Are low *LRs* reliable?

- 1 John S. Buckleton D.Sc., Ph.D.^{1,2}, Simone N. Pugh, M.S.³, Jo-Anne Bright Ph.D.², Duncan
- 2 A Taylor^{4,5}, Ph.D., Ph.D., James M. Curran Ph.D.⁶, Maarten Kruijver Ph.D.², Peter Gill
- 3 Ph.D.^{6,7}, Bruce Budowle Ph.D.⁸ Kevin Cheng MSc^{1,2}

¹ University of Auckland, Department of Statistics, Private Bag 92019, Auckland, New Zealand

² Institute of Environmental Science and Research Limited, Private Bag 92021, Auckland, 1142 New Zealand

³ California Department of Justice, Redding, California

⁴ Forensic Science South Australia, 21 Divett Place, Adelaide, SA 5000, Australia

⁵ School of Biological Sciences, Flinders University, GPO Box 2100 Adelaide SA, Australia

⁶ Department of Forensic Biology, Norwegian Institute of Public Health, Oslo, Norway

⁷ Department of Forensic Medicine, University of Oslo, Oslo, Norway

⁸ Center for Human Identification, Department of Microbiology, Immunology, and Genetics, University of North Texas Health Science Center, 3500 Camp Bowie Blvd., Fort Worth, TX, 76107, USA

Are low *LR*s reliable?

Keywords: Forensic DNA; Probabilistic genotyping; STRmix; Likelihood ratio

Abstract

To answer the question “Are low likelihood ratios reliable?” requires both a definition of reliable and then a test of whether low likelihood ratios (*LR*s) meet that definition. We offer, from a purely statistical standpoint, that reliability can be determined by assessing whether the rate of inclusionary support for non-donors is not larger than expected from the *LR* value. Thus, it is not the magnitude of the *LR* alone that determines reliability. Turing’s rule is used to inform the expected rate of non-donor exclusionary support, where the rate of non-donor inclusionary support is at most the reciprocal of the *LR*, i.e. $\Pr(LR > x \mid H_a) < 1/x$.

There are parallel concerns about whether the value of the evidence can be communicated. We do not discuss that in depth here although it is an important consideration to be addressed with training.

In this paper, we use a mixture of real and simulated data to show that the rate of non-donor inclusionary support is significantly lower than the upper bound given by Turing’s rule. We take this as strong evidence that low *LR*s are reliable.

1.0 Introduction

It is important to demonstrate the validity and reliability of the approaches used in forensic genetics. Concerns have been raised that low-level DNA profiles are not reliable and should therefore not be used as evidence [1-4]. This issue, which has appeared in admissibility hearings in the US and Australia, leads to the question: how do we define “low-level evidence”? “Low-level” has been defined using various criteria including initial template input, signal output, mixture proportion, and/or likelihood ratio (*LR*). More recently, low-level has been described using both template and mixture proportion [1]. However, the pertinent issue is not the definition of “low-level” but rather the value of the evidence when the *LR* is obtained.

The concern is whether low *LR*s are reliable.

Briefly, low *LR*s greater than 1 are typically undervalued or may be interpreted as supporting exclusion [5]. Simply is there some threshold below which DNA evidence is designated low-level and deemed inconclusive?

Dealing with evidence of low evidential value is not a new problem. Early frequentist methods of assigning a match statistic to DNA profiles such as the combined probability of inclusion (CPI) [6] and the random match probability (RMP) [7] would often give high probabilities for frequency of inclusion. These values were easily understood as weak evidence.

The *LR* is a number on a continuous scale in the interval $[0, \infty)$. It is useful to work with the log of the *LR*. $\log_{10}LR$ s are also on a continuous scale in the interval $(-\infty, \infty)$. The familiar statements of “inclusion” and “exclusion” are often still used, but their precise definition is less clear-cut because the scale is continuous. Instead, following advice from Ian Evett (pers comm, 2000), we describe positive $\log_{10}LR$ s as supporting the prosecution’s (or contributor) hypothesis (H_p) of inclusion, and negative $\log_{10}LR$ s as supporting the alternate hypothesis

(H_a) of exclusion. Thus, with LR s, “inclusionary support” and “exclusionary support” are the terms applied. The term “exclusion” is reserved for scenarios where the $\Pr(E|H_p)=0$. That is when the evidence cannot be explained by the contributor hypothesis.

The frequentist approach exists only in the categorical inclusion/exclusion space. Such methods [8] tend not to make the most use of the available data. However, some have advocated that the CPI and RMP are easier to understand and that the apparent simplicity of the frequency reported, if high, warns the decision maker that the evidence should not be overvalued. For low positive $\log_{10}LR$ s, it is important that a similar warning is given. One way to provide this understanding is with a graphical display of the distribution of the LR under the assumption that the alternative proposition, H_a , is true, along with the value for the LR in the case [9]. The existence of positive $\log_{10}LR$ s from non-donors should not be a surprise to any scientist¹; indeed, with the frequentist approach, this observation was captured in the high probability values given in casework. While the inclusion of a non-donor is neither a new issue nor unique to LR s, there is a need to communicate to investigators and the legal arena that positive $\log_{10}LR$ s are expected from non-donors at a rate that depends on the sample quality and the results obtained.

A complexity arises due to the practice of deliberately assigning an LR believed to be at the low end of the plausible distribution. One can think of this in statistical terms as presenting a lower bound (say of a 99% credible interval) rather than the estimated mean. This practice, deemed “conservatism” in forensic genetics, is intentional and often viewed as desirable. There have been arguments against such a practice [10]. Not least among these is that the practice of deliberately understating the value of the evidence may result in wrongful eliminations of true donors. The concept of a plausible distribution introduces the topic of the uncertainty in the LR . Strictly, any uncertainty in the input parameters should be averaged out by integration, resulting in one LR point assignment (supported by a number of authors, see [11-15]). Practice, however, strongly embeds conservatism.

A sub-source LR considers the evidence given propositions or hypotheses about the donor of the DNA. A DNA profile with limited information will not allow high discrimination and will typically give inclusionary support for many genotypes. Thus, as the information in a profile decreases, more adventitious inclusionary support is observed.

The premier guidance setting body in the US is the Scientific Working Group on DNA Analysis Methods (SWGDM). SWGDM has published guidance on reporting LR s. SWGDM’s recommended scale of verbal qualifiers for reporting LR s [16] is shown in Table 1. A similar scale is adopted by European laboratories [17].

Table 1. SWGDM [16] scale of verbal qualifiers for reporting likelihood ratios

LR for H_p Support and $1/LR$ for H_a Support		Verbal Qualifier
As published	As interpreted for this work ²	
1	[1,2)	Uninformative
2 – 99	[2,100)	Limited Support
100 – 9,999	[100,10000)	Moderate Support
10,000 – 999,999	[10000,1000000)	Strong Support
$\geq 1,000,000$	[1000000, ∞]	Very Strong Support

¹ A similar concept that may be more familiar is that a laboratory technique, say a presumptive test, can sometimes give a positive result when in fact the true status is negative.

² Where [means inclusive of and) means not including. Hence a value in [1,2) is greater than or equal to 1 and strictly less than 2 ($LR \in [1,2) \Leftrightarrow 1 \leq LR < 2$)

61

62 Budowle and Bieber (pers comm, 2019) have questioned the use of a verbal scale with
63 defined boundaries, arguing that context is crucial and the ultimate users (investigators,
64 prosecutors, defense attorneys, victims, suspects, families, the public) rightly have very
65 different views on the strength of the evidence. For example, they contrast a closed
66 environment where there are only four valid possible donors with an open environment where
67 anyone with opportunity could be a donor. They point out that, in a closed environment,
68 even low-discriminating evidence may be sufficient. For instance, if three of the possible
69 donors are excluded ($LR = 0$), then any non-zero LR is meaningful evidence towards the
70 fourth person. These observations speak to the prior odds which are often unknown by the
71 scientist [18]. Budowle and Bieber warn against dismissing evidence of low positive
72 inclusionary support without someone, most suitably the fact finders, considering the context.

73 In this paper we examine the rate of adventitious matching for low-level profiles analysed
74 using the probabilistic genotyping (PG) software STRmix. As mentioned above, low-level
75 profiles have alternately been defined using template, signal, mixture proportion, and LR .
76 We use LR as the defining factor in this study, as it is the only one of these measures that is a
77 direct assessment of the information content of the profile.

78 The number of contributors (NoC) to these profiles is assigned subjectively based on the
79 DNA profile, just as it would be in casework. There has been some misinterpretation of the
80 effect of subjective assignment of NoC. It has been asserted that uncertainty in NoC leads
81 directly to uncertainty in the resulting LR [2, 3, 19]. Our position is to the contrary, and LR s
82 can be provided reliably. This issue is addressed in the study herein.

83 In order to answer the question “Are low LR s reliable?” we provide a definition of reliable.
84 Herein, “reliable” is defined as having a rate of adventitious (non-donor) inclusionary support
85 that is as expected or less. The primary test of PG software is its ability to differentiate
86 between true and non-donors. Some non-donors are expected to produce inclusionary LR s by
87 chance. Adventitious matches may also occur when comparing mixtures with individuals
88 who are closely related to contributors or due to inbreeding effects. Turing stated that the
89 expected LR for a false proposition is one if the model is correct [20]. Turing’s rule informs
90 us that the fraction of non-donors producing an $LR \geq x$ is expected to be at most $\frac{1}{x}$.

91 Mathematically stated:

92
$$\Pr(LR \geq x | H_a) \leq x^{-1}, x > 0$$

93 Turing’s rule informs us of this maximum expected rate of non-donor inclusionary support
94 which we term “Turing’s bound.” For example, suppose that $LR = 100$ is calculated for a
95 contributor to a profile. The prior expectation after comparing this profile with a sample of
96 1000 non-contributors would be to find 10 or fewer individuals with LR s that are 100 or
97 higher. This is examined through large H_a true tests.

98 **2.0 Method**

99 Ten profiles were selected from the publicly available Boston University PROVEDIt
100 Initiative (Project Research Openness for Validation with Experimental Data) dataset [21]. A
101 list of the samples used is given in Appendix 1. The mixtures were typed using the
102 GlobalFiler™ kit (Applied Biosystems) with 29 PCR cycles and run on a 3500 capillary
103 electrophoresis instrument (Applied Biosystems). The donors to the PROVEDIt datasets are
104 known (the known donors). Profiles were analysed in GeneMapper ID-X 1.5 with an

105 analytical threshold of 75 rfu (blue), 100 rfu (green), 60 rfu (yellow), 80 rfu (red) and 100 rfu
 106 (purple).

107 Mixtures were selected that appeared to have originated from four individuals as judged by at
 108 least one of two experienced operators who separately viewed the electropherograms.

109 Initially, the point estimates of the LR s with $\theta=0.01$ were assigned for the known
 110 contributors using the FBI Caucasian allele frequencies [22].

111 Propositions used for assignment of the LR s were:

112 H_p : The DNA comes from the person of interest (POI) and $N - 1$ unknown, unrelated persons
 113 H_a : The DNA comes from N unknown, unrelated persons,

114 where the POI is the donor considered and N is the assigned number of contributors.

115 The ten mixtures selected had at least one LR for a known contributor which was less than 25
 116 (see Table 3). Since ground truth is known, we retrospectively know that six of these profiles
 117 were constructed as four-person mixtures and four as five-person mixtures. In Table 2 we
 118 give a brief description of the ten mixtures interpreted within this study.

119 Table 2. A description of the mixtures used in this study. NoC stands for Number of
 120 Contributors. “Unmasked present” means that the allele was present above AT and was not
 121 masked by another contributor’s alleles or stutter of another contributor. “Unmasked absent”
 122 means that the allele was not present above AT and was not masked by another contributor’s
 123 alleles or stutter of another contributor³.

³ We counsel against treating the target NoC and fraction of key reference as ground truth. There are many steps involved in the production of an epg and experience shows that the target fractions seldom translate directly into peak heights.

Code	Target ratios	Target template (ng)	Target NoC	Treatment	Assigned NoC by two analysts		Maximum number of allelic peaks	Donor with lowest <i>LR</i>	Target fraction of donor with lowest <i>LR</i>	Lowest <i>LR</i> donor	
										Unmasked present	Unmasked absent
A10	1;4;4;4;1	0.500	5	DNase I	4	5	6	K35	7.1%	2	4
B04	1;4;4;4	0.195	4	Humic acid	4	4	6	K48	7.7%	1	6
C06	1;4;4;1	0.150	4	DNase I	4	4	6	K36	10.0%	6	2
C07	1;4;4;1	0.150	4	DNase I	4	4	6	K36	10.0%	1	5
C08	1;9;9;9;1	0.750	5	DNase I	4	5	6	K39	3.4%	1	7
D06	1;4;4;4;1	0.190	5	UV	4	5	7	K40	7.1%	3	3
E04	1;4;4;4	0.210	4	UV	3	4	6	K35	7.7%	0	7
E08	1;9;9;1	0.500	4	DNase I	3	4	7	K48	5.0%	4	1
F03	1;4;4;4	0.403	4	Humic acid	4	5	7	K37	7.7%	3	4
F07	1;4;4;4;1	0.434	5	Humic acid	4	5	7	K48	7.1%	3	2

125

126 The profiles were interpreted using STRmix version 2.7 (<https://www.strmix.com/>) using the
 127 settings described by Kelly et al. [23]. The diagnostics were examined by a skilled operator,
 128 as they would be in casework. Random Match Probability (RMP) and Cumulative
 129 Probability of Inclusion (CPI) were also calculated.

130 2.1 Experiment 1: Effect of varying theta on known and non-donors

131 Using the STRmix interpretation with assigned NoC = 4, the point estimate *LR*s for the
 132 known donors to the ten profiles and 1036 known non-donors were assigned with the co-
 133 ancestry coefficient $\theta \in \{0.00, 0.01, 0.03\}$ using the FBI extended Caucasian allele
 134 frequencies [22]. The 1036 genotypes from the NIST revised US population dataset (July
 135 2017) [24] were used as known non-donors. These donors are of differing ethnicities.

136 2.2 Experiment 2: Large H_a true test

137 The STRmix deconvolution for each mixture was compared with 10,000,000 non-donors
 138 whose genotypes had been simulated randomly from the FBI Caucasian allele frequencies.
 139 *LR*s were assigned using $\theta=0$ and the FBI Caucasian allele frequencies using the program
 140 DBLR [25]. In order to manage the data, only *LR*s ≥ 1 were stored, which allows a tabulation
 141 of the *LR*s ≥ 1 but means that the average *LR* cannot be obtained. The resulting *LR*s were
 142 reduced to aggregated counts using the SWGDAM scale of verbal qualifiers.

The $\log_{10}LR$ for the known donors were compared with assigned NoC = 4 using $\theta=0$ and the FBI Caucasian allele frequencies.

2.3 Experiment 4: Large H_p true test

The probative value of each mixture was estimated using the software DBLR. As for experiment 3, 10,000,000 non-contributors were sampled and LR s were assigned with assigned NoC = 4 using $\theta=0$ and the FBI Caucasian allele frequencies. In addition, 10,000,000 known non-donors were sampled according to the weighted genotype combinations of the minor contributor from the deconvolution [25]. H_p true distributions are beneficial to help determine the suitability of a profile or profile component for comparison to any POI profiles. The exceedance probabilities $q_i = \Pr(LR \geq 1 | H)$ were calculated with H replaced with the hypotheses H_p and H_a using both the Monte Carlo estimate and importance sampling as described by Kruijver et al. [25].

2.4 Experiment 5: Effect of varying NoC

As a by-product of the subjective assignment of NoC, there are seven instances where two plausible NoC were assigned for the same profile by the different analysts. Samples E04 and E08 were assigned as either NoC = 3 or NoC = 4, and samples A10, C08, D06, F07, and F03 were assigned as either NoC = 4 or NoC = 5 (Table 2). The LR for the known contributors to these profiles was assigned using $\theta=0$ and the FBI Caucasian allele frequencies. The LR s using NoC = 3 or 5 were compared to the original values where NoC = 4.

3.0 Results

3.1 Experiment 1: Effect of varying θ on known and non-donors

The $LR_{\theta=0.01}$ values for the known contributors are given in Table 3 for the assigned NoC = 4. Tables 4a and 4b show the count of the 1036 non-donors giving $LR_{\theta=0.01}$ in various categories and the probability that an LR is greater than or equal to a nominated LR value for each of the ten profiles, assuming four contributors. A plot of the distribution of non-donor LR s for θ values of 0.00, 0.01, or 0.03 is given in Figure 1.

169 Table 3. $LR_{\theta=0.01}$ values for the known contributors. Assigned NoC = 4. LR s for known donors (K29 – K50) less than 25 are shown in bold.

Sample	K29	K33	K34	K35	K36	K37	K38	K39	K40	K48	K49	K50
A10				0.67	1.2×10^{10}	3.2×10^5	4.1×10^7	77.3				
B04	7.0×10^5									2.2	6.6×10^8	2.5×10^{13}
C06		42	5.3×10^{11}	1.8×10^{13}	14.5							
C07		0.21	1.1×10^{10}	1.7×10^{16}	2.5							
C08					0.005	1.5×10^6	3.4×10^{11}	7.0×10^{10}	2.6			
D06				11.5	1.6×10^{11}	7.3×10^6	9.3×10^5	1.6×10^4				
E04	6.5×10^3									3.8	1.9×10^5	2.6×10^8
E08						6.1	6.4×10^{21}	7.9×10^{21}	7.8×10^5			
F03	5.3×10^9									5.1	5.1×10^7	1.9×10^{15}
F07				1.3×10^4	1.6×10^{10}	6.9×10^4	1.5×10^{16}	3.5				

171 Table 4a. The count of non-donors with $LR_{\theta=0.01}$ between certain values for the 1036 non-
172 donors from the NIST database. Assigned NoC = 4.

LR value	A10	B04	C06	C07	C08	D06	E04	E08	F03	F07
[1,2)	10	5	6	16	2	9	9	24	1	
[2,3)	4	2	5	13		1	9	14	1	
[3,4)	2	2	2	6		3	4	5		
[4,5)				3	1	1	1	4		
[5,10)	3	3	6	6		5	2	13	1	3
[10,15)	2	2	3	2			5	3	1	1
[15,20)		2	1			1		3		
[20,25)			1	2				1		
[25,30)	1									
[30,35)							2			
≥ 35	2	1	2	2	2	1	2	4	1	1
Max $LR_{\theta=0.01}$	45	402	146	45	282	49	99	281	187	332

173

174 Table 4b. The cumulative counts of non-donors with $LR_{\min}^{\theta=0.01} \geq$ certain values for the 1036
175 non-donors from the NIST database. Assigned NoC = 4. Turing's bound is also given as the
176 maximum expected count.

LR value	Turing's bound	A10	B04	C06	C07	C08	D06	E04	E08	F03	F07
≥ 1	1036^4	24	17	26	50	5	21	34	71	5	5
≥ 2	518	14	12	20	34	3	12	25	47	4	5
≥ 3	345	10	10	15	21	3	11	16	33	3	5
≥ 4	259	8	8	13	15	3	8	12	28	3	5
≥ 5	207	8	8	13	12	2	7	11	24	3	5
≥ 10	104	5	5	7	6	2	2	9	11	2	2
≥ 15	69	3	3	4	4	2	2	4	8	1	1
≥ 20	52	3	1	3	4	2	1	4	5	1	1
≥ 25	41	3	1	2	2	2	1	4	4	1	1
≥ 30	35	2	1	2	2	2	1	4	4	1	1
≥ 35	30	2	1	2	2	2	1	2	4	1	1

177

⁴ The bound has limited value for LR s as low as 1 or 2. Consider the cell for $LR \geq 1$. This cell informs us that there must be fewer than 1036 profiles (that is, all of the non-donors) that have an $LR \geq 1$. Since Turing informs us that the average LR should be 1, we cannot have more than 1036 profiles with an $LR \geq 1$. In reality, there are many fewer.

178

179

180 Figure 1. The non-donor distributions of $\log_{10} LR$ for the three values for
 181 $\theta \in \{0.00, 0.01, 0.03\}$ (excluding values where $LR = 0$)

182 3.2 Experiment 2: Large H_a true tests
 183 The results of the $10 \times 10,000,000$ non-donor tests are given in Tables 5a and 5b.
 184 Table 5a. The count of LR s obtained for simulated non-donors in each of the SWGDAM verbal qualifier ranges and some additional categories
 185 for each of 10,000,000 non-donor tests for each mixture using $\theta=0$ and the FBI Caucasian allele frequencies. Also given is the count of
 186 simulated non-donors with LR s that equal or exceed the LR for the POI.

Count of LR	Mixture									
	A10	B04	C06	C07	C08	D06	E04	E08	F03	F07
[1,2)	41,815	66,635	114,711	138,561	14,349	62,202	155,936	25,782	7780	6679
[2,10)	113,810	81,070	287,186	248,252	14207	78286	319,986	28,770	10,576	8148
$[10^2,10^2)$	62,874	32,280	133,268	61,171	3698	33636	120,176	10,683	5602	3427
$[10^2,10^3)$	11,206	4998	16,444	5679	309	5736	13,796	1869	1413	660
$[10^3,10^4)$	1506	572	1256	314	12	721	1120	212	306	87
$[10^4,10^5)$	108	48	58	11		63	52	21	45	6
$[10^5, 10^6)$	12	2	2			3	2	1	5	2
$\geq 10^6$										
$LR_{POI} \theta=0$	3	6	34	282	205	48	7	205	20	15
$\geq LR_{POI}$	186,634	58,227	50,896	1700	130	12,024	195,408	1118	4903	3149

187
 188

189 Table 5b. The count of non-donors *LR*s with values greater than various values for the 10,000,000 simulated non-donors using $\theta=0$ and the
190 FBI Caucasian allele frequencies. Assigned NoC = 4.

Cumulative	Turing's expectation	Mixture									
		A10	B04	C06	C07	C08	D06	E04	E08	F03	F07
≥ 1	10,000,000	231,331	185,605	552,925	453,988	32,575	180,647	611,068	67,338	25,727	19,009
≥ 2	5,000,000	189,516	118,970	438,214	315,427	18,226	118,445	455,132	41,556	17,947	12,330
≥ 10	1,000,000	75,706	37,900	151,028	67,175	4,019	40,159	135,146	12,786	7,371	4,182
≥ 100	100,000	12,832	5,620	17,760	6,004	321	6,523	14,970	2,103	1,769	755
$\geq 1,000$	10,000	1,626	622	1,316	325	12	787	1,174	234	356	95
$\geq 10,000$	1,000	120	50	60	11		66	54	22	50	8
$\geq 100,000$	100	12	2	2			3	2	1	5	2
$\geq 1,000,000$	10										
$LR_{POI} \theta=0$		3	6	34	282	205	48	7	205	20	15
$\geq LR_{POI}$		186,634	58,227	50,896	1700	130	12,024	195,408	1118	4903	3149
Turing's expectation $\geq LR_{POI}$		3,333,333	1,666,667	294,118	35,461	48,780	208,333	1,428,571	48,780	500,000	666,667

3.3 Experiment 3: Large H_a true tests

Figure 2 shows a plot from DBLR of a smoothed histogram of $\log_{10}LR$ s sampled according to expectation for the minor (fourth) contributor of the mixture (H_p) and non-contributors (H_a) for sample B04. In Table 6 the exceedance probabilities are listed for all mixtures for H_p and H_a (using both Monte Carlo estimate and importance sampling). As an example, the results of the simulation testing for sample B04 indicated that $\Pr(LR \geq 1 | H_p) \approx 0.9798$ and $\Pr(LR \geq 1 | H_a) \approx 0.0019$; or in words, the probability of an LR greater than 1 given a true contributor is greater than 97.98% and the probability of an LR greater than 1 given a non-contributor is greater than 0.19%.

Figure 2: Smoothed histogram of $\log_{10}LR$ values for 10,000,000 H_p (H1 true, blue) and H_a (H2 true, red) simulations for sample B04.

Table 6: Summary of exceedance probabilities, $\Pr(LR \geq 1 | H)$, for 10,000,000 H_p and H_a true simulations for all mixtures where assigned NoC=4

Sample	$\Pr(LR \geq 1 H_p)$ Monte Carlo estimate	$\Pr(LR \geq 1 H_a)$ Monte Carlo estimate	$\Pr(LR \geq 1 H_a)$ Importance sampling estimate
A10	0.9834	0.001698	0.001697
B04	0.9798	0.001884	0.001881
C06	0.9544	0.004210	0.004212
C07	0.9458	0.004565	0.004563
C08	0.9938	0.0003500	0.0003529
D06	0.9819	0.001813	0.001817
E04	0.9553	0.004062	0.004056
E08	0.9897	0.0006975	0.0006993
F03	0.9979	0.0002573	0.0002570
F07	0.9977	0.0002002	0.0002015

3.4 Experiment 4: Varying NoC

Figure 3 is a plot of $\log_{10}LR$ for the known donors when two different NoC were assigned using $\theta=0$ and the FBI Caucasian allele frequencies. On the x -axis is the analysis using the target NoC (the NoC used to construct the mixture). On the y -axis is the analysis using either one more (N+1) or one fewer (N-1) contributor. Two comparisons gave $LR = 0$ and are not plotted in Figure 3. The exclusion of one or more contributors (usually the lowest-template donor) is a known event with STRmix and usually occurs because one donor has contributed no meaningful signal [26, 27].

Figure 3. A plot of $\log_{10}LR$ for all known contributors using $\theta=0$ and the FBI Caucasian allele frequencies. On the x -axis is the analysis using the target NoC (the NoC used to construct the mixture). On the y -axis is the analysis using either one more (N+1) or one

fewer (N-1) contributor. The lines $x = y$, $x-1 = y$, and $x+1 = y$ are plotted. Two comparisons gave exclusions for N-1 and are not plotted.

4.0 Discussion

Interpretation of these ten profiles following the manual methods outlined by Bille et al. and Bieber et al. [6, 7] led to all loci in all ten profiles being deemed inconclusive due to potential dropout. PG is the most suitable modern method for interpretation of these types of complex DNA profiles.

Experiment 1 demonstrates the effect of θ on the LR for non-donors. Common values for theta used for casework are 0.01 and 0.03. These values have been shown to be reasonable for many sub-populations including Caucasians, African Americans, and Hispanics [28]. The choice of allele frequencies depends on the jurisdiction and circumstances of the crime. The use of different allele frequencies will lead to different LR s; in this work the FBI Caucasian extended allele frequencies were used for all LR assignments. In practise, many laboratories use allele frequencies from multiple different sub-populations and report either all relevant LR s, some form of stratification across populations, or the lowest assigned LR .

Of note is that even for these very complex and low-level mixtures, the rate and magnitude of inclusionary support for non-donors are low and certainly lower than expected from Turing's rule (see Table 4b). The non-donor distributions for $\theta \in \{0.00, 0.03\}$ differ little (see Figure 1). At most there is a small movement of the distribution to lower $\log_{10}LR$ as θ is increased.

In Experiment 2, LR s for 10 million non-donors were assigned using STRmix with $\theta = 0$. This analysis removes some conservatism to allow a more direct assessment of the rate and magnitude of inclusionary support for non-donors. Simulated non-donors were used, allowing an increase in the number tested from the 1036 individuals tested in Experiment 1. Again, the rate and magnitude of inclusionary support for non-donors are low and also lower than expected from Turing's rule. For example, from Table 5b, 10 non-donors giving $LR \geq 1$ million were predicted, while the actual observation was zero for all of the ten profiles. The observed number is lower than the expected number for all categories for all mixtures. This observation suggests that low LR s assigned by STRmix are reliable when stripped of all their layers of conservativeness. In casework, several practices that have a conservative (i.e. lowering) effect on the LR are commonly and deliberately employed -in addition to the use of θ . Two examples of these conservative practices are the implementation of the Balding-Nichols subpopulation model [29], where the value for θ is usually set at the high end of the plausible range, and the use of the highest posterior density calculation, which uses a 0.99 lower bound on both allele frequencies and Monte Carlo effect. Removing these bounds exposes the software to a more rigorous test. If the software performs adequately in this instance, then it should perform conservatively when the usual layers of conservativeness are added. -Although low LR s should be considered reliable, it is still important that they are effectively expressed as weak supportive evidence (depending on the context in the case).

Experiment 3 gives an example of a smoothed histogram (Figure 2) of $\log_{10}LR$ s sampled according to expectation for the minor (fourth) contributor to the mixture (H_p) and non-contributors (H_a) for sample B04. Table 6 gives $\Pr(LR \geq 1 | H)$ for 10,000,000 H_p and H_a true simulations for all mixtures where assigned NoC = 4. These results are presented to inform the question of whether they have any place in describing a casework comparison

with an alternate perspective, or discussing the Daubert question of the rate of misleading evidence (error rate). Recall, replacing the LR with the exceedance probability loses logical rigor and has a tendency to overstate the value of the evidence [30].

In Experiment 4, mixtures were run using two different NoC. The LR does not change significantly for the higher-level donors (see Figure 3) when the assigned NoC differs from the target NoC. For the trace donors, the change in LR is more marked, often resulting in a false exclusion for N-1. Four LR s have moved downwards for N-1 (two data appear in Figure 3 and two further gave exclusions and are not plotted). This effect is a known phenomenon and should be expected [26, 27, 31]. If there are four known donors and the mixture is treated as NoC = 3, then there are only three “places,” and it is likely that one donor, usually the smallest, has nowhere to align. To emphasize, that the donor that is excluded has so little signal that it was not noticed during assignment of NoC.

The data demonstrate our expectation that a portion of non-contributors will have an $LR > 1$ and the distribution is well within in predicted limits given Turing’s bound. Concern has been expressed that the uncertainty in NoC has an impact on uncertainty in the LR . The important point is that underestimating the NoC tends to cause false exclusion of a low-level minor contributor, and overestimating tends to cause inclusion of non-contributors into low-level positions with appropriately low LR s.

We have empirically shown that the rate of non-donor inclusionary support is markedly below the maximum expected from Turing’s rule. A theoretical analysis gives further insight. A DNA profile analysed using STRmix is assigned a list of weighted genotype combinations at each locus. An LR is assigned by comparing the genotype of a POI to these genotype combinations and computing likelihoods under two alternate propositions. These likelihoods are computed using the weights and estimated genotype probabilities informed from allele frequency tables and a subpopulation correction. The LR distribution for non-donors can be explicitly obtained (for example [25]), and the rate of non-donor inclusionary support can be precisely estimated from the subpopulation genetic model and the mixture deconvolution. The appearance of the resulting LR distributions depends on how well the profile is resolved.

At one extreme, a fully resolved single-source profile has only a single genotype at each locus with a weight of 1.00. Hence, there is a single genotype giving a large LR and every other genotype would give a categorical exclusion. In the absence of subpopulation correction, the true donor would be assigned an LR of $1/RMP$ and the rate of non-donor inclusionary support equals the RMP . Hence, the rate of non-donor inclusions is very small, but if a non-contributor is included, then the evidence will be very strong. In this case, Turing’s bound is an equality for the true donor’s LR but is crude for any smaller LR threshold. On the other hand, if a mixture is very poorly resolved, there will be many genotype combinations with low weights. Accordingly, there will be many non-donor genotypes that yield small inclusionary LR s. Any deconvolution sits between these two extremes. Either there are many genotypes giving weak LR s or few genotypes giving strong LR s. Turing’s bound is crude unless a threshold in the right tail of the LR distribution is chosen. The weighted genotype list guarantees that strong non-donor inclusions are rare.

As mentioned above, the rate of non-donor inclusionary support is bounded by Turing’s rule but can also be precisely estimated from a model. The models that are used in practice

strongly embed conservatism. The experiments have demonstrated that the rate of non-donor inclusion realised in practice is much lower than predicted from a conservative model.

Acknowledgements

Removed for refereeing

References

- [1] Hon. Janet T. Neff. United States v. Gissantaner. 2019. <https://casetext.com/case/united-states-v-gissantaner> Accessed 19 February 2020.
- [2] Stiffelman B. No Longer the Gold Standard: Probabilistic Genotyping is Changing the Nature of DNA Evidence in Criminal Trials. Berkeley Journal of Criminal Law. 2019;24:110-46.
- [3] President's Council of Advisors on Science and Technology. Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods. 2016. https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/PCAST/pcast_forensic_science_report_final.pdf Accessed 22 April 2017.
- [4] President's Council of Advisors on Science and Technology. An addendum to the PCAST report on forensic science in criminal courts. 2016. https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/PCAST/pcast_forensics_addendum_finalv2.pdf Accessed 20 July 2017.
- [5] Eldridge H. Juror comprehension of forensic expert testimony: A literature review and gap analysis. Forensic Science International: Synergy. 2019;1:24-34.
- [6] Bieber FR, Buckleton JS, Budowle B, Butler JM, Coble MD. Evaluation of forensic DNA mixture evidence: protocol for evaluation, interpretation, and statistical calculations using the combined probability of inclusion. BMC Genetics. 2016;17:125.
- [7] Bille T, Bright J-A, Buckleton J. Application of Random Match Probability Calculations to Mixed STR Profiles. Journal of Forensic Sciences. 2013;58:474-85.
- [8] Buckleton J, Curran J. A discussion of the merits of random man not excluded and likelihood ratios. Forensic Science International: Genetics. 2008;2:343-8.
- [9] Schuerman C, Kalafut T, Buchanan C, Sutton J, Bright J-A. Using the Nondonor Distribution to Improve Communication and Inform Decision Making for Low LR_s from Minor Contributors in Mixed DNA Profiles. Journal of Forensic Sciences.n/a.
- [10] Hicks T, Buckleton JS, Bright J-A, Taylor D. A framework for interpreting evidence. In: Buckleton JS, Bright J-A, Taylor D, editors. Forensic DNA Evidence Interpretation. Boca Raton, Florida: CRC Press; 2016.

- [11] Taroni F, Aitken CGG, Garbolino P. De Finetti's subjectivism, the assessment of probabilities and the evaluation of evidence: a commentary for forensic scientists. *Science and Justice*. 2001;41:145-50.
- [12] Biedermann A, Garbolino P, Taroni F. The subjectivist interpretation of probability and the problem of individualisation in forensic science. *Science and Justice*. 2013;53:192-200.
- [13] Biedermann A, Bozza S, Taroni F, Aitken C. The consequences of understanding expert probability reporting as a decision. *Science and Justice*. 2017;57:80-5.
- [14] Biedermann A, Bozza S, Taroni F, Aitken C. The meaning of justified subjectivism and its role in the reconciliation of recent disagreements over forensic probabilism. *Science and Justice*. 2017;57:477-83.
- [15] Biedermann A, Bozza S, Taroni F, Aitken C. Reframing the debate: A question of probability, not of likelihood ratio. *Science and Justice*. 2016;56:392-6.
- [16] Scientific Working Group on DNA Analysis Methods. Recommendations of the SWGDAM Ad Hoc Working Group on Genotyping Results Reported as Likelihood Ratios. 2018. https://docs.wixstatic.com/ugd/4344b0_dd5221694d1448588dcd0937738c9e46.pdf Accessed 8 November 2018.
- [17] Willis SM, McKenna L, McDermott S, O'Donnell G, Barrett A, Rasmusson B, et al. ENFSI guideline for evaluative reporting in forensic science. 2015. http://enfsi.eu/wp-content/uploads/2016/09/m1_guideline.pdf Accessed 23 October 2019.
- [18] Gill P, Hicks T, Butler JM, Connolly E, Gusmão L, Kokshoorn B, et al. DNA commission of the International society for forensic genetics: Assessing the value of forensic biological evidence - Guidelines highlighting the importance of propositions: Part I: evaluation of DNA profiling comparisons given (sub-) source propositions. *Forensic Science International: Genetics*. 2018;36:189-202.
- [19] McNevin D, Wright K, Chaseling J, Barash M. Commentary on: Bright et al. (2018) Internal validation of STRmix™ – a multi laboratory response to PCAST, *Forensic Science International: Genetics*, 34: 11–24. *Forensic Science International: Genetics*. 2019;41:e14-e7.
- [20] Good IJ. *Probability and the Weighing of Evidence*. London: Charles Griffin & Company Limited; 1950. p. 72.
- [21] Alfonse LE, Garrett AD, Lun DS, Duffy KR, Grgicak CM. A large-scale dataset of single and mixed-source short tandem repeat profiles to inform human identification strategies: PROVEDIt. *Forensic Science International: Genetics*. 2018;32:62-70.
- [22] Moretti TR, Moreno LI, Smerick JB, Pignone ML, Hizon R, Buckleton JS, et al. Population data on the expanded CODIS core STR loci for eleven populations of significance

375 for forensic DNA analyses in the United States. Forensic Science International: Genetics.
376 2016;25:175-81.

377 [23] Kelly H, Bright J-A, Kruijver M, Cooper S, Taylor D, Duke K, et al. A sensitivity
378 analysis to determine the robustness of STRmix™ with respect to laboratory calibration.
379 Forensic Science International: Genetics. 2018;35:113-22.

380 [24] Steffen CR, Coble MD, Gettings KB, Vallone PM. Corrigendum to 'U.S. Population
381 Data for 29 Autosomal STR Loci'; [Forensic Sci. Int. Genet. 7 (2013) e82-e83]. Forensic
382 Science International: Genetics. 2017;31:e36-e40.

383 [25] Kruijver M, Bright J-A, Kelly H, Buckleton J. Exploring the probative value of mixed
384 DNA profiles. Forensic Science International: Genetics. 2019;41:1-10.

385 [26] Bright J-A, Curran JM, Buckleton JS. The effect of the uncertainty in the number of
386 contributors to mixed DNA profiles on profile interpretation. Forensic Science International:
387 Genetics. 2014;12:208-14.

388 [27] Bright J-A, Richards R, Kruijver M, Kelly H, McGovern C, Magee A, et al. Internal
389 validation of STRmix™ – A multi laboratory response to PCAST. Forensic Science
390 International: Genetics. 2018;34:11-24.

391 [28] Buckleton J, Curran J, Goudet J, Taylor D, Thiery A, Weir BS. Population-specific F_{ST}
392 values for forensic STR markers: A worldwide survey. Forensic Science International:
393 Genetics. 2013;23:91-100.

394 [29] Curran JM, Buckleton JS, Triggs CM. What is the magnitude of the subpopulation
395 effect? Forensic Science International. 2003;135:1-8.

396 [30] Kruijver M, Meester R, Slooten K. p -Values should not be used for evaluating the
397 strength of DNA evidence. Forensic Science International: Genetics. 2015;16:226-31.

398 [31] Bright J-A, Taylor D, Curran J, Buckleton J. Searching mixed DNA profiles directly
399 against profile databases. Forensic Science International: Genetics. 2014;9:102-10.

400 Appendix 1

401 A list of samples names used in the interpretation derived from Alfonse et al. [21]

402 A10_RD14-0003-35_36_37_38_39-1;4;4;4;1-M2d-0.5GF-Q1.4_01.15sec.hid

403 D06_RD14-0003-35_36_37_38_39-1;4;4;4;1-M2U15-0.21GF-Q1.0_04.15sec.hid

404 F07_RD14-0003-35_36_37_38_39-1;4;4;4;1-M3I22-0.434GF-Q2.0_06.15sec.hid

405 C08_RD14-0003-36_37_38_39_40-1;9;9;9;1-M2c-0.75GF-Q1.6_03.15sec.hid

406 E08_RD14-0003-37_38_39_40-1;9;9;1-M2c-0.5GF-Q0.6_05.15sec.hid

407 C06_RD14-0003-33_34_35_36-1;4;4;1-M2c-0.15GF-Q0.8_03.15sec.hid

408 C07_RD14-0003-33_34_35_36-1;4;4;1-M2e-0.15GF-Q0.9_03.15sec.hid

409 B04_RD14-0003-48_49_50_29-1;4;4;4-M3I22-0.195GF-Q1.7_02.15sec.hid

410 E04_RD14-0003-48_49_50_29-1;4;4;4-M2U105-0.195GF-Q12.1_05.15sec.hid

411 F03_RD14-0003-48_49_50_29-1;4;4;4-M2I15-0.403GF-Q1.1_06.15sec.hid