# Casework applications of probabilistic genotyping methods for DNA mixtures that allow relationships between contributors

Peter J. Green
UTS, Sydney, Australia
University of Bristol, UK.

Julia Mortera
Università Roma Tre, Italy.
University of Bristol, UK.

Lourdes Prieto
Forensic Sciences Institute
University of Santiago de Compostela, Spain
Comisaría General de Policía Científica, DNA Laboratory, Madrid, Spain

January 10, 2021

**Abstract**

In both criminal cases and civil cases there is an increasing demand for the analysis of DNA mixtures involving relationships. The goal might be, for example, to identify the contributors to a DNA mixture where the unknown donors may be related, or to infer the relationship between individuals based on a DNA mixture. This paper applies a new approach to modelling and computation for DNA mixtures involving contributors with arbitrarily complex relationships to two real cases from the Spanish Forensic Police.

*Some key words:* Coancestry, deconvolution, disputed relationship, identity by descent, kinship, DNA mixtures, likelihood ratio.

## 1 Introduction

In both criminal and civil cases relying on inference about relationships there is an increasing demand for the analysis of DNA mixtures where relatives are involved. The goal might be to identify the unknown contributors to a mixture where the donors may or may not be related, or to determine relationships between typed individuals and one (or more) of the contributors to a mixture, also in the case that the mixture contributors themselves are related. Here we use a novel approach that is able to tackle these problems, which to our knowledge have not previously been analysed rigorously in the literature. A new general software `KinMix` `R` package [6] which can handle complex relationships with and between mixture contributors has been developed to make inference in these cases. Inference is not limited to two-way relationships but can be extended to relationships among 3 (or possibly more) contributors to a mixture.

We analyse two real cases from the Spanish Forensic Police. In the first case we wish to identify a missing person through the analysis of DNA mixtures found on personal belongings. In many cases, the genetic profile detected on the objects is not from a single source, but might be a DNA mixture, revealing that the object was used by 2 (or more) people. In addition, very often, the contributors to these mixtures are related, mainly in cases, such as this one, where the missing person shared the

dwelling with relatives. Here, among other analyses, we tackle the novel problem of computing a likelihood ratio that the two unknown contributors to the mixture are related compared to unrelated, testing relationships such as parent-child, sibs, first cousins, etc.

The second case concerns a murder where a man was stabbed in his home. A DNA sample was taken from the murder weapon and appeared to be a DNA mixture from the victim and possibly a close relative of the victim.

Here we use probabilistic genotyping methods for DNA mixtures, under hypotheses about the relationships among contributors to the mixture and to other individuals whose genotype is available. We now briefly summarise these methods and refer to [10] which presents a review on DNA mixtures where further background can be found.

A natural basis for any model-based continuous DNA mixture analysis is a joint model for the peak heights $\mathbf{z}$ in the electropherogram (EPG) and genotypes $\mathbf{n}$, $p(\mathbf{n}, \mathbf{z}|\psi) = p(\mathbf{n}) \times p(\mathbf{z}|\mathbf{n}, \psi)$, with parameters $\psi$ characterising the conditional distribution of peak heights [4]. We base our analysis of DNA mixtures on the model for $p(\mathbf{z}|\mathbf{n}, \psi)$ described in [2]. This model takes fully into account the variation in peak heights and the possible artefacts, like stutter and dropout, that might occur in the DNA amplification process. The model can coherently analyse a combination of replicates, a combinations of different samples and a combinations of different kits.

In the standard case, unknown contributors to the mixture are assumed drawn at random from the gene pool. When the contributors are related, there is positive association between their contributor genotypes. A new model aimed at making inference about complex relationships from DNA mixtures is presented in [8]. This generalises the work in [7] which allowed inference about particular close relationships between contributors to a DNA mixture with unknown genotype and other individuals of known genotype. The new model extends the analysis to different scenarios and allows to specify arbitrary relationships between a set of actors, each of which may be mixture contributors, or have measured genotypes, or both. We can evaluate the likelihood of any such model, and compare models accordingly. A brief description of the key ideas underlying, specifying, modelling and computing relationship inference is given in the Appendix.

The case work examples in Section 2 illustrate some scenarios, where we make inference about two-way relationships between two mixture contributors with and without information about their or their relatives' genotypes.

The software used to analyse the case work examples is the new `KinMix` `R` package [6] that extends the `DNAmixtures` `R` package [4] to allow for modelling DNA mixtures with related contributors.

Among existing published work on relationships and mixtures, [11] presents an empirical study with data known to include known sibs among the reference samples, used to broaden the basis for evaluation of the information gain from using peak height data. Free software is also available to deal with DNA mixtures where contributors can be related [1], but this addresses a different problem: a specific kinship relationship has to be defined and one of the contributors has to be known.

# 2 Results of the analysis of complex DNA mixtures involving relationship testing

In this section we demonstrate the results and performance of our methods on the two case studies. For the first example we used the data gathered on 21 markers included in the GlobalFiler™ Amplification kit (ThermoFisher) and in the second example we also used data on 16 markers in the PowerPlex®16 kit. In all examples we assume known allele frequencies and adopt a threshold of 50 rfus.

## 2.1 Example 1: Identification using personal belongings of a missing person

**Background on the case** Personal belongings such as toothbrushes or razor blades can be used as a source of DNA in missing person cases. In these objects, DNA from the missing person may be found since they may have been frequently used before his/her disappearance. Nevertheless, there is uncertainty about the actual donor of the DNA isolated from these objects, and this is why it is recommended to "validate" the detected profile by using a reference (known) sample from a relative of the missing person. Usually, these profiles (from objects and/or relatives) are then compared with DNA profiles of unidentified bodies that are stored in national databases ("massive comparison"). This is useful to know if the missing person has passed away but his body was not identified. Unfortunately, in some cases, the genetic profile detected on an object is not a single source profile but a DNA mixture, revealing that the object was used by 2 (or more) people. In addition, very often, the contributors to these mixtures are related (mainly in cases where the missing person shared the dwelling with relatives).

In this example, we present a real case related to a missing male. The full anonymised data together with the `R` scripts to compute the results are given in the Supplementary Material web-pages[1]. The data are anonymised to avoid serious privacy and confidentiality concerns. In this case, only a daughter of the missing male was available to donate a DNA sample. This is not the ideal situation since false DNA matches can be found after a massive comparison of profiles in a database when only one relative is available as a reference sample. In order to improve the reference genetic data, a toothbrush and a razor-blade, presumably used by the missing person, were also collected. DNA from both objects was recovered and analysed by using the GlobalFiler kit (Thermo Fisher). The reference sample from the daughter of the missing male was also genotyped with the same kit. Two different DNA mixtures were detected in the two objects. An excerpt of the (anonymised) data is shown in Table 1, showing the alleles and peak heights in the two DNA mixtures found on the toothbrush T and the razor-blade RB. The DNA profile of the daughter, denoted by D, is also shown. The sex-related markers indicated that the mixture was most probably from one female and one male contributor.

**Results** Here we analyse the two DNA mixtures found on the toothbrush T, and a razor-blade RB, presumably used by the missing person (ante-mortem data).

Table 2 shows the estimated parameters $\psi = (\mu, \sigma, \xi, \phi)$ for the analysis of the DNA mixtures found on $T$ and $RB$. We assume there are 2 unknown contributors, denoted $U_1$ and $U_2$, to each of $T$ and $RB$: not necessarily the same individuals in the two cases. We fix on two contributors since the analysis performed for 3 (not shown here) yielded an almost vanishing proportion for the third contributor, $\phi_3 = 0$. Note also that the stutter proportion $\xi$ for sample $T$ is zero indicating that stutter peaks were most probably removed from the data. The estimated proportion of DNA for the two contributors to sample $T$ is large for the major contributor $U_1$, $\phi_{U_1} = 0.93$, whereas, for item $RB$ the estimated proportions of DNA contributed by $U_1$ and $U_2$ are roughly equal, $\phi_{U_1} \simeq \phi_{U_2} = 0.5$, implying they contributed in almost equal proportions to the mixture. As we will see in the latter case the estimation of the LR and other inference is problematic. In these models, the likelihood can have a complicated shape and numerical maximisation can be unreliable. The values in Table 2 are the maximum likelihood estimates as calculated by `DNAmixtures`.

Table 3 shows the LR and $\log_{10} LR$ for testing $\mathcal{H}_\text{p}$: D is the child of $U_1$ (and similarly for $U_2$) *vs.* $\mathcal{H}_0$: no unknown contributors are related to D. For item $T$, $\log_{10}\text{LR} = 10.97$ is large pointing to $U_1$ being a parent of D. It is also substantial for the hypothesis concerning $U_2$ being a parent of D. Could this be due to the fact that the two contributors might be related? We will test this assumption later. For the $RB$ the $\log_{10}\text{LR}$ in Table 3 for $\mathcal{H}_1$ *vs.* $\mathcal{H}_0$ is almost the same when

[1] `https://petergreenweb.wordpress.com/example-1-data-code-and-output/`

Table 1: Example 1: An excerpt of the anonymised data from the toothbrush $T$ and the razorblade $RB$, showing the markers, alleles and relative peak heights. The DNA profile of the daughter D of the missing person is also shown.

| markers | alleles in mixture | toothbrush peak height | razorblade peak height | D |
|---|---|---|---|---|
| Marker 6 | 17 | | 945 | |
| | 19 | 264 | 853 | 19 |
| | 21 | 3664 | 612 | 21 |
| Marker 7 | 13 | 1152 | 245 | |
| | 14 | 126 | 796 | |
| | 15 | 941 | 830 | 15 |
| Marker 14 | 13 | 5158 | 2141 | 13 |
| | 15 | 304 | 1512 | 15 |
| Marker 20 | 13 | 3218 | 334 | |
| | 17 | 3550 | 1795 | 17 |
| | 18 | | 1274 | |

Table 2: Example 1: Estimated parameters based on an analysis of the two mixture samples assuming that the toothbrush $T$ and $RB$ contain DNA from two unknown contributors.

| | $\mu$ | $\sigma$ | $\xi$ | $\phi_{U_1}$ | $\phi_{U_2}$ |
|---|---|---|---|---|---|
| toothbrush | 2381 | 0.0614 | 0 | 0.9262 | 0.0736 |
| razor-blade | 1602 | 0.0504 | 0.0118 | 0.5001 | 0.4999 |

testing whether D is the child of $U_1$ or of $U_2$. This is probably due to the fact that the proportions are almost identical, $\phi_{U_1} \simeq \phi_{U_2} = 0.5$, which makes it extremely difficult to distinguish between the contributors.

We also tested whether the daughter D was a contributor to $T$ or not, similarly for $RB$, and in both cases the logLR was zero, excluding D from being a contributor to either mixture.

In Table 4 we present comparisons with the results of another freely-available package that analyses DNA mixtures involving relatives, `relMix` [9]; this uses allele-presence only, not peak heights. We compare, marker-by-marker, with `KinMix` both with and without peak height information. The results obtained with `relMix` and `KinMix` when not including the peak height information (columns 2 and 3) are quite similar. Small differences between `relMix` and `KinMix` when not including peak heights are to be expected since they are based on different statistical models for the mixture. For the majority of markers, when including peak height information `KinMix` gave a larger $\log_{10}$ LR (10.97 compared with 9.53, corresponding to a LR 27.5 times smaller). For two of the markers, Markers 8 and 10, `relMix` is unable to compute the likelihood, most likely because of excessive storage demands. We can compute "partial" $\log_{10}$ LRs by excluding these 2 markers, and these are

Table 3: Example 1: $\log_{10} LR$ for testing whether in $T$ and $RB$, $\mathcal{H}_p$ contributor ($U_1$ or $U_2$) is a parent of D *vs.* $\mathcal{H}_0$ no contributor is related to D.

| | $\log_{10}$ LR | |
|---|---|---|
| | $U_1$ | $U_2$ |
| toothbrush | 10.974 | 4.531 |
| razor-blade | 8.443 | 8.444 |

Table 4: Example 1: Excerpt of marker-wise LR and overall $\log_{10}$ LR for item $T$, using `relMix` and `KinMix` with and without peak height information, for testing whether in $T$, $\mathcal{H}_p$: $U_1$ is a parent of D *vs.* $\mathcal{H}_0$: $U_1$ and $U_2$ are random members of the population.

| marker | `relMix` | `KinMix` w/o peak heights | `KinMix` with peak heights |
|---|---|---|---|
| Marker 6 | 2.55 | 2.58 | 3.34 |
| Marker 7 | 1.08 | 1.07 | 1.59 |
| Marker 9 | 1.26 | 1.18 | 1.62 |
| Marker 14 | 2.09 | 2.12 | 1.51 |
| partial $\log_{10}$ LR | 8.35 | 8.42 | 9.94 |
| overall $\log_{10}$ LR | | 9.53 | 10.97 |

also shown in the Table.

Table 5: Example 1: For items $T$ and $RB$, $\log_{10}$ LR for $\mathcal{H}_\text{p}$: the two contributors to the mixture are related, *i.e.* $U_1$ has relationship $R$ to $U_2$, *vs.* $\mathcal{H}_0$: the two contributors are unrelated. Several different relationships $R$ are tested.

| Relationship $R$ between $U_1$ to and $U_2$ under $\mathcal{H}_\text{p}$ | $T$ | $RB$ |
|---|---|---|
| | $\log_{10}$ LR | |
| monozygotic twins | $-\infty$ | $-\infty$ |
| parent-child | $-\infty$ | $-\infty$ |
| sibs | $-2.14$ | $-2.85$ |
| double first cousins | $-0.510$ | $-0.657$ |
| quadruple-half-first-cousins | $-0.44$ | $-0.630$ |
| half-sibs | $-0.37$ | $-0.625$ |
| first cousins | $-0.10$ | $-0.148$ |
| half-cousins | $-0.034$ | $-0.037$ |

Table 5 shows the results for testing whether the contributors $U_1$ or $U_2$ to item $T$ and $RB$ are related, *i.e.* $\mathcal{H}_p$*: $U_2$ has relationship $R$ to $U_1$ versus* $\mathcal{H}_0$*: $U_1$ and $U_2$ are unrelated.* The $\log_{10}$ LRs are all negative, implying that the LRs are smaller than 1. Although only a finite set of possible relationships has been considered, these vary widely, and it is overwhelmingly clear there is there is no support for any relationship between the two contributors.

We now consider the toothbrush EPG in more detail, examining the joint relationships between the mixture contributors and the typed daughter D, which clarifies the role of D in validating the mixture profile. Table 6 shows the $\log_{10}$ LR for item $T$ for several hypotheses $\mathcal{H}_\text{p}$ concerning different relationships $R$ among $U_1$, $U_2$ and D, *vs.* the null hypothesis that these individuals are all unrelated. The values of the $\log_{10}$ LR show that there is strong evidence that the two contributors to item $T$ are the missing father of D and D's mother, or at least very close relatives of them. Comparing the first 4 rows of Table 6 confirms that the most likely single possibility is that they are indeed the mother and father. All values in the Table remain unchanged if the sexes of all contributors are reversed; we choose to identify them in the way shown because inference (not shown) also including the Amelogenin locus indicates that is is most likely that the major contributor $U_1$ is female.

If there is interest in comparing two of the models displayed in Table 6, the appropriate $\log_{10}$ LR is simply obtained by calculating the difference beteen the values shown. For example, comparing the first row and the fifth, $17.935 - 10.974 = 6.961$ gives the weight of evidence that $U_2$ is the father of D, given that it is already assumed that $U_1$ is the mother of D. There are too many different such

comparisons that can be made to list them all here.

Some of the specific relationships examined in Table 6 are speculative, but might be of interest in cases where a home is shared by an extended family.

Table 6: Example 1: For item $T$, $\log_{10}\text{LR}$ for several hypotheses $\mathcal{H}_\text{p}$ concerning different relationships $R$ among $U_1$, $U_2$ and D, *vs.* $\mathcal{H}_0$: $U_1$ and $U_2$ and D are unrelated. The results in the lower half of the table can be used as baselines for comparison for those in the upper half. All $\log_{10}\text{LR}$ remain unchanged if the sexes of $U_1$ and $U_2$ are switched.

| $\mathcal{H}_p$ | $\log_{10}\text{LR}$ |
|---|---|
| $U_1$ mother of D and $U_2$ father of D | 17.935 |
| $U_1$ maternal aunt of D and $U_2$ father of D | 14.028 |
| $U_1$ mother of D and $U_2$ paternal uncle of D | 15.579 |
| $U_1$ maternal aunt of D and $U_2$ paternal uncle of D | 11.763 |
| $U_1$ mother of D and $U_2$ unrelated | 10.974 |
| $U_1$ maternal aunt of D and $U_2$ unrelated | 7.452 |
| $U_1$ unrelated and $U_2$ father of D | 4.530 |
| $U_1$ unrelated and $U_2$ paternal uncle of D | 2.796 |

Finally for this example, we consider the two mixture profiles $T$ and $RB$ jointly. What is the strength of evidence that the same individuals have contributed to both mixtures, and if so, are they related to D? To answer such questions, we use `KinMix` to model various scenarios which deal with the two DNA mixture traces simultaneously, with various patterns among the contributors. There are too many permutations to show them all, so in Table 7 we just present some interesting examples. As parameters for these joint peak height model, we copy over the relevant values from Table 2. For full details of these calculations, please consult the codes in the online Supplementary material.

Table 7 shows strong support for the hypothesis that the contributors to $T$ and $RB$ overlap and are mostly likely identical, strengthened further when a common contributor is a parent to D. As in previous analyses, the results are unchanged when sexes are interchanged, and in each hypothesis concerning a parent, the possibility that it is a close relative instead could also be examined.

Table 7: Example 1: $\log_{10}\text{LR}$ for the joint analysis of several hypotheses concerning the identity between contributors to $T$ and $RB$ and whether a common contributor is a parent of D. In all cases, the baseline $\mathcal{H}_0$ states that both contributors and D are unrelated. All $\log_{10}\text{LR}$ remain unchanged if the sexes of the contributors are switched. In the last two rows, the contributors are mentioned in order, major then minor, omitted for brevity.

| $\mathcal{H}_p$ | $\log_{10}\text{LR}$ |
|---|---|
| $T$ and $RB$ have same 2 contributors | 23.56 |
| $T$ and $RB$ have same 2 contributors, first being parent of D | 34.54 |
| $T$ and $RB$ have same major contributor | 16.53 |
| $T$ and $RB$ have same major contributor, being parent of D | 27.50 |
| $T$ has father and mother of D, $RB$ has father and unknown | 34.46 |
| $T$ has mother and father of D, $RB$ has father and unknown | 25.54 |

## 2.2 Example 2: Analyses of a Spanish murder case

**Description of the case** This concerns a murder case where a man was stabbed in his home. There was a knife with blood at the crime scene. The blood was mainly on the blade, but there was also some blood on the handle. The sample from the handle turned out to be a DNA mixture, with a major profile matching the victim. We also wish to test whether the minor profile in the mixture could be a close relative of the victim (possibly a son). The DNA profile of the son was not available. Two EPGs from the mixture were obtained by using two different kits, we denote these by EPG1 and EPG2. The kits have partially overlapping sets of markers, EPG1 was analysed on its 16 markers and EPG2 on its set of 22 markers, both include Amelogenin. Here we assume known allele frequencies taken from the Spanish allele frequency database collected on n = 284 individuals [3].

Months after the murder, a man was arrested for a different crime, drug trafficking, and a reference DNA sample was collected. When his profile was entered in the DNA database several matches were found, among which with the DNA mixture on the handle of the knife. The matches were investigated and the identity of the person (name, date of birth, place of birth, name of the father, name of the mother) was that of the son of the victim. Table 8 gives an excerpt of the data showing the markers, alleles and relative peak heights for EPG1 and EPG2, together with the genotypes of the father (the victim) and the son (the suspect).

Table 8: Example 2: An excerpt of the data showing the markers, alleles and relative peak heights for EPG1 and EPG2, together with the father's and son's genotypes

| marker | allele | EPG1 height | EPG2 height | father | son |
|---|---|---|---|---|---|
| CSF1PO | 10 | 305 | 625 | 10 | 10 |
| | 11 | 240 | 504 | 11 | 11 |
| D10S1248 | 13 | | 6990 | 13 | |
| | 14 | | 2309 | | 14 |
| | 16 | | 7144 | 16 | 16 |
| D7S820 | 9 | 606 | 1136 | 9 | 9 |
| | 10 | | 686 | 10 | |
| TH01 | 9.3 | 863 | 2654 | 9.3 | 9.3 |
| | 10 | 570 | | | 10 |

**Results** We analysed the data from this case to illustrate the different scenarios that can be analysed using the recently developed `Kinmix` code.

In particular we analyse the following different possible scenarios:

**Scenario 1** Here none of the contributors are typed. The analysis is of a 2-person mixture model for a prosecution hypothesis $\mathcal{H}_p$ : being the two unknowns being father and son versus $\mathcal{H}_0$ the two unknown contributors are unrelated.

**Scenario 2** Here only the father (the victim) is typed. The analysis is of a 2-person mixture model, where father has been typed and the prosecution hypothesis is $\mathcal{H}_p$ : son of father and 1 unknown are contributors versus $\mathcal{H}_0$ : no contributor is related to the typed individual (the victim).

**Scenario 3** Both father and son are typed. Here we analyse a 2-person mixture model where $\mathcal{H}_p$: the contributors are victim (father) and son versus $\mathcal{H}_0$ : contributors to the mixture(s) are 2

unknown individuals.

**Scenario 4** Both father and son are typed. Here we analyse a 2-person mixture model where $\mathcal{H}_p$: the contributors are victim (father) and son versus $\mathcal{H}_0$ : contributors to the mixture(s) are the victim and an unknown.

In all scenarios, unless otherwise stated, when considering an unknown contributor to a mixture, he or she is taken to be a random member of the reference population, so unrelated to typed individuals.

For EPG1 the MLEs of the parameters under both $\mathcal{H}_p$ and $\mathcal{H}_0$ are similar and are roughly equal to $\psi = (\mu = 576, \sigma = 0.32, \xi = 0, \phi_{U_1} = 0.88, \phi_{U_2} = 0.12)$. When the victim's genotype is known the estimated proportion contributed to EPG1 is $\phi_v = 0.18$, $\phi_{U_1} = 0.82$. For EPG2 the MLEs of the parameters are roughly equal to $\psi = (\mu = 2542, \sigma = 0.97, \xi = 0, \phi_{U_1} = 0.75, \phi_{U_2} = 0.25)$. When the victim's genotype is known the estimated proportion contributed to EPG1 is $\phi_v = 0.14$, $\phi_{U_1} = 0.86$. In both EPG1 and EPG2 the victim is estimated to be the minor contributor. Note that EPG2 has a higher $\mu$ than EPG1 but this is also accompanied by a larger $\sigma$, so the coefficient of variation is similar in both EPGs. The MLEs of the mean stutter proportion $\xi$ are zero, probably because preprocessing of the data has removed peaks that were classified in the laboratory as stutter. Our models, however, allow for stutter and do not require that the data be preprocessed before analysis.

Table 9 gives the $\log_{10}$ LR for the 4 scenarios when analysing EPG1 and EPG2 separately and jointly. When combining EPGs made from the same DNA extract, as in this case, it is natural to make an assumption that contributors are the same. In [5] we show how results based on a combination of replicates, a combinations of different samples and a combinations of different kits improve the robustness of the analysis and help in fixing any complications relating to degradation. However, when combining profiles from different samples one needs to carefully consider whether there is perhaps only a partial overlap.

Table 9: Example 2: $\log_{10}$ LR for Scenarios 1–4 using EPG1 and EPG2 separately and in combination.

| Scenario | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| Typed actors | none | father | father & son | |
| EPG1 | −0.806 | 5.60 | 22.16 | 22.78 |
| EPG2 | −0.175 | 10.66 | 29.16 | 11.68 |
| EPG1 & EPG2 | 2.49 | 8.26 | 40.17 | 26.20 |

Table 10: Example 2: For item EPG1 and EPG2, $\log_{10}$ LR for $\mathcal{H}_\mathrm{p}$: the two contributors to the mixture are related, *i.e.* $U_1$ has relationship $R$ to $U_2$, *vs.* $\mathcal{H}_0$: the two contributors $U_1$ and $U_2$ are unrelated and are independent of the typed individuals. Several different relationships $R$ are tested.

| | $\log_{10}$ LR | |
|---|---|---|
| Relationship | EPG1 | EPG2 |
| parent-child | −0.806 | −0.175 |
| sibs | −1.270 | −0.940 |
| quadruple-half-first-cousins | −0.316 | −0.022 |
| half-sibs | −0.275 | 0.045 |
| first cousins | −0.108 | 0.059 |
| half-cousins | −0.046 | 0.040 |

Table 10 shows $\log_{10}$ LR for testing whether the two unknown contributors to the DNA mixture are related versus that they are unrelated. For EPG1 the LRs for testing $\mathcal{H}_{\rm p}$ that the $U_1$ has a relationship $R$ to $U_2$, *vs.* $\mathcal{H}_0$: the two contributors $U_1$ and $U_2$ are unrelated and are independent of the typed individuals, vary between 0.16 and 0.9 giving roughly equal weight to $\mathcal{H}_1$ versus $\mathcal{H}_0$. For EPG2 these vary between 0.11 and 0.86.

Table 11 shows the deconvolution for the major contributor to the mixture for the two EPGs. The table only indicates genotype probabilities of at least 0.001, meaning that cells with a probability of less than 0.001 have been suppressed. We have denote by *other* the collection of alleles for which no peak has been observed in the EPG. For EPG1 the highest ranking genotype for the major contributor $U_1$ on all markers has posterior probability greater than 0.99 and coincides with the genotype of the suspect (who is the son of the victim) on all markers. The deconvolution for EPG2 gives a much poorer performance. For example, on marker D7S850 the top ranking genotype for EPG2 is incorrect, the correct genotype (9,9) is ranked 3$^{\rm rd}$ having a small probability of 0.077.

Table 11: Example 2: Predicted genotypes of $U_1$ with corresponding probabilities for EPG1 and EPG2 for an excerpt of the markers. An allele not observed in the EPG is denoted by *other*.

| | EPG1 | | | EPG2 | | |
|---|---|---|---|---|---|---|
| | genotype | | prob. | genotype | | prob. |
| CSF1PO | 10 | 11 | 1 | 10 | 11 | 0.751 |
| | | | | 10 | 10 | 0.097 |
| | | | | 11 | 11 | 0.083 |
| | | | | 10 | *other* | 0.036 |
| | | | | 11 | *other* | 0.033 |
| D13S317 | 12 | 13 | 0.997 | 12 | 13 | 0.576 |
| | 12 | 12 | 0.003 | 12 | 12 | 0.363 |
| | | | | 12 | *other* | 0.043 |
| | | | | 13 | *other* | 0.011 |
| | | | | 13 | 13 | 0.006 |
| D7S820 | 9 | 9 | 1 | 9 | 10 | 0.768 |
| | | | | 10 | 10 | 0.077 |
| | | | | 9 | 9 | 0.077 |
| | | | | 9 | *other* | 0.043 |
| | | | | 5 10 | *other* | 0.034 |
| TH01 | 9.3 | 10 | 1 | 9.3 | 9.3 | 0.812 |
| | | | | 9.3 | *other* | 0.185 |
| | | | | *other* | *other* | 0.003 |

# 3 Conclusions

We have shown that a wide range of relationship inference problems where one or more actors appear only as contributors to a DNA mixture, can be handled coherently. We can make inference about relationships among contributors, and between contributors and typed individuals. We carried out diagnostic plotting (not shown here) as recommended by [5] and found nothing to suggest the model was failing to fit the data.

The new `KinMix` package [6] used in the casework examples illustrated here is a highly flexible modular software package capable of solving much more complex relationships among two or more mixture contributors than those presented here. It is not limited to pairwise relationships. In [8] we show its capabilities of dealing with multi-way relationships in DNA mixtures including cases where the contributors might be inbred.

# Appendix

The key idea that enables the specification, modelling and computation of DNA mixtures with familial relationships among the contributors, and/or between contributors and other typed individuals is the IBD pattern distribution. IBD stands for identity by descent, the phenomenon where two or more related individuals have a common allelic value at a marker, not by the coincidence of several draws from the gene pool giving the same value, but because the allele was passed from parent to child in the process of meiosis. For a given set of related individuals, or 'family', an IBD pattern is a partition of the alleles of the individuals in the family according to their identity by descent. The IBD pattern distribution is simply the probability distribution of this partition induced by repeated application of Mendel's first law.

For just two related contributors, the idea has been in use to quantify relatedness for 80 years, in the form of Cotterham's kappas; for example, the relationship between two full brothers is captured by the probabilities that 0, 1 or 2 alleles are identical by descent: $\kappa_0 = 0.25$, $\kappa_1 = 0.5$, $\kappa_2 = 0.25$. The IBD pattern distribution extends this notion to any number of related individuals, and also deals correctly with inbreeding.

In `KinMix`, the IBD pattern distribution is used not only to specify the relationships in question, but also to model the distribution of the genotype profiles, and as a data structure to drive the computation. With unlinked autosomal STR markers in Hardy-Weinberg equilibrium, the joint distribution of the genotype profiles of the family members is completely determined by the IBD pattern distribution and the allele frequencies for each marker. As in much other recent work on computation for STR probabilistic genotyping methods for mixtures, joint distributions of genotype profiles are implemented using Bayesian networks (BNs), which allow efficient exact computation. The IBD pattern distribution is used directly in building the BN for the related genotypes. Full details are given in [8], and the methodology is implemented in the `R` package `KinMix` [6].

# References


[1] Ø. 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 Science International: Genetics*, 21:35–44, 2016.

[2] R. G. Cowell, T. Graversen, S. L. Lauritzen, and J. Mortera. Analysis of DNA mixtures with artefacts (with discussion). *Journal of the Royal Statistical Society: Series C*, 64:1–48, 2015.

[3] O. García, J. Alonso, J. A. Cano, R. García, G. M. Luque, P. Martín, I. M. de Yuso, S. Maulini, D. Parra, and I. Yurrebaso. Population genetic data and concordance study for the kits Identifiler, NGM, PowerPlex ESX 17 System and Investigator ESSplex in Spain. *Forensic Science International: Genetics*, 6(2):e78–e79, 2012.

[4] T. Graversen. *DNAmixtures: Statistical Inference for Mixed Traces of DNA*, 2013. R package version 0.1-4. `http://dnamixtures.r-forge.r-project.org`.

[5] T. Graversen, J. Mortera, and G. Lago. The Yara Gambirasio case: Combining evidence in a complex DNA mixture case. *Forensic Science International: Genetics*, 40:52–63, 2019.

[6] P. J. Green. *KinMix: DNA mixture analysis with related contributors*, 2020. R package version 2.0, `https://petergreenweb.wordpress.com/kinmix2-0`.

[7] P. J. Green and J. Mortera. Paternity testing and other inference about relationships from DNA mixtures. *Forensic Science International: Genetics*, 28:128–137, 2017. `http://dx.doi.org/10.1016/j.fsigen.2017.02.001`.

[8] P. J. Green and J. Mortera. Inference about complex relationships using peak height data from DNA mixtures, 2020. https://arxiv.org/abs/2005.09365.

[9] E. Hernandis, G. Dørum, and T. Egeland. relMix: An open source software for DNA mixtures with related contributors. *Forensic Science International: Genetics Supplement Series*, 2019. https://doi.org/10.1016/j.fsigss.2019.09.085.

[10] J. Mortera. DNA mixtures in forensic investigations: The statistical state of the art. *Annual Review of Statistics and Its Application*, 7:1–34, 2020. https://doi.org/10.1146/annurev-statistics-031219-041306.

[11] K. Slooten. The information gain from peak height data in DNA mixtures. *Forensic Science International: Genetics*, 36:119–123, 2018.