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

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.

Keywords: DNA evidence; fact-finder; match probability; relatedness; semantics

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

[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 panel discussion bringing together judges, prosecutors, forensic scientists and academics brought out that forensic scientists must improve in expressing clearfully their results in reports and in court hearings, in particular when it comes to competing hypotheses their wording and what they encompass must be transparent for all stakeholders and dispel any blur whether conscious or not.

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 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 who 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

[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.

more genetically distant than close relatives. Take the example of first cousins. Their kinship coefficient[3] (ϕ) is 0.0625. However, in theory 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.

Actually, forensic biologists 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".

Moreover, the concept of unrelatedness, as applied in forensic science, disagrees with 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/used, it is appropriate to apply standard equations based on the Hardy-Weinberg (HW) law without conditioning the weight of evidence on the unrelatedness between the person of interest (e.g. suspect) and the source of the trace.

[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., $\phi = 0.25$ for brothers).

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 brother is the source of the trace", or that "another person other 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 alternative source 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 source is

[4] Concerning the assessment of multiple hypotheses in the LR calculation, we refer the reader to Buckleton *et al.* [17] J.S. Buckleton, C.M. Triggs, C. Champod, An extended likelihood ratio framework for interpreting evidence, Science and Justice 46(2) (2006) 69-78.

unrelated to the suspect, that is $\phi = 0$ (with H_d defined accordingly). However, this is a mistake because the only thing that the calculation in the denominator entails 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 $\Pr(\phi > 0) = 0$, i.e. that individuals are totally unrelated. Strictly speaking, an absence of kinship between individuals is expected only in infinite populations since $\Pr(\phi > 0) \to 0$ when $N \to \infty$ under random mating [19]. Consequently, the absence of information cannot be equated to an absence of kinship since an individual is always related somehow to any other member of the finite population.

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. The theta (θ) parameter of this model corrects for the non-independence of their genetic profiles by incorporating information from population genetic studies. Obviously, this means that $I_R \neq \varnothing$ and that the probability of kinship between the suspect and other persons from the same subpopulation ($\Pr(\phi > 0)$) 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 common admittance, HW or BN equations do provide correct values for the probability of a genetic profile when one admits the possible 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 but wrong acceptance as the "match probability when no relatives exist". A similar reasoning applies to other forms of weight-of-evidence metrics that tend to refer to unrelated persons in their verbal formulations, including LRs of various degree of sophistication.

4. The unrelatedness concept: an unnecessary burden

To circumvent the lack of precision conveyed by reference 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 [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 (i.e., considering that $I_R \neq \varnothing$ at population level) [22]. The STRmix™ software lets the user specify the average number of children per family (i.e., $I_R \neq \varnothing$), 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 weigh of evidence, they do not address the semantic issue of their verbal formulation. Indeed, they do not totally eliminate the use of the word 'unrelated' because an 'unrelated' category may still remains among the several types of relatives considered. What should we do then?

First, we suggest to 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 (the source of the DNA trace under the defence hypothesis) 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. Second, referring simply to an "unknown person" or to the "random individual" is sufficient because one should not (and doesn't need to) discard the possibility that the source is related to the suspect to an

213 unknown degree. Alternatively, a more explicit wording would be “an unknown person, without
214 regard to his relatedness to the suspect”. Again, the important point here is not unrelatedness but
215 the absence of relevant knowledge about relatives ($I_R = \varnothing$), which prevails in most real life
216 caseworks. Critically, in assessing the weight of the DNA evidence with standard metrics, one
217 must nevertheless bear in mind the assumption that the suspect has no more or less chance to
218 have relatives of a given degree than the average person in the population of interest. Therefore,
219 it is still important to specify that potential relatives are included in the list of possible donors,
220 especially when the set of possible suspects is small.

221

222 **5. Conclusion**

223 The arguments presented in this paper call for a change in reporting practices to prevent semantic
224 confusion and potential misinterpretation of DNA evidence by fact-finders and other criminal
225 justice system participants. We suggest avoiding the routine and default use of the word
226 ‘unrelated’, not only in oral communications and expert reports, but also in the forensic literature
227 in general, including guidelines and recommendations. Some might believe that this issue is
228 unlikely to have a big influence on the interpretation of forensic DNA expertises. We doubt it is
229 the case, considering the confusion that exists even among reporting scientists (see section 2).
230 There is a vacuum in the literature about this question that needs to be filled. Thus, we hope this
231 paper will spark discussion, and will be glad to hear what other people think, including scientists,
232 investigators, prosecutors, lawyers and fact-finders supporting or mitigating our concern, whether
233 through formal or informal replies (we opened a web page [www.uqtr.ca/lrc/unrelated] to gather
234 comments from the readers).

235

236 In all cases, future studies in criminology, psychology and law will be essential to better
237 document the variation in the perception, both by scientists and non-scientists, of the ‘random
238 individual’ and unrelatedness concepts, and the impact of this variation on the justice system. The
239 perception of alternative formulations should be compared, such as the one proposed here (“an
240 unknown person, without regard to his relatedness to the suspect”). This calls for an active
241 collaboration between scientists and stakeholders of the criminal justice system to reduce the gap
242 “that exists between questions lawyers are actually interested in, and the answers that scientists

deliver to Courts" [23]. Finally, while this paper focuses on the evaluative phase, it will also be important to assess how various interpretations of the unrelatedness concept could impact decisions and action during the investigative phase of a criminal casework.

6. References

[1] N.R. Council, Strengthening forensic science in the United States: a path forward, Washington D.C., 2009.
[2] P.s.C.o.A.o.S.a. 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, Science and 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, Science and 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 science international 231 (2013) 102-112.
[6] C. Kruse, The Bayesian approach to forensic evidence: Evaluating, communicating, and distributing responsibility, Social Studies of Science 43 (2013) 657-680.
[7] K.A. Martire, R.I. Kemp, B.R. Newell, The psychology of interpreting expert evaluative opinions, Australian Journal of Forensic Sciences 45 (2013) 305-314.
[8] K.A. Martire, R.I. Kemp, M. Sayle, B.R. Newell, On the interpretation of likelihood ratios in forensic science evidence: Presentation formats and the weak evidence effect, Forensic science international 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 and Human Behavior 37 (2012) 197-207.
[10] C. Mullen, D. Spence, L. Moxey, A. Jamieson, Perception problems of the verbal scale. Science and Justice, Science and 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 Science International: Genetics 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", Journal of the Forensic Science Society 32 (1992) 5-14.
[16] J. Buckleton, C.M. Triggs, Relatedness and DNA: are we taking it seriously enough?, Forensic science international 152(2–3) (2005) 115-119.
[17] J.S. Buckleton, C.M. Triggs, C. Champod, An extended likelihood ratio framework for interpreting evidence, Science and Justice 46(2) (2006) 69-78.
[18] J.A. Bright, J.M. Curran, J.S. Buckleton, Relatedness calculations for linked loci incorporating subpopulation effects, Forensic Science International: Genetics 7 (2013) 380-383.
[19] D.L. Hartl, A.G. Clark, Principles of Population Genetics, 4th 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 science international 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 Science International: Genetics 6 (2012) 387-392.
[22] D. Taylor, J.A. Bright, J. Buckleton, J. Curran, An illustration of the effect of various sources of uncertainty on DNA likelihood ratio calculations, Forensic science international. Genetics 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 science international. Genetics 7(4) (2013) 467-70.
[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.

APPENDIX A

Hardy-Weinberg equations, or their derivations (e.g., those incorporating some form or correlation between uniting gametes such as fixation indices), 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 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 in a given offspring generation tends toward 0 as $N \to \infty$. 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, creating unnecessary confusion around the concept of 'unrelatedness' (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. We show that admitting the existence of relatives does not bias 1) genotype frequencies in the population of interest, with respect to expectations from a reference sample, nor 2) the calculation of the match probability.

Admitting relatives does not bias genotype frequencies

This first point is fairly trivial. 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). 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 a reference sample for a population, denoted here by the vector $\mathbf{P}_{\text{ref}}$, must be coherent with this

structure because the true allele frequencies ($\mathbf{P}_{pop}$) necessarily admit all these relatives. Therefore, $E(\mathbf{P}_{ref})$ should equate $\mathbf{P}_{pop}$, where $E(.)$ denotes the expectation. This statement is denied (i.e. $E(\mathbf{P}_{ref}) \neq \mathbf{P}_{pop}$) by the very definition of $\mathbf{P}_{ref}$ as reflecting only the pool of unrelated individuals. Indeed, the word 'unrelated' implies that $\mathbf{P}_{ref}$ is meaningless for forensic purposes because it refers to a non-existent fraction of the population, underscoring the fundamental problem of using this word[5]. 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.

Admitting relatives does not bias the match probability

To illustrate this second 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 $B \in I_R$, where I_R denotes circumstantial information pertaining to the suspect's relatives. 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 $\Pr(E)$[6]:

[5] This also applies to genotypes. 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.

[6] 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.

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

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 $\bar{B}$ then

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

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 assumes that *a priori* 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:

$$\Pr(E|H_d, I_R, I_O) = \Pr(E|H_d, B, B \notin I_R, I_R = \emptyset, I_O) \Pr(B) + \Pr(E|H_d, \bar{B}, \bar{B} \notin I_R, I_R = \emptyset, I_O) \Pr(\bar{B}) \qquad \text{(A.1)}$$

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 (or BN) equations to quantify the probability that a random man is the source of the trace. To demonstrate this, we must consider three possibilities of a match between the suspect and 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)[7].

Random match probability

Let G_{UK} be the profile of the unknown who presumably 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 the brother of a suspect (or person of interest) could be the unknown person who left the DNA trace (the 'random man'), we need to consider the sampling of genotypes in a population. Since we have no knowledge about the suspect's relatives (in our notation: $I_R = \varnothing$), it is assumed 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:

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

Under random mating, and assuming a hypergeometric distribution of genotype frequencies in a population of finite size N, Following Weir [25]:

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

where 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 one

[7] 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.

gamete is drawn from a suspect's parent and the other is drawn from an unknown individual, so that the random man who left the trace would be a half sib (HS) of the suspect. Finally, the last possible outcome is that the two gametes come from two unknown individuals, thus the random man is a "non-sib" (NS). The probability of these two outcomes can also be calculated from the hypergeometric distribution and

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

Note that "non-sib" does not mean 'unrelated'.

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

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

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

- **Model 1=Fixed allele frequencies:** the postulated (reference; $\mathbf{P}_{\mathrm{ref}}$) allele frequencies p_a and p_b for the population of size N are fixed. That is, if $p_a = 0.1$ and $N = 10000$, there are exactly 0.1 x 2 x 10000 = 2000 copies of allele *a* in the population. In such a population, the probability of a *a*/*b* heterozygote will be slightly upwardly biased relative to that in an infinite size population: $2p_a(2N^*p_b)/(2N\text{-}1) > 2p_ap_b$.
- **Model 2=Random allele frequencies:** allele counts in the finite population are a random draw of $2N$ alleles based on the postulated (reference) 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, that is, for two of the individuals who belong to the finite (parental) population, 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 $\Pr(G_{UK}=a/b \,|\, \mathrm{UK}=\mathrm{FS})$,

447 $\Pr(G_{UK} = a/b \mid \mathrm{UK} = \mathrm{HS})$ and $\Pr(G_{UK} = a/b \mid \mathrm{UK} = \mathrm{NS})$ by considering each pair of possible
448 suspect's parent pair, weighted by its probability.

449

450 Table A1 provides examples of the values obtained for the RMP as calculated using standard HW
451 equations compared to those calculated using the approach described here ("RMP brother"). We
452 note that the "RMP brother" is generally equal to the standard RMP for a given set of parameters
453 N, p_a and p_b. The reader will note that the "RMP brother" tends to overestimate very slightly the
454 standard RMP, but the difference is negligible even for very small populations of interest. For
455 instance, in the worse case shown in Table A1 (i.e. under model 2, when N=100, p_a=0.5 and
456 p_b=0.1), "RMP brother"/"RMP std"= 1.000613 instead of 1. This is due to the effect of the
457 knowledge of the suspect's genotype on the RMP for a finite population (which is a different
458 issue than the one addressed here). This effect is due to the negative covariance of genotypic
459 counts and increases with decreasing N. In other words, the observation of G_S update our
460 knowledge of realized genotype frequencies in the population due to the constraint that allele
461 frequencies are either fixed (model 1) or random draw from a very large (infinite) population
462 having the postulated (reference) allele frequencies. Therefore, observing $G_S = a/b$ implies that
463 one of the $2Np_a$ copies of allele a, and one of the $2Np_b$ copies of allele b, in the population, are
464 found together in the suspect, meaning that other genotypes existing in the population must be
465 made from the remaining $2Np_a$ -1 and $2Np_b$ -1 copies, limiting possible values for genotype
466 counts in an increasing manner with decreasing N (independently of the suspect's relatives issue).

467

468

Table A.1 Values obtained for the standard random match probability ("RMP std") and the random match probability accounting for the possibility that the brother is the random man ("RMP brother", which also includes the possibility for half sib), for a heterozygote a/b and various settings of N, p_a and p_b (assuming θ=0). The standard RMP 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].

			Model 1: fixed allele frequencies		**Model 2: random allele frequencies**	
p_a	p_b	N	**RMP std**	**RMP brother**	**RMP std**	**RMP brother**
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
		1,000	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
		1,000	0.10005000	0.10005000	0.09995000	0.09995006
		100	0.10050250	0.10050260	0.09950000	0.09956099