LRmix: statistical specifications

Version 1.0

Hinda Haned

July 4, 2013

Contents

1 Background

Forensic DNA testing of a biological traces consists in the comparison of the profile from the crime-scene (swab from the victim, blood, semen, saliva, etc.) to the DNA profile of one or several persons of interest, such as suspects and victims.

At the HBS department, such traces are generally analyzed with a genetic system called NGM. It consists of 15 genetic locations or loci. For every locus, there can be any number of alleles for traces. A given sample can be analyzed once or several times, with a maximum of four times. We call these repeated analyzes 'replicates'. Table 1 gives an example of such DNA profile.

Sample	Marker	Allele1	Allele2	Allele3
rep1	D3S1358	14	15	16
rep1	VWA	16	19	21
rep1	D16S539	9	10	11
rep1	D2S1338	20	23	24
rep2	D3S1358	14	15	16
rep2	VWA	16	19	21
rep2	D16S539	9	10	11
rep2	D2S1338	20	23	24
rep3	D3S1358	15	16	17
rep3	VWA	16	19	21
rep3	D16S539	9	10	11
rep3	D2S1338	20	23	24

Table 1: Example of a crime-scene DNA profile, at four loci and three replicates.

The DNA profile of the person of interest (a suspect or a victim) is only analyzed once. An example is given in Table 2.

Sample	Marker	Allele1	Allele2
suspect	D3S1358	15	17
suspect	VWA	16	21
suspect	D16S539	9	10
suspect	D2S1338	23	24

Table 2: DNA profile of a suspect. Only four loci are shown.

When comparing crime-sample profiles and reference profiles, the question of interest is: "is this person the donor to the crime-sample?" In order to assess this question, forensic scientists calculate a statistic that translates the weight of the DNA evidence. There are several ways in which such weight can be evaluated, however, the preferred approach is termed the likelihood ratio (LR) (Gill et al., 2006, 2012). Within the likelihood ratio approach, the reporting officer (RO) of the NFI evaluates two alternative hypotheses explaining the origin of the crime-scene profile: the hypothesis of the prosecution (H_p) and the hypothesis of the defense (H_d). The weight of the evidence is then expressed in terms of a likelihood ratio, measuring the relative weight of each hypothesis. The LR can be written as:

$$LR = \frac{Pr(E|H_p)}{Pr(E|H_d)} \tag{1}$$

where E is the profile of the DNA evidence. H_p is the hypothesis of the prosecution, typically of the form: the suspect is the donor to the crime scene profile, while H_d is the hypothesis of the defense, usually of the form: an unknown person is the donor to the crime-scene profile. While the prosecution hypothesis tries to explain the evidence through the profile of the suspect, the defense will argue that an alternative donor, who happens to have the same or a similar DNA profile than the suspect, has contributed to the DNA profile. For example, if a LR=10,000 is obtained for the above hypotheses, the RO would report that "the profile of the examined crime-sample is 10,000 times more likely to be observed if H_p is true, than if H_d is true".

In order to evaluate how likely the evidence is if one or the other of the two hypotheses is true, a probabilistic model, termed as the drop-out/drop-in model, has been introduced at the HBS department. This model, along with its successive enhancements, were described and evaluated in several peer-reviewed papers (Gill et al., 2007; Curran et al., 2005; Gill et al., 2008; Haned et al., 2012).

I have implemented this probabilistic model in the *likEvid* function, written in the R language. This function is available in the Forensim package for the R statistical software[1] (Haned, 2011). *likEvid* is written in R, but in fact it calls a series of functions written in C code. This ensures faster calculations in the R environment. In order to make this function of Forensim more accessible to reporting officers, I created a Tcl/Tk user-friendly interface, called *LRmix*, also available in the Forensim package. *LRmix* is currently used to report likelihood ratios on complex cases at the HBS department of the NFI.

This document intends to be a detailed description of the probabilistic model implemented in *LRmix*. It compiles the information available in the literature, and provides detailed examples on simple cases. Throughout the document, links between the different components of the model, and its implementation within *LRmix* are outlined.

2 Model description

The first step of the analysis is to specify the prosecution and the defense hypotheses. DNA profiles do not always consist of a DNA from a single person, and the more individuals are involved the more complex the hypotheses are. Table 3 gives a number of possible sets of hypotheses.

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

H_p:	The suspect is the donor
H_d:	An unknown person is the donor
H_p:	The suspect and the victim are the donors
H_d:	Two unknowns are the donors
H_p:	The suspect and three unknowns are the donors
H_d:	Four unknowns are the donors

Table 3: Examples of H_p and H_d hypotheses.

The likelihood ratio of a pair hypotheses, such as the ones stated above, are calculated for every locus of a given profile (see Table 1). The Overall LR is simply obtained by taking the product of the per-locus LRs:

$$LR = \frac{Pr(E_1|H_p)}{Pr(E_1|H_d)} \times ... \times \frac{Pr(E_L|H_p)}{Pr(E_L|H_d)} \tag{2}$$

where E_L is the profile of the DNA evidence at locus L. For example, in Table1, $E_{VWA} = \{16, 19, 21; 16, 19, 21; 16, 19, 21\}$, where the three replicates are separated by the ';'.

2.1 Theoretical considerations

The quantity of interest is given in eq. (3): this the the probability of observing the n replicates in the DNA profiles, conditioned on the known and the unknown genotypes, as specified by a given hypothesis:

$$Pr(R_1, ..., R_n|H) = Pr(R_1, ..., R_n|T, V, x) \tag{3}$$

- $R_1, ..., R_n$ are the replicates of the crime-scene profile, for a given locus L[2]. At the NFI, at most four replicates are obtained for every crime-scene profile, the minimum being one.

- T is the list of genotypes from the profiled individuals. For example, the suspect and the victim are profiled, their genotypes are known exactly. There is no limit to the number of genotypes in T, but casework experience shows that this number does not exceed three individuals.

- V is the list of genotypes from the profiled individuals that are known to be non-contributors under a given hypothesis H. For instance, a suspect who has donated DNA under H_p, becomes a known non-contributor under hypothesis H_d. We define this set because genotypes that have been observed will determine how likely it is to see a given genotype for the unknown. Note that V is automatically filled by the *LRmix* program depending on the contributors stated under H_p. All the profiled individuals that are among the hypothesized contributors under the prosecution but not under the defense hypothesis, become known non-contributors under H_d.

[2] I leave out the indexation on L for clarity

- x is the number of unknown contributors to the sample under a given hypothesis H. However, the crime-sample can be a mixture of DNAs from the suspect and one additional person, who is unknown ($x = 1$). The model will then account for this source of uncertainty by going through all possible genotypes for this unknown person.

By definition, the x unknown individuals could have any set of genotypes. We denote U the set of possible genotypes for the x unknown individuals, and we rewrite equation (3) as follows:

$$Pr(R_1, ..., R_n|T, V, x) = Pr(R_1, ..., R_n|T, V, U) \tag{4}$$

The replicates R_i are (conditionally) independent, thus, the conditional probabilities can be evaluated for each genotype in the U set and eq. (4) can be re-written:

$$Pr(R_1, ..., R_n|T, V, U) = \sum_g Pr(U_g|T, V) \prod_i Pr(R_i|T, U_g) \tag{5}$$

Equation (5) helps decompose the probability calculations in two separate components, which can be calculated separately: the replicate probability $Pr(R_i|T, U_g)$ and the genotype probability $Pr(U_g|T, V)$. Note that the replicate probability only depends on the genotypes of the assumed contributors, and the set V is ignored. Table 4 outlines these two components, along with the parameters (input by the user, and later discussed in the document) needed for the calculations. Note that in case there are no unknown individuals, $U = \varnothing$ and eq. (5) simplifies to eq. (6).

$$Pr(R_1, ..., R_n|T) = \prod_i Pr(R_i|T) \tag{6}$$

Probability	**If $x = 0$**	**Parameters**
Replicate	$Pr(R_i\|T)$	Drop-out probability Drop-in probability Allele frequencies
Genotype	1	-

Probability	**If $x \neq 0$**	**Parameters**
Replicate	$Pr(R_i\|T, U_g)$	Drop-out probabilities Drop-in probability Allele frequencies
Genotype	$Pr(U_g\|T, V)$	Allele frequencies θ

Table 4: Probability terms and corresponding parameters following eq. (5). Note that the replicate probability depends only on the genotypes who have contributed, thus, V and $U = \varnothing$ are discarded in the conditioning. The parameters needed for the calculations are discussed later in this document.

Thes replicate and genotype probabilities are evaluated for every element g in set U of the genotypes of the unknown. If there are no unknown contributors ($x = 0$), then the genotype probability is one. Indeed, the genotype probability is only calculated when there are unknown genotypes, however if there are unknown individuals that have not been profiled, we need to factor in the genotype probabilities in the model. In all cases, the replicate probability has to be calculated.

In what follows, I explain how the genotype, and the replicate probabilities are calculated.

Box1
The *likEvid* function in R calls the *LRC* function written in C . This function carries out all the calculations of eq. (5), under a given hypothesis H, defining the content of sets T, V and U. The results are then passed to the R function *likEvid*. Thus, to calculate the LR, *likEvid* is called iteratively for every locus of the crime-scene profile, both under H_p and H_d. The *LRC* function calls two main functions: *genoC* and *repC*, which implement the genotype and the replicate probabilities, respectively. The interaction between the C and R code is outlined in Appendix A.

2.2 Genotype probability

If x $\neq$ 0, $Pr(U_g|T, V)$ has to be calculated. This is the probability of observing a certain combination of genotypes (for the unknown), denoted g, based on the information provided by the profiles in the set of known contributors T and known non-contributors V. The details of the calculations are given below, but first, I explain how the possible genotypes for the unknowns are derived.

2.2.1 Determining the genotypes for the unknowns

While information in T, V is provided by user input in the *LRmix* program, the genotypes of the unknown individuals have to be derived by the program. Since the unknown individuals were not profiled, they could have any genotype, with no restriction. In order to derive these genotypes at a given locus, we have to derive all possible combinations of alleles at that locus.

If a locus has m alleles in the population, then there are $C^2_{m+1} = \frac{(m+1)!}{2(m-1)!}$ ways of choosing two alleles among these m alleles to form a genotype. This corresponds to the number of combinations of two elements among m, with replacement (Curran et al., 2005). For example, at locus FGA, 20 alleles were observed in the Dutch population, thus, at this locus, there are 210 distinct genotypes.

If there is a single unknown, there are C^2_{m+1} possible genotypes for him. If there are more than one unknown, we have to consider all the combinations of x genotypes among the C^2_{m+1} genotypes that are possible at a given locus.

Box 2
The possible combinations for the genotypes are determined in the main C function *LRC*, based on the information provided by the allele frequencies. The combinations for x unknown individuals are derived iteratively, and for each combination, functions *genoC* and *repC* are then called to carry out the calculations outlined in eq. (5). The results are then passed to the R function *likEvid*.

For example, if a locus has only two alleles 13, 14, then there are three possible genotypes: 13/13; 14/14 and 13/14. With two unknown individuals ($x = 2$), and three possible genotypes, there are six unique combinations for two individuals (the number of ways of choosing two genotypes among three with replacements). Table 5 outlines these possibilities.

Combination	Unknown 1	Unknown 2
U_1	13,13	13,13
U_2	13,13	13,14
U_3	13,14	13,14
U_4	13,14	14,14
U_5	14,14	14,14
U_6	13,13	14,14

Table 5: Possible genotypic combinations at a virtual locus with two alleles 13 and 14. Note that the order in which the genotypes are assigned to the unknown individuals is arbitrary.

Note that Table 5 ignores the possibility 13/14; 13/13, which is in fact redundent with 13/13; 13/14. Howvever, both combinations have to be accounted for. Table 6 gives the exhaustive list of possibilities for the two unknowns.

Combination	Unknown 1	Unknown 2
U_1	13,13	13,13
U_2	13,13	13,14
U_2'	13,14	13,13
U_3	13,14	13,14
U_4	13,14	14,14
U_4'	14,14	13,14
U_5	14,14	14,14
U_6	13,13	14,14
U_6'	14,14	13,13

Table 6: Exhaustive list of all the possible genotypic combinations at a virtual locus with two alleles 13 and 14. Note that the combinations that were ignored in Table 5 above, are highlighted here.

For this example, there are nine possible genotypic combinations for the two unknowns. In fact, we have that the genotypic probability is not affected by the

order in which the genotypes are evaluated. Thus, for those combinations containing the same genotypic types, the genotype probability can be evaluated only once. For instance we have:
$Pr(U_2 = \{13/13; 13/14\}|T, V) = Pr(U'_2 = \{13/14; 13/13\}|T, V)$,
this also applies to the replicate probability in equation (5):

$$Pr(R_1, ..., R_n|T, U_2 = \{13/13; 13/14\}) = Pr(R_1, ..., R_n|T, U'_2 = \{13/14; 13/13\})$$

Thus, the calculations can be carried out for the unique pairs of combinations, provided we correct for the ignored (redundant) combinations with an appropriate permutation factor. For every genotypic combination, the permutation factor is simply the number of permutations with replacements of that given combination. Thus, for a given combination, if there are k_1 genotypes of type 1, and k_2 genotypes of type 2, and so on, the permutation factor is given by $\frac{n!}{k!} = \frac{n!}{k_1!...k_y!}$
where n is the total number of genotypes, and k_y is the number of genotypes of type y among the n genotypes. For example, for the second combination in Table 5, $n = 2$ genotypes, each one of them appearing once, thus, $k_1 = 1$ and $k_2 = 1$, this gives us $\frac{2!}{1!1!} = 2$ permutations.

2.2.2 Calculating the genotype probabilities

The probability of any genotypic combination in U, is simply that of its expected frequency in the target population. For a homozygote genotype 13/13, the genotypic probability is p_{13}^2, and for a heterozygote genotype 13/14, the probability is $2p_{13}p_{14}$. These estimates of the genotypic frequencies from the allele frequencies follow from a basic model in population genetics, called the Hardy-Weinberg model (Evett and Weir, 1998). Following this model, probabilities for rare and common genotypes, only depend on the allele frequencies. In fact, if there is a rare allele in the crime-sample, and the suspect happens to have this allele, then the evidence would point quiet strongly towards including him as a possible donor to the crime stain profile. However, the frequencies of the alleles observed at different loci are estimates. Thus, it is possible that the rarity of a profile could be the result of a bad sampling strategy, or, that of the sampling of the wrong population: a profile could be very rare in the whole Dutch population, but very common in the suspect's (undefined) sub-population. Unfortunately, it is hard, if not impossible, to define human sub-populations (Buckleton et al., 2005). In order to correct for this source of uncertainty in the allele frequencies, a common practice consist in using a correction for the allele frequencies, called the θ correction (Balding and Nichols, 1994). θ is a measure of the subdivision of human populations, it translates the extent to which allele frequencies in sub-populations can differ from the frequencies of the population as a whole. Usual values are between 0.01 and 0.05 for human populations. At NFI, this correction is usually set to 0.01. The θ correction is applied through a 'sampling formula' (Balding and Nichols, 1994). For every allele k in a given genotypic combination U_g, the sampling formula is defined as follows:

$$Pr(U_g|T,V) = 2^h \prod_{k \in U_g} \frac{m_k\theta + (1-\theta)p_k}{1+(n_k-1)\theta} \tag{7}$$

where, for every element k of U_g, m_k is the number of alleles of type k in sets U_g, T and V, n_k is the number of alleles that have been sampled when every allele k of U_g is evaluated, these are the alleles in sets T and V (if they are not empty) and the alleles that have already been sampled in U_g. h is the number of heterozygote genotypes U_g[3], and p_k is the frequency of allele k. Importantly, if $\theta = 0$ in eq. (7), the Hardy-Weinberg estimates are recovered.

To illustrate the calculation of the genotype probability using the sampling formula, two examples are given below.

2.2.3 Example 1

In this example, we consider the genotypic combination $U_2 = \{13/13; 13/14\}$, $T = \varnothing$ and $V = \varnothing$.

For every element (allele) in U_2 we apply the formula in eq. 7, and we multiply by the permutation factor. The sampling formula goes through the genotypic combination U_2 allele per allele, as indicated per the blue color below:

- $\boxed{13}$/13;13/14
  - no alleles of type 13 have been sampled yet → $m_k = 0$
  - no alleles in T nor V → $n_k = 0$
  - If we apply the formula, we obtain: p_{13}
- 13/$\boxed{13}$;13/14
  - one allele of type 13 has been sampled already → m_k=1
  - one allele has been sampled (previous step) → n_k=1
  - If we apply the formula, we obtain: $\theta + (1-\theta)p_{13}$
- 13/13;$\boxed{13}$/14
  - two alleles of type 13 have been sampled already → $m_k = 2$
  - two alleles have been sampled (previous step) → $n_k = 2$
  - If we apply the formula, we obtain: $\dfrac{2\theta + (1-\theta)p_{13}}{1+\theta}$
- 13/13;13/$\boxed{14}$
  - no alleles of type 14 have been sampled yet → $m_k = 0$
  - three alleles have been sampled (previous steps) → $n_k = 3$

[3] Note here that the formula goes through U_g, allele per allele, if taken two per two, these alleles form genotypes of the unknown individuals.

  - If we apply the formula, we obtain: $\dfrac{(1-\theta)p_{14}}{1+2\theta}$

To obtain the probability, we multiply the probabilities obtained at each iteration:

$$Pr(U_2 = \{13/13; 13/14\}|T = \varnothing, V = \varnothing) = 4p_{13}(\theta+(1-\theta)p_{13})\frac{2\theta + (1-\theta)p_{13}}{1+\theta}\frac{(1-\theta)p_{14}}{1+2\theta}$$

Note that we multiply by four, this is the product of the permutation factor of two and the correction for the heterozygote genotypes (h in eq. 7).

2.2.4 Example 2

In this second example we consider the genotypic combination $U_2 = \{13/13; 13/14\}$, $T = \{11/12\}$ and $V = \varnothing$.

We start with the first allele 13 (the order does not matter):

- $\boxed{13}$/13;13/14
  - no alleles of type 13 have been sampled yet → $m_k = 0$
  - two alleles in T → $n_k = 2$
  - Applying the formula, we obtain: $\dfrac{(1-\theta)p_{13}}{1+\theta}$
- 13/$\boxed{13}$;13/14
  - one allele of type 13 has been sampled already → $m_k = 1$
  - three alleles have been sampled (previous step) → $n_k = 3$
  - Applying the formula, we obtain: $\dfrac{\theta + (1-\theta)p_{13}}{1+2\theta}$
- 13/13;$\boxed{13}$/14
  - two alleles of type 13 have been sampled already → $m_k = 2$
  - four alleles have been sampled (previous steps) → $n_k = 4$
  - If we apply the formula, we obtain:
    $\dfrac{2\theta + (1-\theta)p_{13}}{1+3\theta}$
- 13/13;13/$\boxed{14}$
  - no alleles of type 14 have been sampled yet → $m_k = 0$
  - five alleles have been sampled (previous steps) → $n_k = 5$
  - If we apply the formula, we obtain: $\dfrac{(1-\theta)p_{14}}{1+4\theta}$

Thus we can write:

$$Pr(U_2 = \{13/13; 13/14\}|T = \{11/12\}, V = \varnothing) =$$
$$4\frac{(1-\theta)p_{13}}{1+\theta}\frac{\theta+(1-\theta)p_{13}}{1+2\theta}\frac{2\theta+(1-\theta)p_{13}}{1+3\theta}\frac{(1-\theta)p_{14}}{1+4\theta}$$

Here again note that we multiply by four, this is the product of the permutation factor of two and the correction for the heterozygote genotypes.

Box 3
The *genoC* function is called sequentially on every combination in U, generated by the main function *LRC*. The function computes the probability according to the formula in eq. (7), then calculates the number of permutations for every genotypic combination U_g of U. In practice, we need to know how many genotypes are repeated within set U_g (that is one possible combination within set U). The program goes through every genotype and records the number of different genotypes, it then applies the permutation factor.

2.3 Replicate probability

The replicate probability determines how likely a given replicate is, conditioned on the genotypes of the donors under H (including the unknown) and on the drop-out and drop-in parameters. Every element in every replicate is compared to the genotypes of T and U_g, thus the probability of replicate R_i must be defined for each combination of T and U_g. For simplicity, we define the set of genotypes G_g, which is the union of sets T and U_g: $G_g = T \cup U_g$. Eq. (6) can be re-written:

$$Pr(R_i|T, U_g) = Pr(R_i|G_g) \tag{8}$$

2.3.1 Drop-out and drop-in: definitions

A drop-out is an allele that is not detected in the crime-sample profile, because of a low quality or quantity of DNA. Although it is present in the sample, this allele may not be detected. A drop-in is an allele that is detected in the crime-sample profile, although it is not related to the crime-sample. Drop-in is due to DNA fragments present in the laboratory environment (e.g plasticwear). Drop-in can be considered as a case of mild contamination, where one or two alleles, not related to the sample being analyzed, appear in the profile of the trace.

2.3.2 Drop-out and drop-in: formalization

Drop-out and drop-in being stochastic events, they may occur in on replicate, but not in the other. Following Haned et al. (2012), the drop-out and drop-in alleles are treated as independent across the replicates. In order to facilitate the calculation of the replicate probability, based on the drop-out and drop-in probabilities, the information provided by each replicate R_i and a particular combination G_g, is partitioned in into three disjoint sets:

- Drop-out set: the set of dropped-out alleles, these alleles are in G_g but not in R_i .

- Non drop-out set: the set of alleles that did not drop-out, these alleles are both in G_g and R_i

- Drop-in set: the set of alleles that have dropped-in, these are alleles in R_i but not in G_g

These sets can be further described as follows:

- A_i: drop-out set, these are the alleles that have dropped-out (whether in one or several copies), these are all alleles in G_g, but not in R_i,

- $B_{1,i}$: set of alleles that have not dropped-out, and are present in one copy only in G_g,

- $B_{2,i}$: set of alleles present in multiple copies in G_g that have not dropped out from R_i,

- C_i: alleles that have dropped-in in replicate R_i.

For every genotypic combination g, the replicate probability is defined by the following equations:

If $C_i \neq \varnothing$

$$Pr(R_i|G_g) = \prod_{a \in A_i} d_a \prod_{b \in B_{1,i}} (1 - d_b) \left(1 - \prod_{b \in B_{2,i}} d_b\right) \prod_{k \in C_i} cp_k \tag{9}$$

If $C_i = \varnothing$

$$Pr(R_i|G_g) = (1 - c) \prod_{a \in A_i} d_a \prod_{b \in B_{1,i}} (1 - d_b) \left(1 - \prod_{b \in B_{2,i}} d_b\right) \tag{10}$$

where d_a is the drop-out indicator for allele a, and c is the drop-in probability[4].

For every replicate i, sets A_i, $B_{1,i}$, $B_{2,i}$ and C_i are defined, and the joint probability for the n replicates is obtained by multiplying the probabilities obtained for each replicate.

[4] in *LRmix*, both probabilities are provided by the user

Box 4

In eq. (9) and (10), every element of every drop-out set has a defined drop-out probability. This probability can in principle vary across genotypes and across loci. In the *likEvid* function, the user has to input the drop-out parameter per contributor and per locus, as well as the drop-in parameter. In practice the *likEvid* function has two arguments related to the drop-out: *prDHet*, for drop-out of the heterozygote genotypes, and *prDHom*, for the drop-out of the homozygote genotypes. If there are four hypothesized contributors under H_p, among which two unknowns, the *prDHet* parameter has to be of length four, every element giving the dropout for the contributors in T (profiled contributors), followed by the drop-out parameter to apply to the two unknown contributors. The drop-in is also input by the user, in the *prC*. In the current version of *LRmix*, the user applies the same drop-out probabilities to all loci (see Appendix B), but in principle, this probability could vary across loci, because some loci can have a higher susceptibility to drop-out than others.

In what follows, I illustrate how the calculations for the replicate probability are carried out.

2.3.3 Example 1

In this example, we consider replicates $R_1 = \{11, 12, 13\}$ and $R_2 = \{11, 12\}$, and a combination G_1 of the known and unknown genotypes G_1=$\{$11/11;12/12;13/14$\}$. As explained in Box 4 above, the drop-out parameters are determined per contributor, and then applied per genotype of every contributor. If the drop-out parameter of a genotype of a given person, say the suspect (first contributor) is d_1, then this means that all alleles from this person, can drop-out form the replicates with a probability that can be written as a function of d_1. Specifically, one allele from a heterozygote genotype (eg. 13/14) drops-out with a probability d_1, while an allele from homozygote genotype (eg. 13/13) drops-out with a probability $d'_1 = d_1^2$. In the G_1 combination, there are three genotypes belonging to three different contributors, this means that there are three heterozygote drop-out parameters, and the homozygote drop-out parameters are derived by simply taking the square of the heterozygote drop-outs:

- Heterozygote drop-out parameters for every genotype in the G_1 combination: d_1, d_2 and d_3,
- Homozygote drop-out parameters for every genotype in the G_1 combination are $d'_1 = d_1^2$, $d'_2 = d_2^2$ and $d'_3 = d_3^2$.

For every replicate R_i, we define the alleles in the sets A_i, $B_{1,i}$, $B_{2,i}$ and C_i, and the corresponding parameters for every allele in these sets:

Replicate R_1	Parameters
$A_1 = \{14\}$	d_3
$B_{1,1} = \{11, 12, 13\}$	d'_1, d'_2, d_3
$B_{2,1} = \varnothing$	-
$C_1 = \varnothing$	$1 - c$

Replicate R_2	Parameters
$A_2 = \{13, 14\}$	d_3, d_3
$B_{1,2} = \{11, 13\}$	d'_1, d'_2
$B_{2,2} = \varnothing$	-
$C_2 = \varnothing$	$1 - c$

Following eq. (10), the replicate probabilities for this example are:

$$\begin{aligned}
Pr(R_1, R_2|G_1) &= \left[(1-c)\prod_{a\in A_1} d_a \prod_{b\in B_{1,1}} (1-d_b)\right]\left[(1-c)\prod_{a\in A_2} d_a \prod_{b\in B_{1,2}} (1-d_b)\right] \\
&= [(1-c)d_3(1-d'_1)(1-d'_2)(1-d_3)]\left[(1-c)d_3^2(1-d'_1)(1-d'_2)\right] \\
&= (1-c)^2 d_3^3(1-d'_1)^2(1-d'_2)^2(1-d_3)
\end{aligned}$$

2.3.4 Example 2

In this example we consider the genotypic combination G_2={11/11; 12/12; 12/14} and the same replicates as in Example 1: $R_1 = \{11, 12, 13\}$, and $R_2 = \{11, 12\}$. Note that in this example, the genotypes in the G_2 set have the shared allele 12. The drop-out parameters remain the same as in Example 1. For every replicate i, we define the sets A_i, $B_{1,i}$, $B_{2,i}$ and C_i as follows:

Replicate R_1	Parameters
$A_1 = \{14\}$	d_3
$B_{1,1} = \{11\}$	d'_1
$B_{2,1} = \{12\}$	d'_2, d_3
$C_1 = \{13\}$	cp_{13}

Replicate R_2	Parameters
$A_2 = \{14\}$	d_3
$B_{1,2} = \{11\}$	d'_1
$B_{2,2} = \{12\}$	d'_2, d_3
$C_2 = \varnothing$	$1 - c$

Applying eq. (9) and (10), the replicate probability is:

$$\begin{aligned}
Pr(R_1, R_2|G_2) = &\left[\prod_{a\in A_1} d_a \prod_{b\in B_{1,1}} (1-d_b)\left(1-\prod_{b\in B_{2,1}} d_b\right)\prod_{k\in C_1} cp_k\right] \times \\
&\left[(1-c)\prod_{a\in A_2} d_a \prod_{b\in B_{1,2}} (1-d_b)\left(1-\prod_{b\in B_{2,2}} d_b\right)\right] \\
&= [d_3(1-d'_1)(1-d'_2d_3)cp_{13}]\,[(1-c)d_3(1-d'_1)(1-d'_2d_3)] \\
&= c(1-c)p_{13}d_3^2(1-d'_1)^2(1-d'_2d_3)^2
\end{aligned}$$

While previous examples served as illustrations for the replicate and genotype probabilities, calculated separately, the next section details the calculation of the likelihood ratio on a simple example.

3 Full example

In this section I illustrate the calculations for a simple single-locus example, with a sample with a single donor. At a given locus, the replicates show alleles as follows: $R_1 = \{13\}$ and $R_2 = \{13, 14\}$. The hypotheses to be evaluated in the likelihood ratio are:

- H_p: the suspect with genotype 13/14 contributed to the sample,
- H_d: an unknown person, unrelated to the suspect, contributed to the sample.

To reduce the complexity of the calculations, I assume that the considered locus shows only three alleles in the Dutch population, 12,13 and 14, with frequencies p_{12}, p_{13} and p_{14}, respectively.

Under H_p Since there are no unknown contributors under H_p, the genotype probability is one. The replicate probabilities are given in eq. (11) and (12).

$$Pr(R_1 = \{13\}|T = \{13/14\}) = d_1(1 - d_1)(1 - c) \tag{11}$$

$$Pr(R_2 = \{13, 14\}|T = \{13/14\}) = (1 - d_1)^2(1 - c) \tag{12}$$

The join replicate probabilities is then simply the product of equations (11) and (12):

$$Pr(R_1, R_2|T = \{13/14\}) = d_1(1 - d_1)^3(1 - c)^2 \tag{13}$$

Since there are no unknwons under H_p, the probablity of the evidence is simply the product of the replicate probabilities:

$$Pr(E|H_p) = d_1(1 - d_1)^3(1 - c)^2 \tag{14}$$

Under H_d Under H_d, there is one unknown, and one known non-contributor (the suspect) with genotype 13/14. Thus, $x = 1$, $T = \varnothing$, and $V = \{13/14\}$. The first step of the analysis, regardless of the profile of the evidence, is to derive the possible genotypes for the unknown individual. The considered locus, has three distinct alleles in the Dutch population. Thus, there are six possible genotypes: 12/12, 13/13, 14/14, 12/13, 12/14, and 13/14. The unknown could have any of these genotypes. Thus, set U contains six possible genotypes. Table 7 below, gives the genotypic probabilities for these six genotypes, when $\theta = 0$ and when $\theta \neq 0$.

U_g	$Pr(U_g \mid T,V)$	
	$\theta = 0$	$\theta \neq 0$
12/12	p_{12}^2	$\frac{(1-\theta)p_{12}(\theta+(1-\theta)p_{12})}{(1+\theta)(1+2\theta)}$
13/13	p_{13}^2	$\frac{(\theta+(1-\theta)p_{13})(2\theta+(1-\theta)p_{13})}{(1+\theta)(1+2\theta)}$
14/14	p_{14}^2	$\frac{(\theta+(1-\theta)p_{14})(2\theta+(1-\theta)p_{14})}{(1+\theta)(1+2\theta)}$
12/13	$2p_{12}p_{13}$	$2\frac{(1-\theta)p_{12}(\theta+(1-\theta)p_{13})}{(1+\theta)(1+2\theta)}$
12/14	$2p_{12}p_{14}$	$2\frac{(1-\theta)p_{12}(\theta+(1-\theta)p_{14})}{(1+\theta)(1+2\theta)}$
13/14	$2p_{13}p_{14}$	$2\frac{(\theta+(1-\theta)p_{13})(\theta+(1-\theta)p_{14})}{(1+\theta)(1+2\theta)}$

Table 7: Genotype probablities under H_d.

The next step, is to derive the replicate probablities for every possible genotype of the unknwon individual. Table 8 gives these probabilities.

G_g	$P(R_1, R_2 \mid Gg)$
12/12	$d_1'^2 c^2 p_{13} p_{14}$
13/13	$(1-d_1')^2 c p_{14}(1-c)$
14/14	$d_1'(1-d_1')c^2 p_{13}^2$
12/13	$d_1^2(1-d_1)^2 p_{14} c(1-c)$
12/14	$d_1^3(1-d_1)(cp_{13})^2$
13/14	$(1-c)^2 d_1(1-d_1)^3$

Table 8: Replicate probabilities for the genotypes of the unknown. The drop-out probabilities of the heterozygote and homozygote alleles are d_1 and d_1' respectively.

Applying the formula in equation (5), and combining the terms in Tables 7 and 8, the probability of the evidence under H_d when $\theta = 0$:

$$\begin{aligned}
&Pr(E|H_d) = d_1'^2 c^2 p_{13} p_{14} \times p_{12}^2 + \\
&(1-d_1')^2 c p_{14}(1-c) \times p_{13}^2 + d_1'(1-d_1')c^2 p_{13}^2 \times p_{14}^2 + \\
&d_1^2(1-d_1)^2 p_{14} c(1-c) \times 2p_{12}p_{13} + c^2 p_{13}^2 d_1^3(1-d_1) \times 2p_{12}p_{14} + \\
&(1-c)^2 d_1(1-d_1)^3 \times 2p_{13}p_{14}
\end{aligned}$$

The LR is then simply obtained by taking the ratio of the probabilities:

$$\begin{aligned}
LR = d_1(1-d_1)^3(1-c)^2 / \big[&d_1'^2 c^2 p_{13} p_{14} \times p_{12}^2 + \\
&(1-d_1')^2 c p_{14}(1-c) \times p_{13}^2 + d_1'(1-d_1')c^2 p_{13}^2 \times p_{14}^2 + \\
&d_1^2(1-d_1)^2 p_{14} c(1-c) \times 2p_{12}p_{13} + c^2 p - 13^2 d_1^3(1-d_1) \times 2p_{12}p_{14} + \\
&(1-c)^2 d_1(1-d_1)^3 \times 2p_{13}p_{14}\big]
\end{aligned} \tag{15}$$

The next section, I verify the results from this simple example, using the *likEvid* function of the Forensim R package.

4 Illustration using the Forensim package

Using the example in section (3), I illustrate how the Forensim R package can be used to calculate likelihood ratios. In what follows, R command lines are encoded in red, while the output in encoded in blue. Appendix B offers more information about how to use R and Forensim.

First, we download the Forensim library, from which the *likEvid* function can be called:

```
> library(forensim)
```

Next, the frequencies for alleles 12, 13, and 14 of the example above need to be defined. The same frequency of 0.15 is set for all three alleles:

```
> #vector of frequencies
> freq<-rep(0.15,3)
> #assign the names to vector freq
> names(freq)<-12:14
```

Using the formula in section (3), I create two functions Hp and Hd, which implement the formula for the probability of the evidence under H_p and under H_d:

```
> Hp<-function(d1,d2,c){d1*(1-d1)^3*(1-c)^2}
```

```
> Hd<-function(d1,d2,c){
 #d1, het, d2 homo
 d2^2*(c*freq['13'])^2*(freq['14']*c)*freq['12']^2 +
 (1-d2)^2*c*freq['14']*(1-c)*freq['13']^2 +
 d2*(1-d2)*(c*freq['13'])^2*freq['14']^2 +
 d1^2*(1-d1)^2*freq['14']*c*(1-c)*2*freq['12']*freq['13'] +
 d1^3 *(1-d1)*(c* freq['13'])^2 * 2*freq['12']*freq['14']+
 (1-c)^2*d1*(1-d1)^3* 2*freq['13']*freq['14']}
```

Calling the functions with parameters d1=0.1 (heterozygote drop-out), d2=0.01 (homozygote drop-out) and a drop-in c=0.05 gives us:

```
> Hd(d1=0.1,d2=0.01,c=0.05)
```

```
0.003120385
```

```
> Hp(d1=0.1,d2=0.01,c=0.05)
```

```
[1] 0.06579225
```

The LR is simply the ratio of these two probabilities:

```
> Hp(d1=0.1,d2=0.01,c=0.05)/Hd(d1=0.1,d2=0.01,c=0.05)
```

```
21.08466
```

The same results are obtained via the *likEvid* function:

```
> # Under Hp
>  likEvid(Repliste=c(13,0,13,14),T=c(13,14),V=0,x=0,theta=0,prDHet=0.1,
  prDHom=0.01,prC=0.05,freq=freq)
[1] 0.06579225
> # Under Hd
> likEvid(Repliste=c(13,0,13,14),T=0,V=c(13,14),x=1,theta=0,prDHet=0.1,
 prDHom=0.01,prC=0.05,freq=freq)
[1] 0.003120385
> # The LR
>  likEvid(Repliste=c(13,0,13,14),T=c(13,14),V=0,x=0,theta=0,prDHet=0.1,
  prDHom=0.01,prC=0.05,freq=freq)/likEvid(Repliste=c(13,0,13,14),T=0,
 V=c(13,14),x=1,theta=0,prDHet=0.1,prDHom=0.01,prC=0.05,freq=freq)
[1] 21.08466
```

Thus, the LR is ≈ 21. This means that the profile is 21 times more likely to be observed if H_p is true than if H_d is true.

5 Additional features

In this section I outline the features that are implemented in *LRmix* and published, but are not described in the document.

5.1 Brief overview of *LRmix*

In *LRmix*, the user inputs two types of data: the DNA profiles of interest (trace and suspects/victims), and the model parameters. First, the user has to upload his data. The first window displayed by *LRmix* when the program is first called, invites the user to proceed in three steps: first load the sample profile (or crime-scene profile), then load the reference profiles; and finally, load the allele frequencies (Figure 1). The required format for these files is specified in the *LRmix* tutorial, in Appendix B.

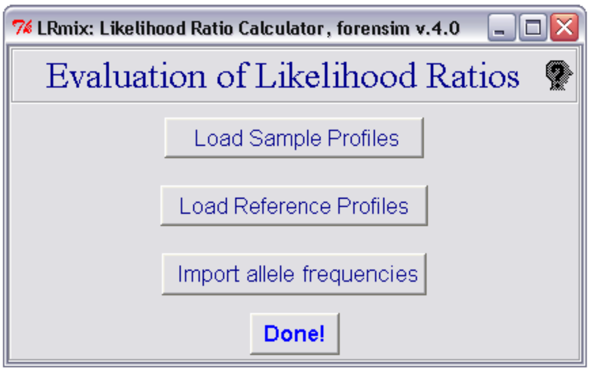
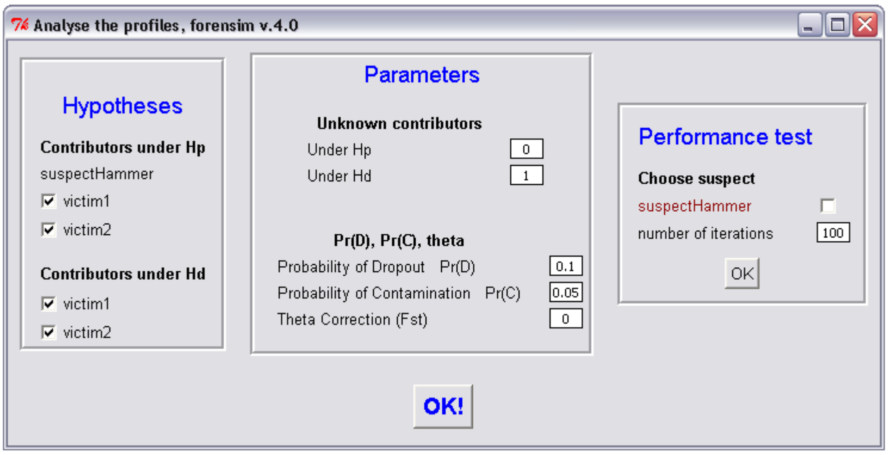

LRmix: Likelihood Ratio Calculator, forensim v.4.0
Evaluation of Likelihood Ratios
Load Sample Profiles
Load Reference Profiles
Import allele frequencies
Done!

Figure 1: First window of LRmix, inviting the user to input the data.

The next step of the analysis, is to choose the individuals contributing under the prosecution (H_p) and the defense hypothesis (H_d), and to input the model parameters (Figure 2).



Analyse the profiles, forensim v.4.0

Hypotheses

Contributors under Hp

suspectHammer

☑ victim1

☑ victim2

Contributors under Hd

☑ victim1

☑ victim2

Parameters

Unknown contributors

Under Hp 0

Under Hd 1

Pr(D), Pr(C), theta

Probability of Dropout Pr(D) 0.1

Probability of Contamination Pr(C) 0.05

Theta Correction (Fst) 0

Performance test

Choose suspect

suspectHammer □

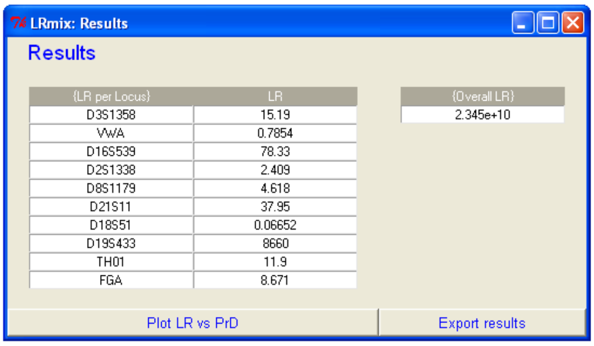
number of iterations 100

Figure 2: Second window of *LRmix*, where the user specified what the hypotheses and the parameters are.

Figure 2 shows that the user has to input several parameters that enable the calculation of the probabilities 2. These parameters are:

- the number of unknown individuals under each hypothesis,
- the probability of drop-out,
- the probability of drop-in,
- theta correction .

Note that at this stage, the allele frequencies have already been imported by the user in the first window of *LRmix*. Using the specified profiles under each hypothesis, and the parameters, *LRmix* calculates the likelihood ratio for this case. The results are then displayed by the program (Figure 3).



LRmix: Results

Results

{LR per Locus}	LR
D3S1358	15.19
VWA	0.7854
D16S539	78.33
D2S1338	2.409
D8S1179	4.618
D21S11	37.95
D18S51	0.06652
D19S433	8660
TH01	11.9
FGA	8.671

{Overall LR}
2.345e+10

Figure 3: LR per locus, for the specified hypotheses and parameters in Figure 2.

5.2 Sensitivity analysis

LRmix offers the possibility to analyze the sensitivity of the likelihood ratio LR to variations of drop-out probability, between zero and one. The sensitivity analysis is carried out by varying the drop-out probability, but keeping all other parameters and hypotheses fixed. For every value of the drop-out probability in the [0.01, 0.99] interval, the product of the LR among all the analyzed loci is taken, and then the final distribution is plotted and displayed (Figure 4).

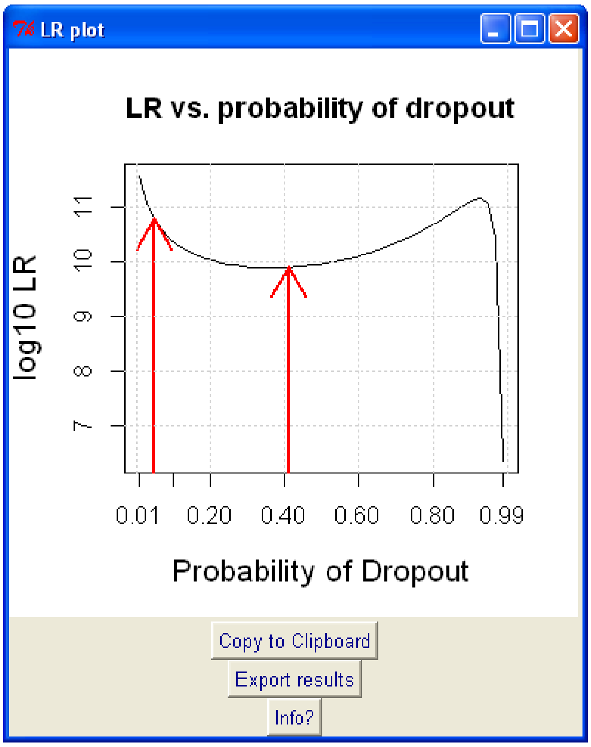


Figure 4: Sensitivity analysis of the LR to the drop-out probability.

5.3 Drop-out estimation

The user has to input all the parameters of the model, the program offers the possibility to derive possible ranges for the probability of drop-out, based on the parameters, and the hypotheses stated by the user. The algorithm behind this estimation procedure is explained in Haned et al. (2012). Based on the total number of alleles observed in the crime scene profile, the algorithm simulates mixtures that have a similar composition to the mixture being observed.

5.4 Performance tests

LRmix offers the possibility to conduct a performance study through a 'non-contributors test' (Gill and Haned, 2013). The principle is simple, for a given set of hypotheses and parameters, the program simulates a large numbers of individual profiles, using the allele frequencies provided by the user. For every simulated individual profile, the LR is calculated, while keeping the parameters and hypotheses fixed, except that the person of interest under H_p (e.g. the suspect) is replaced by the simulated individual. So if 100,000 individuals are simulated, 1000,000 likelihood ratios are obtained by the software. LRmix returns the plot of the empirical distribution function of the yielded distribution (Figure 5).

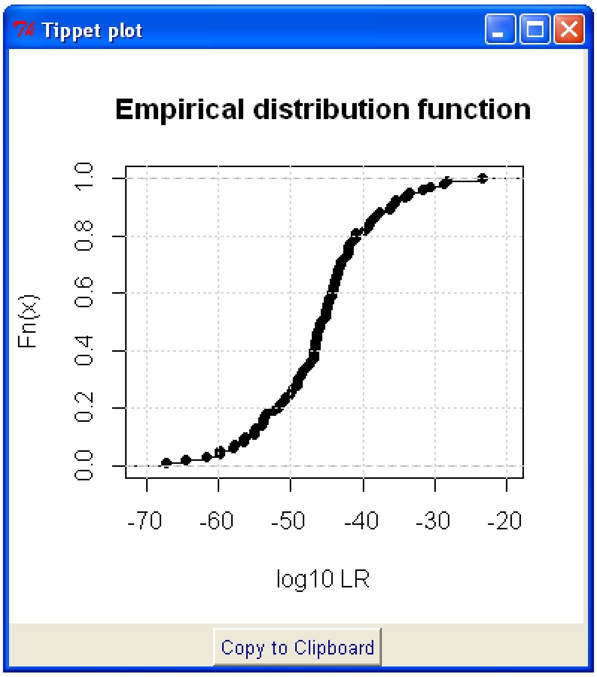


Figure 5: Performance test in *LRmix*.

5.5 Rare alleles

The file of allele frequencies input by the user, has to document all the alleles and their frequencies in the target population. However, some rare alleles can be encountered in human populations, and these may not be documented in the allele frequencies file. When *LRmix* finds a rare alleles, it documents its frequency by $\frac{1}{2 \times N}$, where N is the number of individuals that were sampled in the population study carried out to estimate the allele frequencies. At HBS, $N = 2085$, thus rare alleles are replaced with a frequency of ≈ 0.00024.

Appendix A: *LRmix* diagram

Figure 6 outlines the interaction between the different C functions called by the *likEvid* function.

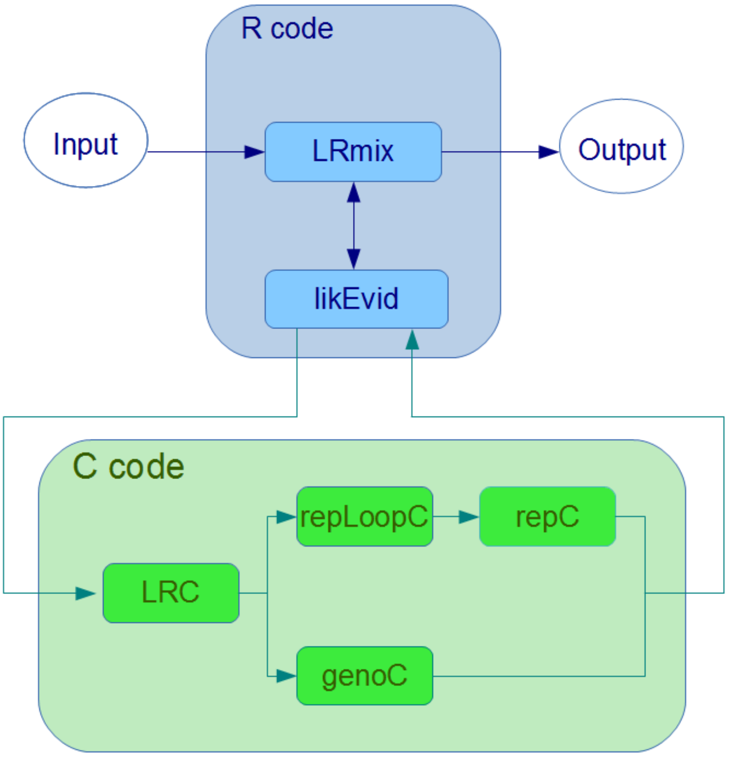


Figure 6: Structure of the *likEvid* function of the *LRmix* program. Elements in green refer to C functions while the blue refers to R functions.

Appendix B: *LRmix* tutorial

The LRmix tutorial details the input/output parameters for the program.

LRmix tutorial, version 4.1

Hinda Haned
Netherlands Forensic Institute, The Hague, The Netherlands
May 2013

Contents

1 What is LRmix?

Forensim is an R-package dedicated to facilitate the statistical interpretation of forensic DNA evidence. It also provides simulation tools made to mimic data from casework. A detailed description of forensim is given in the package tutorial, available from: `http://forensim.r-forge.r-project.org/`. The present tutorial aims at describing one particular module of Forenim, LRmix, which facilitates the calculation of likelihood ratios of LTDNA samples with drop-out, drop-in, any number of contributors and replicates. It is programmed after the model proposed by Curran et al. (2005) and Gill et al. (2007). LRmix is programmed in the R language and offers a user-friendly graphical interface (based on Tcl/Tk) that facilitates the interaction with the program. In order to use LRmix, you first need to install the R software, and then the Forensim package. LRmix and Forensim are available for free, under the GNU General Public licence version ≥ 2. The following section details the installation process. The Hammer case, published in Gill et al. (2007) and available from Forensim website, is used for illustration purposes.

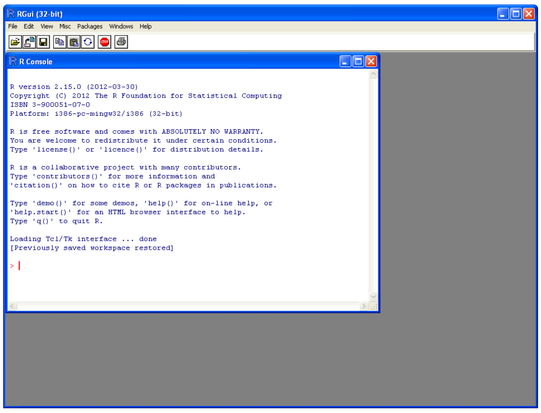
A note on notation A few typographical conventions are used in this tutorial: different colours are used for the R commands and for the R results. A verbatim font is used for `R commands`.

2 Installation

Before we start, make sure you have installed R properly. The R software is available from the Comprehensive R Archive Network (CRAN). Hereafter we explain how the software can be installed.

2.1 Install the R software

- Go to `http://www.cran.r-project.org/`
- Dependent on which operating system you use, click on one of the links:
  - Download R for Linux
  - Download R for MacOSX
  - Download R for Windows
- For Windows, simply follow the link Install R for the first time
- Click the link "Download R 3.0.1 for Windows", run the file and the installation program will start.
- Click on R-3.0.1.exe to install the set-up file
- After installation, a blue colored icon appears on your desktop, click on the icon to launch an R session (Figure 1).



```
R version 2.15.0 (2012-03-30)
Copyright (C) 2012 The R Foundation for Statistical Computing
ISBN 3-900051-07-0
Platform: i386-pc-mingw32/i386 (32-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

Loading Tcl/Tk interface ... done
[Previously saved workspace restored]

> |
```

Figure 1: An R session (Windows)

A note for Mac users The LRmix module has a user graphical interface that relies on the Tcl/Tk language. The Tcl/Tk distribution is provided separately for Mac system, and you will need to download it. Go to: `http://cran.r-project.org/bin/macosx/tools/` and download the tcltk-8.5.5-x11.dmg file and install it on your system.

Once R is downloaded on your system, you have to download Forensim and its dependencies.

2.2 Install the Forensim package

If your computer is connected to the Internet, follow Option 1, otherwise, follow Option 2.

2.2.1 Option 1: install the packages directly from the R environment

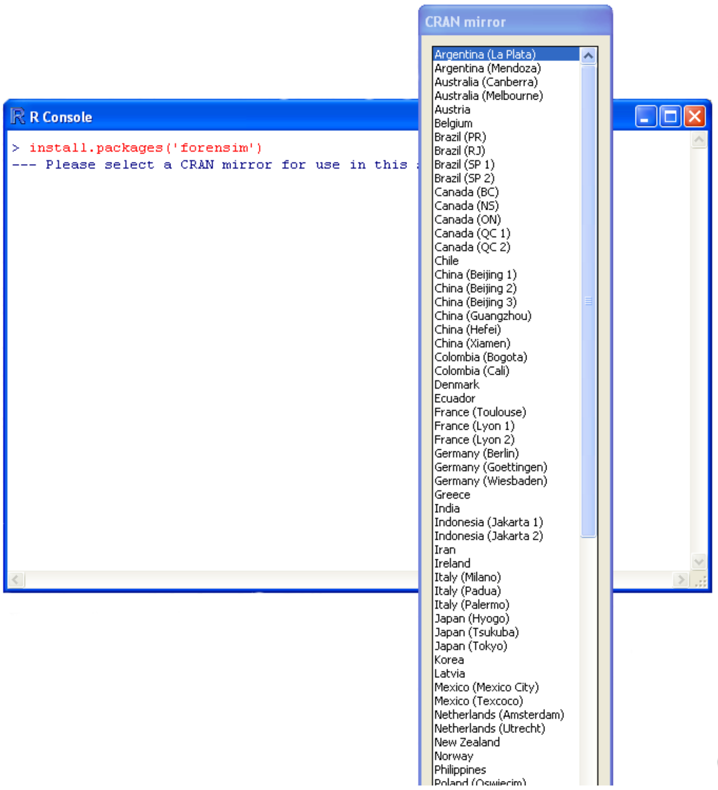
Follow these steps:

1. Open R
2. type the following command in the R console:

   ```
   > install.packages('forensim')
   ```

   This automatically opens a list of 'CRAN mirrors' from which you can install Forensim. You can choose a mirror that it is close to you, thus if you are in France, you can choose Lyon, or if you are in the Netherlands, you can choose Amsterdam (Figure 2).

To make the Forensim package fully functional in R you need some additional packages. Repeat the previous step for the following packages:



```
> install.packages('forensim')
--- Please select a CRAN mirror for use in this :
```

Figure 2: Install packages from the CRAN repository.

1. tcltk2
2. tkrplot.

The forensim package is now ready for use!

2.2.2 Option 2: manual installation

It is possible to install Forensim and its dependencies manually, this is useful if you do not have a connextion to Internet, so you can first download the relevant files and then install them into your computer. Forensim and its dependencies can be found on the CRAN website http://www.cran.r-project.org. In the left menu, under 'Software', click the link 'Packages', then click on 'Table of available packages, sorted by name'. Search for the Forensim package. Click the link with the appropriate file. If you use windows it is the one next to Windows binary, for the Forensim package, it is the forensim_4.1.zip file. Save the file into your working folder. **Do not unzip the file, as this is the required format for R packages**. To make the Forensim package fully functional in R you need some additional packages. Repeat the previous step for the following packages:

1. tcltk2
2. tkrplot.

All downloaded packages now need to be activated in R. Follow these steps:

- Open R
- Install packages using the R function `install.packages`:

```
> install.packages(‘‘forensim_4.1.zip’’, repos=NULL)
```

Do this for every downloaded package. Change the information within the quotation marks according to each package. The forensim package is now ready for use!

Tip for windows users Download all the zip files in the same folder, then click on the Packages tab: install packages from zip files. It is possible to select all the packages at once, and install them at the same time (Figure 3).

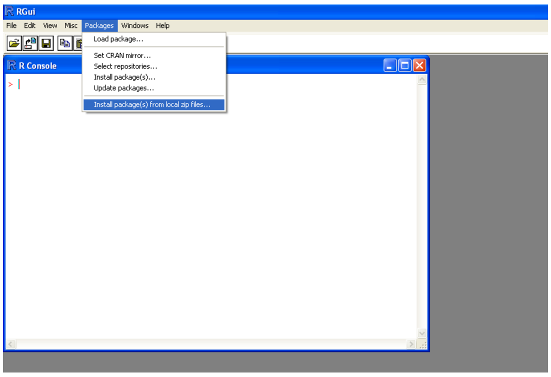


Figure 3: Package installation under Windows.

3 The LRmix module

Forensim implements a number of statistical methods that can be used in the statistical interpretation of evidentiary DNA samples. These methods are documented in the manual of the Forensim package as well as in Haned (2011).

The LRmix module implements a model for the qualitative evaluation of DNA samples. It is a direct implementation of the model described in Curran et al. (2005). The LRmix module allows the calculation of likelihood ratios for different replicates, with any number of contributors, and in case dropout and drop-ins occur. Population substructure is also accounted for using the classical θ correction (Balding and Nichols, 1994).

3.1 Getting started

The first step is to launch R. To do so, simply click on the blue R icon. This should open an R session as shown in Figure 1. The LRmix module is programmed into the R language, and its graphical user interface is programmed in Tcl/Tk.

Load the package forensim to your current R session using the function `library`:

```
> library(forensim)
```

Note! Every time R is closed and opened again a new session starts and the forensim package needs to be loaded again, using the command `library(forensim)`. This command loads the library into your R session, which will enable you to use all the functions available in Forensim. The LRmix module is launched by the LRmixTK command:

```
> LRmixTK()
```

This launches a window that is the main interface to the LRmix module (Figure 4).

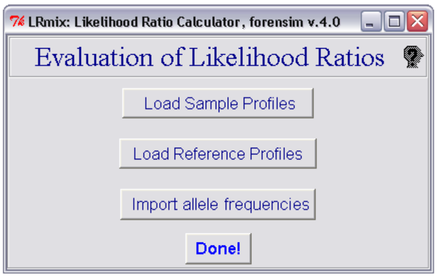

LRmix: Likelihood Ratio Calculator, forensim v.4.0
Evaluation of Likelihood Ratios
Load Sample Profiles
Load Reference Profiles
Import allele frequencies
Done!

Figure 4: LRmix main graphical user interface.

To be able to use the module you have to make sure that your R session is open, but you can minimize the R windows, and continue using the LRmix interface independently. The module has three buttons that correspond to three steps: first, load the sample profiles, second, load the reference profiles, and third, import the allele frequencies.

3.2 Load sample Profiles

This button launches a window that allows you to select the files that contain the profiles of the evidence (Figures 5 and 6).

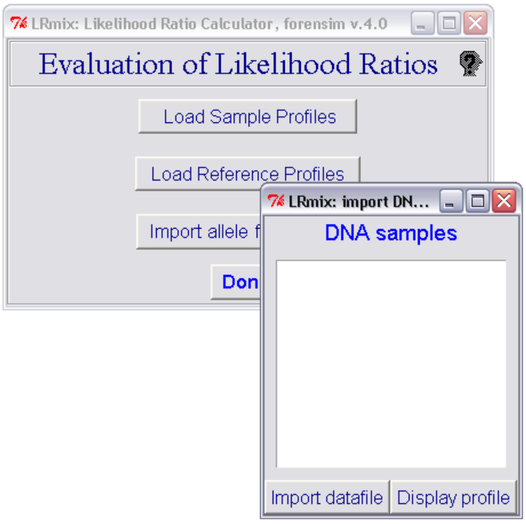
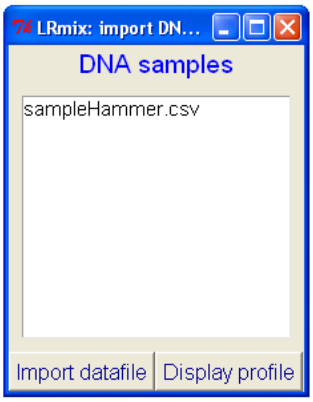


Figure 5: LRmix file upload window for the evidence profile.

The input files can either be text or CSV files. They are typically obtained by exporting your data using genotyping software as text file table. Table 1 gives an example of such file. The names of the replicates must be indicated using the SampleName column. The Marker column indicates the names of the markers. In this example, the user chose to use the data for the first five alleles. In practice, any number of alleles can be provided to the software. Empty or NA columns will be ignored by LRmix.

SampleName	Marker	Allele1	Allele2	Allele3	Allele4
R1	D3S1358	14	16		
R1	VWA	15	16	19	
R1	D16S539	11	13	14	
R1	D2S1338	20	23	24	25
...	...	...	...	...	...
R2	D3S1358	14	16		
R2	VWA	15	16	17	19
R2	D16S539	11	13	14	
R2	D2S1338	20	24	25	

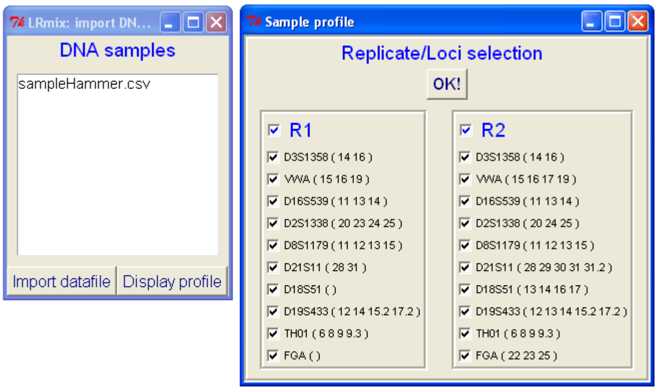
Table 1: Required format for the input file for the evidence profile(s), extract of the Hammer case profiles. Note tha there are two replicates R1 and R2.



sampleHammer.csv

Figure 6: LRmix file upload window for the evidence profile. In this example, the sampleHammer.csv file has been uploaded into LRmix.

Once the file is chosen, the program allows you to see the profiles, and to eventually select the loci as well as the replicates to be analysed (Figure 7). Note that for the purpose of the course, only four replicates can be analysed simultaneously.



sampleHammer.csv

☑ R1

- ☑ D3S1358 (14 16)
- ☑ VWA (15 16 19)
- ☑ D16S539 (11 13 14)
- ☑ D2S1338 (20 23 24 25)
- ☑ D8S1179 (11 12 13 15)
- ☑ D21S11 (28 31)
- ☑ D18S51 ( )
- ☑ D19S433 (12 14 15.2 17.2)
- ☑ TH01 (6 8 9 9.3)
- ☑ FGA ( )

☑ R2

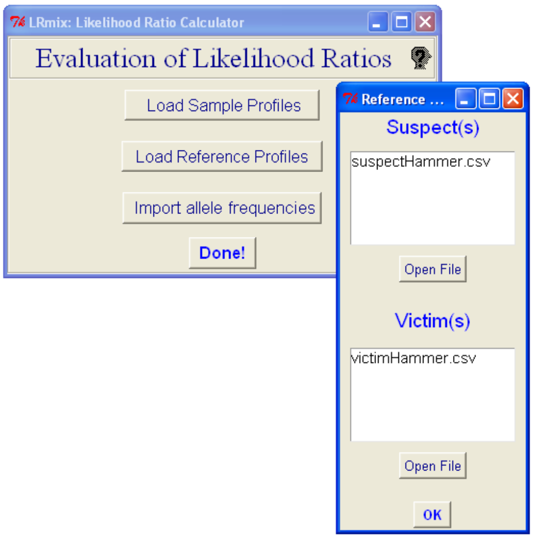
- ☑ D3S1358 (14 16)
- ☑ VWA (15 16 17 19)
- ☑ D16S539 (11 13 14)
- ☑ D2S1338 (20 24 25)
- ☑ D8S1179 (11 12 13 15)
- ☑ D21S11 (28 29 30 31 31.2)
- ☑ D18S51 (13 14 16 17)
- ☑ D19S433 (12 13 14 15.2 17.2)
- ☑ TH01 (6 8 9 9.3)
- ☑ FGA (22 23 25)

Figure 7: DNA profiles from the Hammer case.

The alleles for each replicate are given between brackets for each locus. By default, all loci are included in the calculations, but you can unselect the loci that you want to exclude from the analysis. Note that if there are no alleles at a given locus (see for example at locus FGA in the Hammer case, replicate 1) LRmix displays empty brackets. Once your choice is made, press OK!, this will close the window. At this stage, the program has recorded your preferences.

3.3 Load reference profiles

The next step is to import the reference profiles, namely the suspect and the victim. Press OK when you finish uploading your files (Figure 8).


LRmix: Likelihood Ratio Calculator
Evaluation of Likelihood Ratios
Load Sample Profiles
Load Reference Profiles
Import allele frequencies
Done!
Reference ...
Open File
Open File
OK

Suspect(s)
suspectHammer.csv

Victim(s)
victimHammer.csv

Figure 8: Uploading the reference DNA profiles from the Hammer case.

The selected files should be in the same format as the files used for the sample file (see Table 2). Any number of suspects and victims can be uploaded into the program.

SampleName	Marker	Allele1	Allele2
suspect	D3S1358	14	16
suspect	VWA	15	19
suspect	D16S539	11	14
suspect	D2S1338	24	25
suspect	D8S1179	12	13
suspect	D21S11	28	31
suspect	D18S51	14	17
suspect	D19S433	15.2	17.2
suspect	TH01	9	9.3
suspect	FGA	22	23

Table 2: Required format for the input file for the reference profile(s).

Note that if there two or more suspects, you need to upload a file that contains the profiles of all these suspects. This implies that you want to analyse all the suspects at the same time, if you want to analyse them separately, you need to do separate analyses with different suspect files. It is always compulsory to provide a suspect, but you don't have to provide a victim file. If you have more than one victim, you need to provide the relevant profiles in a single file, where different individuals have different IDs, see the example in Table 3.

SampleName	Marker	Allele1	Allele2
victim1	D3S1358	16	16
victim1	VWA	15	16
victim1	D16S539	13	13
victim1	D2S1338	20	20
victim1	D8S1179	11	15
victim1	D21S11	29	30
victim1	D18S51	17	17
victim1	D19S433	12	14
victim1	TH01	6	8
victim1	FGA	22	25
victim2	D3S1358	15	17
victim2	VWA	16	19
victim2	D16S539	12	13
victim2	D2S1338	18	25
victim2	D8S1179	11	13
victim2	D21S11	29	30
victim2	D18S51	15	17
victim2	D19S433	14	14
victim2	TH01	6	7
victim2	FGA	20	22

Table 3: Required format for the input file for the reference profile(s). The table gives the profiles of two profiled victims, victim 1 and victim2.

3.4 How to import your own allele frequencies

Users can import their allele frequencies. The required format is generally found in forensic journals, and is described below. The files can either be given in CSV format (comma separated values), or in text format (with tab separated values). Table 4 gives an example of such file:

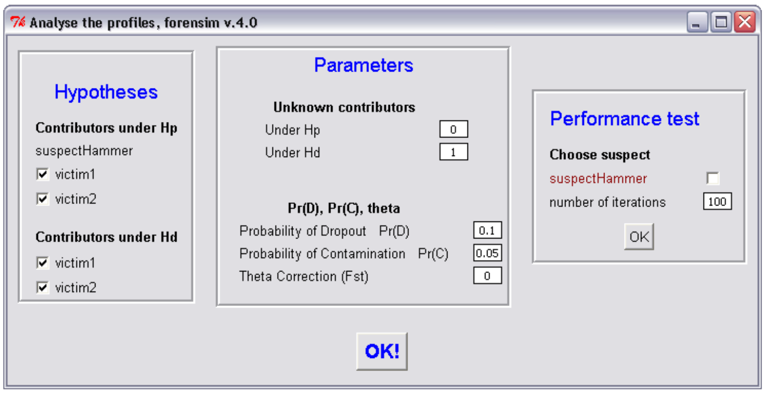
Allele	CSF1PO	FGA	TH01	TPOX	VWA	D3S1358
5			0.002	0.002		
6			0.232	0.002		
7			0.190			
8	0.005		0.084	0.535		
8.1						
...	...	...	...	...	...	...
16.2						
17					0.281	0.215
17.2						
18		0.026			0.200	0.152
18.2						
19		0.053			0.104	0.012
19.2						
20		0.127			0.005	0.002
21		0.185			0.002	
21.2		0.005				
22		0.219				
22.2		0.012				
...	...	...	...	...	...	...

Table 4: Required format for the allele frequencies file. Extract from the Identifiler (Applied Biosystems) allele frequencies (Butler et al., 2003).

The first colum 'Allele' gives the allele lengths (5, 6,...,22.2), the other columns correspond to the loci. The allele frequencies are given in row for each allelic form. Once the file containing the allele frequencies is selected, press the OK! button,

3.5 Analysis

The analysis button launches a window where you have to specify the model parameters.



Hypotheses

Contributors under Hp

suspectHammer

☑ victim1

☑ victim2

Contributors under Hd

☑ victim1

☑ victim2

Parameters

Unknown contributors

Under Hp 0

Under Hd 1

Pr(D), Pr(C), theta

Probability of Dropout Pr(D) 0.1

Probability of Contamination Pr(C) 0.05

Theta Correction (Fst) 0

Performance test

Choose suspect

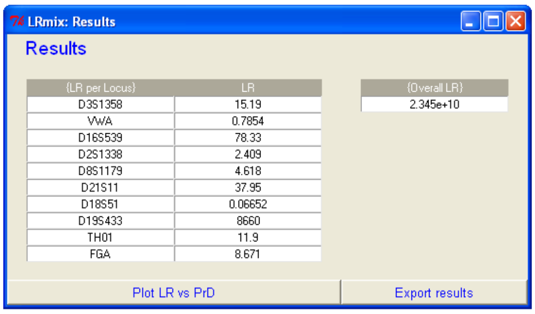
suspectHammer □

number of iterations 100

Figure 9: Analysing the DNA profiles from the Hammer case.

This interface allows you to define the hypotheses that you want to evaluate in the likelihood ratio. By default the model selects the suspect and the victim (if provided) as the contributor(s) under Hp, and the victim(s) as the contributors under Hd. The suspect is automatically non-contributor under Hd. Note that you cannot unselect the suspect under Hp, but you can choose to add the victims as contributors under either Hd or Hp. If you provide more than one suspect, all suspects will be considered under Hp. The unknown numbers of contributors must also be specified under each hypothesis. Finally the probabilities of dropout (PrD) and drop-in must be specified, default values are 0.1 and 0.05 respectively. The theta correction is set to zero by default. The OK button launches the computations and the results are displayed in a separate window. The LR is given per locus and overall loci by multiplying the per-locus values (Figure 10).

Likelihood ratios



Results

{LR per Locus}	LR
D3S1358	15.19
VWA	0.7854
D16S539	78.33
D2S1338	2.409
D8S1179	4.618
D21S11	37.95
D18S51	0.06652
D19S433	8660
TH01	11.9
FGA	8.671

{Overall LR}
2.345e+10

Figure 10: Likelihood ratios of hypotheses Hp and Hd, as specified in Figure 9.

LRmix displays the LRs in a separate window. The results can be saved into a text file (button Export results). The user can also choose to carry on the analysis with a 'sensitivity analysis', this is the exploration of the sensitivity of the likelihood ratios when the dropout probability varies between 0.01 and 0.99. The sensitivity analysis takes longer than the simple evaluation of LRs when a single value of PrD is given. You can follow the progress of the calculations in the R window.

Given the hypotheses and the parameters given in the Analysis window, LRmix tries to find plausible ranges for the probability of dropout following the Monte-Carlo simulation method described in Gill et al. (2007). This qualitative method derives the most plausible ranges of PrD, based on the total number of alleles in the sample profiles. Conditioned on the genotypes specified under each hypothesis, the program simulates a large number of mixtures that have the same composition in alleles than the questioned sample, and looks for the levels of dropout that could have generated a sample with the same number of alleles. Because the method relies on the hypothesised contributors under each hypothesis, the estimation is carried out separately under Hp and under Hd. The minimum and the maximum values obtained across the two analyses, are reported on the sensitivity analysis plot. The results of the sensitivity analysis can be exported as a text file (Export results button), the range of drop-out are given in at the bottom of the file, as the 5% and the 95% percentiles of the empirical distributions of the probabilities of dropout, under Hp and under Hd. These values are also reported on the sensitivity plot (Figure 11).

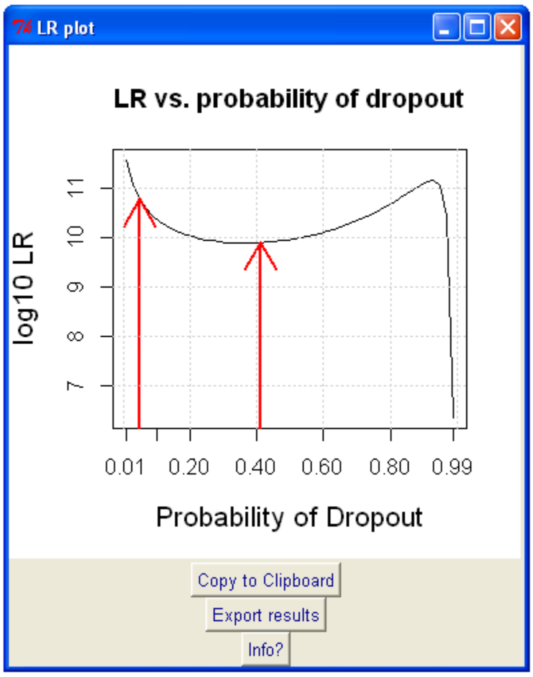


Figure 11: Sensitivity analysis of the LR to variations in the dropout probability. The red arrow correspond to the most plausible ranges for the probabilities of dropout, derived via Monte-Carlo simulations.

3.6 Tippet plots

LRmix offers the possibility to carry out robustness studies, using Tippet plots (Gill et al., 2008). Tippet plots are implemented to enable the evaluation of likelihood ratios when the suspect is substituted with a random man, simulated by randomly drawing alleles from the allele frequencies provided by the user. Only one suspect can be evaluated at a time, thus, if multiple suspects are evaluated, the user has to choose which suspect has to be replaced by a random man. The parameters and hypotheses specified in the Analysis window are the ones used in the LRs calculated in the Tippet plots module.

The user can choose the number of iterations, which corresponds to the number of simulations (i.e. the number of random men) the program has to run in order to build the distribution of LRs. Increasing this number, which is set by default to 100, increases the computation time and may slow down the program. The progress of the computation can be followed in the R console.

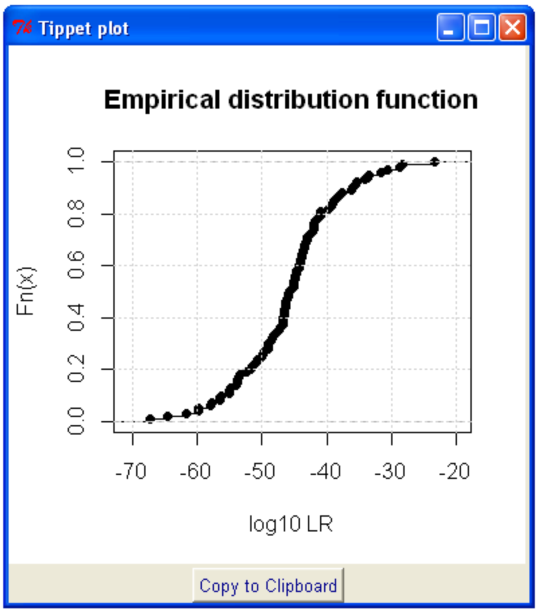


Figure 12: Tippet plot generated by LRmix in the Hammer case.

4 Debugging

4.1 Use this checklist before you load your data into LRmix

If you are using CSV files

- Check that the column names of your files are: SampleName, Marker, Allele1, Allele2, Allele3...
- Check that there are no spaces in the column names
- If you are uploading sample profiles, you can add as many alleles as you want, if you are uploading reference profile, you must only add Allele1 and Allele2
- The field separator must be the comma ','
- The decimal separator must be the dot '.'
- Do not provide the Amel locus
- Make sure that the marker names in your files are consistent with the markers provided in the allele frequencies files

If you are using txt files Please check that:

- Check that the column names of your files are: SampleName, Marker, Allele1, Allele2, Allele3...

- Check that there are no spaces in the column names

- If you are uploading sample profiles, you can add as many alleles as you want, if you are uploading reference profile, you must only add Allele1 and Allele2

- The field separator is the 'tab', as typically obtained from an Excel file.

- The decimal separator must be the dot '.'

Note that it does not matter to LRmix whether there are quotes in your files or not.

4.2 Common errors

- To avoid format errors, refer to the example files given on Forensim website. Typical errors consist in using the comma ',' instead of the dot '.' as a decimal separator in the data files. If you are encountering problems uploading your files into LRmix, open your files under a text editor (e.g. Notepad++) and display the spaces, this may help find the errors.

- Another common error is that the allele provided in the data files are not recovered in the allele frequencies files uploaded by the user, so make sure that in the allele frequencies files, all relevant alleles are listed.

- LRmix notifies the user if it fails to determine the dropout ranges. Keep in mind that this qualitative approach depends on the hypotheses, thus if n alleles are observed and only one contributor is hypothesised under a given hypothesis, the program may fail to derive the ranges, which means that no occurrences of n alleles were found in the Monte-Carlo simulations with the assumed contributor. In this case, reconsider the hypotheses (if relevant) and rerun the porgram.

5 Workshop

Several cases are explored during the practical sessions, and participants are encouraged to analyse their own cases during the course. The Hammer case is provided as an example on Forensim website. The case files are provided both in CSV and txt formats. The case profiles are provided in two zipped folders (in txt and CSV formats). To get the files, simply unzip the folders. It is recommended that you create a working folder for the course, and start R in that folder. Windows users can simply copy the R blue icon in the working folder (shortcut for R), and start R by a double-click. To make sure that R starts in the working folder, right-click on the blue icon, and make sure the "start in" entry is left blank.

Appendix C: relevant literature

Forensic Science International: Genetics 6 (2012) 762–774

Contents lists available at SciVerse ScienceDirect

Forensic Science International: Genetics

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

Exploratory data analysis for the interpretation of low template DNA mixtures

H. Haned [a,*], K. Slooten [a,b], P. Gill [c,d]

[a] Netherlands Forensic Institute, Department of Human Biological traces, The Hague, The Netherlands
[b] VU University Amsterdam, Amsterdam, The Netherlands
[c] Norwegian institute of Public Health, Oslo, Norway
[d] University of Oslo, Norway

ARTICLE INFO

Keywords:
Low template
Mixtures
Drop-out
Drop-in
Likelihood ratios

ABSTRACT

The interpretation of DNA mixtures has proven to be a complex problem in forensic genetics. In particular, low template DNA samples, where alleles can be missing (allele drop-out), or where alleles unrelated to the crime-sample are amplified (allele drop-in), cannot be analysed with classical approaches such as random man not excluded or random match probability. Drop-out, drop-in, stutters and other PCR-related stochastic effects, create uncertainty about the composition of the crime-sample, making it difficult to attach a weight of evidence when (a) reference sample(s) is (are) compared to the crime-sample. In this paper, we use a probabilistic model to calculate likelihood ratios when there is uncertainty about the composition of the crime-sample. This model is essentially exploratory in the sense that it allows the exploration of LRs when two key-parameters, drop-out and drop-in are varied within their plausible ranges of variation. We build on the work of Curran et al. [8], and improve their probabilistic model to allow more flexibility in the way the model parameters are applied. Two new main modifications are brought to their model: (i) different drop-out probabilities can be applied to different contributors, and (ii) different parameters can be used under the prosecution and the defence hypotheses. We illustrate how the LRs can be explored when the drop-out and drop-in parameters are varied, and suggest the use of Monte Carlo simulations to derive plausible ranges for the probability of drop-out. Although the model is suited for both high and low template samples, we illustrate the advantages of the exploratory approach through two DNA mixtures (involving two and at least three individuals) with low template components.

1. Introduction

The interpretation of DNA profiles obtained from low template DNA (LTDNA) samples has proven to be a particularly difficult problem [1,2]. LTDNA samples often comprise DNA from multiple contributors, in different quantities and in limited amounts, which cause PCR-related stochastic effects, such as drop-out (alleles in the sample that fail to PCR-amplify) and drop-in (alleles unassociated with crime-samples that are PCR-amplified) [3,4].

When a reference sample, e.g. from a suspect, is compared to a crime-sample profile, stochastic effects typically create discordances at several loci, making it impossible to use classical methods, such as random man not excluded or the random match probabilities, to report the weight of the DNA evidence. Several models have been proposed in the literature to overcome these issues, but none is in general use or are easily available (free software). They are all anchored in a likelihood ratio (LR) framework, and are traditionally classified in two categories based on the type of information they take into account: (i) continuous models, model the peak heights as continuous variables, and thus account for both the qualitative and quantitative data provided by the electropherograms (epgs) [5–7], and (ii) qualitative models that only use the list of alleles observed in a DNA profile [8–11]. Continuous models consider peak heights to be continuous random variables, and in principle, make the 'best use' of available data. However, when PCR-related stochastic effects such as drop-out and drop-in affect the sample profile (i.e. typical low-template DNA profiles), these models are less efficient because the variability of the signal is exacerbated and the uncertainty in the peak heights is difficult to assess [12]. Comparative studies have not yet been undertaken. Consequently, it is not clear yet how these models behave when applied to low template DNA (LTDNA) in practice, and there is little published on the matter of their robustness when used with these type of samples [7]. Because the utility of peak height information decreases as the amount of template decreases [13], the qualitative and continuous models must eventually converge.

It is possible to evaluate complex mixtures and account for the main stochastic effects related to LTDNA samples, namely,

[*] Corresponding author.
E-mail addresses: h.haned@nfi.minvenj.nl, hi.haned@gmail.com (H. Haned).

1872-4973/$ – see front matter © 2012 Elsevier Ireland Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.fsigen.2012.08.008

drop-out and drop-in, without explicitly modelling the peak heights as continuous variables. This is achieved by adopting a probabilistic model that evaluates likelihood ratios, conditioned on the probability of allelic drop-out and drop-in. Such a model has been described by Curran et al. [8] and Gill et al. [9]. The model enables the computation of LRs for DNA samples with several replicates, which may show drop-out and drop-in alleles, and with multiple contributors. Although this model falls within the qualitative category, it is more accurate to describe it as semi-continuous, since information derived from the epgs is included in the LR to account for uncertainty in the data [8]. In this paper, we improve this model by implementing three major modifications: (i) the probability of drop-out is split per contributor, (ii) the drop-out parameter can vary under the prosecution and the defence hypothesis and (iii) allele masking due to shared alleles between contributors is accounted for. The results of the modified model that we will refer to as the 'SplitDrop' model, are compared to the original 'basic' Curran model, as well as to a newly available software, LikeLTD [14], which also relies on the method described in [8]. The basic model and LikeLTD are essentially the same, but instead of exploring a range of values for these probabilities, likeLTD searches for single drop-out and drop-in estimates which maximise the likelihoods under the defence and the prosecution hypotheses. We illustrate how the SplitDrop model can be applied in practice to typical cases of DNA mixtures reported by the Netherlands Forensic Institute, and the Norwegian Institute of Public Health, and show how it can be employed as an exploratory approach to evaluate the strength of DNA evidence.

2. Theoretical considerations

2.1. The classical likelihood ratio

The classical likelihood ratio (LR) approach consists of a comparison of the likelihood of obtaining the observed DNA profiles given alternative competing hypotheses. The probability of observing the evidence E given hypothesis H, can be computed using probabilistic reasoning. The LR is usually written as:

$$\mathrm{LR} = \frac{Pr(E|H_p)}{Pr(E|H_d)} \tag{1}$$

Fig. 1 shows an example of three epgs of a crime-sample at a single locus. We want to evaluate the following hypotheses, assuming that it has exactly one contributor:

- H_p: the suspect contributed to the sample,
- H_d: an unknown person, unrelated to the suspect, contributed to the sample.

First consider the case where there is sufficient DNA in the sample for the alleles to faithfully reflect the genotype of the donor of the sample. If the observed profile matches that of the person of interest (the suspect in this case), then under H_p, the probability of observing the crime-sample profile is one, since the suspect is assumed to be the contributor. Under H_d, we assume that an unknown person is the contributor of the sample. This person, under the assumption of a single donor trace, needs to match the reference profile. In our example case A, the only 'unknown genotype' that can explain the profile is a heterozygote 9, 10. The probability of observing the evidence, conditioned on an unknown person contributing to the sample is the probability of observing the genotype in the target population. If the target population consists of the general population, unrelated to the offender, with

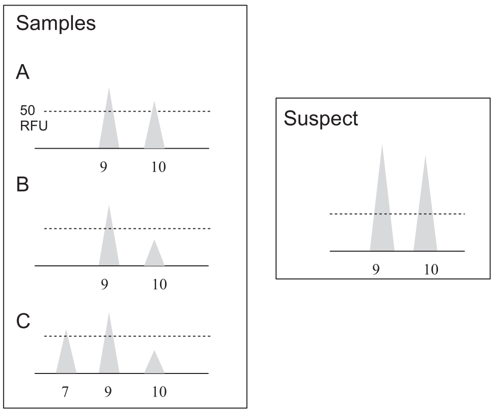


Fig. 1. Single source, single-locus examples. When the suspect is assumed to be the contributor to the samples: case A: no drop-out, no drop-in; case B: one drop-out, no drop-in; case C: one drop-out, one drop-in.

allele frequencies p_9 and p_{10} for alleles 9 and 10, then the LR in Eq. (1) is simply:

$$\mathrm{LR} = \frac{1}{2\,p_9\,p_{10}} \tag{2}$$

Let us assume now that that it is no longer certain that the observed alleles in the sample faithfully reflect the trace donor's genotype, a situation that arises in a low-template crime-sample profile. For example, certain alleles may have failed to PCR-amplify, or there could also be alleles unrelated to the contributor(s) that appear in the sample epg (allele drop-in). In the classical LR approach (unjustly ignoring the uncertainties), the probability $Pr(E|H_p)$ can be zero. This happens if the crime-sample profile cannot be explained by the suspect profile, and one way to deal with this situation is to ignore the problematic locus, and to compute a statistic for loci that do not show drop-out, drop-in or other stochastic effects. However, this approach is biased as it effectively considers evidence to be 'neutral' (LR = 1) and obviously may be very anti-conservative [15]. Models are needed however that are able to fully evaluate any hypothesis.

The Curran et al. [8] model enables unrestricted computation of likelihood ratios when PCR-related stochastic effects such as drop-out and drop-in are possible. In the following section, we illustrate how unrestricted computation of likelihood ratios is enabled when the probability of the evidence is conditioned on the probabilities of allelic drop-out and drop-in.

2.2. Likelihood ratio allowing for drop-out and drop-in

Curran et al. [8] proposed a probabilistic model that enables the evaluation of low template DNA samples. The model is based on simple principles of probabilistic theory, and only makes use of qualitative data.

Suppose n replicates, $R_1, \ldots, R_n$, have been analysed. We want to compute the LR for two competing hypotheses, H_p and H_d, which state the alternative contributors to the crime-sample. To achieve this, we need first to compute $P(R_i|H)$, where H is a hypothesis stating the number of contributors, the genotype of some of these contributors (possibly none), the probabilities of observing each donor's alleles in a replicate (i.e. allele drop-out) and the probability that alleles outside the donor's genotypes are nonetheless observed in a replicate (i.e. allele drop-in). The data may consist of one or more replicates of the DNA sample, and the profiles of the possible contributors, such as the suspect(s) and the victim(s). In probabilistic terms, this is written:

$$Pr(R_1, R_2, ..., R_n|H) \tag{3}$$

Following Curran et al. [8], we consider that all R_i replicates are (conditionally) independent given the drop-out and drop-in rates and the genotypes of all contributors. This means that events of drop-in and drop-out are independent between replicates given this information. Since the replicates are not independent but only conditionally independent when all the contributor's genotypes are known, we can compute the probability of interest as:

$$Pr(R_1, ..., R_i, ..., R_n|H) = \sum_j \left[\prod_i Pr(R_i|U_j, H) \right] Pr(U_j|H) \tag{4}$$

where U_j runs over the possible genotypes of the unknown contributors. Note that the LR evaluation is reduced to two steps:

1. Evaluation of the replicate probabilities: $\prod_i Pr(R_i|U_j, H)$,
2. Evaluation of the genotype probabilities: $Pr(U_j|H)$.

Note that in the original model [8], sets U, V and T, standing respectively for the possible alleles of the unknown contributors, those of known non-contributors and known contributors, are defined. However we will have no need for these notions.

2.2.1. Replicate evaluation

The evaluation of the replicate probabilities $Pr(R_i|H)$ consists of a comparison of the sample profile with the genotypes of the hypothesised contributors. Three possibilities can occur:

-

- the alleles in the replicate are exactly those of the hypothesised contributors.
- some of the alleles of the hypothesised contributors are not recovered in the replicate.
- some alleles in the replicate are not explained by the contributors.

The two latter conditions can be explained by allelic drop-out and drop-in [3]. The last possibility can also be explained by unknown contributors. If positive drop-out and drop-in rates are incorporated into the calculation, then these latter two cases do not lead to zero probability of observing the replicates.

2.2.2. Allelic drop-out

Allele, or locus, drop-out is defined as a signal that is below the limit of detection threshold (LOD). It occurs when either one or both alleles of a heterozygote fail to PCR-amplify. Homozygote drop-out is treated as a special case, since two identical alleles must simultaneously drop-out in order for the allelic drop-out to occur. Because of the dosing effect, homozygote drop-out is less likely than allele drop-out [10]. Drop-out is often considered as a possibility if there is an allele missing in the crime-sample that is visualised in the reference sample. For example, allele 10 in Fig. 1B illustrates this point. In practice, the possibility of allelic drop-out or any other stochastic effect, is evaluated by the expert before any comparison with reference profiles. If there is uncertainty in the crime-sample profile, then we suggest that a computation of the LR needs to incorporate the possibility of drop-out for alleles in reference profiles that are not observed in the crime-sample.

2.2.3. Homozygote vs. heterozygote drop-out

Denoting d the probability of drop-out of a heterozygote allele, and d' that of a homozygote allele, it can be assumed that $d' < d^2$, since alleles amplify independently of each other. In addition, it is clear that if $d = 1$ then $d' = 1$: if a heterozygote allele cannot be sufficiently PCR-amplified, then neither can a homozygote allele. Balding and Buckleton [10] propose $d' = \alpha d^2$ for $0 < \alpha < 1$, but this correction does not imply $d' = 1$ when $d = 1$, hence this correction cannot be used for all d in the [0,1] interval. In reality, it is expected that the d^2 approximation is a lower bound for d' because a homozygote peak is the combination of two coinciding heterozygote peaks, each of which may separately be below the detection threshold while the total is above it. However, if the probability of this event is small enough, then $d' = d^2$ seems to be a reasonable approximation. Hence, throughout the manuscript we use the $d' = d^2$ correction.

2.2.4. Drop-in

We follow the definition of the DNA commission of the International Society for Forensic Genetics [16] and define allele drop-in as an allele that is not associated with the crime-sample, and remains unexplained by the contributors stated under either H_p or H_d. In the original Curran model, a drop-in event of allele i was assigned probability $c \times p_i$, where c is the drop-in probability, and p_i frequency of allele i. When no drop-in occurs, probability $1 - c$ applies to the whole replicate [8]. In the probabilistic model provided in Supplementary Section I, we write the LR as a function of drop-out and drop-in, and we explicitly show how the formulae derived by Curran et al. [8] can be viewed as approximations to the ones obtained within this model. The exact calculations for the single-locus examples in Fig. 1 when drop-in is considered are also given in Supplementary Section II. In these examples, the differences between the likelihood ratios obtained with the Curran method and the likelihood ratios obtained with our model are very small, provided that the drop-in parameter c is small. Therefore, in the following, we retain the simple Curran approximation, but we consider the drop-in parameter c to be small, at most in the order of 0.05. This value corresponds to one expected drop-in allele per 20 loci, which may coincide with the real contributor's alleles.

2.2.5. Genotype evaluation

When unknown contributors are involved in the conditioning of a given hypothesis, the probability that they have each of the possible genotypes must be evaluated. A key step in the model presented here, is the enumeration of all possible genotypes for unknown contributors. Given that alleles can drop-out and/or drop-in, this leads to a much larger number of potential genotypes for the unknown contributors, compared to the case where genotypes of the unknown are constrained to the visualised alleles observed in the sample epgs [8]. In theory, unknown contributors can have any genotype, including those not supported by peak height data. Gill et al. [9] introduced the Q designation to refer to those alleles that can be of any type, except those already observed in the crime-sample, across all replicates. The set of unknown genotypes, U, can include alleles that have been observed, and alleles that have not been observed because drop-out has occurred (Q alleles).

When drop-out is deemed possible, then the unknown contributors could have genotypes that have completely or partially dropped out from the crime-sample. Formally, an allele Q stands for an allele outside of the set S that contains the alleles observed in the crime-sample. Based on the Q designation, the unknown contributors could either be homozygotes or heterzoygotes. Such homozygote genotypes are denoted QQ with corresponding genotypic frequency of p_{QQ}, and heterozygotes genotypes are denoted QQ', with genotypic frequency $p_{QQ'}$, where $Q \neq Q'$. In the example given in Fig. 1A, the observed alleles in the crime-sample are 9 and 10, hence, the virtual Q allele can be any allele at this hypothetical locus, except 9 and 10, i.e. $S = \{9, 10\}$. In practice, the evaluation of genotype probabilities can be carried out for each putative genotype using the simple product rule, hence we have in this example: $p_{QQ} = \sum_{i \notin S} p_i^2$, where p_i is the frequency of the ith allele. And for heterozygotes, denoted QQ', the genotypic probability is: $p_{QQ'} = \sum_{i \notin S;\ j \notin S} 2\, p_i\, p_j$.

The frequency of heterozygote genotypes where only one allele has dropped out is denoted p_{iQ} (with i an allele in S). The frequency of such genotypes is given by: $p_{iQ} = 2p_i \sum_{j \notin S} p_j$. In the example Fig. 1A, the frequency of the heterozygote genotype p_{9Q} is: $p_{9Q} = 2p_9 \sum_{j \notin S} p_j$. Note that $p_{QQ} + p_{QQ'}$ is the frequency of genotypes that do not contain alleles in S, as a consequence we have: $p_{QQ} + p_{QQ'} = (1 - \sum_{i \in S} p_i)^2$.

2.2.6. Mathematical justification of the probabilistic model

The simple formalization, based on Curran's formula [8] has two advantages, first, it greatly simplifies the calculations, through the use of different sets of alleles, and second, it provides a simple tool to facilitate the use of the model to analyse real casework (as demonstrated in Section 3). We provide a mathematical proof and general formula for our probabilistic model, in order to provide justification for our calculus in Supplementary Section I. In the following examples, we first illustrate some of the model features by demonstrating the calculation of likelihood ratios for single-locus examples compiled in Fig. 1, then we use multi-loci examples to further illustrate the principles of the approach.

2.3. Single locus examples

2.3.1. Case B

In the single-locus example in Fig. 1B, the crime-sample profile is 9. Under H_p, the suspect 9, 10 is assumed to have contributed to the sample. Since only allele 9 is recovered in the sample drop-out must be invoked in order to explain the sample profile. Hence we record one drop-out and one allele that has not dropped out. Under H_d, an unknown person is conditioned to have contributed to the sample, in theory this unknown genotype could be either 9, 9 or 9Q, with Q being any allele in the population at the considered locus, except allele 9. If drop-in is deemed possible, two additional genotypes, QQ and QQ', with $Q \neq Q'$ which are not concordant with the sample profile, have to be considered. Note in this example, in addition to assigning a drop-out probability d when a drop-out occurs, it is necessary to assign the non-dropout probability, $1 - d$, when an allele is observed in the sample. It is not strictly necessary to consider drop-in to explain the profile in this example, however, we evaluate the effect of incorporating the drop-in parameter in this case. Table 1 summarises the formulae.

Table 1
Replicate and genotype probabilities for the single locus case, when drop-in is either considered as impossible ($c = 0$) or possible ($c \neq 0$).

	Contributors	Replicate prob.		Genotype prob.
		No drop-in	Drop-in	
Under H_p	9,10	$d(1-d)$	$d(1-d)(1-c)$	1
Under H_d	9,9	$(1-d')$	$(1-d')(1-c)$	p_9^2
	9,Q	$d(1-d)$	$d(1-d)(1-c)$	p_{9Q}
	QQ	0	$d'cp_9$	p_{QQ}
	QQ'	0	d^2cp_9	$p_{QQ'}$

Table 2
Likelihood ratios for case B.

Classical approach	LR = 0
Drop-out, no drop-in	$\mathrm{LR} = \frac{d(1-d)}{(1-d')\,p_9^2 + d(1-d)\,p_{9Q}}$
Drop-out and drop-in	$\mathrm{LR} = \frac{d(1-d)(1-c)}{(1-c)\left[(1-d')\,p_9^2 + d(1-d)\,p_{9Q}\right] + c\,p_9\left[d'\,p_{QQ} + d^2\,p_{QQ'}\right]}$

Table 2 summarises the likelihood ratios calculated for case B, and Fig. 2 displays the sensitivity analysis relative to the probability of drop-out d.

Although we need to consider drop-out in order to include the suspect, if the drop-out probability is very low ($d \approx 0$) then the heterozygote genotype of the suspect is no longer supported under H_p. If drop-in is accounted for, then the LR decreases when drop-in increases from 0 to 0.05. This can also be explained intuitively: if drop-in is more likely, then the suspect alleles that are recovered in the sample are more likely to be drop-in alleles rather than alleles from the suspect, the evidence will then tend to be weighted more towards the defence hypothesis of exclusion. The probabilities of the data under H_p and under H_d are shown in the electronic Supplementary Figs. 4 and 5.

2.3.2. Case C

Case C (Fig. 1C) differs from case B in that we need to assume an allele drop-in to fully explain the profile in the epg under the prosecution hypothesis. Under H_p, the suspect contributed to the sample, in this case, there is one drop-out (allele 10) and one drop-in (allele 7). Under H_d, an unknown person contributed to the sample. The unknown genotype can either be a combination of alleles in the sample, or Q alleles (any allele except 7 and 9) that have dropped out. Table 3 gives the calculation details.

In case C, we need both drop-in and drop-out to explain the profiles under H_p. Using 0.01 as the drop-in probability c, yields an LR in favor of the defence hypothesis H_d, for most of the range of variation of d ($d < 0.98$, Fig. 3). Indeed, small values of c penalise the H_p proposition, since drop-in is needed to include the suspect (Supplementary Fig. 6). Under H_d, increasing c increases the weight

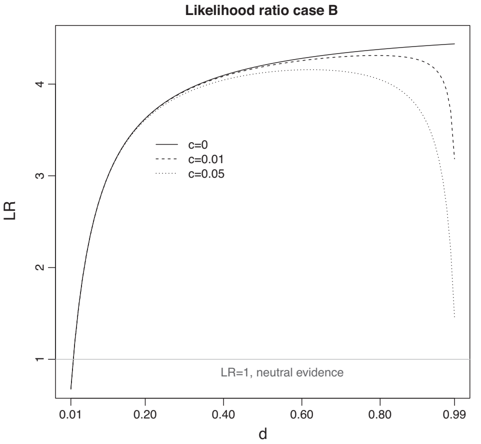


Fig. 2. Sensitivity plot of the LR to the drop-out probability d and to the drop-in probability c. d varies in [0.01, 0.99], while c varies in $\{0, 0.01, 0.05\}$. The crime-sample is 9, and the suspect is 9, 10. Frequency of allele 9 is taken as 0.11. The homozygote drop-out is taken as $d' = d^2$.

Table 3
Replicate and genotype probabilities for case C.

	Contributors	Replicate prob.	Genotype prob.
Under H_p	9,10	$d(1-d)p_7c$	1
Under H_d	7,9	$(1-d)^2(1-c)$	$2p_7p_9$
	7,7	$(1-d')p_9c$	p_7^2
	9,9	$(1-d')cp_7$	p_9^2
	7Q	$(1-d)dp_9c$	p_{7Q}
	9Q	$(1-d)dp_7c$	p_{9Q}
	QQ	$d'p_7p_9c^2$	p_{QQ}
	QQ′	$d^2p_7p_9c^2$	$p_{QQ'}$

of those genotypes that are not recovered in the sample, these are rendered even more likely when drop-out probability tends towards 1 (Supplementary Fig. 7).

In the following sections, we turn to multi-loci profiles, comprising two or three contributors. Intuitive interpretations such as those presented above, become more complex to assess when multiple loci are analysed simultaneously. We exemplify the calculations for few loci and use graphical representation to summarise multi-dimensional sensitivity analyses.

3. Casework examples

Unbalanced DNA mixtures, with one major component and one or more minor components are commonly encountered in casework. In these types of cases, it is typically observed that the major contributor is often a complete profile, whereas the minor contributor is partial. Consequently, it is appropriate to consider different drop-out probabilities for alleles from different contributors, since alleles from minor and major contributors are unlikely to drop-out with the same probabilities. We modify the model of [8,9], to allow application of different drop-out rates per contributor. This translates, for a vast majority of cases, into associating a low drop-out probability with the major contributor(s), and a higher drop-out probability for the minor contributor(s). We first evaluate a two-person mixture, where the major contributor is identified as the victim. We then analyse a mixture of at least three individuals, with one major contributor corresponding to a known victim, and possibly two minor contributors. For both cases, a sensitivity analysis of the likelihood ratios to the probability of drop-out is carried out. In order to assess the impact of the improvements to the model of Curran et al., we compare the likelihood ratios obtained with two models:

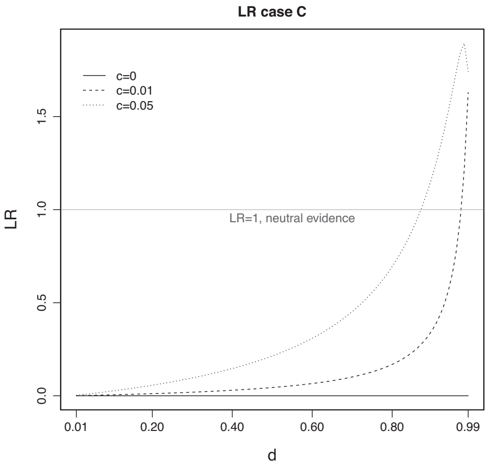


Fig. 3. Sensitivity plot for case C. Probability d varies in [0.01, 0.99], while c varies in $\{0, 0.01, 0.05\}$. The crime stain is 7,9, and the suspect is 9, 10. Frequency of alleles 7 and 9 are taken as 0.32 and 0.11 respectively. The homozygote drop-out is taken as $d' = d^2$.

-

- the ‘basic model’: the basic model involves a single drop-out probability applied to all hypothesised contributors and with no correction for allele sharing,
- the ‘SplitDrop’ model: different drop-out probabilities are applied to different contributors, and allele sharing is accounted for.

Both models were programmed into the the Forensim package for the R software [17,18], and all the calculations and the plots presented in the paper are generated using Forensim.

3.1. Example 1: a two-person mixture

In this example (Table 4), the crime-sample profile is an unbalanced DNA mixture analysed with the Applied Biosystems Next Generation Multiplex (NGM). The expert assessed the profile as most likely consisting of two contributors - one major contributor and one minor. The major contributor corresponds to an identified victim (from which the sample was collected), and a suspected individual is detained by the police. Consequently, we apply two drop-out parameters: d_1 and d_2 for the major (victim) and the minor contributor respectively. The probabilities for alleles from homozygous profiles are denoted d'_1 and d'_2 respectively. All parameters are kept constant under H_p and H_d. The following hypotheses are evaluated for this case:

-

- Under H_p: the suspect and the victim are the contributors.
- Under H_d: the victim and one unknown, unrelated to the suspect, are the contributors.

3.1.1. One drop-out, locus FGA

	FGA
Sample	20,23
Victim	20,23
Suspect	23,24

Under H_p

We assume that the suspect and the victim left the trace. While the victim's alleles are recovered in the sample, one of the suspect's alleles has dropped-out. A consideration of drop-out is used to explain the sample profile under H_p.

Since suspect and victim share allele 23, and because we have no information about the number of copies of allele 23 in the sample, we have to consider the possibility that there is either one or two copies of this allele in the sample. Table 5 shows how the probabilities of drop-out are assigned to each possible genotype.

By summing the probabilities for all possibilities (column four of Table 5), we derive $Pr(E|H_p) = (1-d_1)(1-d_1d_2)d_2$.

Under H_d

The victim and one unknown person contributed to the sample. The victim's alleles are recovered in the sample and have thus not dropped out. The unknown contributor can have any combination of alleles among those observed in the sample, but could also have alleles that have dropped out (Q alleles). If we ignore the possibility of drop-in, then, the genotype of the unknown contributor and the

Table 4
Case 1: A two-person mixture analysed with the NGM system (one replicate). The table displays alleles one to four (A1,...,A4) and their corresponding peak heights (H1,...,H4).

Marker	A1	A2	A3	A4	H1	H2	H3	H4
D10S1248	13	14	15		140	82	3016	
vWA	16	18	19		1562	1778	193	
D16S539	9	11			2981	67		
D2S1338	19	22	23	24	56	973	973	91
D8S1179	10	11	15		110	107	2900	
D21S11	28	29	30	31.2	94	1268	201	1116
D18S51	11	16	17		1094	1102	194	
D22S1045	11	15	16		1557	1378	163	
D19S433	13	14	15		69	80	2933	
TH01	7	9	9.3		119	97	3174	
D2S441	11	14	15	16	1388	90	119	1177
D3S1358	15	16	17		1279	1486	71	
D1S1656	11	12	13		1404	209	1278	
D12S391	17	18.3	20		1012	110	928	
FGA	20	23			1092	1295		

victim profile must explain the sample profile. Table 6 details the calculations of the probability of the evidence under H_d. We correct for allele sharing directly similarly to Table 5.

By summing over all probabilities in Table 6, the LR for locus FGA is finally:

$$\mathrm{LR}_{FGA} = (1-d_1)(1-d_1d_2)d_2/$$
$$\{(1-d_1d_2)^2 2\,p_{20}\,p_{23} + (1-d_1d_2)(1-d_1)d_2(\,p_{23Q} + p_{20Q})+$$
$$(1-d_1d'_2)(1-d_1)(\,p_{20}^2 + p_{23}^2) + (1-d_1)^2(d'_2\,p_{QQ} + d_2^2\,p_{QQ'})\}$$

3.1.2. *No drop-out, locus D8S1179*

	D8S1179
Sample	10,11,15
Victim	15,15
Suspect	10,11

Under H_p

The suspect and the victim are the contributors to the trace. Since alleles 10, 11 and 15 are observed in the sample, no drop-out has occurred. Alleles 10 and 11 are recovered in the sample and this has probability $(1-d_2)^2$, and the homozygous genotype 15,15 is also observed in the profile, and this has probability: $(1-d'_1)$. The replicate probability is then:

$Pr(E|Hp) = (1-d'_1)(1-d_2)^2$, when drop-in is not considered a possibility (c = 0).

Under H_d

The major contributor, (15,15) and an unknown person left the trace. Since drop-in is ignored here, there is only one possible genotype for the unknown contributor and that is 10,11. The replicate probability in this case is thus the same than under H_p. The genotype probability is that of genotype 10,11 in the target population: $2p_{10}p_{11}$. This leads to: $Pr(E|H_d) = (1-d'_1)(1-d_2)^2 2\,p_{10}\,p_{11}$. The LR at locus D8S1179 is thus:

$$\mathrm{LR}_{D8S1179} = \frac{(1-d'_1)(1-d_2)^2}{(1-d'_1)(1-d_2)^2 2\,p_{10}\,p_{11}} = \frac{1}{2\,p_{10}\,p_{11}}$$

This result could have been deduced directly, since we can completely deconvolve the mixture at this locus.

3.1.3. *No drop-out, one shared allele: locus D1S1656*

	D1S1656
Sample	11,12,13
Victim	11,13
Suspect	11,12

Under H_p

The alleles in the sample can be explained completely by the profiles of the suspect and the victim. Note that at this locus, suspect and victim share one allele (11). No drop-out events are recorded at this locus, still, we do not know the exact number of copies of allele 11 in the sample. Similarly with locus FGA (under H_p Table 6), we must ensure that all possibilities are accounted for, since there could either be two copies or one copy of the allele present. The former corresponds to no drop-out at all of allele 11. The latter corresponds to one drop-out, either from the victim, or the suspect. Summing the probabilities across each possibility yields the formula shown in Table 7.

Under H_d

The major contributor and an unknown person left the trace, Table 7 displays all the possible genotypes for the unknown contributor for which the replicate probability is non-zero.

The likelihood ratio at locus D1S1656 is finally:

$$\mathrm{LR}_{D1S1656} = (1-d_1d_2)(1-d_1)(1-d_2)/$$
$$\Big\{(1-d_1)^2(1-d'_2)\,p_{12}^2 + 2\,p_{12}(1-d_1)(1-d_1d_2)(1-d_2)(\,p_{13} + p_{11})+$$
$$\Big(1-d_1)^2(1-d_2)d_2\,p_{12Q}\Big\}$$

Fig. 4 shows the sensitivity analysis of the likelihood ratios for loci FGA, D8S1179, and D1S1656, and for the overall LR for the 15 NGM loci.

Overall, regardless of the drop-out values, the LRs obtained from both analyses ranged between ≈10^{13} and ≈10^{14}. Higher LRs were obtained with the SplitDrop model ($d_1 \neq d_2$) for drop-out values lower than 0.76, however the difference between the two models is always below one unit (on $\log_{10}$ scale). Essentially, both models provide very similar results, illustrating that it is uneccesary to specify different drop-out rates per contributor. The LR was stable over the reasonable range of variation of the drop-out of the minor contributor and the value of ≈10^{13} (d = 0.01), could be used as a lower bound to assist reporting officers to assess the strength of the evidence. Note that this bound was obtained under the assumption that the drop-out probabilities were the

Table 5
Replicate probability for locus FGA, under the prosecution hypothesis H_p. All drop-out events relate to heterozygote genotypes.

Genotype		Dropouts	Non-dropouts	Probability
23 dropped out from the suspect				
Victim	20,23	0	2	$(1-d_1)^2 \times d_2^2$
Suspect	23,24	2	0	
23 dropped out from the victim				
Victim	20,23	1	1	$(1-d_1)d_1 \times (1-d_2)d_2$
Suspect	23,24	1	1	
23 did not drop out				
Victim	20,23	0	2	$(1-d_1)^2 \times (1-d_2)d_2$
Suspect	23,24	0	1	

same under H_p and H_d. We now investigate the effect of varying the drop-out probabilities between these two hypotheses.

Recall that we evaluated the hypotheses: H_p, the victim and the suspect were the source of the crime sample vs. H_d, the victim, and one unknown were the source of the sample. We have previously applied the drop-out probability $d_1 = 0$ to the victim, and the same probability to the suspect under H_p and the unknown under H_d, which varies in [0.01,0.99] (Fig. 4). Given the level of the peak heights in the sample profile, it seems reasonable to apply the same drop-out probability to the suspect and the unknown under H_p and H_d. However, there may be debate over the use of the same parameters under different propositions, and recently authors have suggested that the likelihoods of H_d and H_p should be evaluated separately, using different drop-out parameters [14]. The SplitDrop model allows this flexibility, since different drop-out probabilities can be applied to different hypothesised contributors. As a consequence, another dimension can be added to the sensitivity analysis in Fig. 4, in order to explore the effect of varying drop-out parameters between the suspect under H_p, and the alternative unknown contributor under H_d. If $d_{suspect} \neq d_{unknown}$ we need to compute the LRs for all possible combinations of values for ($d_{suspect}$, $d_{unknown}$). For each value of $d_{suspect}$ in [0.01,0.99], we calculate the LRs when d_{unkown} varies in the whole [0.01,0.99] range. If k values are explored in [0.01,0.99], then the sensitivity analysis yields a $k \times k$ matrix, where the columns are the values of $d_{suspect}$ varying in [0.01,0.99] and the rows are $d_{unknown}$ also varying in [0.01,0.99]. One way to represent these data, is to use a heatmap plot, as shown in Fig. 5.

Varying the drop-out probabilities separately under H_p and H_d, results in a huge variation in the LRs, which range from 10^{-20} ($d_{suspect} = 0.99$, $d_{unknown} = 0.01$) to 10^{40} ($d_{suspect} = 0.01$, $d_{unknown} = 0.99$). It is essential therefore to report an interval of LRs, corresponding to the most reasonable values of the probabilities of drop-out. The procedure followed in this study is basically the same as that described in Gill et al. [9], except that the drop-out estimates are determined separately for H_p and H_d. We further describe this procedure as follows.

Table 6
Replicate and genotype probabilities under H_d for locus FGA. d_1 and d_2 are the heterozygote drop-out probabilities for the victim and the suspect, respectively. d'_1 and d'_2 are the corresponding homozygote probabilities.

Victim; unknown	Replicate probability	Genotype probability
20,23; 20,23	$(1-d_1d_2)^2$	$2p_{20}p_{23}$
20,23; 20,20	$(1-d_1d'_2)(1-d_1)$	p_{20}^2
20,23; 23,23	$(1-d_1d'_2)(1-d_1)$	p_{23}^2
20,23; 20Q	$(1-d_1d_2)(1-d_1)d_2$	p_{20Q}
20,23; 23Q	$(1-d_1d_2)(1-d_1)d_2$	p_{23Q}
20,23; QQ	$(1-d_1)^2d'_2$	p_{QQ}
20,23; QQ'	$(1-d_1)^2d_2^2$	$p_{QQ'}$

Table 7
Replicate and genotype probabilities under H_d, locus D1S1656. The replicate probability under H_p can be deduced from the combination (11,13; 11,12), with a genotype probablity of one.

Victim; Unknown	Replicate probability	Genotype probability
11,13; 12,12	$(1-d_1)^2(1-d'_2)$	p_{12}^2
11,13; 12,13	$(1-d_1)(1-d_1d_2)(1-d_2)$	$2p_{12}p_{13}$
11,13; 11,12	$(1-d_1)(1-d_1d_2)(1-d_2)$	$2p_{11}p_{12}$
11,13; 12Q	$(1-d_1)^2(1-d_2)d_2$	p_{12Q}

Defining plausible ranges for the probability of drop-out via Monte Carlo simulation

It is possible to define plausible ranges for the probability of drop-out based on the hypothesised number of contributors (and their profiles) and the observed number of alleles in the DNA profile [9,19]. Gill et al. [9] suggested a maximum likelihood approach to estimate probabilities of drop-out that maximise the probability of observing x alleles in the sample. This can be carried out either via exact calculations, or, via approximations using Monte Carlo simulations (see for example Gill et al. [9], Appendix A). Another possibility, is the use of experimental data sets. For example, Perez et al. [19] followed an experimental approach, where they assessed the levels of drop-out, based on large sets of DNA mixtures obtained in different conditions. These methods derive estimates for the drop-out probabilities, which do not rely on the questioned sample, but rather on a population of samples from which the crime-sample could have originated from. In this paper, we suggest a different approach that relies on the crime-sample itself, rather than simulated or experimental samples: the crime-sample is re-simulated n times, at each iteration, a random sampling of the alleles is applied, in order to select the alleles that will drop-out from the sample. Since the true probabilities of drop-out are unknown, different drop-out probabilities, ranging from zero to one, are applied. The rationale behind this procedure is to explore the range of probabilities of drop-out that could have led to the crime-sample of interest. The simulations ultimately yield an empirical distribution for the probabilities of drop-out, and the most plausible values for these probabilities are the ones that lead to the same number of alleles that are observed in the crime-sample under investigation. The advantages of such approach, is that the ranges of the drop-out probability can be evaluated separately under H_p and H_d, and that we avoid reporting values of drop-out that are supported by one hypothesis but not by its alternative. Monte-Carlo simulations are carried out to estimate the outcomes of $d_{suspect}$ and $d_{unknown}$ in the range [0.01, 0.99].

Using the *simumix* function of the Forensim package, 10,000 two-person mixtures were simulated, where each mixture was composed of the genotypes of the victim and the suspect under H_p, and the victim and one unknown under H_d. The genotype of the unknown was randomly generated by sampling alleles at their proportions in the target population (Dutch NGM allele frequencies). Once a mixture was generated, its alleles were sampled for drop-out, using two drop-out probabilities: $d_{suspect}$ and $d_{unknown}$, which vary in [0.01,0.99] (note that the probability for the victim was set to zero for reasons previously explained). To continue the example, each allele from the suspect has a probability $1 - d_{suspect}$ of being recovered in the mixture, and a probability of $d_{suspect}$ of dropping out. The quantity of interest is the total number of alleles in the sample profile, and this number is computed separately under each hypothesis. Once the simulation procedure is carried out, two distributions (one for H_d, and one for H_p) of the number of alleles observed in the profile are obtained. Computing the 5% and 95% percentiles of these distributions yields plausible ranges for the probabilities of drop-out under each hypothesis. In the two person-mixture (Table 4), 46 alleles were recovered in the

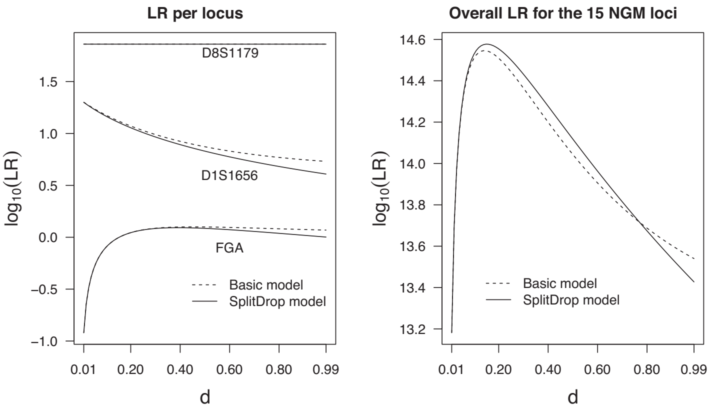


Fig. 4. Sensitivity analysis of the LR to the probability of drop-out for case 1. All LR values are reported on a $\log_{10}$ scale. Continuous curves correspond to the SplitDrop model: LRs are generated with the major contributor having a different drop-out probability than the other hypothesised contributors (suspect under H_p and unknown under H_d), which corresponds to $d_1 = 0$ and d_2 varying between 0.01 and 0.99. Dashed curves correspond to the basic model: LRs are computed with $d_1 = d_2$. The homozygote drop-out is taken as $d' = d^2$.

crime-scene sample. The following results were obtained with the aforementioned simulation procedure:

	Drop-out probabilities (percentiles)	
	5%	95%
H_p	0.03	0.26
H_d	0.01	0.22

Superimposing these values on the heatmap Fig. 5 narrowed down the range of variation for the likelihood ratios to the [10^8, 10^{11}] interval.

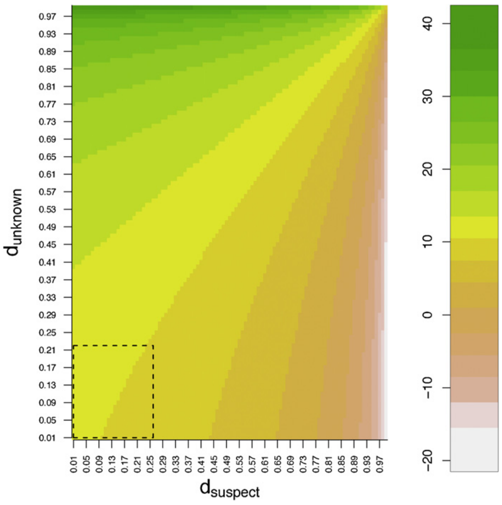


Fig. 5. Heatmap for the two-person mixture example depicting the sensitivity of the likelihood ratio to variations in the drop-out probabilities under H_p ($d_{suspect}$) and under H_d ($d_{unknown}$). The rectangle in the lower-left corner corresponds to the most likely ranges for the drop-out probability. LRs are given on a $\log_{10}$ scale.

3.2. Example 2: a three-person mixture

The analysis of LTDNA mixtures is further complicated where there is an indication that more than two people could have contributed to the sample. In complex mixtures, there may be more than one 'minor component'. A typical example would comprise clear major and clear minor contributors. Major contributors are non-problematic since they can all be assigned a probability $d \approx 0$ based on an assessment of their peak heights. Conversely all minor contributors are characterised by a drop-out probability strictly greater than zero and lower than one. We propose to analyse such profiles by combining the low-template components – usually, they will not be distinguishable – into a 'minor contributors category'. This category will often consist of alleles belonging to unknown individuals, who contributed low amounts of DNA to the sample. The sensitivity analysis consists of varying the drop-out probabilities of the combined minor contributors. To demonstrate the 'categorical' approach we use a more complex mixture where at least three people have contributed to the sample (Table 8). The case involved a female victim and a male suspect detained by the police. A blood stain recovered from the victim was analysed, and three replicates were obtained from the same DNA extract that were amplified and characterised by the SGM+ kit (Applied Biosystems). In the following example, we applied the model to three replicates simultaneously, and we compare the LRs obtained by the Curran model (basic model) to the modified model, where the drop-out probability is split between contributors (SplitDrop model).

The following hypotheses were evaluated:

–

- H_p: the victim, the suspect and one unknown person, unrelated to them, have contributed to the mixture.
- H_d: the victim and two unknowns have contributed to the mixture.

Note that for the three-person mixture, H_d assumes that there maybe more than one unknown person contributing to the crime-sample. As a consequence, in comparison to the first example, there are more genotypic combinations to be considered

Table 8
Case 2: A three-person mixture analysed in three-replicates. The table displays alleles one to five (A1,...,A5) and their corresponding peak heights (H1,...,H5).

Marker	A1	A2	A3	A4	A5	H1	H2	H3	H4	H5
vWA	12	18	19			125	1059	913		
	12	18	19			199	1531	1245		
	12	18	19			64	1115	987		
TH01	7	9.3				139	1326			
	7	8	9.3			317	84	2125		
	7	9.3				65	1119			
FGA	20	23	25	26		800	85	67	616	
	20	22	23	25	26	1302	72	171	107	1295
	20	25	26			707	55	741		
D8S1179	10	14				1101	895			
	10	14				1647	1288			
	10	14				1078	679			
D3S1358	14	15	16	17	18	75	1143	353	303	1034
	14	15	16	17	18	58	1708	452	213	1311
	15	16	17	18		1097	348	319	1099	
D2S1338	18	24	25			132	658	475		
	18	24	25			217	984	964		
	18	24	25			121	838	602		
D21S11	28	29	31	33.2		1305	72	1492	227	
	28	30	31	33.2		1626	109	1557	148	
	28	29	31	33.2		1059	59	1083	62	
D19S433	14	15	16			925	339	914		
	14	15	16			1462	611	1019		
	14	15	16			727	421	863		
D18S51	12	14	17	18		68	982	806	80	
	12	14	17			124	1361	1105		
	12	14	17			150	837	724		
D16S539	11	12	13	14		822	863	109	228	
	11	12	13	14		1201	1192	203	269	
	11	12	13	14		849	954	95	129	

for the unknown contributors, which further complicates the computation of the LR. For a three-person mixture with more than one unknown contributor, the computational details become complex to derive by hand and the use of a software is desirable. Thus, the computational details are left out in this section and further emphasis is put on the sensitivity analyses. As shown by Eq. (4), multiple replicates are evaluated simultaneously in this model. Once the genotypes of the unknown individuals are derived, conditioned on the data observed in the three replicates, the replicate probabilities can be multiplied together. However, this is different from simply deriving the LR for each replicate separately, and taking the product, since n replicates are simultaneously conditioned on the genotypes of the hypothesised contributors. From Table 8, it is reasonable to evaluate a unique drop-out probability under H_p, since both suspect and one extra unknown contributor can be both assigned to the 'minor contributors category'. The same rationale applies under H_d, where the hypothesised two unknowns can be assigned to the same minor contributors category. In the sensitivity analysis carried out Fig. 6, the LRs obtained using a single drop-out parameter for all hypothesised contributors (basic model: $d_{victim} = d_{suspect} = d_{unknown}$), is compared to LRs obtained when the probability of drop-out for the major contributor is set to zero (SplitDrop model: $d_{victim} = 0$, $d_{suspect} = d_{unknown}$). Similar ranges of variation are obtained with the two models as the LRs vary from $\approx 10^7$ to $\approx 10^{10}$.

Unlike the first case, the overall LR is rather unstable for this three-person mixture, and reporting the lower bound of 10^7 implies a very high drop-out probability that is not supported by the data. Indeed, a probability of drop-out of 0.97 for the suspect implies that most of his alleles have dropped-out, which is not supported by the peak height data. Consequently, it is desirable in this case to obtain bounds on the probability of drop-out. This will enable reporting ranges for the LRs that are actually supported by the observed data, under both H_d and H_p. The plausible ranges for the drop-out probabilities are evaluated via the same simulation procedure explained in Section 3.1. Since all contributors in this case either belong to the 'minor' or the 'major' category, the simulation procedure used to determine plausible ranges for drop-outs can be carried out with a single parameter for both the suspect and the unknown under H_p and another single parameter is used

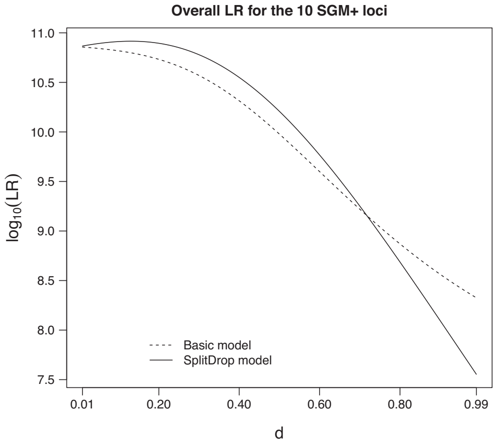


Fig. 6. Sensitivity analysis of the LR to the probability of drop-out for case 2. All LR values are reported on a $\log_{10}$ scale. The continuous curve corresponds to LRs where the major contributor does not have the same drop-out probability as the other hypothesised contributors (suspect under H_p and unknown under H_d), which corresponds to $d_{victim} = 0$ and $d_{suspect} = d_{unknown}$ varying between 0.01 and 0.99, dashed curves correspond to LRs computed with probabilities $d_{victim} = d_{unknown} = d_{suspect}$. The homozygote drop-out is taken as $d' = d^2$.

for the two unknowns under H_d. Under both hypotheses, the drop-out probability for the victim is set to zero. An average of 33 alleles are observed across the three replicates for case 2 (Table 8). The following percentiles for the probability of drop-out were obtained under H_p and under H_d:

	Drop-out probabilities (percentiles)	
	5%	95%
H_p	0.22	0.58
H_d	0.17	0.58

We superimpose these values on the heatmap of the likelihood ratios obtained when varying drop-out probabilities under H_p and under H_d (Fig. 7).

Note that varying drop-out probabilities under H_p and H_d led to great variation of the LR. The LR ranges between $\approx 10^{-56}$ and $\approx 10^{74}$, the minimum value for the LR was obtained for a drop-out probability under H_p of 0.99 and 0.27 under H_d. The maximum value is obtained for a drop-out probability under H_p of 0.25 and under H_d, $d = 0.99$. Note that the LRs given by the SplitDrop model (Fig. 6) can be recovered on the diagonal of the matrix plotted in the heatmap in Fig. 7. Applying the 5–95% percentiles of the (empirical) drop-out distribution obtained via the simulation procedure, narrowed the range of variation of the LRs down to the [10^7, 10^{13}] interval. A lower bound of 10^7 could therefore be reported for this case, when different drop-out levels are considered under H_p and H_d.

4. Alternative models

Likelihood-ratios, closely rely on a probabilistic model that defines how the probability of the data, conditioned on a given hypothesis, is computed. Since there is no single true model to evaluate the likelihood of a given hypothesis, there is no single true LR. However, the robustness and efficiency of different models can be discussed and tested on large sets of complex LTDNA cases [20]. Another way to assess the robustness of the yielded LRs, is to compare the LRs from different models. We compared the LRs given by the SplitDrop model to the LRs obtained via the LikeLTD program provided by Balding [14]. This program generates LRs based on the maximum likelihood principle: using a simulation annealing algorithm, the program (written in the R language) determines the drop-out probabilities, among other parameters, which maximise the probability of the evidence per locus. The search for the maximum likelihood estimates is carried out separately for H_p and for H_d (analogous to the SplitDrop model principle except that the final LR is the ratio of the maximum likelihoods). In the following sections, the LRs per locus for the two- and the three-person mixtures, were evaluated and compared using the two models. The possibility of allele drop-in was ignored in both models, and the θ correction was set to 0.02.

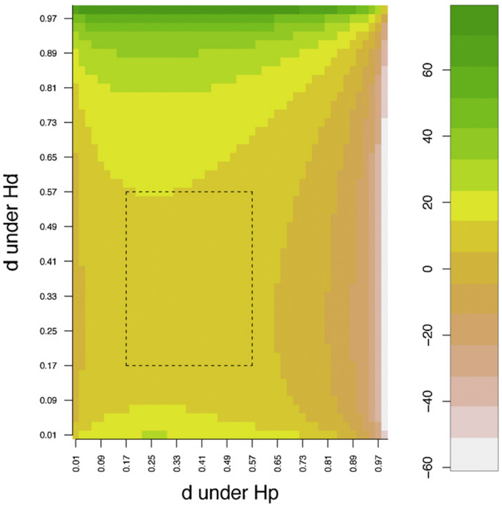


Fig. 7. Heatmap for the three-person mixture depicting the sensitivity analyses when different drop-out probabilities are explored under H_p and under H_d in the [0.01,0.99] interval, using the categorical model (major/minor). The rectangle drawn in dashed lines shows the bound of the $\log_{10}$ LR when the ranges for the drop-out probabilities are reported.

4.1. Two-person mixture case

The likelihood ratios obtained via the exploratory approach of the SplitDrop model, described previously, were compared to the LR values obtained via the LikeLTD model. The drop-out values used to carry out the comparison between the models is shown in Table 9, and the results of the comparison of the LRs are shown in Table 10.

The two models rely on different statistical approaches, but we show that the LRs are comparable and this gives confidence that the results are not ‘over-dependent’ upon the modelling assumptions.

4.2. Three-person mixture

The same comparison between the exploratory SplitDrop model and LikeLTD was carried out for the three-person mixture case previously analysed in Section 3. The results of the comparison of the LRs are shown in Table 12 and the drop-out probabilities used in each model are given in Table 11.

Table 12.

The difference between the LRs obtained with both models was more pronounced for the three-person mixture, with a difference of three units (on the $\log_{10}$ scale). Lower LRs were obtained with the exploratory approach, due to the difference in the estimated drop-out probabilities. In fact, greater differences between the models were to be expected for the three-person mixture than for the two-person mixture. Indeed, in the exploratory model, the suspect was a minor contributor. In the LikeLTD model, the drop-out probabilities estimates are simulated in order to maximise the probability of the data under the defence and prosecution hypotheses, we can therefore expect lower drop-out values under H_p, because most of the suspect's alleles are recovered in the crime-sample. The comparison shows that the LRs can differ greatly between models. The differences are due to the underlying methods used to generate the likelihoods of the evaluated hypotheses. Although the probabilistic models underlying the SplitDrop and the LikeLTD models are conceptually similar, their implementation differs in several ways. While the LRs obtained with the LikeLTD model correspond to ratio of the maximum

Table 9
Summary of the estimates for the drop-out probabilities obtained by the SplitDrop model and the LikeLTD model.

	SplitDrop model	LikeLTD model
d_{victim}	0	0
$d_{suspect}$	$\{0.03, 0.26\}$	0.072
$d_{unknown}$	$\{0.01, 0.22\}$	0.052

Table 10
Likelihood ratios obtained via different models. The values of LRs obtained via the SplitDrop model are displayed (on the linear scale) next to the LRs yielded by the maximum likelihood method of the LikeLTD program (version 4.1). Note that $LR_{5\%}$ (resp. $LR_{95\%}$) correspond to the LRs computed with the 5% (resp. 95%) quantile of d estimated via the Monte Carlo simulation procedure. The θ value is taken in both models as 0.02.

	SplitDrop model		LikeLTD model
	$LR_{5\%}$	$LR_{95\%}$	LR
D10S1248	5.19	4.87	7.70
vWA	8.46	6.74	4.53
D16S539	4.50	3.10	6.97
D2S1338	29.18	27.35	12.96
D8S1179	53.45	50.11	50.58
D21S11	10.80	10.13	12.51
D18S51	14.99	10.75	7.67
D22S1045	2.01	2.15	3.82
D19S433	6.33	5.93	6.95
TH01	18.27	17.13	16.82
D2S441	26	24.38	23.73
D3S1358	3.62	3.10	5.11
D1S1656	14.24	8.56	13.95
D12S391	1.37	7.35	8.08
FGA	0.28	1.18	1.48
Overall LR	5.88×10^{12}	1.85×10^{13}	2.43×10^{13}

likelihoods obtained under each hypothesis, the SplitDrop model (implemented in the LRmix module, see Section 5) considers ranges of drop-out probabilities that do not necessarily correspond to the highest likelihoods for each of the evaluated hypotheses. Another difference resides in the correction used for the homozygote drop-out probability. Balding [14] uses a correction based on the logistic model of Tvedebrink et al. [21], the homozygote drop-out is defined as $d' = d \times 2^{-4.35}/(1 + d \times (2^{-4.35} - 1))$, while in our model, $d' = d^2$. Our correction gives higher homozygote drop-out probabilities, which can lead to important differences in the LRs, for instance if the suspect is a homozygote and a drop-out is observed under H_p, for a drop-out probability of 0.58 (95% percentile suspect drop-out, Table 11), our correction gives a homozygote drop-out probability of 0.33, against 0.063 for LikeLTD. Also note that LikeLTD corrects for size bias, while the SplitDrop model does not.

Table 11
Summary of the estimates for the drop-out probabilities obtained by the SplitDrop and the LikeLTD models, under H_p and H_d. For the LikeLTD model, estimates are given per replicate and per contributor, for the SplitDrop model, the same drop-out probabilities are used for all three replicates.

Contributor	SplitDrop model	LikeLTD model
Drop-out probabilities under H_p		
Victim	0	$d_{Rep1} = 0.0001$
	–	$d_{Rep2} = 0.0001$
	–	$d_{Rep3} = 0.001$
Suspect	$\{0.22, 0.58\}$	$d_{Rep1} = 0.0085$
	–	$d_{Rep1} = 0.0081$
	–	$d_{Rep1} = 0.024$
Unknown	$\{0.22, 0.58\}$	$d_{Rep1} = 0.55$
	–	$d_{Rep2} = 0.54$
	–	$d_{Rep3} = 0.78$
Drop-out probabilities under H_d		
Victim	0	$d_{Rep1} = 0.011$
	–	$d_{Rep2} = 0.008$
	–	$d_{Rep3} = 0.034$
Unknown 1	$\{0.17, 0.58\}$	$d_{Rep1} = 0.045$
	–	$d_{Rep2} = 0.034$
	–	$d_{Rep3} = 0.0015$
Unknown 2	$\{0.17, 0.58\}$	$d_{Rep1} = 0.74$
	–	$d_{Rep2} = 0.68$
	–	$d_{Rep3} = 0.89$

Table 12
Likelihood ratios for the three-person mixture obtained via different models. The values of LRs obtained via the exploratory approach are displayed (on the linear scale) next to the LR yielded by the maximum likelihood method of the LikeLTD program (version 4.1). Note that $LR_{5\%}$ (resp. $LR_{95\%}$) correspond to the LRs computed with the 5% (resp. 95%) quantile of d estimated via the Monte Carlo simulation procedure. The θ value is taken in both models as 0.02.

	SplitDrop model		LikeLTD model
	$LR_{5\%}$	$LR_{95\%}$	LR
vWA	30	25.26	44.96
TH01	3.05	1.64	6.47
FGA	3.48	1.13	10.88
D8S1179	7.44	4.31	5.82
D3S1358	2.09	2.69	14.25
D2S1338	8.51	6.74	4.75
D21S11	8.41	3.11	32.92
D19S433	3.73	5.45	2.85
D18S51	14.38	11.22	9.94
D16S539	12.87	20.03	36.08
Overall LR	2.44×10^{8}	1.39×10^{7}	4.8×10^{10}

Although a given model could have appealing properties for a particular application, when compared to another model, it is important to stress that there is no method better than another. However, the choice of a model must be guided by its performances and its robustness using real cases [13]. These principles are important to consider, and they will be expanded and explained further in future work.

5. Implementation and availability

The basic and the SplitDrop models were written in the R language. The heatmap plots were generated with the *myImagePlot* R function available from http://www.phaget4.org/R/image_matrix.html. The basic model has been made available within the LRmix module of the Forensim package (version 3.0) [18,22] for the R statistical software [17]. LRmix is a user-friendly graphical interface, which allows the computation of LRs for any number of contributors (known and unknown). Users can import their data files, for the crime-sample profile and the hypothesised contributors, as well as their own allele frequencies. The sensitivity plots can also be performed and the results stored in Excel files. Forensim was used to generate all the figures in this paper. The underlying R code of the LRmix module has been checked extensively but comes with no guarantee of accuracy under the GNU general public license (version ≥ 2).

6. Discussion

In this paper, we extend the model of Curran et al. [8] to complex DNA mixtures, with different levels of contributions (major and minor components). By incorporating the probabilities of drop-out and drop-in in the evaluation of likelihood ratios, this probabilistic model offers a flexible framework in which the uncertainty that lies within the data can be explored. Whereas the drop-in probability is straightforward to estimate from negative controls [9], the drop-out parameter is more difficult to assess. The possibility of investigating the effect of the drop-out parameter on the LRs is thus an appealing feature of the model. The SplitDrop model allows greater flexibility in the evaluation of likelihood ratios, and the uncertainty in the data can be explored by means of graphical sensitivity analyses. A significant difference, with Curran et al.'s model, is the possibility of applying different probabilities of drop-out to different contributors. The comparison of a single generalised drop-out parameter to one parameter by contributor did not lead to important differences in the likelihood ratios. However, varying the probabilities of drop-out between the two hypotheses, H_p and H_d, led to dramatic changes in the LRs, this shows that the chosen values for the probabilities of drop-out are a very critical part in the implementation of the model in casework. We suggested the use of Monte Carlo simulations to derive plausible ranges for the probability of drop-out, which narrows down the range of variation of the likelihood ratios.

Our model allows in principle, the use of any value for the drop-in probability. However, the approximations used in this paper (and those derived in the mathematical formalization in Supplementary Section I) assume that the drop-in levels are low, possibly in the order of one to five percent. In practice, we suggest the levels of drop-in should be estimated from negative controls [16]. Internal validation of the NGM STR system (Applied Biosystems) at the Netherlands Forensic Institute showed that drop-in levels were extremely low even with LTDNA samples. Observations at the higher range of 5% were only obtained when techniques were employed to increase the sensitivity of analysis (such as increasing the number of PCR cycles). Consequently, if several extra (unexplained) alleles are observed, then it is preferred that an extra contributor should be considered in the formulation of the hypotheses [9]. The drop-in parameter is not intended to deal with multiple drop-in events.

Ideally, any probabilistic model used for the calculation of likelihood ratios, should rely on all available data. In principle, models that use peak heights extract more information from the epgs than the qualitative model discussed in this paper, and it is expected that they would yield higher likelihood ratios. However, quantitative models rely on distributional assumptions and parameters that depend on the STR system in use. Hence a necessary step before the implementation of such models, is the analysis of the stochastic variation in single-source and mixed DNA samples, such studies were carried out for example by Bright et al. [23] and Perez et al. [19]. The results from these studies give valuable insight into the nature of variation of the peak heights, but each laboratory has to conduct these experiments in order to characterise the variability of the peak heights according to the laboratory own standards and practices. Ultimately, the performance of these models have to be assessed on both high and low-template DNA samples. For the latter, robustness studies are still needed to assess the performance of quantitative models when the stochastic effects are exacerbated.

Another aspect that is important to consider in all models, is the homozygote drop-out. The correction for homozygote drop-out used throughout the paper follows a probabilistic logic, the drop-out of a homozygote is equivalent to the drop-out of two heterozygote alleles. However, as pointed out by [10], homozygote drop-out can occur, and still a signal can be detected. Tvedebrink et al. [24] have discussed the issue of the α correction and the authors show that $\alpha = 0.5$ underestimates the homozygote probability of drop-out. However, in this study the true homozygote distribution of drop-out, which was determined via simulation, depends on the underlying simulation model used to generate the peak heights intensities. Although the $\alpha = 0.5$ correction seems inappropriate, it is not clear how these results can be extended to casework samples without the analysis of large number of cases. In fact, a key issue in modelling homozygote drop-out, is understanding the relationship between the number of template DNA molecules and the peak heights intensity: how does the peak signal vary if double the number of molecules are detected? The answer depends on the machinery used and the standards in place. PCR simulation models such as those described in Gill et al. [13] and Weusten and Herbergs [25] can help understand homozygote drop-out, and serve as a fist step in designing experiments that will lead to establishing amore satisfying relationship than $d' = d^2$.

To conclude, we would like to point out that likelihood-ratios rely on the model used to generated them, and we showed that an alternative model (LikeLTD) leads to different results with complex cases. Although different models can prove useful in different contexts, it is important for each forensic laboratory to derive their own guidelines and justify their choice of particular approach over another [16]. The exploratory approach of the basic and the SplitDrop models discussed in this paper can serve such a purpose. Although the two models yielded similar results on the two mixtures analysed in this paper, we advocate the use of the SplitDrop model, because it is more powerful and flexible, and enables a thorough exploration of the cases under each hypothesis. Establishing this basic qualitative model allows evaluative comparison with more complex models such those developed by Cowell et al. [26] and Perlin et al. [7]. The next step will be to carry out comparative studies and to discover limitations of modelling assumptions by carrying out extensive test of robustness. Availability in an open-source platform ensures transparency of the underlying code and guarantees the possibility to all users to test the robustness of the model.

Conflict of interest statement

None declared.

Acknowledgements

This study was partly supported by a grant from the Netherlands Genomics Initiative/Netherlands Organization for Scientific Research (NWO) within the framework of the Forensic Genomics Consortium Netherlands. Peter Gill has received funding support from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 285487. The authors thank Ankie van Gorp (NFI) and Jerien Klaver (NFI) for helpful guidance on casework data and two anonymous reviewers who helped improve the paper.

Appendix A. Supplementary Data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.fsigen.2012.08.008.

References

[1] J. Buckleton, P. Gill, Low Copy Number in Forensic DNA Evidence Interpretation, CRC Press, pp. 275–297.
[2] C. Benschop, H. Haned, T. Blaei, A. Meulenbroek, T. Sijen, Assessment of mock cases involving complex low template DNA mixtures: a descriptive study, Forensic Sci. Int. Genet. (2012), http://dx.doi.org/10.1016/j.fsigen.2012.04.007.
[3] P. Gill, J. Whitaker, C. Flaxman, N. Brown, J. Buckleton, An investigation of the rigor of interpretation rules for STRs derived from less than 100 pg of DNA, Forensic Sci. Int. 112 (1) (2000) 17–40.
[4] J. Butler, Forensic DNA Typing, Academic Press London, 2001.
[5] R.G. Cowell, S.L. Lauritzen, J. Mortera, Identification and separation of DNA mixtures using peak area information, Forensic Sci. Int. 166 (2007) 28–34.
[6] J.M. Curran, A MCMC method for resolving two person mixtures, Sci. Justice 48 (2008) 168–177.
[7] M. Perlin, M. Legler, C. Spencer, J. Smith, W. Allan, J. Belrose, B. Duceman, Validating Trueallele® DNA mixture interpretation, J. Forensic Sci. 56 (6) (2011) 1430–1447.
[8] J.M. Curran, P. Gill, M.R. Bill, Interpretation of repeat measurement DNA evidence allowing for multiple contributors and population substructure, Forensic Sci. Int. 148 (2005) 47–53.
[9] P. Gill, A. Kirkham, J.M. Curran, LoComatioN: a software tool for the analysis of low copy number DNA profiles, Forensic Sci. Int. 166 (2-3) (2007) 128–138.
[10] D.J. Balding, J. Buckleton, Interpreting low template DNA profiles, Forensic Sci. Int. Genet. 4 (2009) 1–10.
[11] H. Kelly, J.A. Bright, J.M. Curran, J. Buckleton, The interpretation of low level DNA mixtures, Forensic Sci. Int. Genet. 6 (2) (2012) 191–197.
[12] R.G. Cowell, Validation of an STR peak area model, Forensic Sci. Int. 3 (2009) 193–199.

[13] P. Gill, J.M. Curran, K. Elliot, A graphical simulation model of the entire DNA process associated with the analysis of short tandem repeat loci, Nucleic Acids Res. 33 (2) (2005) 632–643.

[14] D. J. Balding, LikeLTD: Likelihoods for Low-template DNA Profiles, May 2012, https://sites.google.com/site/baldingstatisticalgenetics/.

[15] J.M. Curran, J. Buckleton, Inclusion probabilities and dropout, J. Forensic Sci. 55 (2010) 1171–1173.

[16] P. Gill, C.H. Brenner, J.S. Buckleton, A. Carracedo, M. Krawczak, W.R. Mayer, N. Morling, M. Prinz, P.M. Schneider, B.S. Weir, DNA commission of the International Society of Forensic Genetics: Recommendations on the interpretation of mixtures, Forensic Sci. Int. 160 (2–3) (2006) 90–101.

[17] R Development Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2012, ISBN:3-900051-07-0, http://www.r-project.org.

[18] H. Haned, Forensim: an open source initiative for the evaluation of statistical methods in forensic genetics, Forensic Sci. Int. Genet. 5 (4) (2011) 265–268.

[19] J. Perez, A.A. Mitchell, N. Ducasse, J. Tamariz, T. Caragine, Estimating the number of contributors to two-, three-, and four-person mixtures containing DNA in high template and low template amounts, Croat. Med. J. 52 (3) (2011) 314–326.

[20] G. Morrison, Measuring the validity and reliability of forensic likelihood-ratio systems, Sci. Justice (2011).

[21] T. Tvedebrink, P.S. Eriksen, H.S. Mogensen, N. Morling, Estimating the probability of allelic drop-out of str alleles in forensic genetics, Forensic Sci. Int. Genet. 3 (4) (2009) 222–226.

[22] H. Haned, P. Gill, Analysis of complex DNA mixtures using the Forensim package, Forensic Sci. Int. Genet. Supp. Series. 3 (1) (2011) e79–e80.

[23] J.A. Bright, K. Manus, S. Harbison, P. Gill, J. Buckleton, A comparison of stochastic variation in mixed and unmixed casework and synthetic samples, Forensic Sci. Int. Genet. 6 (2) (2012) 180–184.

[24] T. Tvedebrink, P.S. Eriksen, M. Asplund, H. Mogensen, N. Morling, Allelic drop-out probabilities estimated by logistic regression – Further considerations and practical implementation, Forensic Sci. Int. Genet. 6 (2) (2012) 263–267.

[25] J. Weusten, J. Herbergs, A stochastic model of the processes in PCR based amplification of STR DNA in forensic applications, Forensic Sci. Int. Genet. 6 (1) (2011) 17–25.

[26] R.G. Cowell, S.L. Lauritzen, J. Mortera, Probabilistic expert systems for handling artifacts in complex DNA mixtures, Forensic Sci. Int. Genet. 5 (3) (2009) 202–209.

SUPPLEMENTARY MATERIAL

I. Mathematical formalization of the model

1. Model definition

In the main body of this paper we have described (by means of generic examples) the evaluation of likelihood ratios based on the qualitative data available from DNA profiles, namely the set of observed peaks. This evaluation is based on a model that was first described in Curran et al. [1], but the treatment of drop-out is not the same as in that article. Indeed, in [1] any allele that is present in the genotypes of at least one of the contributors drops out with probability D and does not drop out with probability $1 - D$. We have modified this to take into account homo- and heterozygosity and different drop-out probabilities per contributor. In the main body, this was illustrated on practical cases, without specifying the underlying probabilistic model. This specification is what we will do in this section by providing a general formula for the probability of observing a given allele a, conditioned on the genotypes of the contributors under a hypothesis H as well as on their drop-out probabilities. We will also point out that by doing so we actually define a probability space, since the sum of the probabilities of observing any specific outcome is one. We also illustrate in this section how good approximations can be obtained by hand. Let us also mention at this point that the formalization that we propose here treats drop-in in a different way than Curran et al. [1].

We will assume that all the loci that we consider can be treated independently of each other. Hence, suppose that we work on locus L, with allelic ladder $L = \{a_1, \ldots, a_t\}$. We need to define, for every allele a_i on the allelic ladder, the probability of observing allele a_i in the crime-sample profile. We assume that the number of contributors to the trace is known and equal to $n \geq 1$. Furthermore, we suppose that the donor's genotypes are all known and equal to $g_1 = (a_{1,1}, a_{1,2}), \ldots, g_n = (a_{n,1}, a_{n,2})$. Finally, we suppose that for contributor i, there is a number $0 \leq d_i \leq 1$ and a number $0 \leq D_i \leq d_i^2$, which we call the heterozygous, respectively homozygous, drop-out probability for contributor i. d_i and D_i are defined in the following way: if the crime-sample has a unique contributor i (with the same amount of DNA as in the actual sample), then the probability to observe $a_{i,1}$ would be $1 - d_i$ if $a_{i,1} \neq a_{i,2}$ and $1 - D_i$ if $a_{i,1} = a_{i,2}$. Hence, d_i represents the probability, for each heterozygous allele, that it drops out and D_i represents the probability that a homozygous allele would drop out.

We now return to the original situation with n contributors with known genotypes and drop-out probabilities for each. Let $n_{i,a}$ be the number of alleles that contributor i has of type a. Thus, $0 \leq n_{i,a} \leq 2$. Let $\mathtt{R}$ represent the DNA profile obtained from the crime-sample. Recall that we view $\mathtt{R}$ as a random subset of L: it contains the alleles that have been detected as the result of a stochastic process. Thus, $\mathtt{R}$ is a random variable taking values in the subsets of L and we can now define its probability distribution.

For each $a \in L$, we can now define the probability that it appears in the DNA profile. For convenience, we consider n as fixed (or clear from the context) and summarize $\vec{g} = (g_1, \ldots, g_n)$ and similarly for $\vec{d}$ and $\vec{D}$.

$$P(a \notin \mathtt{R} \mid \vec{g}, \vec{d}, \vec{D}) = \prod_{i: n_{i,a}=2} D_i \cdot \prod_{i: n_{i,a}=1} d_i, \tag{1}$$

where the empty product equals 1 by definition.

To arrive at the probability that R is a specific subset of L, we use (1) as follows. For a subset $R \subset L$, we have

$$P(\mathtt{R} = R \mid \vec{g}, \vec{d}, \vec{D}) = \prod_{a \in R} P(a \in \mathtt{R} \mid \vec{g}, \vec{d}, \vec{D}) \cdot \prod_{a \notin R} P(a \notin \mathtt{R} \mid \vec{g}, \vec{d}, \vec{D}). \tag{2}$$

This defines a probabilistic model, since

$$\sum_{R \subset L} P(\mathtt{R} = R \mid \vec{g}, \vec{d}, \vec{D}) = 1. \tag{3}$$

1.1. Uncertainty

In (2), the probability is conditioned on the genotypes of the contributors, and on their drop-out probabilities. In reality of course we do not know these. However, if we instead have a joint probability distribution for the number of contributors $\mathtt{N}$, their genotypes $\mathtt{G}_1, \ldots, \mathtt{G}_n$ and their drop-out probabilities $\mathtt{d}_i, \mathtt{D}_i$, then integrating over this distribution yields other probabilities of interest such as $P(\mathtt{R} = R \mid \mathtt{N} = 2, \mathtt{G}_1 = g_1)$, the probability that the sample's DNA profile (or rather, the set of detected alleles) is equal to R given that there are two contributors (one of whom has genotype g_1 and the other unknown) and a joint probability distribution for the drop-out probabilities of each contributor and the genotype of the second contributor.

In all our applications we consider the $\mathtt{G}_i$ as independent from each other and from the drop-out probabilities, and let $P(\mathtt{G}_i = g)$ be the population frequency of genotype g. However, the treatment of related contributors is not conceptually more challenging, only practically so. We will also assume that the number of contributors is specified exactly by the hypotheses for which we compute a likelihood ratio. Again, strictly speaking this is not necessary: for example, one may define a joint probability distribution such that the DNA profile is most likely to be obtained if it has a small number of contributors with small drop-out probability or a higher number of contributors with higher drop-out probabilities.

1.2. Multiple replicates

If several DNA profiles have been generated from the same trace, then one can extend the model (2) in various ways. The simplest and in our opinion also the most natural way, which we have therefore adopted in this paper, is to model each replicate as conditionally independent of all other replicates, given the contributors' genotypes and drop-out probabilities. In that case, the k-th replicate is modelled by random variable $\mathtt{R}_k$ and the $\mathtt{R}_k$ are independent, identically distributed copies of $\mathtt{R}$. Thus
$P(\mathtt{R}_1 = R_1, \ldots, \mathtt{R}_k = R_k \mid \vec{g}, \vec{d}, \vec{D}) = \prod_{j=1}^{k} P(\mathtt{R} = R_j \mid \vec{g}, \vec{d}, \vec{D})$.

1.3. Drop-in

In the illustration section of this paper, an approximation is used, based on Curran et al. [1], in order to account for the possibility of drop-in. In this section, we provide a justification for such approximation and demonstrate that it is a good estimate for the exact probabilities.

Note that according to (1), $P(\mathtt{R} = R \mid \vec{g}, \vec{d}, \vec{D}) = 0$, unless $R \subset \cup_i \{a_{i,1}, a_{i,2}\}$, i.e., the DNA profile obtained from the trace can only contain alleles present in at least one of the contributors. One may wonder whether it is necessary to refine this model such that there is a non-zero probability for an allele to be in $\mathtt{R}$ even when none of the contributors have it. We will do so below since it is a recurring feature in the literature. However, it can also be argued that such an extension of the model is not always necessary. It first has to be defined exactly what is meant by a drop-in, in this paper we follow the definition of the DNA commission for drop-in [2]: allele drop-in is a " Contamination from a source unassociated with the crime stain manifested as one or two alleles". But in any definition the drop-in alleles ultimately come from some human individual, usually one of no criminological interest. In case there are several replicates, there is a difference between this uninteresting low-level contributor being present in the sample itself or not. If so, one may simply add to the contributors 'of interest another contributor (or, possibly, even several) with a very high drop-out probability. For example, one may believe unprofessional handling of the sample during its collection may have added tiny amounts of DNA of a police worker to the sample, in which case the extra individual will be fixed (but unknown) throughout. As a consequence no more than two drop-in alleles can be explained by this model. If on the other hand, we think of a drop-in as the result of a random selection from the allelic population, dropping in from the air (e.g. plastic-ware), then we may see no reason to expect the drop-in alleles to be correlated over the various replicates, and model each replicate with a new extra individual. In that case one can also explain at most two alleles per locus per extra individual.

To take drop-in into account without having to resort to extra individuals, but still taking allele frequencies into account, we propose to modify (1) into

$$P(a \notin \mathtt{R} \mid \vec{g}, \vec{d}, \vec{D}) = (1 - cp_a) \prod_{i:n_{i,a}=2} D_i \prod_{i:n_{i,a}=1} d_i, \tag{4}$$

where $0 \leq c \leq 1$. In this case as well, by applying (2) to this model (4), we retain the property (3) that we are working within a model where the sum of the probabilities of obtaining each of the possible subsets of L equals one.

But now, for an allele a that none of the contributors possesses, i.e., $n_{i,a} = 0$ for all i, we have $P(a \in \mathtt{R} \mid \vec{g}, \vec{d}, \vec{D}) = cp_a$. So, for $n = 0$ and $c = 1$ we obtain that the probability $P(a \in \mathtt{R} \mid \vec{g}, \vec{d}, \vec{D}) = p_a$ and the expected number of drop-in alleles in that case is equal to one. However, any number of drop-in alleles has nonzero probability of occurring.

1.4. Likelihood ratios

The drop-in model (4) yields likelihood ratios

$$LR = \frac{P(\mathtt{R} = R \mid H_p)}{P(\mathtt{R} = R \mid H_d)}.$$

Such a likelihood ratio is a function of allele frequencies, the drop-out probabilities of the contributors and the drop-in variable c. Since c is very small, it is convenient to write the likelihood ratio as a power series

$$LR = \sum_{i=0}^{\infty} h_0 + h_1 c + h_2 c^2 + ..., \tag{5}$$

in the drop-in variable c, since it then suffices to compute only the first few terms h_i to reach a good approximation to the likelihood ratio. To calculate these coefficients h_i, we note that the likelihood ratio is obtained as the quotient of two probabilities, each of which we can write as a power series in c with the equation (4): let

$$P(\mathtt{R} = R \mid H_p) = \sum_{i=0}^{\infty} f_i c^i \tag{6}$$

and

$$P(\mathtt{R} = R \mid H_d) = \sum_{i=0}^{\infty} g_i c^i, \tag{7}$$

then

$$LR = \frac{f_0 + f_1 c + f_2 c^2 + \dots}{g_0 + g_1 c + g_2 c^2 + \dots}.$$

We can now obtain the coefficients h_i (which will be functions of the drop-out probabilities and allele frequencies) from

$$\sum_{i=0}^{\infty} h_i c^i = \frac{\sum_{i=0}^{\infty} f_i c^i}{\sum_{i=0}^{\infty} g_i c^i} \iff \sum_{i=0}^{\infty} f_i c^i = (\sum_{i=0}^{\infty} g_i c^i)(\sum_{i=0}^{\infty} h_i c^i)$$

by working out the product on the right hand side and comparing coefficients. This yields

$$f_i = \sum_{k=0}^{i} g_k h_{i-k},$$

from which the h_i can be obtained recursively. In particular the first few coefficients are

$$h_0 = \frac{f_0}{g_0}, \tag{8}$$

$$h_1 = \frac{f_1 g_0 - g_1 f_0}{g_0^2}, \tag{9}$$

$$h_2 = \frac{f_2 g_0^2 - f_1 g_1 g_0 + f_0 g_1^2 - f_0 g_2 g_0}{g_0^3}. \tag{10}$$

Hence, we have the following expressions for the likelihood ratio:

$$\frac{f_0}{g_0} + O(c),$$

$$\frac{f_0}{g_0} + \frac{f_1 g_0 - g_1 f_0}{g_0^2} c + O(c^2),$$

and

$$\frac{f_0}{g_0} + \frac{f_1 g_0 - g_1 f_0}{g_0^2} c + \frac{f_2 g_0^2 - f_1 g_1 g_0 + f_0 g_1^2 - f_0 g_2 g_0}{g_0^3} c^2 + O(c^3),$$

where $O(c^k)$ denotes terms that only involve powers of c that are of order k or higher.

Because the model is quite crude and we suppose that c is very small, there is limited interest in the higher order terms, and we will only consider the approximation

$$LR = \frac{f_0}{g_0} + \frac{f_1 g_0 - g_1 f_0}{g_0^2} c + O(c^2). \tag{11}$$

From this we see that if $f_0 = 0$, meaning that under H_p a drop-in allele is needed in order for the replicate to have nonzero probability of being observed, then we obtain

$$LR = \frac{f_1}{g_0} c + O(c^2). \tag{12}$$

This means that we need only to compute the terms f_1 and g_0, which correspond (respectively) to the probability of the replicate being observed under H_p if one drop-in allele has been observed, and the probability of the replicate being observed under H_d without a drop-in allele having been observed.

2. Illustration

2.1. Case B

We illustrate the principles exposed above with the example in Figure 1-B. In this case, $\mathtt{R} = (9)$, only allele 9 has been observed in the crime-sample (we do not take the peak into account that is visible at allele 10 but is below the detection threshold), and there is a suspect whose genotype is $(9, 10)$. We write p_i for the allele frequency of allele i. Since a drop-in allele is needed neither under H_p nor under H_d, it is clear that the LR will depend only very weakly on c. Therefore the most appropriate approximation to it would be

$$LR \approx \frac{f_0}{g_0},$$

which amounts to the LR obtained with $c = 0$. This can be calculated easily. Indeed, under H_p the donor's genotype is known and yields the observed replicate with probability $f_0 = d(1-d)$ in the absence of drop-in, whereas under H_d the contributor needs to have allele 9, and we obtain $g_0 = p_9^2(1-d^2) + 2p_9(1-p_9)d(1-d)$, hence

$$LR = \frac{d}{p_9^2(1+d) + 2p_9(1-p_9)d} + O(c),$$

which is the same as in Table 2 (taking $d' = d^2$).

However for illustration purposes, we now compute the LR up to first order of c.

2.1.1. Under H_p

Under the hypothesis H_p, the trace has one contributor, namely the suspect, whose drop-out probability for a heterozygous allele is d.

It follows from the application of the drop-in model (4) that

$$P(9 \in \mathtt{R} \mid H_p) = 1-d+dcp_9, P(10 \notin \mathtt{R} \mid H_p) = (1-cp_{10})d, P(x \notin \mathtt{R} \mid H_p) = 1-cp_x,$$

where $x \notin \{9, 10\}$. Hence,

$$P(\mathtt{R} = (9) \mid H_p) = (1-d+dcp_9)d(1-cp_{10}) \prod_{x \notin \{9,10\}} (1-cp_x).$$

An algebraic manipulation, collecting power of c, yields

$$P(\mathtt{R} = (9) \mid H_p) = d(1-d) + d(d+p_9-1)c + O(c^2).$$

If $c = 0$ we retrieve $P(\mathtt{R} = (9) \mid H_p) = d(1-d)$.

2.1.2. Under H_d

Under H_d, the trace comes from an unknown contributor U with heterozygous drop-out probability d, homozygous drop-out probability $d' = d^2$, whose alleles are drawn at random from the population according to the allele frequencies. Note that we take the drop-out probability for a heterozygous allele to be the same as it was under H_p, which is not necessary but seems reasonable and simplifies the computation. The assumption $d' = d^2$ is also made for such reasons. We have, abusing our notation somewhat,

$$P(\mathtt{R} = (9) \mid H_d) = \sum_{(xy)} P(\mathtt{R} = (9) \mid H_d, U = (xy))P(U = (xy) \mid H_d).$$

Now, according to the drop-in model, and analogously to what we found under H_p, it is not hard to see that

$$\begin{aligned} P(\mathtt{R} = (9) \mid U = (9,9), H_d) &= (1 - d^2 + cp_9 d^2) \prod_{x \neq 9} (1 - cp_x) \\ &= 1 - d^2 + (d^2 + p_9 - 1)c + O(c^2), \end{aligned}$$

$$\begin{aligned} P(\mathtt{R} = (9) \mid U = (9,a), H_d) &= (1 - d + cp_9 d)d(1 - cp_a) \prod_{x \notin \{9,a\}} (1 - cp_x) \\ &= (1 - d)d + d(d + p_9 - 1)c + O(c^2), \end{aligned}$$

$$P(\mathtt{R} = (9) \mid U = (a,b), H_d) = cp_9 d^2 + O(c^2),$$

where a, b are alleles different from 9.

Summing over the possible genotypes yields $P(\mathtt{R} = (9) \mid H_d)$.

2.1.3. Likelihood ratio

Putting everything together, we get (after an algebraic manipulation)

$$LR = \frac{P(\mathtt{R} = (9) \mid H_p)}{P(\mathtt{R} = (9) \mid H_d)} = \frac{d}{p_9(p_9 - d(p_9 - 2))} + \frac{d^2(d(p_9 - 1)^2 - p_9^2)}{p_9(d - 1)(p_9 - d(p_9 - 1))^2} c.$$

In the below Figure 1, we plot for $c = 0$ and $c = 0.1$, and $p_9 = 0.11$ the likelihood ratio as a function of d. The dotted lines represent the above linear approximation, whereas the red lines correspond to the exact solution based on (4). We notice that the linear approximation is very good for small values of d, but as $d \to 1$, it diverges from the exact solution.

Indeed, as $d \to 1$, the contributor's alleles cannot be detected any more, hence the observed allele must be a drop-in, and cannot be informative about H_p and H_d any more.

Looking at our expression as the likelihood ratio as a power series (5) in c, this behaviour is understandable. Indeed, in (7) the term g_0 contains a factor $(1-d)$, whereas the terms f_i in (6) for $i > 0$ do not. Therefore, all coefficients h_i except h_0 will have a pole at $d = 1$, meaning that the series does not converge any more; and for d sufficiently close to one, more terms would be needed.

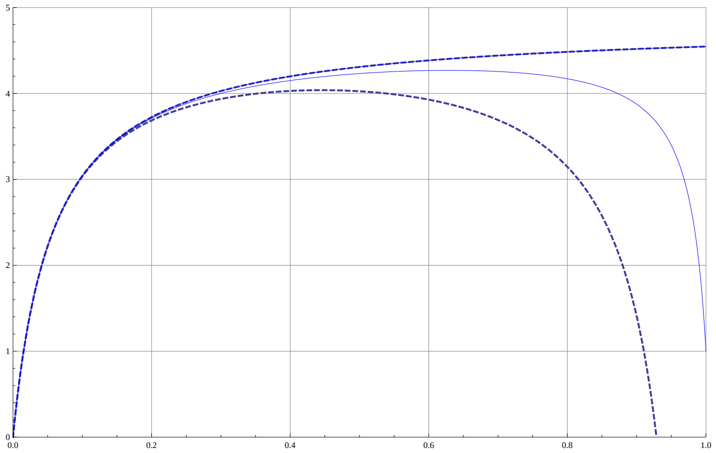


Figure 1: Exact LR and linear approximation (dotted) in c, for $c = 0$ and $c = 0.1$. For $c = 0$, the curves coincide. For $c = 0.1$, the approximation is best for small d.

2.2. Case C

We now turn to the case C (Figure 1-C), where $R = (7, 9)$ and the suspect is the same as in Case B, having genotype $(9, 10)$. Therefore, if the suspect has to explain the trace (i.e., under H_p), a drop-in allele is needed. If we write $P(\mathtt{R} = (7,9) \mid H_p) = \sum_{i=0}^{\infty} f_i c^i$ and $P(\mathtt{R} = (7,9) \mid H_p) = \sum_{i=0}^{\infty} g_i c^i$, we have $f_0 = 0$ and therefore we can write, as first order approximation,

$$LR = \frac{f_1}{g_0} c + O(c^2).$$

This is easy to compute, since f_1 corresponds to the situation where there is one drop-in allele under H_p (which must be allele 7) and g_0 corresponds to the situation where there is no drop-in allele under H_d, and therefore the donor of the trace must be of genotype $(7, 9)$. Thus, $f_1 = p_7 d(1-d)$ and $g_0 = 2p_7p_9(1-d)^2$, and therefore

$$LR = \frac{d}{2p_9(1-d)}c + O(c^2). \tag{13}$$

For small c, this will be a satisfactory approximation to the likelihood ratio; but for illustration purposes we briefly indicate how to obtain the second order approximation

$$LR = \frac{f_1}{g_0}c + \frac{f_2g_0 - f_1g_1}{g_0^2}c^2 + O(c^3).$$

We are therefore going to compute the terms f_1, f_2, g_0, g_1.

2.2.1. Under H_p

Along the same lines as for case B, we get

$$P(\mathtt{R} = (7,9) \mid H_p) = cp_7(1-d+dp_9)d(1-cp_{10}) \prod_{x \notin \{7,9,10\}} (1-cp_x).$$

This yields

$$f_0 = 0, f_1 = p_7d(1-d), f_2 = dp_7(p_9 + (1-p_7)(d-1)).$$

2.2.2. Under H_d

As for case B, here we need to compute the probability to observe $\mathtt{R} = (7,9)$ for every possible genotype of the donor separately. Since we are only interested in the terms with at most one drop-in allele, we need only consider those genotypes that involve allele 7 or 9. Of these genotypes, (7,9) is the only one where no drop-in is needed, from which we see that, as has been already mentioned,

$$g_0 = 2p_7p_9(1-d)^2.$$

To compute g_1, we reason among similar lines as for case B. This yields, after straightforward algebraic manipulations,

$$g_1 = (d-1)p_7p_9(2 - 3p_7 - 3p_9 + 3d(-2 + p_7 + p_9)).$$

For the LR, this means that

$$\begin{aligned} LR &= \frac{f_1}{g_0} + \frac{f_2 g_0 - f_1 g_1}{g_0^2} c^2 + O(c^3) \\ &= \frac{d}{2p_9(1-d)} c + \frac{d(-p_7 - p_9 + d(-4 + p_7 + 3p_9))}{4(d-1)^2 p_9} c^2 + O(c^3). \quad (14) \end{aligned}$$

2.2.3. Evaluation

To verify in which range of d the approximations are good enough, we plot the likelihood ratio as a function of d for $c = 0.01$ and $c = 0.1$ calculated according to three choices: (1) the first order approximation (13), (2) the second order approximation (14) and (3) the exact likelihood ratio according to the drop-in model.

The results are in the Figure below.

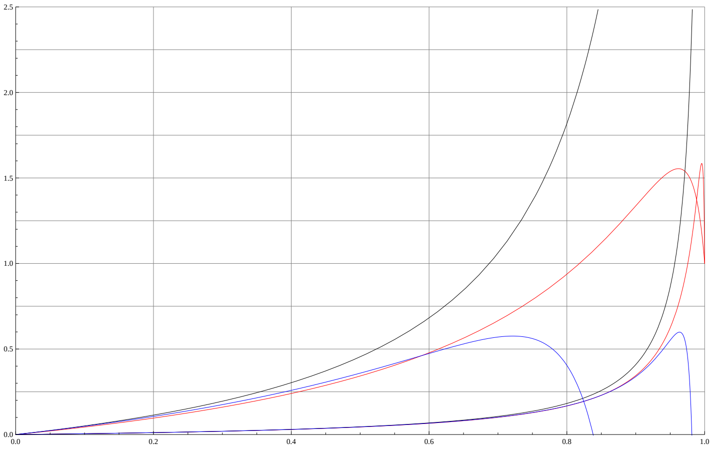


Figure 2: Linear (black) and quadratic (blue) approximations to the LR as well as exact LR (red) for $c = 0.01$ (lower three lines) and $c = 0.1$ (upper three lines).

Note that for small drop-out probabilities, the exact likelihood ratio almost coincides with both approximations and the LR can be very well estimated as f_1/g_0c, which can be easily calculated by hand. For large d however, the first order approximation keeps growing whereas the second order approximation yields absurd results, i.e., negative likelihood ratios. This is due to the fact that g_0 contains a factor $d-1$ which persists in all the h_i for $i > 0$. Using the exact likelihood ratio, we see that it tends to one as $d \to 1$. Indeed, if drop-out is certain, then the profile must be the result of a drop-in and cannot be informative about hypothesis that state its contributor.

2.2.4. Comparison to Curran model

Finally, we compare the likelihood ratio as obtained in Table 3, to the exact likelihood ratio calculated according to (4). The result is, for $c = 0.1$, presented in the Figure below, from which it is clear that for small values of d the likelihood ratios almost coincide, but for large d the likelihood ratio obtained from Table 3 (obtained from the Curran model) tends to slightly overestimate the evidence.

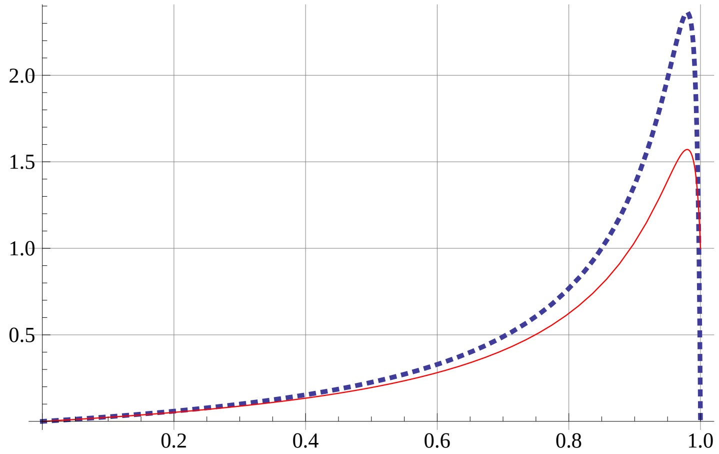


Figure 3: Exact (red) and Curran's LR (blue, dotted) for $c = 0.1$.

2.3. Evaluation

From these examples it is clear that model (4), which is easy to program into a computer and is a consistent probabilistic model, can also be used to obtain good approximations to the exact likelihood ratio which can be done by hand. Indeed, in Case B (where no drop-in event was needed under

H_p, the approximation f_0/g_0 is very satisfactory for small d. Similarly for case C, where one drop-in event is needed under H_p and none under H_d, the approximation f_1/g_0c is very satisfactory. Since the terms f_1 and g_0 are very easy to obtain, a good approximation to the likelihood ratio can be obtained without much effort.

Finally we also observe that the exact likelihood ratios computed with (4) are, in the examples Case B and Case C, well approximated by the expressions obtained in the main body of the article based on the formulas from Curran et al. However, since for Case C we have observed that the latter formulas can overestimate the LR, it would be advisable to perform the exact calculation, especially for elevated drop-out probabilities.

II. Supplementary Figures

Case B

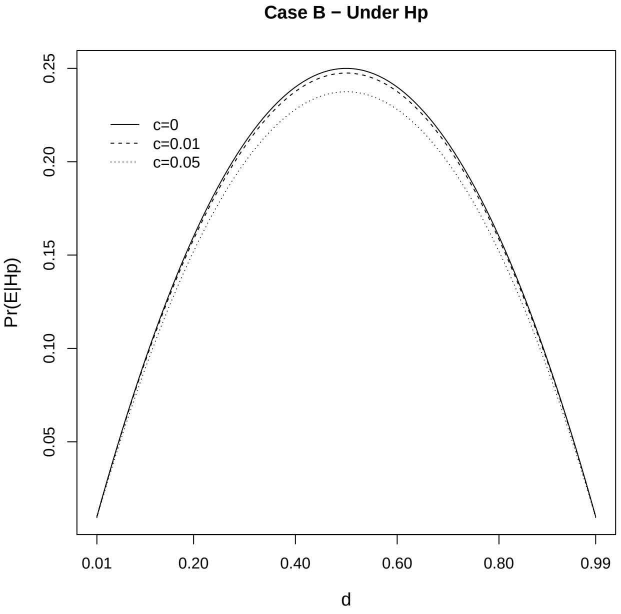


Figure 4: Effect of drop-in on the probability of the evidence under H_p for case B. The drop-in probability c varies in $\{0, 0.01, 0.05\}$.

As shown in Figure 4, the sensitivity analysis under H_p yields a bell-shaped curve and it is symmetric, the maximum value is reached for $d = 0.50$, when drop-out and non-drop-out are equally likely. If drop-out decreases below 0.5, this means that the evidence points away from the heterozygote profile of the suspect, the same thing happens if we increase d above 0.5, the heterozygote genotype is no longer supported. Incorporating the c parameter for drop-in decreases the probability of seeing the evidence if the heterozygote suspect contributes to the sample. When comparing the dashed lines (drop-in) to the solid line (no drop-in), the drop-in parameter c acts like a scaling factor, decreasing the probability of the evidence.

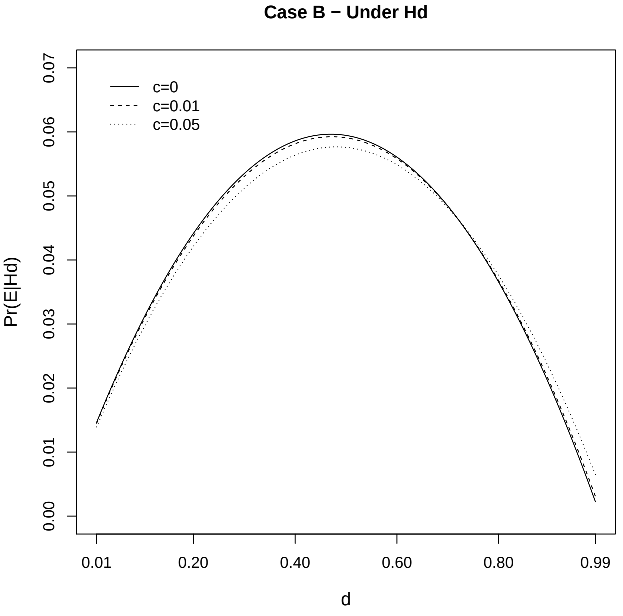


Figure 5: Effect of drop-in on the probability of the evidence under H_d for case B. The drop-in probability c varies in $\{0, 0.01, 0.05\}$.

Under H_d, the replicate probabilities are weighted by the genotypic probabilities[1]. As shown under H_p, increasing the drop-in probability render the

[1] Which is also the case under H_p, except that in this single-source case, there are no unknowns under H_p, and hence the genotype probabilities are always 1.

heterozygote genotype less likely. Conversely, if drop-in is more likely, then genotypes such as QQ, and QQ', have a higher probability, this will inflate the likelihood under H_d, compared to the case where $c = 0$. Taking the ratio, we see why the LR decreases when c increases.

Case C

In this case, we need to incorporate the drop-in parameter in order to explain the sample profile (see Figure 1-C). Under H_p, there is one drop-out (allele 10) and one drop-in (allele 7). The sensitivity plot under H_p is shown in Figure 6.

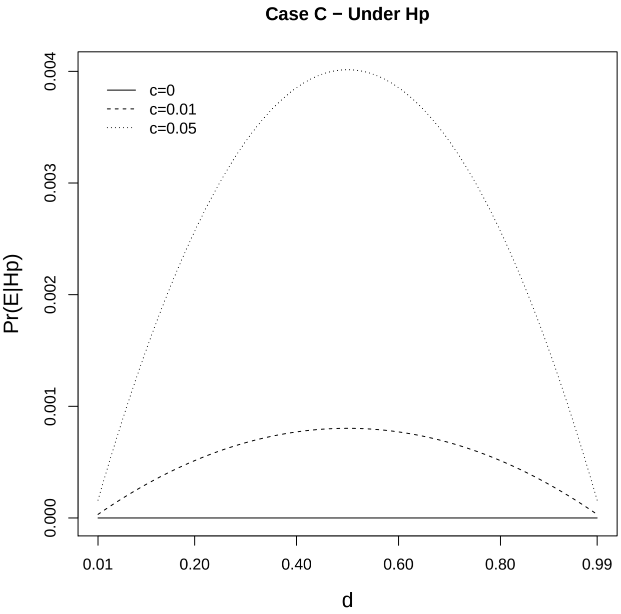


Figure 6: Effect of increasing the drop-in on the probability of the evidence under H_p for case C. The drop-in probability c varies in $\{0, 0.01, 0.05\}$.

The probability of the evidence is 0 under H_p if the drop-in parameter is set to 0 (solid line), increasing the probability of drop-in yields a bell-shaped curves, but note that the probabilities are much lower than for case B under H_p (Figure 4).

Under H_d, the maximum probability value is obtained when $d = 0.01$ and $c = 0$ (Figure 7). This corresponds to having a random person wit genotype 7,9 contributing to the sample (see Table 3) . The probability tends to 0 when d tends to 1, since plausible genotypes under H_d become less likely if drop-out decreases.

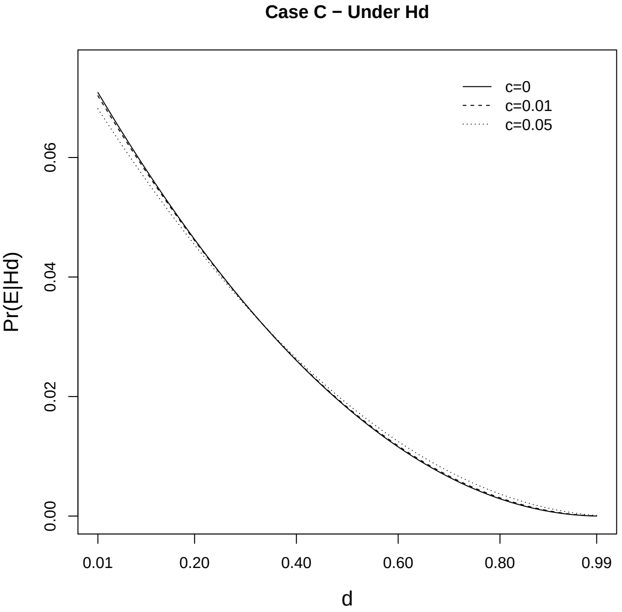


Figure 7: Effect of increasing the drop-in on the probability of the evidence under H_d for case C. The drop-in probability c varies in $\{0, 0.01, 0.05\}$.

References

[1] J. M. Curran, P. Gill, M. R. Bill, Interpretation of repeat measurement DNA evidence allowing for multiple contributors and population substructure, Forensic Sci. Int. 148 (2005) 47–53.

[2] P. Gill, C. H. Brenner, J. S. Buckleton, A. Carracedo, M. Krawczak, W. R. Mayer, N. Morling, M. Prinz, P. M. Schneider, B. S. Weir, DNA commission of the International Society of Forensic Genetics: Recommendations on the interpretation of mixtures, Forensic Sci. Int. 160(2-3) (2006) 90–101.

References

P. Gill, C. H. Brenner, J. S. Buckleton, A. Carracedo, M. Krawczak, W. R. Mayer, N. Morling, M. Prinz, P. M. Schneider, and B. S. Weir. DNA commission of the International Society of Forensic Genetics: Recommendations on the interpretation of mixtures. *Forensic Sci. Int.*, 160(2-3):90–101, 2006.

P. Gill, L. Gusmao, H. Haned, W.R. Mayr, N. Morling, W. Parson, L. Prieto, M. Prinz, H. Schneider, P.M. Schneider, and B.S. Weir. Dna commission of the international society of forensic genetics: Recommendations on the evaluation of str typing results that may include drop-out and/or drop-in using probabilistic methods. *Forensic Sci. Int. Genet.*, 6(6):679 – 688, 2012.

P. Gill, A. Kirkham, and J. M. Curran. LoComatioN: A software tool for the analysis of low copy number DNA profiles. *Forensic Sci. Int.*, 166(2-3):128–138, 2007.

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

P. Gill, J. M. Curran, C. Neumann, A. Kirkham, T. Clayton, J. Whitaker, and J. Lambert. Interpretation of complex dna profiles using empirical models and a method to measure their robustness. *Forensic Sci. Int. Genet.*, 2:91–103, 2008.

H. Haned, K. Slooten, and P. Gill. Exploratory data analysis for the interpretation of low template dna mixtures. *Forensic Science International: Genetics*, 6(0): 762–774, 2012.

H. Haned. Forensim: an open source initiative for the evaluation of statistical methos in forensic genetics. *Forensic Sci. Int. Genet.*, 5(4):265–268, 2011.

I. W. Evett and B. S. Weir. *Interpreting DNA evidence: statistical genetics for forensic scientists*. Sinauer Associates Sunderland, Mass, 1998.

J. Buckleton, C. M. Triggs, and S. J. Walsh. *Forensic DNA evidence interpretation*. CRC PRESS, 2005.

D. J. Balding and R. A. Nichols. DNA profile match probability calculation: how to allow for population stratification, relatedness, databse selection and single bands. *Forensic Sci. Int.*, 64:125–140, 1994.

P. Gill and H. Haned. A new methodological framework to interpret complex dna profiles using likelihood ratios. *Forensic Sci. Int. Genet.*, 7(2):251 – 263, 2013.