AUTHORS' POSTPRINT
Bozza, S., Scherz, V., Greub, G., Falquet, L. and Taroni, F.
A probabilistic approach to evaluate salivary microbiome in forensic science when the Defense says: 'It is my twin brother'. Forensic Science International: Genetics 57 (2022) article 102638. doi.org/10.1016/j.fsigen.2021.102638

A probabilistic approach to evaluate salivary microbiome in forensic science

Abstract

Salivary microbiota profiles may represent a valid contribution to forensic investigation when standard DNA genotyping methods fail. Starting from questioned and control materials in the form of saliva, the evidence can be expressed by means of a distance between those materials taking into account specific aspects of the microbiota composition. The value of the evidence for forensic discrimination purposes is quantified by means of a Bayes' factor, that allows one to overcome the major limitations and pitfalls of intuition connected to the use of cut-off values as a mean of decision.

Key words: Bayes' factor; Cut-off; Discrimination; Monozygotic twins; Salivary microbiome; Similarity score; Evaluation of evidence

1. Introduction

The suitability of salivary microbiota composition to support discrimination between closely related individuals has been investigated in [16]. Microbiota profiles have shown high individuality and stability over time and may represent a valid and precious contribution for forensic investigations when standard genotyping DNA methods fail (e.g. due to degraded material or to close relatedness among individuals, such as homozygous twins). Previous studies have used microbiomes to differentiate an individual from a limited group of people (see, e.g. [6]).

Beta-diversity (β-diversity) indices (such as the Jaccard distance) are commonly used in microbiota studies to highlight taxonomical differences between pairs of samples [10]. They have been implemented to overcome the dimensionality problem that renders the quantification of the ressemblance between microbiota profiles problematic. Their adequacy and feasibility for discrimination between hypotheses of forensic interest have been discussed in [16].

The assessment of the value of a score quantifying the similarity between measurements originating from questioned material and from control material for discrimination purposes is an open problem, as reiterated by [6]. The use of a *cut-off* value against which such score is compared in order to support or even suggest a conclusion regarding their origin (e.g. whether compared saliva traces originate from the same individual or from closely related individuals) is often suggested. Routine examples come mainly from the toxicology domain where a scientist is often asked to classify a controlled substance in a blood or hair item into a predefined category (e.g. doping athletes, abusive drinkers, and so on). However, as it will be discussed in Section 5, such an approach presents severe limitations and leaves the problem of forensic interpretation open. Most importantly, it renders a binary response, no matter the effective similarity (as measured by a beta-diversity indece) between the compared material. In the current paper, a probabilistic approach to evidence evaluation is proposed. The most efficient way to provide a quantitative measure of the support of the evidence (i.e., a β-diversity distance) to competing hypotheses about the origin of a given salivary material is given by the Bayes' factor that expresses, in probabilistic form, the ratio between the probability to observe the evidence given each of the hypotheses of interest.

The paper is structured as follows. The scenario of interest and the proposed probabilistic approach are introduced in Section 2. Available data and the chosen statistical models are described in Section 3, while the results of statistical analyses are presented and discussed in Section 4. Section 5, finally, concludes the paper.

2. A probabilistic approach

Consider a case where a saliva trace is collected from a receptor (e.g. a given object on a crime scene). The salivary microbiome of the trace, as well as that of a saliva sample taken from a known source (say, Mr. X) for comparative purposes, is thus analyzed with the aim of discriminating between the following two competing hypotheses:

H_p: the saliva trace originates from Mr. X;

H_d: the saliva trace originates from the twin brother of Mr. X,

where the lower-case letter p stands for *prosecution* and the lower-case letter d stands for *defence*. Note that the defence proposition could also be specified as: H_d: 'the saliva trace originates from an individual unrelated to Mr. X', or H_d: 'the saliva trace originates from the twin brother of Mr. X or from an individual unrelated to Mr. X'. Here, one is faced with what is known as 'multiple hypotheses' scenario.

The scientific outcome, denoted here by the letter E, $E = \{y, x\}$, is given by the measurements y (i.e. the salivary microbiome) on the questioned material (i.e. the saliva trace recovered at the crime scene) and the measurements x (i.e. the salivary microbiome) on the control material of known source (i.e. a sample taken from Mr. X).

A probabilistic method for the evaluation of the strength of the evidence when it is the result of a comparative analysis between control and recovered materials was proposed by Lindley [13] and is given by the so-called *Bayes' factor* (BF):

$$\mathrm{BF} = \frac{f(y, x \mid H_p)}{f(y, x \mid H_d)}. \tag{1}$$

The use of the Bayes' factor (often referred in forensic science applications as the likelihood ratio) as a metric to assess the probative value of forensic traces (e.g. fingerprints, glass fragments, biological stains) is largely supported in different forensic disciplines [18]. A detailed presentation can be found in Aitken et al. [1].

A number of issues arise when one is faced with how such a Bayes' factor can be computed in practice. Some of these issues are related to the specification of a probabilistic model $f(\cdot)$ in cases where measurements are obtained using high-dimensional techniques, e.g. for fingermarks (using complex sets of variables) or for biological traces that may be described by several chemical components. In such situations where the risk is to make unfounded assumptions about the underlying mechanism that is at the origin of the evidence E, a *score-based* approach where the evidence is the result of a comparative analysis between questioned and control material may represent a viable or even the sole alternative. This is what happens in the current case where analytical results are over 7000-dimensional [**?**].

Let $\delta(x, y)$ denotes the *similarity score* quantifying the distance between the compared material. A *score-based* Bayes factor (sBF) can be obtained as:

$$\mathrm{sBF} = \frac{g(\delta(x, y) \mid H_p)}{g(\delta(x, y) \mid H_d)}. \tag{2}$$

The score-based Bayes factor (sBF) in (2) replies to the following questions: how much more likely is the score between the salivary microbiomes coming from the recovered trace and from the person of interest (say, Mr. X) to arise under the hypothesis that the two salivary samples originate from the same individual (H_p holds) or from alternative individuals - a twin brother or/and an unrelated person - (H_d holds).

The choice of a score-based approach leaves other questions that need to be tackled: (i) the choice of the metric $\delta(\cdot)$, and (ii) the choice of the probability distribution $g(\cdot)$ under the competing propositions H_i, $i = \{p, d\}$. These aspects will be discussed in Section 3. A comment on the use of score-based BFs in forensic science applications is given in Neumann and Ausdemore [14].

3. Material and models

3.1. Data

A longitudinal study on 30 pairs of homozygous twins (29 couples and 1 triplet) has shown the potential of salivary microbiota-based profiles for discrimination between closely related individuals [16]. Four salivary samples have been collected from each participant: the first sample was taken during the first inclusion visit, and the rest in correspondence of the follow-up visits that have been planned 1, 12 and 13 months after inclusion. DNA was then extracted from the native saliva samples with a targeted time to extraction less than 8 days. The intra-individual variability over time was computed comparing all four samples of each individual to each others. Similarly, the intra-pair variability was obtained comparing samples from one individual to the samples from their twin brother. Finally, all samples have been compared to samples from the other pairs (inter-individual variability).

In Scherz et al. [16] three similarity scores (i.e. Jaccard, Bray-Curtis and Jensen-Shannon distances) have been compared to analyze the β-diversity within different samples from the same individual (*Intra-individual*), within different samples from related individuals (*Intra-pair*) and between unrelated individuals (*Inter-individual*). The Jaccard distance has shown a better discrimination once dealing with native samples, and it is retained in this study. In Figure 1 there are shown the Jaccard distances calculated for samples originating from the same individual (*Intra-individual*), from couples of siblings (*Intra-pair*, or *Twins*) and from couples of unrelated persons (*Inter-individual*, or *Unrelated*).

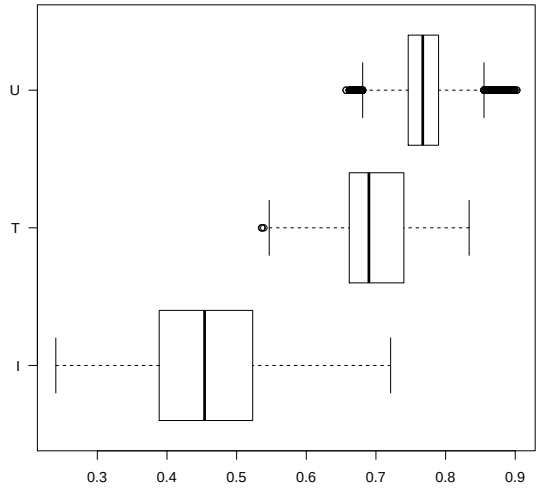

Figure 1: Jaccard distances for salivary microbiota compositions of pairs of samples from individual persons (Intra-individual, I), pairs of related persons (Twins, T) and pairs of unrelated persons (Unrelated, U).

A Bayesian analysis of variance has been conducted in order to study the effect of several variables in explaining the β-diversity. This allows to overcome limitations of standard analysis of variance in presence of non-Normal dependent data. For each variable (effect) there are reported the posterior means and the 95% credible intervals. An effect is considered significant whenever the associated 95% credible interval does not contain zero. Moreover, it is also reported the effect size, measuring the amount of variability explained by a given variable.

Consider first the intra-individual β-diversity. The following variables have been considered: the time interval between visits (less (equal to)/more than 1 month, variable Group 1), the age of participants (younger (equal to)/older than 30, variable Age) and the difference in time of collection (less (equal to)/more than 4 hours between the hour of the day of saliva collection of compared samples, variable 4h). The results of the (Bayesian) analysis of variance are displayed in Table 1. The intra-individual salivary microbiome composition diverged over time between collection of saliva traces (variable Group 1) with an effect size equal to 0.35 (roughly). The time interval between visits turned out to be the variable with the most significant effect, followed by the age (older individuals tend to have less changes in their microbiome). No significant effect has been observed with reference to the time interval (variable 4h). No interactions amongst variables have been observed (results are not reported). The intra-individual Jaccard distances are represented in Figure 2 (left) with a distinction based on time interval between visits, and in Figure 3 with a distinction based on age of individuals (left) and daily time interval between visits (right).

Secondly, in order to account for the intra-pair β-diversity, the Jaccard distances are calculated between measurements from salivary samples originating from homozygous twins. In this case it has also been considered whether twins had been sharing the same house (variable Home). Variables Group 1 and Age have been defined earlier. The results of the Bayesian analysis of variance are displayed in Table 2. Note that in this case the time interval between

Source	Table of coefficients		Table of effect sizes	
	Mean	95% credible interval	Effect size	95% credible interval
Group 1	-0.0559	(-0.0669,-0.0433)	0.3497	(0.331,0.359)
Age	0.0221	(0.0110,0.0335)	0.0801	(0.046,0.109)
4h	0.0009	(-0.0218,0.0007)	0.0000	(-0.027,0.001)

Table 1: Intra-individual Jaccard distance: Bayesian analysis of variance. Table of coefficients (columns 2-3); Table of effect sizes (columns 4-5).

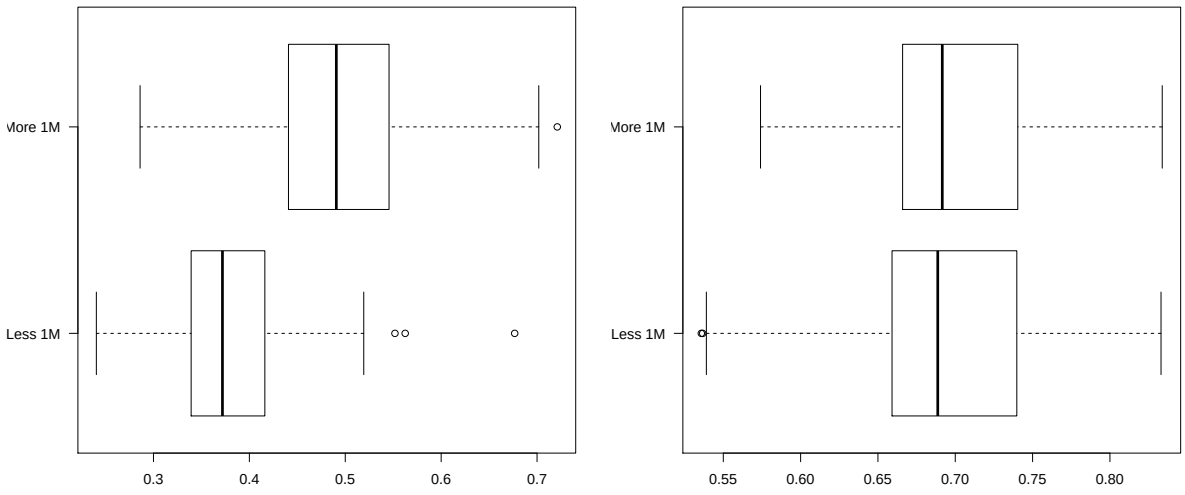

Figure 2: Jaccard distances for salivary microbiome compositions of pairs of samples from the same individual (left) and from couples of twins (right). A distinction is based on time interval between visits (not greater or greater than one month).

visits (variable Group 1) is not significant and will not be taken into account in what follows (see also Figure 2 (right)). The intra-pair salivary microbiome composition diverged for twins sharing and not sharing the same house (variable Home) with an effect size equal to 0.05. The intra-pair Jaccard distances are represented in Figure 4 with a distinction based on the age of participants (left) and on the fact that twins share/do not share the same home (right).

Source	Table of coefficients		Table of effect sizes	
	Mean	95% credible interval	Effect size	95% credible interval
Group 1	-0.0034	(-0.0115 ,0.0041)	0.0000	(-0.026,0.004)
Age	-0.0069	(-0.0161,0.0032)	0.0142	(-0.017,0.052)
Home	0.0142	(0.039,0.0252)	0.0515	(0.005,0.111)

Table 2: Intra-pair Jaccard distance: Bayesian analysis of variance. Table of coefficients (columns 2-3); Table of effect sizes (columns 4-5).

Table 3, finally, contains the results of comparisons between intra-individual distances and intra-pair distances (variable Group 2), and between intra-individual distances and inter-individual distances (variable Group 3). Intra-individual distances are significantly smaller than intra-pair distances with an effect size equal to 0.72; intra-individual distances are significantly smaller than inter-individual distances with an effect size equal to 0.34.

3.2. Models

Denote by $\{z_{lij},\, l = 1,2,\, i = 1,\ldots,m_p,\, j = 1,\ldots,n_l\}$ the intra-individual distances, where $l = 1(2)$ for comparisons at a time interval not greater (greater) than 1 month, $m_p = 61$ is the number of participants and $n_{1(2)} =$

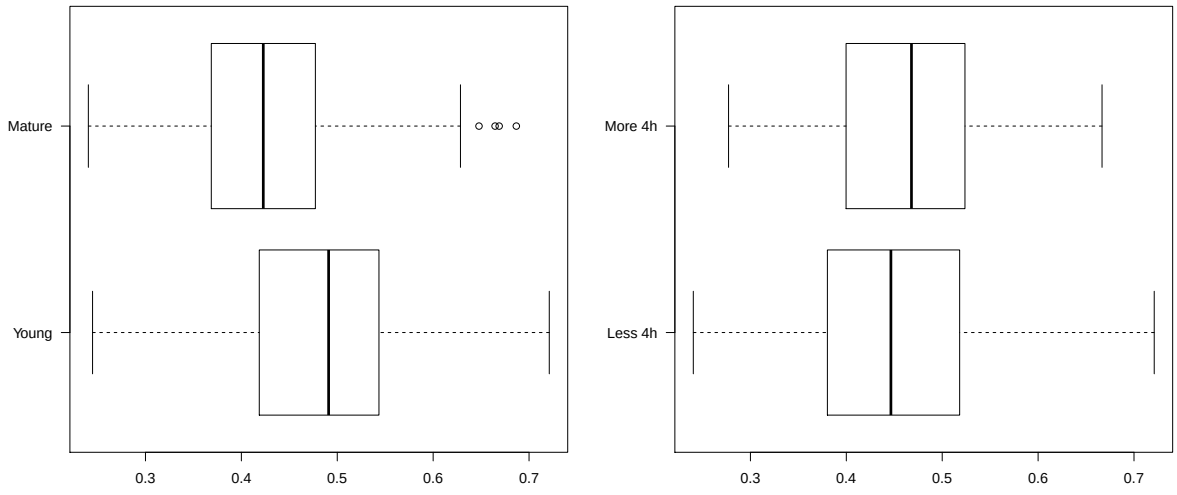

Figure 3: Jaccard distances for salivary microbiome compositions of pairs of samples from the same individual. Left: variable Age (individuals younger (Young) or older (Mature) than 30); right: variable 4h (daily time interval (in hours) between compared samples less (equal to)/more than 4 hours) (right).

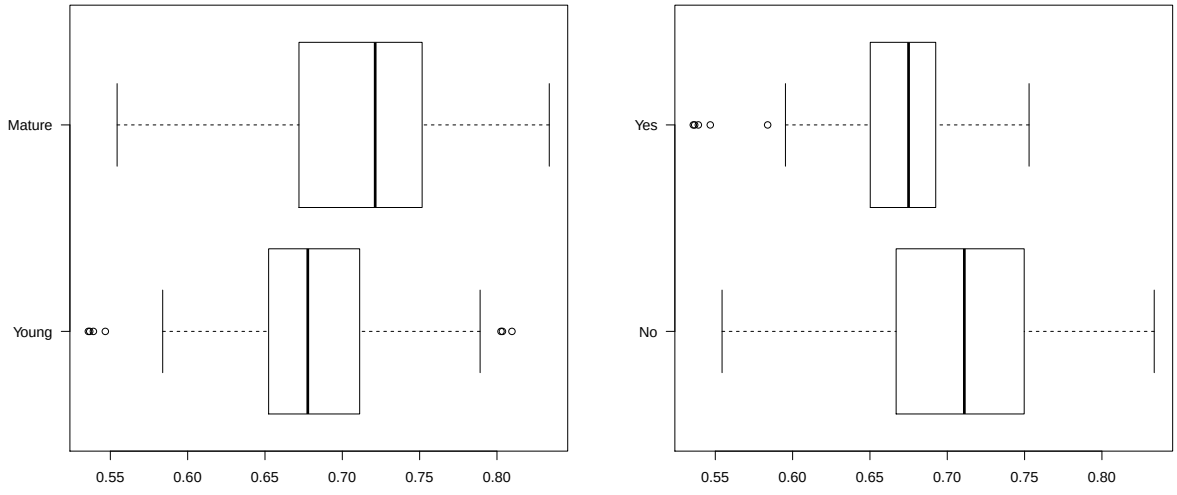

Figure 4: Jaccard distances for salivary microbiome compositions of pairs of samples from couples of twins. Left: variable Age (individuals younger (Young) or older (Mature) than 30); right: variable Home (individual sharing (Yes) or not sharing (No) the same home).

Source	Table of coefficients		Table of effect sizes	
	Mean	95% credible interval	Effect size	95% credible interval
Group 2	-0.1204	(-0.1265,-0.1146)	0.7266	(0.725,0.727)
Group 3	-0.1554	(-0.1570,-0.1536)	0.3454	(0.345,0.345)

Table 3: Comparison between Intra-individual, Intra-pair and Inter-individual Jaccard distance: Bayesian analysis of variance. Table of coefficients (columns 2-3); Table of effect sizes (columns 4-5).

$2(4)$ is the available number of intra-volunteer comparisons. Denote by $\{t_{ij},\, i = 1, \ldots, m_t,\, j = 1, \ldots, n_t\}$ the intra-pair distances, where $m_t = 30$ are the number of couples/triplets and $n_t = 16$ is the minimum number of intra-pair comparisons (n_t is augmented to 48 in correspondence of the triplet). Finally, denote by $\{u_{ij},\, i = 1, \ldots, m_p,\, j = 1, \ldots, n_u\}$ the inter-individual distances, where n_u is the available number of inter-individual comparisons.

A probability distribution must be fitted for the numerator and for the denominator of the (score-based) Bayes factor in (2). The (score-based) Bayes' factor must in fact be calculated as the ratio of two probability density functions $g(\cdot)$ evaluated at the evidence score $\delta(x, y)$ (e.g. the Jaccard distance between salivary microbiota composition of the compared samples) if proposition H_p (H_d) holds. The two density functions can be assessed using many sample scores produced under the competing propositions and the Bayes factor can be obtained as

$$\text{sBF} = \frac{\hat{g}(\delta(x, y) \mid x, H_p)}{\hat{g}(\delta(x, y) \mid y, H_d)}. \tag{3}$$

This would amount to generate many scores for comparisons between: (i) the measurements x on the control material (whose source is known) and measurements on other salivary samples originating from the same source (numerator); (ii) the measurements y on the recovered material and measurements on other salivary samples taken randomly from the available database (denominator), where proposition H_d is formulated as 'The saliva trace originates from an individual unrelated to Mr. X'. Ideally, one should dispose of many samples from the same individual in order to be able to assess a probability distribution at the numerator that is specific for a person of interest. In the current study, however, a total number of 6 intra-volunteer comparisons are available for each participant. These reduce to 2 if the retained time interval between the deposit of the compared material does not exceed one month (i.e., the comparison between material collected at the beginning of the study and at the first follow-up visits planned after one month, and that between material collected at the 12^{th} month after inclusion and at the end of the study). In the same way, the assessment of the probability distribution at the denominator may be problematic whenever the alternative proposition H_d is formulated as 'The saliva trace originates from the twin brother of Mr. X'. In this latter case, in fact, the measurements y on the recovered material should be compared on other salivary materials originating from the twin brother of Mr. X, and the available samples may be limited.

A feasible alternative is to perform a so-called *common-source* approach and compare the following pair of propositions:

H_p: the saliva traces originate from the same individual;

H_d: the saliva traces originate from an individual and their twin brother.

The Bayes factor is therefore calculated as

$$\text{sBF} = \frac{\hat{g}(\delta(x, y) \mid H_p)}{\hat{g}(\delta(x, y) \mid H_d)}. \tag{4}$$

This amounts to adopt a so-called *non-anchored* approach at the numerator and at the denominator.

Consider the numerator first. Two sources of variation are taken into account: that between measurements (i.e. scores) within the same individual (known as the *within-source* variation), and that between individuals (known as *between-source* variation). For the within-source variation, denote the mean distance within individual i for measurements related to material collected at a time interval l by θ_{li} (where $l = 1(2)$ denotes that the similarity distance $\delta(x, y)$ has been calculated for salivary material x, y collected at a time interval not greater (greater) than one month) and the variance by σ_l^2. The distribution of Z_{lij} is taken to be Normal, $(Z_{lij} \mid \theta_{li}, \sigma_l^2) \sim \text{N}(\theta_{li}, \sigma_l^2)$, $l = 1, 2$,

$i = 1, \ldots, m_p$ and $j = 1, \ldots, n_l$. For the between-source variation, the inspection of Figure 5 (left) suggests that a lognormal distribution may be assumed for either the salivary comparisons taken at a time interval not greater (top left) and greater (bottom left) than 1 month. However, a deviation in the tails is observed in both cases Figure 5 (right), and a kernel density estimate is suggested (Figure 6).

Starting from the available measurements, the probability density function for the between-source variation can be estimated as:

$$\begin{aligned} f(\theta_l \mid \bar{z}_{l1}, \ldots, \bar{z}_{lm_p}) &= \frac{1}{m_p} \sum_{i=1}^{m_p} \mathrm{K}(\theta_l \mid \bar{z}_{li}, \tau_l^2, h_l) \\ &= \frac{1}{m_p} \sum_{i=1}^{m_p} \frac{(2\pi)^{-1/2}}{h_l \tau_l} \exp\left(-\frac{(\theta_l - \bar{z}_{li})^2}{2h_l^2 \tau_l^2}\right), \end{aligned} \tag{5}$$

where $\bar{z}_{li} = \frac{1}{n_l} \sum_{j=1}^{n_l} z_{lij}$ are the sample means, h_l is the smoothing parameter and the kernel density function $\mathrm{K}(\cdot)$ is a Normal density function centered at $\bar{z}_{li}$ with variance equal to $h_l^2 \tau_l^2$.

The probability density function of the evidence $\delta(x, y)$ in the numerator is given by

$$\hat{g}_l(\delta(x,y) \mid H_p) = \int_{\theta_l} g(\delta(x,y) \mid \theta_l, \sigma_l^2) f(\theta_l \mid \bar{z}_{l1}, \ldots, \bar{z}_{lm_p}, \tau_l^2, h_l) d\theta_l. \tag{6}$$

The integral in (6) has an analytical solution and the probability density in the numerator takes the form

$$\hat{g}_l(\delta(x,y) \mid \sigma_l^2, \bar{z}_{l1}, \ldots, \bar{z}_{lm_p}, \tau_l^2, h_l) = \frac{1}{m_p} \sum_{i=1}^{m_p} \frac{1}{\sqrt{2\pi(\sigma_l^2 + h_l^2 \tau_l^2)}} \exp\left\{-\frac{1}{2(\sigma_l^2 + h_l^2 \tau_l^2)} \left(\delta(x,y) - \bar{z}_{li}\right)^2\right\}. \tag{7}$$

Note that a distinction has been made to take into account the time interval ($l = 1, 2$) between the deposit of the compared material. However, it must be acknowledged that this information is often unavailable. In that case, the probability density function in the numerator can be obtained as a weighted average as follows:

$$\hat{g}(\delta(x,y) \mid H_p) = \sum_{l=1}^{2} w_l \, \hat{g}_l(\delta(x,y) \mid \sigma_l^2, \bar{z}_{l1}, \ldots, \bar{z}_{lm_p}, \tau_l^2, h_l), \tag{8}$$

where the weights w_l represent the probabilities that the time interval is not greater (w_1) or greater (w_2) than 1 month.

As far as the denominator, two sources of variation are taken into account: that between measurements (i.e. scores) within the same couple (within-pair), and that between different couples (between-pair). For the within-pair variation, denote the mean distance within couple i by θ_i and the variance by σ_t^2. The distribution of T_{ij} is taken to be Normal, $(T_{ij} \mid \theta_i, \sigma_t^2) \sim \mathrm{N}(\theta_i, \sigma_t^2)$. The between-source distribution is bimodal, as it can be observed from the inspection of Figure 7. A kernel density estimate is proposed as in (5) and is depicted in Figure 7 (right). The probability density function of the evidence in the denominator $\hat{g}(\delta(x,y) \mid H_d)$ is obtained analogously as in (7), and will be denoted $\hat{g}(\delta(x,y) \mid \sigma_t^2, \bar{t}_1, \ldots, \bar{t}_{m_t}, \tau_t^2, h_t)$.

The Bayes factor can then be obtained as in (2):

$$\frac{\hat{g}_l(\delta(x,y) \mid \sigma_l^2, \bar{z}_{l1}, \ldots, \bar{z}_{lm_p}, \tau_l^2, h_l)}{\hat{g}(\delta(x,y) \mid \sigma_t^2, \bar{t}_1, \ldots, \bar{t}_{m_t}, \tau_t^2, h_t)} = \frac{\sum_{i=1}^{m_p} \frac{1}{\sqrt{2\pi(\sigma_l^2 + h_l^2 \tau_l^2)}} \exp\left\{-\frac{1}{2(\sigma_l^2 + h_l^2 \tau_l^2)} \left(\delta(x,y) - \bar{z}_{li}\right)^2\right\} / m_p}{\sum_{i=1}^{m_t} \frac{1}{\sqrt{2\pi(\sigma_t^2 + h_t^2 \tau_t^2)}} \exp\left\{-\frac{1}{2(\sigma_t^2 + h_t^2 \tau_t^2)} \left(\delta(x,y) - \bar{t}_i\right)^2\right\} / m_t}. \tag{9}$$

Note that, following the observations made in Section 3, the time interval between visits has not taken into account for intra-pair comparisons.

Clearly, if the alternative hypothesis supported by the defence states that saliva traces originate from unrelated individuals, the distances from unrelated individuals, u_{ij}, must be taken into account in order to obtain the probability distribution at the denominator. Inspection of Figure 8 suggests a deviation from Normality in the right tail. A kernel density estimate can be fitted as above. The probability density in the denominator, now denoted $\hat{g}(\cdot \mid \sigma_u^2, \bar{u}_1, \ldots, \bar{u}_{m_p}, \tau_u^2, h_u)$ can be obtained as in (7).

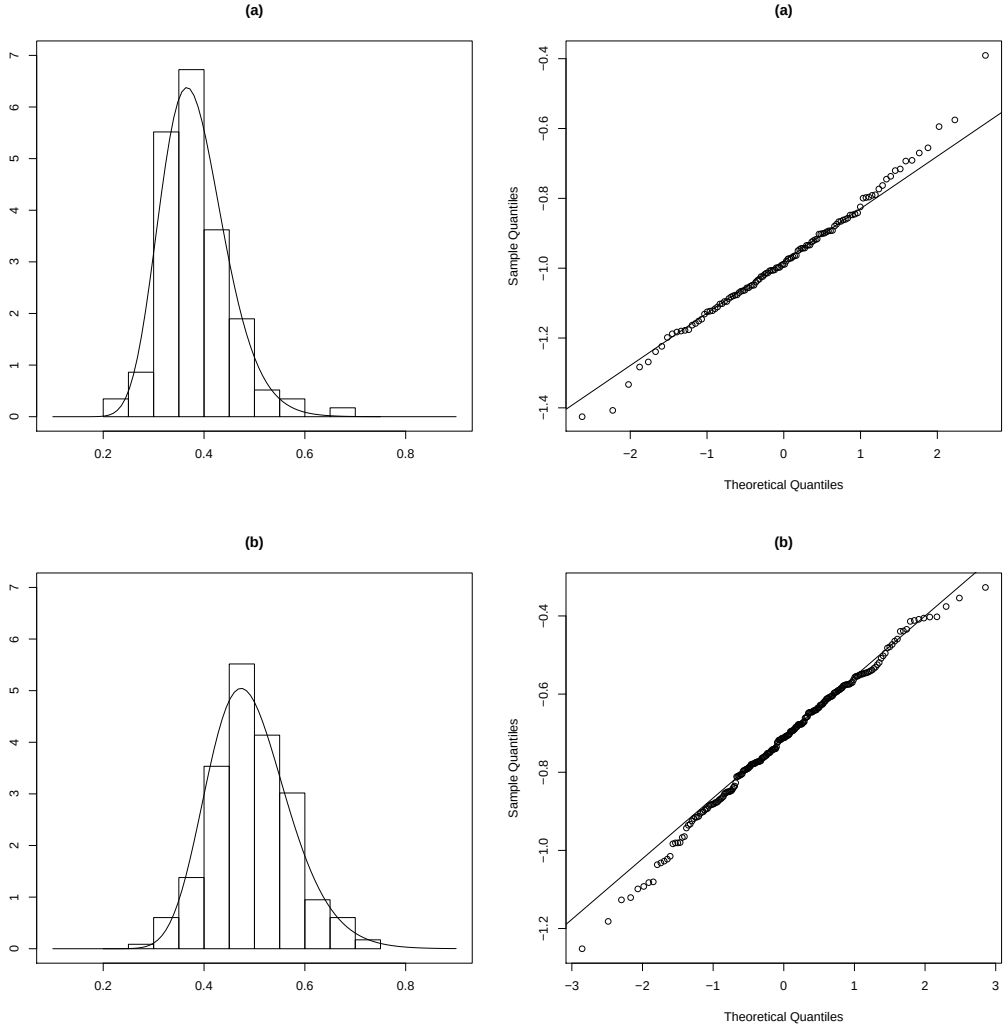

Figure 5: Intra-individual distances. (a) Scores obtained considering only visits at a time interval not greater than 1 month: overlay between the histogram of available data and a fitted lognormal distribution (left); quantile-quantile plot of the logarithm of available data; (b) Scores obtained considering only visits at a time interval greater than 1 month: overlay between the histogram of available data and a fitted lognormal distribution (left); quantile-quantile plot of the logarithm of available data.

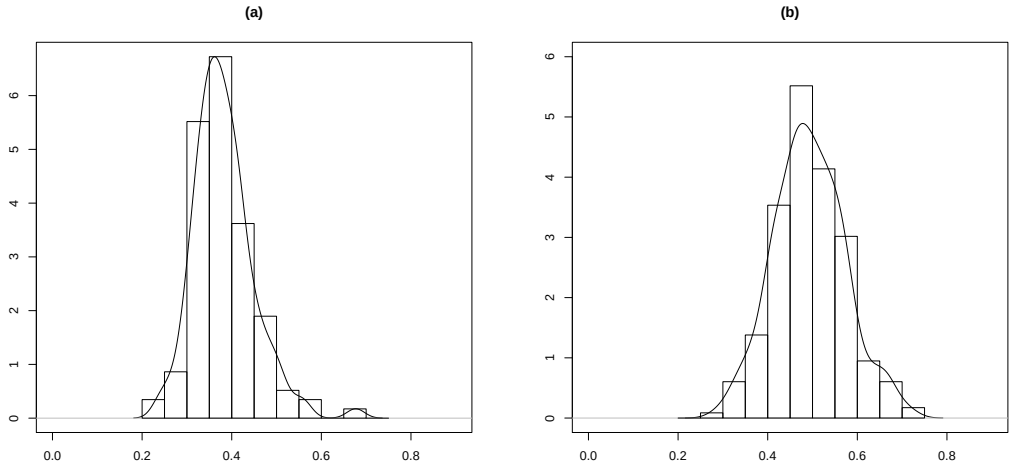

Figure 6: Intra-individual distances. Scores obtained considering only visits at a time interval not greater than 1 month: overlay between the histogram of available data and a kernel density estimate (a); Scores obtained considering only visits at a time interval greater than 1 month: overlay between the histogram of available data and a kernel density estimate (b).

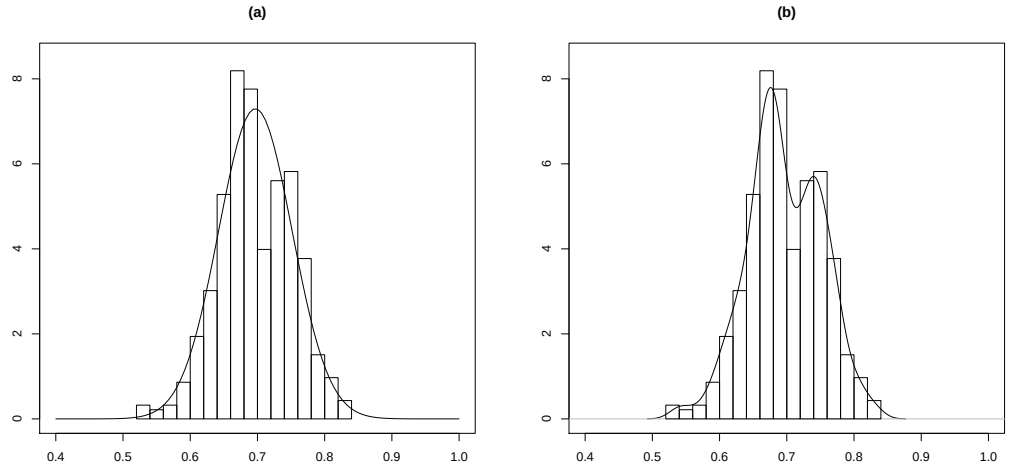

Figure 7: Scores obtained considering intra-pair comparisons: overlay between the histogram of available data and a fitted Normal distribution (a); overlay between the histogram of available data and a kernel density estimate (b).

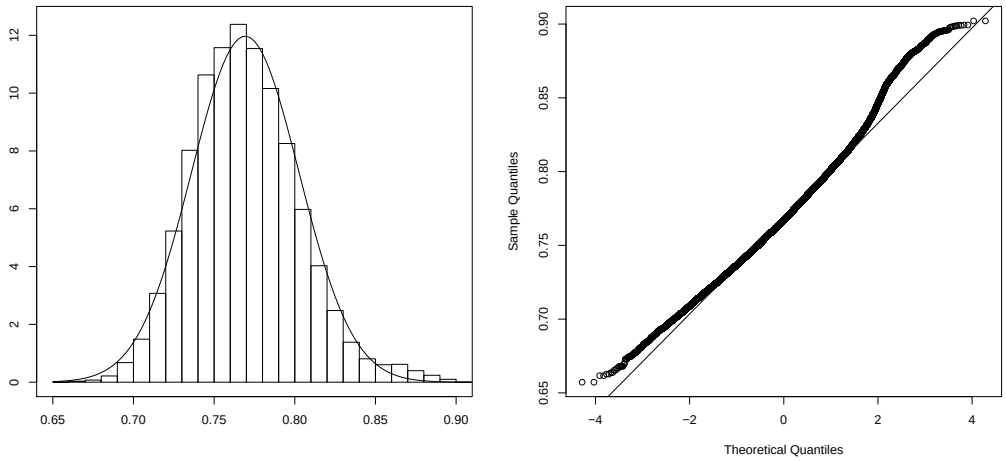

Figure 8: Scores obtained considering inter-volunteers comparisons: overlay between the histogram of available data and a fitted normal distribution (left); quantile-quantile plot of available data (right).

There are scenarios where the alternative hypothesis, i.e. the defence hypotheses, may be thought of as a composite hypothesis. There is more than one possible alternative to the first hypothesis (the salivary microbiota comes from the same person). Under the defence hypothesis, H_d, someone else is the source of the salivary microbiota, the term 'someone else' refers to a set of donors as the potential source. It might be a twin brother or an unrelated person.

Scientific literature pointed out that there are situations where there will be three or more propositions. This can happen with DNA mixtures, for example, where the number of contributors to the mixture is in dispute [5] or in cases involving the DNA profile of a single stain of body fluids when relatives of the defendant are considered as potential donors of the recovered stain. Discussion on this aspect dated back to Lempert [12], Evett [8], Balding [2] and Buckleton and Triggs [4]. This is the situation we are concerned. Consider, for sake of illustration, the following competing propositions:

H_p: the saliva trace originates from Mr. X;

H_d: the saliva trace originates from the twin brother of Mr. X or from an unknown individual unrelated to Mr. X.

The Bayes' factor can then be obtained as presented in Aitken et al. [1]

$$\frac{\hat{g}_l(\delta(x,y) \mid \sigma^2_l, \bar{z}_{l1}, \ldots, \bar{z}_{lm_p}, \tau^2_l, h_l) \sum_{i=\{t,u\}} p_i}{\hat{g}(\delta(x,y) \mid \sigma^2, \bar{t}_1, \ldots, \bar{t}_{m_t}, \tau^2_t, h_t)p_t + \hat{g}(\delta(x,y) \mid \sigma^2, \bar{u}_1, \ldots, \bar{u}_{m_p}, \tau^2_u, h_u)p_u}, \tag{10}$$

where p_t and p_u are the prior probabilities of the hypothesis that the questioned material originates from the twin brother or from an unrelated, respectively, and $p_i = \{p_t, p_u\}$, and $\sum_{i=\{t,u\}} p_i < 1$.

4. Data analysis

All comparisons, graphical representations and statistical analyses were conduced in RStudio (Version 1.2.5033), supported by the BANOVA package [7].

4.1. Summary statistics

The model parameters at the numerator can be estimated from the available data $\{z_{ij}\}$ by

$$\hat{\mu} = \bar{z} = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} z_{ij} \quad (11)$$

$$\hat{\sigma}^2 = \frac{1}{m(n-1)}\sum_{i=1}^{m}\sum_{j=1}^{n}(z_{ij}-\bar{z}_i)^2 \quad (12)$$

$$\hat{\tau}^2 = \frac{1}{m-1}\sum_{i=1}^{m}(\bar{z}_i-\bar{z})^2 - \frac{\hat{\sigma}^2}{n}, \quad (13)$$

$$\hat{h} = 1.06m^{1/5}, \quad (14)$$

where $\bar{z}_i = \sum_{j=1}^{n} z_{ij}$ and the the smoothing parameter h is estimated as in Silverman [17].

The model parameters at the denominator can be estimated analogously using either the database $\{t_{ij}\}$ or $\{u_{ij}\}$ of intra-pair and inter-individual distances.

4.2. Results

To study the distribution of Bayes' factor results under the competing propositions, a leave-one-out method has been used. A Bayes' factor has been calculated for measurements originating from each couple in the database, while the remaining data have been used to fit the probabilistic distribution at the numerator and the denominator.

To test the hypothesis H_p : 'the saliva traces originate from the same individual', the salivary microbiome characterizing saliva traces collected from the same individual have been selected to act as control (x) and recovered (y) data, respectively. The (Jaccard) distance $\delta(x, y)$ is therefore calculated between the salivary composition characterizing the selected samples.

To test the hypothesis H_d : 'the saliva traces originate from an individual and their twin brother', the salivary microbiome characterizing saliva traces collected from one individual and from their twin brother have been selected to act as control and recovered data, respectively.

Table 4 shows the values of the Bayes' factor (within intervals) when the evidence comes from the same individual and the time interval between the collection of compared salivary profiles is not greater than 1 month (column 2) and greater than 1 month (column 3). The correct classification rate is at roughly 98% when considering samples collected at a time interval not greater than one month; it is slightly worse (roughly 92%) when considering samples collected at a greater time interval. This is not surprising as the overlap between the intra-individual and the intra-pair distributions increases.

Table 5 shows the values of the Bayes' factor (within intervals) when the evidence comes from monozygous twins. Results in column 2 have been obtained using, at the numerator, the fitted probability density function for intra-individuals similarity distances at a time interval between visits not greater than 1 month. Analogously, in column 3 there are reported the results that have been obtained using, at the numerator, the fitted probability density function for intra-individuals similarity distances at a time interval between visits greater than 1 month. Again, performances are slightly worse when a bigger time interval between the collection of compared material is retained.

Other variables may have an impact in explaining the diversity between compared material. For example, it has been observed in Section 3 that the intra-pair diversity decreases for pairs of twins sharing the same house and (slightly) increases for older pairs of twins (see Table 2 and Figure 4). The model performance increase when the variable Home (pairs sharing/not sharing the same house (Table 6), or the variable Age of the compared couples (Table 7) are considered.

BF	Time interval between visits	
	≤ 1 month	> 1 month
$< 10^{-3}$	1	0
$10^{-3} - 10^{-2}$	0	0
$10^{-2} - 10^{-1}$	0	5
$10^{-1} - 1$	1	14
$1 - 10$	1	37
$10 - 10^{2}$	4	36
$10^{2} - 10^{3}$	4	33
$10^{3} - 10^{4}$	4	36
$10^{4} - 10^{5}$	4	5
$10^{5} - 10^{6}$	11	19
$10^{6} - 10^{7}$	9	14
$10^{7} - 10^{8}$	10	6
$10^{8} - 10^{9}$	12	4
$10^{9} - 10^{10}$	9	5
$10^{10} - 10^{15}$	39	8
$10^{15} - 10^{20}$	7	0
$> 10^{20}$	0	0
Total	116	232
Correct (%)	98	92

Table 4: Assessment of the performance of the probabilistic approach for the evaluation of the similarity of salivary microbiome when the samples originate from the same source (H_p). The value smaller than 10^{-3} is of the order of 10^{-5}; no extreme values were observed.

BF	Time interval between visits	
	≤ 1 month	> 1 month
$< 10^{-6}$	0	0
$10^{-6} - 10^{-5}$	7	0
$10^{-5} - 10^{-4}$	12	0
$10^{-4} - 10^{-3}$	130	3
$10^{-3} - 10^{-2}$	212	76
$10^{-2} - 10^{-1}$	85	218
$10^{-1} - 1$	12	144
$1 - 10$	6	18
$10 - 10^{2}$	0	5
$> 10^{2}$	0	0
Total	464	464
Correct (%)	99	95

Table 5: Assessment of the performance of the probabilistic approach for the evaluation of the similarity of salivary microbiome when the samples originate from monozygous twins (H_d); no extreme values were observed.

BF	Cohabitants		Non cohabitans	
	≤ 1 month	> 1 month	≤ 1 month	> 1 month
$< 10^{-3}$	0	0	0	0
$10^{-3} - 10^{-2}$	1	0	0	0
$10^{-2} - 10^{-1}$	0	3	0	2
$10^{-1} - 1$	1	11	0	2
$1 - 10$	1	19	0	15
$10 - 10^2$	3	22	0	15
$10^2 - 10^3$	2	23	1	11
$10^3 - 10^4$	1	26	3	5
$10^4 - 10^5$	3	15	2	6
$10^5 - 10^6$	8	12	1	0
$10^6 - 10^7$	7	15	0	5
$10^7 - 10^8$	8	3	2	2
$10^8 - 10^9$	9	6	2	1
$10^9 - 10^{10}$	10	5	3	0
$10^{10} - 10^{15}$	25	8	10	0
$10^{15} - 10^{20}$	4	0	6	0
$10^{20} - 10^{25}$	1	0	1	0
$> 10^{25}$	0	0	1	0
Total	84	168	32	64
Correct (%)	98	92	100	94

Table 6: Assessment of the performance of the probabilistic approach for the evaluation of the similarity of salivary microbiome when the samples originate from the same source (H_p) and the variable Home is taken into account.

BF	Young		Mature	
	≤ 1 month	> 1 month	≤ 1 month	> 1 month
$< 10^{-3}$	0	0	0	0
$10^{-3} - 10^{-2}$	1	0	0	0
$10^{-2} - 10^{-1}$	0	2	0	1
$10^{-1} - 1$	1	9	1	5
$1 - 10$	0	28	0	6
$10 - 10^2$	3	32	1	6
$10^2 - 10^3$	3	25	0	11
$10^3 - 10^4$	3	11	0	21
$10^4 - 10^5$	4	8	2	10
$10^5 - 10^6$	1	9	4	8
$10^6 - 10^7$	6	2	4	15
$10^7 - 10^8$	5	2	6	2
$10^8 - 10^9$	6	0	3	6
$10^9 - 10^{10}$	8	0	7	5
$10^{10} - 10^{15}$	19	0	19	7
$10^{15} - 10^{20}$	3	0	4	1
$10^{20} - 10^{25}$	1	0	1	0
$> 10^{25}$	0	0	0	0
Total	64	128	52	104
Correct (%)	97	91	98	94

Table 7: Assessment of the performance of the probabilistic approach for the evaluation of the similarity of salivary microbiome when the samples originate from the same source (H_p) and the variable Age is taken into account.

5. Conclusion

The discrimination between individuals based on their salivary microbiome profiles has been investigated. Individual profiles have shown high individuality and stability. The fact that self comparisons characterizing the same individual tend to have a modest distance if compared with intra-pair (or inter-individuals) comparisons does not however allow one to conclude that a low value of such distance is suggestive to say that the compared material originate from the same individual. This can be observed in Figure 9, where it is clear that the distribution of the distances originating from either self-comparisons or intra-pair comparisons overlap. A low value of the similarity score between compared material does not allow to conclude that the saliva trace originates from the same individual, and vice versa. This is one of the reasons why the use of cut-offs to reply to the questions of interest is often misleading and inadequate (see discussion in Biedermann et al. [3] and Aitken et al. [1]).

The introduction of a probabilistic approach, and in particular the assignment of a Bayes' factor, represents the correct way to quantify the value of the available evidence. The performance of the proposed model has been assessed in terms of correct classification rate under competing hypotheses (H_p and H_d). The high values of correct classification rates clearly support the use of such a probabilistic metrics for the evaluation of this peculiar type of evidence. In addition to desiderata upon which scientific and legal literature converge (notably the desirable BF's properties of balance, transparency, robustness, added value, flexibility and logic [e.g. 9, 11], such performance results clearly indicate that Bayes' factor is able to discriminate between hypothesis regarding highly related individuals (i.e. monozygous twins) based on microbiota salivary data. BF avoids logical problems related to inductive (and decisional) reasoning that the widespread use of cut-off in some forensic related fields such as toxicology misinterpret. BF is able to assess the value for every quantified scores so to avoid the so-called *fall-off the cliff effect* [15] that classify two proxy measurements into two extremely different categories (e.g. same individuals versus a twin). A measurement just below or just above the cut-off value will be classified into different categories provoking an under- or over-evaluation of the observed score. It seems important to restate that the BF measures the effect of scientific findings (i.e. the similarity score) on the probabilities of the two (or more) hypotheses of interest. It measures the amplitude of the change from the probabilities of the hypotheses before the acquisition of the evidence to that after such acquisition. The BF does not allow a scientist to quantify probabilities related to the hypotheses and it plays a role in the decisional process about whether a given individual should be classified into one or the other of the categories, without permitting one to express a categorical opinion. Scientific observations alone provide an incomplete knowledge basis for decision. The use of the BF just recalls the forensic users to respect their role in the judicial procedure.

Finally, it is important to note that the models were developed by totally ignoring the potential biological limitations coming from saliva low quantities, degradation or potential mixture situations. Any future research should also consider these added factors.

Acknowledgment

References

[1] C G G Aitken, F Taroni, and S Bozza. *Statistics and the Evaluation of Evidence for Forensic Scientists*. John Wiley & Sons, Chichester, 3rd edition, 2021.

[2] D J Balding. Interpreting DNA evidence: can probability theory help? In J L Gastwirth, editor, *Statistical Science in the Courtroom*, pages 51–70. Springer-Verlag, New York, 2000.

[3] A Biedermann, F Taroni, S Bozza, , M Augsburger, and C G G Aitken. Critical analysis of forensic cut-offs and legal thresholds: A coherent approach to inference and decision. *Forensic Science International*, 288:72–80, 2018.

[4] J S Buckleton and C M Triggs. Relatedness and DNA: are we taking it seriously enough? *Forensic Science International*, 152:115–119, 2005.

[5] J S Buckleton, Evett I W, and B S Weir. Setting bounds for the likelihood ratio when multiple hypotheses are postulated. *Science & Justice*, 38:23–26, 1998.

[6] T H Clarke, H Gomez, S Singh, K E Nelson, and L M Brinkac. Integrating the microbiome as a resource in the forensics toolkit. *Forensic Science international: Genetics*, 30:141–147, 2017.

[7] C Dong, M Wedel, and A Kopyakova. Package 'BANOVA', 2021. https://cran.r-project.org/web/packages/BANOVA/BANOVA.pdf.

[8] I W Evett. Evaluating DNA profiles in the case where the defence is 'it was my brother'. *Journal of the Forensic Science Society*, 35:5–14, 1992.

[9] I W Evett. Expert evidence and forensic misconceptions of the nature of exact science. *Science & Justice*, 36:118–122, 1996.

[10] F Finotello, E Mastrorilli, and B Di Camillo. Measuring the diversity of the human microbiota with targeted next-generation sequencing. *Briefings in Bioinformatics*, 19:679–692, 2018.

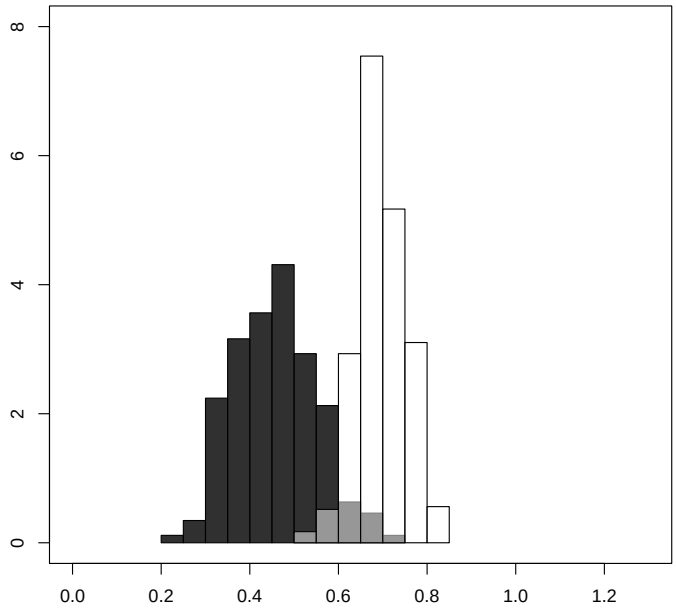

Figure 9: Overlay between the histogram of same-source comparisons (black) and intra-pair comparisons (white).

[11] G Jackson. The scientist and the scales of justice. *Science & Justice*, 40:81–85, 2000.

[12] R Lempert. Some caveats concerning DNA as criminal identification evidence: with thanks to the Reverend Bayes. *Cardozo Law Review*, 13:303–341, 1991.

[13] D V Lindley. A problem in forensic science. *Biometrika*, 64:207–213, 1977.

[14] C Neumann and M Ausdemore. Defence against the modern arts: the curse of statistics: Part II: 'score-based likelihood ratios'. *Law, Probability and Risk*, 19:21–42, 2020. doi: https://doi.org/10.1093/lpr/mgaa006.

[15] B Robertson and G A Vignaux. *Interpreting Evidence. Evaluating Forensic Science in the Courtroom*. John Wiley & Sons, Chichester, 1995.

[16] V Scherz, C Bertelli, S Bozza, S Aeby, O Opota, L Falquet, F Taroni, and G Greub. When the defense says: "no, it is my twin brother": a salivary microbiota-based identification of monozygotic twins. Technical report, University of Lausanne, Institute of Microbiology, April 2021.

[17] B W Silverman. *Density estimation*. Chaptman & Hall, London, 1986.

[18] SM Willis, L McKenna, S McDermott, G O'Donell, A Barrett, B Rasmusson, A Nordgaard, CEH Berger, MJ Sjerps, JJ Lucena-Molina, G Zadora, CCG Aitken, T Lovelock, L Lunt, C Champod, A Biedermann, TN Hicks, and F Taroni. ENFSI guideline for evaluative reporting in forensic science, Strengthening the evaluation of forensic results across Europe (STEOFRAE). Dublin, 2015.