**Title:** The use of Bayesian Networks and simulation methods to identify the variables impacting the value of evidence assessed under activity level propositions in stabbing cases.

**Authors:** Samie L, Champod C, Taylor D, Taroni F

**Journal:** Forensic science international. Genetics

**Year:** 2020 Jun 11

**Issue:** 48

**Pages:** 102334

**DOI:** 10.1016/j.fsigen.2020.102334

# The use of Bayesian Networks and simulation methods to identify the variables impacting the value of evidence assessed under activity level propositions in stabbing cases

## 1. Introduction

Technical analytical developments have made it possible to analyse very low amounts of DNA. The drawback of this progress is that evaluation of the results has become increasingly complex when activity level propositions are taken into account. Indeed, low quantity DNA traces can be the result of a secondary transfer or even a tertiary transfer. Variables related to transfer, persistence and recovery are becoming increasingly important considerations in evidence evaluation, as the relevant question arising in Court is no longer "who is the source of this DNA?" but rather "how did this DNA get there?" [1-2]. These considerations were at the basis of the ENFSI guideline for evidence reporting [3] and the advice given to DNA reporting officers to systematically report trace DNA evidence considering activity level propositions adopting a likelihood ratio (LR) approach. This article is focused on how to interpret traces with small amounts of DNA, typically found following the touching of an object. This is not a simple task as it requires the probabilistic consideration of multiples variables such as transfer, persistence, recovery and background level of DNA. Many experts face practical difficulties when assessing these variables due to the perception that every case has a unique set of circumstances and numerical assignations are critically dependent on the specificities of the case. Biedermann *et al.* [4] explained further why forensic scientists may struggle with these assessments. A common argument put forward is: "because each case has its own feature, the use of numerical values from experimental studies performed under controlled (laboratory) conditions cannot be used for evaluation in real life case". Besides, experiments can only cover a limited number of options for any particular variable, so it is difficult to envisage experiments taking into account all possible variations. However, although each case may have different data for these variables, this does not mean that the LR would be affected by all possible variations of these variables. Identifying the variables that impact on the LR will help forensic scientists to focus on a limited number of variables of interest in order to limit time and cost of the required data acquisition. The key task that will be explored in this paper is to identify the variables that have a significant impact on the weight to be assigned to the DNA findings.

The objective of this paper is to present a methodology to support forensic scientists in the evaluation of their results given activity level propositions. This study will show how to identify

the variables impacting on the LR taking advantage of simulation methods. The purpose is to show how to reduce the number of variables that require consideration. Once this is done then this provides a focus for further data collections which can be used to inform distributions that can be used for evaluative court-going purposes.

Taylor *et al*. [5] built a Bayesian Network (BN) [6] allowing the transfer mechanisms to be considered. They offer a way to compute the LR associated with the DNA findings considering activity level propositions. However, in order to be used in casework, the BN nodes need to be quantified by probabilities, also informed with adequate data. The number of variables in the BN from [5] is large and a case-specific data acquisition to inform the parameters on all of them would be out of reach. This paper provides a method on how to identify the key variables that truly impact on the LR by performing simulations based on the BN. It is an extension of the simulation approach adopted by Taylor *et al*. [7] on body fluids attribution. Because the BN construction in [5] is the basis of the present contribution, the reader is advised to refer to it as we will not fully describe here its construction. In this paper, we will limit the description of the BN to the few modifications that we have introduced, the BN parametrisation and the simulations techniques that we will use.

To illustrate this method, the scenario of a stabbing attack with a knife is used. It can be summarized as follows: A victim is found dead in his flat following a stabbing incident carried out by an offender using a knife. The exact day of the stabbing is uncertain. At the crime scene, a knife is recovered and believed to be the attacker's weapon. The knife had been properly secured. It is believed by the investigators, in line with pathologist's assessment, that the stabbing occurred about 2 days (± 0.5 day) before the discovery of the crime scene.

Based on the elements of the investigation, not related to DNA evidence, a person of interest (POI) is arrested and suspected to be the offender who stabbed the victim. Two days after the reporting of the incident, a DNA swab is taken from the unstained smooth plastic handle of the knife with a view to detect potential trace DNA left by the offender.

The prosecution's proposition (denoted $H_p$) is that the POI was the person who used the knife to stab the victim.

In this paper, to gain some generalisation, two options are studied to reflect upon the defence point of view (denoted $Hd_1$ and $Hd_2$ respectively):

– In the first defence proposition, $Hd_1$, the possibility of a secondary transfer is explicitly considered. The POI claims that, he shook hands with an unknown person, probably few hours before the

stabbing, and it is that unknown person who is the real offender (who will be called alternative offender AO).

– In the second defence proposition, $Hd_2$, it is alleged that the POI didn't stab the victim, but someone else unrelated to him did it (AO). In addition, the POI denies any prior encounter with either the victim, the knife or the AO. This represents a situation in which the POI is claiming no direct link with the offence.

Exploring these two defence's propositions will allow to show that, depending on the propositions of interest, the variables that have the strongest impact on the LR might be different.

The above-described circumstances provide the elements of what we will call the "initial case". Again, with a view to explore further that this specific set of circumstances, we will develop two additional cases. The first will adopt circumstances favouring the transfer of the POI's DNA (a "high transfer case"). The second adopts circumstances that are less favourable to the transfer of POI's DNA (an "low transfer case"). These sets of circumstances are further described in section 2.4.

## 2. Methodology

The methodology adopted for this research is decomposed into several stages. First, we constructed a Bayesian network (BN) allowing a scientist to assess DNA findings associated with the stabbing attack scenario (including the possibility to change the defence proposition). The construction is mainly based on Taylor *et al.* [5] but has been adapted in order to carry out simulations. Then the conditional probability tables (CPTs) for each node of the network have been informed based on the data available from the literature. The parameters have been modelled in a way to (a) allow a Bayesian update in the light of new data and (b) to reflect when applicable the fact that the amount of data may be sparse or limited.

As such the constructed BN can be used to compute a likelihood ratio for a given case as done in [5]. However, we would like to go further by exploring impact and limitations of data on the LR values. A similar approach was adopted by [7]. This will be done using simulations resampling from the underpinning distributions used to inform the CPTs. The sampling will be carefully chosen by exploring one variable after the other. That is done in order to identify which variable (or node in the BN) has the most significant impact on the variations observed on the LRs. The

isolation of the variables that have the most bearing on the variation of the LR from one simulation to the other is important to inform a data acquisition strategy. As it will transpire later, the developed BN has multiple variables and forensic laboratories will certainly not be able to systematically investigate the underpinning data for all of them. We will show that this simulation methodology allows us to successfully identify the key variables that should be the focus of future data acquisition. Such methods have already been applied successfully in areas such as DNA and fibres [8, 9]. Here, it is proposed to adopt such techniques to establish a baseline to inform future data acquisition campaigns.

In the next section, we will firstly present the development of the Bayesian network on the basis of what we call the "initial case". It will be used as the starting point from which all simulations will be carried out. Then, the method used to perform the simulations, from this "initial case", will be presented.

### *2.1. The Bayesian network, the underpinning data to inform the CPTs and variable instantiations for the "initial" case*

In this section, we present the BN, its variables, their states and the data used to inform the CPTs. In addition, we will use this BN with specific variable instantiations representing the circumstances of a case. We have called it the "initial case". Initially, it is given this set of circumstances that we will explore how the variables of the BN impact on the LR values. The other two cases (high transfer and low transfer) will be dealt with in a second step of the study.

The Object-Oriented Bayesian Network (OOBN), illustrated in Figure 8 of Taylor *et al.* [5] is reproduced here in Figure 1 and will be used in our study. It was adapted to be easily used in simulations.

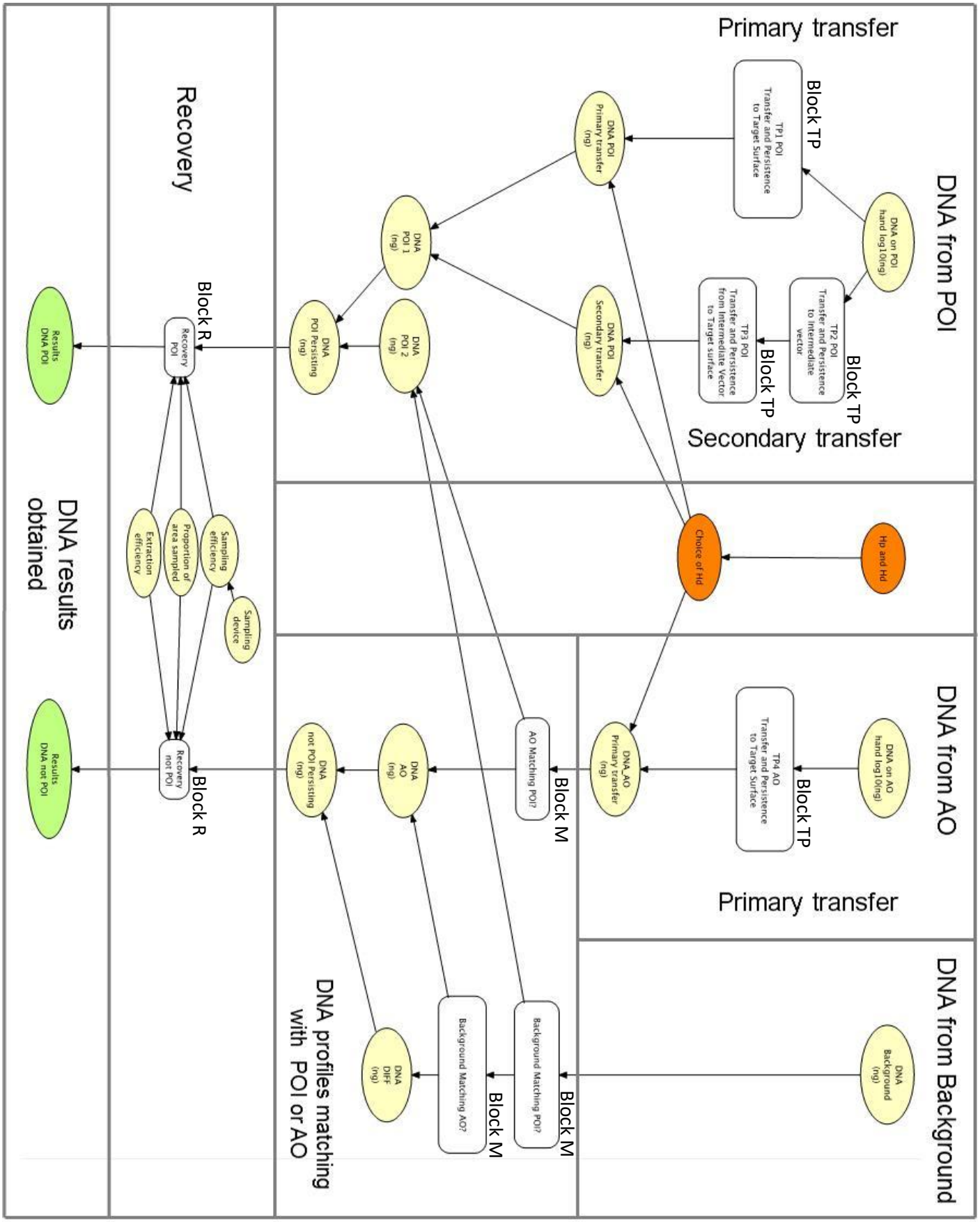



*Figure 1: Bayesian network used to evaluate the findings under activity level propositions involving primary vs secondary transfer events adapted from [5].*

Each TP block contains variables involved in transfer and persistence event, more specifically:

- Block TP1 contains variables involved in the transfer and persistence of POI's DNA from POI's hand to the knife handle that occurs during the stabbing.
- Block TP2 contains variables involved in the transfer and persistence of POI's DNA from POI's hand to AO's hand that occurs during the handshake.
- Block TP3 contains variables involved in the transfer and persistence of POI's DNA from AO's hand to the knife handle that occurs during the stabbing.
- Block TP4 contains variables involved in the transfer and persistence of AO's DNA from AO's hand to the knife handle that occurs during the stabbing.

Variables included in each block called TP (for Transfer Persistence), M (for matching) and R (for recovery) are described in Table 1. The BN has been developed using Hugin Researcher (version 8.6, www.hugin.com).

| **Block** | **Variable** |
|---|---|
| Block TP | Proportion of Contact |
| | Transfer proportion |
| | Nature of surface (for both target and shedding surface) |
| | Type of Contact |
| | Environment |
| | Days |
| | Decay factor of the loss of DNA (node named 'alpha2'), function of Environment. |
| Block M | Match Probability |
| Block R | Sampling efficiency |
| | The proportion of area sampled |
| | Extraction efficiency |

*Table 1: Variables taken into account in the blocks of the BN described in Figure 1.*

Regarding the states of the variables, they can be instantiated depending on the case information available.

Table 2 details the states for each variable according to the case circumstances of the "initial case" and the type of sampling used by the laboratory.

| Variable | Possible states | Instantiated state for "initial case" | Explanation for the choice of the instantiated state |
|---|---|---|---|
| Nature of target surface | Hand<br>Smooth<br>Rough | Smooth | TP1-TP3-TP4: The surface of the knife handle is smooth. |
| | | Hand | TP2: It's a handshake |
| Nature of shedding surface | Hand<br>Smooth<br>Rough | Hand | TP1-TP2-TP3-TP4: The hand is the primary source of DNA |
| Vigour of contact | Passive<br>Pressure<br>Friction | Friction | TP1-TP3-TP4: It is assumed that type of contact is friction when you stab a person. |
| | | Pressure | TP2: It is assumed that the type of contact is pressure when you shake hands. |
| Environment | Favourable<br>Poor | Favourable | TP1-TP3-TP4: The target surface is the knife. It is assumed that the knife being kept in a paper bag by the crime scene investigators shortly after it was seized. The environment (the paper bag) is considered favourable in the sense that DNA will be preserved. |
| | | Poor | TP2: The intermediate surface is a hand. It is assumed that this surface can be considered as an unfavourable environment because of the high risk of contact with other surfaces, resulting in a loss of DNA. |
| Sampling device | Tapelift<br>Swab | Swab | A swab was used to take samples from the knife |

*Table 2: Variables with their associated states and the instantiated state corresponding to the circumstances of the "initial case".*

Table 3 presents the non-instantiated variables and the data that will be used to inform their CPTs.

| Variable | Data informing the variable |
|---|---|
| Days | Data 1 |
| Proportion of area sampled | Data 2 |
| Transfer proportion | Data 3 |
| Sampling efficiency | |
| Extraction efficiency | |
| DNA quantity on hands | |
| Background | |

*Table 3: Non-instantiated variables and corresponding data used to inform their CPTs.*

The data that will be used to inform their CPTs are referred to as Data 1 to Data 3 are detailed hereinafter.

The outcomes of the DNA analysis are set in two variables that gives the amount of DNA in ng, respectively for the POI and for not POI (Table 4).

| Variable | Possible states (ng) |
|---|---|
| Results DNA POI | interval node;<br>0 to 0.1 in steps of 0.01<br>0.1 to 1 in steps of 0.1<br>1 to 5 in steps of 0.5<br>5 to 10 in steps of 1<br>10 to 25 in steps of 5<br>25 to 1000<br>1000 to inf |
| Results DNA not POI | interval node;<br>0 to 0.1 in steps of 0.01<br>0.1 to 1 in steps of 0.1<br>1 to 5 in steps of 0.5<br>5 to 10 in steps of 1<br>10 to 25 in steps of 5<br>25 to 1000<br>1000 to inf |

*Table 4: Possible results for the amount of DNA in ng respectively for POI and not POI.*

### *2.2. Representing our lack of knowledge in the BN*

Within the nodes in the BN there is uncertainty as to what the state of nature was in any particular instance. This can be thought of as comprising two distinct parts, the uncertainty surrounding the state that applies to a specific case, and the uncertainty surrounding the suitability of our data to model the world. The node 'DNA on hands' is a good example to further explain this idea.

In the stabbing scenario being considered there is uncertainty surrounding the amount of DNA on the hands of the POI (or AO) and we deal with this in the BN by treating the 'DNA on hands' node as a distribution that is meant to reflect our uncertainty through the amount of DNA that the general population will possess on their hands. This distribution reflects our prior belief on the amount of DNA that the suspect (or AO) had on their hands at the time of the offence. Within the BN architecture the POI and AO have separate 'DNA on hands' nodes to reflect the fact that they are different people and can have different amounts of DNA on their hands (the posterior distribution of which would be obtained after instantiation of case information). In order to model the distribution of DNA on hands in the population, we must take a sample from the population and measure the amount of DNA on peoples' hands. The second component of

our uncertainty is then how representative our sample is of the general population. It is only this second component of uncertainty that will benefit from additional research and sampling i.e. will potentially reduce the sensitivity of the LR to the data underlying these nodes. Consider a scenario where the results have very little power to distinguish between the propositions (and hence a LR close to one would be obtained). If this is due to the shape of the distributions used to model the population (and not how representative the sample is of the general population), then no amount of additional sampling will help to increase discrimination.

In our sampling schemes (which we list below) there are different aspects of case information and experimental data that we focus on during our simulations that represent both aspects of uncertainty.

Data 1: Days

To account for the uncertainty on the number of days between the stabbing and the crime scene attendance, and the number of days (or hours) between the stabbing and the alleged handshake, the variable "Days" is modelled by a Gamma distribution Ga(α, ß) that has been discretized on a range of possibilities (from 0 to 31 days). The choice for a Gamma distribution allowed to easily model events from 0 to infinity. The parameters α and ß (shape and rate) are calculated using [10] based on a mean and a variance (set by the circumstances) (Table 5). The variance is set to 0.5 in order to account for an uncertainty of about two days maximum. Note that in [5] we modelled the persistence of DNA (or the reduction of the amount of DNA over time) using an exponential decay curve set by a variable called 'alpha2'. The parameter of the decay curve depends on the state of the variable "Environment" (either 'poor' or 'favourable') and is based on [Raymond et al, 2009]. The parameter 'alpha2' is set to 0.022 if the environment is favourable or to 0.052 if the environment is poor.

| Variable | Discretized states | Mean and variance | Explanation | Ga(α, ß) modelling the variable |
|---|---|---|---|---|
| Days | interval node; 0, 0.5 then 1 to 31 in steps of 1 | $\mu=2$<br>$\sigma^2=0.5$ | TP1-TP3-TP4: The item was examined around two days after the offence. | Ga(8, 4) |
| | | $\mu=0.5$<br>$\sigma^2=0.5$ | TP2: It is assumed that the handshake was made less than 12 hours but more than 2 hours before the stabbing. | Ga(0.5, 1) |

*Table 5: Variable "Days" with its states and parameters (mean and variance and associated Gamma distributions) used to inform the CPTs.*

Data 2: Proportion of area sampled

This variable is modelled using a Beta distribution Be(α, ß), whose parameters are estimated from a mean and variance for the proportion. It represents the proportion the touch surface that is sampled (using a swab or a tapelift). It allows to account for the fact that the whole (100%) of the touched surface may not be sampled. Note that it does account for the uptake efficiency of the swab or tapelift. The variability on the latter is accounted directly in the variable "Transfer proportion". The parameters α, ß are computed from mean and variance according to [11]. Table 6 summarizes these parameters. As for "Days", this distribution is used here to reflect an uncertainty regarding the circumstances of the case. If we had no doubt regarding them, it would be unnecessary to carry out such simulations.

| **Variable** | **Discretized states** | **Mean and variance of data** | **Explanation** | **Be(α, ß) based on mean and variance** |
|---|---|---|---|---|
| Proportion of area sampled | interval node; values from 0 to 1 in steps of 0.1 | $\mu$=0.95<br>$\sigma^2$=0.001 | We have assumed that almost the entire knife handle is sampled, representing about 95% of its surface. | Be(44.17, 2.32) |

*Table 6: Variable "Proportion of area sampled" with the explained mean and variance associated to the data used to inform the CPTs, with the parameters of the beta distribution based on the mean and variance.*

Data 3: Transfer proportion, Sampling efficiency, Extraction efficiency, DNA quantity on hands and Background

We have adopted a full Bayesian strategy to inform the parameters associated with these variables. It means that we have initially set a prior probability distribution for the variables. Then, based on data from the scientific literature, we have updated these distributions, leading to posterior distributions that will be used to inform the CPTs associated with them.

The prior distributions set for each variable are presented in Table 7.

For the variables "Transfer proportion", "Sampling efficiency" and "Extraction efficiency", a so-called flat prior Be(1,1) has been chosen to start from a uninformed situation between 0 and 1.

The variable "DNA quantity on hands" is transformed in log10 (from –inf to +inf) with a prior distribution based on a mean quantity of 2ng of DNA and a large variance (variance of 1). The mean quantity of 2ng is what appears to be a reasonable amount of DNA. A large variance was chosen to account for the paucity of data available at this stage. Note that the above method with these parameters led to only positive values.

The variable "Background" is modelled using a Gamma distribution whose parameters have been estimated [10] from a mean quantity of background DNA (0.5 ng) and an associated variance (variance of 5). The mean of 0.5 ng is viewed as a reasonable upper bound quantity. The large variance has been chosen to be large reflecting the paucity of data available. Ten prior observations have been drawn from that Gamma distribution to act as our prior data counts. We will then update these counts with the data obtained from the literature. 10 was chosen to reflect the paucity of the sample sizes available in the literature. Indeed, we couldn't claim more in the prior counts than what is actually published.

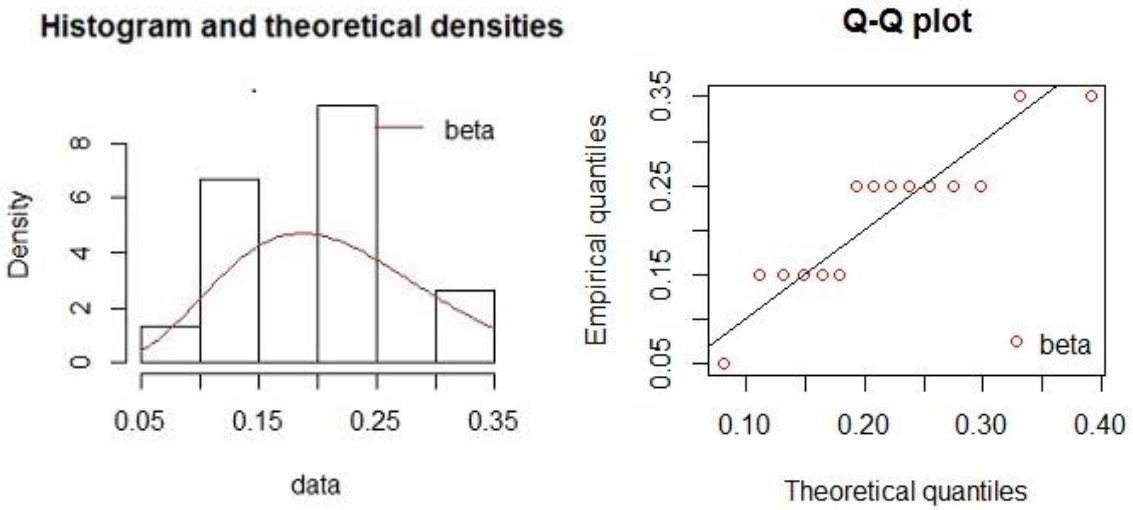
The data used to update the above prior distribution for the variable "Transfer proportion" are based on [5] where the authors estimated the parameters of the Beta distribution modelling "Transfer proportion" based on simulations from the original data from Daly [12], Bontadelli [13] and Goray [14]. To reflect the fact that a laboratory will conduct only a limited number of experiments to inform that variable, we randomly selected 51 data from the beta distribution obtained in [5] to act as the dataset that we will use to update our prior distribution. 51 is the number of experiments done by Daly [12].

For the variable for the "Sampling Efficiency", we will distinguish the "swab" from the "Tapelift". In [5], the parameters Be(25, 20) of the beta distribution representing the data for "Swab" condition were based on [15]. As before, we will randomly draw a limited sample (21) from that distribution to carry out the Bayesian update of our prior distribution. Indeed, 21 experiments were done in [15]. For the "Tapelift" condition, we have used directly the data from [15].

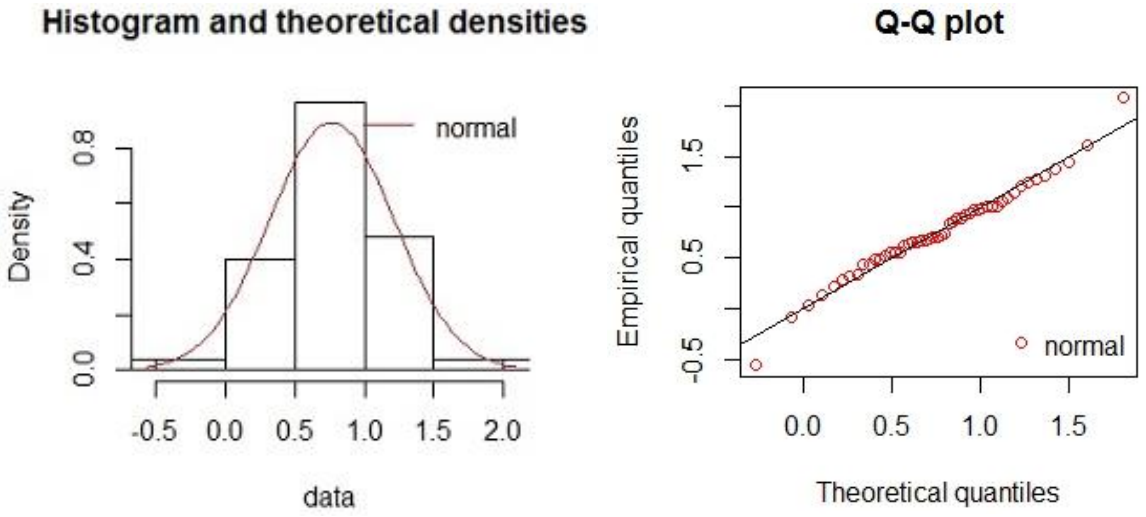
For the variables "Extraction efficiency", "DNA quantity on hands" and "Background", we have used directly the data from the literature (Table 7). We have used these data to fit, respectively, a Beta, a Normal and a Gamma distribution (Figure 2). Note that only 15 data points from [17] have been used to inform the variable "Extraction efficiency", hence the poor

quality of the fit (Figure 2a). It will serve as a starting point, keeping in mind that this variable will probably be flagged up following the simulations as a variable requiring more data to inform it.

(a) Extraction efficiency




(b) Log10 of Quantity of DNA on hands




(c) Background

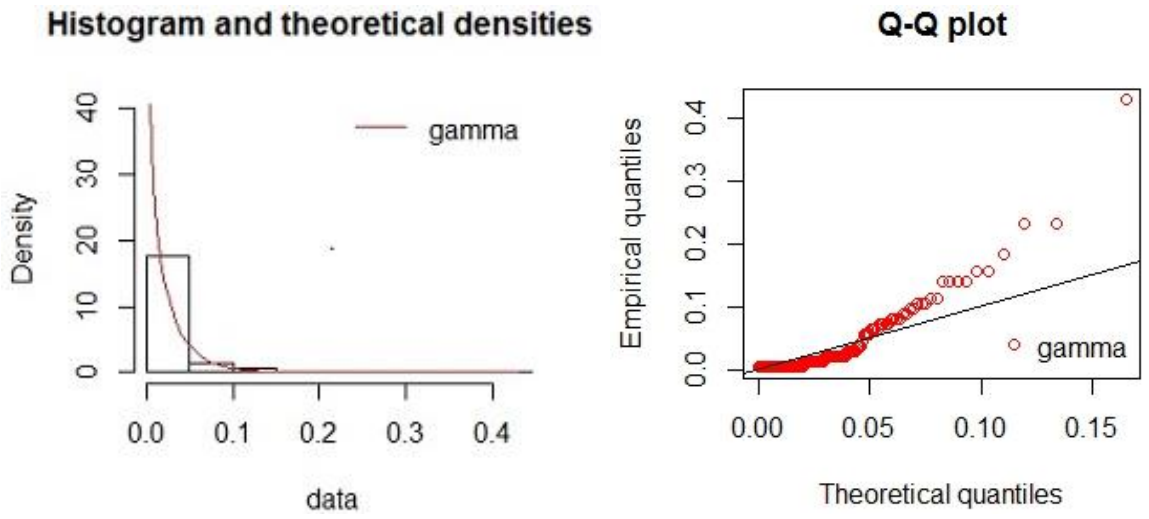



*Figure 2: Histogram and theoretical densities and Q-Q plots for data of the variables "Extraction efficiency" (a), "DNA quantity on hands" (b) and "Background" (c).*

Beta, Normal and Gamma distributions were updated using standard Bayesian methods [16]

| Variable | Discretized states | Prior distribution | Data Source | Number of data points | Posterior distribution |
|---|---|---|---|---|---|
| Transfer proportion | interval node; from 0 to 1 in steps of 0.05 | Be(1,1) | 51 random sample from the Beta distribution defined in [5] | 51 | Hand(rough)/passive: Be(1.89,3.83)<br><br>Hand(rough)/pressure: Be(1.99,2.78)<br><br>Hand(rough)/friction: Be(2.02, 2.98)<br><br>Smooth/passive: Be(1.55,32.65)<br><br>Smooth/pressure: Be(1.73,33.43)<br>Smooth/friction: Be(1.43, 2.10) |
| Sampling efficiency | interval node; from 0 to 1 in steps of 0.05 | Be(1,1) | State "Tapelift":[15]<br>State "Swab": 21 random sample from beta distribution from [5] | 21 | Swab: Be(14.36,11.79)<br><br>Tapelift: Be(3.15,17.11) |
| Extraction efficiency | interval node; from 0 to 1 in steps of 0.05 | Be(1,1) | [17] | 15 | Be(5.79,18.43) |
| Log10(DNA quantity on hands) | interval node;<br>-inf to -1.5<br>-1.5 to 3.5 in steps of 0.1<br>3.5 to inf | N(0.3,1) | [13] | 50 | N(0.764,0.004) |
| Background | interval node;<br>0 to 0.1 in steps of 0.01<br>0.1 to 1 in steps of 0.1<br>1 to 5 in steps of 0.5<br>5 to 10 in steps of 1 | 10 data randomly selected from the distribution Ga(0.08,0.16) | [18] | 301 | Ga(0.6, 30.15) |

*Table 7: Variables, discretized states, prior distribution, data source to update them and posterior distribution used to inform their CPTs.*

### *2.3. Methods used to perform simulations from the "initial case"*

In the previous section, we have presented the BN that captures the "initial case". As such it can be used to compute a LR for any DNA outcome (Table 4) in terms of quantity of DNA corresponding to the POI and non-corresponding to the POI. For a given outcome, we will obtain one LR that encapsulates the knowledge that has been used to inform the CPTs.

Using the Bayesian update mechanism presented before (under Data 3), we have used specific data sets from published studies. The numbers of data points used in these studies are rather sparse and can be seen as small subsets from larger and unknown populations. Through simulations, we would like to show the impact (if any) on the LR of such limited samples. To do that, we have resampled with replacement the data points that have been used to carry out the Bayesian update. The simulation method is detailed later in this section. At each simulation then, the above-described posterior distributions are recomputed, the BN CPTs are updated and LRs obtained for any DNA outcome. We carried out that task 100 times. Hence for a given DNA outcome, we have now 100 LRs. These LRs will have a range that can be characterized [e.g. IQR, min to max]. Note that these 100 LRs represents 100 slightly different scenarios.

A first set of results out of these simulations will show these ranges (one per possible DNA outcome and for each of the defence propositions). They reflect how the LRs vary based on the state of knowledge and understanding for sets of data points used to inform the CPTs.

The second question we will address trough simulation is to identify which variables impact the most on the LR. This identification of impacting variables will allow prioritisation of further data acquisitions. That approach stems from the realisation that, given the complexity of the problem, we cannot expect systematic acquisition of large datasets for all the variables identified in the BN. To make that selection, we will carry out resampling on a variable by variable basis (keeping all the other variables constant). For each set of simulations (100 simulations per variable), we will measure the ranges (for each DNA outcome and considering each of the defence propositions). These simulations on a variable per variable basis allow pinpointing of the variables that have the most impact on the LRs. These shall then constitute the focus for further data acquisition, because additional datasets have the potential to reduce the observed ranges of LRs and improve on their robustness.

The simulations were performed in Rstudio (Version 1.1.463) [19] with R (version 3.5) [20] combined with RHugin (version 8.4) [21]. The library RHugin specifically allows to liaise directly with the Hugin inference engine. It allows then to load the BN as in Hugin and to interact directly with it without resorting to the Hugin GUI. It means that all the captured dependencies between the variables in the BN are duly maintained and are part of the computation.

Table 8 presents the variables that will be used for the simulation exercise (either jointly or separately) with an indication of the method of simulation that will be applied. These three simulation methods are described below.

| **Variable** | **Simulation method** |
|---|---|
| Days | 1 |
| Proportion of area sampled | |
| Transfer proportion | 2 |
| Sampling efficiency | |
| Extraction efficiency | |
| DNA quantity on hands | |
| Background | |
| Match probability | 3 |

*Table 8: Each variable associated with a method of simulation.*

*Simulation method 1*: For each simulation, the mean for each variable ("Days" and "Proportion of area sampled") is resampled from a Normal distribution with a mean set as per the initial case but allowing a variance, respectively of 0.5 and 0.01 around it. For instance, the mean for the number of days in the TP1 block is 2. For each simulation then, the mean will be obtained by randomly selecting from a sample from a N(2,0.5), until the mean is positive (For "Proportion of area sampled", the mean is resampled until a value between 0 and 1 is obtained). From that mean, and keeping the variance constant, we estimate, as before, the parameters of the Gamma distribution (for "Days"). The values for the variance were chosen because, in the authors' opinion, they adequately reflect the amount of uncertainty surrounding the timeframes.

*Simulation method 2*: For each of these variables that were subjected to the Bayesian update, the simulations are carried out by resampling (with replacement) from original data presented in the previous section.

*Simulation method 3*: The block M of the BN, presented in Figure 3, shows a dependence between the node "Match probability" and the node "Quantity of DNA" as in [5].

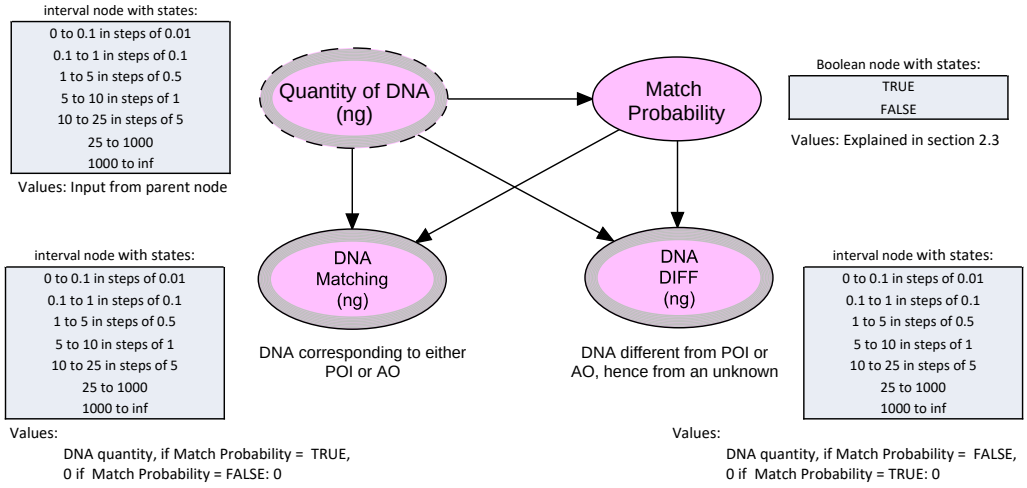



*Figure 3: Block M with its four nodes and associated states and values.*

In that block, only the node "Match Probability" requires input probabilities (the node "Quantity of DNA" being set by its parent). Its state is TRUE when the DNA profile is matching, FALSE otherwise. The probability of being TRUE is set by the match probability which in turn depends on the quantity of DNA. Linked to the quantity of DNA is the number alleles that the profile will show. To compute the match probability, we have accounted for that relationship between the quantity of DNA and the number of alleles.

Hence, we will first establish the relation between quantity of DNA and the number alleles and then propose a way to compute match probability as a function of the number of alleles. The number of alleles corresponding to a given quantity of DNA is modelled based on the empirical observations [22] made between the quantity of DNA and the minimum and maximum numbers of detected alleles (Table 9). A Gamma distribution is used to represent the numbers of detected alleles as a flexible modelling distribution for counts between 0 and infinity. The parameters of the Gamma distribution are informed on the mean and variance obtained from [22]. These Gamma distributions will be used at each simulation to generate a given number of alleles corresponding to a given quantity of DNA (sampled from its own distribution).

| **Quantity (ng)** | **Minimum to maximum numbers of detected alleles** | **Mean and variance** | **Ga(κ, ϴ) modelling the Number of alleles** |
|---|---|---|---|
| 0-0.01 | 0 to 1 | $\mu=1$; $\sigma^2=0.1$ | Ga(10, 10) |

| | | | |
|---|---|---|---|
| 0.01-0.02 | 0 to 2 | $\mu=1$; $\sigma^2=0.4$ | Ga(2.5, 2.5) |
| 0.02-0.03 | 1 to 3 | $\mu=2$; $\sigma^2=0.4$ | Ga(10, 5) |
| 0.03-0.04 | 2 to 4 | $\mu=3$; $\sigma^2=0.4$ | Ga(22.5, 7.5) |
| 0.04-0.05 | 3 to 5 | $\mu=4$; $\sigma^2=0.4$ | Ga(40, 10) |
| 0.05-0.06 | 4 to 6 | $\mu=5$; $\sigma^2=0.4$ | Ga(62.5, 12.5) |
| 0.06-0.08 | 5 to 7 | $\mu=6$; $\sigma^2=0.4$ | Ga(90, 15) |
| 0.08-0.1 | 6 to 8 | $\mu=7$; $\sigma^2=0.4$ | Ga(122, 17.5) |
| 0.1-0.2 | 9 to 22 | $\mu=15$; $\sigma^2=1.5$ | Ga(150, 10) |
| 0.2-0.3 | 20 to 30 | $\mu=25$; $\sigma^2=1.4$ | Ga(446, 17.9) |
| 0.3-0.4 | 29 to 32 | $\mu=31$; $\sigma^2=0.5$ | Ga(1922, 62) |
| >0.4 | 32 | the number of alleles is 32 | the number of alleles is 32 |

*Table 9: Range of quantities of DNA (ng) with the associated minimum and maximum numbers of alleles observed empirically. For each number of alleles (min-max), a mean and variance is set to reflect these ranges and the Gamma distributions parameters are obtained.*

To obtain the match probability for a given number of alleles (associated with a given quantity of DNA), we proceeded as follows:

(1) 5 million full DNA profiles (32 alleles in total) are randomly generated based on allelic frequencies of the NGMSElect kit [23] for the Swiss Caucasian population.
(2) From the above profiles, 27046 partial DNA profiles are created using the allelic degradation model from Hicks et al [24]. It means that for each number of alleles (1 to 31), we have a collection of partial DNA profiles (from 6 to 1000) depending on the number of alleles.
(3) For each of these profiles, their match probability (MP) is computed with a $\theta$ of 0.02 using the allele frequencies from [23].
(4) For a given number of alleles, at each simulation, we draw the match probability from the collection of match probabilities associated with the drawn number of alleles (corresponding to a given quantity of DNA).

During the simulation process, the number of alleles and the associated MPs are resampled only if this node is taken into consideration. Otherwise, its CPTs is set once for all simulations.

### *2.4. Method used to perform simulations beyond the "initial case"*

The "initial case" represents the circumstances of the case, typically set by the chosen states in the nodes that have been instantiated. The BN also allows us to explore other sets of circumstances. Hence, we can repeat the simulation process to explore how each node impacts

on the LRs under any set of circumstances. We have chosen to investigate two additional scenarios deviating from the conditions of the "initial case". The first will adopt circumstances that will favour the transfer of the POI DNA (a rough surface to increase the deposition, a short time delay between the stabbing and the collection of the swabs, and favourable environmental conditions). The second adopts circumstances that are less favourable to the transfer of DNA. The states of the nodes for each scenario are presented in Table 10. The other nodes related to the case circumstances were kept constant with the same states as for the "initial case".

| **Node** | **"Initial case" Instantiated state or mean** | **"High transfer case" Instantiated state or mean** | **"Low transfer case" Instantiated state or mean** |
|---|---|---|---|
| Nature of target surface | TP1-TP3-TP4: Smooth<br>TP2: Hand | TP1-TP3-TP4: Rough<br>TP2: Hand | TP1-TP3-TP4: Smooth<br>TP2: Hand |
| Days between both transfers or transfer and recovery | TP1-TP3-TP4: 2<br>TP2: 0.5 | TP1-TP3-TP4: 0.5<br>TP2: 0.5 | TP1-TP3-TP4: 20<br>TP2: 0.5 |
| Environmental conditions | TP1-TP3-TP4: Favourable<br>TP2: Unfavourable | TP1-TP3-TP4: Favourable<br>TP2: Unfavourable | TP1-TP3-TP4: Unfavourable<br>TP2: Unfavourable |

*Table 10: Choice for node instantiations and mean for the "initial case", for the "high transfer case" and for the "low transfer case".*

### *2.5. Method used for the analysis of the simulation results*

Following a set of simulations (one for each of the initial, the high transfer and the low transfer case), we obtain 100 LRs for each combination of results (quantity of POI and not POI DNA – 36 x 36 possibilities), each proposition retained for the defence ($Hd_1$ and $Hd_2$) and for each node considered (10 nodes and the case with all simulated nodes considered jointly). It represents a total of 28,512 combinations and 2,851,200 LRs.

A dedicated Shiny application (https://lydie-samie.shinyapps.io/DNA_Activity/) has been designed to allow the visualisation of these results for each possible combinations of variables.

To explore which node (or variable) has impact on the LRs, we ordered them by range. That will be done by conditioning on the defence proposition. It is important to stress that with this approach, we aim at identifying the variables that have the most impact across all possible outcomes. The variable by variable analysis allows us to identify which variables contribute significantly to the whole. By significant, we mean that the range of log10 (LRs) (meaning the difference between the maximum log10 (LR) and the minimum log10 (LR)) produced by

sampling a variable exceeds 1 order of magnitude of log10 (LR). The range was chosen to cover up to extreme situations even if their probability of occurrence is low. At this stage, the issue to screen among variables with a rather wide net, instead of being too strict in their selection by adopting a more stringent criterion such as the interquartile range.

### *2.6. Performing simulations adapting the number of data points informing the conditional probability tables*

As proposed in [7] we will mimic an increase of knowledge, to do so we have increased the counts informing the CPTs by a constant factor (retaining the observed proportions).

The variables ("Background", "DNA quantity on hands", "Sampling efficiency", "Extraction efficiency" and "Transfer Proportion") that proved to be significant based on the method described in section 2.4, have been studied using simulations. To simulate an increase in the size of datasets, we have adapted the simulation process (*Simulation method 2)* by tripling the number of data points resampled and then used to inform the corresponding CPTs. The original numbers of datapoints used for each variable are given in Table 7. It means that the relative proportions associated with each state of each variable remain the same, but the counts are multiplied by 3 as if the study had been conducted on a larger sample. The factor of 3 is what we considered reasonable for an operational laboratory.

## 3. Results and Discussion

### *3.1. Ranges of simulated LRs obtained under $Hd_1$*

$Hd_1$ stipulates that the defence alleges that the DNA corresponding to the POI's profile is the consequence of a secondary transfer. The DNA corresponding to POI's profile is the contribution of the POI following the secondary transfer and the DNA different from POI's profile is to the potential joint contribution of the background and the DNA of the alternative offender (AO). Figure 4 illustrates the ranges of log10 (LRs) for all outcomes (i.e. all considered amounts of POI and not POI's DNA, each possibility gives a range of LRs after 100 simulations) considering respectively all variables jointly and then each variable simulated in turn. The results are shown for the three cases considered: "initial", "high transfer" and "low transfer"). Regardless of the case, the significant variables (with an impact of more than an order of magnitude) are: "DNA quantity on hands", "Extraction efficiency", "Background",

and “Sampling efficiency”. Two of these variables “Extraction efficiency” and “Sampling efficiency” relates to the laboratory choices regarding their sampling devices and extraction techniques. The results obtained for the three cases are similar, particularly the results when “Initial case” and “High transfer case” are considered.

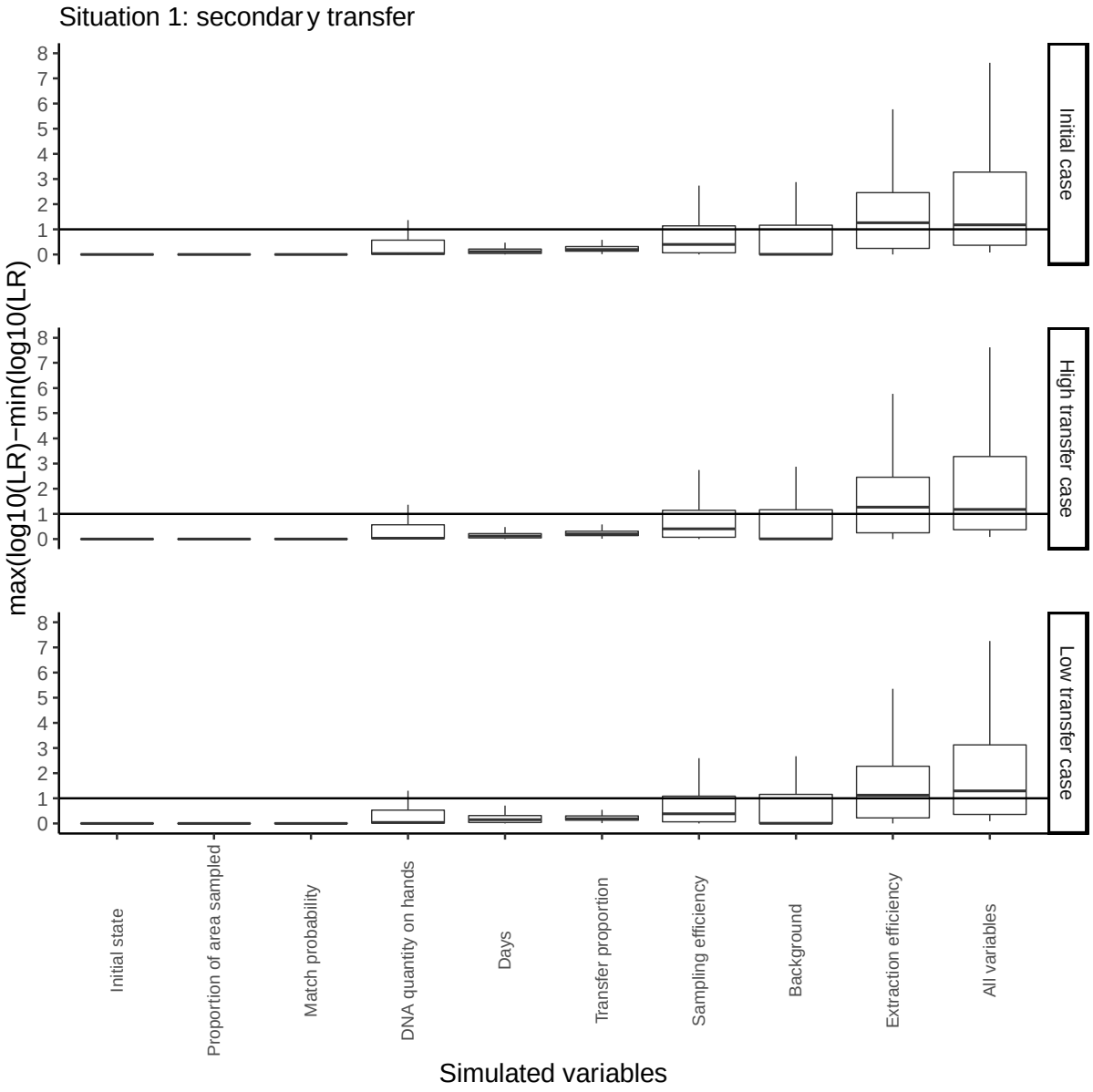



*Figure 4: Boxplots presenting the ranges (min-max) expressed in log10 of the LRs obtained following 100 simulations under $Hd_1$. The panels present the results for the three cases considered. The horizontal line drawn at the difference of 1 (in log10) set the limit above the variable considered will be declared as having a significant effect on the global variability shown when “all variables” are resampled jointly.*

Hence, when secondary transfer is alleged, the current state of knowledge regarding the transfer and persistence of DNA is sparse and induces a large variability on the simulated LRs. One way to overcome this problem would be to increase the underpinning data (see section 3.4 below). It is rather easy for a laboratory to increase its knowledge base associated with some of the

variables involved (see [25] for example regarding the acquisition of data in relation to sampling and extraction efficiency).

### *3.2. Ranges of simulated LRs obtained under situation 2 ($Hd_2$)*

In situation 2 the defence alleges that the DNA corresponding to POI's profile is due to the contribution of an unknown alternative offender (AO) ($Hd_2$). In this situation, the possibility for a secondary transfer for the POI is not retained. Figure 5 illustrates the ranges of log10 (LRs) plotted considering the same variables as before. The variable by variable analysis allows to identify the variables contributing to more than one order of magnitude. They are the following: "DNA quantity on hands", "Match probability", "Extraction efficiency", "Proportion of transfer", "Sampling efficiency" and "Background", regardless of the scenario considered.

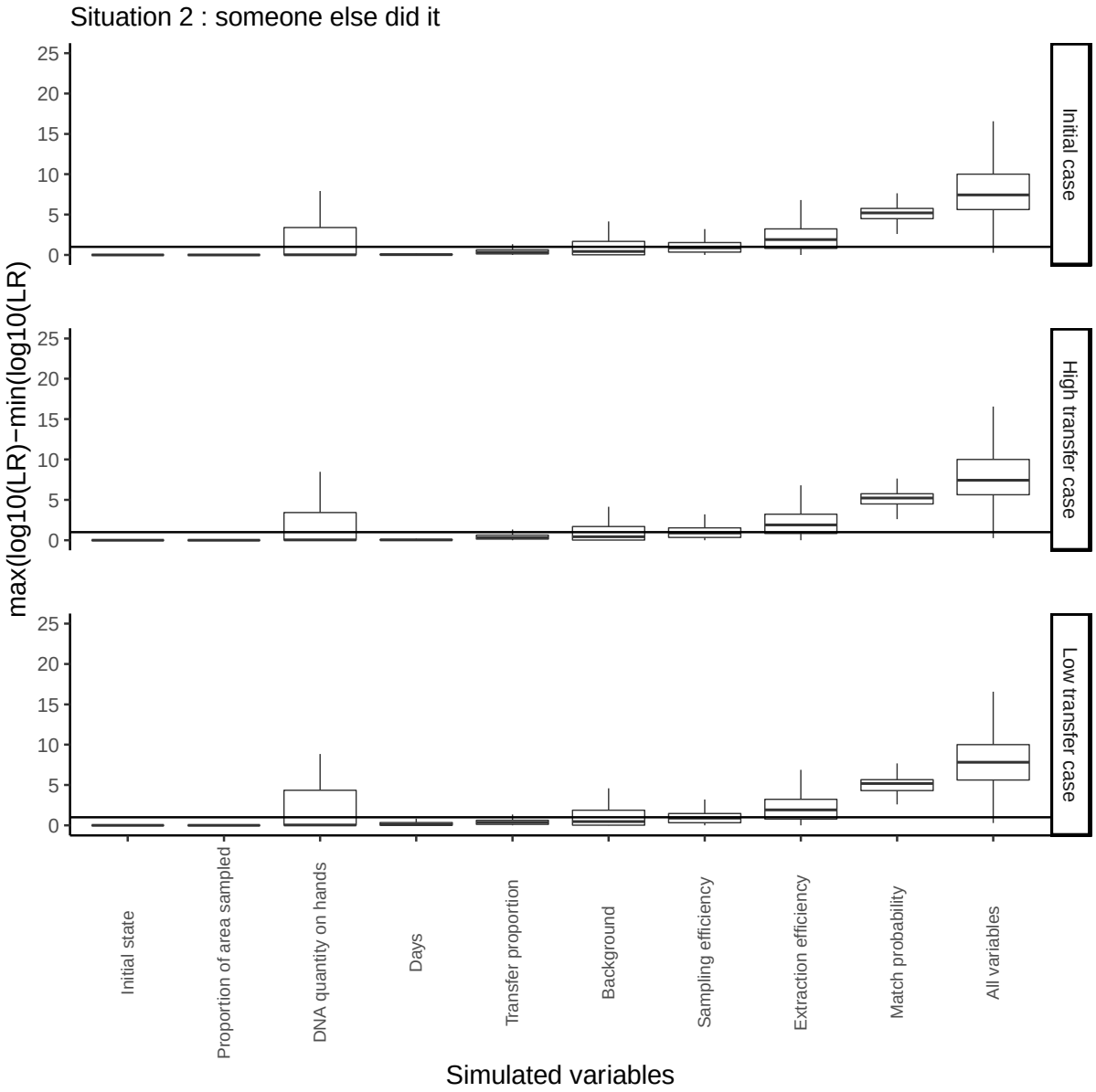



*Figure 5: Boxplots presenting the ranges (between min-max) expressed in log10 of the LRs obtained following 100 simulations under $Hd_2$ The panels present the results for the three cases considered. The horizontal line drawn at the difference of 1 (in log10) set the limit above the variable considered will be declared as having a significant effect on the global variability shown when "all variables" are resampled jointly.*

In addition to the variables identified for $Hd_1$, the variables "Match probability" and "Transfer proportion" comes into play under $Hd_2$. This is due to the fact that under that defence proposition, the DNA corresponding potentially to the POI may arise from the background level of DNA on the surface. That was much less critical under $Hd_1$ as the secondary transfer was dominating the considerations compared to the background.

The above results are showing the most impacting variables in a global sense, regardless of the outcome observed in a given case (i.e. the respective amounts of DNA corresponding to POI and to the not POI). It helps us to identify the variables that ought to receive attention if we

want to treat the problem globally considering all possible outcomes. Also the case (initial, high transfer and low transfer) considered as no impact on the impacting variables.

However, in a given case with observed outcomes, the variables that will have the most impact will vary and could be different from the above globally selected variables. To illustrate this point, we present one example (with 0.5-0.6 ng of POI's DNA with 0.03-0.04 ng of not POI's DNA) leading to a different selection of the most contributing variables (Figure 6). The reader can refer to the Shiny application to select any set of outcomes and explore the most impacting variables. In this chosen particular case, under $Hd_1$, the results show that the most significant variables are "DNA quantity on hands", "Background" and "Extraction Efficiency", whereas under $Hd_2$, the most impacting variable is the "match probability". As before, a variable is declared to be significant (boxplots coloured in green in Figure 6) if the range (the whole height of the boxplots shown in Figure 6) is above 1 order of magnitude of log10 (LR).

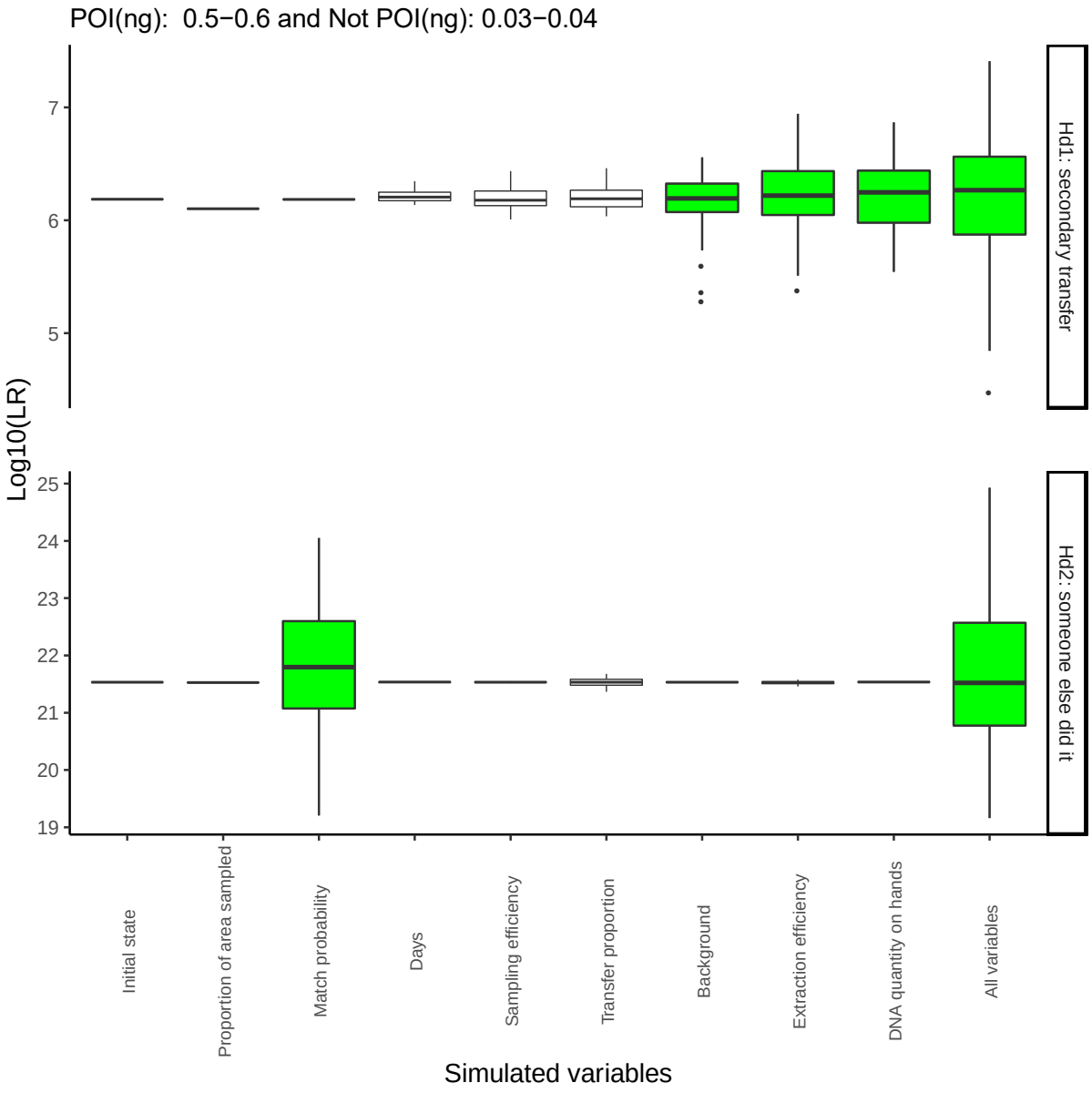



*Figure 6: Boxplots of the LRs (expressed in Log10) obtained after 100 simulations of each variable in both situations ($Hd_1$ and $Hd_2$) with a quantity of POI's DNA between 0.5 and 0.6 ng with between 0.03 and 0.04 of not POI's DNA. The boxplots corresponding to the significant variables are coloured in green.*

The purpose of setting up a rather complex simulation regime was twofold: (1) facing a large networks of connected variables, the simulations allow us to identify where there is a knowledge gap by identifying the most impacting variables and (2) to appraise the range of variations the computed LRs may take given the limited data points that have informed each conditional probability table in the Bayesian network.

For the example in Figure 6, we note that when considering $Hd_1$ the variables significantly impacting are only 2 and relates to transfer mechanisms. Under $Hd_2$, the match probability variable is dominating on all the other variables, making variables in relation to transfer less impacting in favour of the variable in relation to the source of the DNA. This is not surprising

as $Hd_2$ stipulated that the POI has no previous encounter with the knife and should his/her DNA profile be present, it must be from the AO whose profile, by chance, matches the POI or some background DNA which, by chance, matches the POI. Given that we are considering a relatively large amounts of DNA for the POI (0.5-0.6 ng), the chance of an adventitious correspondence is very small. The range of LRs observed stems from the fact that the simulation process accounts for the distribution of the number of alleles that could be derived from that quantity of DNA and for each of them the match probabilities that could potentially be associated. In other words, it accounts for a large range of potential matching DNA profiles each of them having different match probabilities depending on the number of alleles and their designations. In this paper the match probabilities are used as a plug-in to explore the relative importance of the variables, they do not guide precisely the sub-source level likelihood ratio. They are reasonable in terms of order of magnitude for single profiles, but they fail to account for the impact of mixtures on the sub-source level likelihood ratio. In a given case, with a specified DNA allelic profile, that variability will be extremely reduced as the match probability node will be directly informed with the sub-source level likelihood ratio computed for the case at hand. It will lead to different overall likelihood ratio compared to the values reported in Figure 6. When mixtures are involved, the node "Match probability" will be informed by 1 over the LR obtained for that mixture considering sub-source propositions.

In this example, when considering $Hd_1$ and $Hd_2$, the overall range of the LRs is of 2 orders of magnitude but for a median log10(LR) of about 6 under $Hd_1$ and about 22 under $Hd_2$. The 2 orders of magnitude have then a greater impact under $Hd_1$ rather than $Hd_2$ with regards to how it might influence the decision making of the recipient. We shall see below (in section 3.4) that one way to reduce the range is to increase the number of data points informing the impacting variables, that would be worth the investment for a case involving $Hd_1$ and not for a case involving $Hd_2$.

Exploring the results in the Shiny application allows one to observe that, for some specific cases (a given amount of POI's DNA and not-POI's DNA), other variables may have a higher impact, hence be more critical to the case. That observation calls for a case specific approach if needed once a specific outcome has been observed in a given case. The general trend though is given by the variables identified globally.

### *3.3. Impact on the LR of the quantities of DNA, POI and not POI respectively*

To show the respective impact of the quantity of POI's DNA and not POI's DNA on the LR obtained (summarized by their median), we can show two situations linked to the above example. In the first, we will maintain the quantity of not POI's DNA (0.03-0.04 ng) and vary the amount of POI's DNA. In the second, we will keep fixed the amount of POI's DNA (0.5-0.6 ng) and vary the not POI's DNA quantity. Both are shown in Figure 7.

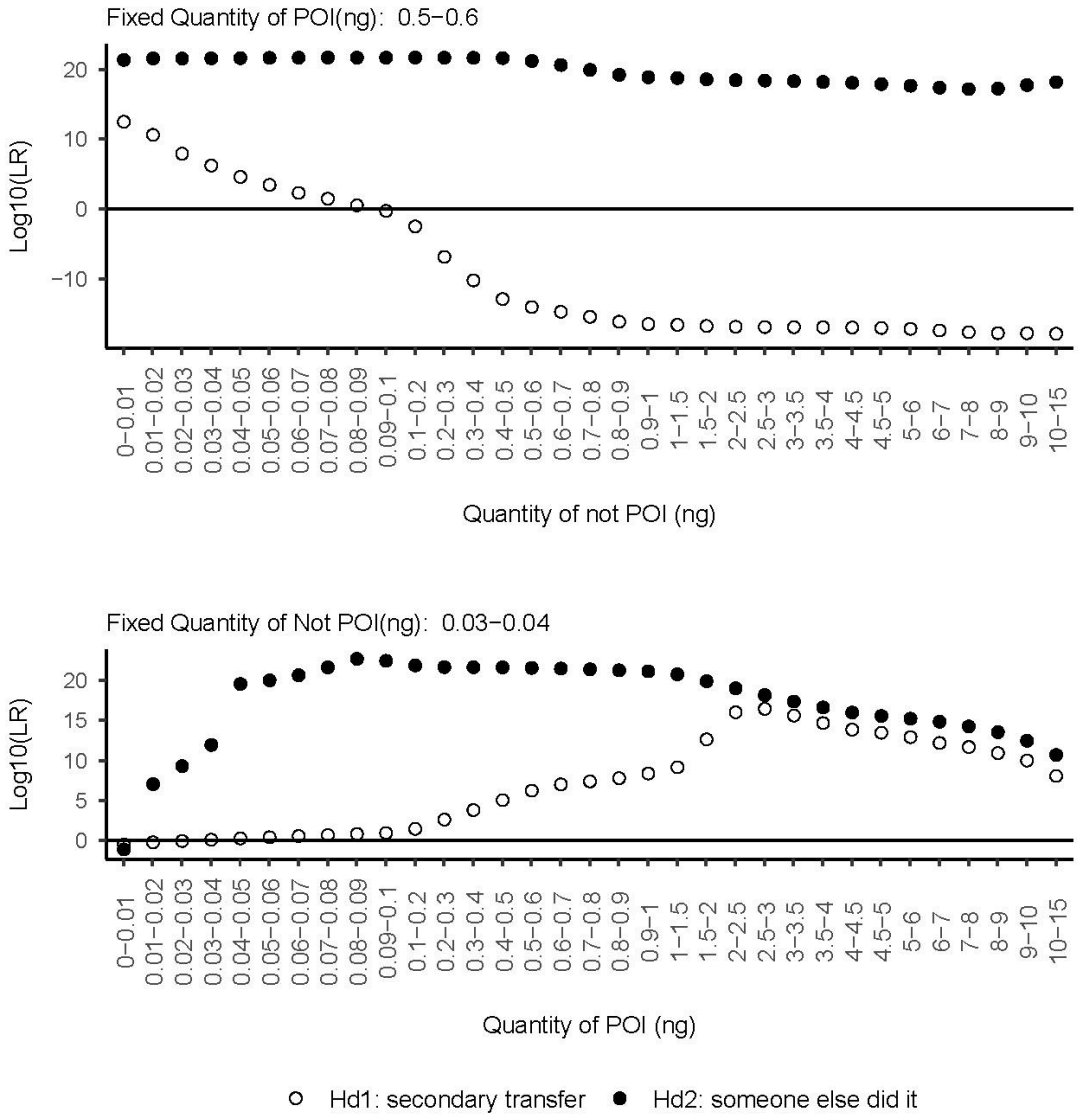



*Figure 7: LRs obtained when the amount of DNA (POI and non-POI) are respectively varied for a given amount of DNA that remain fixed.*

When the quantity of DNA of the POI is kept fixed (top graph of in Figure 7), under $Hd_2$, the presence of non-matching DNA has little impact on the LR. It is the fact that $Hd_2$ stipulates that the POI has no link whatsoever of with the knife, hence the probability that some background DNA or AO DNA, would correspond to the POI, is the key consideration. The amounts have almost no effect. Under $Hd_1$, however, the LR will be above 1 (the findings will provide support for Hp) with a maximum above one billion, but lower that 1 over the match probability (value under $Hd_1$)[1]. When the quantity of not POI's DNA is increased, the LR drops gradually. The clinching point (when the LR is equal to 1 or 0 on a log10 scale) is with 0.07−0.08 ng of non-matching DNA. Above that amount, the non-matching DNA becomes more compatible with the stabbing activity, hence the findings overall would lend support for the defence.

When the quantity of DNA of the not POI is kept fixed and we increase the amount the quantity of DNA corresponding to the POI (bottom graph of in Figure 7), under $Hd_2$, the likelihood ratio gradually increases with the increased amount DNA corresponding to the POI. The maximum LR is obtained when the quantity is the most expected when handling a knife (0.07−0.08 ng). Above that quantity, the LR will gradually reduce. Under $Hd_1$, the LR will increase only when the quantity of DNA corresponding to POI reaches the point that it is more compatible with the POI stabbing scenario than under the POI handshaking scenario. That clinching point is at the same quantity as the quantity of not POI's DNA (0.03−0.04 ng). That is logical because we have modelled the transfer probabilities in the same way for both POI and AO. Then, adding quantity of DNA corresponding to the POI will gradually increase the LR up until the quantity that is best expected under the primary transfer and not under the scenario of a secondary transfer. Then it will reduce again. The LRs lending support for the defence spans over a large range of POI corresponding DNA quantities despite the presence of non-matching DNA. It stems from the fact that the non-matching DNA is a quantity that is more compatible with background level than with a quantity you would expect following primary transfer.

The above considerations will change as a function of the choice of quantities, hence in the Shiny application, the one can find the same representations as in Figure 7 but for any given choice of POI or not POI quantities of DNA.

[1] Interestingly, while this is true of the values we found, there is nothing stopping the LR from being greater than 1 over the match probability (MP). If the probabilities of the transfers are extremely low, they can quite legitimately (in theory) lead to an LR that exceeds 1/MP.

### *3.4. Increasing the datasets informing impacting variables*

Should we judge that the overall variation on the ranges of LRs obtained is too high (such in the example shown in Figure 6 under $Hd_1$), one way to reduce it is to increase the number of data points informing the most impacting variables. Indeed, large variations in LRs translate limitations on the size of the datasets constituting the knowledge used to inform the CPTs. But before rushing into conducting specially designed experiments to acquire additional data to inform the relevant conditional probability tables, we can again take advantage of the simulation strategy developed in this research. It will allow to assess if the analytical investment if worth the effort, hence lead to a reduction of the ranges of LRs that will be beneficial for the case at hand.

We could proceed considering any combination of results (POI and not POI's DNA) or jointly for all of them. In Figure 8 we show the results for the case presented before (POI's DNA of 0.5-0.6 ng and not POI's DNA of 0.03-0.04 ng).

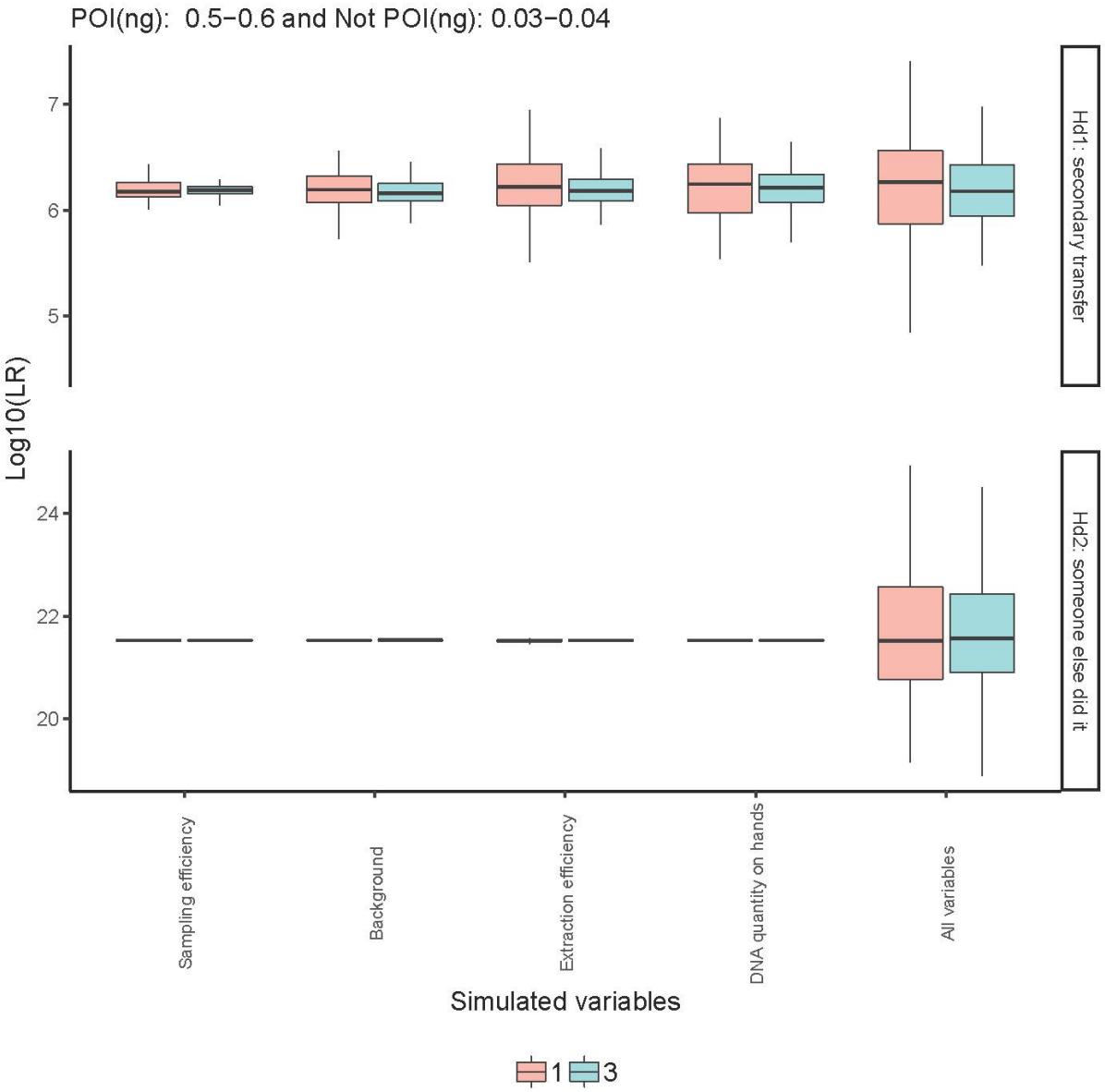



*Figure 8: Boxplots of the LRs (expressed in Log10) obtained after 100 simulations of each variable in both situations ($Hd_1$ and $Hd_2$) with a quantity of POI's DNA between 0.5 and 0.6 ng with between 0.03 and 0.04 of not POI's DNA. The boxplots in red are with the actual data, those in blue are obtained after increasing the underpinning data counts by a factor of 3 on the significant variables only.*

As expected, when probabilities are informed by an increased number of experiments, the ranges of likelihood ratios decrease. Under Hd1, the reduction is greater than under Hd2 due to the fact that the increase of data affected the significant variables (identified in Figure 4). For Hd2 however, the increase of data does not have a strong impact on the range of LR observed for "All variables". This is due to the dominant impact on the LR of the match probability that is not affected by this increase in data. If Hd2 is the defence proposition, there is no benefit in acquiring more data on the other variables that were significant under Hd1.

These simulations allow an assessment, in advance of investing in a large number of experiments to inform the CPTs, if the benefits, in terms of reduction of LR range, would be meaningful for the case at hand.

### *3.5. The discretisation of continuous variables*

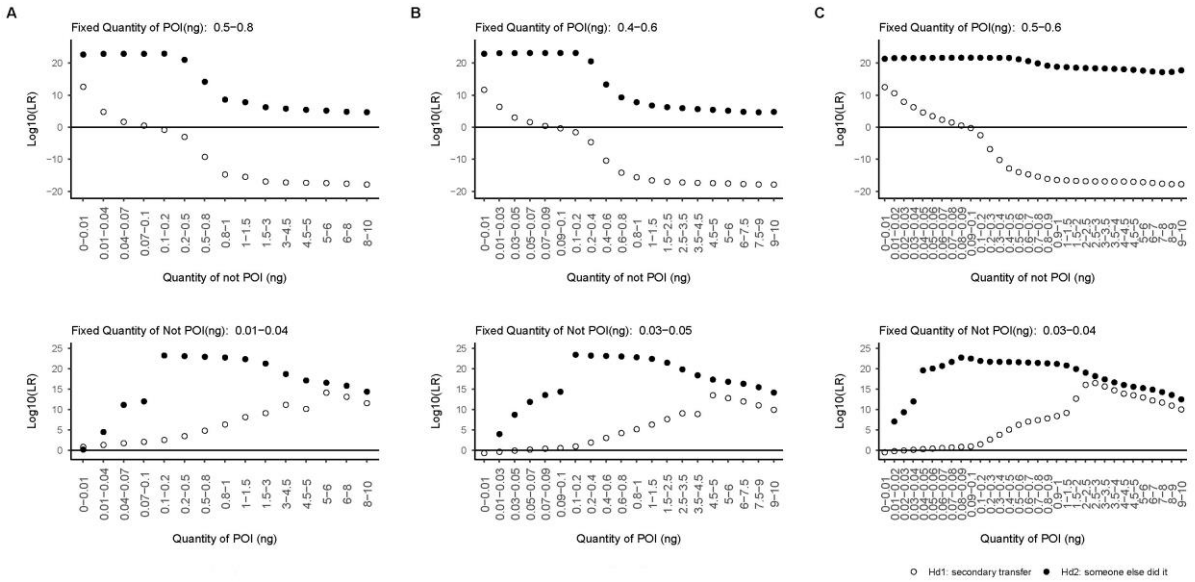
During the analysis setting up of nodes within the BN and the assignment of probabilities we have discretised the continuous variables, such as those dealing with DNA amounts, proportions of transfer, etc. This is a small step away from a fully Bayesian analyses of the data (and the idea of having parameters for the continuous distributions of variables or their priors), which would keep all continuous variables as continuous distributions. This is largely due to the limitations of the BN software. It would be possible to maintain continuous variables, and perhaps utilise stochastic sampling techniques, in a more customisable software, at the cost of a loss of comprehensibility. The discretisation itself could have an effect on the sensitivity of the LR to the data. As an initial investigation into the potential effects of different discretizations we performed simulations of the initial case using a BN whose states of the variables associated with the quantity of DNA (in ng) were discretized in three different ways as described in Table 11, called "New discretization 1" and "New discretization 2". Moving from the initial discretization, to New discretization 1, then to New discretization 2 represents increasing coarseness in describing the data.

| Variable | Possible states for the new discretization 1 | Possible states for the new discretization 2 |
|---|---|---|
| Results DNA POI | interval node;<br>0 to 0.01<br>0.01 to 0.09 in steps of 0.02<br>0.09 to 0.1 | interval node;<br>0 to 0.01<br>0.01 to 0.1 in steps of 0.03<br>0.1 to 0.2 |
| Results DNA not POI | 0.1 to 0.2 | 0.2 to 0.8 in steps of 0.3 |
| DNA Matching | 0.2 to 1 in steps of 0.2<br>1 to 1.5 | 0.8 to 1<br>1 to 1.5 |
| DNA DIFF | 1.5 to 4.5 in steps of 1<br>4.5 to 5 | 1.5 to 4.5 in steps of 1.5<br>4.5 to 5 |
| Background | 5 to 6<br>6 to 9 in steps of 1.5 | 5 to 6<br>6 to 10 in steps of 2 |
| Quantity on hands (ng) | 9 to 10<br>10 to 15<br>15 to 25<br>25 to 1000<br>1000 to inf | 10 to 15<br>15 to 25<br>25 to 1000<br>1000 to inf |

*Table 11: New discretizations of the states for the non-instantiated interval variables.*

The results show the following trends regarding the LRs obtained (Figure 9):

- In situation 1 (Hd1: secondary transfer), for a fixed DNA quantity of POI (Figure 9 top plots), or not POI (Figure 9, bottom plots), the trends taken by the LRs are similar, regardless the chosen discretization as shown by the white dots.
- For situation 2 (Hd2: someone else did it, black dots in in Figure 9), we have similar observations when the quantity of POI rises for a fixed quantity of not POI (bottom plots). However, for a fixed quantity of POI (top plots), we observed that the presence of non-matching DNA has little impact on the LR for the initial discretization (plot C). This is due to the fact that $Hd_2$ stipulates that the POI has no link whatsoever of with the knife, hence the probability that some background DNA or AO DNA, would correspond by chance to the POI, is the key consideration that is not impacted by the quantity of not POI. For coarser discretization (plot A and B), the log10 (LR) decreases more drastically when an increased quantity of not POI is obtained.
- We note that at the junction between bins of different sizes the LRs may drop or increase abruptly. For example, bottom plots A under Hd2 (black dots), when moving from 0.07-0.1 ng with the next bin at 0.1-0.2 ng, we observe an increase of the LR. The same is seen on plots under B (bin 0.09-0.1 to bin 0.1-0.2).




613

614 *Figure 9: LRs obtained when the amount of DNA (POI and non-POI) are respectively varied for a given amount*
615 *of DNA that remain fixed with the BN whose states of the variables were discretized with the new discretization 2*
616 *(A) or the new discretization 1 (B) or the initial discretization (C). The top plots refer to the situation where the*
617 *quantity that is varied is the DNA from not POI and the quantity of POI is fixed (the states varies only because of*
618 *the discretization). The bottom plots give the LRs for a fixed quantity of not POI and a varying quantity of DNA*
619 *corresponding to the POI.*

620

621 During our work on discretization, we noted that care must be exercised when changing bin
622 sizes. The relative sizes of adjacent bins have an impact. This is due to the fact the probabilities
623 for the nodes associated with the quantity of transferred DNA and the quantity of recovered
624 DNA are obtained by multiplying the previous quantities by a discount factor linked to the loss
625 of DNA. These probabilities, obtained though multiplication, will be affected by the sizes of
626 adjacent bins. In some cases (not reported in this paper), our discretization choices led to LRs
627 that were misleading (cases under Scenario 2 with LRs below 1). These results highlight the
628 need for a careful choice of the states. There is a need in the future to investigate how BN can
629 be constructed in a way that is less affected by multiplication factors.

630

631 To further illustrate the impact of discretization on the obtained LRs, we compared, in Table
632 12, the log10(LR) for four contrasting outcomes, in both scenarios:

633 • High quantity of POI (2 ng) with high quantity of not-POI (2 ng)
634 • High quantity of POI (2 ng) with low quantity of not-POI (0.03 ng)
635 • Low quantity of POI (0.03 ng) with high quantity of not-POI (2 ng)
636 • Low quantity of POI (0.03 ng) with low quantity of not-POI (0.03 ng).

| Situation | Outcomes | Log10 (LR) New discretization 2 | Log10 (LR) New discretization 1 | Log10 (LR) Initial discretization |
|---|---|---|---|---|
| Situation 1 (secondary transfer) | High quantity of POI (2 ng) High quantity of not-POI (2 ng) | -15 | -16 | -16 |
| | High quantity of POI (2 ng) Low quantity of not-POI (0.03 ng) | 9 | 12 | 18 |
| | Low quantity of POI (0.03 ng) High quantity of not-POI (2 ng) | -18 | -18 | -18 |
| | Low quantity of POI (0.03 ng) Low quantity of not-POI (0.03 ng) | 1 | 1 | 0.8 |
| Situation 2 (someone else did it) | High quantity of POI (2 ng) High quantity of not-POI (2 ng) | 21 | 20 | 18 |
| | High quantity of POI (2 ng) Low quantity of not-POI (0.03 ng) | 21 | 21 | 19 |
| | Low quantity of POI (0.03 ng) High quantity of not-POI (2 ng) | -18 | -17 | -13 |
| | Low quantity of POI (0.03 ng) Low quantity of not-POI (0.03 ng) | 4 | 8 | 12 |

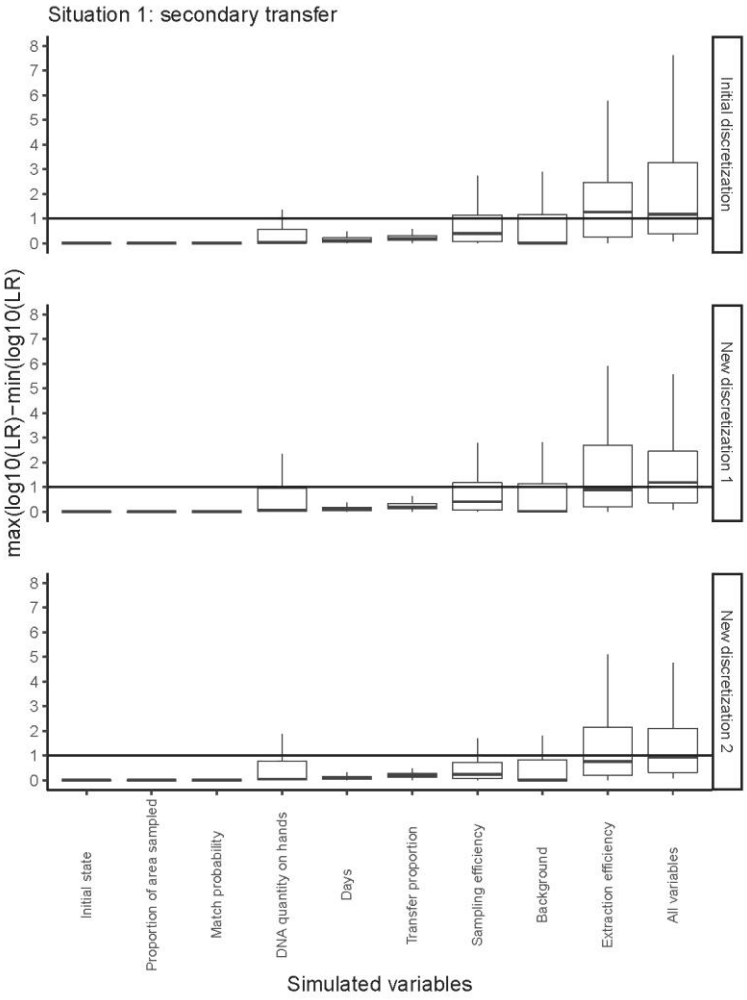
*Table 12: LRs (in log10) obtained for four outcomes, depending on the situation and the chosen discretization of the states.*

We observe that the choice of discretization can have an impact on the value of the LRs, depending on the choice of DNA quantities as findings. The fewer states there are, the less discrimination of the propositions is achieved, which means that smaller LRs can be obtained.

For example, when a low quantity of not-POI DNA is considered, coarser discretization (moving from initial to new discretization 2), leads to lower LR because of the loss of discrimination capability between the propositions.

Regarding the selection of the impacting variables, we observe an impact of discretization that is linked to the loss of discrimination capability when coarser discretization is adopted (Figure 10). This is especially the case for the new discretization 2. For all discretizations, the same variables are passing the threshold set to be declared significant, but we can expect that if an even coarser discretization would be chosen, some significant variables would drop below the threshold because of the loss of discrimination.

654


Situation 1: secondary transfer
max(log10(LR))−min(log10(LR))
Initial discretization
New discretization 1
New discretization 2
Initial state
Proportion of area sampled
Match probability
DNA quantity on hands
Days
Transfer proportion
Sampling efficiency
Background
Extraction efficiency
All variables
Simulated variables

655

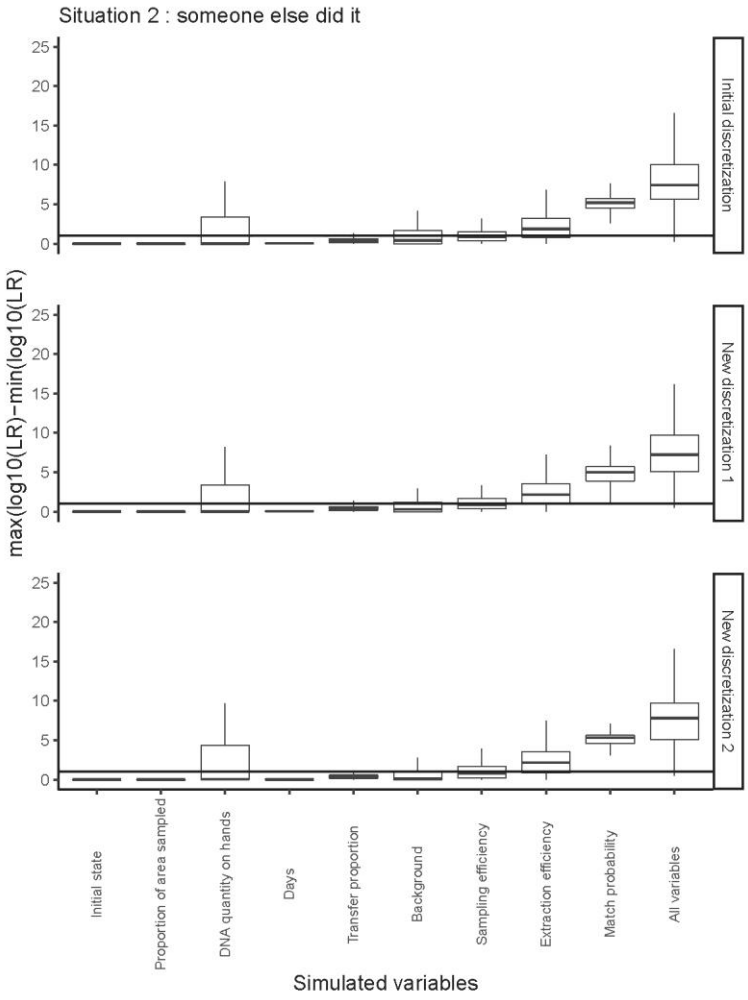



*Figure 10: Boxplots presenting the ranges (min-max) expressed in log10 of the LRs obtained following 100 simulations under Situation 1 and Situation 2. The panels present the results for the initial case using either the initial discretization of the states or the two new discretizations presented in this part. The horizontal line drawn at the difference of 1 (in log10) set the limit above the variable considered will be declared as having a significant effect on the global variability shown when "all variables" are resampled jointly.*

## 4. Conclusion

The ENFSI guideline [3] is advising forensic scientist to evaluate biological traces with a low level of DNA considering activity level propositions. For that task, specific variables such as transfer, persistence, recovery and background need to be considered. Many experts face practical difficulties when considering these variables for various perceived reasons. Typically, expert will indicate that:

– the number of variables at play is overwhelming and unmanageable;

- every case represents a unique set of circumstances and any numerical assignment of probabilities is critically dependant on the specificities of the case;

- the paucity of current published studies to inform the parameters associated with these variables cannot reasonably be compensated by reasonable data acquisition campaigns.

In this study, we show that Bayesian networks can handle this complexity efficiently, when coupled with simulation techniques, they can be used to identify the most impacting variables, hence reducing the data acquisition burden by directing the laboratory to the key issue.

The method has been applied to a scenario involving trace DNA recovered from knife handles where the prosecution alleges that the person of interest (POI) stabbed a victim. The findings considered take the form of given quantities of DNA (in ng) corresponding or not the POI. As a general tendency, regardless of the findings, we showed that when the defence claims that the POI has nothing to do with the incident, the match probability associated with the POI will dictate most of the weight to be assigned to the findings. If the POI defence is invoking the possibility of secondary transfer, the key variables are associated with the sampling, the extraction efficiency, the background and the quantity of DNA on the hands.

Simulation techniques can also be used to assess the merit of increasing the knowledge base (in terms of size of studies carried out) when the significant variables had been identified. We presented preliminary results on the impact of the choice of discretization of the variables. Discretization can have an impact on the LRs and potentially on the choice of impacting variables, mainly due to the loss of discriminability between the propositions when a too coarse discretisation is adopted. In our view, the number states and their ranges should be chosen carefully in a way that avoids losing information (e.g. merging states to a point where discrimination is lost).

Finally, we noted that the identification of significant variables depends on the obtained DNA results and this selection may be refined on a case by case basis. To allow one exploring all possibilities, a dedicated Shiny application has been designed (https://lydie-samie.shinyapps.io/DNA_Activity/).

## 5. Acknowledgements

We would like to express our gratitude to Marco De Donno for all his assistance in performing the simulation.

## 6. Bibliography

[1] F. Taroni, A.Biedermann, J. Vuille, N. Morling, Whose DNA is this? How relevant a question? (a note for forensic scientists), Forensic Sci. Int. Genet. 7 (2013) 467–470.

[2] C. Champod, DNA transfer: informed judgment or mere guesswork?, Frontiers in Genetics. 4 (2013) 300 in A. Biedermann, J. Vuille, F. Taroni, DNA, Statistics and the Law: A Cross-Disciplinary Approach to Forensic Inference, Frontiers Media. 4 (2014) 22-24.

[3] S. Willis et al., ENFSI guideline for evaluative reporting in forensic science, European Network of Forensic Science Institutes, Dublin (2015). http://enfsi.eu/wp-content/uploads/2016/09/m1_guideline.pdf (accessed May 6, 2019).

[4] A. Biedermann, C. Champod, G. Jackson, P. Gill, D. Taylor, J. Butler, N. Morling, T. Hicks, J. Vuille, F. Taroni, Evaluation of Forensic DNA Traces When Propositions of Interest Relate to Activities: Analysis and Discussion of Recurrent Concerns, Frontiers in Genetics. 7 (2016) 1–12.

[5] D. Taylor, A. Biedermann, L. Samie, K.-M. Pun, T. Hicks, C. Champod, Helping to distinguish primary from secondary transfer events for trace DNA, Forensic Sci. Int. Genet. 28 (2017) 155-177.

[6] F. Taroni, A. Biedermann, S. Bozza, P. Garbolino, C.Aitken, Bayesian Networks for Probabilistic Inference and Decision Analysis in Forensic Science, 2nd edition, Statistics in Practice, Chichester: John Wiley & Sons, Ltd (2014).

[7] D. Taylor, T. Hicks, C. Champod. Using Sensitivity Analyses in Bayesian Networks to Highlight the Impact of Data Paucity and Direct Future Analyses: A Contribution to the Debate on Measuring and Reporting the Precision of Likelihood Ratios. Science & Justice, 56 (2017) 402-410.

[8] P. Ka-Man, Interprétation des profils génétiques obtenus à partir des traces de contact, PhD thesis, School of Criminal Justice, University of Lausanne, 2016.

[9] R. Palmer, The Evaluation of Fibre Evidence in the Investigation of Serious Crime, PhD thesis, School of Criminal Justice, University of Lausanne, 2016.

[10] M. Khodabina, A. Ahmadabadi, Some properties of generalized gamma distribution, Mathematical Sciences 4 (2010) 9-28

[11] D. Dufresne, The Beta Product Distribution with Complex Parameters, Commun. Stat. —Theory

728 Methods. 39:5 (2010) 837–854

729 [12] D.J. Daly, C. Murphy, S.D. McDermott, The transfer of touch DNA from hands to glass, fabric and wood,
730 Forensic Sci. Int. Genet. 6 (2013) 41–46.

731 [13] L. Bontadelli, Study of DNA Shedder Quality, MSc thesis, School of Criminal Justice, University of
732 Lausanne, Lausanne, 2009.

733 [14] M. Goray, R.J. Mitchell, R.A.H.V. Oorschot, Investigation of secondary transfer of skin cells under
734 controlled conditions, Leg. Med. 12 (2010) 117–120.

735 [15] T.J. Verdon, R.J. Mitchell, R.A.H.V. Oorschot, Evaluation of tapelifting as a collection method for
736 contact DNA, Forensic Sci. Int. Genet. 8 (2014) 179–186.

737 [16] J.M. Bernado and A.F.M Smith, Bayesian Theory, John Wiley & Sons, Inc (2000)

738 [17] E. Butts, Exploring DNA Extraction Efficiency Forensics@NIST 2012 Meeting, Gaithersburg
739 https://www.nist.gov/sites/default/files/documents/oles/3_Butts-DNA-extraction-2.pdf (2017).

740 [18] G.E. Meakin, E. V. Butcher, R.A.H. van Oorschot, R.M. Morgan, The deposition and persistence of
741 indirectly-transferred DNA on regularly-used knives, Forensic Sci. Int. Genet. Suppl. Ser. 5 (2015) e498-
742 e500.

743 [19] RStudio Team (2015). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA URL
744 http://www.rstudio.com/.

745 [20] R Core Team (2019). R: A language and environment for statistical computing. R Foundation for
746 Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

747 [21] Kjell Konis. (2017). RHugin: RHugin. R package version 8.4. http://rhugin.r-forge.r-project.org] to link
748 with the developed Hugin network.

749 [22] L. Samie, T. Hicks, V. Castella, F. Taroni, Stabbing simulations and DNA transfer, Forensic Science
750 International: Genetics 22 (2016) 73–80.

751 [23] C. Gehrig, B. Balitzki, A. Kratzer, C. Cossu, N. Malik, V. Castella, Allelic proportions of 16 STR loci –
752 including the new European Standard Set (ESS) loci – in a Swiss population sample, Int. J. Legal Med.
753 128 (2013) 461–465.

754 [24] T. Hicks, F. Taroni, J. Curran, J. Buckleton, O. Ribaux, V. Castella, Forensic Science International :
755 Genetics Use of DNA profiles for investigation using a simulated national DNA database : Part I . Partial
756 SGM Plus 1 profiles, Forensic Sci. Int. Genet. 4 (2010) 232–238.

[25] Samie L., Champod C., Glutz V., Garcia M., Castella, V. & F. Taroni, The efficiency of DNA extraction kit and the efficiency of recovery techniques to release DNA using flow cytometry, Science & Justice, 59 (2019) 405-410