Forensic Science International: Genetics 40 (2019) 52–63

Contents lists available at ScienceDirect

Forensic Science International: Genetics

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

Research paper

The Yara Gambirasio case: Combining evidence in a complex DNA mixture case

Therese Graversen[a], Julia Mortera[b,*], Giampietro Lago[c]

[a] University of Copenhagen, Denmark
[b] Universitá Roma Tre, Italy
[c] Raggruppamento Carabinieri Investigazioni Scientifiche, Parma, Italy

ARTICLE INFO

Keywords:
Combining evidence
DNA mixtures
Forensic statistics
Kinship
Likelihood ratio
Mixture deconvolution
Probabilistic genotyping

ABSTRACT

Here, we illustrate how statistical methods can help extract information from mixed DNA profiles pertaining to an Italian case, referred to by the media as *The murder of Yara Gambirasio*. We base the analysis on a model for DNA mixtures that takes fully into account the peak heights and possible artefacts, like stutter and dropout that might occur in the DNA amplification process. We show how to combine the evidence from multiple samples and from different marker systems all within the model framework. The combined evidence is used for deconvolution, where the focus is to find likely profiles for the donors to the sample. We also show how a mixture can be used to establish familial relationships between a reference profile and a donor to the mixed DNA sample. We compare results based on a single mixed DNA profile, combination of replicates, combinations of different samples and combinations of different kits. Based on the Yara case, we discuss just a few of the plethora of possibilities of combining evidential information.

1. Introduction

Through discussing three different aspects of identification, we wish to illustrate how statistical methods can help extract information from mixed DNA profiles. One common use of statistical models for mixtures focuses on computing a likelihood ratio that quantifies the evidence for the presence of DNA from a specific person of interest. A second use is for deconvolution, where the focus is more investigative and aims to find likely profiles for the donors to the sample. As a third aspect we illustrate how a mixture can be used to establish familial relationships between a reference profile and a donor to the sample. All these analyses can be based on different combinations of evidence. The simplest analyses are based on a single mixed DNA profile, or possibly a set of replicates thereof. For more complex analyses, we may wish to use a combination of mixed DNA profiles taken from different samples, or profiles that are typed using different kits. Using evidence from a real case, we discuss just a few of the plethora of possibilities of combining evidential information.

Benschop et al. [1] and Steele et al. [19] discuss the potential gain in information achieved by splitting a sample pre-extraction and thus basing the analysis on multiple EPGs from subsamples rather than on a single EPG. Using `LRmix` for the analysis, Benschop et al. [1] found a gain in information about the major contributor, but a loss of information about the minor contributor. The analysis based on peak heights showed no systematic gain or loss of information due to replication [19]. That a potential gain in information may be achieved by combining DNA evidence from multiple samples is also mentioned in [15].

Problems concerning kinship in DNA mixtures have been investigated in [14,7,2,16,18], but most of these methods are based solely on the qualitative information about the detected alleles in the mixture, whereas we use the quantitative information carried in the peak heights. Here we use the methods in [13] which build on the model in [3], that takes peak height information fully into account.

In this paper we discuss part of the evidence relating to an Italian case referred to by the media as *The murder of Yara Gambirasio*. The paper is organised as follows. After a brief overview of the Yara case and the evidence pertaining to our analyses we give an overview of our statistical model for one or more DNA mixtures and the framework implemented in the `DNAmixtures` software. We then discuss how to identify the genotypes of the donors to the mixed profiles. Finally we discuss how an extended model can be used to evaluate evidence for potential relationships based on mixed DNA profiles using the `KinMix` software.

[*] Corresponding author.
E-mail addresses: graversen@math.ku.dk (T. Graversen), julia.mortera@uniroma3.it (J. Mortera), giampietro.lago@carabinieri.it (G. Lago).

https://doi.org/10.1016/j.fsigen.2018.12.010
Received 27 October 2018; Received in revised form 11 December 2018; Accepted 21 December 2018
Available online 05 January 2019

1.1. The murder of Yara Gambirasio

On Friday 26 November 2010 at 17:15 13-year-old Yara Gambirasio left home in Brembate di Sopra, a small town near Milan, Italy, to go to the gym. An hour and a half later she left the gym never to return home. Three months later her body was found in an abandoned field in an industrial area 10 km south of Brembate di Sopra. She had suffered multiple injuries from a sharp weapon, which had pierced her clothing at various points. It seemed that she had been attacked and abandoned. She had died slowly from hemorrhage and hypothermia.

The DNA from the genetic material that was taken from the victim's clothes was analysed. The DNA extracted from the front and the waistband of her underpants showed the presence of male DNA. The analysis of these DNA profiles that we will show here, lead to the profile of an unknown contributor, referred to by the media as *Ignoto 1*. It was assumed that this profile was from the murderer, who had left his DNA on the girl's underpants.

The profile of *Ignoto 1* was compared to 18,000 DNA samples taken from relatives of the deceased and from many thousands of male individuals who were either local or known to have been in the area around the time of Yara's disappearance. Comparisons were previously also made with Interpol criminal databases, but had not given any leads.

Familial search showed that two brothers, who were visitors to a nearby nightclub and unrelated to the crime, shared many alleles with *Ignoto 1* and could therefore potentially be related to the murderer. A DNA sample from their mother revealed that she shared no alleles with *Ignoto 1*. The brothers' father, GG, was a bus driver who had died in 1999, eleven years before the crime. A DNA profile was at first retrieved from a stamp he had licked, and in March 2013 DNA was extracted from his exhumed body. The resulting profile identified him as overwhelmingly likely to be the father of *Ignoto 1*.

However, it was apparently totally unknown to anyone that GG had any other children, so it was hypothesized that GG had an illegitimate child. The investigators then decided to screen women who potentially could have borne him a child decades earlier. By combing through the population registers of the time they found a woman, EA. Before moving to Brembate di Sopra she had, in fact, lived in the same village as GG and an analysis of EA's DNA showed that it was compatible with that of the mother of *Ignoto 1*. Thus EA's son, MGB, became the chief suspect.

MGB was sentenced to life imprisonment on 1 July 2016 for the murder of Yara. On 18 July 2017 the appeal court upheld the life sentence. On 12 October 2018 the *Corte di Cassazione*, the Italian Supreme Court, confirmed the sentence to life imprisonment.

2. Materials and methods

2.1. Crime scene profiles

A thorough analysis of Yara's clothes revealed the presence of male biological material in an area on her underpants (exhibit 31). This area was further inspected through 24 virtual grid cells (G1 to G24). Each grid cell was split into two parts.

Serological analyses were made on one part of the grid cell, in order to determine the biological nature of the male contribution (e.g. blood, saliva, or sperm). All serological tests available were applied and, in all cases, the tests were positive for blood, and negative for other tissues. However, these negative results are not conclusive due to the probable presence of blood from the victim, as also observed from samples with only her DNA. Thus, a technical diagnosis of the biological nature of the male contribution was not possible since none of the blood tests can discriminate between blood/blood and blood/other tissues. After examination of the dimension, shape, type, amount and type of diffusion of the biological material, the biologists indicated that the more likely hypothesis was that the male contribution was made from blood.

The other part of each grid cell was used for quantitation and successive DNA profiling with at least three different kits among NGM, Identifiler, PowerPlex ESI16, and PowerPlex ESX16. Some grid cells, *e.g.* G20, were also analysed with Argus X, Y-Filer, and PowerPlex 16 in order to obtain information about the X- and Y-chromosomes and the two Penta loci.

In this paper, we only analyse the following pieces of evidence pertaining to Yara's underpants, but emphasise that many more pieces of evidence played a part in resolving the case.

- An NGM profile from each of the grid cells G13-G16 and G20;
- an ESX16 profile from grid cell G14; and
- two replicates (R1 and R2) of an Identifiler profile obtained from grid cell G20.

An excerpt of the profiles can be seen in Fig. 1, which shows all the observed peaks above 150RFU. Some loci, e.g. D18 for sample G13, exhibit complete dropout at this level of detection. An excerpt of GG, EA, MGB and the victim's genotypes is given in Table 14, Section 5.4.

2.2. A statistical model for mixed DNA profiles

The statistical model for peak heights was developed in [4,5,3]. We base our analyses on the model specified in [3], which takes fully into account the peak height information. We give a brief summary of the main features of the model and refer to [3] for details. An in-depth interpretation of the model and its parameters can also be found in [9].

Our statistical framework is a model that describes both the genotypes of the donors and the set of peak heights that we may observe in a crime scene profile. Such a model can be specified in two components, where one describes the variability that may be observed in the crime scene profile for a specific set of DNA profiles for the contributors, and the other describes the uncertainty about the possible DNA profiles that the unknown contributors may have. The following two sections describe these two components of the model.

2.2.1. Peak height model for fixed genotypes

We first consider the situation where we know the DNA profiles of all contributors to the mixture. This situation implies we know exactly the number of a alleles n_{ia} that contributor i possesses; where n_{ia} may be 0, 1, or 2.

In a case where the hypothesis of interest concerns unknown contributors, the analysis generally involves looking into all possible configurations of DNA profiles for the unknown contributors. A hypothesis with only known contributors also arises naturally when further conditioning on a specific configuration of profiles for the unknown contributors.

Thinking of a mixed profile in terms of a set of peaks and their corresponding heights, we use a statistical model to capture the variability in the peaks that we would see if the sample were imagined to be repeatedly re-analysed under the exact same conditions and new mixed profiles were produced.

Naturally the characteristics of a mixed profile depends on various quantities. For instance in a k-person mixture, we need to know the proportions $\phi = (\phi_1, \ldots, \phi_k)$ in which the DNA is mixed.

As is customary, our model assumes that the variability at an allelic position is independent of the variability at other allelic positions when the model parameters and genotypes are considered fully known.

Let us therefore consider a single allelic position a and consider the variability of the corresponding peak height Z_a. We use a gamma

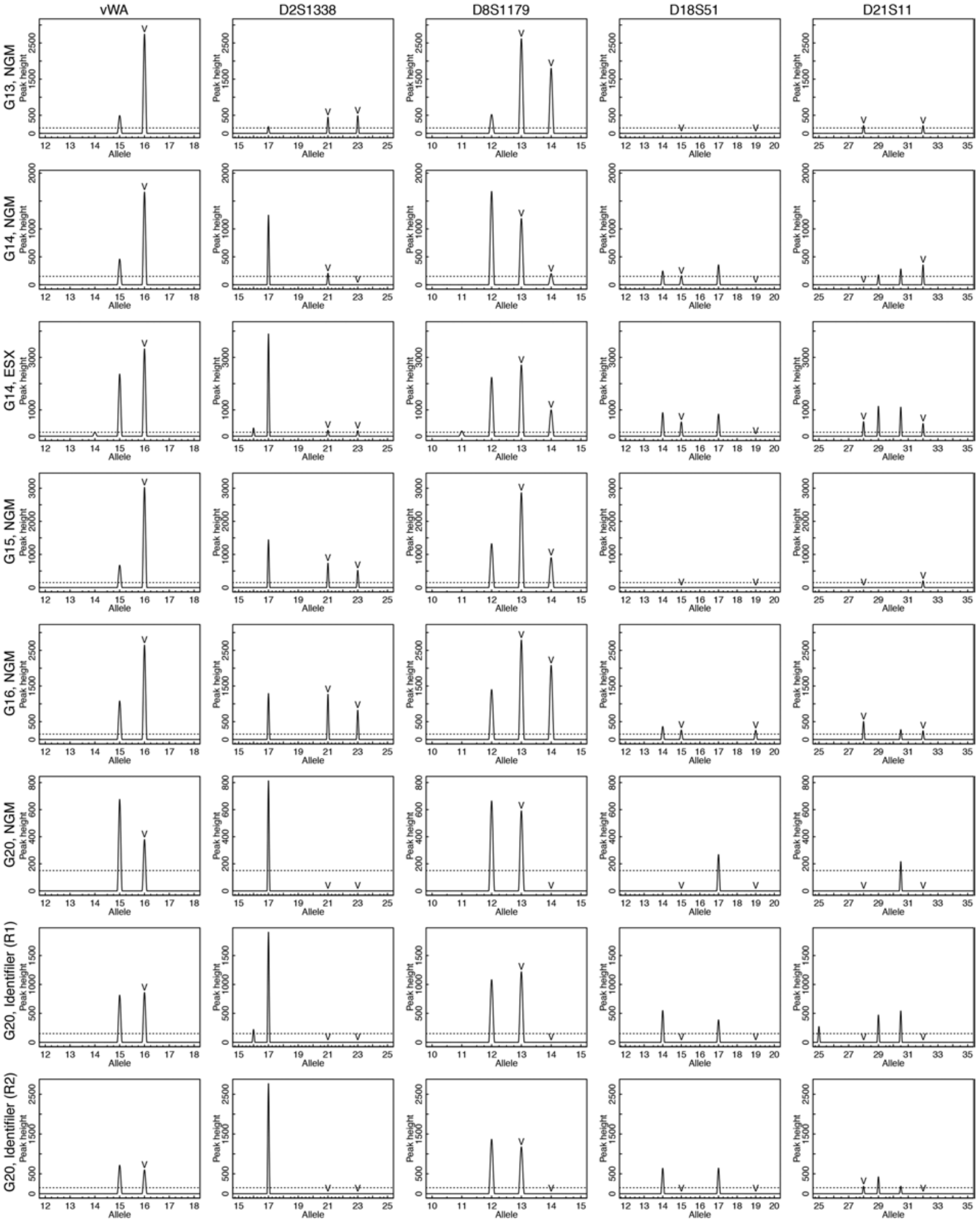

Fig. 1. Peak heights and victim's genotype across five different loci for the mixed profiles used in this paper.

distribution with shape and scale parameters given in (1) to capture the variability of the peak height

$$Z_a \sim \Gamma\left\{\frac{1}{\sigma^2}\left((1-\xi)\sum_i \phi_i n_{\mathrm{ia}} + \xi \sum_i \phi_i n_{i,a+1}\right), \mu\sigma^2\right\}. \tag{1}$$

In a sample from a single donor where no dropout or stutter has occurred, we may interpret μ directly as the mean peak height for a heterozygous allele and σ as the coefficient of variation for the peak height. For example, $\sigma = 0.67$ corresponds to the standard deviation of the peak being 67% of its mean μ. The model captures back-stutter through the parameter ξ, which determines the mean proportion of stutter that may be observed in the allelic position one repeat less. Here ξ is the ratio of the stutter peak with respect to the parent plus the stutter peak, rather than the more commonly used ratio between the stutter peak and the parent peak.

In principle the gamma distribution in (1) captures peak heights down to a level of zero RFU; however a detection threshold can be applied, so that peaks are only considered present when they exceed this threshold. Thereby allelic dropout is represented in our model as the consequence of the peak failing to appear above the applied detection threshold, C. For the analyses here, we adopt a detection threshold of 150RFU. The profiles exhibit a lot of baseline noise and this choice of threshold enables the statistical analysis to take into account the uncertainty about whether any signal below this level should be considered a peak or not.

Since we assume that peak heights are conditionally independent given the full DNA profiles of all contributors and the values of the model parameters $\psi = (\phi, \mu, \sigma, \xi)$, we have that the joint distribution (2) for the set of peak heights, $\mathbf{z}$, across all allelic positions in the EPG, for a given hypothesis H, is obtained as a product of distributions for the individual peak heights z_a.

$$\Pr(\mathbf{z}|\psi, H) = \prod_a \Pr(z_a|\psi, H). \tag{2}$$

2.2.2. Model for genotypes

In Section 2.2.1 we considered the contributors genotypes as known, but when a hypothesis involves unknown contributors, we need to take into account the uncertainty about their genotypes.

For the analysis in this paper, we assume that all unknown contributors are unrelated to each other and to any known contributors to the sample. Further, we assume that they come from the same population and that this population may be suitably characterised through the allele frequencies in the Italian Caucasian reference population. These allele frequencies are taken to be fixed and known quantities with no further uncertainty and no correction for distant relatedness.

We emphasise that not accounting for distant relatedness, sampling uncertainty of allele frequencies, or heterogeneity in the reference population, is simply a model choice made for the specific analyses here and not a limitation of our framework. For the purpose of this paper, we wish to make all analyses comparable by using the exact same assumptions about unknown contributors, and the specific choice on how to take relatedness into account is usually very context specific. For instance, in computing LRs one needs to assume relatedness to a specific person of interest, so the model for genotypes depends directly on the person under consideration.

2.2.3. Estimation of model parameters

We follow [3] in estimating the model parameters by maximum likelihood. Unless otherwise stated, we use the likelihood for the hypothesis under consideration.

The evidence E in a crime scene profile consists of the observed set of peak heights, $\mathbf{z}$, together with the genotypes of any known individuals. For a given hypothesis H, the likelihood is obtained by summing over all possible combinations of genotypes $\mathbf{n}$ with probabilities $\mathrm{P}(\mathbf{n}|H)$ associated with H:

$$\Pr(E|\psi, H) = \sum_{\mathbf{n}} \Pr(\mathbf{z}|\psi, \mathbf{n})\mathrm{P}(\mathbf{n}|H). \tag{3}$$

An inherent difficulty in analysing mixed samples is that the likelihood may well have multiple modes – this is particularly the case as the number of unknown contributors is increased. This creates an imminent risk of obtaining only a local maximum of the likelihood, and to alleviate this problem we have used several starting points for the maximisation procedure.

Any hypothesis that involves several unknown contributors to the mixture has a vast number of possible genotype configurations, rendering the computation of the sum (3) computationally demanding. However, it can be calculated efficiently using the techniques described in [11]; these computational techniques are also at the core of the DNAmixtures software.

2.3. DNAmixtures: a software for complex statistical analyses of mixtures

The DNAmixtures software[1] [10] is a statistical tool that allows the expert to perform a very detailed statistical analysis of mixed DNA profiles. It has been used for evidential calculations in court both in the UK and in Denmark. It was also used in the investigative phase of the case discussed in this paper.

Founded directly on statistical principles, the framework implemented in DNAmixtures is extremely flexible and versatile in its potential for application whilst maintaining a consistency between all parts of the analysis. This is achieved by having a fully-specified joint model for contributors genotypes and the quantities measured – the set of observed peaks and their corresponding heights. By basing all analyses on the same statistical framework, the results obtained are guaranteed to be coherent.

Another advantage of having a fully specified statistical framework for mixed profiles is that it can be used as a building block to elaborate the models so as to allow for other types of evidence. For example, in the case of a kinship analysis based on one or more mixed profiles, the evidence E includes both the peak height information from the mixed profiles and the reference profile of the potential relative.

DNAmixtures is an open-source software and is available online from R-forge as a library for the statistical software R. Extensive tutorials on how to use DNAmixtures are available in [9], in the supporting information to [3], and through the help-pages for the library. Note that the use of DNAmixtures requires an installation of the full HUGIN API.[2] An additional suite of functions, KinMix, is available online.[3] It extends the capabilities of DNAmixtures to allow for the analysis of relationships between unknown contributors to a mixture and individuals with known DNA profiles.

In this paper we use the default settings of DNAmixtures: model parameters are taken to be the same across loci, but are allowed to differ between the crime scene profiles (EPGs) included in the analysis. Unknown contributors are ordered according to decreasing estimated contributions to the first EPG. Tables 1–5 show the order in which the EPGs are included for the different combinations of EPGs discussed in this paper.

2.4. Combining evidence from multiple samples or different marker systems

As detailed in Section 2.1 there are multiple crime scene profiles available in the Yara case. These are either replicates or originate from different samples. Furthermore, the samples were also typed with different kits.

There are many reasons for combining evidence, one important one being that it strengthens the information about the profiles of any shared contributors. Our model and software readily allows a combined analysis of multiple crime scene profiles.

Combining the information in multiple profiles requires a slightly more complex analysis than that of single profiles, since it is now necessary to make assumptions about which – if any – contributors may be in common. When combining replicates it is natural to make an assumption that contributors are the same, however when combining profiles from different samples one needs to carefully consider whether there is perhaps only a partial overlap. However, once a hypothesis describing the contributors is formulated, the mathematical details in extending a peak height model from one to multiple crime scene profiles are completely straightforward.

Imagine a pool of k persons, who are all (proposed) contributors to one or more of the profiles that we wish to analyse. Our model then describes the DNA profiles of the set of k persons and the associated

[1] http://dnamixtures.r-forge.r-project.org/.

[2] www.hugin.com.

[3] https://people.maths.bris.ac.uk/~mapjg/KinMix/.

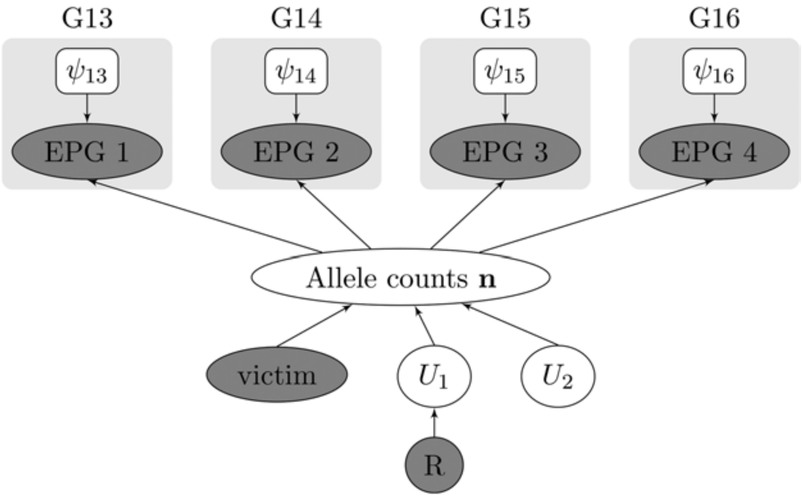

Fig. 2. Pictorial DAG representation of the combination of different types of evidence: four crime scene profiles (EPG 1–4) for samples G13-16 as well as the reference profiles of the victim and a putative parent of U_1. The hypothesis under investigation here assumes a total of at most three contributors to the four samples. Each EPG has its own set of model parameters $\psi = (\phi, \mu, \sigma, \xi)$.

peak heights across the set of profiles. Each profile (EPG) e now has an individual set of model parameters that determines the variability of the peaks that we may observe just for this profile. In particular, ϕ_{ei}, specifies the proportion of DNA that person i has contributed to profile e. This also covers the situation where a person is not a contributor to that profile; it simply corresponds to $\phi_{ei} = 0$.

We assume that, conditionally on the DNA profiles of the entire pool of contributors (and the model parameters), the peak heights in one EPG are independent of the peak heights in the other EPGs. The variability of the peak heights for each profile is then described by the gamma distribution in (1) as before. For an in-depth description of the model combining profiles, see [3]. A pictorial representation of the combination of peak height information from four crime scene profiles can be seen in Fig. 2. Ignoring the node R, the figure depicts how the peak-height information is conditionally independent across the four EPGs given the full DNA profiles of all the contributors to the mixture. Here the set of contributors is assumed to consist of the victim and two unknown contributors, U_1 and U_2.

3. Identifying unknown donors to a mixed sample

A common approach to addressing the problem of identifying unknown donors to a mixed sample, *i.e.* deconvolving the mixture, is to use the posterior distribution of the contributors' DNA profiles given the observed evidence. It is customary to assign a distribution – based on the allele frequencies in a reference population – for the DNA profiles of unknown contributors and, in the context of a Bayesian analysis, this may be seen as a prior distribution. The posterior distribution for DNA profiles is readily specified through our statistical model, since the model jointly describes both the observed profile(s) and the genotypes of unknown contributors and thus, in turn, also the conditional distribution of genotypes given the observed profile(s). Note that the analysis of the profiles – including various posterior distributions for DNA profiles – will naturally depend on the choice of a prior distribution. It is therefore important to assess the suitability of the prior distribution; we discuss possible ways of doing so in Section 6.3.

3.1. Posterior most likely DNA profiles

In order to deconvolve the mixtures, first we need to formulate a hypothesis H that suitably characterises the set of contributors and the (prior) distribution of their DNA profiles. An example would be *"H: The DNA comes from the victim and two unknown contributors, who are biologically unrelated to the victim and to each other"*. This characterises a set of three contributors, where the prior distribution for their DNA profiles assumes that the DNA profiles for the contributors are independent, the victim has a fixed known profile (*i.e.* the reference profile) and each unknown contributor has two alleles that are sampled from a multinomial distribution according to the Italian Caucasian population frequencies.

We estimate the model parameters by maximum likelihood under the selected hypothesis H and, conditional on these and the observed evidence, we can identify a list of most likely configurations of genotypes for contributors ranked by their probability. We emphasise that, given the parameter estimates, all our computed probabilities are exact unlike probabilities computed by MCMC or other types of approximative methods.

3.2. Summarize the posterior distribution

Once the model parameters have been estimated, we compute the posterior distribution of DNA profiles for some or all contributors. Conditional on the estimated parameters, the markers are independent, so the probability of any profile can be obtained as the product of probabilities of the genotypes for each marker separately.

The space of genotypes is rather large, and for informative samples most of the genotypes have a tiny probability. To give a clearer summary of the important genotypes, we report only genotypes with a probability larger than 0.001.

We also introduce a "compound allele", *other*, to denote the collection of alleles for which no peak has been observed in any of the samples analysed. If the person is predicted to have an allele *other*, it means that the person has an allele that has dropped out in all the samples. A motivation for introducing this "compound allele" is that making a prediction among alleles that are not seen in any sample under consideration can feel somewhat speculative, and one could argue that such a prediction reflects the population allele frequency, rather than the information in the observed profiles.

We note that the "compound allele" *other* is introduced only to report the results more clearly and it has no impact on the computations. It is a simple matter to obtain the probability for any genotype, should one wish to investigate further the possible configurations that contributors might have among this collection of alleles.

Another – slightly more intuitive – way of defining the "compound allele" would be by considering the set of samples on which the prediction is based and include only alleles that are not seen in any of these. However, this would not allow a direct comparison of the prediction results from different sets of samples, because the set of observed alleles would vary depending on the samples considered. As an example, imagine that allele 17 has been observed in one sample (S1, say), but not in the other (S2, say), whereas allele 18 has been seen in S2, but not in S1. Using our definition, neither allele 17 nor 18 are included in the "compound allele" *other*. Thus, the genotype (17, 18) will figure on our list explicitly regardless of whether the analysis is based on S1, S2, or both. Using the profile-specific definition, the genotype should be listed as (17, *other*) for S1, (18, *other*) for S2, and (17,18) for a combined analysis of S1 and S2.

It is generally challenging to usefully summarise the posterior probabilities and the choice of how to do it will naturally depend on the focus of the analysis. A detailed discussion on predicting profiles using the model and `DNAmixtures` may be found in [9].

4. Inference about relationships from DNA mixtures

In the Yara case it is of interest to assess whether a contributor to a mixture is a child of GG (the putative father) or of EA (the putative mother), who both have known genotypes.

Green and Mortera [13] develop methods for inference about relationships between contributors to a DNA mixture and other individuals of known genotype. The evidence for the relationship is evaluated as the likelihood ratio, LR, for H_p, which claims a specified relationship between a typed individual and an unknown contributor, versus the alternative H_0, which claims that the unknown contributor to the mixture is a random member of the population. Note that when formulating the two alternative hypotheses, H_p and H_0, we implicitly assume that the set of contributors has already been described in some detail. For instance, in the Yara case, we may base a paternity analysis on the assumption that a mixture consists of DNA from the victim and two unknown contributors. Then the likelihood ratio comparing H_p to H_0 addresses the further suggestion that GG may be the father of a specific individual among this set of contributors.

Following [13] we let R denote the genotypes of the measured individuals, GG or EA, and we let $U_i = U$ be a specified contributor to the mixture. The evidence $E = \{R, \mathbf{z}\}$ for which to evaluate the likelihood now consists of both the peak height information (quantified through $\mathbf{z}$) in the EPG(s) and the measured genotypes R. The likelihood ratio is

$$\mathrm{LR} = \frac{\mathrm{P}(E|H_p)}{\mathrm{P}(E|H_0)}, \tag{4}$$

where for the Yara case we shall always take $U = U_1$.

Fig. 2 shows how the evidence is combined under the hypothesis that a specific person whose genotype is among those measured in R is a parent of U_1. It is a pictorial DAG representation of the problem that illustrates the conditional independence assumptions among the components of the model as represented by the nodes. The first and second row of nodes represent the DNA mixture models for the samples G13-G16, whereas the last two rows of nodes denote the potential relationship between the mixture contributors and potential relatives of known genotypes R. The DAG shows, for example, that the genotypes in R are conditionally independent of the peak heights $\mathbf{z}$ in the EPGs given U_1.

In order to compute the likelihood ratio, we adopt the exact additional likelihood nodes (ALN) method described in Section 3.2 of [13]. We use the same model parameters for both hypotheses, *i.e.* the MLE found under the assumption of non paternity (non maternity). This gives a lower LR than if we were to estimate the parameters by maximising the likelihood under each hypothesis separately. Assuming that the model parameters are the same under H_p and H_0, we may rewrite the LR by

$$\mathrm{LR} = \sum_{\mathrm{Ugt}} \mathrm{LR}_{\mathrm{Ugt}} \times \mathrm{P}(\mathrm{Ugt}|\mathbf{z}), \tag{5}$$

where $\mathrm{P}(Ugt|\mathbf{z})$ is the posterior probability of U_1 having genotype Ugt after taking into account the peak height information, and $\mathrm{LR}_{Ugt} = \mathrm{P}(R|Ugt, H_p)/\mathrm{P}(R|H_0)$ is the likelihood ratio conditional on U_1 having genotype Ugt. For a specific parent genotype in R, pgt, this conditional likelihood ratio is given as

$$\begin{aligned}\mathrm{LR}_{\mathrm{Ugt}} &= \mathrm{P}(\mathrm{pgt}|\mathrm{Ugt}, H_\mathrm{p})/\mathrm{P}(\mathrm{pgt}|H_0)\\ &= \begin{cases} n_{\mathrm{ia}}/2q_a & \text{if pgt} = \{a, a\}\\ n_{\mathrm{ia}}/4q_a + n_{\mathrm{ib}}/4q_b & \text{if pgt} = \{a, b\}, \quad a \neq b, \end{cases}\end{aligned} \tag{6}$$

where q_a and q_b are the allele frequencies and n_{ia} and n_{ib} the allele counts for Ugt.

The ALN method of [13] computes the LR by introducing an additional likelihood node, based on the relationship under question, into the Bayesian network created by `DNAmixtures`. The calculations are performed in `R` using the open-source suite of functions `KinMix`,[4] which extends the functionality of `DNAmixtures`.

5. Results

In this section we show the results on the parameter estimation, the deconvolution and the relationship identification from a DNA mixture on the case presented in Section 1.1. All samples are typed with NGM unless otherwise stated and the Italian allele frequencies database is used.

5.1. Estimated model parameters

5.1.1. Single profiles

Analysing each profile separately, we firstly posit that each corresponding mixture is composed of the victim and at most 3 unknown contributors U_1, U_2 and U_3. Table 1 shows the estimated parameters for this scenario. Each profile may, in fact, be explained by the victim and just a single unknown, and from Table 1 it is quite apparent that contributors U_2 and U_3 have very minor contributions for all of the samples. Table 2 shows the estimated parameters after reducing the number of contributors to two per sample. The reduction in the number of contributors is supported by the fact that the maximised likelihood (not shown) of the observed profile is left practically unchanged. Further, model checking methods [11] confirm that two contributors can adequately explain the profiles; Examples of some of the graphical assessments carried out are given in Section 6.3 (Figs. 3 and 4).

For sample G13, there is a well-determined contribution of 85% from the victim, $\phi_v = 0.85$. The contributions of the unknown donors simply split the proportion of DNA not accounted for by the victim – thus in Table 1 we see U_1, U_2 and U_3 accounting for roughly the same percent of the DNA as just U_1 in Table 2 .

This phenomenon of equal estimated proportions for unknown contributors is encountered particularly often in models with several very minor unknown donors. We may intuitively think of it as the data not carrying enough information for the analysis to pick up a difference in the proportion of DNA from each unknown donor. Mathematically, it is a natural consequence of the symmetrical nature of the role that unknown contributors play in computing the likelihood and the local or global modes arising because of these symmetries. We refer to [9] for a detailed description of the phenomenon, which will occur for any probabilistic genotyping model. To make sure that a maximisation of the likelihood resulting in such equal contributions from unknown contributors is not due to encountering only a local mode of the likelihood function, we have used multiple starting points for all of the maximisations. To explore only the relevant parts of the parameter space for mixing proportions, we have used starting points where the victim is assigned the larger proportion.

[4] Available at https://people.maths.bris.ac.uk/~mapjg/KinMix/.

Table 1
Estimated parameters based on an analysis of the individual samples assuming that each sample contains DNA from the victim and three unknown contributors. Note that the labels U_1, U_2, and U_3 may not refer to the same three individuals across samples.

Sample	μ	σ	ξ	ϕ_v	ϕ_{U_1}	ϕ_{U_2}	ϕ_{U_3}
G13	1028	0.761	0.119	0.833	0.056	0.056	0.056
G14	966	0.628	0.140	0.424	0.576	0.000	0.000
G15	1371	0.736	0.118	0.722	0.278	0.000	0.000
G16	1931	0.554	0.078	0.738	0.262	0.000	0.000
G20	561	0.639	0.000	0.000	1.000	0.000	0.000

Table 2
Estimated parameters based on an analysis of individual samples assuming that each sample contains DNA from the victim and one unknown contributor U_1.

Sample	μ	σ	ξ	ϕ_v	ϕ_{U_1}
G13	1025	0.763	0.126	0.850	0.150
G14	966	0.628	0.140	0.424	0.576
G15	1371	0.736	0.118	0.722	0.278
G16	1931	0.554	0.078	0.738	0.262
G20	561	0.639	0.000	0.000	1.000

Table 3
Estimated parameters based on a joint analysis of samples G13-16 using NGM and assuming that each of the samples contains DNA from one or more individuals among the victim and three unknown donors.

Sample	μ	σ	ξ	ϕ_v	ϕ_{U_1}	ϕ_{U_2}	ϕ_{U_3}
G13	1022	0.729	0.079	0.752	0.248	0.000	0.000
G14	956	0.531	0.080	0.324	0.676	0.000	0.000
G15	1367	0.631	0.067	0.571	0.429	0.000	0.000
G16	1931	0.471	0.047	0.619	0.357	0.012	0.012

Table 4
Estimated parameters for sample G14 based on NGM and ESX16

Sample	μ	σ	ξ	ϕ_v	ϕ_{U_1}	ϕ_{U_2}	ϕ_{U_3}
NGM	956	0.518	0.082	0.319	0.681	0.000	0.000
ESX	3651	0.430	0.083	0.243	0.757	0.000	0.000

Table 5
Estimated parameters for G20 based on NGM and replicates R1 and R2 produced with Identifiler. The estimates are based only on loci included in the NGM system.

Sample	μ	σ	ξ	ϕ_v	ϕ_{U_1}	ϕ_{U_2}	ϕ_{U_3}
G20, NGM	559	0.609	0.000	0.000	1.000	0.000	0.000
G20, Identifiler (R1)	1157	0.461	0.060	0.000	0.929	0.012	0.059
G20, Identifiler (R2)	1241	0.507	0.075	0.000	0.954	0.046	0.000

5.1.2. *Combined analysis of profiles*

We now turn to analyses that combine multiple mixed profiles. From the analyses of each profile separately, we have already seen that in all samples there is at least one unknown contributor in addition to the victim. The unknown contributor may or may not be common across different samples, but as the mixed profiles do appear highly similar (see Fig. 1), it is natural to investigate the possibility of shared unknown contributors to each sample.

We consider a total of at most three unknown contributors in addition to the victim, which allows the possibility that some samples do not share the same contributor. As is evident from Table 3 the estimated parameters do indeed point to a shared contributor. We note that further investigations into this could be done through statistical methods as in [3,9], if desired, however we shall not pursue this as it is not central to our analysis.

The estimated μ, σ and ξ parameters in Table 3 are similar to those in Table 1 where estimates were made separately on each sample. However, the estimated proportions of DNA, ϕ_v and ϕ_{U_1}, are noticeably different. Again, the unknown contributors U_2 and U_3 have very minor contributions. However, the likelihood does increase by a factor around 300 when removing the two unknown contributors, so their contributions are not entirely redundant. In Table 3, based on a combination of samples, the three unknown contributors no longer have equal proportions of DNA for sample G13 as was the case in Table 1 . This is not entirely unexpected since the other samples now lend information about the identity of the profiles of unknown contributors to sample G13. In contrast to the single-sample analysis in Table 1, sample G16 now appears to have very small contributions from U_2 and U_3. Of course, the parameter estimates are associated with some uncertainty and any in-depth interpretation of the estimates should take this into consideration. For instance, the estimated contributions of DNA may appear very different, however if their uncertainty is large then the difference may not be statistically significant.

As our discussion revolves around identifying the (generally high-level) contributor U_1, it is not of great importance whether low-level peaks or small peak imbalances in the profile are explained by stutter (via a higher ξ), higher variability (via a higher σ), or minor contributors such as U_2 or U_3 (via higher ϕ_{U_2} and ϕ_{U_3}). Generally we see very similar parameter estimates across all the analyses, which indicates little sensitivity to the specific hypothesis under consideration.

5.1.3. *Replicates from different marker systems*

Table 4 shows estimated model parameters for G14 based on both NGM and ESX. Compared to the estimates for G14 in Table 1 based on NGM only, we see that U_1's estimated proportion ϕ_{U_1} increases from around 58% to about 68%. The parameter μ is about four times higher for ESX, reflecting that the average peak height in the ESX profile is about four times higher. As expected from the replicates, the two profiles have a very similar composition with U_2 and U_3 having only tiny contributions.

Table 5 gives the G20 estimated model parameters for NGM and two replicates R1 and R2 using Identifiler. For Identifiler we have included only the ten loci (plus Amelogenin) that are also in the NGM profile, though the loss of information about common parameters across loci will result in slightly less precise estimates. We do this mainly for simplicity, since the victim's profile is only available for the NGM loci. However, the framework of `DNAmixtures` does allow the use of all available information in profiles from kits with partially overlapping sets of loci. As in Table 1 the estimates point to U_1 as a clear major contributor and no contribution from the victim.

5.2. *Deconvolution: identification of U_1*

The analyses in Section 5.1 point to a common (male) donor for many of the samples and, further, indicate that this person provides the largest proportion of DNA among the unknown contributors. In this section we investigate possible profiles of the unknown contributor U_1. For all sample combinations, we take the set of contributors to the profiles to consist of the victim, an unknown person of interest U_1, and a further two unknown contributors. We use the parameter estimates obtained in Section 5.1.

Note that the person referred to as U_1 could, in principle, be a different person when a different set of samples is analysed. However, it is unlikely to be the case here, because our analysis points to the profiles sharing an unknown contributor and further indicates that this person has the highest proportion of DNA among the unknown contributors.

5.2.1. *Most likely genotypes*

Tables 6–10 show the posterior probabilities for the genotype of U_1 for a selection of markers. We remind the reader that the tables only indicate genotype probabilities of at least 0.001, meaning that a blank cell represents a probability of less than 0.001. The genotypes in each table are ordered according to decreasing probabilities from left to right.

Table 6
Posterior probabilities for the genotype of U_1 at locus vWA. Any allele not in {14, 15, 16} is denoted by *other*.

Allele 1	Allele 2	G20 (NGM, Identifiler x 2)	G14 (NGM and ESX)	G13-16	G20	G14
15	16	1.000	0.999	0.979	1.000	0.335
16	16			0.017		0.351
15	15			0.004		0.016
16	*other*					0.185
15	*other*					0.067
14	16					0.023
14	15					0.007
other	*other*					0.006
14	*other*					0.001

Table 7
Posterior probabilities for the genotype of U_1 at locus D2S1338. Any allele not in {16, 17, 21, 23} is denoted by *other*.

Allele 1	Allele 2	G20 (NGM, Identifiler x 2)	G14 (NGM and ESX)	G13-16	G20	G14
17	17	1.000	1.000	0.996	0.665	0.582
17	21			0.003	0.010	0.026
17	*other*				0.261	0.339
17	23				0.045	0.030
16	17				0.019	0.020

Table 8
Posterior probabilities for the genotype of U_1 at locus D8S1179. Any allele not in {11, 12, 13, 14} is denoted by *other*.

Allele 1	Allele 2	G20 (NGM, Identifiler x 2)	G14 (NGM and ESX)	G13-16	G20	G14
12	13	1.000	0.996	0.973	1.000	0.636
12	12		0.004	0.023		0.094
12	14			0.003		0.052
13	13					0.140
13	14					0.029
12	*other*					0.021
13	*other*					0.015
11	12					0.004
11	13					0.003

Table 9
Posterior probabilities for the genotype of U_1 at locus D18S51. Any allele not in {14, 15, 17, 19} is denoted by *other*.

Allele 1	Allele 2	G20 (NGM, Identifiler x 2)	G14 (NGM and ESX)	G13-16	G20	G14
14	17	1.000	0.999	0.994	0.170	0.687
15	17		0.001		0.138	0.118
14	*other*			0.005		0.022
17	*other*				0.566	0.123
17	17				0.080	0.038
17	19				0.045	0.005
15	*other*					0.004
other	*other*					0.001

Table 10
Posterior probabilities for the genotype of U_1 at locus D21S11. Any allele not in {25, 28, 29, 30.2, 32} is denoted by *other*.

Allele 1	Allele 2	G20 (NGM, Identifiler x 2)	G14 (NGM and ESX)	G13-16	G20	G14
29	30.2	1.000	1.000	0.936	0.196	0.710
30.2	*other*			0.063	0.611	0.119
28	30.2				0.160	
30.2	32				0.014	
30.2	30.2				0.012	
25	30.2				0.007	
29	*other*					0.147
other	*other*					0.025

5.2.2. *Single vs. multiple replicates*

It is evident from Tables 6–10 that the combination of NGM and ESX for G14 is more informative than using a single profile for G14.

Although G20 can be viewed as a single source sample, and thus simpler to analyse, the poor quality of the sample means that U_1 is still not fully identified when analysing just a single profile. Consider for example Table 7 for marker D2S1338. When analysing the combination of three profiles for G20 – an NGM profile and two Identifiler replicates R1 and R2 – it is fully determined that U_1 is homozygous (17,17) whereas the probability of this scenario drops to about 67% using a single profile. The second most likely scenario here is that U_1 is heterozygous with a dropped-out allele. All five analyses agree that the donor has at least one allele 17.

5.2.3. *Combination of samples*

The analysis based on a combination of the four samples G13-16 is very informative having the highest ranking genotype for U_1 on all markers with posterior probability greater than 0.94. For locus D2S1338 (Table 7) we see again that there is certainty about one allele (17) and only slight uncertainty about the other allele. Recall from Table 3 that the donor was estimated to account for 67.6% of the DNA in sample G14 and only between 24.8 and 42.9% for the three other samples. This illustrates that making a joint analysis of several samples can help increase the information about non-major donors.

5.2.4. *Summary of most likely DNA profiles*

Table 11 gives the posterior probabilities for the three most likely DNA profiles across all loci combined for five different sample choices. Note that these three profiles may not be the same for all five analyses.

Moving from left to right along the columns of Table 11, we clearly see that the distribution changes from highly concentrated to highly dispersed, meaning that we experience a loss of information about the donor. For the analysis based on a single EPG, for sample G14, we see that there is almost no information about the full DNA profile of the donor; indeed the three most likely DNA profiles have vanishingly small probabilities. We emphasise that a very flat distribution of full DNA

Table 11
Posterior probabilities for the three most likely profiles for each of five choices of samples to include in the analysis. Moving from left to right along the columns, we see clearly that the distribution changes from highly concentrated to highly dispersed.

Rank	G20 (NGM, Identifiler x 2)	G14 (NGM and ESX)	G13-16	G20	G14
1	0.999985	0.724825	0.408036	0.020769	0.000013
2	0.000010	0.173731	0.112242	0.017225	0.000012
3	0.000002	0.023686	0.087681	0.016881	0.000012
Total	0.999998	0.922243	0.607959	0.054875	0.000036

profiles is often caused by a few extremely uninformative loci, and in such cases it is still possible to obtain a well-determined partial profile.

In the context of the Yara case, the most likely genotype for each of the loci, shown in the first row of Tables 6–10, coincides with the genotype of the suspect MGB, given in the second last column of Table 14. Looking across all loci for the most probable DNA profile, we found that the less informative analyses, *i.e.* those based on a single replicate (G14 or G20), do not yield MGB's profile as the most likely. In contrast to this, the analyses based on combinations of replicate profiles or multiple samples all suggested MGB as the most likely donor.

5.3. Connecting posterior most likely profiles and evidential calculations

Section 5.2 presents an analysis of the posterior distribution of profiles for U_1 based on various combinations of evidence and indicated which DNA profile U_1 might have. All computations are based on the hypothesis *H: The sample(s) contains DNA from the victim and three unknown contributors* and parameters estimated under *H*.

Now consider the weight of evidence against some person of interest, *K*. This could, for instance, be evaluated by using the hypothesis for the deconvolution as a defence hypothesis, H_d: *The sample(s) contains DNA from the victim and three unknown contributors*, and then forming a prosecution hypothesis by replacing one of the unknown contributors with *K*'s profile, H_p: *The sample(s) contains DNA from the victim, K, and two unknown contributors*. The weight of evidence WoE is then defined as $\log_{10}$LR, where LR $= \mathrm{P}(E|H_p, \psi)/\mathrm{P}(E|H_d, \psi)$. Parameters ψ are estimated under H_d and the same parameter estimates are used for H_p, setting $\phi_{\mathrm{MGB}} = \phi_{U_1}$. This ensures that $\mathrm{P}(E|H_p) = \mathrm{P}(E|U_1 = K, H_d)$.

Cowell et al. [3] introduced the concept of a loss in evidential value for quantifying the difference in $\log_{10}$LR that results from basing an evidential calculation on the mixture rather than on a single-source profile with no uncertainty about the alleles of the contributor.

Following [3], we define the maximal weight of evidence, maxWoE, as the reciprocal of the match probability, $1/\pi_K$ on a $\log_{10}$. scale, $-\log_{10}\pi_K$, with reference to the Italian allele frequency database. For any person of interest, K, we can compute their evidential loss, WL (*K*), as

$$\mathrm{WL}(K) = \mathrm{maxWoE}(K) - \mathrm{WoE}(K) = \mathrm{maxWoE}(K) - \log_{10}\frac{\mathrm{P}(E|U_1 = K, H_d)}{\mathrm{P}(E|H_d)}. \tag{7}$$

Cowell et al. [3] (see Eq. (2) on page 9) further introduced the concept of a *generic loss*. The loss is generic in that it quantifies evidential loss WL(K^*) for the evidence against the posterior most likely person K^*; the loss of evidential value for any other specific profile, such as that of MGB, will be greater. Thus, it may be thought of as a measure of the information in the mixture(s) about contributor U_1.

To see that the generic loss is the *minimal loss*, note that the evidential loss against a person *K* is exactly the reciprocal of the posterior profile probability ($\log_{10}$ scale) from the deconvolution:

$$\begin{aligned}
\mathrm{WL}(K) &= -\log_{10}\mathrm{P}(U_1 = K|H_d) - \log_{10}\left\{\frac{\mathrm{P}(E \mid H_p)}{\mathrm{P}(E \mid H_d)}\right\}\\
&= -\log_{10}\left\{\frac{\mathrm{P}(E \mid H_p)\mathrm{P}(U_1 = K \mid H_d)}{\mathrm{P}(E \mid H_d)}\right\}\\
&= -\log_{10}\left\{\frac{\mathrm{P}(E \mid U_1 = K, H_d)\mathrm{P}(U_1 = K \mid H_d)}{\mathrm{P}(E \mid H_d)}\right\}\\
&= -\log_{10}\left\{\frac{\mathrm{P}(E, U_1 = K \mid H_d)}{\mathrm{P}(E \mid H_d)}\right\}\\
&= -\log_{10}\mathrm{P}(U_1 = K|E, H_d)
\end{aligned} \tag{8}$$

Thus, the posterior most likely profile will always have the smallest possible evidential loss.

Table 13
Loss of evidential value in computing the weight of evidence against MGB.

	G20 (NGM, Ident.)	G14 (NGM, ESX)	G13-16	G20	G14
Posterior probability	0.999985	0.724825	0.408036	0.016881	0.000003
Maximal WoE	20.0	20.0	20.0	20.0	20.0
WoE	20.0	19.8	19.6	18.2	14.4
Evidential loss	0.0	0.1	0.4	1.8	5.6

Table 12 shows the generic loss for the five different analyses carried out in Section 5.2. It is clear that there is an increasing loss of information as we move from left to right; from the analysis of three profiles from the essentially single-source sample G20 to a single profile from sample G14. As the posterior most likely profile based on either of the three combinations of evidence is actually the profile of MGB, we may in these cases further conclude that a LR evaluating the evidence for the presence of DNA from MGB in sample G20 based on three replicates will be virtually maximal.

As we now have MGB's profile, we may wish to compute the evidence that he is a contributor to a particular mixture or set of mixtures. The WoE against MGB compares the prosecution proposition H_p: *The sample(s) contains DNA from MGB, the victim and two unknown contributors* to the defense proposition H_d: *The sample(s) contains DNA from the victim and three unknown contributors*. The results are given in Table 13 . As expected, we see that the loss in weight of evidence for MGB based on the three first combinations of evidence corresponds to the generic loss in Table 12 . Further, the evidential loss we get when computing the WoE based on a single profile for either G20 or G14 is indeed larger than the generic loss in Table 12, since for these profiles MGB is not the most likely contributor.

5.4. Evidential value for familial relationships

The analyses of Section 5.2 aim to establish the DNA profile of the individual U_1; however they do not provide information as to who this specific person might be. Following the events of the Yara case described in Section 1.1, the analyses of the present section addresses the question of connecting the profile of U_1 to other persons pertaining to the investigation through establishing familial relationships. Table 14 shows, for a subset of markers, the DNA profiles of the deceased GG, the

Table 12
Generic loss of evidential value in computing the weight of evidence against a person that has the posterior most likely profile for U_1.

	G20 (NGM, Ident.)	G14 (NGM, ESX)	G13-16	G20	G14
Posterior probability	0.999985	0.724825	0.408036	0.020769	0.000013
Generic loss	0.0	0.1	0.4	1.7	4.9

Table 14
Extract of GG, EA, MGB and the victim's genotypes for a subset of markers.

Marker	GG	EA	MGB	Victim
D2S1338	17 24	17 20	17 17	21 23
D8S1179	12 12	13 13	12 13	13 14
D18S51	12 14	17 19	14 17	15 19
D21S11	24.2 30.2	28 29	29 30.2	28 32
⋮	⋮	⋮	⋮	⋮
vWA	15 16	16 18	15 16	16 16

Table 15
The LR, $\log_{10}$LR, and posterior probability $P(H_p|E)$ for paternity, based on single profiles and joint analyses of G13-16 and G14 (NGM/ESX).

EPG	LR	$\log_{10}$LR	$P(H_p\|E)$
G13	5.26	0.72	0.84014
G14	7,821	3.89	0.99987
G15	18	1.26	0.94813
G16	1,087	3.04	0.99908
G20	90,196	4.96	0.99999
G13-16	226,218	5.35	1.00000
G14 (NGM/ESX)	269,407	5.43	1.00000

local woman EA, the suspect MGB and the victim.

All the analyses shown here are based on assuming that the victim and three unknowns contributed to the mixtures. The data from the Y-filer markers (not shown here) indicated that the unknown contributor is male and the deconvolution analysis of Section 5.2 confirms that U_1 is indeed a male contributor. We will thus condition the analysis on the evidence that U_1 is male. This additional information has, however, no effect on the parameter estimates which are as in Tables 1, 3 and 4 .

5.4.1. Evidential value for U_1 being the child of GG

Having identified the deceased GG as the potential father of *Ignoto 1* (in our analysis taken to be the individual U_1), it is natural to quantify the evidence for this hypothesis through a likelihood ratio. Formally, we wish to compute the likelihood ratio for individual U_1 being the child of GG, *i.e.* comparing the hypothesis H_p: *U_1 is the child of GG* to the alternative hypothesis H_0: *U_1 is unrelated to GG*.

Table 15 shows the results for the analyses based on single profiles as well as those for joint analyses based on samples G13-16 and G14 (NGM/ESX), respectively. The results are reported in three ways: the LR, the corresponding $\log_{10}$LR, and the posterior probability of paternity $P(H_p|E)$ based on assuming uniform prior probabilities, $P(H_p) = P(H_0) = 0.5$ [8].

All LRs point to paternity. The samples G13 and G15 give weak evidence in favour of paternity, as the proportion contributed by U_1 is around 6% for G13 and 28% for G15 (see Table 1). Sample G20, where the victim's DNA is not present, gives a high $\log_{10}\text{LR} = 4.96$, as would be expected, yet the joint G13-16 and G14 (NGM/ESX) analyses yield a LR almost 3 times greater, $\log_{10}$LR of 5.35 and 5.43, respectively.

5.4.2. Evidential value for U_1 being the child of EA

Based on a working hypothesis that *Ignoto 1* might be an illegitimate son of the deceased GG, a further investigation pointed to EA as a potential mother of *Ignoto 1*. Analogously to the paternity tests of Section 5.4.1 we may formally assess the evidence for maternity through a likelihood ratio comparing hypotheses H_m: *U_1 is the child of EA* and H_0: *U_1 is unrelated to EA*.

Table 16
The LR, log $_{10}$LR, and posterior probability $P(H_m|E)$ for maternity, based on single profiles and joint analyses of G13-16 and G14 (NGM/ESX).

EPG	LR	$\log_{10}$LR	$P(H_m\|E)$
G13	1.43	0.16	0.58854
G14	112	2.05	0.99114
G15	85	1.93	0.98832
G16	2.44	0.39	0.70896
G20	605	2.78	0.99835
G13-16	1,407	3.15	0.99929
G14 (NGM/ESX)	1,812	3.26	0.99945

Table 17
Standard tests for paternity and maternity.

Test	LR	$\log_{10}$LR	$P(H_m\|E)$
Paternity	325,522	5.51	0.999997
Maternity	1,626	3.21	0.999385

In Table 16 all LRs point to maternity, although the evidential value for maternity is substantially weaker than that in Table 15 for paternity. Samples G13 and G16 give weak evidence in favour of maternity. Recall from Table 1 that the proportions contributed by U_1 are around 6% and 26% for G13 and G16, respectively. Sample G20, where the victim's DNA is not present, gives quite a high LR, but again the analysis for the joint G13-16 and G14 (NGM/ESX) give a much higher weight of evidence in favour of maternity.

5.4.3. Evidence for a familial relationship between known profiles

The tests for paternity (maternity) carried out so far are based on incomplete information about the DNA profile of the alleged child as obtained through one or more DNA mixtures. However, in the Yara case the investigation eventually lead to MGB being identified as a specific person of interest, which enables a simple paternity (maternity) test to be performed.

Table 17 shows the paternity (maternity) test obtained by using MGB's full DNA profile. The pair of propositions considered for the paternity test is H_p: *MGB is the child of GG* and H_0: *MGB is unrelated to GG*. Similarly, we test the maternity propositions H_m: *MGB is the child of EA* with H_0: *MGB is unrelated to EA*. These classical tests correspond to computing the LR in (5) where Ugt is substituted by the known genotype of the alleged child, expressing that there is no uncertainty about the DNA profile of the alleged child. Note that since other profiles, rather than that of MGB, may make the alleged relationship more likely, an analysis based on mixtures where there is some uncertainty about the DNA profile of the contributor to the mixture may result in a higher LR than the test based on no uncertainty about the DNA profiles. The results for classical parentage testing, *i.e.* H_p: *MGB is the child of GG and EA* versus H_0: *MGB is unrelated to both GG and EA*, yield a $\text{LR} = 2.87 \times 10^{11}$ and $\log_{10}\text{LR} = 11.45$ so the posterior probability of parentage is basically $P(H_p|E) = 1$.

6. Discussion

6.1. Quantifying the weight of evidence against a person

One should bear in mind that the generic evidential loss of Section 5.3 is computed relative to a particular choice of hypotheses and parameters. Of course, for a specific evidential evaluation against a person, we may choose to compute a slightly different LR. For instance, we may opt for using different parameters for the two hypotheses and also standard adjustments based on the particular person of interest. In this case, the generic loss no longer constitutes the smallest evidential loss that we could get.

For LRs based on a combined analysis of replicates (where the DNA has the same origin) and for the LRs based on single profiles it is uncontroversial to use a pair of hypotheses where, in the defense hypothesis, H_d, we substitute an unknown contributor for MGB. However, before putting forward an LR based on a combination of profiles from different samples, it is imperative to carefully consider whether the additional unknown contributor in H_d is thought to be the same or a different individual in each sample – the LR and corresponding maximal weight of evidence can be *very* different depending on the choice of H_d. However, our analysis in Section 5.1 indicated a common contributor U_1 for all samples, so here it is natural to substitute a single common unknown contributor for MGB.

6.2. Quantifying uncertainty

For all analyses, we have used the maximum likelihood estimates for model parameters without further incorporating the uncertainty in these estimates.

If there is little information about the parameters in the mixed profiles, the likelihood function is quite flat and this results in a greater variance of the maximum likelihood estimator. However, the value of the likelihood function (as used for the numerator and the denominator in a LR) remains roughly the same across the supported range of parameters. This means that generally the LR will not be affected much by taking into account the uncertainty about estimated parameters.

However, for a deconvolution analysis where parameters such as the mixture proportions play a larger role, we may well see that conclusions prove more sensitive to the specific choice of parameters. A further investigation into this would be very interesting and could be partially done in a simple fashion through a sensitivity analysis based on the standard error for the estimated parameters.

It would be useful to also incorporate uncertainty in allele frequencies, kinship corrections based on the possibility of alleles being identical by descent and population heterogeneity as in [12]. This however, is beyond the scope of the present paper.

6.3. Does the model capture the evidence?

Model checking methods suggest that the model captures well the the pattern of peaks for each profile, as described by the set of alleles observed in the profile. The set of model checking methods applied here to the statistical model described in Section 2.2 are developed in [11] and are available in `DNAmixtures`. Two model checking examples for samples G13-G16 of the Yara case are shown in Figs. 3 and 4. The two figures each address different aspects of the modelling and jointly enable a thorough assessment of the joint analysis.

From the probability plot in Fig. 3, we see that the model adequately captures the overall variability of peak heights above the detection threshold of 150RFU. When the model fits well, the points should resemble the diagonal line as they do here. Fig. 4 shows a prequential monitor plot [17], which assesses the ability of the model to predict whether or not a peak is observed at the next allelic position in the EPG. When the model fits well, the monitor should stay below the two upper prediction limits. We see that the model captures these aspects of the data very well.

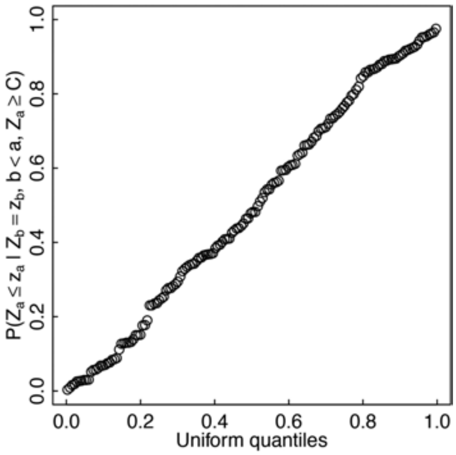

Fig. 3. Probability plot assessing the variability of peak heights for all peaks above the detection threshold of 150RFU. When the model fits well, the points should follow the diagonal as they do here.

Both figures are based on the prequential framework of [6] that, in our context, describes the distribution of a single peak conditionally on the observed peaks up to its position in the EPG. An important reason to base Fig. 3 on the prequential framework is that the transformed peak heights are then not only uniformly distributed, but indeed also statistically independent. The prequential distribution for the peak height at a specific allelic position depends on the ordering in which the allelic positions are considered (both within each EPG and across all EPGs included in the analysis). This may be chosen arbitrarily, though it will affect the trajectory of the monitor in Fig. 4. However, as the height of the monitor at the last point will always be the same, the overall conclusion that the model fits well is not dependent on the chosen order.

Through the graphical inspections carried out, we can also confirm that the samples are generally well described by the proposed set of contributors using the Italian allele frequencies, whether this be with the hypothesis *H: victim and 1 unknown contributor* or *H: victim and 3 unknown contributors*. The prior distribution for unknown genotypes is an important component of the analysis, and Fig. 4 directly enables an assessment of the suitability of this distribution.

Graphical inspections at a more detailed level (not shown here) suggest that the profiles exhibit more degradation than is explicitly modelled by the default settings of `DNAmixtures`. However, we note that this is automatically compensated for through a higher estimated variability of the peak heights (as quantified by parameter σ). Note also that graphical inspections such as that shown in Fig. 4 indicate that the model does indeed suitably capture the pattern of presence and absence of peaks in the profiles, so we have no reason to believe that the risk of any non-captured degradation will give misleading conclusions. A standard example of a consequence of degradation is when a heterozygote contributor appears to be homozygote with one of the two alleles having dropped out. By adapting the model to explicitly allow for degradation, we would expect the mixtures to be more informative about the profile of U_1.

An important feature of combining profiles from different kits is that we instantly improve the robustness of analysis towards any complications relating to degradation. This is achieved because the fragment lengths of alleles and the dye used for the PCR differs between the kits and so the informative (good-quality) markers in one sample can lend information to markers with less information in the other sample.

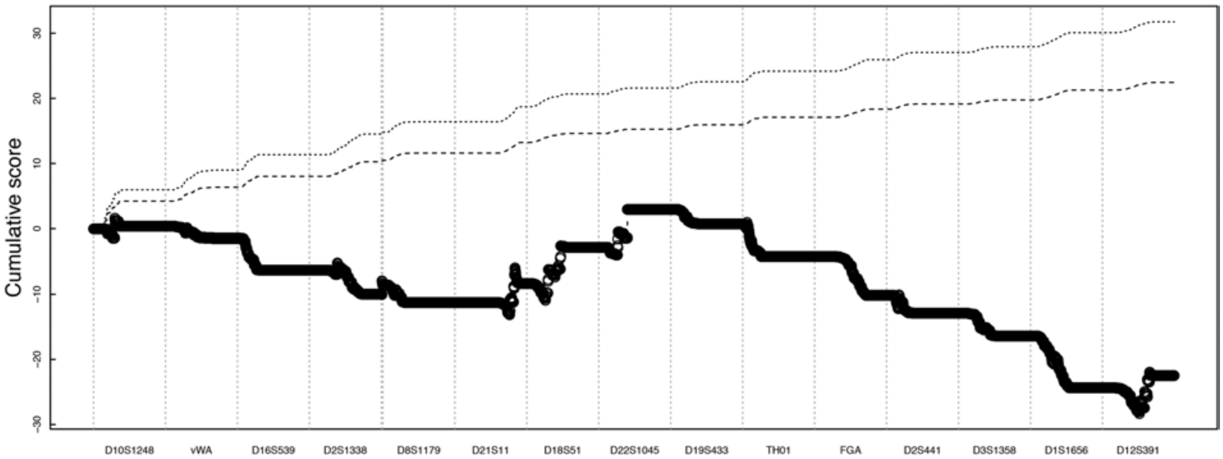

Fig. 4. Prequential monitor plot confirming that the model is able to predict at which allelic positions a peak is observed. When the model fits well, the monitor should stay below the two upper prediction limits (dashed 95%, dotted 99%).

Acknowledgements

The authors would like to thank the Isaac Newton Institute for Mathematical Sciences, Cambridge, for support and hospitality during the programme *Probability and Statistics in Forensic Science* which was supported by EPSRC grant number EP/K032208/1. Graversen is funded by the Danish Council for Independent Research under grant number DFF-6111-00557. We would like to thank Helle Smidt Mogensen for her help in exporting the profiles. We would also like to thank the staff of the Biology Section of RIS Parma in particular Nicola Staiti and Fabiano Gentile.

References

[1] C.C. Benschop, S.Y. Yoo, T. Sijen, Split DNA over replicates or perform one amplification? Forensic Sci. Int. Genet. Suppl. Ser. 5 (2015) e532–e533.

[2] Y. Chung, Y.Q. Hu, W. Fung, Familial database search on two-person mixture, Comput. Stat. Data Anal. 54 (2010) 2046–2051.

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

[4] R.G. Cowell, S.L. Lauritzen, J. Mortera, A gamma model for DNA mixture analyses, Bayesian Anal. 2 (2) (2007) 333–348.

[5] R.G. Cowell, S.L. Lauritzen, J. Mortera, Probabilistic expert systems for handling artefacts in complex DNA mixtures, Forensic Sci. Int. Genet. 5 (2011) 202–209.

[6] A.P. Dawid, Present position and potential developments: some personal views. Statistical theory. The prequential approach (with discussion), J. R. Stat. Soc. Ser. A 147 (1984) 278–292.

[7] G. Dørum, N. Kaur, M. Gysi, Pedigree-based relationship inference from complex DNA mixtures, Int. J. Legal Med. 131 (2017) 629–641, https://doi.org/10.1007/s00414-016-1526-x.

[8] E. Essen-Möller, Die Beweiskraft der Ähnlichkeit im Vaterschaftsnachweis. Theoretische Grundlagen, Mitteilungen der Anthropologischen Gesellschaft 68 (1938) 9–53.

[9] T. Graversen, Statistical and Computational Methodology for the Analysis of Forensic DNA Mixtures with Artefacts, DPhil, University of Oxford, 2014, https://ora.ox.ac.uk/objects/ora:9362.

[10] T. Graversen, DNAmixtures: Statistical Inference for Mixed Traces of DNA. R package version 0.1-4, (2015) http://dnamixtures.r-forge.r-project.org.

[11] T. Graversen, S. Lauritzen, Computational aspects of DNA mixture analysis, Stat. Comput. 25 (2015) 527–541.

[12] P.J. Green, J. Mortera, Sensitivity of inferences in forensic genetics to assumptions about founder genes, Ann. Appl. Stat. 3 (2009) 731–763.

[13] P.J. Green, J. Mortera, Paternity testing and other inference about relationships from DNA mixtures, Forensic Sci. Int. Genet. 28 (2017) 128–137, https://doi.org/10.1016/j.fsigen.2017.02.001.

[14] N. Kaur, M. Bouzga, G. Dørum, T. Egeland, Relationship inference based on DNA mixtures, Int. J. Legal Med. 130 (2016) 323–329.

[15] M.W. Perlin, Combining dna evidence for greater match information, Forensic Sci. Int. Genet. Suppl. Ser. 3 (1) (2011) e510–e511 Progress in Forensic Genetics 14.

[16] K. Ryan, D.G. Williams, D.J. Balding, Encoding of low-quality DNA profiles as genotype probability matrices for improved profile comparisons, relatedness evaluation and database searches, Forensic Sci. Int. Genet. 25 (2016) 227–239.

[17] F. Seillier-Moiseiwitsch, A.P. Dawid, On testing the validity of sequential probability forecasts, J. Am. Stat. Assoc. 88 (1993) 355–359.

[18] K. Slooten, Identifying common donors in DNA mixtures, with applications to database searches, Forensic Sci. Int. Genet. 26 (2017) 40–47.

[19] C. Steele, M. Greenhalgh, D. Balding, Evaluation of low-template DNA profiles using peak heights, Stat. Appl. Genet. Mol. Biol. 15 (2016) 431–445.