Forensic Science International: Synergy 2 (2020) 35–40

Contents lists available at ScienceDirect

Forensic Science International: Synergy

journal homepage: https://www.journals.elsevier.com/forensic-science-international-synergy/

Must the random man be unrelated? A lingering misconception in forensic genetics

Emmanuel Milot [a, b, *], Simon Baechler [a, c, d], Frank Crispino [a, b]

[a] Laboratoire de recherche en criminalistique, Department of Chemistry, Biochemistry and Physics, Université du Québec à Trois-Rivières, Trois-Rivières, Québec, G9A 5H7, Canada
[b] Centre interuniversitaire de criminologie comparée, Université du Québec à Trois-Rivières, Trois-Rivières, Québec, G9A 5H7, Canada
[c] Ecole des Sciences Criminelles, University of Lausanne, UNIL-Batochime, 1015, Lausanne, Switzerland
[d] Service forensique, Police neuchâteloise, Rue des Poudrières 14, 2002, Neuchâtel, Switzerland

ARTICLE INFO

Article history:
Received 30 June 2019
Received in revised form
17 October 2019
Accepted 11 November 2019
Available online 12 November 2019

Keywords:
DNA evidence
Fact-finder
Match probability
Relatedness
Semantics

ABSTRACT

A nearly universal practice among forensic DNA scientists includes mentioning an unrelated person as the possible alternative source of a DNA stain, when one in fact refers to an *unknown* person. Hence, experts typically express their conclusions with statements like: "The probability of the DNA evidence is *X* times higher if the suspect is the source of the trace than if another person *unrelated* to the suspect is the source of the trace." Published forensic guidelines encourage such allusions to the unrelated person. However, as the authors show here, rational reasoning and population genetic principles do not require the conditioning of the evidential value on the unrelatedness between the unknown individual and the person of interest (e.g., a suspect). Surprisingly, this important semantic issue has been overlooked for decades, despite its potential to mislead the interpretation of DNA evidence by criminal justice system stakeholders.

1. Introduction

Forensic science has been the target of severe critiques, in particular through the reports of the National Research Council in 2009 [1] and the President's Council of Advisors on Science and Technology in 2016 in the USA [2]. DNA typing was relatively spared by that storm, largely due to its strong grounding in probabilistic models to assess the weight of evidence. Nevertheless, the rendering of the weight of DNA evidence may mask fundamental interpretation issues for fact-finders, where semantics and communication are of prime importance. As highlighted by a growing body of research [3–10], communication between scientists and non-scientists is far from straightforward and may cause unconscious misunderstandings. Each word is important and the burden is on forensic scientists to convey their message in an accurate, transparent, readable and efficient way. Many debates between law and forensic science experts[1] underline the semantic issues and call to set up solutions for a clear communication that removes any ambiguity, a sort of common language between science and justice.

One semantic issue that has lingered ever since the introduction of trace DNA analyses in criminal investigations pertains to a very widespread practice: the concept of the 'unrelated person'. Experts typically express their conclusions about the weight of DNA evidence with statements like: "The probability of the evidence is *X* times higher under the hypothesis that the suspect is the source of the trace than under the hypothesis that another person unrelated to the suspect is the source of the trace." The word 'unrelated' has spreaded across the forensic literature since its tentative appearance in Jeffreys et al.'s initial paper on DNA fingerprints [11]. Nowadays, the word is almost always present in expert reports, scientific papers, textbooks and, importantly, forensic guidelines

* Corresponding author. Laboratoire de recherche en criminalistique, Department of Chemistry, Biochemistry and Physics, Université du Québec à Trois-Rivières, Trois-Rivières, Québec, G9A 5H7, Canada.
E-mail addresses: emmanuel.milot@uqtr.ca (E. Milot), simon.baechler@unil.ch (S. Baechler), frank.crispino@uqtr.ca (F. Crispino).

[1] Such as for instance the interdisciplinary symposium held during the 69th American Academy of Forensic Sciences conference in New Orleans (USA) in 2017. In this symposium, presentations and a discussion panel bringing together judges, prosecutors, forensic scientists and academics concluded that forensic scientists must improve in expressing clearfully their results in reports and court hearings, in particular the wording of competing hypotheses and what they encompass must be transparent for all stakeholders and dispel any blur whether conscious or not.

https://doi.org/10.1016/j.fsisyn.2019.11.003

and recommendations. For instance, in the *ENFSI Guideline for Evaluative Reporting in Forensic Science*, DNA case examples mention alternative propositions considering "an (unknown) unrelated person" [[12], pp. 34, 40]. Likewise, in its latest recommendations the DNA commission of the International Society of Forensic Genetics mentions "it is standard to apply the 'unrelated' caveat" (see footnote 6 in Ref. [13]).

While there is an abundant literature about the problem of how to deal with relatives in forensic genetics, curiously we found no published reference that fundamentally addresses the interpretation of the concept of 'unrelatedness'. This issue is semantic in nature and does not challenge the validity of the mathematical models that are applied to assign the probability of DNA evidence in everyday casework. However, we are concerned about the confusion that the routine and default usage of the word 'unrelated' can cause among an audience of investigators, lawyers, prosecutors or fact finders over the correct meaning of calculations pertaining to DNA evidence.

2. Confusion over the 'unrelated'

All individuals have relatives. This is a consequence of the finite size ($N < \infty$) of populations. Thus, suspects have relatives too. The more genes they share with them, the more challenging it may be to make conclusive inferences about the source of DNA traces. This explains why forensic experts tend to specify that the reported weight of evidence holds only if the source of the trace is unrelated to the suspect or, equivalently, that the suspect's relatives are excluded from the pool of individuals that may be randomly drawn from the population of interest. However, since an individual is always related to any other member of the population – whether their most recent common ancestor lived one generation or thousands of years ago – conditioning on unrelatedness implies that the weight of evidence strictly applies to a non-existent fraction of the population. No doubt that forensic scientists have a more practical definition in mind when they use the word 'unrelated', such as "not closely related to the suspect" or "not related to a degree close enough to bias substantially the calculation of the weight of evidence". Yet, such fuzzy definitions can be misleading.

First, referring to a person unrelated to the suspect may be perceived as if the population of interest excluded (close) relatives, in a sense a form of covert exoneration.[2] This is because, in such a case, the set of people encompassed by the prosecution and the defence hypotheses excludes relatives, which may give the impression that both sides do not consider them as relevant. Second, one may think that relatives compromise the value of evidence. For instance, as suggested by a reporting scientist with whom we discussed the issue, one may wonder if the use of the word 'unrelated' in the alternative proposition means that if the suspect has a brother, the weight of evidence is meaningless and the DNA evidence useless. Third, non-geneticists may think that two persons that do not fall under a usual "close relationship" category are necessarily more genetically distant than close relatives. Take the example of first cousins. Their kinship coefficient[3] (φ) is 0.0625. However, there are a plethora of pedigree relationships that can lead to the exact same kinship level when two persons share several but more remote ancestors, especially in endogamous populations.

Moreover, forensic biologists themselves do not seem to agree on the correct interpretation of 'unrelated'. The issue arouse independently to authors of this paper in different contexts in Europe and North America, demonstrating similar concerns about the word 'unrelated' shared by practitioners and researchers in various countries. For example, in a 2012 international workshop on forensic DNA, one of us suggested that the word 'unrelated' should not be used anymore in expert reports. The discussion that followed among reporting scientists showed that they diverge over the interpretation and implications of this term. The issue was also brought forward in 2017 within a Swiss working group dedicated to interpreting forensic evidence and expressing conclusions. Despite admitting discomfort when asked to justify the default use of the word 'unrelated', the members decided to keep using it until the scientific literature addresses the question because, if questioned, they must refer to "the scientific state of knowledge".

Furthermore, as applied in forensic science the concept of unrelatedness appears to be an incorrect interpretation of population genetic principles. Essentially, the problem arises when the *absence of knowledge about the relatives* of a person of interest leads the scientist to transform the 'unknown person' (the classical 'random man') into an 'unrelated person' upon reporting a random match probability, a likelihood ratio, or any other quantitative assessment of the DNA evidence. However, the key point for the correct interpretation of the weight of DNA evidence is not the existence of relatives *per se* but rather the information that one has or not about them and about their potential involvement in the case at hand. As we show in the next section, when no information about relatives is available, one should not interpret Hardy-Weinberg (HW) equations, or their derivations (e.g., those incorporating some form of coancestry), as conditioning the weight of evidence on the unrelatedness between the person of interest (e.g., suspect) and the unknown source of the trace.

3. All is relative

Consider two competing hypotheses about the source of a trace, H_p and H_d, respectively proposed by the prosecution and the defence [14]. In a Bayesian framework, the strength of our belief in favour of one hypothesis over the other before observing the DNA evidence (i.e. the ratio of their prior odds), is given by $\Pr(H_p|I)/\Pr(H_d|I)$, where I is any other relevant (e.g., circumstantial) information available about/for the casework. After observing the DNA evidence (E), the posterior odds become

$$\frac{\Pr(H_p|E,I)}{\Pr(H_d|E,I)} = \frac{\Pr(E|H_p,I)}{\Pr(E|H_d,I)} \times \frac{\Pr(H_p|I)}{\Pr(H_d|I)} \quad (1)$$

where $\Pr(E|H_p,I)/\Pr(E|H_d,I)$ is the likelihood ratio (LR). In equation (1), case information available about relatives is a component of I and we will designate it by I_R. A classical example is when the suspect has a brother who is assumed to belong to the population of interest. In such a case, H_p usually remains unchanged (e.g. "the suspect is the source of the trace") while H_d could be that "his

[2] Indeed, background case information is most of the time insufficient or unavailable to assume such exclusion. This is not the role of the forensic scientist alone, who is left in most cases with a great deal of uncertainty about the relatedness factor. It may be tempting to reduce uncertainty by gathering circumstantial information about existing relatives through further investigations, by querying administrations or by asking directly the suspect. However, such information can rarely be considered as fully reliable and comprehensive. For instance, the suspect could state in interviews that he has brothers when in fact he has none. Administrative registers, when they exist, may be incomplete in particular about foreigners, and they provide official family relationships that do not always reflect biological relationships (e.g., illegitimate children) and certainly do not cover the full range of close to remote relationships. Finally, putting aside the impact on efficiency and timeliness for the case in process, one may also claim against bias of the forensic scientist's interpretation when gathering further circumstantial information.

[3] The kinship coefficient is defined as the probability of randomly drawing from different individuals two alleles that are identical-by-descent (IBD), i.e. due to common ancestry (e.g., $\varphi = 0.25$ for brothers).

brother is the source of the trace", or that "another person than the suspect, not excluding his brother, is the source". In either case the calculation of the LR denominator must be adjusted appropriately [15]. Therefore, changing I_R can modify or refine both the set of hypotheses to be evaluated and the calculation of the LR, in agreement with these hypotheses [16].[4] Now, since this is true for any defence hypothesis admitting any specified relatives as the potential donor of the DNA stain [15,18], we will not limit our consideration to the sole brother case and refer more generally to the kinship coefficient φ, which has a value for every degree of genetic relationship (see footnote 3).

When the reporting scientist has no knowledge about the existence of relatives, then $I_R = \varnothing$ (empty set). In this case, it is generally assumed that the calculation of $\Pr(E|H_d, I)$ in equation (1), which is based on Hardy-Weinberg law in the simplest model, holds only when the donor is *unrelated* to the suspect, that is $\varphi = 0$. However, this has no resonance for stakeholders of the justice system that have to deal with the real world, where crimes occur in populations composed of many kinds of relatives. Actually, the only thing that the denominator should entail is that the reporting scientist incorporates no relevant information about the kinship of the suspect to other persons in the population. That is, $I_R = \varnothing$ does not imply that potential donors are totally unrelated to the suspect (i.e. that $\Pr(\varphi > 0) = 0$). Strictly speaking, an absence of kinship between individuals is expected only in infinite size populations since $\Pr(\varphi > 0) \to 0$ when $N \to \infty$ under random mating [19] (see also Appendix A). Consequently, the absence of information about relatives should not be equated to an absence of kinship.

The use of the word 'unrelated' is even more problematic under the Balding-Nichols (BN) model [20], which is routinely applied by forensic labs in place of the HW model. This model postulates that relatedness does exist between the suspect and the source of the trace due to population subdivision, such that individuals from the same subpopulation share a common ancestry (and assuming that the suspect and the donor belong to the same subpopulation). The theta (θ) parameter of this model corrects for the non-independence of their genetic profiles by incorporating information from studies on the genetic structuring of human populations. Obviously, this means that the kinship between the suspect and other persons from the same subpopulation is greater than zero. Consequently, it is incoherent to use the word 'unrelated' in the formulation of the weight of evidence based on this model.

Moreover, contrary to a widespread idea, HW or BN equations do provide correct values for the probability of a genetic profile when one admits the inclusion of the suspect's relatives in the population of interest, as long as no information about these relatives is available, as demonstrated for the HW case in Appendix A. Hence, the standard random match probability must be understood as the match probability in the absence of knowledge about relatives, rather than in its common acceptance as the "match probability when relatives are excluded from the population of interest", which is equivalent to make the unreasonable assumption that the suspect has no relatives! A similar reasoning applies to other weight-of-evidence metrics that tend to refer to unrelated persons in their verbal formulations, including likelihood ratios of various degrees of sophistication.

4. The unrelated man: an unnecessary burden

To circumvent the lack of realism conveyed by references to unrelated individuals, some authors proposed to change the calculation and presentation of the weight of evidence. In their "call for a re-examination of reporting practice", Buckleton and Triggs [16] concluded that "it is time that the match probabilities for a sibling are reported in all casework involving many loci where the suspect has a non-excluded sibling" – a call that however appears to have had little effect on current common DNA reporting practice (see Ref. [21] for a similar argument). Likewise, Taylor *et al.* [22] proposed a unified LR that accounts for potential relatives and "removes the need to stipulate that the alternative donor is unrelated when forming the propositions" [22, p. 57]. Basically, LRs considering different types of relatives are calculated, weighted by the postulated frequency of each type of relative, and then summed up [22]. The STRmix™ software implements a different approach by letting the user specify the average number of children per family, to better reflect the composition of the population of interest (see http://strmix.esr.cri.nz/#home for a list of publications relative to the methods implemented in STRmix™). As far as the assumptions about the relatedness structure are made explicit, above approaches have the advantage of considering populations that are more realistic of human mating systems than the classic 'random mating' scheme. However, while they address the problem of how to best quantify the weight of DNA evidence, they do not fully address the semantic issue of its verbal formulation, because an 'unrelated' category may still remain among the several types of relatives considered. What should reporting scientists do then?

We suggest referring simply to an 'unknown person' or to the 'random individual' is sufficient because one should not, and does not need to, discard the possibility that the source is related to the suspect to an unknown degree. Alternatively, a more explicit wording would be 'an unknown person, without regard to his relatedness to the suspect'. Again, the important point here is not unrelatedness but the absence of relevant knowledge about relatives ($I_R = \varnothing$), which prevails in most real life casework. From this perspective, one can simply consider that if the unknown individual who left a DNA trace happened to be the brother or the cousin of the suspect, this would be a sort of ancillary consequence, a way by which we categorize and name one among many possible genetic outcomes of a random draw in a finite population. This way of expressing the relatedness avoids the pitfalls associated with the choice of an arbitrary definition of 'unrelated' within the forensic context. Critically, in assessing the weight of the DNA evidence with standard metrics, one must nevertheless bear in mind the assumption that the suspect has no more or less chance to have relatives of a given degree than the average person in the population of interest. Therefore, it is still important to specify that potential relatives are included in the list of possible donors, especially when the set of possible suspects is small.[5]

5. Prospective

The arguments presented in this paper call for a change in reporting practices to prevent semantic confusion and potential misinterpretation of DNA evidence by fact-finders and other criminal justice system participants. We suggest avoiding the routine and default use of the word 'unrelated', not only in oral communications and expert reports, but also in the forensic literature in general, including guidelines and recommendations. Some

[4] Concerning the assessment of multiple hypotheses in the LR calculation, we refer the reader to Buckleton *et al.* [17].

[5] In particular when the population of interest is exceptionally limited by circumstances – such as a crime committed in a prison or on a boat, or when the frame of suspects is restricted to a small town, a remote area or a family. In such crimes committed "behind closed doors", it appears reasonable for the forensic scientist to suggest to pursue the investigations in order to determine which relative (if any) is actually included in the limited suspect population, and unless they have a verified alibi, to test them for DNA and eventually exclude them as a possible source.

might believe that this issue is unlikely to have a big influence on the interpretation of forensic DNA expertises, but the confusion that exists even among reporting scientists (see section 2) casts doubt on such an assumption. For clarity and robust reporting practices in forensic genetics, there is a vacuum in the literature about this question that needs to be filled. Thus, we hope this paper will spark discussion, and will be glad to hear what other people think, including scientists, investigators, prosecutors, lawyers and fact-finders supporting or mitigating our concern (a web page [www.uqtr.ca/lrc/unrelated] has been opened to gather comments from the readers).

In all cases, future studies in criminology, psychology and law will be essential to better document the variation in the perception, both by scientists and non-scientists, of the unrelatedness concept, and the impact of this variation on the justice system. The perception of alternative formulations should be compared, such as the one proposed here ('an unknown person, without regard to his relatedness to the suspect'). This calls for an active collaboration between scientists and stakeholders of the criminal justice system to reduce the gap "that exists between questions lawyers are actually interested in, and the answers that scientists deliver to Courts" [23]. Finally, while this paper focuses on evaluative reporting, it will also be important to assess if and how various interpretations of the unrelatedness concept could impact decisions and action in the course of criminal investigations.

Conflict of Interest

The authors declare no conflict of interest.

Declaration of competing interest

We declare that this research involves no competing interests.

Acknowledgement

We wish to thank Franco Taroni, Tacha Hicks Champod, Alex Biedermann, John Buckleton, Ian Evett and two anonymous reviewers for providing comments that helped to improve the manuscript. Note that the opinions expressed in this paper are ours and may or may not be shared by people who commented on the manuscript.

APPENDIX A

Hardy-Weinberg equations, or their derivations (e.g., those incorporating some form of coancestry), are routinely applied to quantify the weight of DNA evidence. This appendix demonstrates why these equations provide correct probabilities of genetic profiles when one admits the possible inclusion of the suspect's relatives. Strictly speaking, the Hardy-Weinberg law holds when there is no random genetic drift, which is the case when the population size N tends toward infinity. Standard forensic calculations assume that N is large enough to make negligible any bias caused by the fact that a real population is actually not of infinite size. From this perspective, it might seem logical to consider that calculations based on HW equations admit only 'unrelated' individuals in the population of interest, since the average kinship of two persons in a given offspring generation tends toward 0 as $N \rightarrow \infty$. However, this has no resonance for stakeholders of the justice system that have to deal with the real world (see section 2 of the main text). The perspective adopted here is different. We consider that in a finite population, Hardy-Weinberg equations provide the probability, averaged over all possible relatedness degrees, to randomly draw a given genotype. Taking the brother case as an example, we illustrate that there is no difference between the values given by the standard random match probability (RMP), i.e. the one typically associated with 'unrelated man'-type verbalizations, and a RMP that explicitly incorporates the possibility that the unknown is a brother of the suspect, but assuming that the scientist has no knowledge about what relatives the suspect may have ($I_R = \varnothing$, where I_R denotes circumstantial information pertaining to the suspect's relatives).

Admitting relatives does not bias genotype frequencies

Since the estimation of populational genotype frequencies is central in the assessment of the weight of DNA evidence, especially when the defence hypothesis involves unknown individuals, it is important to first explain why admitting the existence of a suspect's relatives in the population of interest does invalidate genotype frequencies estimated from allele counts in reference samples used by forensic labs. When the population is of finite size, as in real life, it will occur that two gametes will be drawn from the same reproducing individual, with a probability that is inversely proportional to the population size (all else being equals).[6] When these two gametes carry the same parental gene copy, this will generate identity-by-descent (IBD) alleles carried by different offspring. It will also occur that gametes are drawn from individuals that are related because they share IBD alleles due to reproduction in previous generations. The current generation is thus composed of individuals related to diverse degrees as a result from the genealogical structure that has developed over time. In all logic, allele frequencies estimated from allelic counts in a reference sample for a population, denoted here by the vector $\mathbf{P}_{\text{ref}}$, must be coherent with this structure because the true (but unknown) allele frequencies ($\mathbf{P}_{\text{pop}}$) necessarily reflect the existence of all these relatives. In other words, the underlying assumption made by forensic labs is that $E(\mathbf{P}_{\text{ref}})$ equates $\mathbf{P}_{\text{pop}}$, where $E(.)$ denotes the expectation. Consequently, it is incoherent to consider, on one hand, that $\mathbf{P}_{\text{ref}}$ constitutes a valid approximation of $\mathbf{P}_{\text{pop}}$, and, on the other hand, that genotype frequencies estimated from $\mathbf{P}_{\text{ref}}$ reflect only a pool of unrelated individuals, i.e. a non-existent fraction of the population.[7] This comes down to the issue of what is the basal population. As underscored by Lynch and Walsh [24], "Technically speaking, all members of a species or population are related to each other to some degree for the simple reason that they contain copies of genes that were present in some remote ancestor in the phylogeny. We avoid this problem by letting the reference population be the base of an observed pedigree". While these authors raised this issue within the context of quantitative genetics, the reasoning remains true for the problem addressed here.

[6] Note that under the random mating model one expects many more half sibs than full sibs in a population. While this is generally unrealistic for human populations, it is nevertheless the model underlying 'random man' type calculations for finite populations.

[7] A key point to consider here is the following: under random mating, when assessing a genotype probability, it is irrelevant to consider whether or not the two gametes drawn from the parental generation to form the zygote were previously drawn from the same parents to create other offspring. In other words, the simple fact of having a brother does not influence the probability of drawing randomly one's genotype from the same parental population. From a forensic perspective, this implies that *when knowledge about the brother is not available*, then the genotype probability is solely based on postulated allele frequencies. The reasoning holds for more remote degrees of relationships than brothers, such as cousins. When gametes are drawn randomly to create a new generation, the major parameters are the frequency of alleles in the parental generation and the mating system. Whether some of the parental alleles are IBD (implying related individuals) or simply identical-in-state (IIS) due to recurrent mutations is irrelevant.

Admitting relatives does not bias the match probability

To illustrate this point, we will consider the case where the suspect *may* have a brother. For the sake of simplicity we assume again a random mating population with no subdivision (i.e. HW model) although the reasoning holds under the Balding-Nichols model [20]. First, let's postulate that the suspect *has* a brother who is member of the population of interest, an event that we denote by B. From equation (1) in the main text, this postulate amounts to consider that $B \in I_R$. Then including the possibility for the brother in the defence hypothesis (H_d) and conditioning the probability of the trace DNA profile (E) on B makes sense because B may be informative of $\Pr(E)$[8]:

$$\Pr(E|H_d, I_R, I_O) = \Pr(E|H_d, B \in I_R, I_R,\ I_O)$$

Here I_O refers to any other circumstantial information not pertaining to the suspect's relatives (i.e. $I = I_R \cup I_O$). If, instead, we postulate that the suspect *has no* brother, an event denoted $\overline{B}$, then

$$\Pr(E|H_d, I_R, I_O) = \Pr(E|H_d, \overline{B}, \overline{B} \in I_R, I_R,\ I_O).$$

Now, consider the case where the suspect *may have* a brother but that we have no information about whether he does. That is $I_R = \varnothing$, which *a priori* assumes that the suspect is not more or less likely to have a brother than the average individual in the population. In such a case, H_d would refer to an 'unknown person' and can be expressed as the sum of the probabilities of the trace under both possibilities that the suspect has and does not have a brother:

$$\begin{aligned}\Pr(E|H_d, I_R, I_O) = {} & \Pr(E|H_d, B, B \notin I_R, I_R = \varnothing,\ I_O)\Pr(B) \\ & + \Pr(E|H_d, \overline{B}, \overline{B} \notin I_R, I_R = \varnothing,\ I_O)\Pr(\overline{B}) \qquad \text{(A.1)}\end{aligned}$$

In the absence of knowledge about a suspect's relative, recognizing the possibility that he may have a brother ($\Pr(B) > 0$) does not invalidate the use of HW equations to quantify the probability that an unknown man is the source of the trace. To demonstrate this, we must consider three possibilities of a match between the suspect's and the trace DNA profiles, under the defence hypothesis. Thus either:

1. The suspect has a brother who carries the same genotype as him, and the brother is the unknown individual who left the DNA trace;
2. The suspect has a brother but another unknown individual carrying the same genotype as the suspect left the DNA trace;
3. The suspect has no brother and an unknown individual carrying the same genotype as the suspect left the DNA trace.

Summing up probabilities for these three events recovers the genotype probability expected under HW, at least when assuming the typical hypergeometric distribution of genotype frequencies ([25]; see next section). In other words, the brother could be the unknown man who left the DNA trace. This would not bias the calculation because this hypothesis is not explicitly evaluated with HW equations (and assuming that the reporting scientist doesn't know about his existence or non-existence).

Random match probability

Let G_{UK} be the genetic profile of the unknown who left the DNA trace (under the defence hypothesis) and G_S that of the suspect. For convenience, we can equate G_{UK} with the random match probability (RMP) since the observation of the first copy of the genotype does not change the probability of observing the second copy under the Hardy-Weinberg model. To assess the impact of admitting that an unknown brother of a suspect could be the person who left the DNA trace, we need to consider the sampling of genotypes in a finite size population, assuming that the probability that the suspect has a brother is the same as that for the average person in the population. For commodity and without loss of generality, we consider that the probability that the unknown (UK) is a brother (or full sib (FS)) of the suspect is equivalent to the probability of randomly drawing two gametes from the parental population, one from each of the suspect's parent[9]:

$\Pr(\text{UK=FS}) = \Pr(1$ gamete is from the suspect's mother $\cap$ 1 gamete is from suspect's father).

Under random mating in a population of finite size N:

$$\Pr(\text{UK} = \text{FS}) \sim h(k = 2, N, K = 2, n = 2)$$

where $h(.)$ denotes the hypergeometric distribution, k is the number of success (i.e. drawing a gamete from a suspect's parent), K is the number of parents of the suspect and n is the number of draws. Another outcome possible is that only one gamete is drawn from a suspect's parent and the other allele from another individual, giving a half sib (HS) of the suspect. Finally, the last possible outcome is that none of the two gametes come from the suspect's parents, giving a "non-sib" (NS). The probability of these two outcomes can also be calculated from the hypergeometric distribution and

$$\Pr(\text{UK} = \text{FS}) + \Pr(\text{UK} = \text{HS}) + \Pr(\text{UK} = \text{NS}) = 1$$

As an example, let's suppose that the suspect has the heterozygous profile a/b. We need to evaluate the following expression:

$$\begin{aligned}\text{RMP} \;=\; \Pr(G_{UK} = G_S = a/b) = {} & \Pr(G_{UK} = a/b|\text{UK} = \text{FS}, G_S = a/b)\Pr(\text{UK} = \text{FS}) \\ & + \Pr(G_{UK} = a/b|\text{UK} = \text{HS}, G_S = a/b)\Pr(\text{UK} = \text{HS}) \\ & + \Pr(G_{UK} = a/b|\text{UK} = \text{NS}, G_S = a/b)\Pr(\text{UK} = \text{NS})\end{aligned}$$

We considered two different models and performed RMP calculations independently under each of these models.

- **Model 1 = Fixed allele frequencies:** the postulated (reference; $\mathbf{P}_{\text{ref}}$) allele frequencies p_a and p_b for the population of size N are fixed. That is, if $p_a = 0.1$ and $N = 10{,}000$, there are exactly 0.1×2 x $10{,}000 = 2000$ copies of allele a in the population. In such a

[8] This is particularly true under the BN model, where knowledge of any genotype from the same subpopulation update the information about allele frequencies for that subpopulation. For the HW model, the brother's genotype is informative of $\Pr(E)$ only if the brother is suspected more strongly than other members of the population of interest.

[9] Reminder: Once again this is justified by the fact that we no know nothing about the suspect having a brother ($I_R = \varnothing$). In the opposite case, usual identity-by-descent calculation would apply.

population, the probability of a a/b heterozygote will be slightly upwardly biased relative to that in an infinite size population: $2p_a\,(2\,N^{*}p_b)/(2N\text{-}1) > 2p_ap_b$.

- **Model 2 = Random allele frequencies:** allele counts in the finite population are a random draw of 2 N alleles based on the postulated (reference; $\mathbf{P}_{\text{ref}}$) allele frequencies. In other words, the population of size N behaves as a random sample (one possible realization) from a very large (infinite) population having the postulated allele frequencies. In such a population of size N, the probability of the a/b heterozygote is slightly biased downwardly due to the negative covariance of allele counts: $E(2p_ap_b) = 2p_ap_b - p_ap_b/N$ [25].

Given the suspect's genotype, the possible genotypic states for his parents are limited to those that can give birth to an a/b offspring (e.g. mother a/a – father b/b, mother a/x – father b/x, where x is any allele different from a and b). Thus, the approach used here is to evaluate $\Pr(G_{UK} = a/b|\text{UK} = \text{FS})$, $\Pr(G_{UK} = a/b|\text{UK} = \text{HS})$ and $\Pr(G_{UK} = a/b|\text{UK} = \text{NS})$ by considering each of possible suspect's parent pair, weighted by its probability. This way of calculating the RMP admitting siblings will be denoted RMP_{sib} herein.

Table A1 provides examples of the values obtained for the RMP as calculated using standard HW equations (RMP_{std}) compared to those calculated using the approach described here (RMP_{sib}). Note that we are not suggesting that the latter should be used to assess the weight of evidence in everyday caseworks. Again, we use it to illustrate that there is no difference between the values given by RMP_{std} and a RMP that explicitly incorporates the possibility that the unknown is a sibling of the suspect. Indeed, Table A1 shows that the RMP_{sib} is generally equal to RMP_{std} for a given set of parameters N, p_a and p_b. The reader will note that RMP_{sib} tends to overestimate very slightly RMP_{std}, but the difference is negligible even for very small populations of interest. For instance, in the worse case shown in Table A1 (i.e. under model 2, when $N = 100$, $p_a = 0.5$ and $p_b = 0.1$), $\text{RMP}_{\text{sib}}/\text{RMP}_{\text{std}} = 1.000613$ instead of 1. This is due to the effect of the knowledge of the suspect's genotype on the RMP for a finite population (which is a different issue than the one addressed here). This effect arises from the negative covariance of genotypic counts, and increases with decreasing N. In other words, the observation of G_S update our knowledge of realized genotype frequencies in the population due to the constraint that allele frequencies are either fixed (model 1) or randomly drawed from population having the postulated (reference; $\mathbf{P}_{\text{ref}}$) allele frequencies. Therefore, observing $G_S = a/b$ implies that one of the $2Np_a$ copies of allele a, and one of the $2Np_b$ copies of allele b, in the population, are found together in the suspect, meaning that other genotypes existing in the population must be made from the remaining $2Np_a - 1$ and $2Np_b - 1$ copies, limiting possible values for genotype counts in an increasing manner with decreasing N (independently of the suspect's relatives issue).

Table A.1
Values obtained for the standard random match probability (RMP_{std}) and the random match probability accounting for the possibility that suspect's siblings may exist in the population (RMP_{sib}), for a heterozygote a/b and various settings of N, p_a and p_b (assuming no coancestry due to population subdivision, i.e. $\theta = 0$). RMP_{std} for model 2 integrates the expected difference in the genotype frequencies in a finite population ($2p_ap_b - p_ap_b/N$) that is a random draw from an infinite population ($2p_ap_b$) [25].

p_a	p_b	N	Model 1: fixed allele frequencies		Model 2: random allele frequencies	
			RMP_{std}	RMP_{sib}	RMP_{std}	RMP_{sib}
0.1	0.1	∞	0.02000000	0.02000000	0.02000000	0.02000000
		1,000,000	0.02000001	0.02000001	0.01999999	0.01999999
		10,000	0.02000100	0.02000100	0.01999900	0.01999900
		1000	0.02001001	0.02001001	0.01999000	0.01999020
		100	0.02010050	0.02010071	0.01990000	0.01992015
0.5	0.1	∞	0.10000000	0.10000000	0.10000000	0.10000000
		1,000,000	0.10000010	0.10000010	0.09999995	0.09999995
		10,000	0.10000500	0.10000500	0.09999500	0.09999501
		1000	0.10005000	0.10005000	0.09995000	0.09995006
		100	0.10050250	0.10050260	0.09950000	0.09956099

References

[1] National Research Council, Strengthening Forensic Science in the United States: a Path Forward, Washington D.C, 2009.
[2] President's Council of Advisors on Science and Technology, Forensic Science in Criminal Courts: Ensuring Sceintific Validity of Feature-Comparison Methods, 2016.
[3] E. Arscott, R. Morgan, G. Meakin, J. French, Understanding forensic expert evaluative evidence: a study of the perception of verbal expressions of the strength of evidence, Sci. Justice 57 (2017) 221–227.
[4] L.M. Howes, The communication of forensic science in the criminal justice system: a review of theory and proposed directions for research, Sci. Justice 55 (2015) 145–154.
[5] L.M. Howes, K.P. Kirkbride, S.F. Kelty, R. Julian, N. Kemp, Forensic scientists' conclusions: how readable are they for non-scientist report-users? Forensic Sci. Int. 231 (2013) 102–112.
[6] C. Kruse, The Bayesian approach to forensic evidence: evaluating, communicating, and distributing responsibility, Soc. Stud. Sci. 43 (2013) 657–680.
[7] K.A. Martire, R.I. Kemp, B.R. Newell, The psychology of interpreting expert evaluative opinions, Aust. J. Forensic Sci. 45 (2013) 305–314.
[8] K.A. Martire, R.I. Kemp, M.A. Sayle, B.R. Newell, On the interpretation of likelihood ratios in forensic science evidence: presentation formats and the weak evidence effect, Forensic Sci. Int. 240 (2014) 61–68.
[9] K.A. Martire, R.I. Kemp, I. Watkins, M.A. Sayle, B.R. Newell, The expression and interpretation of uncertain forensic science evidence: verbal equivalence, evidence strength, and the weak evidence effect, Law Hum. Behav. 37 (2012) 197–207.
[10] C. Mullen, D. Spence, L. Moxey, A. Jamieson, Perception problems of the verbal scale, Sci. Justice 54 (2014) 154–158.
[11] A.J. Jeffreys, V. Wilson, S.L. Thein, Individual-specic 'fingerprint' of human DNA, Nature 316 (1985) 76–79.
[12] ENFSI, ENFSI Guideline for Evaluative Reporting in Forensic Science, 2010.
[13] P. Gill, T. Hicks, J.M. Butler, E. Connolly, L. Gusmão, B. Kokshoorn, N. Morling, O. Van, W. Parson, M. Prinz, P.M. Schneider, T. Sijen, D. Taylor, DNA commission of the ISFG: 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.
[14] I.W. Evett, B.S. Weir, Interpreting DNA Evidence: Statistical Genetics for Forensic Scientists, Sinaur Associates, Sunderland, 1998.
[15] I.W. Evett, Evaluating DNA Profiles in a case where the defence is "it was my brother", J. Forensic Sci. Soc. 32 (1992) 5–14.
[16] J.S. Buckleton, C.M. Triggs, Relatedness and DNA: are we taking it seriously enough? Forensic Sci. Int. 152 (2005) 115–119.
[17] J.S. Buckleton, C.M. Triggs, C. Champod, An extended likelihood ratio framework for interpreting evidence, Sci. Justice 46 (2006) 69–78.
[18] J.A. Bright, J.M. Curran, J.S. Buckleton, Relatedness calculations for linked loci incorporating subpopulation effects, Forensic Sci. Int. Genet. 7 (2013) 380–383.
[19] D.L. Hartl, A.G. Clark, Principles of Population Genetics, fourth ed., Sinauer, Sunderland, 2007.
[20] D.J. Balding, R.A. Nichols, DNA profile match probability calculation: how to allow for population stratification, relatedness, database selection and single bands, Forensic Sci. Int. 64 (1994) 125–140.
[21] T. Tvedebrink, P.S. Eriksen, J.M. Curran, H.S. Mogensen, N. Morling, Analysis of matches and partial-matches in a Danish STR data set, Forensic Sci. Int. Genet. 6 (2012) 387–392.
[22] D. Taylor, J.A. Bright, J.S. Buckleton, J.M. Curran, An illustration of the effect of various sources of uncertainty on DNA likelihood ratio calculations, Forensic Sci. Int. Genet. 11 (2014) 56–63.
[23] F. Taroni, A. Biedermann, J. Vuille, N. Morling, Whose DNA is this? How relevant a question? (a note for forensic scientists), Forensic Sci. Int. Genet. 7 (2013) 467–470.
[24] M. Lynch, B. Walsh, Genetics and Analysis of Quantitative Traits, Sinauer, Sunderland, Massachuset, 1998.
[25] B.S. Weir, Genetic Data Analysis II, Sinauer, Sunderland, 1996.