

Research paper

Source level interpretation of mixed biological stains using coding region SNPs

Guro Dørum^{a,*}, Øyvind Bleka^b, Peter Gill^{b,c}, Cordula Haas^a^a Zurich Institute of Forensic Medicine, University of Zurich, Switzerland^b Forensic Genetics Research Group, Department of Forensic Sciences, Oslo University Hospital, Norway^c Department of Clinical Medicine, University of Oslo, Norway

ARTICLE INFO

Keywords:

Forensic Science
Body fluid identification
mRNA
Coding region SNP (cSNP)
Body fluid mixtures
Source level propositions
Likelihood ratio

ABSTRACT

The association of body fluids/cell types and donors in mixed biological traces is an important, but challenging task required to evaluate the value of evidence given forensic propositions concerning the source of the DNA. The linking of a DNA profile with evidence from presumptive tests or RNA analysis is not straightforward. Coding region SNPs (cSNPs) are a novel type of evidential markers that are both cell type specific and individual specific. They thereby provide a direct link between a donor and a body fluid in mixed biological stains. In this proof-of-concept paper we consider the evaluation of cSNP profiles given source level propositions and explore the use of the open-source software EuroForMix to compute likelihood ratios. The discrimination power of the cSNPs for various body fluids is investigated with simulations. We provide case examples where the type of biological material is questioned and where cSNP profiles can be used to assign a donor to a body fluid, and discuss how the results can be reported in court.

1. Introduction

In mixed biological stains from a crime scene, it may be crucial not only to identify the body fluids present but also to assign each cell type to the correct donor. Following the hierarchy of propositions framework [1], courts are mainly interested in propositions at the activity level, such as “Mr X sexually assaulted Ms Y”. However, the forensic geneticist is usually restricted to make statements about the evidence given propositions at the sub-source level such as “Mr X is the donor of the DNA”, where the strength of evidence is based upon analysis of DNA profiles (usually STRs). The focus of this paper is evaluation of evidence given source level propositions which are concerned with the cell type that the DNA is derived from, such as “Mr X is the donor of blood”. The DNA commission of the ISFG [2] provides guidelines for formulation and evaluation of evidence related to propositions at different levels, and points out that a likelihood ratio calculated given propositions at the sub-source level cannot be carried over to propositions at the source level. This means that a likelihood ratio calculated for a DNA profile to assess the value of the evidence in support of the ‘identity’ of the DNA donor, cannot be used in statements about the donor of blood, unless it is obvious that the DNA originates from blood and not some other cell type. In a case presented by Gill [3] (wrongful arrest of Adam Scott) a man was accused of rape because his

DNA profile, originating from an undetected saliva contamination, was automatically assumed to derive from the detected semen in the stain.

Traditionally, presumptive tests have been used to identify the presence of certain body fluids. Some of these tests are of limited specificity [4] and there are no reliable tests for vaginal secretion and menstrual blood, two commonly observed forensic body fluids. Over the last few years RNA has been shown to be a good predictor of the body fluids present in a stain. mRNA profiling relies on the differential expression of mRNAs in different tissues. A number of mRNA markers have been identified for forensically relevant body fluids/cell types, i.e. blood, saliva, semen, vaginal secretion, menstrual blood, sweat, nasal mucosa, nasal blood and skin [5–16]. Expression of body fluid specific transcripts is abundant in the respective body fluid, but not necessarily absent in others. Therefore, scoring systems or statistical prediction tools are used to interpret the RNA results [11,17–19].

Although RNA or presumptive tests may contribute important evidence regarding the source of the DNA, when there is a mixture, the task of assigning a donor to a specific body fluid is not straightforward. Hartevelde et al. [20] investigated whether the height of DNA and RNA signals may guide association of donor and cell type in a mixture. Clearly, the gender-specificity of certain body fluids (semen, vaginal secretion, menstrual blood) can be instructive. However, the authors discourage associating cell types and donors based on signal

* Correspondence to: University of Zurich, Zurich Institute of Forensic Medicine, Forensic Genetics, Winterthurerstrasse 190/52, CH-8057 Zurich, Switzerland.
E-mail address: guro.dorum@irm.uzh.ch (G. Dørum).

<https://doi.org/10.1016/j.fsigen.2022.102685>

Received 4 August 2021; Received in revised form 1 March 2022; Accepted 4 March 2022

Available online 20 March 2022

1872-4973/© 2022 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

Table 1
Probability of identity (PI) for each body fluid, where L is the number of cSNPs.

Blood ($L = 11$)	Saliva ($L = 3$)	Semen ($L = 8$)	Vaginal ($L = 3$)	Menstrual ($L = 3$)	Skin ($L = 7$)
3.88×10^{-4}	0.258	1.65×10^{-3}	0.176	0.136	2.66×10^{-3}

heights when performing combined RNA and DNA analyses. Their conclusion was also supported by our own experiments using read counts from massively parallel sequencing (MPS). We observed that the mixture ratios could vary greatly between DNA and RNA [21]. Several publications propose the use of Bayesian networks to combine DNA and body fluid evidence when propositions at the source level are considered [22–24]. These studies mainly focus on the use of presumptive tests to identify body fluids, and in addition a large number of casework samples are needed to assign the conditional probabilities for such a network.

In a previous paper [21] we introduced a set of 35 coding region SNPs (cSNPs) in body fluid specific transcripts analysed with MPS to assign a body fluid in a mixture to a specific individual. These cSNPs were chosen specifically for each of the six forensically interesting body fluids/cell types blood, saliva, semen, vaginal secretion, menstrual blood and skin, with the aim of being highly discriminating among individuals. cSNPs can be derived both from DNA in reference profiles and from RNA in crime stains. The cSNPs in a stain can be compared to the cSNPs in a reference profile, and consequently there is a direct link between body fluid and donor that can be quantified.

Here we consider the evaluation of cSNP profiles given source level propositions, both in scenarios where the donors have contributed different body fluids, and where they have contributed the same body fluid. We explore the use of the STR mixture software EuroForMix [25] for computation of likelihood ratios to quantify the evidence provided by the cSNP results. EuroForMix takes peak heights/read counts into account and has previously been successfully applied to SNP mixtures analysed with MPS [26]. Since the number of cSNP markers and their sensitivity and specificity differ among the six body fluids, we use simulations to highlight the theoretical discriminatory power of the different body fluids. Finally, we compute likelihood ratios for a number of real stains and present some mock case examples.

2. Methods

2.1. cSNPs

Our cSNP panel consists of 35 coding region SNPs situated in body fluid/cell type specific transcripts [21]. Among the 35 cSNPs there are 11 blood, 8 semen, 3 saliva, 3 vaginal secretion, 3 menstrual blood and 7 skin specific markers (Supplementary Table S1). A cSNP profile derived from RNA in a crime stain should only expose cSNPs specific for the body fluid(s) present in the stain. In addition to being body fluid specific, the cSNPs were chosen to be discriminatory among individuals. A link between a body fluid and a donor can be made by comparing a reference cSNP profile with the stain cSNP profile. Reference cSNP profiles are derived directly from DNA and consequently show genotypes in all body fluids. We hereby refer to the stain profile as the RNA-cSNP profile and the reference profile as the DNA-cSNP profile. The allele frequencies for the 35 cSNPs are estimated from 188 individuals as described in [21]. Table 1 shows the probability of identity (PI), i.e. the probability that two unrelated individuals have the same genotype, for the set of cSNPs for each body fluid. The PI is calculated per locus as the sum of the squared genotype frequencies, and multiplied over all loci for the given body fluid [27,28].

Fig. 1 shows a constructed example of a 1:1 mixture where donor 1 has contributed saliva and donor 2 has contributed vaginal secretion. Only the saliva and vaginal secretion specific markers are expressed in the RNA-cSNP profile from the stain, hence only these markers are

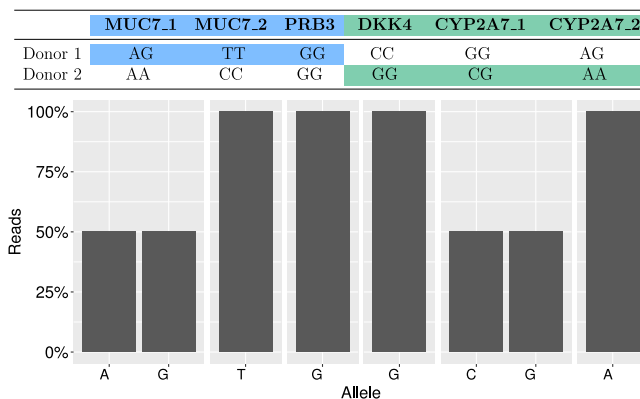


Fig. 1. A toy example of a saliva–vaginal secretion mixture. The table displays the DNA-cSNP profiles of the body fluids in question for the two donors. Donor 1 has contributed saliva (blue markers) and donor 2 has contributed vaginal secretion (green markers). The barplot shows the corresponding read counts in the RNA-cSNP profile relative to the total number of reads per marker. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

shown. The donors can be distinguished based on their DNA-cSNP reference profiles. Fig. 2 shows a constructed example where both donors in a 1:1 mixture have contributed semen. The read counts for the RNA-cSNP profile indicate that it is a mixture; the heterozygous markers show allele ratios that deviate from the expected 1:1 in a single donor stain.

Figs. 1 and 2 show idealised mixture stains; the mixture proportion (M_x) is 0.5, all donor alleles are present and there are no unspecific reads or heterozygous imbalance. In our previous paper [21] we noted that some cSNPs were prone to allelic dropout in the RNA-cSNP profile. In addition, some drop-in alleles were observed.

2.2. Evaluation of RNA-cSNP profiles

We consider individually the two scenarios where: (1) the donors have contributed different body fluids and (2) the donors have contributed the same body fluid. In the first scenario where there is one donor per body fluid, the RNA-cSNP profile can be evaluated similarly to a single source STR profile. In the second scenario where there are two or more donors of one body fluid, the RNA-cSNP profile can be evaluated similarly to a mixed STR profile. Since the RNA-cSNP markers are specific to a body fluid/cell type, we evaluate each body fluid separately. We assume throughout this paper that the markers carry independent information, however in the discussion we consider the issue with linked markers.

2.2.1. Conditioning on sub-source results

In casework, before the results of tests are known, the activity level propositions are set by the mandating authorities [29]. In order to help the court address the activity level, it is necessary to evaluate the evidence at both the sub-source and source levels (here we do not consider the activity level itself). Before source level propositions are formulated, we assume that the court has agreed with sub-source level propositions regarding the identity of the DNA donors [2]. As an example, assume that the sub-source level propositions are H_p : The DNA came from the victim and the suspect and H_d : The DNA came from the victim and an unknown. Based on the STR results a very high LR is calculated, and H_p is accepted by both parties. This means that the suspect and victim are accepted as the only donors to the stain and we can discount any unknowns in the source level propositions. On the other hand, if the sub-source propositions are H_p : The DNA came from the suspect and an unknown and H_d : The DNA came from two unknowns, and H_p is accepted by both parties, the source level propositions will

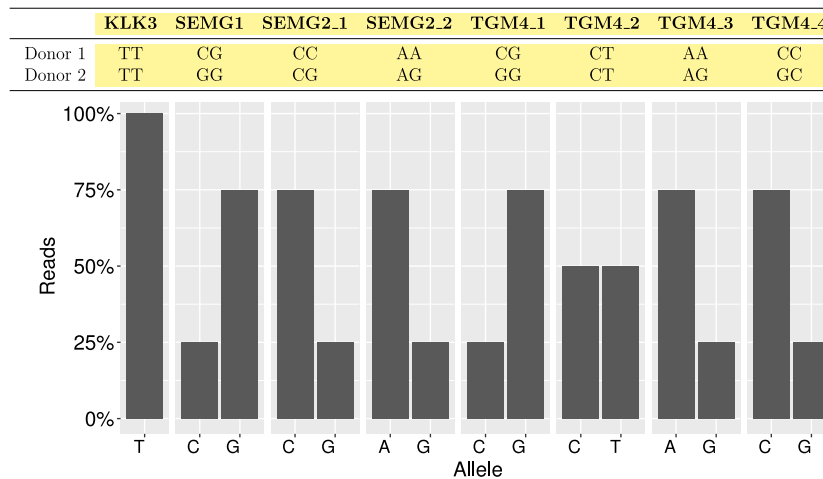


Fig. 2. A toy example where both donors have contributed semen. The table shows the DNA-cSNP profiles of the donors in the semen markers, and the barplot shows the read counts per allele in the RNA-cSNP profile relative to the total number of reads per marker.

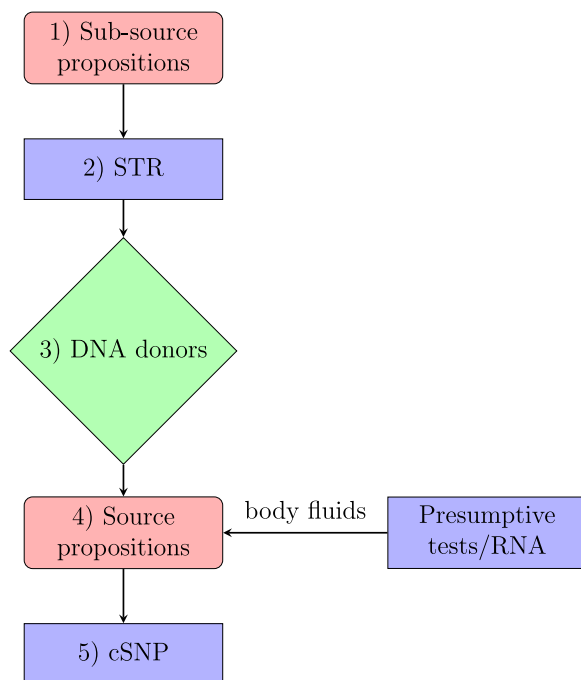


Fig. 3. Hierarchy of propositions and evaluated data. (1) Sub-source propositions are defined. (2) The STR profile is evaluated given the sub-source propositions. (3) The DNA donors present in the stain, i.e. the number of contributors and any known donors, are agreed upon. (4) Source propositions are defined conditioned on the DNA donors, and alternatively on body fluid information from presumptive tests/RNA analysis. (5) The cSNP profile is evaluated given the source propositions.

include an unknown. The results at sub-source level together with body fluid information from mRNA analysis or presumptive tests and/or other case specific information, form the basis for the source level propositions. See Fig. 3 for an overview of the hierarchy.

Another important assumption we make, based on our experience, is that cSNPs are less sensitive than STRs. Although two DNA donors are accepted at sub-source level, we may observe a cSNP profile for only one of the donors. The second DNA donor may have contributed a body fluid/tissue not among the six defined in our cSNP assay, or the body fluid is present in such small amounts that it is not detected. On the other hand, we may also observe cSNP profiles for three body

fluids in a two-person DNA mixture, if one (or both) of the donors has contributed several body fluids.

2.2.2. Source level likelihood ratio

Let R denote the RNA-cSNP profile extracted from a stain. R can be divided into six body fluid specific profiles (blood, saliva, semen, vaginal secretion, menstrual blood and skin), where R_j denotes the profile for body fluid j . Assuming no unspecific reads, R will be empty in markers specific to body fluids not present in the stain. Further, let g denote the DNA-cSNP reference profile of a person of interest. Since g is derived from DNA it contains genotype information in all the body fluid specific markers, however we only consider the markers specific to the body fluid in question, denoted g_j . The DNA-cSNP profiles of any undisputed donors of the body fluid are denoted collectively by I_j . Given two source level propositions H_p and H_d , the likelihood ratio that evaluates the cSNP profile for body fluid j is defined as

$$LR_j = \frac{P(R_j, g_j | H_p, I_j)}{P(R_j, g_j | H_d, I_j)} = \frac{P(R_j | H_p, g_j, I_j)P(g_j | H_p, I_j)}{P(R_j | H_d, g_j, I_j)P(g_j | H_d, I_j)} \quad (1)$$

Assuming that information on genotyped individuals does not vary between the prosecution and defence propositions (individuals are assumed to be unrelated), the LR simplifies to

$$LR_j = \frac{P(R_j | H_p, g_j, I_j)}{P(R_j | H_d, g_j, I_j)} \quad (2)$$

Typically, the propositions dispute the donor of a given body fluid, e.g.

H_p : The person of interest contributed body fluid j ,

H_d : An unknown individual contributed body fluid j .

If it has been concluded at sub-source level that the person of interest and one or more unknown individuals have contributed, then H_d may postulate that one (or more) of the unknowns contributed this body fluid (and the person of interest did not). If there are, besides the person of interest, other known individuals whose contribution is undisputed, then H_d may postulate that one (or more) of them contributed body fluid j (and the person of interest did not).

With a continuous model, R_j will be accompanied by read counts for the observed alleles. Following the approach in Bleka et al. [25] we analysed the cSNP profiles with the continuous model in EuroForMix v3.1.0 [26] without a degradation or stutter model. Likelihood ratios were calculated with the maximum likelihood approach with $F_{st}=0.01$, a drop-in probability of 0.05, and the parameter for modelling the reads of drop-in alleles (λ) set to 0.01. The drop-in values were set to the default values in EuroForMix since we lack statistics on drop-in events in cSNP data. In addition, the step tolerance parameter was set

to $1e-6$ (default value is $1e-3$) to make the optimisation more robust. A minimum detection threshold of 50 read counts was applied.

Theoretically, a set of propositions could be formulated and an LR could be computed for each body fluid present in an RNA-cSNP profile, and the court could be presented with a number of body fluid specific likelihood ratios. A more plausible scenario, however, is that one specific body fluid is of interest to the case.

2.2.3. Confirmatory tests in the likelihood ratio framework

In certain types of cases the likelihood ratio cannot be quantified, however it can be used as a confirmatory test. Confirmatory tests are used to positively identify a body fluid, see e.g. [4] for a review. Although these tests are able to ascribe the presence of a body fluid, they cannot assign the body fluid to a given individual in mixed samples. The advantage of cSNPs is the ability to associate a given body fluid to a given contributor.

One example of confirmatory tests arises in single source stains where it is accepted at sub-source level that the person of interest is the only contributor. Another example is a two-person mixture where a male and a female donor are accepted as the only contributors, but the body fluid of interest is gender specific. In both scenarios there are no alternative donors of the body fluid, and the propositions would dispute the body fluid rather than the donor, e.g.:

H_p : The person of interest contributed body fluid j ,

H_d : The person of interest contributed some other body fluid/cell type.

If the cSNPs are highly body fluid specific, meaning that detection of an R_j happens with probability almost zero if no one has contributed body fluid j , then these can be used to conclude the presence of a body fluid, when detected. Hence, if the propositions differ in whether or not body fluid j was contributed to the stain, then the LR approaches infinity in favour of the proposition that states that body fluid j is present.

2.2.4. Example 1: Confirmatory test

The following example is based upon the case of Regina v. Weller [30] described in Gill [3, pp. 43–46]. The suspect (Mr S) and victim (Ms V) were at a party. Ms V became drunk and was partially incapacitated. She claimed that she was digitally penetrated by Mr S. He denied this, saying that he only looked after her while she was ill. A mixed profile was found under the fingernails of the suspect's left hand comprising the suspect and victim (the latter was a minor contributor). There was no evidence of the body fluid type; at court the possible methods of transfer were discussed:

- contact with the hair of the victim,
- touching the victim whilst putting her to bed,
- insertion of fingers into vagina.

The prosecution asserted that the latter proposition was most likely whilst the defence asserted (a) and (b). If the source of the victim's DNA was not vaginal cells, then (c) cannot be supported. In the absence of a test, evaluation was carried out according to the expectations of finding a given quantity of DNA, if the activities mentioned above were true. No direct evidence was adduced to show that vaginal cells were present; the presentation was non-probabilistic, relying upon the scientist's experience instead. The defendant was found guilty.

Here we rework the Weller case to show how the formulation of propositions would follow with source level. The evaluation is based on constructed cSNP data. The following activity level propositions are considered:

H_p : Mr S sexually assaulted Ms V by digital penetration and had social interaction,

H_d : Mr S did not assault Ms V, he only had social interaction and helped her when she was ill.

Table 2

Constructed cSNP data in the reworked Weller case example showing the cSNP profile recovered from underneath fingernails of Mr S (RNA-cSNP) and reference profiles (DNA-cSNP) of Ms V and Mr S.

Marker	Body fluid	RNA-cSNP		DNA-cSNP	
		Allele	Read count	Ms V	Mr S
CYP2A7_1	Vaginal	C	5241	C/C	A/C
CYP2A7_2	Vaginal	T	4829	T/T	C/T
DKK4	Vaginal	A/G	2698/2417	A/G	G/G
COL17A1_1	Skin	C	5241	C/C	C/C
COL17A1_2	Skin	C	4829	C/C	C/C
COL17A1_3	Skin	A/G	2698/2417	A/A	A/G
KRT77_1	Skin	A/C	2377/2581	A/C	A/C
KRT77_2	Skin	C/T	2304/1900	T/T	C/T
LCE1C_1	Skin	A	5391	A/A	A/A
LCE1C_2	Skin	G	5267	A/G	G/G

Upon analysis, a full mixed STR profile was found underneath the fingernails of Mr S that supported the proposition that the donors were Mr S and Ms V (the latter was the minor contributor). However, the sub-source statistic is of little benefit, since it was common ground that social interaction had occurred at the party. Therefore, by itself, this does not help the court. The next level in the hierarchy of propositions is the source level. Since Mr S and Ms V were accepted as the only contributors at sub-source level, Ms V is the only possible vaginal secretion donor. This leads to the propositions:

H_p : Vaginal cells from Ms V were recovered from underneath fingernails of Mr S,

H_d : Some other cell type from Ms V was recovered from underneath fingernails of Mr S.

The source level describes the critical difference of positions between prosecution and defence.

Table 2 presents the RNA-cSNP profile obtained from underneath the fingernails of Mr S, and DNA-cSNP reference profiles of both Ms V and Mr S. Note that Mr S's reference profile also includes the vaginal secretion markers since the reference profiles are based on DNA. The RNA-cSNP profile shows genotypes for the vaginal secretion and skin markers, which suggests the presence of both cell types. No other body fluids appear in the profile. Ms V's reference profile 'matches' the vaginal secretion markers, but not the skin markers. Mr S's profile 'matches' the skin markers. Let R_{vag} denote the alleles in the vaginal secretion markers in the RNA-cSNP profile and g_{vag} denote Ms V's DNA-cSNP profile in the vaginal markers. There are no undisputed donors present in the profile, i.e. I_{vag} is empty. The source level likelihood ratio is

$$LR_{vag} = \frac{P(R_{vag} | H_p, g_{vag})}{P(R_{vag} | H_d, g_{vag})}. \quad (3)$$

Since Ms V is the only possible vaginal secretion donor, it follows that $P(R_{vag} | H_p, g_{vag}) \approx 1$ and $P(R_{vag} | H_d, g_{vag}) \approx 0$. The LR approaches infinity and the cSNPs confirm Ms V as the donor of vaginal fluid.

Note that in this example it was a prerequisite that the number of contributors to the stain was restricted to the two known contributors Mr S and Ms V, because this was agreed upon by both parties at sub-source level. If the sub-source results had indicated the presence of an unknown female contributor, a likelihood ratio could have been computed, as is demonstrated in Section 2.2.5.

The source level does not address the activity level propositions, i.e. the LR calculated at source level cannot be carried over. This is because the activity level incorporates probabilities of transfer, persistence and recovery of body fluids. In particular, we would need to consider the probability of secondary transfer and/or contamination. This evaluation will lead to a different LR that is unrelated to the source level. In order to carry out such an assessment, different data are needed, e.g. incidence of the body fluid as background on the hands or clothing of the victim and an assessment of whether secondary transfer to the fingers of Mr S is possible. This is beyond the scope of this paper.

2.2.5. Example 2: Likelihood ratio — one donor per body fluid

We extend the example in Section 2.2.4 to assume that a mixture of three people was found underneath the suspect's fingernails: Mr S, Ms V and an unknown female. Mr S claims that he had consensual sexual activity with a girlfriend prior to the party, but this girlfriend is not available for a reference profile. The sub-source propositions:

H_p : The DNA came from Mr S, Ms V and an unknown contributor,

H_d : The DNA came from Mr S and two unknown contributors,

resulted in an LR of 1 billion. Assuming the court has accepted the H_p sub-source proposition, there is no dispute about the presence of Ms V's DNA. That leads to the following source level propositions:

H_p : Vaginal cells from Ms V only were recovered from underneath fingernails of Mr S,

H_d : Vaginal cells from an unknown female were recovered from underneath fingernails of Mr S. Ms V contributed some other cell type.

We used the data as shown in Table 2. The allele ratios of about 1:1 in the heterozygous markers suggest only one donor of vaginal secretion, however the donor could be the unknown female, as is proposed by the defence. Let R_{vag} denote the alleles (with corresponding read counts) in the vaginal markers in the RNA-cSNP profile and g_{vag} denote the DNA-cSNP profile of Ms V in the vaginal markers. There are no undisputed persons present in the profile, i.e. I_{vag} is empty. The likelihood ratio of interest is

$$LR_{vag} = \frac{P(R_{vag} | H_p, g_{vag})}{P(R_{vag} | H_d, g_{vag})}.$$

For a single source profile with no dropout and drop-in, the LR is approximately equal to 1 divided by the random match probability, i.e. the probability that a random person in the population has this particular vaginal secretion cSNP profile. The analysis with EuroForMix concluded that it is 38 times more likely to observe the cSNP vaginal secretion profile if Ms V is the donor rather than if an unknown female is the donor (Supplementary Figure S1).

Note that a separate likelihood ratio could be computed with the skin markers as evidence, given a suitable set of propositions. However, the vaginal secretion markers are the incriminating evidence, and the presence of Ms V's skin cells would not diminish the value of this evidence.

2.2.6. Example 3: Likelihood ratio — two donors per body fluid

A woman (Ms V) was raped during a party at her boyfriend's flat. She was intoxicated and could not remember details from the assault. The boyfriend's roommate (Mr S) is a suspect, but he denies the incident and claims that he only had social interaction with her. Before the party Ms V had consensual intercourse with her boyfriend (Mr B). The activity level propositions are:

H_p : Ms V had consensual intercourse with Mr B at time t_1 and was raped by Mr S at time t_2 ,

H_d : Ms V had consensual intercourse with Mr B at time t_1 and was raped by an unknown at time t_2 . Ms V and Mr S only had social interaction.

The offence was reported two days after the crime, so no vaginal swab was taken. However, a stain on the victim's underwear gave a positive result for semen with a presumptive test. The Y-STR profile recovered from the stain showed a mixture of three males: a major profile which 'matched' the Y-STR profile of Mr B, a minor profile that 'matched' that of Mr S and a minor unknown profile. The sub-source propositions:

H_p : The DNA came from Mr B, Mr S and an unknown contributor,

H_d : The DNA came from Mr B and two unknown contributors,

resulted in a high LR. Provided that the court accepts the evidence, the presence of Mr S's DNA is not disputed. The following source level propositions were considered:

H_p : Mr B and Mr S contributed semen,

H_d : Mr B and an unknown contributed semen; Mr S contributed skin cells or another body fluid.

Table 3

cSNP profile in the rape case example showing the alleles and read counts in the stain (RNA-cSNP) and the reference profiles of Mr B and Mr S (DNA-cSNP).

Marker	Body fluid	RNA-cSNP		DNA-cSNP	
		Allele	Read count	Mr B	Mr S
KLK3	Semen	G	4949	G/G	G/G
SEMG1	Semen	A/T	854/4146	T/T	A/T
SEMG2_1	Semen	A/C	2626/2418	A/C	A/C
SEMG2_2	Semen	A	4679	A/A	A/A
TGM4_1	Semen	G/T	2930/2054	G/T	G/G
TGM4_2	Semen	G	4439	G/G	G/G
TGM4_3	Semen	C/G	2487/2447	C/G	C/G
TGM4_4	Semen	A/G	3021/1791	A/G	A/A

Note that under H_p , the unknown donor is assumed to have contributed some other body fluid than semen.

Table 3 shows constructed cSNP data for this example: DNA-cSNP reference profiles for Mr B and Mr S in addition to the RNA-cSNP profile from the victim's underwear. The RNA-cSNP profile only shows genotypes in the semen cSNPs. Thus, if Mr S has not contributed semen to the stain, the source of his DNA must be an undetected body fluid or cell type. The same is implied for the unknown donor. The table also displays the RNA read counts for each observed allele. The allele ratios in the heterozygous markers differ from the expected 1:1 for a single source profile, which suggests more than one donor. Let R_{sem} denote the RNA-cSNP profile in the semen markers, g_{sem} denote Mr S's genotype and I_{sem} denote Mr B's profile who is an undisputed donor. The source level likelihood ratio was calculated as

$$LR_{sem} = \frac{P(R_{sem} | H_p, g_{sem}, I_{sem})}{P(R_{sem} | H_d, g_{sem}, I_{sem})}.$$

The likelihood ratio evaluates the probability that a random person in the population has a semen cSNP profile that fits with the RNA-cSNP profile. Analysis with EuroForMix concluded that it is 4291 times more likely to observe the cSNP semen profile if the stain is a mixture of Mr B and Mr S rather than if it is a mixture of Mr B and an unknown contributor (Supplementary Figure S2). Under H_p , the mixture proportions were estimated as 0.72:0.28 for Mr B:Mr S.

2.3. Simulations

We used simulations to investigate the distribution of likelihood ratios that can be expected for cSNP profiles, with the purpose of highlighting the theoretical discriminatory power of the different body fluids. Each body fluid was considered separately. To carry out the simulations we made the simplifying assumption that all cSNPs within a body fluid have the same expected read counts. No unspecific reads were simulated. We used the function `genDataset()` in the EuroForMix R package to simulate cSNP profiles based on real allele frequencies. This function generates random read counts from the gamma distribution based on the provided expected read count (μ) and coefficient of variation (σ). These parameters were estimated from the heterozygous single-contributor alleles in the 29 mixture stains in Table S2 (described in Section 2.4). In the simulations we used a drop-in probability of 0.05, λ for modelling the reads of drop-in alleles set to 0.01, and no stutter or degradation. The same parameter values were used in the computation of likelihood ratios. A minimum detection threshold of 50 read counts was applied and the step tolerance parameter was set to 1e-6. Fst correction was not applied because the simulation function does not have the option to simulate data with population substructure.

To evaluate the different body fluids' ability to discriminate between the propositions H_p and H_d , we used Receiver Operating Characteristics (ROC) curves [31]. For a chosen threshold t , a true positive is defined as the event where $LR \geq t$ when H_p is true. Likewise, a false

positive is defined as the event where $LR \geq t$ when H_d is true. The true positive rate (TPR) is calculated as

$$TPR(t) = \frac{1}{N} \sum_{i=1}^N I(LR_i \geq t) \quad (4)$$

where I is 1 if $LR_i > t$ and 0 otherwise, and LR_i is the likelihood ratio computed from simulation $i = 1, \dots, N$. Since events under H_d where $LR \geq t$ are rare for large values of t , we used importance sampling to compute the false positive rate (FPR):

$$FPR(t) = \frac{1}{N} \sum_{i=1}^N I(LR_i \geq t) \cdot \frac{1}{LR_i}, \quad (5)$$

where LR_i is the likelihood ratio computed from simulation $i = 1, \dots, N$ under H_p [32]. One ROC curve was constructed per body fluid. For each ROC curve we estimated the area under the curve (AUC). AUC = 1 indicates a perfect discrimination between propositions, while AUC = 0.5 indicates that we are not able to discriminate between propositions with the given body fluid markers.

2.3.1. One donor per body fluid

Read counts for $N = 1000$ one-contributor RNA-cSNP samples were simulated separately for each body fluid. We assume that it has been accepted at sub-source level that there is a second, unknown DNA donor in the stain. For each sample we calculated the likelihood ratio for the following set of propositions:

- H_p : The true donor contributed the body fluid,
- H_d : The unknown second donor contributed the body fluid.

2.3.2. Two donors per body fluid

Read counts for $N = 1000$ two-contributor RNA-cSNP mixtures were simulated for each body fluid. The expected read count was set to 2μ . The mixture ratios used were 1:1 (both donors are equal contributors), 3:1 (donor 1 is a major contributor and donor 2 is a minor contributor) and 1:3 (donor 1 is a minor contributor and donor 2 is a major contributor). We assume that it has been accepted at sub-source level that there is a third, unknown DNA donor in the stain. For each sample we computed the likelihood ratio for the following set of source level propositions:

- H_p : Donor 1 and donor 2 contributed the body fluid,
 - H_d : Donor 1 and the unknown third donor contributed the body fluid.
- Note that for the samples with mixture ratio 3:1 we condition on the major contributor in the propositions, and with mixture ratio 1:3 we condition on the minor contributor.

2.4. Real data

The real data samples analysed here are described in Ingold et al. [21]. The data set consists of RNA-cSNP profiles from 29 two-person mixtures where the donors have contributed different body fluids and 15 two-person mixtures where the donors have contributed the same body fluid (Supplementary Tables S2 and S3). These mixture stains consist of varying amounts of blood, semen, saliva, vaginal secretion, menstrual blood and skin. Reference DNA-cSNP profiles were generated from buccal swabs or from other body fluids. Sequencing was carried out on an Illumina MiSeq FGx platform and the results were presented as read counts per cSNP position.

2.4.1. One donor per body fluid

For the 29 two-person mixtures with one donor per body fluid (Supplementary Table S2), each body fluid component was analysed separately as a single source profile in EuroForMix. Let donor 1 be the contributor of component 1, and donor 2 the contributor of component 2. We assume that it was accepted at sub-source level that donor 1, donor 2 and a third, unknown donor contributed DNA. For each body fluid component we evaluated the data given two sets of source level propositions. For component 1, where donor 1 is the true donor, the source level propositions were:

- H_p : Donor 1 contributed the body fluid,
 - H_d : Donor 2 contributed the body fluid,
- and
- H_p : Donor 1 contributed the body fluid,
 - H_d : The unknown third donor contributed the body fluid.

We refer to the likelihood ratios as LR_{12} and LR_{1u} , respectively. For component 2, where donor 2 is the true donor, the source level propositions were

- H_p : Donor 2 contributed the body fluid,
 - H_d : Donor 1 contributed the body fluid,
- and
- H_p : Donor 2 contributed the body fluid,
 - H_d : The unknown third donor contributed the body fluid,
- resulting in the likelihood ratios LR_{21} and LR_{2u} , respectively.

2.4.2. Two donors per body fluid

The 15 mixtures with two donors per body fluid (Supplementary Table S3) were evaluated as two-person mixtures in EuroForMix. Let donor 1 and donor 2 denote the two contributors. We assume that it was accepted at sub-source level that donor 1, donor 2 and a third, unknown donor contributed DNA. The following sets of source level propositions were considered:

- H_p : Donor 1 and donor 2 contributed the body fluid,
 - H_d : Donor 1 and the unknown third donor contributed the body fluid,
- and

- H_p : Donor 1 and donor 2 contributed the body fluid,
- H_d : Donor 2 and the unknown third donor contributed the body fluid.

We denote the corresponding likelihood ratios LR_{21} and $LR_{1|2}$, where $LR_{i|j}$ is the LR for donor i conditioned on the presence of donor j .

3. Results

3.1. Simulations

Based on real data we estimated the following expected read counts (μ) per body fluid: blood = 1200, saliva = 830, semen = 1245, vaginal secretion = 460, menstrual blood = 2650 and skin = 1080. The estimated coefficients of variation (σ) were: blood = 1.75, saliva = 1.89, semen = 1.32, vaginal secretion = 2.47, menstrual blood = 1.09 and skin = 1.15. These values were used as input in the simulations. Many of the generated profiles had a considerable amount of dropout; the vaginal secretion cSNPs showed a large number of dropout alleles even in the major contributor in the two-person mixtures (see Supplementary Tables S4 and S5).

3.1.1. One donor per body fluid

Fig. 4 shows the ROC curve for each body fluid based on the simulations with one donor per body fluid. At threshold $t = 1$, blood (11 markers), semen (8 markers) and skin (7 markers) have a true positive rate close to 1 and a false positive rate below 0.05. Saliva and vaginal secretion (3 markers each) have a true positive rate of 0.86 and 0.75, respectively, and a false positive rate of around 0.35 at this threshold. At $t = 10$, the true positive rate for blood, semen and skin is still around 0.9, while for saliva and vaginal secretion it has dropped considerably, as has the false positive rate. At $t = 100$, blood, semen and skin still have a true positive rate in the range 0.59–0.68, while at $t = 1000$ the rates have dropped to 0.14–0.18.

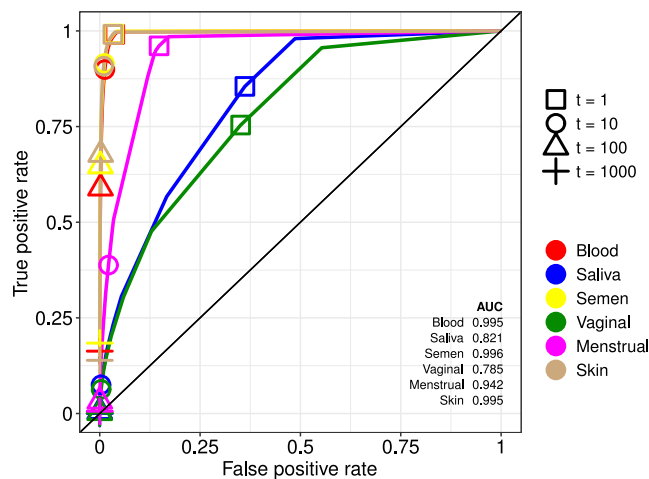


Fig. 4. ROC curves and AUC values for simulations of one donor per body fluid. For a range of thresholds t , the true positive rate refers to the proportion of $LR \geq t$ given H_p true, and the false positive rate refers to the proportion of $LR \geq t$ given H_d true. Thresholds $t = 1$, $t = 10$, $t = 100$ and $t = 1000$ are indicated in the curves, and exact values are given in Supplementary Table S6. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

3.1.2. Two donors per body fluid

Fig. 5 shows ROC curves based on simulations with two donors per body fluid, in different mixture proportions. As expected, the ability to discriminate between H_p and H_d for these two-person mixtures is lower than for the single source stains in the previous section. For the 1:1 mixtures, the three best performing body fluids, blood, semen and skin, all have true positive rates above 0.9 at $t = 1$, however the false positive rates are also high (around 0.2). LRs above 100 are rather unlikely for all body fluids. With mixture ratio 3:1, where we conditioned on the major contributor, the three best body fluids obtain true positive rates of around 0.8 at $t = 1$, but the false positive rates are high. Both saliva and vaginal secretion have AUC values just above 0.5. With mixture ratio 1:3, where we conditioned on the minor contributor, blood, semen and skin obtain true positive rates around 0.95 and false positive rates around 0.12. Saliva and vaginal secretion have true positive rates of 0.56 and 0.54 and false positive rates of 0.37 and 0.39. With $t = 10$ the false positive rates are low for all body fluids, and blood, semen and skin all have true positive rates around 0.65.

3.2. Real data

3.2.1. One donor per body fluid

Table 4 shows the results of the 29 two-person mixtures with one donor per body fluid analysed with EuroForMix. In the table, a dropout is defined as an allele in the reference profile with less than 50 reads in the RNA-cSNP profile. Dropout in homozygous markers counts as two alleles. Since the number of markers differs between the body fluids, a per sample dropout value was calculated as $x/(2L)$ where x is the number of dropped out alleles and L is the number of loci. For the mixtures that contained gender specific body fluids and a male and a female donor, the likelihood ratio functioned as a confirmatory test. Note that some LRs are 'NA', which means that EuroForMix did not return a result. Mainly, the reason was that all alleles for this component had dropped out (as indicated by 'DO = 1'). For semen in sample 14 and menstrual blood in sample 21 the model did not converge.

Some false negative results ($LR < 1$) were observed. For many samples this can be explained by a large proportion of dropout alleles (samples 1, 2, 4, 7 and 23). For the body fluids with only three markers (saliva, vaginal secretion and menstrual blood) even one dropout is a considerable loss of information (samples 3 and 19). The very low

Table 4

Results of EuroForMix analysis of the 29 samples with one donor per body fluid (Supplementary Table S2). 'Component 1' and 'Component 2' correspond to the first and second body fluid component of the mixture (red = blood, blue = saliva, yellow = semen, green = vaginal secretion, magenta = menstrual blood, brown = skin). Donor 1 is the true donor of component 1 and donor 2 is the true donor of component 2. 'DO' is the proportion of dropout alleles for the true donor. ' LR_{12} ' compares donor 1 (true donor) to donor 2, and ' LR_{1u} ' compares donor 1 to an unknown. ' LR_{21} ' compares donor 2 (true donor) to donor 1, and ' LR_{2u} ' compares donor 2 to an unknown. For certain components, one donor could be excluded because of gender (denoted by '-'). 'NA' denotes that EuroForMix did not give a result. For the mixtures of blood and menstrual blood (marked with *) we expect the menstrual blood donor to also appear in the blood markers together with the blood donor. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Sample	Component 1			Component 2		
	DO	LR_{12}	LR_{1u}	DO	LR_{21}	LR_{2u}
1	0.59	2.45	1.38e-05	0	1	2.35
2	0	5.14e+18	3,505	0.33	3.21	0.04
3	0	1.39e+27	8,347	0.17	7.39e+18	0.21
4	0.55	1.35e+17	3.21	0.31	4.83e+33	0.15
5	0	5.31e+29	2,409	0	-	403
6	1	NA	NA	0.81	39.3	4.23
7	0.59	2.74	0.46	1	NA	NA
8	0.86	5.15	1.87	1	NA	NA
9	0.27	0.11	3.43e-09	1	NA	NA
10*	0.36	130,538	2.54	1	NA	NA
11*	0	2.76e-05	9.18e-26	0	1	4.53
12*	0.59	0.41	2.34	0	1.96e+14	10.3
13	0.67	2.37	1.33	0.06	7.88e+50	90.1
14	0.67	1	1.33	0.06	-	NA
15	0.67	2.38	1.34	0.12	6.11e+18	4.51
16	1	NA	NA	0.67	1	1.09
17	1	NA	NA	0	26.4	5.37
18	0.5	0.28	1.04	0.17	0.50	0.42
19	0.33	1	1.71	0.17	20.8	0.18
20	0.33	1	1.83	0	-	13.2
21	0.67	1	2.84	0	NA	NA
22	0	1	2.33	0.71	5.27	2.67
23	0.94	-	0.66	0.67	-	1.09
24	0.12	-	9.48	0.67	-	1.09
25	0.12	-	0.29	0.67	-	1.09
26	0.44	-	3.08	1	-	NA
27	0.62	-	2.18	1	-	NA
28	0.25	-	6.36	0.33	-	2.14
29	0.19	-	0.53	0	-	4.56

value for LR_{1u} for the blood component in sample 1, is caused by one allele in the locus AMICA1_2 with 1800 reads that is not found in the donor's genotype. If a DNA contamination has occurred, the contaminating profile must by chance have the exact same genotype as the donor in all the other blood markers except AMICA1_2, which seems very unlikely. This is essentially a drop-in we have no explanation for. The blood component in sample 9 fits better as a mixture of the blood donor and the vaginal secretion donor, which could be explained if the vaginal secretion sample was contaminated with blood. Samples 11 and 12 are mixtures of blood and menstrual blood; both have $LR_{12} < 1$, indicating that the blood markers fit better to donor 2 (the menstrual blood donor) than to donor 1 (the blood donor). It is expected that the menstrual blood donor appears in the blood markers together with the blood donor. For sample 12, the high likelihood ratio for the menstrual blood component (LR_{21}) indicates that menstrual blood is consistent with donor 2. In samples 25 and 29 the low LR is likely caused by some unbalanced alleles. In addition, these samples have several dropout alleles in the markers TGM4_2, TGM4_3 and TGM4_4.

Moreover, Table 4 shows that in the three mixtures with blood and vaginal secretion, all the vaginal secretion alleles dropped out. Also in mixtures with semen, vaginal secretion had a large number of dropout alleles. Further, there are two saliva-vaginal secretion mixtures where all saliva alleles dropped out. In two of the menstrual blood components all menstrual blood alleles dropped out, one in a mixture with blood and one in a mixture with semen. In one blood-semen mixture all blood alleles dropped out, however, semen also had a large proportion of dropout alleles in this sample.

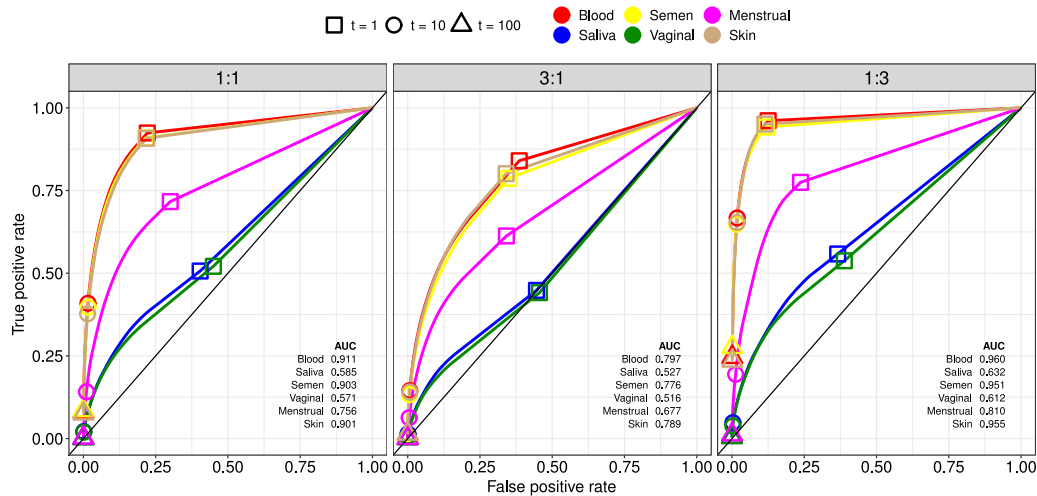


Fig. 5. ROC curves and AUC values based on the simulations of two donors per body fluid, with mixture proportions 1:1, 3:1 (LR conditioned on major contributor) and 1:3 (LR conditioned on minor contributor). Plotted for a range of thresholds t , the true positive rate is the proportion of $LR \geq t$ given H_p true, and the false positive rate is the proportion of $LR \geq t$ given H_d true. Thresholds $t = 1$, $t = 10$ and $t = 100$ are indicated in the curves and exact values are presented in Supplementary Table S7. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 5

Results of EuroForMix analysis of the 15 samples with two donors per body fluid (Supplementary Table S3). ' M_x ' is the designed mixture proportion for donor1:donor2, while 'Est M_x ' is the estimated mixture proportion. ' $LR_{2|1}$ ' compares 'donor 1 + donor 2' to 'donor 1 + unknown', while ' $LR_{1|2}$ ' compares 'donor 2 + donor 1' to 'donor 2 + unknown'. 'DO' denotes the proportion of dropout alleles in donor1:donor2. The samples are coloured according to body fluid (red = blood, blue = saliva, yellow = semen). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Sample	M_x	Est M_x	DO	$LR_{2 1}$	$LR_{1 2}$
30	0.5:0.5	0.34:0.66	0.09:0	104	71
31	0.5:0.5	0.54:0.46	0:0	138	614
32	0.5:0.5	0.48:0.52	0.18:0.14	93	38
33	0.1:0.9	0.32:0.68	0.18:0.14	1.25	1.98
34	0.1:0.9	0.17:0.83	0:0	2311	247
35	0.5:0.5	1:0	0.67:0.83	1.00	1.23
36	0.5:0.5	0.95:0.05	0.33:0.33	1.00	3.45
37	0.5:0.5	0.34:0.66	0.17:0	0.93	0.89
38	0.1:0.9	0.26:0.74	0.17:0	1.22	1.13
39	0.1:0.9	NA:NA	0.67:0.67	NA	NA
40	0.5:0.5	0.52:0.48	0.12:0.12	9.77e-05	4.38e-03
41	0.5:0.5	0.4:0.6	0.19:0.12	3.08e-03	2.36
42	0.5:0.5	NA:NA	0.06:0	NA	NA
43	0.1:0.9	0.08:0.92	0.19:0	312	54
44	0.1:0.9	0:1	0.62:0.38	7.04	1.00

3.2.2. Two donors per body fluid

Table 5 shows the results of the 15 mixtures with two donors per body fluid analysed with EuroForMix. Only blood, semen and saliva are represented among these samples. Sample 39 (saliva) only had one read count above the threshold, and EuroForMix returned no result. For sample 42 (semen), the model did not converge. Overall, the saliva samples had the most dropout alleles. This body fluid also had the most imprecise mixture proportions and lowest LR_s, which is expected since this is the body fluid with the fewest markers. A few false negatives were observed. In sample 37 (saliva) there was one dropout among only three markers. For sample 40 (semen), an allele present in both donors had dropped out in the markers TGM4_2 and TGM4_3. For sample 41 (semen) where $LR_{2|1} < 1$, the allele balance for some of the markers did not fit very well with donor 2's profile.

3.3. Mock casework examples

3.3.1. Sexual assault

This mock example is based on the semen-saliva mixture in sample 13 (Supplementary Table S2).

Table 6

Data from the mock casework semen-saliva mixture showing the alleles and read counts in the stain (RNA-cSNP), and the reference profile of Mr S (DNA-cSNP). The RNA-cSNP profile only had reads in the saliva and semen markers. Alleles in parentheses fall below the threshold of 50 read counts and are not considered in the LR calculation.

Marker	Body fluid	RNA-cSNP		DNA-cSNP
		Allele	Read count	Mr S
MUC7_1	Saliva	(C)	(40)	C/T
MUC7_2	Saliva	C	1697	C/G
PRB3	Saliva	(G)	(2)	G/G
KLK3	Semen	A/G	7462/53	A/A
SEMG1	Semen	A/T	3020/3477	A/T
SEMG2_1	Semen	A/C	1843/529	A/C
SEMG2_2	Semen	A/G	1923/874	A/G
TGM4_1	Semen	G	1416	G/G
TGM4_2	Semen	G	3269	A/G
TGM4_3	Semen	G	1719	G/G
TGM4_4	Semen	G	4108	G/G

A man is accused of sexual assault, but he denies the offence and says that he only spoke to the 'victim' Ms V. The activity level propositions are:

H_p : The suspect sexually assaulted Ms V,

H_d : An unknown sexually assaulted Ms V, the suspect only talked to her.

The evidence is a stain from the victim's T-shirt. A presumptive test for semen was positive and Y-STR analysis revealed a mixture of two males. The sub-source level propositions:

H_p : The DNA came from Mr S and an unknown contributor,

H_d : The DNA came from two unknown contributors, resulted in a high LR. Assuming that the court has accepted the H_p sub-source proposition, and since semen is the body fluid of interest, the following source level propositions are formulated:

H_p : Mr S contributed semen,

H_d : The second (unknown) donor contributed semen; Mr S contributed saliva or another body fluid/cell type.

A cSNP analysis was performed, and read counts for all the 35 cSNPs in the RNA-cSNP profile can be seen in Fig. 6. The reads are predominantly in the semen markers, but also one saliva marker has a considerable amount of reads. All other body fluids can be disregarded when we use a minimum threshold of 50 read counts.

Table 6 displays the read counts in the RNA-cSNP profile and the suspect's DNA-cSNP profile for the saliva and semen markers. If

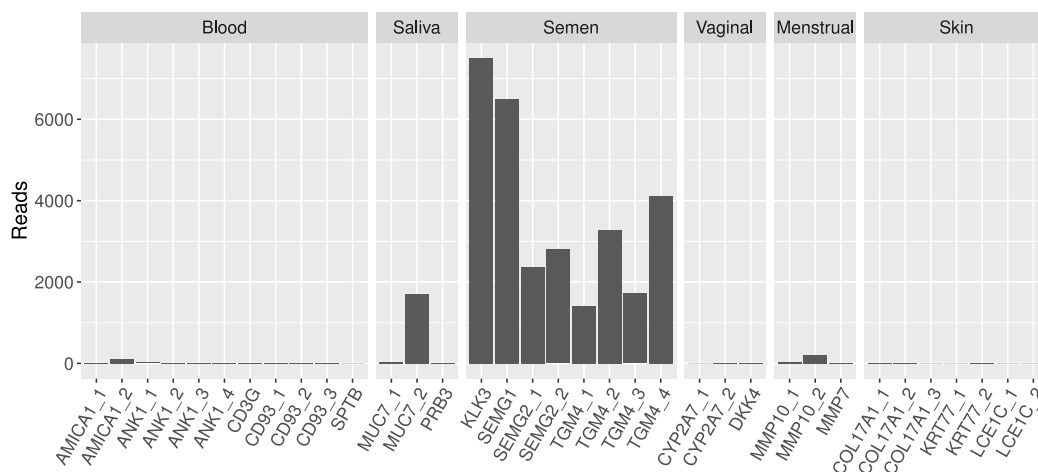


Fig. 6. Total read counts per marker in the RNA-cSNP profile of the mock casework semen-saliva mixture.

the suspect is the semen donor, there must have been a dropout in marker TGM4_2 (allele A), and a drop-in in KLK3 (allele G with 53 reads). Evaluation of the cSNP profile for the eight semen markers in EuroForMix resulted in an LR of 90. The evidence supports the source level proposition that the semen came from Mr S rather than the unknown donor. To address the activity level propositions, a different calculation would be required that is not considered here.

3.3.2. Physical assault

This mock casework example is based on the blood-blood mixture in sample 32 (Supplementary Table S3).

Mr Y is accused of violence against Mr X. He did not see his attacker, but Mr Y is a suspect because they had a heated argument earlier that day. The following activity level propositions are formulated:

H_p : Mr Y assaulted Mr X,

H_d : Mr X was assaulted by someone other than Mr Y; Mr Y and Mr X only had social interactions earlier that day.

STR analysis of a blood stain from Mr X's clothes revealed a three-person mixture. The sub-source propositions:

H_p : The DNA came from Mr X, Mr Y and an unknown contributor,

H_d : The DNA came from Mr X and two unknown contributors, resulted in an LR larger than 1 billion. The presence of Mr Y's DNA is thus not disputed, however there is an expectation under H_d that the DNA contribution from Mr Y came from saliva due to talking or another cell type via social interaction. The following propositions are considered at source level:

H_p : Mr X and Mr Y contributed blood,

H_d : Mr X and an unknown contributed blood; Mr Y contributed saliva or another body fluid/cell type.

Fig. 7 shows read counts in the RNA-cSNP profile from the stain for all the 35 cSNPs. The read counts are mainly in the blood markers, however semen and saliva also have spurious reads. The data for the 11 blood markers are given in Table 7. Using a minimum read count threshold of 50, both donors have a dropout in ANK1_4 (allele G) and the marker CD93_3 has no read counts. Evaluation of the cSNP blood markers with EuroForMix resulted in an LR of 93. The estimated mixture ratio was 0.48:0.52 for Mr X:Mr Y. The evidence supports the proposition that both Mr X and Mr Y donated blood, rather than Mr X and an unknown being the blood donors. The activity level assessment is a separate analysis, not attempted here.

4. Discussion

We have investigated the interpretation of coding region SNPs for the association of body fluids and donors in mixed biological stains in

Table 7

Data from the mock casework blood-blood mixture showing the alleles and read counts in the stain (RNA-cSNP), and the reference profiles of Mr X and Mr Y (DNA-cSNP). The marker CD93_3 had no reads. Only the blood markers are shown since this is the contested body fluid. Alleles in parentheses fall below the threshold of 50 read counts and are not considered in the LR calculation.

Marker	Body fluid	RNA-cSNP		DNA-cSNP	
		Allele	Read count	Mr X	Mr Y
AMICA1_1	Blood	G/A	5764/2221	G/G	A/G
AMICA1_2	Blood	T/C	5591/2382	C/T	T/T
ANK1_1	Blood	G/(A)	7979/(3)	G/G	G/G
ANK1_2	Blood	G/A	3332/693	G/G	A/G
ANK1_3	Blood	T/C	1470/624	C/T	C/C
ANK1_4	Blood	A/(G)	955/(6)	G/G	A/G
CD3G	Blood	A/G	2722/1153	A/A	A/G
CD93_1	Blood	G/(A)	2753/(4)	G/G	G/G
CD93_2	Blood	C/T	2284/683	C/T	C/C
CD93_3	Blood	-	-	A/G	A/A
SPTB	Blood	T/C	676/585	C/T	C/C

a likelihood ratio framework. RNA-cSNP data provides information on the origin of the biological fluid through the presence of body fluid specific markers, but also on the identity of the donor through comparison with a DNA-cSNP reference profile. cSNPs therefore provide a direct link between a donor and a body fluid.

4.1. Computational considerations

Since the interpretation of cSNP profiles is similar to that of STR profiles, we used the STR mixture software EuroForMix to calculate likelihood ratios for the cSNP mixtures. With EuroForMix we are constrained to consider source level propositions and cSNP data for only one body fluid at a time. If we would consider different body fluids under H_p and H_d , different cSNP markers would have to be analysed under each proposition, which would not be a valid comparison. On the other hand, if the whole RNA-cSNP profile was to be used as evidence in EuroForMix, all the markers that are not expressed would be considered as dropout. This would again reduce the expected signal, which would not be an appropriate assumption. For source level propositions where there are no alternative donors of the body fluid in question, we showed that we could assign a donor to a body fluid with a confirmatory test.

When simulating cSNP read counts we made two simplifying assumptions. First, we assumed that all cSNPs within a body fluid have the same expected read count, meaning that they are equally 'good' markers. From our experience with real data we note that this is not the case [21]. Nevertheless, without making this assumption we would have to estimate the average read count and variance for each marker

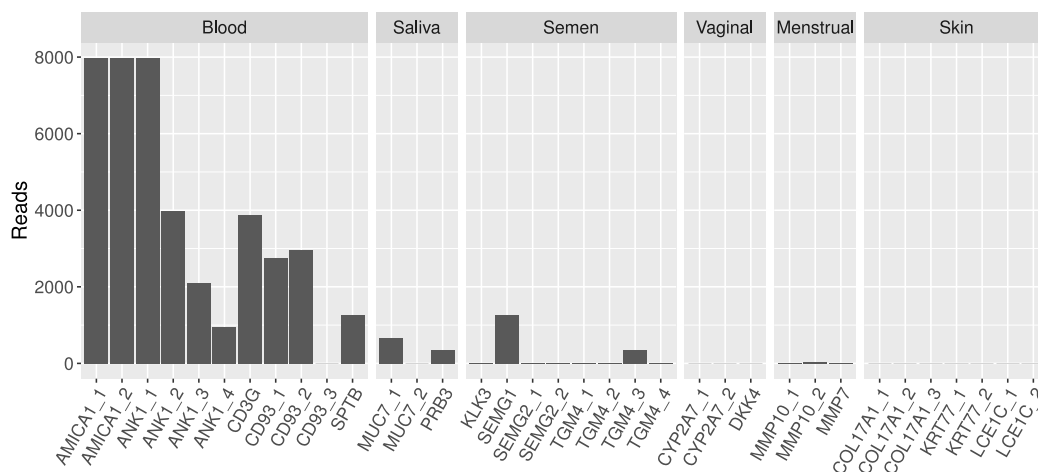


Fig. 7. Total read counts per marker in the RNA-cSNP profile of the mock casework blood-blood mixture.

separately, which we do not have sufficient data for. In addition, large variation between markers within the same body fluid was reflected in the estimated variance within a body fluid. The other assumption was that the cSNPs are completely specific for one body fluid. This implied that we only simulated read counts for the cSNPs associated with the body fluid in question. This does not fully reflect reality as we expect some non-specific reads in the cSNP profiles. However, since we analyse only the cSNPs for the body fluid involved in the propositions, read counts in cSNPs associated with other body fluids would not influence the likelihood ratio.

4.2. Discrimination power and specificity

The simulations reflected that the number of markers, the expected read counts and the variation between and within markers for the same body fluid all influence the discrimination power of the different body fluids (Figs. 4 and 5). Blood, semen and skin, the body fluids with the most markers, performed best; they were most likely to provide an LR that supported the true hypothesis. Saliva and vaginal secretion have only three markers each, and performed rather poorly for the mixtures with two donors per body fluid. Vaginal secretion had the lowest estimated read count and the highest variation, and also the lowest discrimination power. Although menstrual blood only has three markers, it had a higher estimated read count and lower variance, and performed considerably better than saliva and vaginal secretion.

Similar results were observed with the real data. For the mixtures with one donor per body fluid (Table 4), blood and semen were the body fluids with the fewest inconclusive results (NA or LR=1), nevertheless some false negatives (LR<1) could be observed. Vaginal secretion, saliva and menstrual blood all had a large number of inconclusive results, however, menstrual blood had no false negatives. For the mixtures with two donors per body fluid (Table 5), the best results were obtained with blood. All the blood samples gave support to the true hypothesis, while the saliva and semen samples showed varying results. For a few of the semen stains in Tables 4 and 5, EuroForMix was not able to fit a model to the data, although the read counts were high. A plausible explanation is that the very small variation in read counts between the markers resulted in a very small estimate of the coefficient of variation of the read counts (σ). In this case the model becomes very strict in accepting which genotype combinations are allowed. Semen, vaginal secretion and menstrual blood donors can in some cases be excluded because of gender.

The comparison of EuroForMix on real and simulated data indicates that the simulations overestimate the discrimination power; the real data LRs were lower than we would expect from the ROC curves.

For instance, the ROC curve for one-person blood samples (Fig. 4) estimates the true positive rate (TPR) at $t = 1$ to be 0.99. For the blood components in the real samples with one donor per body fluid (Table 4), only 7 out of 12 LRs that compare the true donor to an unknown (LR_{1u}) are above 1. This corresponds to an observed TPR of 0.58. Similarly, the estimated TPR for semen at $t = 1$ is 0.99, while the observed TPR in the real semen samples is only $8/13 = 0.62$. Menstrual blood has an estimated TPR of 0.96 and an observed TPR of $5/9 = 0.56$. Some inconsistencies between the observed and estimated discrimination power could also be seen in the samples with two donors per body fluid. For 1:1 semen mixtures the estimated TPR at $t = 1$ is 0.91 (Fig. 5), while only 4 out of the 10 LRs that compare both donors to one donor and an unknown ($LR_{1|2}$ and $LR_{2|1}$) are above 1. This corresponds to an observed TPR of $4/10 = 0.4$. For the blood samples the opposite could be observed. The estimated TPR at $t = 10$ for 1:1 blood mixtures is 0.41, while 8 out of the 10 LRs were above 10, resulting in an observed TPR of 0.8. We note that the amount of dropout is significantly larger in the real samples compared to the simulated samples (Supplementary Tables S4 and S5), with the exception of the two-person blood samples. One explanation for these inconsistencies would be that the simulated data does not mimic the observed data very well, since they are based on the EuroForMix model.

Due to the natural composition of menstrual blood, we expect reads also in vaginal secretion and blood cSNPs for stains where menstrual blood is present. This could be observed in some of the real mixtures of blood and menstrual blood in Section 3.2.1 where the menstrual blood donor also showed expression of blood. Therefore, in a stain where one donor has contributed blood and the other has contributed menstrual blood, it may be more appropriate to evaluate the evidence in the blood cSNPs as a two-person mixture, assuming that both the blood donor and the menstrual blood donor is present. The same applies for the evidence in the vaginal secretion cSNPs in mixtures of vaginal secretion and menstrual blood.

Although the cSNPs are body fluid specific, they are not as specific as RNA markers designed merely for the purpose of discriminating between body fluids [21]. Some of the discriminatory body fluid specific mRNA markers do not contain suitable cSNPs, or the cSNPs are located deep inside a large exon, and therefore no RNA-specific primers (spanning an intron or located on an exon-exon-boundary) can be designed. However, body fluid specific and cSNP mRNA markers can be incorporated into one targeted MPS assay. Nevertheless, RNA analysis or presumptive tests for body fluid identification is an important step before cSNP analysis, as the results can assist in the formulation of source level propositions.

4.3. cSNP artefacts

The analysis of two-component body fluid mixtures in Table 4 suggested that there may be an interaction and/or interference between different body fluids in a mixture, which may cause one body fluid to be dominant. We observed for instance that when vaginal secretion was mixed with blood, all vaginal secretion alleles had dropped out. In mixtures with saliva, the vaginal secretion markers were more visible. A similar phenomenon was observed in Dørum et al. [19] where a prediction model was used to identify body fluid components in mRNA mixtures. Several of the mRNA markers used there overlap with our cSNP markers (Supplementary Table S8), although different primer sets are used to incorporate the cSNPs. The reasons for this phenomenon are unknown but may be due to a combination of biological and technical factors. Biochemical interactions could involve differences in the composition and stability of the cellular and extra-cellular RNA present in the constituent body fluid samples comprising the mixture. It may be that one of the body fluid samples has a higher concentration of extracellular RNA transcripts that are more susceptible to hydrolytic degradation compared to the other sample and/or that one of the body fluids has a higher concentration of extracellular nucleases. Variation in the expression of tissue specific transcripts between different body fluids as well as inter- and intra-individual differences in gene expression of the same body fluid may also contribute to finding an apparent interaction between body fluids. Sample preparation effects may also confound the expected results with mock mixtures. For example, even if mixtures are designed to be 1:1, e.g. 50 µl of blood and 50 µl of saliva, they may contain very different amounts of RNA. Or if the major component is present in high excess, titration of the critical library preparation and PCR reagents may result in the second (minor) component not being detected. Note that this possible interaction between body fluids was not taken into account in the simulations, however it could have been incorporated by e.g. assuming different expected read counts for vaginal secretion when mixed with semen and when mixed with saliva. The estimation of such mixture specific read counts would require a larger data set.

Another phenomenon observed in the analysis of real data was drop-in and dropout alleles at the RNA level. A reason for dropout in the RNA results could be the monoallelic or preferential expression of one allele [33,34]. In particular, this was observed in the cSNPs TGM4_2, TGM4_3, and TGM4_4 (also previously reported [21]). We suppose that this is a biological rather than a technical issue [35]. Monoallelic expression differs from regular dropout in that it is not related to low read counts. It would make the (true) heterozygous genotype less likely, thus reducing the LR. This phenomenon is not modelled in EuroForMix. Further, allelic drop-in is possibly due to stochastic sample amplification during PCR. In sample 1 (Table 4) there was a blood marker with a large unexplained drop-in. More investigations are needed, but it is possible that the drop-in distribution in EuroForMix needs adjustments to accommodate cSNP specific artefacts.

We have assumed throughout this paper that all the cSNPs carry independent information, while in fact several of the 35 cSNPs are situated on the same genes. Linkage may have an effect on match probability calculations for close relatives [36]; but this is only relevant if kinship is in the propositions. A further result of closely linked loci may be linkage disequilibrium (LD), which can be observed as non-random association between alleles at different loci on a population level. In our previous paper [21] we found LD between several of the cSNPs. Laurie and Weir [37] showed that by ignoring LD one underestimates the match probability. A conservative subpopulation correction may however prevent overestimation of the likelihood ratio [36]. Alternatively, one could consider haplotype frequencies rather than allele frequencies when estimating match probabilities. Further, one could envisage a continuous likelihood ratio model that handles linked markers by modelling the peak heights/read counts of haplotypes rather than alleles, which is not possible in EuroForMix. However,

this is beyond the scope of this paper where the focus is on establishing a general framework for evaluation of cSNP profiles. There is currently ongoing work to improve the set of cSNP markers, and the final set of cSNPs may not have the same issues with linkage and LD as the preliminary set of markers presented here.

4.4. Conditioning on sub-source results

Thus far we have only considered LR calculations for cSNP profiles where the number of contributors is the same under both propositions. However, if we regard the propositions as logical consequences of scenarios put forward by prosecution and defence, it may be that the scenarios happen to have a different number of contributors. This point of view was also supported by the recommendations on the interpretation of mixtures by the ISFG DNA commission [38]. Consider the case example presented in Section 2.2.6 where we evaluated a semen RNA-cSNP profile. Assume that the STR profile only showed a two-person mixture, and that it was accepted at sub-source level by both the prosecution and defence that Mr B and Mr S were the only contributors to the stain. Following this, the natural source level propositions become

H_p : Mr B and Mr S contributed semen,

H_d : Mr B only contributed semen; Mr S contributed skin cells or a different body fluid.

Analysis of the semen cSNPs with EuroForMix results in an LR of 363 million. The large likelihood ratio reflects the fact that the RNA-cSNP profile does not resemble a single contributor profile as is conditioned on under H_d , and the model describes a two-person profile better. However, both parties had already accepted at sub-source level that Mr B and Mr S were the only contributors to the stain, so under H_d there was no conceivable second semen contributor. The likelihood under H_d approaches 0 and the LR becomes a confirmatory test. If we would have included an unknown semen donor under H_d , we would implicitly assume that it is possible to detect a cSNP profile for a contributor that was not detected in the STR profile. This contradicts with our assumption that the STRs are more sensitive than the cSNPs (as stated in Section 2.2.1).

On a different note, if analysis of the semen cSNPs with EuroForMix resulted in a low LR, thus supporting the proposition that Mr S did not contribute semen, then the LR for semen is not, on its own, the whole source-LR. If a sufficient amount of DNA supports his STR profile but none of the five other body fluids were detected, the most plausible scenario is that Mr S has left some other body fluid/cell type not covered by the cSNPs. Although the cSNPs do not contribute a direct link between a body fluid and the suspect in this case, they still carry some evidential value.

4.5. Hierarchy of propositions framework

The ISFG DNA commission has published recommendations regarding the evaluation of evidence [2,39]. The hierarchy of propositions framework is fundamental, and it is pertinent to carry out a brief review of the framework here. The hierarchy of propositions is divided into three main levels: sub-source, source and activity. The sub-source relates to the DNA profile itself; source level addresses the body fluid and the activity level describes the 'how' and 'when' the body fluid became evidential. The highest-level propositions deal with offence, however findings at this level are not addressed by the scientist. It is an important precept that each level in the hierarchy is a separate consideration that requires a different LR.

Addressing source level in casework adds significant value to the interpretation of a case. Several examples are provided in Sections 2.2 and 3.3. One of the described cases (Regina v. Weller [30], Section 2.2.4) was considered by the appeal court of England and Wales, where source level was the main consideration. There was alleged sexual assault by digital penetration; contact between the suspect and

victim was not denied. Consequently, there is an a priori expectation to find epithelial cells from the victim underneath the fingernails of the suspect. In court, the prosecution successfully argued that the evidence, based upon the likelihood of observing a full DNA profile from the victim and the scientist's experience with such cases, supported their contention that the DNA was from vaginal cells, rather than from epithelial cells or vomit. However, there was no quantitative assessment at source level, i.e. no evidence that vaginal cells were actually present, hence the assessment was subjectively based. In contrast, the application of cSNP profiling removes the need to rely upon the scientist's 'experience'. By providing a quantitative assessment of the evidence, it would greatly assist a court with deliberations in complex cases such as Weller, that are dependent upon source level assessment.

Taylor et al. [22] introduced a Bayesian network to answer questions like these using criteria such as visual appearance of a stain, quantification of DNA, results of presumptive tests such as Hema-Stix, and the value of the evidence at sub-source level. Because many factors are considered, the resulting Bayesian networks are complex. Conversely, the advantage of cSNP profiling is that it offers a much-simplified approach, that requires fewer assumptions, and is therefore both more robust and arguably easier to adopt in casework. If laboratories adopt the same procedures, then standardisation follows. This has been greatly assisted by several EDNAP/EUROFORGEN studies that show the successful application of mRNA profiling for body fluid identification [40–45] and cSNP profiling [35] in the participating laboratories.

4.6. Reporting

A consideration of a level in the hierarchy of propositions requires acceptance of the previous level by the court [2]. To explain, a caveat is needed in the statement to go from sub-source to source level:

“Provided that it is accepted that the DNA came from Mr S, I can consider the evidence of the body fluid attribution”.

A similar caveat will be needed if activity level is addressed. It is important to note that there is an essential distinction to make between source level and activity level reporting. Source level will address propositions relating to the cell type in a stain and can address its association with a particular DNA profile. Therefore, in the Weller example above, where samples are taken from underneath the suspect's fingernails, source level propositions can be formulated as follows in a statement:

“I have been asked to consider two alternative propositions:

H_p : Vaginal cells from Ms V were recovered from underneath fingernails of Mr S,

H_d : Some other cell type from Ms V was recovered from underneath fingernails of Mr S.

Provided that it is accepted by the court that Ms V was a contributor of DNA to the sample, I can carry out an evaluation of the evidence in relation to the body fluid source. To do this I carried out a confirmatory test using cSNP profiling. The confirmatory test provides either a positive or negative result (note that a negative result can occur if there is insufficient body fluid present to test — it does not definitively ‘exclude’ its presence). My conclusion is that the cSNP evidence supported the proposition that vaginal cells from Ms V were recovered from underneath fingernails of Mr S”.

In the extended version of this case presented in Section 2.2.5, there was also an unknown donor present under the fingernails of the suspect. In this case the likelihood ratio could be quantified, and the results could be reported somewhat differently:

“I have considered two alternative propositions:

H_p : Vaginal cells from Ms V were recovered from underneath fingernails of Mr S,

H_d : Vaginal cells from an unknown female were recovered from underneath fingernails of Mr S. Ms V contributed some other cell type.

The evidence is X times more likely if the first proposition is true rather than if the alternative were true”.

However, this does not directly address the ‘activity level’. The distinction is subtle, but it does require a separate assessment of the evidence. Hence, in this case example, the activity level propositions are the same as described in Section 2.2.4:

H_p : Mr S sexually assaulted Ms V by digital penetration and had social interaction,

H_p : Mr S did not assault Ms V, he only had social interaction and helped her when she was ill.

To assess the activity level propositions requires an understanding of the various modes that DNA transfer can occur and relevant data. If the prosecution proposition is true, then digital penetration occurred. However, if the defence proposition is true, it is necessary to consider the alternatives. These are:

- Secondary/tertiary transfer: The suspect's contact with the defendant's hands, clothing of the victim, or from the local environment e.g. her bed.
- Contamination: accidental transfer of materials from one item to another during collection and analysis (e.g. miscarriage of justice of Farah Jama [3, pp. 27–30].

Whereas secondary transfer has been addressed in the literature (see reviews [46,47]) with respect to DNA, there is a gap in literature with respect to mRNA i.e. there are no studies that ascertain levels of prevalent (known individuals) and background (unknown individuals) mRNA in the environment. Until such studies become available, reporting scientists will need to explain the limitations of the evidence to avoid inadvertent carry-over of the source level LR to the activity level. A caveat can be applied to the statement:

“My assessment is limited to a consideration of the value of the evidence relating to the body fluid/cell type taken from the item. I have not addressed the activity that led to its deposition. This is a separate consideration that requires a different analysis. To evaluate, I would need to take account of possible alternative methods of transfer of DNA and mRNA, such as contamination and levels of background RNA in the environment. Currently, there are insufficient data to assist me with this task”.

Once data become available it will be possible to report activity level using Bayesian networks such as those described by Gill et al. [39].

5. Conclusion

The aim of this paper was to establish a framework for evaluation of cSNP data given source level propositions in a likelihood ratio setting. More specifically, we wanted to explore the use of the continuous model in EuroForMix for analysis of cSNP mixtures. The results indicate that saliva, vaginal secretion and menstrual blood, with only three markers each, are not informative enough to properly assess the applicability of EuroForMix. Blood, semen and skin have more cSNPs and better discrimination power, but based on the inconsistencies between results from simulated and real data, more experiments should be conducted to assess if EuroForMix can be applied to cSNP data. Phenomena in cSNPs that do not occur in STRs, such as unspecific reads, monoallelic expression and the fact that cSNPs for some body fluids appear to be silenced by other body fluids, need to be better investigated and understood, and different models should be tested to find the optimal way of evaluating these data. Nevertheless, we have demonstrated that cSNPs can contribute valuable information in assigning a donor to a body fluid in mixed biological stains in a range of different scenarios.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The generation of the MPS data used in this study was funded by the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 285487 (EUROFORGEN-NoE). The authors thank Jack Ballantyne for reviewing the manuscript and contributing to the discussion and Thore Egeland for discussions on the topic. We would also like to thank two anonymous reviewers that helped improve the manuscript.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at <https://doi.org/10.1016/j.fsigen.2022.102685>.

References

- [1] R. Cook, I. Evett, G. Jackson, P. Jones, J. Lambert, A hierarchy of propositions: deciding which level to address in casework, *Sci. Justice* 38 (4) (1998) 231–239, [http://dx.doi.org/10.1016/S1355-0306\(98\)72117-3](http://dx.doi.org/10.1016/S1355-0306(98)72117-3).
- [2] P. Gill, T. Hicks, J.M. Butler, E. Connolly, L.G. ao, B. Kokshoorn, N. Morling, R.A. van Oorschot, W. Parson, M. Prinz, P.M. Schneider, T. Sijen, D. Taylor, DNA commission of the International society for forensic genetics: Assessing the value of forensic biological evidence - Guidelines highlighting the importance of propositions: Part I: evaluation of DNA profiling comparisons given (sub-) source propositions, *Forensic Sci. Int.: Genet.* 36 (2018) 189–202, <http://dx.doi.org/10.1016/j.fsigen.2018.07.003>.
- [3] P. Gill, *Misleading DNA Evidence: Reasons for Miscarriages of Justice*, Elsevier, 2014.
- [4] K. Virkler, I.K. Lednev, Analysis of body fluids for forensic purposes: from laboratory testing to non-destructive rapid confirmatory identification at a crime scene, *Forensic Sci. Int.* 188 (1) (2009) 1–17.
- [5] J. Juusola, J. Ballantyne, Multiplex mRNA profiling for the identification of body fluids, *Forensic Sci. Int.* 152 (1) (2005) 1–12, <http://dx.doi.org/10.1016/j.forsciint.2005.02.020>.
- [6] D. Zubakov, I. Kokmeijer, A. Ralf, N. Rajagopalan, L. Calandro, S. Wootton, R. Langit, C. Chang, R. Lagace, M. Kayser, Towards simultaneous individual and tissue identification: A proof-of-principle study on parallel sequencing of STRs, amelogenin, and mRNAs with the ion torrent PGM, *Forensic Sci. Int.: Genet.* 17 (2015) 122–128.
- [7] C. Haas, B. Klessner, C. Maake, W. Baer, A. Kratzer, mRNA profiling for body fluid identification by reverse transcription endpoint PCR and realtime PCR, *Forensic Sci. Int.: Genet.* 3 (2) (2009) 80–88, <http://dx.doi.org/10.1016/j.fsigen.2008.11.003>.
- [8] R.I. Fleming, S. Harbison, The development of a mRNA multiplex RT-PCR assay for the definitive identification of body fluids, *Forensic Sci. Int.: Genet.* 4 (4) (2010) 244–256.
- [9] F. Kohlmeier, P.M. Schneider, Successful mRNA profiling of 23 years old blood stains, *Forensic Sci. Int.: Genet.* 6 (2) (2012) 274–276, <http://dx.doi.org/10.1016/j.fsigen.2011.04.007>.
- [10] A. Lindenbergh, M. de Pagter, G. Ramdayal, M. Visser, D. Zubakov, M. Kayser, T. Sijen, A multiplex (m)RNA-profiling system for the forensic identification of body fluids and contact traces, *Forensic Sci. Int.: Genet.* 6 (5) (2012) 565–577, <http://dx.doi.org/10.1016/j.fsigen.2012.01.009>.
- [11] A. Lindenbergh, P. Maaskant, T. Sijen, Implementation of RNA profiling in forensic casework, *Forensic Sci. Int.: Genet.* 7 (1) (2013) 159–166.
- [12] E. Hanson, S. Ingold, C. Haas, J. Ballantyne, Messenger RNA biomarker signatures for forensic body fluid identification revealed by targeted RNA sequencing, *Forensic Sci. Int.: Genet.* 34 (2018) 206–221.
- [13] K. Sakurada, T. Akutsu, H. Fukushima, K. Watanabe, M. Yoshino, Detection of dermcidin for sweat identification by real-time RT-PCR and ELISA, *Forensic Sci. Int.* 194 (1–3) (2010) 80–84.
- [14] M. van den Berge, B. Bhoelai, J. Hartevelde, A. Matai, T. Sijen, Advancing forensic RNA typing: on non-target secretions, a nasal mucosa marker, a differential co-extraction protocol and the sensitivity of DNA and RNA profiling, *Forensic Sci. Int.: Genet.* 20 (2016) 119–129.
- [15] K. Sakurada, T. Akutsu, K. Watanabe, M. Yoshino, Identification of nasal blood by real-time RT-PCR, *Leg. Med.* 14 (4) (2012) 201–204.
- [16] E. Hanson, C. Haas, R. Jucker, J. Ballantyne, Specific and sensitive mRNA biomarkers for the identification of skin in 'touch' DNA evidence, *Forensic Sci. Int.: Genet.* 6 (5) (2012) 548–558.
- [17] A.D. Roeder, C. Haas, mRNA profiling using a minimum of five mRNA markers per body fluid and a novel scoring method for body fluid identification, *Int. J. Legal Med.* 127 (4) (2013) 707–721, <http://dx.doi.org/10.1007/s00414-012-0794-3>.
- [18] J. de Zoete, J. Curran, M. Sjerps, A probabilistic approach for the interpretation of RNA profiles as cell type evidence, *Forensic Sci. Int.: Genet.* 20 (2016) 30–44, <http://dx.doi.org/10.1016/j.fsigen.2015.09.007>.
- [19] G. Dørum, S. Ingold, E. Hanson, J. Ballantyne, L. Snipen, C. Haas, Predicting the origin of stains from next generation sequencing mRNA data, *Forensic Sci. Int.: Genet.* 34 (2018) 37–48, <http://dx.doi.org/10.1016/j.fsigen.2018.01.001>.
- [20] J. Hartevelde, A. Lindenbergh, T. Sijen, RNA cell typing and DNA profiling of mixed samples: can cell types and donors be associated? *Sci. Justice* 53 (3) (2013) 261–269.
- [21] S. Ingold, G. Dørum, E. Hanson, J. Ballantyne, C. Haas, Assigning forensic body fluids to donors in mixed body fluids by targeted RNA/DNA deep sequencing of coding region SNPs, *Int. J. Legal Med.* 134 (2) (2020) 473–485.
- [22] D. Taylor, D. Abarno, T. Hicks, C. Champod, Evaluating forensic biology results given source level propositions, *Forensic Sci. Int.: Genet.* 21 (2016) 54–67.
- [23] D. Taylor, Probabilistically determining the cellular source of DNA derived from differential extractions in sexual assault scenarios, *Forensic Sci. Int.: Genet.* 24 (2016) 124–135.
- [24] J. de Zoete, W. Oosterman, B. Kokshoorn, M. Sjerps, Cell type determination and association with the DNA donor, *Forensic Sci. Int.: Genet.* 25 (2016) 97–111.
- [25] Ø. Bleka, M. Eduardoff, C. Santos, C. Phillips, W. Parson, P. Gill, Open source software EuroForMix can be used to analyse complex SNP mixtures, *Forensic Sci. Int.: Genet.* 31 (2017) 105–110.
- [26] Ø. Bleka, G. Storvik, P. Gill, Euroformix: an open source software based on a continuous model to evaluate STR DNA profiles from a mixture of contributors with artefacts, *Forensic Sci. Int.: Genet.* 21 (2016) 35–44.
- [27] R. Fisher, Standard calculations for evaluating a blood-group system, *Heredity* 5 (1) (1951) 95.
- [28] G. Sensabaugh, in: S. R. (Ed.), *Biochemical Markers of Individuality*. In: *Handbook of Forensic Science*, Prentice Hall Inc, Upper Saddle River, 1982.
- [29] S. Willis, ENFSI Guideline for the Formulation of Evaluative Reports in Forensic Science. Monopoly Project MP2010: The Development and Implementation of an ENFSI Standard for Reporting Evaluative Forensic Evidence, Technical report, European Network of Forensic Science Institutes, 2015.
- [30] Weller, r. v [2010] EWCA Crim 1085 (04 March 2010), <http://www.bailii.org/cgi-bin/markup.cgi?doc=/ew/cases/EWCA/Crim/2010/1085.html>.
- [31] M.H. Zweig, G. Campbell, Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine, *Clin. Chem.* 39 (4) (1993) 561–577.
- [32] M. Kruijver, Efficient computations with the likelihood ratio distribution, *Forensic Sci. Int.: Genet.* 14 (2015) 116–124, <http://dx.doi.org/10.1016/j.fsigen.2014.09.018>.
- [33] A. Gimelbrant, J.N. Hutchinson, B.R. Thompson, A. Chess, Widespread monoallelic expression on human autosomes, *Science* 318 (5853) (2007) 1136–1140.
- [34] P.R. Buckland, Allele-specific gene expression differences in humans, *Hum. Mol. Genet.* 13 (suppl_2) (2004) R255–R260.
- [35] S. Ingold, G. Dørum, E. Hanson, D. Ballard, A. Berti, K.B. Gettings, F. Giangasparo, M.-L. Kampmann, F.-X. Laurent, N. Morling, et al., Body fluid identification and assignment to donors using a targeted mRNA massively parallel sequencing approach – results of a second EUROFORGEN/EDNAP collaborative exercise, *Forensic Sci. Int.: Genet.* 45 (2020) 102208.
- [36] P. Gill, C. Phillips, C. McGovern, J.-A. Bright, J. Buckleton, An evaluation of potential allelic association between the STRs vWA and D12S391: implications in criminal casework and applications to short pedigrees, *Forensic Sci. Int.: Genet.* 6 (4) (2012) 477–486.
- [37] C. Laurie, B. Weir, Dependency effects in multi-locus match probabilities, *Theor. Popul. Biol.* 63 (3) (2003) 207–219, Uses of DNA and genetic markers for forensics and population studies, [https://doi.org/10.1016/S0040-5809\(03\)00002-9](https://doi.org/10.1016/S0040-5809(03)00002-9).
- [38] P. Gill, C. Brenner, J. Buckleton, A. Carracedo, M. Krawczak, W. Mayr, N. Morling, M. Prinz, P. Schneider, B. Weir, DNA commission of the international society of forensic genetics: Recommendations on the interpretation of mixtures, *Forensic Sci. Int.* 160 (2006) 90–101, <http://dx.doi.org/10.1016/j.forsciint.2006.04.009>.
- [39] P. Gill, T. Hicks, J.M. Butler, E. Connolly, L. Gusmão, B. Kokshoorn, N. Morling, R.A. van Oorschot, W. Parson, M. Prinz, et al., DNA commission of the international society for forensic genetics: Assessing the value of forensic biological evidence – Guidelines highlighting the importance of propositions. Part II: Evaluation of biological traces considering activity level propositions, *Forensic Sci. Int.: Genet.* 44 (2020) 102186.

- [40] C. Haas, E. Hanson, M.J. Anjos, W. Baer, R. Banemann, A. Berti, E. Borges, C. Bouakaze, A. Carracedo, M. Carvalho, V. Castella, A. Choma, G. De Cock, M. Doetsch, P. Hoff-Olsen, P. Johansen, F. Kohlmeier, P.A. Lindenbergh, B. Ludes, O. Maronas, D. Moore, M.-L. Morerod, N. Morling, H. Niederstaetter, F. Noel, W. Parson, G. Patel, C. Popielarz, E. Salata, P.M. Schneider, T. Sijen, B. Sviezena, M. Turanska, L. Zatkalikova, J. Ballantyne, RNA/DNA co-analysis from blood stains - results of a second collaborative EDNAP exercise, *Forensic Sci. Int.: Genet.* 6 (1) (2012) 70–80, <http://dx.doi.org/10.1016/j.fsigen.2011.02.004>.
- [41] C. Haas, E. Hanson, M.J. Anjos, R. Banemann, A. Berti, E. Borges, A. Carracedo, M. Carvalho, C. Courts, G. De Cock, M. Doetsch, S. Flynn, I. Gomes, C. Hollard, B. Hjort, P. Hoff-Olsen, K. Hribikova, A. Lindenbergh, B. Ludes, O. Maronas, N. McCallum, D. Moore, N. Morling, H. Niederstaetter, F. Noel, W. Parson, C. Popielarz, C. Rapone, A.D. Roeder, Y. Ruiz, E. Sauer, P.M. Schneider, T. Sijen, D.S. Court, B. Sviezena, M. Turanska, A. Vidaki, L. Zatkalikova, J. Ballantyne, RNA/DNA co-analysis from human saliva and semen stains - Results of a third collaborative EDNAP exercise, *Forensic Sci. Int.: Genet.* 7 (2) (2013) 230–239, <http://dx.doi.org/10.1016/j.fsigen.2012.10.011>.
- [42] C. Haas, E. Hanson, M.J. Anjos, K.N. Ballantyne, R. Banemann, B. Bhoelai, E. Borges, M. Carvalho, C. Courts, G. De Cock, K. Drobic, M. Doetsch, R. Fleming, C. Franchi, I. Gomes, G. Hadzic, S.A. Harbison, J. Hartevelde, B. Hjort, C. Hollard, P. Hoff-Olsen, C. Huels, C. Keyser, O. Maronas, N. McCallum, D. Moore, N. Morling, H. Niederstaetter, F. Noel, W. Parson, C. Phillips, C. Popielarz, A.D. Roeder, L. Salvaderi, E. Sauer, P.M. Schneider, G. Shanthan, D.S. Court, M. Turanska, R.A.H. van Oorschot, M. Vennemann, A. Vidaki, L. Zatkalikova, J. Ballantyne, RNA/DNA co-analysis from human menstrual blood and vaginal secretion stains: Results of a fourth and fifth collaborative EDNAP exercise, *Forensic Sci. Int.: Genet.* 8 (1) (2014) 203–212, <http://dx.doi.org/10.1016/j.fsigen.2013.09.009>.
- [43] C. Haas, E. Hanson, R. Banemann, A.M. Bento, A. Berti, A. Carracedo, C. Courts, G. De Cock, K. Drobic, R. Fleming, C. Franchi, I. Gomes, G. Hadzic, S.A. Harbison, B. Hjort, C. Hollard, P. Hoff-Olsen, C. Keyser, A. Kondili, O. Maronas, N. McCallum, P. Miniati, N. Morling, H. Niederstaetter, F. Noel, W. Parson, M.J. Porto, A.D. Roeder, E. Sauer, P.M. Schneider, G. Shanthan, T. Sijen, D.S. Court, M. Turanska, M. van den Berge, M. Vennemann, A. Vidaki, L. Zatkalikova, J. Ballantyne, RNA/DNA co-analysis from human skin and contact traces – results of a sixth collaborative EDNAP exercise, *Forensic Sci. Int.: Genet.* 16 (2015) 139–147, <http://dx.doi.org/10.1016/j.fsigen.2015.01.002>.
- [44] M. van den Berge, A. Carracedo, I. Gomes, E. Graham, C. Haas, B. Hjort, P. Hoff-Olsen, O. Maroñas, B. Mevåg, N. Morling, et al., A collaborative European exercise on mRNA-based body fluid/skin typing and interpretation of DNA and RNA results, *Forensic Sci. Int.: Genet.* 10 (2014) 40–48.
- [45] S. Ingold, G. Dørum, E. Hanson, A. Berti, W. Branicki, P. Brito, P. Elsmore, K. Gettings, F. Giangasparo, T. Gross, et al., Body fluid identification using a targeted mRNA massively parallel sequencing approach – results of a EU-ROFORGEN/EDNAP collaborative exercise, *Forensic Sci. Int.: Genet.* 34 (2018) 105–115.
- [46] R.A. van Oorschot, B. Szkuta, G.E. Meakin, B. Kokshoorn, M. Goray, DNA transfer in forensic science: a review, *Forensic Sci. Int.: Genet.* 38 (2019) 140–166.
- [47] D. Taylor, B. Kokshoorn, A. Biedermann, Evaluation of forensic genetics findings given activity level propositions: a review, *Forensic Sci. Int.: Genet.* 36 (2018) 34–49.