the contributors, which reduced the number of alleles observed. It is more difficult to explain why the models sometimes overestimated the number of contributors. However where there are few dropouts and/or there are drop-ins/stutters combined with little allele sharing between the contributors this will be a factor.

Estimating the number of contributors using the quantitative model was sometimes difficult when there were several minor components together with one major component (i.e. three-person samples with 250:50:50 and 500:50:50 in DNA amount (pg)). Here the peak height levels of the minor components are small leading to high interpretation uncertainty, as they are inseparable. The manual inspection did not overestimate the number of contributors since a visual inspection would tend to be biased towards a small amount of drop-out.

The numbers of contributors inferred using the quantitative model was further used for likelihood ratio calculations and estimation of the profile genotypes for the unknown components using *EuroForMix*, while the numbers of contributors inferred using the qualitative model was used for the likelihood ratio calculations using *LRmix*.

4.2 Using receiver operating characteristic (ROC) plots to compare the efficiency of different models

The plots in Figure 1 show the relationship between likelihood ratios (LRs) obtained by *LRmix* and *EuroForMix*, comparisons between siblings 9A and 10B were omitted. The corresponding number of points which fall below/above $LR = 1$ (this corresponds to $log_{10}LR = 0$) for each of the methods for the plots in Figure 1 are given in Table 2. When the POI is the true contributor (the left hand plots where H_p is true), the LR values from *EuroForMix* (quantitative model) are almost always greater than those from *LRmix* (qualitative model). For these comparisons using the MLE method, there were 28 cases with *LRmix* and five cases with *EuroForMix* (out of a total of 228) where the POI was below $LR = 1$. With the conservative method, these numbers increased to 67 and 11, respectively. Considering the cases where the POI is a non-contributor (the right hand plots where H_d is true), it was shown that most of the non-contributors are below $LR = 1$ for both models and methods. However for the MLE method there are a number of small positive values (LR values just above one), 17 with *LRmix* and 121 with *EuroForMix* (out of a total of 2634). When the conservative method was used, these numbers were reduced downwards to four and five, respectively.

An alternative way to represent the information in Figure 1 is to create receiver operating characteristic (ROC) plots as shown in Figure 2 (comparisons between samples regarding the siblings 9A and 10B omitted). These are created by plotting the true positive rate versus the false positive rate at various threshold settings relative to $LR = t$. This corresponds to a decision rule where H_d is rejected if $LR > t$. For instance if $t = 1$, and H_p is true while giving $LR < 1$, this is defined as a false negative.

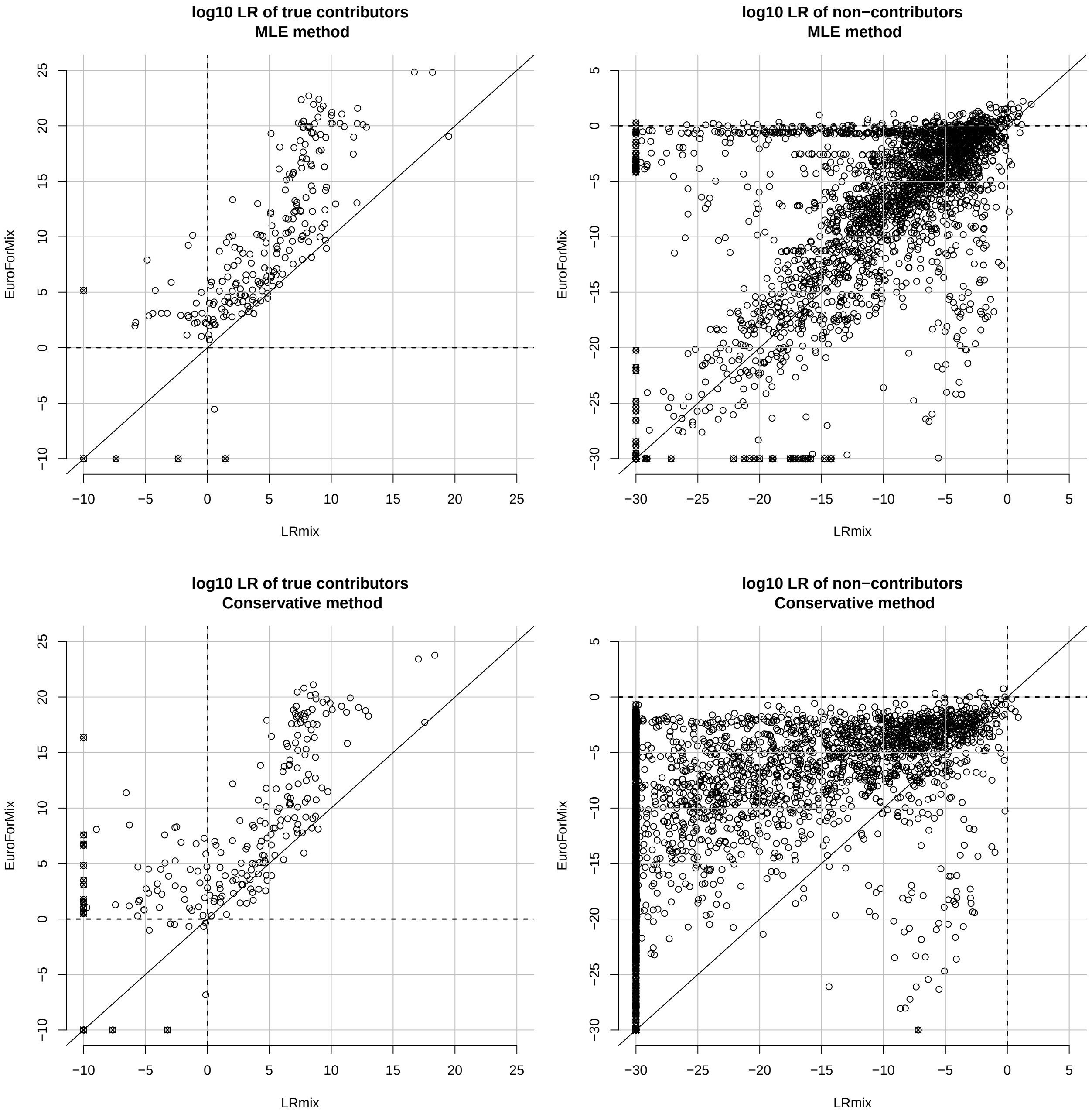

Figure 1: The plots show the likelihood ratio (LR) quantity (logarithmic scale with base 10) comparing *LRmix* (qualitative) along the horizontal axis and *EuroForMix* (quantitative) along the vertical axis. Left plots are H_p is true, while the right plots are H_d is true. The crossed points indicate that at least one of the methods had a value $LR < 10^{-10}$. The two upper plots show the MLE based method, while the two lower plots show the conservative based method.

Conversely, a true positive occurs if $LR > 1$. A false positive is defined if H_p is false and $LR > 1$. Figure 2 shows the proportion of false positives along the x-axis and the proportion of true positives along the y-axis.

In Figure 2, threshold $LR = 1$ is denoted as a cross on each curve with the precise number of instances counted in Table 2. For the MLE method (with $t = 1$), the false positive rate is 0.006 and 0.046 for *LRmix* and *EuroForMix*, respectively. The corresponding true positive rates are 0.88 for *LRmix* and 0.98 for *EuroForMix.* For the conservative method the false positive rate is 0.002 for both models, however with a true positive rate of 0.71 for *LRmix* and 0.95 for *EuroForMix.* An ideal model gives a ROC plot which simultaneously shows a false positive rate equal zero and a true positive rate equal one for some values of t (yielding a point in the upper left corner in the ROC curve). In Figure 2

Method	Truth		$LR_{LRmix} < 1$	$LR_{LRmix} \geq 1$	Total
MLE	H_p	$LR_{EFM} < 1$	3	2	5
		$LR_{EFM} \geq 1$	25	198	223
		Total	28	200	228
	H_d	$LR_{EFM} < 1$	2509	4	2513
		$LR_{EFM} \geq 1$	108	13	121
		Total	2617	17	2634
CONS	H_p	$LR_{EFM} < 1$	11	0	11
		$LR_{EFM} \geq 1$	56	161	217
		Total	67	161	228
	H_d	$LR_{EFM} < 1$	2625	4	2629
		$LR_{EFM} \geq 1$	5	0	5
		Total	2630	4	2634

Table 2: The counts of the number of observations where LR is smaller or greater than one for *LRmix* (given by LR_{LRmix}) and *EuroFormix* (given by LR_{EFM}). The table shows the number of observations when either H_p is true (i.e. considering true contributors) or H_d is true (i.e. considering non-contributors) for the maximum likelihood LR based method (MLE) or the conservative LR based method (CONS).

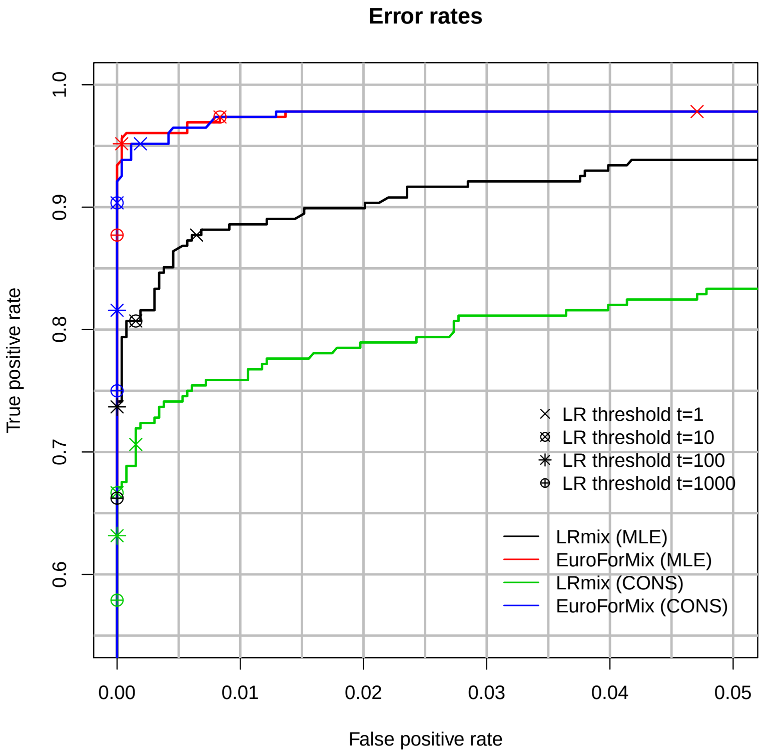

Figure 2: Receiver operating characteristic (ROC) plot where the rate of false positives (FP) (along horizontal axis) and true positives (TP) (along vertical axis) are plotted as a function of LR thresholds. The plot shows the results for the maximum likelihood estimation method (MLE) and the conservative method (CONS) for both *LRmix* and *EuroForMix*. The points on the curves show the FP and TP rates for different LR thresholds.

this condition is approached more effectively with *EuroForMix* than with *LRmix*. Furthermore, MLE performed better than the conservative method. In casework, we would always prefer to minimize the number of false positives, hence the conservative method would apparently be preferred, even though the incidence of false negatives is increased. Note however that this could also be obtained using MLE by changing the threshold.

False positives ($LR > 1|H_p$ is FALSE)

Table 3 describes the false positives observed with *EuroForMix* and *LRmix* (using the conservative method). Notice here that reference 9A is a true contributor in the samples "9.x", while reference 10B is a true contributor in the samples "10.x". Recall that samples 9A and 10B were siblings, sharing 19 out of 30 alleles. *LRmix* does not take into account peak heights, hence high LR values were recorded when a sibling of the true contributor was compared. There was a single high ($LR > 200$) *EuroForMix* result for sample '10.5' where the POI 9A was low level with 15 allele drop-outs making the comparison very unreliable. The conservative method reduced most of the LR values considerably, at the cost of introducing additional six false negatives for *EuroForMix* and 39 for *LRmix*. *LRmix Studio* (`www.lrmixstudio.org`) has the possibility of replacing one unknown under the defense hypothesis by a relative of the POI. When this is carried out for the samples where the POI is a sibling to one of the true contributors (e.g. for sample '9.2' H_p: "10B is contributor" versus H_d: "A sibling of 10B is contributor") the LR values are close to 1, which means that the sample can be explained by either the POI or a sibling. For cases where the POI was not a sibling to one of the true contributors, the false positive matches for *EuroForMix* all gave $LR < 200$, and for *LRmix* $LR < 100$, when the MLE method was used, and $LR < 10$ for both when the conservative method was used.

Table (A): False positives, siblings

Sample	POI|cond	#d	Model	K	$\hat{K}$	MLE LR	Cons LR	Cons sibling LR
9.3	10B	3	*LRmix*	3	3	959356	70	3
9.2	10B	8	*LRmix*	3	3	236768	161385	1e-5
9.5	10B	5	*LRmix*	3	4	165044	138318	12
10.6	9A|10A	6	*LRmix*	3	4	55972	3193	1
10.6	9A	6	*LRmix*	3	4	1327	9	0.02
10.5	9A	15	*EuroForMix*	3	3	288	8	na
10.2	9A	13	*EuroForMix*	3	3	72	2	na
9.3	10B	3	*EuroForMix*	3	3	23	1	na
10.3	9A	7	*EuroForMix*	3	3	20	1	na

Table (B): False positives, non-related

Sample	POI|cond	#d	Model	K	$\hat{K}$	MLE LR	Cons LR	Cons sibling LR
8.7d	3C	17.7	*EuroForMix*	3	3	162	6	na
8.7d	6B	19.3	*EuroForMix*	3	3	92	1	na
8.5	10B|8A	10	*LRmix*	3	4	85	2	0.7
8.5	9A|8A	9	*EuroForMix*	3	4	41	1	na
3.3	14C|3B	11	*EuroForMix*	3	3	35	3	na
2.1	1A	7	*EuroForMix*	3	3	30	2	na
3.2	11B	8	*LRmix*	3	3	14	8	0.8
8.5	1B|8A	8	*LRmix*	3	4	9	2	0.2
11.2	8C	7	*LRmix*	3	4	6	4	0.7

Table 3: Table (A) and (B) list LRs for all false positive errors with threshold $t = 1$ when the conservative method was used. Table (A) shows the false positive results where the POI is sibling to one of the true contributors in the sample, while Table (B) shows the false positive results where the POI is non-related to the contributors in the sample. "POI|cond" is the compared person of interest with possible conditional reference. #d is the number of dropouts for the person of interest (average if replicates where available). K and $\hat{K}$ are the true and the predicted number of contributors, respectively. "Cons sibling LR" indicates the conservative LR, computed using *LRmix Studio*, where under H_d :"A sibling of POI is contributor".

False negatives ($LR < 1|H_p$ is TRUE)

The left hand plots of Figure 1 show LRs where the POI is a true contributor. All false negatives (i.e. $LR < 1|H_p$ is 'TRUE') are listed in Table S6 and Table S7 in supplementary material section G.1: "False negative results". The number of false negatives are lowest when the MLE method is utilised (28 for *LRmix* and five for *EuroForMix*), and increased when the conservative method is used (67 for *LRmix* and 11 for *EuroForMix*). However, all the false negatives encountered can be characterised as originating from a particular category of mixtures. Minor components with less than or equal

50 pg contributions, except for the very degraded sample '9.6d', accounted for all situations of the false negatives. Considering the MLE method, the smallest observed number of dropouts for the POI was four (in average across replicates) for *LRmix* and 12 for *EuroForMix*; for the conservative method, there were two and eight observations, respectively. This information could be used to derive a complexity threshold. For example, a guideline may state: "if less than x alleles match the POI, the sample is regarded too complex for further statistical analysis using the conservative method of *LRmix*".

LR values as a function of allele drop-outs

Mixture interpretation is often considered to be at the borderline when proportions of 10:1 are encountered, the peak heights of the minor contributor are close to the detection threshold and stutters from the major contributor are similar in peak height. Figure 3 shows how the LR values based on the MLE method using *LRmix* and *EuroForMix* are related to the number of drop-outs (for the POI) for instances where the POI was a true contributor or not. For true contributors we also indicated whether it was minor. From the plot we observed (for both *LRmix* and *EuroForMix*) that false positives were occasionally observed with six or more dropout events but the corresponding LR was always low (< 200). When H_p is true, the limit for observing $LR > 1$ with *LRmix* was up to three allele drop-outs before false negatives $LR < 1$ were recorded; for *EuroForMix* up to 11 allele drop-outs, before false negatives $LR < 1$ were recorded.

4.3 Comparison of deconvolution methods

In the first part of the deconvolution study we predict the genotypes of the major component (in specific samples). Both *LoCIM-tool* and *EuroForMix* are suitable for this part. In the second part we predict the genotypes also for non-major components. Here we only investigate the performance of *EuroForMix*, since *LoCIM-tool* is not suitable for this part.

LoCIM-tool and *EuroForMix* were used to compare the most likely deconvolved profile genotype (i.e. the set of predicted locus genotypes over all markers) with the corresponding 'true' reference profile in the dataset. A comparison between the predicted locus genotype and the genotype of the 'true' reference profile returned either a 'full match', meaning that the two genotypes were the same, a 'partial match', meaning that only one allele was shared between the two genotypes, and 'no match', meaning that no alleles were shared between the two genotypes. We classified a predicted locus genotype as either 'certain' or 'uncertain', where the criterion of this differed for the two methods. For *EuroForMix*, a prediction of a locus genotype was 'certain' when the most likely deconvolved locus genotype had a probability at least twice as large as the second likeliest locus genotype. For *LoCIM-tool*, every locus is classified into marker type '1', type '2' or type '3', which represent classes of increasing complexity for inferring a locus genotype for the major contributor. For the purpose of this study, we regarded the type '1' and type '2' markers as 'certain' and type '3' markers as 'uncertain' for locus genotype predictions. Note that such strict per marker criteria are not recommended for casework, where forensic scientists also examine the overall profile, together with the total number of markers per marker type, and the independent replicates, if available. The use of 'certain'/'uncertain' classification of locus genotype prediction enables us to compare performances between the two methods, and to check if our criteria introduce any incorrect 'certain' predictions. A method with good performance will return as many full matches for 'certain' predictions as possible, but zero 'partial' or 'no match' results.

The relative amount of DNA between each true contributor is known beforehand[4] (Table 1). However since PCR for small amounts of DNA causes high stochastic variation in the end result (i.e. the peak heights), the corresponding relative peak height levels may not be the same. As a compliment to the relative DNA amounts, estimation of the relative peak height levels between the contributors

[4] this information is used to decide which samples that are suitable for each of the software

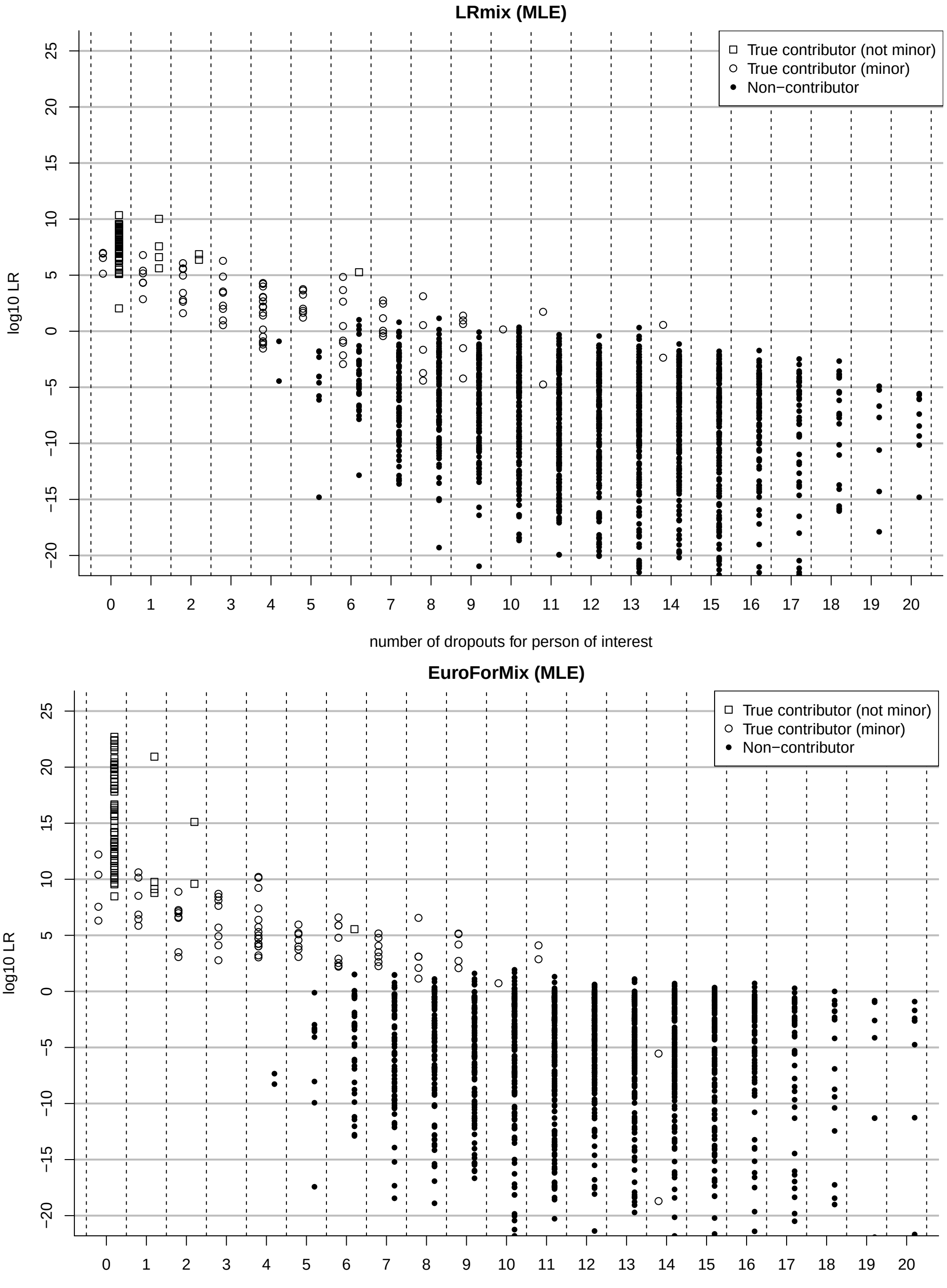

Figure 3: The plots show how the LR values using the MLE method for different POIs are related to the number of dropouts for the corresponding POI, where no conditional reference was assumed and the replicated samples and the sibling comparisons were omitted. “Major” is a contributor having 100 pg, 250 pg or 500 pg amount of DNA. “Minor” is a contributor with only 50 pg DNA.

were carried out using maximum likelihood estimation with the *EuroForMix* model. Importantly, we expect these estimates to influence the performance of the deconvolution.

Comparisons where both *LoCIM-tool* and *EuroForMix* are suitable

Out of the total 59 samples there are 48 which have a major contributor (i.e. the largest component has more DNA than the other components). For these samples we used *LoCIM-tool* to predict the profile genotype of the major components (the degraded sample ’8.7d’ and ’9.6d’ were left out since they had too much dropout for the major contributor). Table 4 shows the overall number of certain predicted locus genotypes for the major component which are are given as 'full match', 'partial match' or 'no match' with the true major profile. For all the 720 comparisons involved (48 × 15 markers), using *EuroForMix*, no instances were observed where a 'certain' predicted locus genotype had 'no match'. For *LoCIM-tool* this occurred for one marker in a profile with 13 'uncertain' markers that had contributors with a ratio of 2:1:1. A partial match means that the predicted locus genotype shares some of its alleles with the true contributor (but not all). This occurred 11 times for *LoCIM-tool* and 12 times for *EuroForMix*. Table 5 gives a summary of the number of samples which returned a given number of markers with 'full match', 'partial match' and 'no match' (for certain predictions). For *EuroForMix* there were five samples with one partial match; two samples with two partial matches; one sample with three partial matches for 'certain' locus genotype predictions. For *LoCIM-tool* three samples had two partial matches and five samples had one partial match for 'certain' predictions. The number of 'full match' is most often different for the two methods. We also found that *EuroForMix* correctly predicted the major profile as certain for 21 samples, whereas *LoCIM-tool* only did so for two.

	EuroForMix	*LoCIM-tool*
Full match	468	369
Partial match	12	11
No match	0	1
Total number of 'certain' markers	480	381
Total predicted markers	720	720

Table 4: The table shows the overall number of certain predicted locus genotypes for the major component which are are given as 'full match', 'partial match' or 'no match' with the true major profile.

Full match	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	0	13	11	7	4	3	2	1	13	11	3	1	0	7	1
Partial match	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	2	2	2	2	2	3	0
No match	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
EuroForMix	21	3	0	1	0	0	1	0	0	1	0	0	1	0	5	7	2	1	1	0	0	0	1	1	1	0	0	0	1	0
LoCIM-tool	2	6	9	0	3	1	2	1	3	0	1	3	3	1	3	1	0	0	0	1	1	2	1	0	0	1	1	1	0	1

Table 5: The table gives a summary of the number of samples (given for each software) where the certain predicted locus genotypes for the major component had a given combination of the number of 'full match', 'partial match' and no match with the true major profile (per sample). For instance, two of the samples for *EuroForMix* had 14 certain predicted locus genotypes which gave 13 "full matches" and one "partial match" with the true major profile.

By investigating the relationship between the number of 'full matches' for 'certain' predictions and the estimated relative peak height levels (see Table S9 in supplementary material section G.3: "Deconvolution results"), we found that *EuroForMix* did well when the estimated level between the non-minor components was at least a two-fold contribution (2:1) in difference. We could also see that the two methods were less certain about a correct prediction when the largest component was close to the two-fold contribution difference. For sample types "x.1" (2:1:1 ratio), only sample '1.1' performed well because of sufficient difference in peak height levels. The samples of type "x.2" have 5:1:1 ratios. Here we found that the samples '6.2' and '11.2' did not perform as well as the other samples because the two largest components in these samples where inferred to have only a 1.4-fold and 2.3-fold contribution difference. As expected, the sample type "x.5" having 10:1:1 ratios performed best, except for sample 11.5 where the major component was inferred to have about a three-fold larger contribution compared to the second largest component. For five of the samples of type "x.6" (10:5:1 ratio), the second largest component was inferred to have the same peak height level as the largest component (i.e. the samples '2.6', '3.6', '6.6', '10.6' and '12.6'). For the replicated samples, *EuroForMix* and *LoCIM-tool* correctly predicted most of the locus genotypes of the major contributor as 'certain'. However, *LoCIM-tool* flagged 30 locus genotype predictions as 'uncertain', while *EuroForMix* flagged only one.

The results presented for *EuroForMix* were based on the most likely locus genotype being at least twice as likely than the second likeliest locus genotype. By increasing the criterion to being at least ten times more likely, ten of the twelve partial matches vanished (not for sample '11.5' and '11.6'). However, 64 'full match' situations for 'certain' locus genotype predictions became 'uncertain'.

Detailed results can be found in Table S9 in the supplementary material section G.3.

Comparisons where only *EuroForMix* is suitable

In this section we considered other cases where *EuroForMix* has a potential to predict the locus genotypes other than for the single major contributors. The prediction results of sample types "x.3" (5:5:1 ratio) were left out in the previous section since the two major contributors had the same DNA amount level. Hence these samples would require that one of the major contributors was conditional (assumed known) in order to achieve any 'certain' predicted locus genotypes. For sample type "x.6" (10:5:1 ratio) only the profile genotypes of the major contributor were compared in the previous section. In this section we additionally predict the profile genotypes of the second largest component. The detailed results are found in Table S10 in the supplementary material section G.3: "Deconvolution results".

For sample type "x.6" (10:5:1 ratio), we found that between 12 and 15 locus genotypes of the second largest component could be correctly predicted (as 'certain') for the following examples: 1) The corresponding peak height level was at least two-fold different from the levels of the two other components (see sample '1.6' and '8.6') 2) A major contributor was conditioned and its corresponding peak height level was at least 1.3-fold different from the other contributors. We found that for sample type "x.3" (5:5:1 ratio), conditioning one of the major contributors was necessary in order to correctly predict at least 12 locus genotypes of the other major component (as 'certain').

5 Discussion

We have investigated the performance of two different open-source models, *LRmix* (a qualitative model) and *EuroForMix* (a quantitative model) to carry out weight-of-evidence calculations and deconvolution (genotype estimation). A dataset of two- to three-person laboratory prepared mock-case mixtures (Benschop and Sijen [18], Haned et al. [17]) was used. The data are made available at `www.euroformix.com/data` so that others may carry out their own comparisons. The mixtures were prepared from known contributors with different quantities of DNA: 50, 100, 250 or 500 pg DNA. Knowing the ground-truth is the only way to carry out performance checks.

The first step was to predict the number of contributors in the sample profiles. For *EuroForMix* and *LRmix* we used the the maximum log-likelihood value together with a penalization term to decide the optimum model. The penalization term for *EuroForMix* was selected as the Akaike information criterion ([9]), while for *LRmix* we used an *ad hoc* value which performed well in a simulation study (section F.1 in the supplementary material). These two methods were compared with a manual interpretation (MI) method where the number of contributors were predicted by a forensic scientist, based on the number of alleles and their relative peak heights using their expert opinion. All three methods were broadly comparable in performance: Out of 59 samples, all three methods underestimated the number of contributors for three of the samples, while *LRmix* and *EuroForMix* overestimated the number of contributors for five and four of the samples, respectively. The fewest errors were with the MI method since it tended to be biased towards underestimating. *LRmix* and *EuroForMix* sometimes overestimated the number of contributors as four instead of three contributors, since it was inefficient to separate minor components that had the same level of peak heights. Interestingly, we did not expect the qualitative model to perform as well in predicting the number of contributors for non-replicates because of the lack of dropout information. However we found that this method was almost as effective as the quantitative model when we followed a maximum likelihood estimation based criterion.

The second step was to identify the likelihood ratio quantity for a broad range of case stain comparisons, and to investigate this as an output for data with different properties. ROC curves were useful to directly compare the efficiency of the different statistical models, where this is defined as simultaneous minimization of false negative and false positive results using $LR = 1$ as the threshold. In this respect, the MLE method was shown to be the most efficient method for both *LRmix* and *EuroForMix* (Figure 2). However, a maximum false positive rate $LR < 200$ was observed (Table 3). This level could be implemented as a reporting limitation. The picture was changed when the conservative method was used, as the maximum false positive rate reduced to $LR < 10$ for both methods, a level that is close to neutral. This was achieved at the expense of an increased false negative rate (Figure 2). In summary, for exploratory purposes it is more efficient to use the MLE approach to compare different models. But for reporting purposes, the conservative approach is preferable, since we consider it prudent to minimise the false positive rate, at the expense of slightly increasing the false negative rate.

When a mixture was evaluated against a non-contributor sibling to the true contributor, the peak height information was useful to reduce the LR to be exclusionary, provided that sufficient DNA was present.

LRmix still gave a high LR for true contributors up to four dropouts for a person of interest (POI) in a three-person mixture. However, the main benefit of *EuroForMix* was with the interpretation of major/minor mixtures where the minor was evidential. Here up to 11 allele dropouts for the POI in a three-person mixture could provide probative evidence, whilst *LRmix* may return a much lower LR or a false negative result. The two models are expected to return similar LR results when contributors have equal mixture proportions or for mixtures of higher order.

A comparative study of the GeneMapper® ID-X Software stutter filter versus the *EuroForMix* stutter model was carried out in section E.1 in the supplementary material. The results showed that in general the two methods compared favorably, the former was useful to remove forward stutter (since *EuroForMix* does not currently accommodate this).

The third step was to carry out deconvolution. *EuroForMix* was compared with *LoCIM-tool* to see how they performed in predicting locus genotypes of unknown components in a sample profile. For *EuroForMix* it was required that the probability of the most likely locus genotype was at least twice as large as the second most likely genotype, otherwise the prediction was flagged as 'uncertain'. This gave 12 situations where a 'certain' predicted locus genotype (480 in total) was only a partial match, but no situations where the predicted locus genotype was completely wrong. There was not much gain in increasing the criterion for flagging a locus genotype prediction as 'uncertain', since the number of partial matches reduced from 12 to two and at the same time the number of full matches reduced with 64. For *LoCIM-tool* there were 11 situations where a 'certain' predicted genotype (381 in total) was only a partial match, whereas there was one situation where a 'certain' predicted genotype had no match. The utility of the two methods was consistent most of the time, however we showed that *EuroForMix* is sometimes preferred in the situations where the components were estimated to have at least a two-fold in peak height level differences. Ideally, using the STR typing kit and settings for PCR and CE as applied to the samples in this study, the major contributor requires at least 250 pg DNA with 50 pg DNA for the minor contributor, to be completely successful otherwise the major contributor may be ranked lower in the list of possibilities.

6 Conflict of interest

None.

7 Acknowledgements

We would like to thank Professor Peter de Knijff from Leiden University Medical Center, The Netherlands, for the trace samples used in this study, and to make these accessible. We would also like to thank Hinda Haned for helpful discussions. ØB, PG and CB received funding support from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement number 285487 (EUROFORGEN-NoE).

8 Supplementary material

Supplementary material can be found in the online version.

References

[1] P. Gill, L. Gusmão, H. Haned, W.R. Mayr, N. Morling, W. Parson, L. Prieto, M. Prinz, H. Schneider, P.M. Schneider, and B.S. Weir. DNA commission of the International Society of Forensic Genetics: Recommendations on the evaluation of STR typing results that may include drop-out and/or drop-in using probabilistic methods. *Forensic Science International: Genetics*, 6(6):679 – 688, 2012. ISSN 1872-4973.

[2] P. Gill, A. Kirkham, and J. Curran. LoComatioN: A software tool for the analysis of low copy number DNA profiles. *Forensic Science International: Genetics*, 166, 2007.

[3] Hinda Haned and Peter Gill. Analysis of complex DNA mixtures using the Forensim package. *Forensic Science International: Genetics Supplement Series*, 3(1):e79 – e80, 2011.

[4] Adele A. Mitchell, Jeannie Tamariz, Kathleen O'Connell, Nubia Ducasse, Zoran Budimlija, Mechthild Prinz, and Theresa Caragine. Validation of a DNA mixture statistics tool incorporating allelic drop-out and drop-in. *Forensic Science International: Genetics*, 6(6):749 – 761, 2012. ISSN 1872-4973.

[5] D. J. Balding. Evaluation of mixed-source, low-template DNA profiles in forensic science. *Proceedings of the National Academy of Sciences of the United States of America*, 110(30):12241–12246, 2013.

[6] D. Balding, A. Timpson, C. D. Steele, M. d'Avezac, and J. Hetherington. likeLTD: Tools to Evaluate DNA Profile Evidence. R-package, 2016. URL `https://cran.r-project.org`.

[7] Keith Inman, Norah Rudin, Ken Cheng, Chris Robinson, Adam Kirschner, Luke Inman-Semerau, and Kirk E. Lohmueller. Lab Retriever: a software tool for calculating likelihood ratios incorporating a probability of drop-out for forensic DNA profiles. *BMC Bioinformatics*, 16(1):1–10, 2015. ISSN 1471-2105.

[8] D. Taylor, J. A. Bright, and J. Buckleton. The interpretation of single source and mixed DNA profiles. *Forensic Science International: Genetics*, 7:516–528, 2013.

[9] O. Bleka, G. Storvik, and 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.*, 21:35–44, 2016.

[10] R. G. Cowell, T. Graversen, S. L. Lauritzen, and J. Mortera. Analysis of forensic DNA mixtures with artefacts. *Appl. Statist.*, 64(1):1–32, 2015.

[11] M. W. Perlin, M. M. Legler, C. E. Spencer, J. L. Smith, W. P. Allan, J. L. Belrose, and B. W. Duceman. Validating TrueAllele DNA Mixture Interpretation. *Journal of Forensic Sciences*, 56: 1430–1447, 2011.

[12] R. Puch-Solis, L. Rodgers, A. Mazumder, S. Pope, I. W. Evett, J. Curran, and D. Balding. Evaluating forensic DNA profiles using peak heights, allowing for multiple donors, allelelic dropout and stutters. *Forensic Science International: Genetics*, 7:555–563, 2013.

[13] Harish Swaminathan, Catherine M. Grgicak, Muriel Medard, and Desmond S. Lun. NOCIt: A computational method to infer the number of contributors to DNA samples analyzed by STR genotyping. *Forensic Science International: Genetics*, 16:172 – 180, 2015. ISSN 1872-4973.

[14] J. A. Bright, D. Taylor, and J. Curran, J.and Buckleton. Searching mixed DNA profiles directly against profile databases. *Forensic Science International: Genetics*, 9:102–110, 2014.

[15] Mark W. Perlin, Kiersten Dormer, Jennifer Hornyak, Lisa Schiermeier-Wood, and Susan Greenspoon. TrueAllele Casework on Virginia DNA Mixture Evidence: Computer and Manual Interpretation in 72 Reported Criminal Cases. *PLoS ONE*, 9(3):1–15, 2014.

[16] Susan A. Greenspoon, Lisa Schiermeier-Wood, and Brad C. Jenkins. Establishing the Limits of TrueAllele® Casework: A Validation Study. *Journal of Forensic Sciences*, 60(5):1263–1276, 2015. ISSN 1556-4029.

[17] Hinda Haned, Corina C.G. Benschop, Peter D. Gill, and Titia Sijen. Complex DNA mixture analysis in a forensic context: Evaluating the probative value using a likelihood ratio model. *Forensic Science International: Genetics*, 16:17 – 25, 2015. ISSN 1872-4973.

[18] Corina C.G. Benschop and Titia Sijen. LoCIM-tool: An expert's assistant for inferring the major contributor's alleles in mixed consensus DNA profiles. *Forensic Science International: Genetics*, 11:154 – 165, 2014. ISSN 1872-4973.

[19] Corina C.G. Benschop, Hinda Haned, Loes Jeurissen, Peter D. Gill, and Titia Sijen. The effect of varying the number of contributors on likelihood ratios for complex {DNA} mixtures. *Forensic Science International: Genetics*, 19:92 – 99, 2015. ISSN 1872-4973.

[20] H. Haned. Forensim: an open-source initiative for the evaluation of statistical methods in forensic genetics. *Forensic Science International: Genetics*, 5:265–268, 2011.

[21] Antoinette A. Westen, Thirsa Kraaijenbrink, Elizaveta A. Robles de Medina, Joyce Harteveld, Patricia Willemse, Sofia B. Zuniga, Kristiaan J. van der Gaag, Natalie E.C. Weiler, Jeroen Warnaar, Manfred Kayser, Titia Sijen, and Peter de Knijff. Comparing six commercial autosomal STR kits in a large Dutch population sample. *Forensic Science International: Genetics*, 10:55 – 63, 2014. ISSN 1872-4973.

[22] Peter Gill, Jonathan Whitaker, Christine Flaxman, Nick Brown, and John Buckleton. An investigation of the rigor of interpretation rules for STRs derived from less than 100 pg of DNA. *Forensic Science International*, 112(1):17 – 40, 2000. ISSN 0379-0738.

[23] P. S. Walsh, N. J. Fildes, and R. Reynolds. Sequence analysis and characterization of stutter products at the tetranucleotide repeat locus vWA. *Nucleic Acids Research*, 24(14):2807–2812, Jul 1996.

[24] Andrew J. Gibb, Andrea-Louise Huell, Mark C. Simmons, and Rosalind M. Brown. Characterisation of forward stutter in the AmpF/STR® SGM Plus® PCR. *Science & Justice*, 49(1):24 – 31, 2009. ISSN 1355-0306.

[25] Antoinette A. Westen, Laurens J.W. Grol, Joyce Harteveld, Anuska S. Matai, Peter de Knijff, and Titia Sijen. Assessment of the stochastic threshold, back- and forward stutter filters and low template techniques for NGM. *Forensic Science International: Genetics*, 6(6):708 – 715, 2012. ISSN 1872-4973. Analysis and biostatistical interpretation of complex and low template {DNA} samples.

[26] Peter Gill, Roberto Puch-Solis, and James Curran. The low-template-DNA (stochastic) threshold - Its determination relative to risk analysis for national DNA databases. *Forensic Science International: Genetics*, 3(2):104 – 111, 2009. ISSN 1872-4973.

[27] D.J. Balding and R.A. Nichols. DNA profile match probability calculation: how to allow for population stratification, relatedness, database selection and single bands. *Forensic Science International*, 64:125–140, 1994.

[28] R. Puch-Solis. A dropin peak height model. *Forensic Sci. Int. Genet.*, (11):80–84, 2014.

[29] P. Gill and H. Haned. A new methodological framework to interpret complex DNA profiles using likelihood ratios. *Forensic Science International: Genetics*, 7:251–263, 2013.

[30] J. Curran, P. Gill, and R. Bill, M. Interpretation of repeat measurement DNA evidence allowing for multiple contributors and population substructure. *Forensic Science International*, 148:47–53, 2005.

[31] H. Haned, K. Slooten, and P. Gill. Exploratory data analysis for the interpretation of low template DNA mixtures. *Forensic Science International: Genetics*, 6:762–774, 2012.

[32] J. Buckleton and J. Curran. Sampling effects. In *Forensic DNA Evidence Interpretation*, CRC Press, chapter 6, pages 197–216. Boca Raton, Florida, 2005.

[33] Therese Graversen. *Statistical and Computational Methodology for the Analysis of Forensic DNA Mixtures with Artefacts*. PhD thesis, University of Oxford, 2014.

[34] H. Haned, L. Péne, J.R. Lobry, A.B. Dufour, and D. Pontier. Estimating the number of contributors to forensic DNA mixtures: does maximum likelihood perform better than maximum allele count? *J Forensic Sci.*, 56(1):3–8, 2011.

[35] T. Egeland, I. Dalen, and P.F. Mostad. Estimating the number of contributors to a DNA profile. *Int J Legal Med*, 117:271–275, 2003.

[36] C. D. Steele and D. J. Balding. Statistical Evaluation of Forensic DNA profile evidence. *Annual Review of Statistics and Its Application*, 1:361–384, 2014.

A Appendix: Model specifications

In this section we describe how $P(E_l|\mathbf{S}_l)$, as part of equations (2) and (3), are defined for *LRmix* and *EuroForMix*. For this, let $\mathbf{S}_l = (S_{l,1}, ..., S_{l,K})$ be the set of locus genotypes for all contributors at marker l, and let $n_{l,a,k}$ denote the number of allele type a in locus genotype $S_{l,k}$ for contributor k, possibly being zero, one or two. For simplification we don't assume any model for drop-in in this section. See Bleka et al. [9] for how this is implemented in *EuroForMix* and Curran et al. [30] for how this is implemented in *LRmix*.

A.1 *LRmix*

LRmix only utilizes the binary information $y_{l,a} \geq T$, where $y_{l,a}$ is the peak height of allele a at marker l. The statistical model behind *LRmix* introduces for each contributor a parameter d_k which is defined to be the probability of drop-out of an allele for a contributor k. In general, every contributor can have different drop-out parameters, $d_1, ..., d_K$, so that

$$P(E_l|\mathbf{S}_l) = \left[\prod_{y_{l,a} \geq T} (1 - \prod_{k=1}^{K} d_k^{n_{l,a,k}})\right] \left[\prod_{y_{l,a} < T} \prod_{k=1}^{K} d_k^{n_{l,a,k}}\right]. \tag{5}$$

Here we have followed the same assumptions as in [31] by assuming that homozygous genotypes drops out with probability d_k^2 and that d_k is constant across all markers.

A.2 *EuroForMix*

EuroForMix is based on the model developed by Cowell et al. [10] and Bleka et al. [9] who specifies a density function directly on the peak height information. Defining $y_{l,a}$ to be the peak height of allele a at marker l, we have

$$P(E_l|\mathbf{S_l}) = \left[\prod_{y_{l,a}\geq T} f_a(y_{l,a}|\mathbf{S}_l,\theta)\right]\left[\prod_{y_{l,a}<T}\int_{x=0}^{T} f_a(x|\mathbf{S}_l,\theta)dx\right] \tag{6}$$

where $f_a(x|\mathbf{S}_l,\theta)$ is the density function of the peak heights, containing model parameters θ, defined as a gamma density function with shape and scale parameter as arguments, parameterized so that μ and σ is the expectation and coefficient of variation of a heterozygote peak height if only one contributor is present (for no degradation):

$$f_a(x|\mathbf{S}_l,\theta) = gamma(x|\beta^{r_a}\sigma^{-2}\sum_{k=1}^{K}\pi_k n_{l,a,k}, \mu\sigma^2). \tag{7}$$

The parameters $\pi_1,..,\pi_K$ are the mixture proportions for each contributor and parameter β models an exponential decaying degradation slope as a function of fragment length r_a. *EuroForMix* also incorporates a model for back-stutters (-4bp) by assuming that some proportion of the original peak height at allele $a+1$, $y_{l,a+1}$, is moved to allele a, also being gamma distributed. The extension of the model becomes

$$f_a(x|\mathbf{S}_l,\theta) = gamma(x|\beta^{r_a}\sigma^{-2}[(1-\xi)\sum_{k=1}^{K}\pi_k n_{l,a,k} + \xi\sum_{k=1}^{K}\pi_k n_{l,a+1,k}], \mu\sigma^2) \tag{8}$$

where parameter ξ is the expected stutter proportion. It is assumed that all parameters are constant across all markers. See Bleka et al. [9] for more details.

Supplementary Materials

This supplementary material includes details of the studies carried out in the paper "A comparative study of qualitative and quantitative models used to interpret complex STR DNA profiles". It contains the following sections:

B Details about data

Table S1 shows a total of 4 two-person mixtures and 55 three-person mixtures which were generated using known reference profiles of 33 individuals (subset described by Benschop and Sijen [18] and Haned et al. [17])[5]. Two siblings were included in the study, references 9A and 10B. Contributors typically consisted of a moderate-template component (i.e. a component with at least 100 picogram (pg) amount of DNA) together with one or more low-template component(s) i.e. components with 30-50 pg of DNA. The column labeled as 'Contributors' denotes the true contributors to each sample(s), with their corresponding amount of DNA given in the column labeled as 'DNA'. The number of allele dropout events is provided in the column labeled as 'Dropout', determined by counting the number of alleles in the reference that has corresponding peak height below 50 RFU (homozygotes were counted twice). For replicates, this number was summed up across all samples. The column labeled as 'Above T_s' denotes which references had most of their peak heights above a stochastic threshold of T_s=175 RFU. The stochastic threshold is an estimated RFU where one of the alleles in a heterozygote pair drops out with a defined probability. With the method tested here the probability of allele dropout is less than 0.01 when the remaining allele has peak height equal or greater than 175 RFU. See Gill et al. [26] for a method of determination. Some samples were replicated, originating from separate amplifications of the same DNA extract. Samples '0.5', '0.9', '0.24', '0.28', '0.6', '0.7', '0.10' and '0.11' had four non-degraded replicates, while samples '8.7d' and '9.6d' contained three very degraded replicates (i.e. the peak heights decreased as the fragment lengths increased). All the other samples were non-replicates and were degraded by varying degrees. The maximum allele count (MAC) is the maximum number of alleles in any of the loci and total allele count (TAC) is the total number of alleles across all markers. These two quantities are simple and useful for deciding the number of contributors (see section F.1: "Estimating number of contributors and drop-out parameter in *LRmix*" in supplementary material for an illustration through a simulation study).

[5] All data can be found in the zip-file "NFIdata" at `www.euroformix.com/data`

Sample(s)	Contributors	DNA (pg)	Dropout	Above T_s	MAC	TAC
0.5.(1-4)	(0A,0C)	(150,30)	(4,43)	none	3/3/3/4	35/35/34/40
0.9.(1-4)	(0A,0C)	(300,30)	(0,37[34])	0A	4/4/3/4	40/41/37/36
0.24.(1-4)	(0A,0C)	(30,150)	(62,0)	0C	3/3/3/3	32/29/32/33
0.28.(1-4)	(0A,0C)	(30,300)	(49[44],0)	0C	3/4/4/3	30/38/36/35
0.6.(1-4)	(0A,0B,0C)	(150,6,30)	(6,59,48)	none	4/3/3/4	34/31/37/39
0.7.(1-4)	(0A,0B,0C)	(150,30,30)	(5,42,48)	none	4/3/4/4	37/37/38/43
0.10.(1-4)	(0A,0B,0C)	(300,6,30)	(0,45[42],39[37])	0A	5/4/4/4	44/40/39/42
0.11.(1-4)	(0A,0B,0C)	(300,30,30)	(0,30,33[32])	0A	5/5/5/4	48/42/48/43
8.7d.(2-4)	(8A,8B,8C)	(500,250,250)	(42,52,54)	none	5/5/3	32/27/16
9.6d.(2-4)	(9A,9B,9C)	(500,250,50)	(38,50,51)	none	4/6/5	32/27/26
1.1	(1A,1B,1C)	(100,50,50)	(0,4,1)	none	5	63
2.1	(2A,2B,2C)	(100,50,50)	(1,7,2)	none	5	55
3.1	(3A,3B,3C)	(100,50,50)	(0,0,3)	none	5	46
6.1	(6A,6B,6C)	(100,50,50)	(6,10,6)	none	4	41
8.1	(8A,8B,8C)	(100,50,50)	(0,6,3)	none	5	59
9.1	(9A,9B,9C)	(100,50,50)	(0,6,1)	none	6	59
10.1	(10A,10B,10C)	(100,50,50)	(1,7,4)	none	5	52
11.1	(11A,11B,11C)	(100,50,50)	(0,4,2)	none	6	58
12.1	(12A,12B,12C)	(100,50,50)	(1,2,2)	none	6	56
14.1	(14A,14B,14C)	(100,50,50)	(0,6,0)	none	6	55
1.2	(1A,1B,1C)	(250,50,50)	(0,11,1)	1A	5	55
2.2	(2A,2B,2C)	(250,50,50)	(0,9,4)	2A	5	52
3.2	(3A,3B,3C)	(250,50,50)	(0,4[3],5)	none	5	40
6.2	(6A,6B,6C)	(250,50,50)	(0,5,9)	none	4	46
8.2	(8A,8B,8C)	(250,50,50)	(0,6[5],3)	8A	6	59
9.2	(9A,9B,9C)	(250,50,50)	(1,14,14)	none	4	40
10.2	(10A,10B,10C)	(250,50,50)	(0,9,2)	none	5	53
11.2	(11A,11B,11C)	(250,50,50)	(2,8,7)	none	5	48
12.2	(12A,12B,12C)	(250,50,50)	(0,6,4)	none	5	51
14.2	(14A,14B,14C)	(250,50,50)	(0,7,0)	none	5	54
2.3	(2A,2B,2C)	(250,250,50)	(0,0,3[2])	2B	6	62
3.3	(3A,3B,3C)	(250,250,50)	(0,0,4)	3B	5	45
6.3	(6A,6B,6C)	(250,250,50)	(0,0,3)	none	5	58
8.3	(8A,8B,8C)	(250,250,50)	(0,0,4)	8A	6	64
9.3	(9A,9B,9C)	(250,250,50)	(0,0,6)	none	6	60
10.3	(10A,10B,10C)	(250,250,50)	(0,2,5)	none	5	57
11.3	(11A,11B,11C)	(250,250,50)	(0,0,9)	none	5	55
12.3	(12A,12B,12C)	(250,250,50)	(0,0,1)	12B	6	60
14.3	(14A,14B,14C)	(250,250,50)	(0,0,2)	14A	6	59
1.5	(1A,1B,1C)	(500,50,50)	(0,4,4[3])	1A	5	59
2.5	(2A,2B,2C)	(500,50,50)	(0,3,0)	2A	6	62
3.5	(3A,3B,3C)	(500,50,50)	(0,1[0],4)	3A	4	44
6.5	(6A,6B,6C)	(500,50,50)	(0,5,9)	none	5	46
8.5	(8A,8B,8C)	(500,50,50)	(0,5[3],5[4])	8A	6	58
9.5	(9A,9B,9C)	(500,50,50)	(0,8,7)	none	6	51
10.5	(10A,10B,10C)	(500,50,50)	(0,11,7)	none	5	46
11.5	(11A,11B,11C)	(500,50,50)	(0,6,4)	none	5	54
12.5	(12A,12B,12C)	(500,50,50)	(0,3[2],5)	12A	6	53
14.5	(14A,14B,14C)	(500,50,50)	(0,8,3)	14A	5	50
1.6	(1A,1B,1C)	(500,250,50)	(0,0,8[7])	1A	5	60
2.6	(2A,2B,2C)	(500,250,50)	(0,0,4[2])	2A	6	62
3.6	(3A,3B,3C)	(500,250,50)	(0,0,2)	3A	5	47
6.6	(6A,6B,6C)	(500,250,50)	(0,0,7)	6A	4	53
8.6	(8A,8B,8C)	(500,250,50)	(0,0,4[2])	8A	6	64
9.6	(9A,9B,9C)	(500,250,50)	(0,0,1)	9A	6	65
10.6	(10A,10B,10C)	(500,250,50)	(0,0,4)	10A	6	61
11.6	(11A,11B,11C)	(500,250,50)	(0,0,8)	11A	5	56
12.6	(12A,12B,12C)	(500,250,50)	(0,0,2)	12A	6	59
14.6	(14A,14B,14C)	(500,250,50)	(0,0,4)	14A	5	57

Table S1: The table gives an overview of contributing individuals for all samples considered, with corresponding amounts of DNA (quantified in picograms). The bracketed information in the 'Sample(s)' column denotes the replicate number, e.g. (2-4) means the replicates '2', '3' and '4'. The first eight samples include components that are low-template (i.e. less than 50 pg) where combinations of the references 0A, 0B and 0C are the true contributors. The next two samples, '8.7d', and '9.6d' have components with more than 50 pg but are greatly degraded and have several drop-outs. The rest of the samples consist of one replicate, but with different amounts of DNA. Column 'Dropout' shows the number of alleles from the corresponding contributors which are not present in the sample (i.e. having peak height below the detection threshold of 50 RFU). The number of drop-outs were counted across all replicates. The dropout number within square brackets is the number of dropouts observed when a stutter-filter was *not* applied (given in bracket only if this value deviated from that obtained when the stutter filter was applied). 'Above T_s' denotes references with their peak heights above the stochastic threshold 175 RFU. 'MAC' is maximum allele count and 'TAC' is total allele counts (excluding amelogenin).

C The sampling effect of allele frequencies

The frequency database is based on the allele set of 2085 Dutch individuals. When the relative proportions of the alleles are used as probabilities for the rarity of the genotypes in the population, this may introduce uncertainty in the likelihood ratio (LR) quantity as a function of the samples (because of the limited number of sampled individuals). Here we will investigate the effect of this uncertainty by assuming that the number of observed alleles of different types follows a multinomial distribution. For this we used 10 of the samples from the data set samples and compared the LR for six references for each of the samples (three true contributors and three non-contributors), hence 60 comparisons in total. The samples were chosen such that the true contributors were different for all samples (we used the samples '1.1', '2.3', '3.5', '6.5', '8.6', '9.1', '10.3', '11.2', '12.6', '14.2'). We used the "Highest posterior density interval" method where we simulated 100 allele frequency sets from the Dirichlet-multinomial posterior model with a flat prior (see [32]). For each of the 100 allele frequency sets we calculated a LR for a given reference compared to a given sample. This was done both for *EuroForMix* and *LRmix* using the maximum likelihood estimation (MLE) method where we assumed three contributors, sub-population structure with $F_{st} = 0.01$, degradation and the drop-in model (with drop-in probability $C = 0.00036$ and $\lambda = 0.02$). For a given reference compared to a given sample we estimated the 0.05 and 0.95 quantiles of the LRs (given as $\widehat{LR}_{0.05}$ and $\widehat{LR}_{0.95}$) across the 100 allele frequency sets.

For each of the 60 comparisons we recorded the estimated distance between the 0.05 and 0.95 quantile of the LRs (on logarithmic scale), given by $b = x_{95} - x_{05}$ where $x_{05} = \log_{10} \widehat{LR}_{0.05}$ and $x_{95} = \log_{10} \widehat{LR}_{0.95}$. Figure S1 shows a histogram of b for the 60 comparisons for both *EuroForMix* and *LRmix*. From the figure we can see that most of the values of b are between 0.1 and 0.25 for *LRmix*, and between 0.05 and 0.55 for *EuroForMix*. This indicates that it is not large effect in the allele frequency sampling when drawing 2085 individuals.

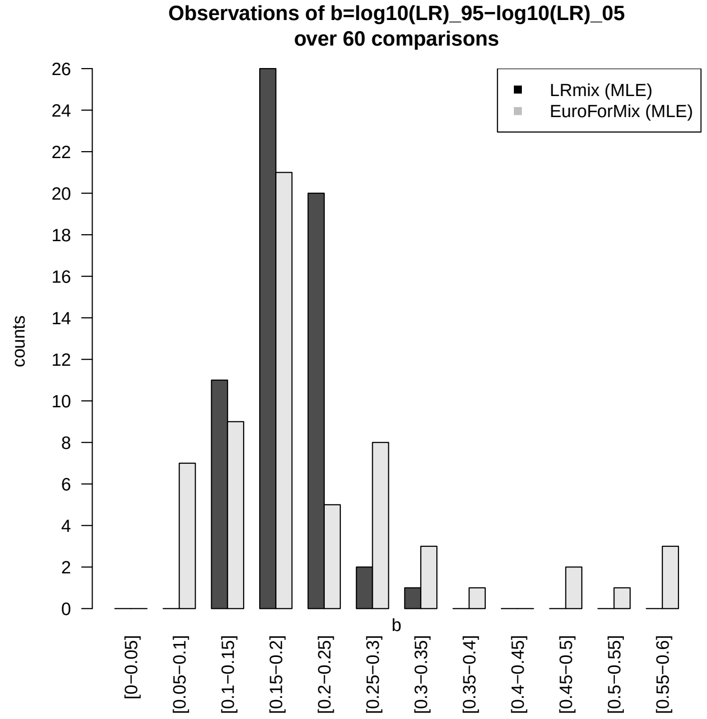

Figure S1: The figure shows the number of comparisons where the estimated width of a 90% interval of the LR values (on logarithmic scale) falls within a certain interval (given on x-axis). The width is estimated as $b = \log_{10} \widehat{LR}_{0.95} - \log_{10} \widehat{LR}_{0.05}$ by generating 100 allele frequency samples from a multinomial-dirichlet posterior model with a flat prior. All the 60 comparisons involved comparing three true contributors and three non-contributors for ten samples from the data set.

D Validation data

In the two following subsections the aim is to check the goodness of fit of the underlying peak height model in *EuroForMix* by comparing it with the peak height variation and drop-outs of single source profiles for different small amounts of DNA.

D.1 Peak height and drop-out data

30 replicated samples of reference DNA007 (is part of an internal validation data set) were obtained for 30, 25 and 20 pg. These samples were not degraded and without stutters. The peak height output above 50 RFU from reference 007 is assumed to be distributed as

$$y_a|\theta \sim gamma(\sigma^{-2}, \mu\sigma^2) \tag{9}$$

and that the probability for a peak height falling below 50 RFU (i.e. allele dropout) is

$$P(dropout|\theta) = \int_0^{50} gamma(x|\sigma^{-2}, \mu\sigma^2)dx. \tag{10}$$

With 15 markers for each sample, having 2 alleles each (only heterozygote markers), this gives 30*15*2=900 number of data points. Let $Y = (y_1, ..., y_m, y_{m+1}, ..., y_n)$ be the set consisting of the m drop-out data $y_1, ..., y_m$ and where the n-m remaining data is the non-drop-out data. With the specified model assumptions we have that the likelihood is given as

$$l(\theta|Y) = m \log P(dropout|\theta) + \sum_{j=m+1}^{n} \log gamma(y_j|\theta). \tag{11}$$

Maximum likelihood estimates of θ are obtained by optimizing the likelihood function in equation (11).

From Figure S2 we see that the distribution of the peak heights for the three dilutions are well explained with the theoretical model. From the 1000 reference samples we also see that the observed cumulative drop-out lies within the simulated ones for all the dilutions.

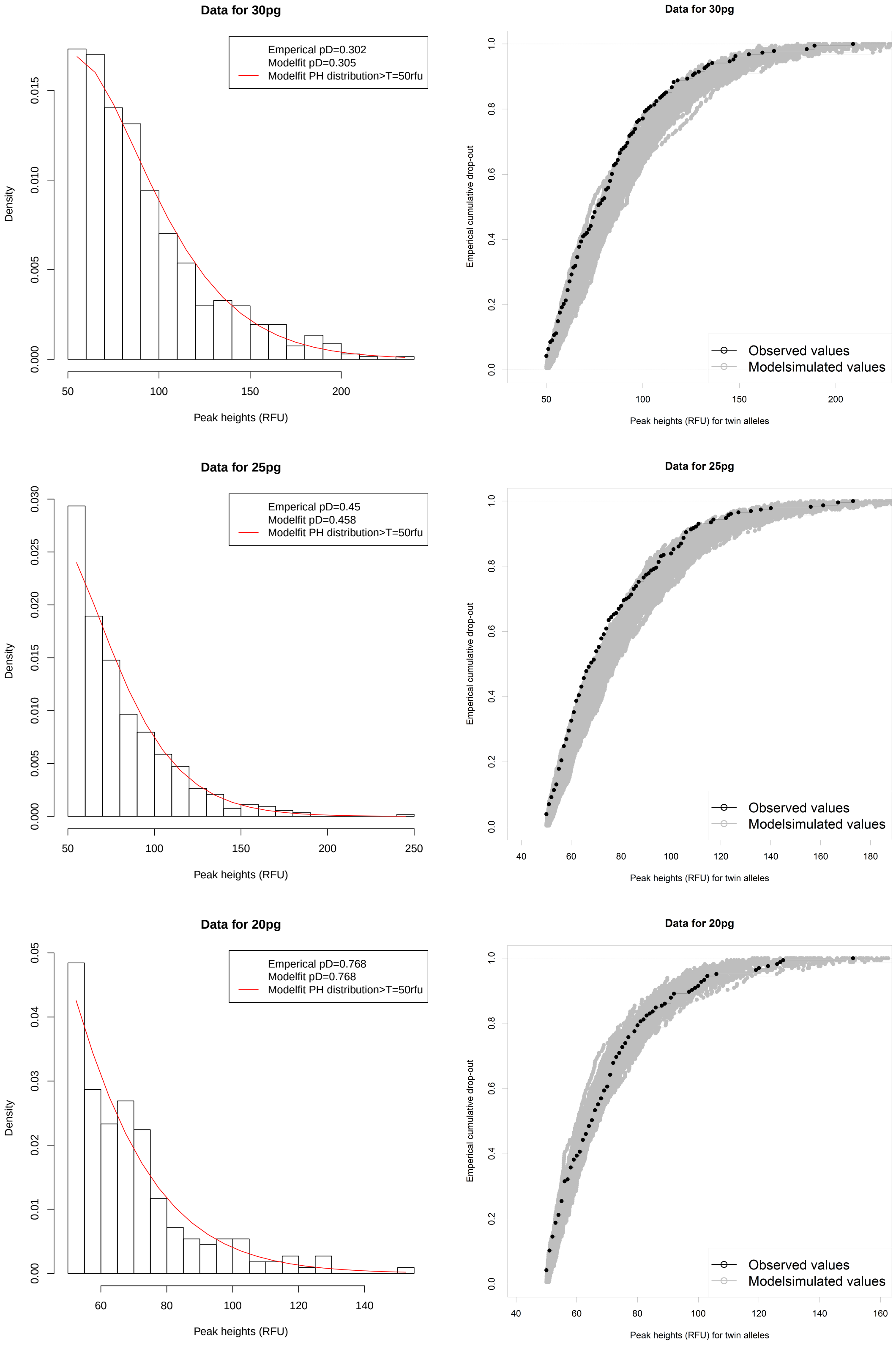

Figure S2: The plots show the observed drop-out data against the model fitted drop-out properties. The left panels show the fitted gamma model (conditional on peak heights above 50 RFU) after optimizing the likelihood function in equation (11). The right panels show the cumulative peak heights for all remaining alleles where the other one has dropped out. Based on the fitted gamma model and the same number of dropouts as in the observations, 1000 reference samples were simulated, each giving a cumulative distribution in gray.

D.2 Heterozygote data

An observation of 173 heterozygote balance data $H_b = \frac{y_a}{y_b}$ for different amount of DNA amount (from reference DNA007) were considered. With the gamma model given in equation (9) in the article, we have that (also pointed out in Graversen [33]), that the heterozygote balance $H_b = \frac{y_a}{y_b}$ for $a \neq b$ is distributed as a F-distribution

$$H_b|\sigma \sim F(2\sigma^{-2}, 2\sigma^{-2}) \tag{12}$$

With this assumption it is easy to obtain the maximum likelihood estimate of σ by assuming the H_b observations as independent (for each amount of DNA). Figure S3 shows a histogram of the observed heterozygote balance data H_b categorized for different amount DNA (31 pg, 63 pg, 125 pg, 250 pg, 500 pg and 750 pg). For each category we estimated σ and used it as a plug-in value to equation (12) to create a superimposed density curve in the figure. Even though there are not so many data points (up to 32 observations in each category), we see in Figure S3 that the underlying gamma model seems to follow the empirical observations quite well, except from the outlier at 250 pg.

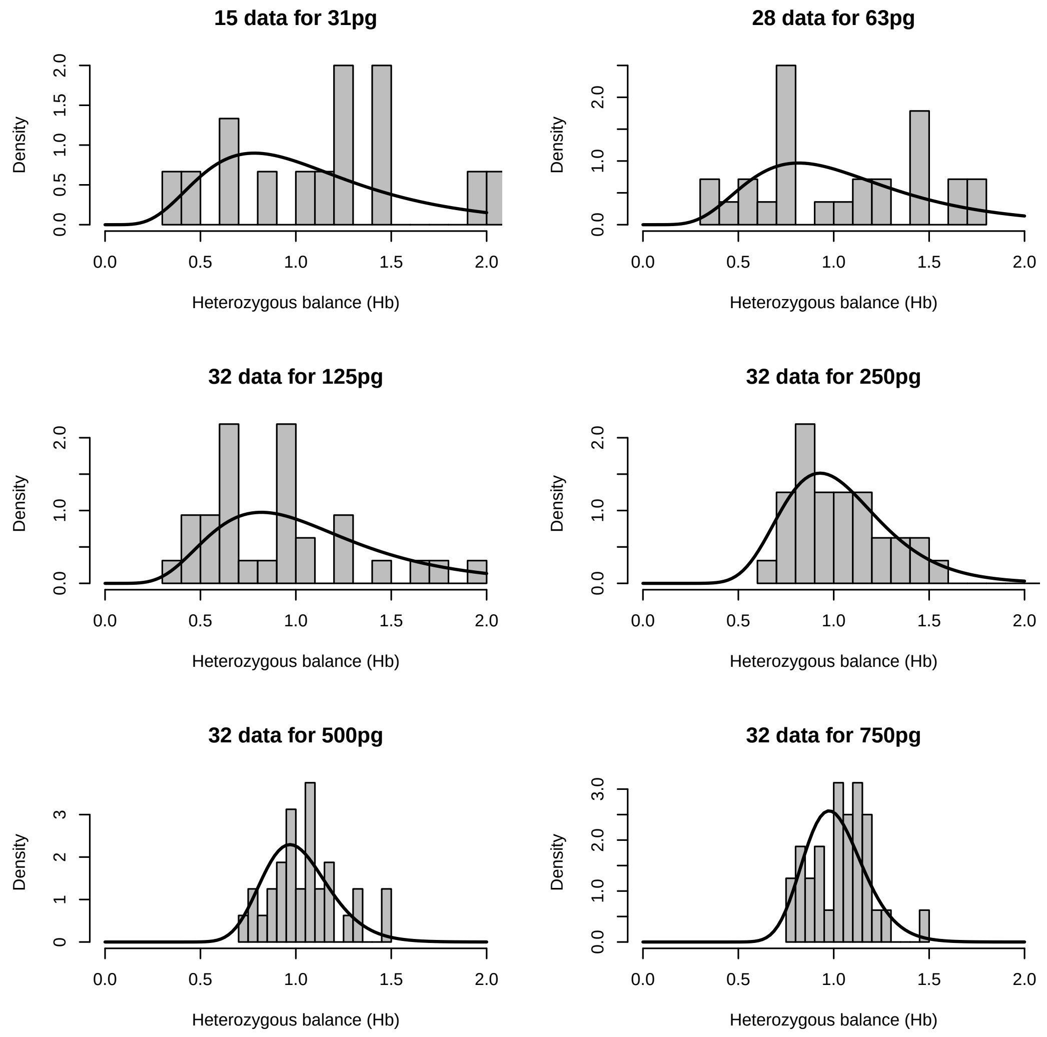

Figure S3: The figure shows several plots with different amount of DNA where emperical data of heterozygote balance is compared (given as a black curve) with $F(2\sigma^{-2}, 2\sigma^{-2})$. The value of σ was inserted as the maximum likelihood estimate based on the observations for the given DNA amount.

D.3 Drop-in data

In this subsection the aim is to check the goodness of fit of the underlying drop-in peak height model in *EuroForMix*, and other alternatives, by comparing them with the peak heights from negative controls.

A total of 14757 negative controls generated with injection time 3k 5s, 29 PCR cycles and detection threshold 50 RFU were gathered. There were 80 false positives alleles found. From the data it was observed that the relative frequency of drop-in per marker were 0.00036. Figure S4 shows the distribution of allele peak heights (represented in left figure as a histogram; and right figure as an empirical cumulative function). We investigated how well three different models fitted these peak heights: the exponential distribution, the Pareto distribution and a variant of the Matern function was used. From Table S2 we found that the Pareto distribution was the best fitting model (as the best AIC score was obtained). However, we see from the curves in the right plot in Figure S4 that the difference between the exponential and Pareto distribution is small. Hence using the exponential distribution should be satisfying.

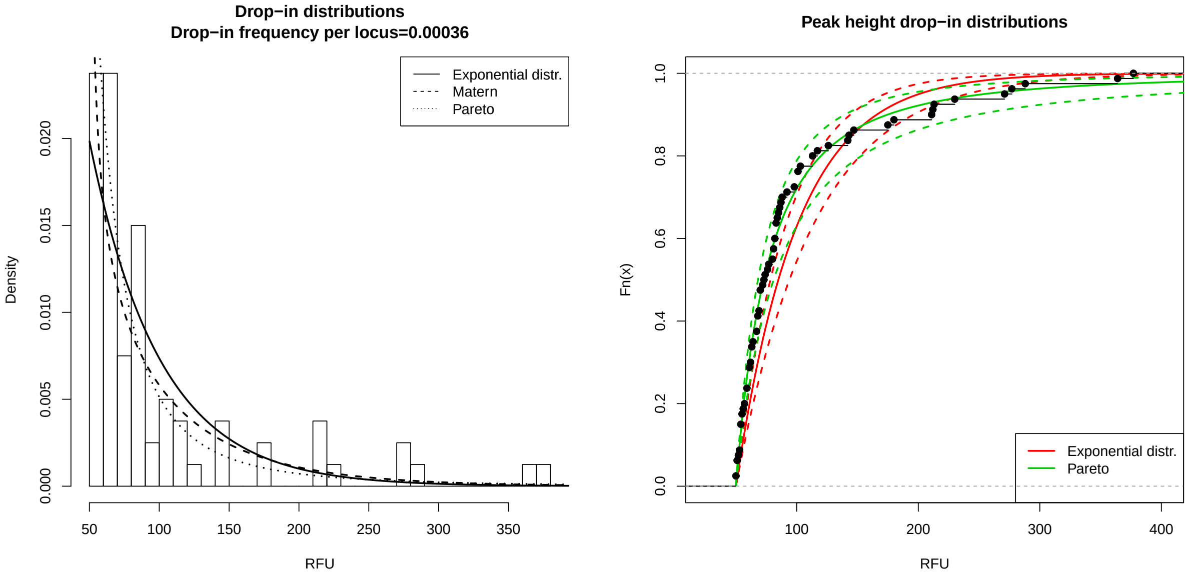

Figure S4: The figures show the distribution of the allele peak heights of the 80 drop-in data points. Left figure shows the peak heights in a histogram, superimposed with MLE fitted models. The right figure shows the peak heights as an empirical cumulative function, superimposed with the MLE fitted models with whole lines supplied with the 95% confidence interval lines as dashed lines.

Model	CDF $F(x\|\theta)$	#param	$\hat{\theta}$	l_{max}	AIC
Pareto	$1-(\frac{x}{50})^{\alpha}$	1	1.84	-387.7	-777.4
Matern	$1-M(x-50\|\theta_1,\theta_2)$	2	(79,0.31)	-387.9	-779.8
Exponential	$1-e^{-\theta(x-50)}$	1	0.02	-393.5	-789.0

Table S2: The table shows the model candidates for modeling the drop-in peak heights. “CDF” mean the cumulative density function taking value x given model parameter θ. “#param” is number of parameters for the model, l_{max} is the maximum likelihood value for the observed data, while AIC$= 2l_{max} - 2$#param, the Akaike information criterion. The model Matern consists of the Matérn covariance function M.

E The effect of including a stutter or drop-in model

E.1 Comparison of the stutter model in *EuroForMix* versus GeneMapper stutter filter

Maximum likelihood estimation (MLE) is used to determine the distribution of the back-stutter proportion in *EuroForMix* (assumed to be the same for all markers), and this distribution is used in the LR calculation (as described in Appendix section A.2). Currently we do not evaluate forward stutter. A consideration of stutter is especially important when major/minor profiles are considered with the POI corresponding to the minor with allele peak heights with approximately the same size as stutter peaks of the major contributor. The stutter model in *EuroForMix* was compared with application of the stutter filter in GeneMapper described in section 2.2 (different per marker). We used the model selection framework from section 3.2 to predict whether a stutter-model should be utilized after applying the stutter filter. The correct number of contributors were applied for all comparisons.

Figure S5 shows that the method using the GeneMapper stutter-filter can improve the performance. It was observed that for the samples '1.6', '3.6', '8.6', '8.3', '3.3' and '3.5', the LRs of the minor contributors increased substantially when the GeneMapper filter was used, as a forward-stutter at allele 17 in D22S1045 was removed in each case. This improved the fit of the model to the data (forward-stutters are not currently modeled in *EuroForMix*). The LR was increased using the GeneMapper stutter filter (compared to the stutter model in *EuroForMix*) even if some low-level minor alleles of the minor POI contributor were removed. Conversely, for samples '3.2', '2.3', '2.6', '10.6' and '3.5' there were no forward-stutters in the original data. Therefore, for these samples the LRs decreased significantly since one or two alleles were removed for each. Details about how the LR changed for each sample when the stutter filter was applied is shown in Figure S6.

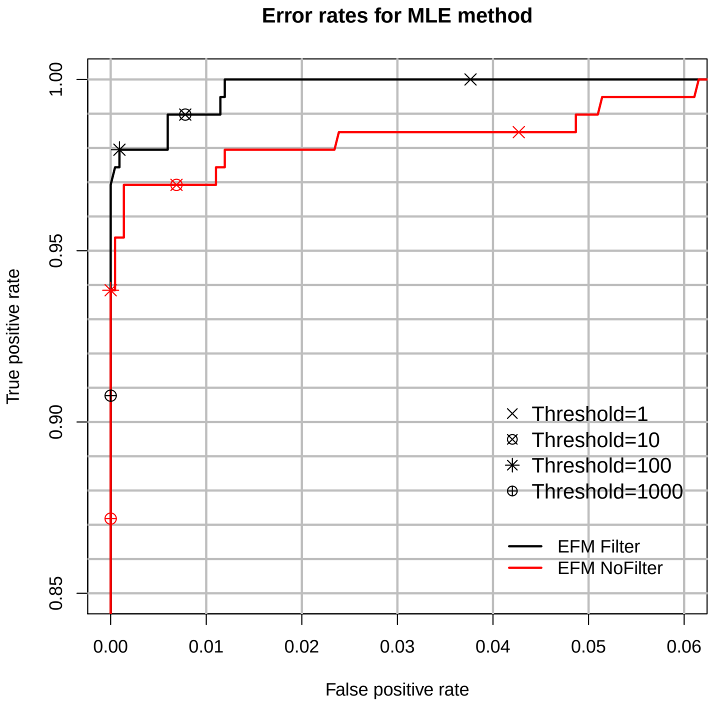

Figure S5: Receiver operating characteristic (ROC) plot where the number of false positives (along horizontal axis) and true positives (along vertical axis) are plotted as a function of LR thresholds. Results shown for *EuroForMix* using the maximum likelihood estimation method based on either stutter-filtered data (EFM Filter) or non-filtered data (EFM NoFilter), where the samples with replicates are omitted.

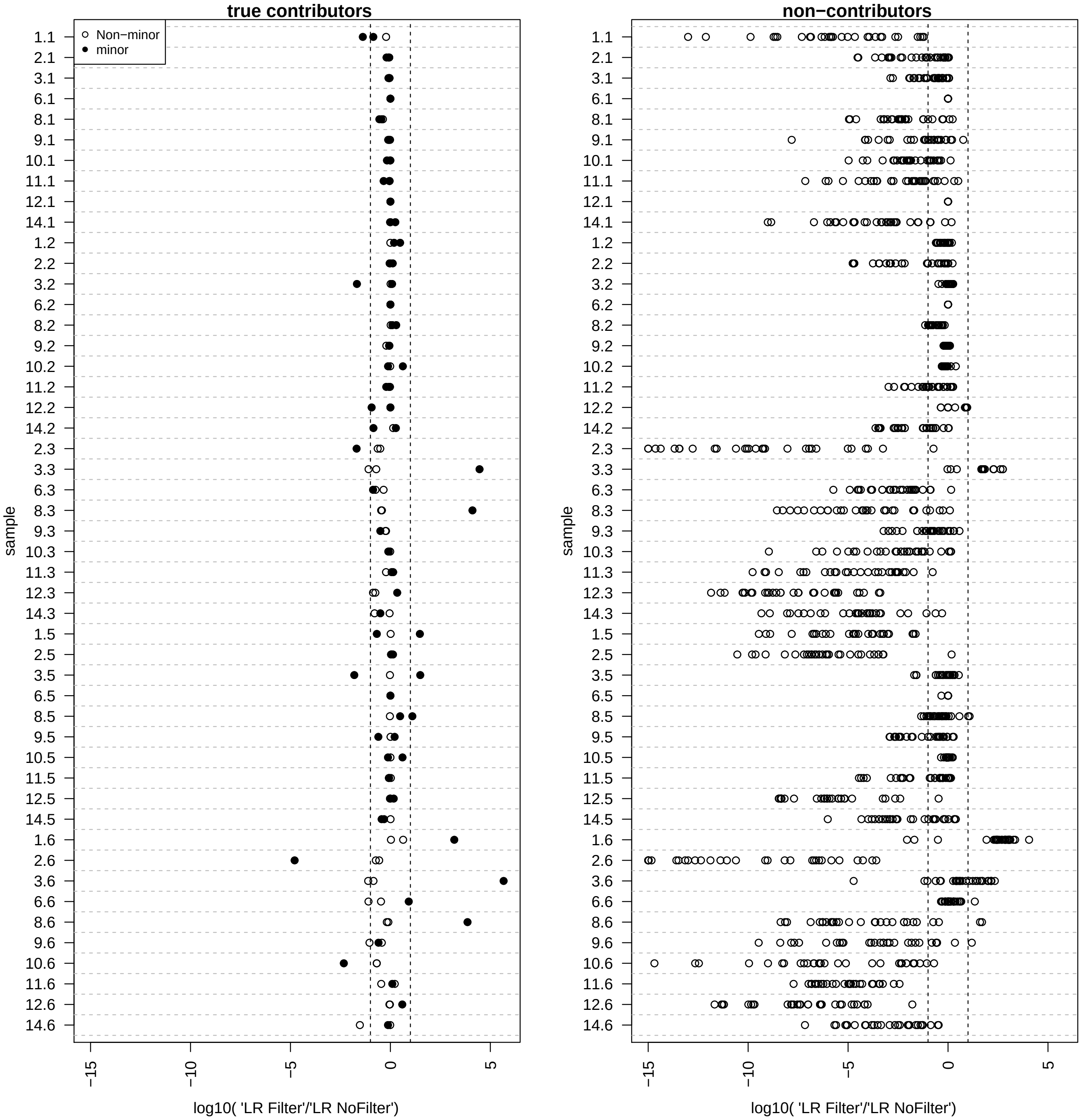

Figure S6: The plots show the difference of $\log_{10} LR$ values between applying the Genemapper stutter-filter and the stutter model within *EuroForMix* (left plot shows true contributors, right plot shows non-contributors). Replicates-samples were not considered. The dashed vertical line indicates $LR = 10^{\pm 1}$.

E.2 Application of the drop-in model

The drop-in models in *EuroForMix* and *LRmix* require a model parameter C, and are implemented as following: When an observed allele a can only be explained as a drop-in (i.e. when there are no contributors to explain the allele), the likelihood term is multiplied with Cp_a, where p_a is the allele proportion of allele a in the population. Hence it denotes the probability for that a specific allele drops in. If no drop-ins are considered, then the likelihood is multiplied by $1 - C$. Importantly, *EuroForMix* and *LRmix* assumes that no more than one allele has dropped in per marker, implying that C is equal the relative frequency of allele drop-ins (per marker). See section 3.1 in the article for how this is estimated.

As presented in Bleka et al. [9], a model for the non-explained peak heights y are distributed as $exp(y - T|\lambda)$, where parameter λ denotes the steepness of the exponential curve and is estimated by $\hat{\lambda} = n/\sum_{i=1}(y_i - T)$. In the drop-in validation we compared the exponential decay curve for modeling the drop-in peak heights with the Pareto and Matérn function model. Here we found that the Pareto model fitted the drop-in peak height best, but did not differ very much from the exponential curve. The effect on the LR when the exponential and Pareto models are compared is investigated in the next section.

E.3 The effect of applying the drop-in model to accommodate an extra allele

If a sample can be explained under H_p without having to consider the possibility of drop-in, the effect of using a drop-in model in the calculation is negligible. The drop-in model is important when spurious (extra) alleles are present that cannot be explained by the conditional contributors in the model.

In this subsection we investigated the effect of applying the fitted drop-in model from the validation data (i.e. $\lambda = 0.02$, $C = 0.00036$) for situations where a spurious allele is included in the data set. We focus on the two samples '12.3' and '12.5' that both have 6 alleles in marker D1S1656, coming from the true contributors, but differ in mixture proportions (see Figure S7). To determine how the assumed drop-in model in *EuroForMix* handles spurious alleles, we inserted an extra allele a with peak height y_a in marker D1S1656, providing a total of 7 alleles. This was carried out with the allele $a = 14$, with frequency 0.09, and the rarer allele 20.3, with frequency 0.002. Based on each of the two samples, we calculated the MLE based likelihood ratio (LR) using *EuroForMix* for each true contributor (reference 12A, 12B and 12C) and the non-contributor 8C (having locus genotype '14/17.3') from the data set used in the paper. The final model was used from the model selection assuming three contributors and no stutter model.

Figure S8 shows the effect on the LR when inserting the spurious allele for each of the samples with different levels for the corresponding peak height y_a. We observed that the drop-in model handles the spurious allele well since the LR remains almost constant (however slightly decreasing) between 50 RFU and 200 RFU for all true contributors, even for minor contributors. From 200 RFU and greater, the high peak height level of drop-in greatly reduced the LR for reference 12A in sample '12.3' which has peak heights around 250 RFU. This effect was also observed when the drop-in level was increased to 350 RFU for sample '12.5' where reference 12A had this level of peak heights. For the non-contributor 8C (having locus genotype '14/17.3') we observed that the LR decreased slightly when a drop-in at allele 20.3 with peak height less than 100 RFU was introduced (but increased when the drop-in peak height was increased further). When a drop-in at allele 14 was introduced (including all alleles of 8C) the LR increased by $10,000$ for drop-in peak height up to 300 RFU (but decreased when the drop-in peak height was increased further).

Hence we observe that the implemented drop-in model in *EuroForMix* accommodates spurious alleles very efficiently - there is a small decrease in the LR. As expected, the larger the peak height, the greater the reduction in LR, because it impacts on heterozygote balance with other alleles.

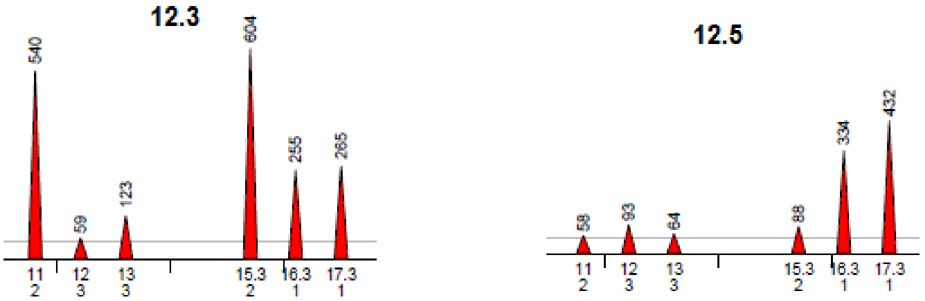

Figure S7: The plot shows the data at marker D1S1656 in the samples '12.3' (left) and '12.5' (right) which both have six alleles included. The alleles labeled as "1" belong to reference 12A, as "2" to reference 12B and "3" to reference 12C.

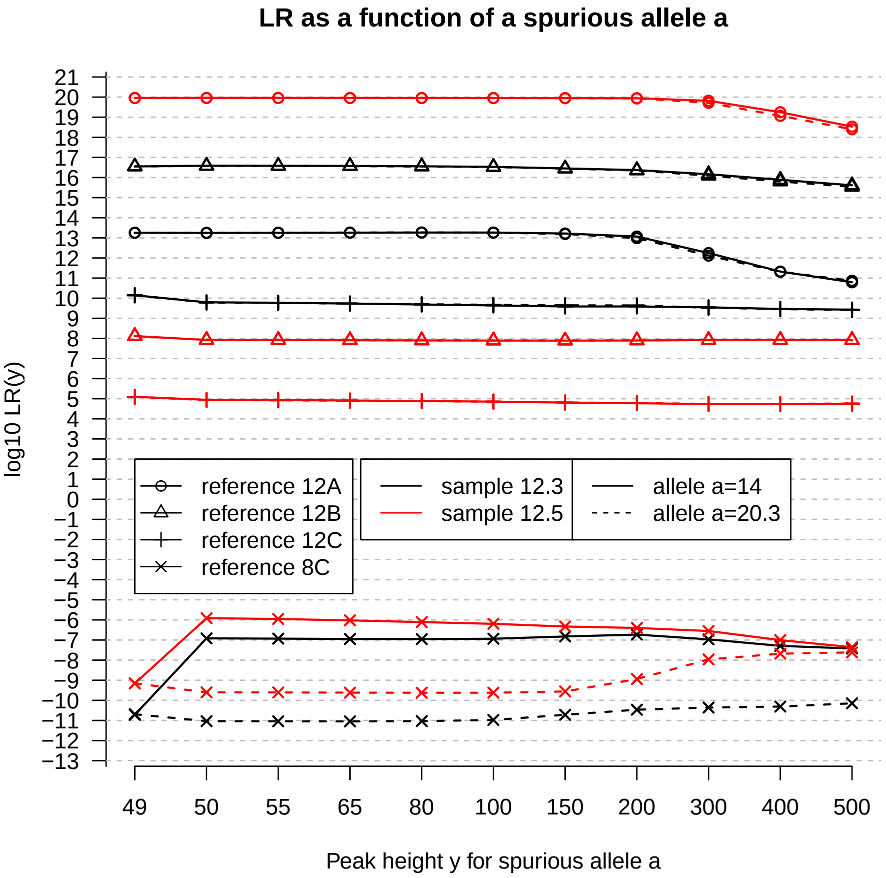

Figure S8: The plot shows the MLE based likelihood ratio (LR) (y-axis) for *EuroForMix* where a spurious allele a was included with peak height y_a (x-axis) into marker D1S1656 for the samples '12.3' (black) or '12.5' (red). The inserted alleles were 14 (whole line) or 20.3 (dashed line). The persons of interest where true contributors were 12A, 12B, 12C and the non-contributor 8C. The reference 8C has the locus genotype '14/17.3'. The model assumed three contributors with no stutters, but assumed the fitted drop-in model from the validation data.

F Estimation of number of contributors

F.1 Estimating number of contributors and drop-out parameter in *LRmix*

The number of unique alleles in a sample depends on the number of contributors, the amount of drop-out and the level of allele sharing. Here we evaluate different methods to predict the number of contributors and amount of drop-out in a sample. In order to do this a simulation study was carried out where we present different quantities for different number of contributors and amount of dropout.

We let K_0 be the true number of contributors in a sample. Each profile for a contributor is drawn from the Dutch population, with alleles likely to dropout with probability d_0 (d_0^2 for homozygotes). Based on this sample we assumed a qualitative model with the true K_0 contributors (see the *BasicDrop* model specification in Appendix A.1). The inference on d_0 is accommodated in two ways:

1 The maximum likelihood estimate of d_0 is given by $\hat{d}_{mle}$, taken as the maximum of the likelihood function of the qualitative model (i.e. $\arg\max_d L(d)$ where L is the likelihood function as a function of the dropout parameter d).

2 Using the median quantile, the average or the mode from the drop-out distribution (requiring at least 1000 samples) conditioning on the total number of alleles in the sample giving either $\hat{d}_{med}$, $\hat{d}_{avg}$ or $\hat{d}_{mod}$ as estimates of d_0 (the mode was estimated by partitioning the drop-out probability into {0,0.01,0.02,...,0.99,1} and returning the cell with the most values).

This is repeated $M = 1000$ times, providing 1000 evidence samples for the different values of $K_0 \in \{1, 2, 3\}$ and $d_0 \in \{0, 0.05, 0.1, ..., 0.45, 0.5\}$.

In addition we stored the following values for each generated sample based on different values of K_0 and d_0:

MAC : the maximum allele counted for any marker

TAC : the total number of alleles across all markers

$l_{max}(K)$: the maximum log-likelihood value assuming $K = 1, 2, 3, 4$ number of contributors.

Estimating number of contributors

Estimating the number of contributors in an evidence sample is very challenging when the amount of allele-dropout is unknown. There are several methods proposed for the qualitative model where the likelihood function is used[34, 35]. However, these does not consider partial profiles where some of the components are dropping out. The conventional approach is to count the maximum number of alleles observed at any one marker, divide by two and then round up to closest integer. Hence observing MAC=3,4 means that the number of contributors is estimated as two, for MAC=5,6 the estimate is three etc. From the upper panel in Figure S9 this simple procedure works well for two-person samples, and for three-person samples without much dropout. However for four-person samples and three-person samples with much dropout, this method will underestimate the number of contributors with high probability. An alternative is to use the total number of alleles across all markers to predict the number of contributors, putting a calibrated threshold on what values of TAC estimates the number of contributors as one, two, three etc. However, from the lower panel in Figure S9 it was observed that there was a lot of overlap across the true number of contributors such that classification becomes very difficult.

Another method uses the maximum likelihood value across the proposed number of contributors (i.e. $\arg\max_K L_{max}(K)$). From the upper panel in Figure S10 it is observed this would work well for two-person mixtures, but the method sometimes overestimates the number of contributors. An *ad hoc* way to correct for this is to penalize the maximum log-likelihood value with the number of assumed contributors (we used $\arg\max_K\{l_{max}(K) - K\}$). From the lower panel in Figure S10, this repairs the overestimation, but conversely the method is more likely to underestimate three and four contributor samples.

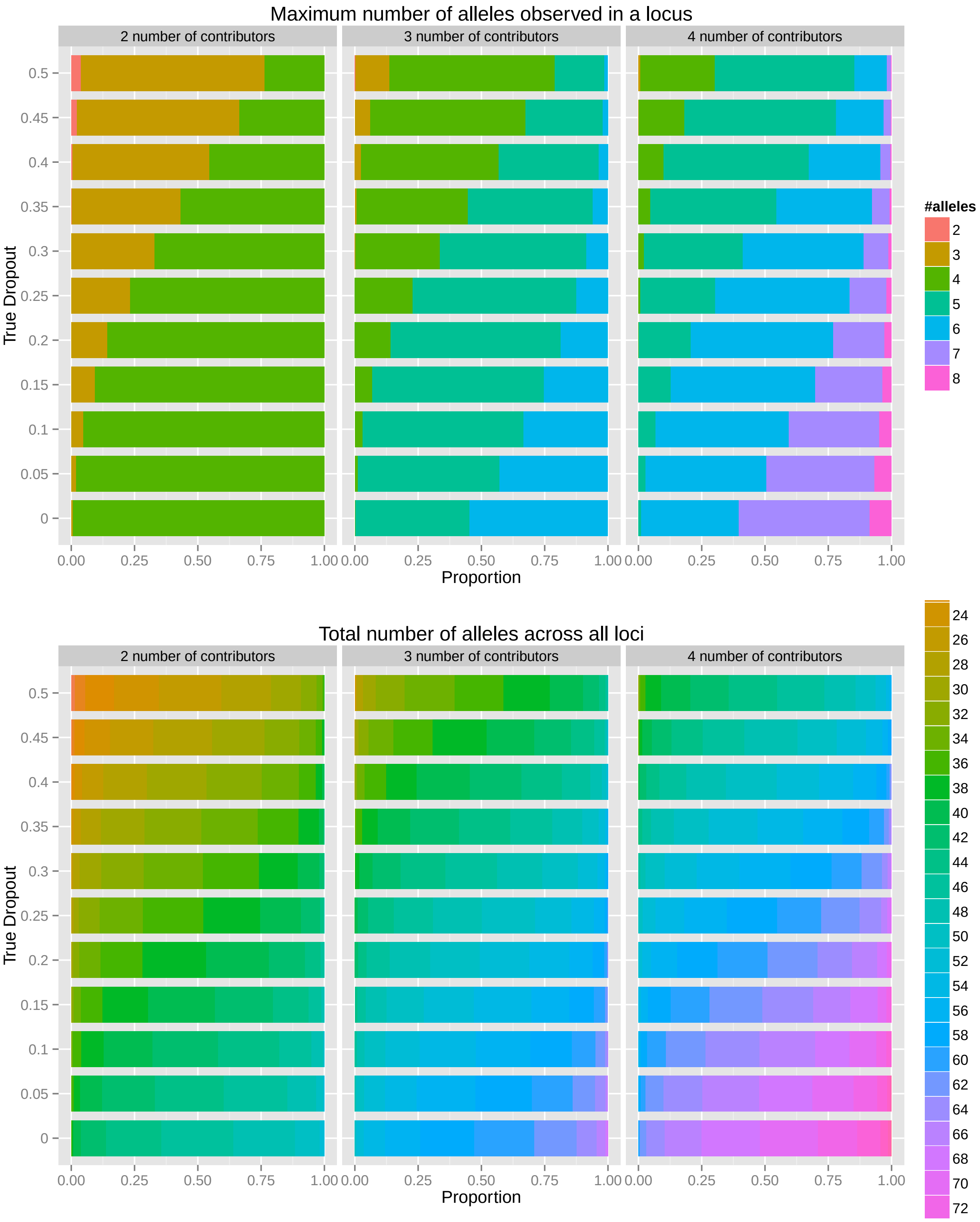

Figure S9: The upper panel shows the maximum number of observed alleles in any marker (MAC), while the lower panel shows the total number of observed alleles across all markers (TAC). The relative proportions of the numbers are based on 1000 evidence samples generated per true values of number of contributors and per drop-out probability value. All the generated samples where based on the 15 NGM markers using the Dutch allele frequencies database.

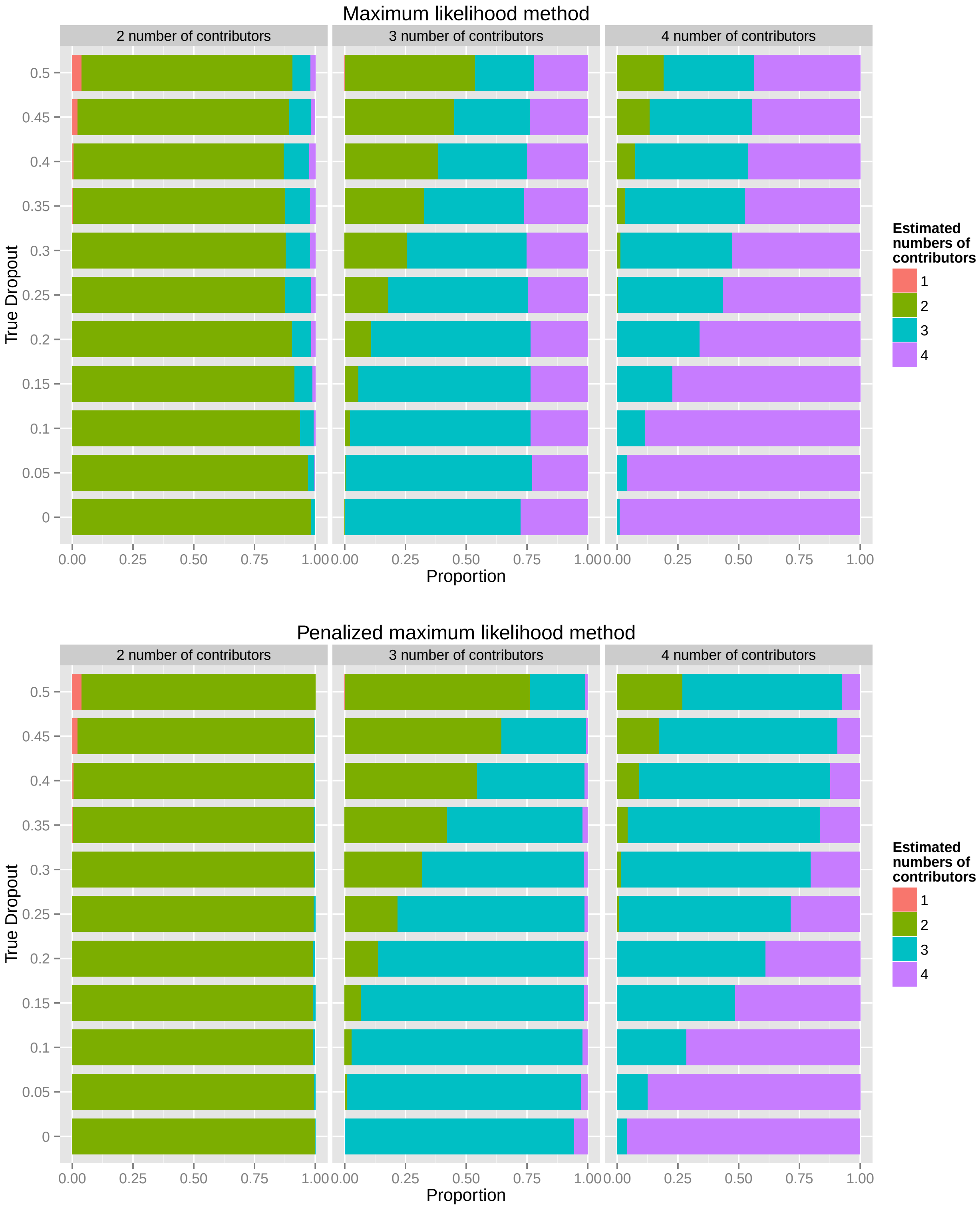

Figure S10: The figure shows the proportion of the predicted number of contributors for different true values of number of contributors and drop-out probability (based on 1000 evidence samples). The upper panel shows the predicted number of contributors using maximum likelihood estimation without penalization (i.e. $\arg\max_K l_{max}(K)$), while the lower panel shows the predicted number of contributors using maximum likelihood estimation penalized with number of contributors in the model (i.e. $\arg\max_K \{l_{max}(K) - K\}$). All the generated samples where based on the 15 NGM markers using the Dutch allele frequencies database.

Estimating the drop-out parameter

In this section a simulation study was carried out in order to clarify differences of using a quantile from the drop-out distribution or using the maximum likelihood estimate directly. The results from Figure S11 showed that the MLE method performed better than the other methods for small drop-out values, but tended to underestimate the dropout for two-person mixtures with dropout more than 35%, whereas the other methods do not. The mean and average methods perform in a very similar way, they both have much greater bias but smaller variance than the MLE method for small dropout values (up to 10%). The mode method did not perform better than the other methods (since its bias and/or variance tended to be larger).

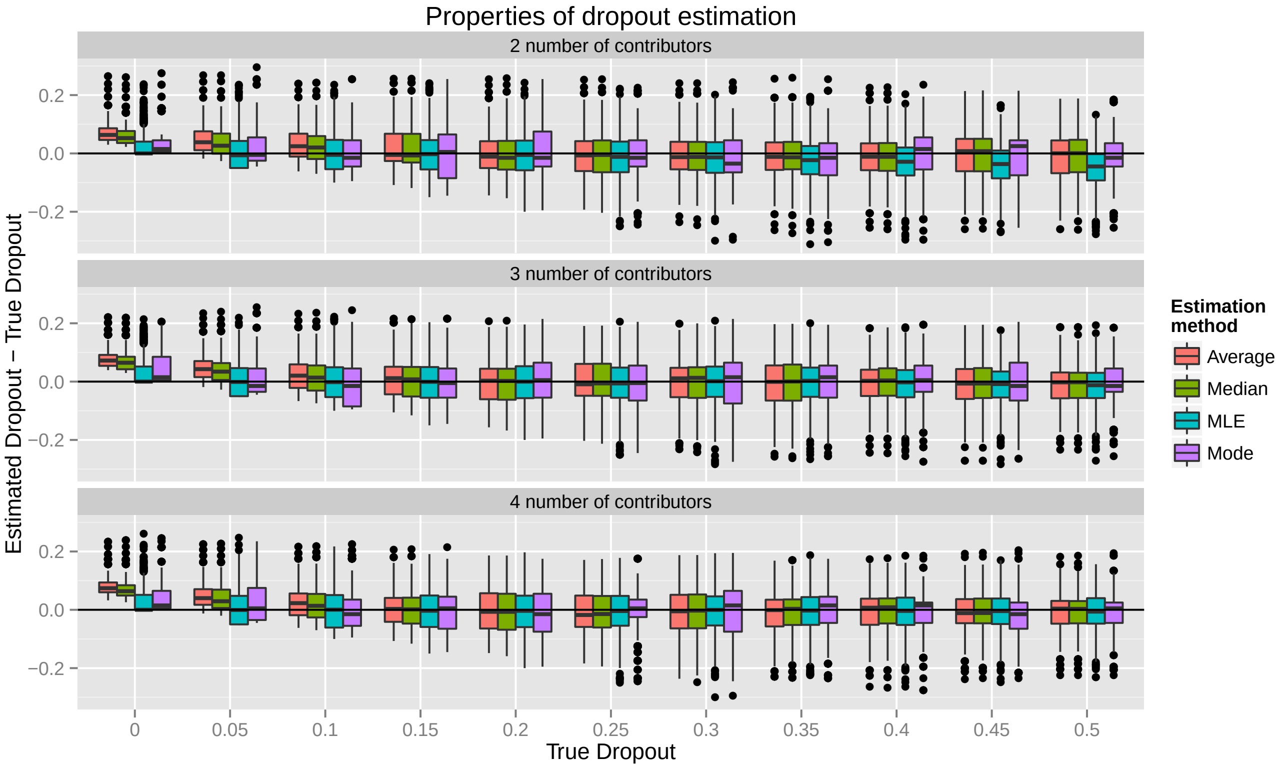

Figure S11: The figure shows the estimated dropout probability minus the true dropout probability (along the vertical axis) for different true values of number of contributors (for each panel) and drop-out probability (along the horizontal axis). The results are based on 1000 evidence samples generated with the 15 NGM markers using the Dutch allele frequencies database. The true drop-out probability is the underlying probability that an allele of a generated contributor drops out. The estimation methods considered are the median quantile, average or mode from the drop-out distribution, and the maximum likelihood estimate using the likelihood function of the qualitative model.

F.2 Inferred number of contributors for all samples

The calculation of the likelihood ratio (LR) quantity requires the number of contributors to be specified. As mentioned in the paper, possible ways to predict the number of contributors are either with manual inspection or by using a model selection framework. In the article we use a model selection framework based on Akaike information criterion (AIC), where the likelihood functions for different numbers of contributors are compared. In this section we present the results for all considered methods: The manual inspection, the qualitative AIC method and the quantitative AIC method with/without sub-population structure assumed ($F_{st} = 0/F_{st} = 0.01$). Table S3 shows detailed information about the situations where the number of contributors where incorrectly estimated. Table S4 and Table S5 show the predicted number of contributors for different types of models.

Sample	Method	$\hat{K}$	MAC, TAC	Comment
0.6	qualitative, quantitative	2	4/3/3/4, 34/31/37/39	Almost all non-sharing alleles of contributor 0B dropped out (only three non-sharing alleles in replicate 4 are present). The peak height balance in marker D16S539 indicate that this is a 3-person mixture.
0.7	quantitative	2	4/3/4/4, 37/37/38/43	vWA has five unique alleles over all replicates. However the model assumes that allele 17 is a stutter from a dropping out allele. The decision between a two and three-person mixture were very close (0.17 in AIC difference).
9.6d	quantitative	4	4/6/5, 32/27/26	This strongly degraded sample was estimated to have a very small fourth contributor (with estimated mixture proportion 0.032), with 0.5 in AIC difference over the three-person model.
2.5	quantitative	4	6, 62	Incorrectly determined even for the situation where contributor 2A was known as a true contributor. In this sample only contributor 2B dropped out with three alleles. The decision between a three and four person mixture were very close (0.17 in AIC difference).
3.5	qualitative	2	4, 44	The total number of alleles was 44 and maximum 4 alleles were observed in any of the markers. From the simulation in Figure S10 in the supplementary material we infer that with low amount of drop-out this is most likely a two-person mixture. The information that contributor 3A is a major contributor led to a correct decision.
6.1	MI	2	4, 41	High peak height uncertainty. Recommended to make replicates.
6.2	MI	2	4, 46	High peak height uncertainty. Recommended to make replicates.
6.6	qualitative	2	4, 53	The total number of alleles were 53 and maximum 4 alleles were observed in any of the markers. From the simulation in Figure S10 in the supplementary material we infer that with low amount of drop-out this is likely to be a three-person mixture. But the qualitative model assuming three-persons were only 0.3 greater in log-likelihood value than the two-person model. However, the information that contributor 6A is a major contributor led to a correct decision.
8.1	qualitative	4	5, 59	Incorrectly determined even when no stutter alleles were present. Hypothetic cause: Large total number of alleles.
8.2	quantitative	4	6, 59	Incorrectly determined even for the situation where reference 8A was known as a true contributor. A fourth contributor fits in as one of three equal minors, getting 1.3 better AIC than for three-persons. In addition, a stutter model was preferred even though no stutter alleles were present. Applying the stutter filter did not change the decision.
8.5	qualitative, quantitative	4	6, 58	The large number of alleles causes both the qualitative model (assumes 0.3 in drop-out probability) and the quantitative model to infer the sample as a four person mixture.
9.2	quantitative, MI	2	4, 40	A very degraded sample which was predicted to be a two-person mixture by manual inspection and the quantitative model. The minor contributors 9B and 9C dropped out with 14 of their alleles, each. High uncertainty in the peak heights. Recommended to generate replicates. The decision between a two and three-person mixture were very close (0.08 in AIC difference)
9.5	qualitative	4	6, 51	The total number of 51 alleles does not indicate that this is a four-person mixture with little drop-out. However the estimated drop-out probability 0.42 assuming four-persons, gives 1.3 better log-likelihood value than for three-persons (with drop-out probability 0.26) .
10.6	qualitative	4	6, 61	The sample is very degraded and has one stutter allele. The combination of a large number of alleles and a estimated drop-out probability of 0.24 gives a 1.6 better log-likelihood value than for three-persons (with drop-out probability 0.04)
11.2	qualitative	4	5, 48	The sample is very degraded and does not include any stutter alleles. The estimated drop-out probability becomes 0.44 assuming four-persons, gives a 1.05 better log-likelihood value than for three-persons (with drop-out probability 0.30).

Table S3: The table shows all instances where the number of contributors were incorrectly predicted as $\hat{K}$ (instead of a three-person mixture), for the qualitative model (for *LRmix*), quantitative model (for *EuroForMix*) or manual inspection (MI). MAC is maximum allele count and TAC is total allele counts (excluding amelogenin).

Estimation of number of contr for qualitative (Qual) and quantitative (Quan) model

Sample(s)	Cond	Contr.		Dropout/Stutter		FST		Contr. next		AIC diff	
		Qual	Quan	Qual	Quan	Qual	Quan	Qual	Quan	Qual	Quan
1.1		3	3	3.8e-07	FALSE	FST=0	FST=0	4	4	-1.8	-2
2.1		3	3	0.14	FALSE	FST=0	FST=0	4	4	-1.7	-0.26
3.1		3	3	0.25	FALSE	FST=0.01	FST=0.01	4	4	-2.4	-2
6.1		3	3	0.42	TRUE	FST=0	FST=0.01	4	4	-1.2	-2
8.1		4	3	0.27	FALSE	FST=0	FST=0.01	3	4	-0.75	-2
9.1		3	3	0.12	FALSE	FST=0	FST=0	4	4	-0.042	-1.3
10.1		3	3	0.25	FALSE	FST=0	FST=0	4	4	-0.51	-2
11.1		3	3	0.015	FALSE	FST=0	FST=0	4	4	-2.3	-2
12.1		3	3	0.06	TRUE	FST=0.01	FST=0.01	4	4	-2.6	-2
14.1		3	3	0.065	FALSE	FST=0.01	FST=0	4	4	-2.4	-2
1.2		3	3	0.096	TRUE	FST=0.01	FST=0	4	4	-3.2	-2
1.2	1A	3	3	0.21	TRUE	FST=0.01	FST=0	4	4	-2.6	-1.2
2.2		3	3	0.17	FALSE	FST=0.01	FST=0	4	4	-2.4	-2
2.2	2A	3	3	0.34	FALSE	FST=0	FST=0	4	4	-2.2	-2
3.2		3	3	0.39	TRUE	FST=0.01	FST=0.01	4	4	-1.9	-2
6.2		3	3	0.34	TRUE	FST=0	FST=0.01	2	2	-1.2	-0.76
8.2		3	4	0.063	TRUE	FST=0	FST=0	4	3	-0.36	-1.3
8.2	8A	3	4	0.15	TRUE	FST=0	FST=0	4	3	-0.092	-1.3
9.2		3	2	0.45	FALSE	FST=0	FST=0	2	3	-1.1	-0.076
10.2		3	3	0.18	TRUE	FST=0.01	FST=0	4	4	-2.1	-2
11.2		4	3	0.45	FALSE	FST=0	FST=0	3	4	-0.045	-2
12.2		3	3	0.2	TRUE	FST=0.01	FST=0	4	2	-3.1	-1.4
14.2		3	3	0.12	FALSE	FST=0.01	FST=0	4	4	-3.2	-2
2.3		3	3	2e-07	FALSE	FST=0	FST=0	4	4	-1.3	-1.9
2.3	2B	3	3	3.7e-07	FALSE	FST=0	FST=0	4	4	-2.3	-1.8
3.3		3	3	0.27	TRUE	FST=0.01	FST=0.01	4	4	-2.4	-1.9
3.3	3B	3	3	0.35	TRUE	FST=0.01	FST=0.01	4	4	-2.5	-1.9
6.3		3	3	0.075	FALSE	FST=0.01	FST=0	4	4	-2.9	-2
8.3		3	3	2.1e-07	FALSE	FST=0	FST=0	4	4	-0.27	-0.94
8.3	8A	3	3	7.5e-07	FALSE	FST=0	FST=0	4	4	-0.8	-0.81
9.3		3	3	0.075	FALSE	FST=0	FST=0	4	4	-0.79	-2
10.3		3	3	0.15	FALSE	FST=0	FST=0	4	4	-0.018	-2
11.3		3	3	0.11	FALSE	FST=0	FST=0	4	4	-2.3	-2
12.3		3	3	2.6e-06	FALSE	FST=0	FST=0	4	4	-3.1	-2
12.3	12B	3	3	0.002	FALSE	FST=0.01	FST=0	4	4	-3.2	-2
14.3		3	3	2.5e-06	FALSE	FST=0	FST=0	4	2	-2.3	-29
14.3	14A	3	3	0.00058	FALSE	FST=0	FST=0	4	4	-2.9	-2
1.5		3	3	0.057	FALSE	FST=0	FST=0	4	4	-1.5	-2
1.5	1A	3	3	0.12	FALSE	FST=0	FST=0	4	4	-1.6	-2
2.5		3	4	2.2e-07	FALSE	FST=0	FST=0	4	3	-3	-0.17
2.5	2A	3	4	5.8e-07	FALSE	FST=0	FST=0	4	3	-3.5	-0.17
3.5		2	3	0.022	FALSE	FST=0	FST=0	3	4	-2.1	-2
3.5	3A	3	3	0.47	FALSE	FST=0	FST=0	4	4	-2.1	-2
6.5		3	3	0.34	TRUE	FST=0	FST=0.01	4	4	-0.97	-2
8.5		4	4	0.3	TRUE	FST=0	FST=0	3	3	-1.8	-0.33
8.5	8A	4	4	0.44	TRUE	FST=0	FST=0	3	3	-1.6	-0.33
9.5		4	3	0.42	FALSE	FST=0	FST=0	3	4	-0.54	-2
10.5		3	3	0.35	TRUE	FST=0	FST=0	4	4	-1.2	-2
11.5		3	3	0.12	FALSE	FST=0.01	FST=0	4	4	-3.2	-2
12.5		3	3	0.19	FALSE	FST=0	FST=0	4	4	-1.3	-2
12.5	12A	3	3	0.29	FALSE	FST=0	FST=0	4	4	-1.2	-2
14.5		3	3	0.23	FALSE	FST=0	FST=0.01	4	4	-1	-2
14.5	14A	3	3	0.35	FALSE	FST=0	FST=0.01	4	4	-1.2	-2

Table S4: The table (part 1) shows an overview of the predicted number of contributors for the qualitative method ('Qual') and the quantitative method ('Quan') based on the AIC method. The corresponding best model is either no sub-population ($F_{st} = 0$) or with sub-population ($F_{st} = 0.01$). 'Dropout' is the estimated drop-out probability parameter for the qualitative model, while 'Stutter' is TRUE or FALSE depending on whether including a stutter-model was best or not. 'Contr.next' is the second best number of contributors with corresponding difference in AIC to the best given in the column 'AIC diff'. Incorrect estimates are indicated in red color.

Estimation of number of contr for qualitative (Qual) and quantitative (Quan) model

Sample(s)	Cond	Contr.		Dropout/Stutter		FST		Contr. next		AIC diff	
		Qual	Quan	Qual	Quan	Qual	Quan	Qual	Quan	Qual	Quan
1.6		3	3	0.0088	TRUE	FST=0.01	FST=0	4	4	-3.5	-1.9
1.6	1A	3	3	0.037	TRUE	FST=0.01	FST=0	4	4	-3.6	-2
2.6		3	3	2.1e-07	FALSE	FST=0	FST=0	4	4	-2.8	-2
2.6	2A	3	3	4.5e-07	FALSE	FST=0	FST=0	4	4	-3.7	-2
3.6		3	3	0.23	FALSE	FST=0.01	FST=0.01	4	4	-2.4	-2
3.6	3A	3	3	0.36	FALSE	FST=0	FST=0.01	4	4	-1.8	-2
6.6		2	3	1.4e-07	FALSE	FST=0	FST=0	3	4	-1.3	-1.6
6.6	6A	3	3	0.24	FALSE	FST=0.01	FST=0	4	4	-3.3	-2
8.6		3	3	1.9e-07	FALSE	FST=0	FST=0	4	4	-1.5	-0.46
8.6	8A	3	3	4.2e-07	FALSE	FST=0	FST=0	4	4	-2.2	-0.48
9.6		3	3	5.1e-07	FALSE	FST=0	FST=0	4	4	-1.5	-2
9.6	9A	3	3	3.6e-07	FALSE	FST=0	FST=0	4	4	-2.5	-2
10.6		4	3	0.26	FALSE	FST=0	FST=0	3	4	-1.2	-2
10.6	10A	4	3	0.36	FALSE	FST=0	FST=0	3	4	-0.86	-1.9
11.6		3	3	0.063	FALSE	FST=0.01	FST=0	4	4	-3.5	-2
11.6	11A	3	3	0.15	TRUE	FST=0.01	FST=0	4	4	-3.5	-2
12.6		3	3	2.2e-06	FALSE	FST=0	FST=0	4	4	-3.6	-2
12.6	12A	3	3	0.017	FALSE	FST=0	FST=0	4	4	-3.2	-2
14.6		3	3	0.027	FALSE	FST=0.01	FST=0	4	4	-3.2	-2
14.6	14A	3	3	0.051	FALSE	FST=0.01	FST=0	4	4	-3.6	-2
0.5		2	2	0.25	FALSE	FST=0.01	FST=0	3	3	-14	-2
0.9		2	2	0.22	FALSE	FST=0.01	FST=0	3	3	-8.8	-2
0.9	0A	2	2	0.56	FALSE	FST=0.01	FST=0	3	3	-10	-2
0.24		2	2	0.28	FALSE	FST=0.01	FST=0	3	3	-36	-2
0.24	0C	2	2	0.78	FALSE	FST=0.01	FST=0	3	3	-6.2	-2
0.28		2	2	0.27	FALSE	FST=0.01	FST=0	3	3	-23	-2
0.28	0C	2	2	0.63	FALSE	FST=0	FST=0	3	3	-10	-2
0.6		2	2	0.26	FALSE	FST=0.01	FST=0	3	3	-21	-2
0.7		3	2	0.36	TRUE	FST=0.01	FST=0	4	3	-18	-0.17
0.10		3	3	0.33	TRUE	FST=0.01	FST=0	4	2	-11	-0.75
0.10	0A	3	3	0.67	TRUE	FST=0.01	FST=0.01	4	2	-5.5	-0.91
0.11		3	3	0.28	FALSE	FST=0.01	FST=0	4	4	-9.9	-2
0.11	0A	3	3	0.53	FALSE	FST=0.01	FST=0	4	2	-11	-23
8.7d		3	3	0.69	FALSE	FST=0	FST=0	4	4	-3	-2
9.6d		3	4	0.65	FALSE	FST=0	FST=0	4	3	-4.3	-0.47

Table S5: The table (part 2) shows an overview of the predicted number of contributors for the qualitative method ('Qual') and the quantitative method ('Quan') based on the AIC method. The corresponding best model is either no sub-population ($F_{st} = 0$) or with sub-population ($F_{st} = 0.01$). 'Dropout' is the estimated drop-out probability parameter for the qualitative model, while 'Stutter' is TRUE or FALSE depending on whether including a stutter-model was best or not. 'Contr.next' is the second best number of contributors with corresponding difference in AIC to the best given in the column 'AIC diff'. Incorrect estimates are indicated in red color.

G Detailed results

G.1 False negative results

All the false negative instances (i.e. $LR < 1|H_p$ is 'TRUE') are listed in the Tables S6 and S7. These all concern instances where a minor contributor is the person of interest (POI). From the table we observed that there were many more false negatives using *LRmix* compared to *EuroForMix* (28 versus five for MLE method and 67 versus 11 for the conservative method). From Table 2 in the article we have that the number of instances where *EuroForMix* gave a true positive when *LRmix* gave a false negative were 25 using the MLE method and 56 using the conservative method. Opposite, the number of instances where *LRmix* gave a true positive when *EuroForMix* gave a false negative were two using the MLE method and zero using the conservative method. From the Tables S6 and S7 we observed that the LR values were much higher for *EuroForMix* than *LRmix* for both the MLE method and the conservative method for almost all situations (except where *EuroForMix* underestimated the number of contributor as two-persons instead of three-persons). From the estimated relative DNA amount between the contributors (i.e. last column), it was observed that *EuroForMix* was able to estimate the major components in almost all situations. This introduced fewer possibilities of the genotype of the non-major components, which increased the discriminatory power. In section G.2 we investigate closer the five comparisons where where the LR of the true donors were greater for *LRmix* using conservative method than for *EuroForMix* using MLE.

MLE method (Table A)

Sample	POI|cond	DNA (pg)	#d	K	*LRmix*	K_Q	*EuroForMix*	K_C	$\hat{\pi}_1/\hat{\pi}_2/.../\hat{\pi}_K$
8.6	8C	500:250:**50**	4	3	0.06	3	1e+10	3	0.66/0.28/0.06
8.3	8C	250:250:**50**	4	3	0.03	3	2e+09	3	0.45/0.45/0.1
0.24	0A	**30**:150	15.5	2	1e-05	2	8e+07	2	0.87/0.13
9.3	9C	250:250:**50**	6	3	0.001	3	760000	3	0.44/0.44/0.12
6.6	6C	500:250:**50**	7	3	3e-23	2	150000	3	0.45/0.45/0.1
11.3	11C	250:250:**50**	9	3	6e-05	3	140000	3	0.43/0.43/0.14
14.6	14C	500:250:**50**	4	3	0.3	3	97000	3	0.64/0.18/0.18
11.1	11B	100:50:**50**	4	3	0.1	3	10000	3	0.33/0.33/0.33
11.6	11C	500:250:**50**	8	3	0.0002	3	1300	3	0.53/0.37/0.1
1.6	1C	500:250:**50**	8	3	4e-05	3	1200	3	0.7/0.26/0.04
1.6	1C|1A	500:250:**50**	8	3	0.0006	3	1300	3	0.7/0.26/0.04
14.2	14B	250:50:**50**	7	3	0.4	3	1300	3	0.65/0.18/0.18
2.6	2C	500:250:**50**	4	3	0.09	3	1100	3	0.45/0.45/0.1
8.2	8B	250:50:**50**	6	3	0.007	3	830	4	0.73/0.09/0.09/0.09
1.2	1B	250:50:**50**	11	3	2e-05	3	740	3	0.75/0.12/0.12
1.2	1B|1A	250:50:**50**	11	3	0.03	3	770	3	0.75/0.12/0.12
2.2	2B	250:50:**50**	9	3	0.03	3	520	3	0.72/0.14/0.14
14.1	14B	100:50:**50**	6	3	0.2	3	190	3	0.38/0.31/0.31
0.6	0C	150:6:**30**	12	3	2e-06	2	180	2	0.76/0.24
9.5	9C	500:50:**50**	7	3	0.7	4	180	3	0.78/0.11/0.11
11.5	11B	500:50:**50**	6	3	0.09	3	160	3	0.64/0.22/0.15
9.6d	9C	500:250:**50**	17	3	0.8	3	150	4	0.51/0.24/0.24/2e-08
0.10	0B	300:**6**:30	11.3	3	1e-06	3	91	3	0.88/0.09/0.03
14.5	14B	500:50:**50**	8	3	0.02	3	14	3	0.78/0.11/0.11
8.7d	8C	500:250:**250**	18	3	0.3	3	10	3	0.56/0.22/0.22
9.2	9B	250:50:**50**	14	3	4	3	3e-06	2	0.78/0.22
0.7	0B	150:**30**:30	10.5	3	27	3	2e-14	2	0.71/0.29
9.2	9C	250:50:**50**	14	3	0.004	3	2e-19	2	0.78/0.22
0.7	0C	150:30:**30**	12	3	4e-08	3	2e-20	2	0.71/0.29
0.6	0B	150:**6**:30	14.75	3	2e-41	2	5e-41	2	0.76/0.24

Table S6: The table shows all examples where the LR of the true contributors was less than 1 when using the MLE method. All examples where x are shown. 'POI|cond' is the person of interest (POI) with possible conditional reference. 'DNA' is amount of DNA for for each contributors, with the boldface indicating POI. K is the true number of contributors, while K_Q and K_C are the predicted number of contributors for *LRmix* and *EuroForMix*, respectively. #d is the number of dropouts (averaged if replicates where available). $\hat{\pi}$ is the MLE of the relative amounts of DNA for the unknown contributors using *EuroForMix* (under H_d).

Conservative method (Table B)

Sample	POI|cond	DNA (pg)	#d	K	*LRmix*	K_Q	*EuroForMix*	K_C	$\hat{\pi}_1/\hat{\pi}_2/.../\hat{\pi}_K$
0.5	0C	150:**30**	13.25	2	8e-14	2	2e+16	2	0.76/0.24
0.28	0A	**30**:300	12.25	2	3e-07	2	2e+11	2	0.93/0.07
8.6	8C|8A	500:250:**50**	4	3	0.003	3	2e+08	3	0.66/0.28/0.06
0.10	0C	300:6:**30**	9.75	3	1e-09	3	1e+08	3	0.88/0.09/0.03
0.10	0C|0A	300:6:**30**	9.75	3	0.002	3	2e+08	3	0.88/0.09/0.03
8.3	8C|8A	250:250:**50**	4	3	0.0003	3	4e+07	3	0.51/0.4/0.09
0.11	0C	300:30:**30**	8.25	3	0.6	3	2e+07	3	0.8/0.14/0.06
2.3	2C	250:250:**50**	3	3	0.007	3	8e+06	3	0.44/0.44/0.12
2.3	2C|2B	250:250:**50**	3	3	0.1	3	6e+06	3	0.48/0.4/0.12
0.24	0A|0C	**30**:150	15.5	2	4e-12	3	4500000	3	0.87/0.13
11.5	11C	500:50:**50**	4	3	0.7	3	730000	3	0.63/0.22/0.15
1.5	1B	500:50:**50**	4	3	0.003	3	170000	3	0.82/0.09/0.09
9.1	9B	100:50:**50**	6	3	0.0004	3	110000	3	0.37/0.37/0.26
14.3	14C	250:250:**50**	2	3	0.9	3	52000	3	0.33/0.33/0.33
0.11	0B	300:30:**30**	7.5	3	2e-06	3	51000	3	0.8/0.14/0.06
0.11	0B|0A	300:30:**30**	7.5	3	0.04	3	28000	3	0.8/0.14/0.06
11.6	11C|11A	500:250:**50**	8	3	2e-05	3	32000	3	0.53/0.41/0.06
10.3	10C|	250:250:**50**	5	3	0.0002	3	30000	3	0.39/0.39/0.23
8.2	8C	250:50:**50**	3	3	0.17	3	21000	4	0.73/0.09/0.09/0.09
14.6	14C|14A	500:250:**50**	4	3	0.0007	3	7200	3	0.64/0.18/0.18
12.5	12C	500:50:**50**	5	3	0.3	3	1800	3	0.76/0.12/0.12
10.2	10B	250:50:**50**	9	3	9e-05	3	1500	3	0.62/0.19/0.19
2.5	2B	500:50:**50**	3	3	0.003	3	980	4	0.73/0.09/0.09/0.09
1.5	1C	500:50:**50**	4	3	0.01	3	480	3	0.82/0.09/0.09
1.1	1B	100:50:**50**	4	3	2e-05	3	220	3	0.51/0.25/0.25
2.6	2C|2A	500:250:**50**	4	3	0.0002	3	170	3	0.56/0.35/0.09
10.1	10B	100:50:**50**	7	3	0.6	3	83	3	0.33/0.33/0.33
3.5	3C	500:50:**50**	4	3	0.02	3	57	3	0.84/0.12/0.04
2.2	2B|2A	250:50:**50**	9	3	0.2	3	13	3	0.72/0.14/0.14
8.2	8B|8A	250:50:**50**	6	3	0.0001	3	11	4	0.73/0.09/0.09/0.09
8.1	8B	100:50:**50**	6	3	0.04	4	10	3	0.33/0.33/0.33
2.1	2B	100:50:**50**	7	3	8e-06	3	6	3	0.33/0.33/0.33
8.7d	8B	500:**250**:250	17.3	3	0.06	3	6	3	0.56/0.22/0.22
0.10	0B|0A	300:**6**:30	11.25	3	2e-15	3	3	3	0.88/0.09/0.03
6.2	6C	250:50:**50**	9	3	0.7	3	0.4	3	0.47/0.34/0.19
9.6d	9C	500:250:**50**	17	3	0.001	3	0.4	4	0.51/0.24/0.24/2e-08
8.7d	8C	500:250:**250**	18	3	0.002	3	0.3	3	0.56/0.22/0.22
6.1	6B	100:50:**50**	10	3	0.5	3	0.2	3	0.37/0.37/0.26
14.5	14B|14A	500:50:**50**	8	3	0.03	3	0.2	3	0.78/0.11/0.11
14.5	14B	500:50:**50**	8	3	2e-05	3	0.1	3	0.78/0.11/0.11
9.2	9B	250:50:**50**	14	3	0.7	3	1e-07	2	0.78/0.22
0.7	0B	150:**30**:30	10.5	3	2e-08	3	7e-16	2	0.71/0.29

Table S7: The table shows all examples where the LR of the true contributors was less than 1 when using the conservative method (where results from Table A are omitted). 'POI|cond' is the person of interest (POI) with possible conditional reference. 'DNA' is amount of DNA for for each contributors, with the boldface indicating POI. K is the true number of contributors, while K_Q and K_C are the predicted number of contributors for *LRmix* and *EuroForMix*, respectively. #d is the number of dropouts (averaged if replicates where available). $\hat{\pi}$ is the MLE of the relative amounts of DNA for the unknown contributors using *EuroForMix* (under H_d).

G.2 Comparisons where *LRmix* gave a higher LR than *EuroForMix*

Table S8 (A) shows the five comparisons (out of 228 possible) where the conservative method for *LRmix* gave higher LR values compared to the MLE method for *EuroForMix*, when the person of interest was the true donor (H_p is 'TRUE'). The current method for *LRmix* is the conservative method, while *EuroForMix* follows [10] and uses the MLE approach. Hence it is relevant to compare these two sets of data. The samples '0.7' and '9.2' were incorrectly interpreted as a two-person mixture instead of a three-person mixture when following the AIC model selection approach using *EuroForMix* because of the large amount of drop-out (the references 9B and 9C had 14 allele drop-out each in sample '9.2', while the references 0B and 0C had 42 and 48 allele drop-outs across all replicates in sample '0.7'). By assuming the correct number of contributors, only two instances remained where the *LRmix* LR (conservative) was larger than the *EuroForMix* LR (MLE) (i.e. reference 3C compared with sample '3.3' and '3.2'), but the differences were within an order of magnitude (on base 10 scale). Steele and Balding [36] propose that differences of one order of magnitude are inconsequential, hence we conclude that these two samples appear comparable.

Table A: Number of contributors based on the model selection procedure

Sample	POI|cond	LRmix MLE	LRmix CONS	K_Q	EuroForMix MLE	EuroForMix CONS	K_C
3.3	3C|3B	75000	51000	3	30000	540	3
3.2	3C	5700	5000	3	1200	16	3
9.2	9B	4	0.7	3	3e-06	1e-07	2
0.7	0B	27	2e-8	3	2e-14	7e-16	2
9.2	9C	0.004	0.0006	3	2e-19	7e-21	2

Table B: Changed results when the correct number of contributors were assumed

Sample	poi|cond	LRmix MLE	LRmix CONS	K_Q	EuroForMix MLE	EuroForMix CONS	K_C
0.7	0B	27	2e-8	3	250000	7400	3
9.2	9B	4	0.7	3	3000	17	3
9.2	9C	0.004	0.0006	3	5	0.06	3

Table S8: Table A shows all situations where the LR of the true donors were greater for "*LRmix*" using conservative method than for "*EuroForMix*" using MLE. 'POI|cond" is the compared person of interest (POI) with conditioned reference, where indicated. Table B shows the LR values when the correct number of contributors were assumed. K_Q and K_C is the number of contributors assumed in *LRmix* and *EuroForMix*, respectively.

G.3 Deconvolution results

This section provides detailed results for the study carried out in section 4.3: “Comparison of deconvolution methods” in the article.

Sample(s)	Compared	DNA	#d	Ratio	$\widehat{\text{Ratio}}$	Certain matches (Full/Partial/No)	
						EFM	LOC
0.5.(1-4)	0A	150	4	5:1	3.1 : 1	15/0/0	7/0/0
0.9.(1-4)	0A	300	0	10:1	12 : 1	15/0/0	15/0/0
0.24.(1-4)	0C	150	0	5:1	7 : 1	15/0/0	13/0/0
0.28.(1-4)	0C	300	0	10:1	13 : 1	15/0/0	13/0/0
0.6.(1-4)	0A	150	6	5:1:0.2	3.2 : 1	14/0/0	9/0/0
0.7.(1-4)	0A	150	5	5:1:1	2.4 : 1	15/0/0	5/0/0
0.10.(1-4)	0A	300	0	10:0.2:1	10 : 0.4 : 1	15/0/0	14/0/0
0.11.(1-4)	0A	300	0	10:1:1	13 : 2.3 : 1	15/0/0	14/0/0
1.1	1A	100	0	2:1:1	2.1 : 1 : 1	11/2/0	15/0/0
2.1	2A	100	1	2:1:1	1 : 1 : 1	0/0/0	3/2/0
3.1	3A	100	0	2:1:1	3 : 3 : 1	1/0/0	3/0/0
6.1	6A	100	6	2:1:1	1.5 : 1.5 : 1	0/0/0	0/2/0
8.1	8A	100	0	2:1:1	1 : 1 : 1	0/0/0	2/0/0
9.1	9A	100	0	2:1:1	1.5 : 1.5 : 1	1/0/0	1/0/0
10.1	10A	100	1	2:1:1	1 : 1 : 1	0/0/0	0/0/0
11.1	11A	100	0	2:1:1	1 : 1 : 1	0/0/0	1/0/1
12.1	12A	100	1	2:1:1	1 : 1 : 1	0/0/0	2/1/0
14.1	14A	100	0	2:1:1	1.2 : 1 : 1	2/0/0	2/1/0
1.2	1A	250	0	5:1:1	6.1 : 1 : 1	15/0/0	13/0/0
2.2	2A	250	0	5:1:1	5.1 : 1 : 1	15/0/0	10/0/0
3.2	3A	250	0	5:1:1	8.7 : 2.2 : 1	15/0/0	13/0/0
6.2	6A	250	0	5:1:1	2.5 : 1.8 : 1	6/0/0	3/0/0
8.2	8A	250	0	5:1:1	8.6 : 1 : 1 : 1	15/0/0	13/0/0
9.2	9A	250	1	5:1:1	3.6 : 1	12/0/0	11/0/0
10.2	10A	250	0	5:1:1	3.4 : 1 : 1	13/1/0	3/0/0
11.2	11A	250	2	5:1:1	2.3 : 1 : 1	7/1/0	4/0/0
12.2	12A	250	0	5:1:1	3.5 : 1 : 1	13/1/0	7/0/0
14.2	14A	250	0	5:1:1	3.7: 1 : 1	14/0/0	9/0/0
1.5	1A	500	0	10:1:1	9.3 : 1 : 1	15/0/0	14/0/0
2.5	2A	500	0	10:1:1	8.4 : 1 : 1 : 1	15/0/0	11/0/0
3.5	3A	500	0	10:1:1	18 : 2.6 : 1	15/0/0	13/0/0
6.5	6A	500	0	10:1:1	6.8 : 1 : 1	15/0/0	14/0/0
8.5	8A	500	0	10:1:1	18 : 1 : 1 : 1	15/0/0	13/0/0
9.5	9A	500	0	10:1:1	7 : 1 : 1	15/0/0	13/0/0
10.5	10A	500	0	10:1:1	7.7 : 1 : 1	15/0/0	14/0/0
11.5	11A	500	0	10:1:1	4.3 : 1.5 : 1	11/1/0	4/1/0
12.5	12A	500	0	10:1:1	6.2 : 1 : 1	15/0/0	14/0/0
14.5	14A	500	0	10:1:1	6.7 : 1 : 1	15/0/0	11/0/0
1.6	1A	500	0	10:5:1	17 : 6.2 : 1	15/0/0	13/0/0
2.6	2A	500	0	10:5:1	4.3 : 4.3 : 1	1/0/0	3/1/0
3.6	3A	500	0	10:5:1	8.8 : 8.8 : 1	1/0/0	4/0/0
6.6	6A	500	0	10:5:1	4.6 : 4.6 : 1	1/1/0	1/1/0
8.6	8A	500	0	10:5:1	16 : 7 : 1 : 1	14/0/0	7/0/0
9.6	9A	500	0	10:5:1	2.1 : 1 : 1	9/0/0	4/0/0
10.6	10A	500	0	10:5:1	3 : 3 : 1	1/0/0	1/0/0
11.6	11A	500	0	10:5:1	5.4 : 3.8 : 1	7/3/0	1/2/0
12.6	12A	500	0	10:5:1	4.2 : 4.2 : 1	0/0/0	1/0/0
14.6	14A	500	0	10:5:1	3.6 : 1 : 1	13/2/0	8/0/0

Table S9: The table shows the deconvolution results for *EuroForMix* (EFM) and *LoCIM-tool* (LOC), where references in the column 'Compared' were compared with the first ranked profile genotype. The results are given as the number of markers where the comparison gave 'Full', 'Partial' or 'No' match, where 'Full' means that the two locus genotypes where equal, 'Partial' means that only one allele between the locus genotypes were equal, and 'None' means that no alleles between the locus genotypes where equal. A summation was used to classify the locus genotype as either 'certain' or 'uncertain'. Column "DNA" refers to the amount of DNA of reference. The column '#d' denotes number of dropout of “Compared”, “Ratio” and $\widehat{\text{Ratio}}$ provides the true and estimated (MLE using *EuroForMix*) relative amounts of DNA between the unknown contributors. Markers where predictions involved drop-outs were removed.

Sample(s)	Compared	DNA	#d	Ratio	$\widehat{\text{Ratio}}$	Certain matches (Full/Partial/No)
8.7d(2-4)	8A	500	42	10:5:5	10 : 4 : 4	5/2/0
9.6d(2-4)	9A	500	38	10:5:1	13 : 7.7 : 7.7 : 1	4/0/0
2.3	2A	250	0	5:5:1	3.6 : 3.6 : 1	1/0/0
2.3	2A|2B	250	0	5:5:1	3.4 : 4 : 1	15/0/0
2.3	2B	250	0	5:5:1	3.6 : 3.6 : 1	1/0/0
3.3	3A	250	0	5:5:1	14 : 14 : 1	0/0/0
3.3	3A|3B	250	0	5:5:1	19 : 13 : 1	13/0/0
6.3	6A	250	0	5:5:1	2.6 : 1 : 1	2/8/0
6.3	6B	250	0	5:5:1	2.6 : 1 : 1	5/5/0
8.3	8B	250	0	5:5:1	4.7 : 4.7 : 1	0/0/0
8.3	8B|8A	250	0	5:5:1	5.5 : 4.3 : 1	15/0/0
9.3	9A	250	0	5:5:1	3.7 : 3.7 : 1	0/0/0
9.3	9B	250	0	5:5:1	3.7 : 3.7 : 1	0/0/0
10.3	10A	250	0	5:5:1	1.7 : 1.7 : 1	0/1/0
10.3	10B	250	2	5:5:1	1.7 : 1.7 : 1	0/1/0
11.3	11A	250	0	5:5:1	3.1 : 3.1 :1	0/1/0
11.3	11B	250	0	5:5:1	3.1 : 3.1 : 1	0/1/0
12.3	12A	250	0	5:5:1	3.9 : 5.2 : 1	4/1/0
12.3	12A|12B	250	0	5:5:1	4 : 6.4 : 1	12/1/0
12.3	12B	250	0	5:5:1	3.9 : 5.2 : 1	7/1/0
14.3	14A	250	0	5:5:1	1 : 1 : 1	1/0/0
14.3	14B	250	0	5:5:1	1 : 1 : 1	1/0/0
14.3	14B|14A	250	0	5:5:1	1.9 : 1 : 1	4/0/0
1.6	1B	250	0	10:5:1	17 : 6.2 :1	14/0/0
1.6	1B|1A	250	0	10:5:1	17 : 6.2 : 1	14/0/0
2.6	2B	250	0	10:5:1	4.3 : 4.3 : 1	1/0/0
2.6	2B|2A	250	0	10:5:1	5.9 : 3.7 : 1	15/0/0
3.6	3B	250	0	10:5:1	8.8 : 8.8 : 1	0/1/0
3.6	3B|3A	250	0	10:5:1	11 : 7.4 : 1	14/1/0
6.6	6B	250	0	10:5:1	4.6 : 4.6 : 1	1/1/0
6.6	6B|6A	250	0	10:5:1	6 : 3.9 : 1	12/0/0
8.6	8B	250	0	10:5:1	16 : 7 : 1 : 1	13/0/0
8.6	8B|8A	250	0	10:5:1	17 : 6.9 : 1 : 1	15/0/0
9.6	9B	250	0	10:5:1	2.1 : 1 : 1	0/0/0
9.6	9B|9A	250	0	10:5:1	2.6 : 1 : 1	0/0/0
10.6	10B	250	0	10:5:1	3 : 3 : 1	1/0/0
10.6	10B|10A	250	0	10:5:1	4.2 : 2.8 :1	13/0/0
11.6	11B	250	0	10:5:1	5.4 : 3.8 : 1	6/3/0
11.6	11B|11A	250	0	10:5:1	9.4 : 7.2 : 1	14/1/0
12.6	12B	250	0	10:5:1	4.2 : 4.2 : 1	0/0/0
12.6	12B|12A	250	0	10:5:1	5.1 : 3.9 : 1	15/0/0
14.6	14B	250	0	10:5:1	3.6 : 1 : 1	2/1/0
14.6	14B|14A	250	0	10:5:1	4.3 : 1.4 : 1	8/2/0

Table S10: The table shows the deconvolution results for *EuroForMix*, where the compared reference is given by “Compared” which were compared with the first ranked profile genotype. The results are given as number of markers where the comparison gave ’Full’, ’Partial’ or ’No’ match, where ’Full’ means that the two genotypes where equal, ’Partial’ means that only one allele between the genotypes where equal, and ’None’ means that no alleles between the genotypes where equal. This was summed up for the situation where the predicted genotype was classified as either certain or uncertain. Column ”DNA” refers to the amount of DNA for “Compared”. The column ’#d’ denotes number of dropout of “Compared”, “Ratio” and $\widehat{\text{Ratio}}$ provides the true and estimated (MLE using *EuroForMix*) relative amounts of DNA for the unknown contributors.