Article:

Bright, J.A., Taylor, D., Curran, J., & Buckleton, J. (2014). **Searching mixed DNA profiles directly against profile databases**. *Forensic Science International: Genetics*, *9*, 102–110.

INSTITUTE OF ENVIRONMENTAL SCIENCE AND RESEARCH LIMITED
Kenepuru Science Centre: 34 Kenepuru Drive, Kenepuru, Porirua 5022 | PO Box 50348, Porirua 5240, New Zealand
T: +64 4 914 0700 **F:** +64 4 914 0770
www.esr.cri.nz

Searching mixed DNA profiles directly against profile databases

Jo-Anne Bright[1,2*], Duncan Taylor[3], James Curran[2], John Buckleton[1]

[1] *Institute of Environmental Science and Research Limited, Private Bag 92021 Auckland 1142, New Zealand*

[2] *Department of Statistics, University of Auckland, Private Bag 92019, Auckland 1142, New Zealand*

[3] *Forensic Science South Australia, 21 Divett Place, Adelaide, SA 5000, Australia*

* Corresponding author at: Institute of Environmental Science and Research Limited, Private Bag 92021, Auckland, 1142, New Zealand. Email address: Jo.bright@esr.cri.nz .

DNA databases have revolutionised forensic science. They are a powerful investigative tool as they have the potential to identify persons of interest in criminal investigations. Previously, a DNA profile generated from a crime sample could only be searched for in a database of individuals if the stain was from single contributor (single source) or if a contributor could unambiguously be determined from a mixed DNA profile. This meant that a significant number of samples were unsuitable for database searching.

The advent of continuous methods for the interpretation of DNA profiles offers a way to circumvent this restriction. Using these methods, each profile on the database may be considered a possible contributor to a mixture and a likelihood ratio (*LR*) can be formed. Those profiles which produce a sufficiently large *LR* can serve as an investigative lead.

In this paper empirical studies are described to determine what constitutes a large *LR*. We investigate the effect on a database search of complex mixed DNA profiles with contributors in equal proportions with dropout as a consideration, and also the effect of an incorrect assignment of the number of contributors to a profile. In addition, we give, as a demonstration of the method, the results using two crime samples that were previously unsuitable for database comparison. We show that effective management of the selection of samples for searching and the interpretation of the output can be highly informative.

Keywords: Forensic DNA; Database; Continuous models; Likelihood ratio

Introduction

DNA databases can be powerful tools in the identification of individuals of interest during a criminal investigation. Typically, DNA databases consist of two sub databases; one containing profiles from known individuals who have either volunteered or been compelled to provide a sample (the database) and the other is a database of profiles collected from samples associated with crime scenes [1] (the crime sample database). The records in the separate databases can be compared to each other to link individuals with crime scenes. This comparison process typically takes a crime scene profile and compares it to each database sample in turn. Often a count is made of concordant and non-concordant alleles. A wild card designation may be included in the crime sample profile or more rarely in the database profile. The wildcard is deemed to be concordant with any allele. Most search algorithms are set up to compare two alleles per locus from the crime sample profile with the two alleles per locus from the database profiles. This approach restricts profiles suitable for searching to single source profiles or a single source component inferred, either completely or partially, from a mixed DNA profile.

This investigative intelligence is provided to investigators to assess in conjunction with the wider case information.

If both the crime sample profile and the database profile are full multilocus profiles then the chance of an adventitious match is small. Adventitious matches are more likely with low level, partial or mixed profiles. Many databases now include profiles from superseded multiplexes which may have as few as six loci scored. Adventitious matches, although expected, can reduce the credibility of the databank operation or even the forensic use of DNA. As an example, the discovery of a number of partial matches in the Arizona database led to considerable discussion including some adverse comment even though these partial matches occurred at approximately the expected rate (see Mueller [2])

The quality of the database can be ensured by legislation such as restricting the type of sample, setting a minimum number of alleles required for database entry, by mandatory participation in quality assurance programmes (as in the USA), and by participation in external audits and proficiency tests [3].

Whilst it is relatively easy to meet a very high standard for the database profiles, the profiles from crime scenes however are frequently compromised in quality.

The likelihood ratio (*LR*) is generally accepted as the most powerful and relevant statistic for the calculation of the weight of the DNA evidence [4]. It is the ratio of the probability of the observed crime stain (*O*) given each of two competing hypotheses, H_1 and H_2, and given all the available information, *I.* Mathematically, we express this as:

$$LR = \frac{\Pr(O \mid H_1, I)}{\Pr(O \mid H_2, I)}$$

Typically database search algorithms do not calculate an *LR* but simply report the number of concordant and non-concordant alleles. However for unresolvable or low level mixtures the use of an *LR* confers considerable advantages.

Stochastic events such as heterozygote imbalance, allelic dropout, locus dropout, and allelic drop in can complicate interpretation [5-7]. The uncertainty in the numbers of contributors and

stutter, a by-product of the PCR process, can further complicate profile interpretation whenever stutter peaks are of a similar height to the minor allelic peaks in mixed DNA profiles.

The number of contributors to a mixed DNA profile is easily determined if the number of alleles is known. It is the step of inferring how many alleles are present from the peaks that is the source of uncertainty. Some peaks are not allelic (for example artefacts or stutter peaks) and some represent contributions from two or more alleles from the same or different individuals superimposed. Some alleles may not have produced a peak due to dropout. At high sensitivity it is possible that some peaks are formed by alleles from the laboratory environment, termed drop-in. Information from replicate amplifications and in certain situations Y STR analysis can be helpful in providing a reasonable estimate of the number of contributors. Statistical methods such as maximum likelihood [8] or Bayesian networks [9] are more statistically sound, and can compensate for artefacts such as stutter, and dropout.

The suggestion that there is a correct number of contributors for every profile would seem self-evidently true but overlooks the fact that this number is inherently unknown and that it is conditioned on what is known about the profile. It should be noted that there is no reason for the number of contributors to be the same under the hypotheses H_1 and H_2. However, proposing an unreasonable number of contributors under the defence hypothesis and holding the number under the prosecution hypothesis at a reasonable assignment will increase the *LR*, favouring the prosecution hypothesis [10].

Complications in profile interpretation have led to a recent push for forensic laboratories to introduce improved models for DNA interpretation. This is motivated by the difficulties traditional methods have with the interpretation of complex profiles [11, 12]. The traditional methods of interpretation are described as binary which describes the fact that the probability of the genotype combination under consideration is assigned as zero or one (hence binary) [13]. Following Kelly et al. [14] we denote the genotype of the observed crime stain as O, and the genotypes of proposed donors as G_i for donor i. For an N donor mixture there are N proposed genotypes, G_i for each proposed combination. The j^{th} set of N genotypes is denoted S_j. Binary models assign the values zero or one to the unknown probability $\Pr(O \mid S_j)$ based on heuristics such as heterozygote balance and mixture proportion, the reasonable values of which are informed by empirical data. Essentially, $\Pr(O \mid S_j)$ is assigned a value of zero if the genotype combination falls outside of these heuristics. $\Pr(O \mid S_j)$ is assigned a value of one if it falls within. These binary methods are slowly being replaced by more advanced interpretation methods, such as the semi-continuous models likeLTD and LRmix and continuous models which can take into account stochastic events. STRmix [15, 16], TrueAllele [17] and LiRa Ht [18], are examples of software that employ a continuous model for DNA profile interpretation.

A continuous model uses the quantitative information from an electropherogram (epg) such as peak heights, to calculate the probability of the peak heights given all possible genotype combinations, assigning a value or weight (w_i) to the normalised probability $\Pr(O \mid S_j)$. Continuous models can remove some of the qualitative thresholds such as heterozygote balance and may remove some of the subjective decisions required within a binary model. A discussion of the merits of the different interpretation models can be found in Kelly et al. [14].

STRmix assigns a relative weighting to the probability of the epg given each possible genotype combination at a locus. The weights across all combinations at that locus sum to one. Therefore, a single unambiguous genotype combination at any locus would be assigned a weighting of one.

Good quality single source DNA profiles, where stochastic effects are not an issue, are likely to result in a profile of sufficient quality for entry to a crime sample database regardless of the interpretation method used. However mixed profiles, or single source profiles subject to stochastic effects, may not result in a profile suitable for entry to a database using traditional binary methods. Interpretation of these profiles using a continuous model may result in

improved profile information and therefore permit database entry. Unless the weight for any given genotype combination is one, assessing the 'quality' of a profile for its suitability for comparison to a database is not straightforward. A guideline for database entry based on some assessment of the risks of loading an incorrectly inferred profile may be employed where the genotype combination of a contributor is ambiguous, such as $w_i > 0.99$. If an individual's profile cannot be reasonably inferred from a DNA mixture, regardless of the interpretation method, then it is unsuitable for entry to a database using traditional database methods.

The number of contributors to a mixed DNA profile (*N*) cannot be known with certainty. It may be the case that the same electropherogram can be interpreted as having come from several different numbers of contributors. Assigning the probable numbers of contributors to a mixed DNA profile is more complicated with low level profiles. Uncertainty is increased when peaks are close to the limit of detection or there are additional peaks just below the analytical threshold. These cases might invoke the addition of a contributor to a profile. Overestimating the number of contributors to a profile has the potential to generate an *LR* that favours inclusion of known non-contributors, whereas underestimating the number of contributors has the potential to generate an *LR* that favours exclusion of a known contributor. Neither of these outcomes is desirable.

The number of contributors must be specified when using current likelihood ratio implementations for profile interpretation [19]. Direct comparison of mixed DNA profiles, where there are multiple possible genotype combinations at one or more loci, to profiles of individuals within a database, can be undertaken using the output of a continuous method of interpretation with a modified search algorithm using a likelihood ratio framework.

In this paper, a method for database entry and comparison to the New Zealand DNA Profile Databank (DPD)[1] of previously unsuitable mixed DNA profiles is described. We examine the efficacy of the method using artificially prepared low level, mixed DNA profiles where the individual contributor profiles are known. We also report the results of two case examples.

Method

Database profiles were blood samples or saliva stains on FTA® Classic or Elute card (Whatman, Maidstone, England). The method for processing is described in Bright et al. [21].

Eight artificial mixed DNA profiles were prepared by amplifying extracted DNA from three known sources with the approximate mixture proportions of 10:5:1 (referred to as major:minor:trace) in varying contributor orders. DNA from the eight prepared mixtures and two case examples was extracted using Promega's DNA IQ™ magnetic bead extraction chemistry (Madison, WI) and quantified using Applied Biosystems Quantifiler™ real time PCR quantitation kit (Life Technologies). A target of 1.5 ng of DNA was amplified using Applied Biosystems' Identifiler™ multiplex (Life Technologies, Carlsbad CA) on an Applied Biosystems 9700 thermal cycler with a

1 The NZ DPD was established in 1996 [20] S. A. Harbison, J. F. Hamilton, S. J. Walsh, The New Zealand DNA databank: its development and significance as a crime solving tool, Science and Justice. 41 (2001) 33-37. and comprises DNA profiles amplified using the Second Generation Multiplex (SGM, Forensic Science Services, UK), and Applied Biosystems' SGMPlus™ and Identifiler™ multiplexes (Life Technologies, Carlsbad CA). As at April 2013 the DPD comprised 8,860 SGM profiles, 65,568 SGMPlus™ profiles and 69,543 Identifiler™ profiles.

silver block as per manufacturer's recommendations [22]. Amplified products were separated on an Applied Biosystems' 3130xl Genetic Analyser and data was analysed using Applied Biosystems' GeneMapper™ ID version 3.2.1 using a 50 relative fluorescent unit (rfu) analytical threshold. Prior to interpretation, the heights of all peaks within the epg of the eight artificial mixtures were halved in order to mimic low level profiles or further modified as described in each experimental method below. Peaks that subsequently fell below 50 rfu were removed prior to interpretation.

In addition, an artificial two person mixture was created by combining the alleles from two known individuals in the proportion 1:1. In one replicate, the peaks were set to a height where dropout would not be a consideration (called two person without drop) and in another replicate the peaks were lowered to a height where dropout was very likely (two person with dropout). An artificial three person mixed DNA profile was created in a similar fashion with three known DNA profiles in the proportion 1:1:1. Peaks heights were adjusted where dropout was not a consideration (three person without dropout) and where dropout was expected (three person with dropout).

All profiles were interpreted using STRmix [23].

Experiment 1 – testing the effect of contributors in the same mixture proportions

Four artificial mixed DNA profiles (one two and one three contributor mixture with and without dropout) were interpreted assuming the known number of contributors. The profiles were compared with 145,470 profiles on the NZ DPD plus the profiles of the known contributors. An LR was calculated using the continuous method (LRC) described in Taylor et al. [15] for each profile from the DPD and the known contributors. Each of the individuals on the database and the known contributors were considered as a potential contributor in turn under the following two hypotheses:

H1: Database individual and N - 1 unknown contributors

H2: N unknown contributors

where N is the number of contributors under consideration. As this search is undertaken during the investigative phase, no subpopulation correction was used and the product rule was calculated. A side benefit is reduced computational effort when searching.

A population database comprising allele frequencies of the four major subpopulations within NZ in their approximate proportions as determined in the 2006 NZ Census was used to generate the LR.

The LR for all known contributors and all adventitious matches (known non-contributors) was recorded.

Experiment 2- testing the effect of overestimating the number of contributors

Eight mixed profiles consisting of DNA from known contributors were interpreted as originating from both three (correct) and four (incorrect) contributors. The profiles were compared with 145,470 profiles on the NZ DPD plus the three profiles of the known contributors and an LRC calculated as described in experiment 1 above. Each of the individuals on the database and the three known contributors were considered as a potential contributor as in experiment 1..

The LR for all known contributors and all adventitious matches (known non-contributors) was recorded. This experiment allows a determination of the LRs for the known contributors and

known non-contributors using both the correct and an incorrect number of contributors in the interpretation. It therefore examines the behaviour of the process if the number of contributors is wrongly assessed as one more than the true number.

Experiment 3 – testing the effect of low level profiles

The eight profiles were further reduced in rfu scale to 10% of the original heights and stochastic effects introduced by the random addition of rfu. This was designed to mimic extremely low level profiles. After reduction in height, all peaks were below 800 rfu with the majority under 400 rfu. All profiles appeared as having only two contributors based on allele count. The profiles were interpreted assuming both two and three contributors. In order to manage run times a 'high risk' database was created by pooling all profiles where the LRC from experiment 2 was above 100. Within the high risk database were 595 Identifiler, 742 SGMPlus and 92 SGM profiles. The proportions of the different multiplexes within the new high risk profile database were approximately the same as the original database. Each of the individuals on the new high risk database (N = 1429) and the three known contributors were considered as a potential contributor and the LRC calculated as described for experiment 1 above.

Experiment 4 - testing the effect of underestimating the number of contributors

A four donor profile was artificially constructed from the 50% reduced known three person mixture (Profile 1) used in experiment 2 by adding a fourth contributor in such a way that allele count would not indicate the presence of the fourth contributor. The fourth contributor was added at the same height as the known trace third contributor. Three additional alleles were added to the profile. Where the fourth contributor shared alleles with the known three contributors or had peaks in stutter positions these peak heights were also increased proportionally. Each of the individuals on the original database (N = 145,470) and the four known contributors were considered as a potential contributors and the LR calculated as described for experiment 1 above.

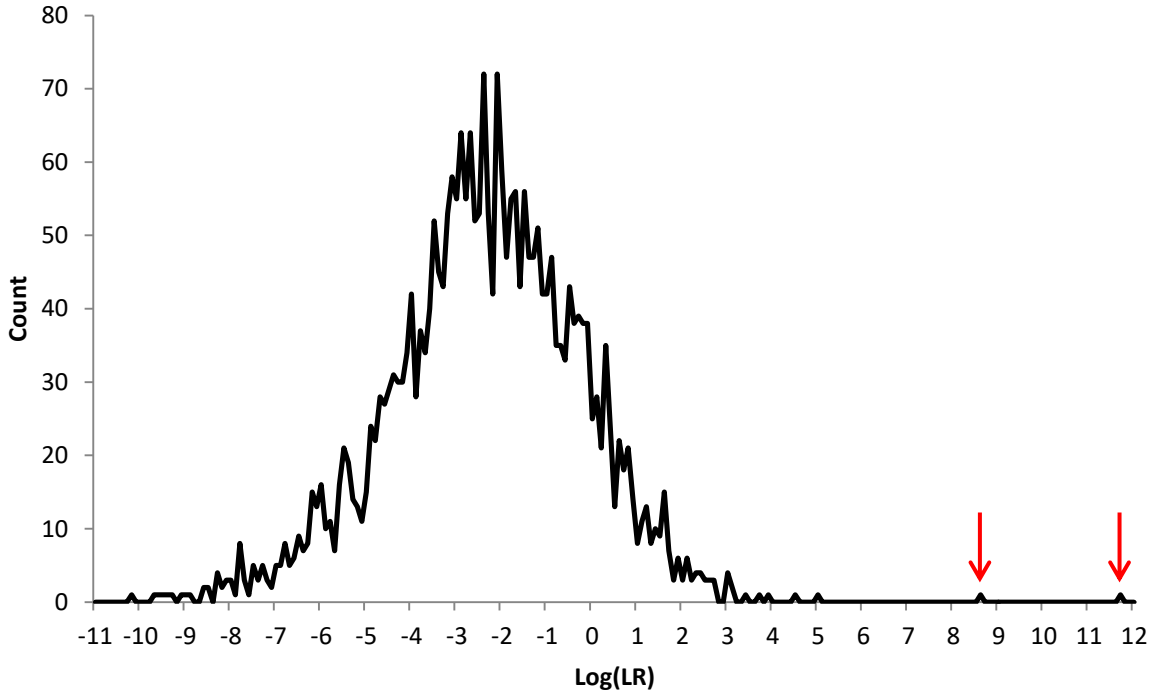
Results

Experiment 1

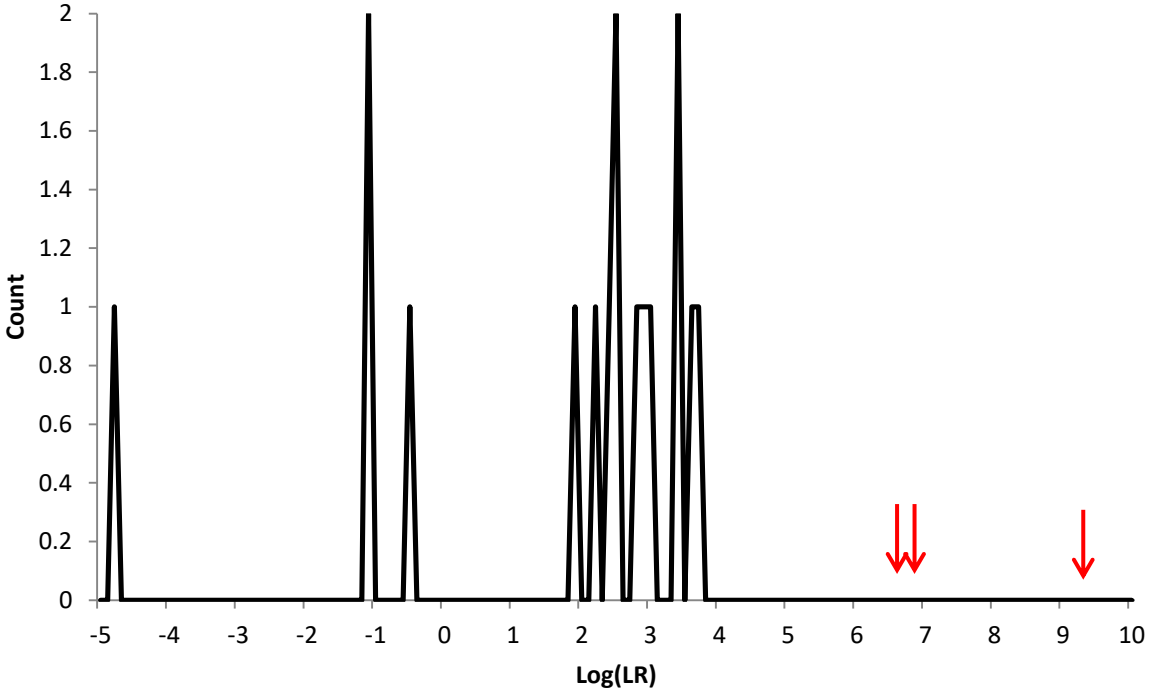
The two person artificial profile without dropout resulted in an LRC value above zero for only the two known contributors within the database. There were no adventitious matches to known non contributors. The two known contributors to the two person profile with dropout resulted in the highest LRC values. 2801 individuals within the database also provided adventitious links to this artificial mixed DNA profile, with LRC values above zero. The highest observed LRC for an adventitious match was 93,665. The counts of the log(LRC) values for all matches are provided as a summary in Figure 1, panel A.

The three person artificial profile both without and with dropout matched the three known contributors with the highest LRC as expected. The LRC values for all adventitious matches (N = 16 for no dropout and N = 111,638 for dropout) above zero are summarised Figure 1, panel B and panel C for no dropout and with dropout, respectively. The highest observed LRC for an adventitious match was 5,189 for the three person profile without dropout and 15,141 for the three person profile with dropout.

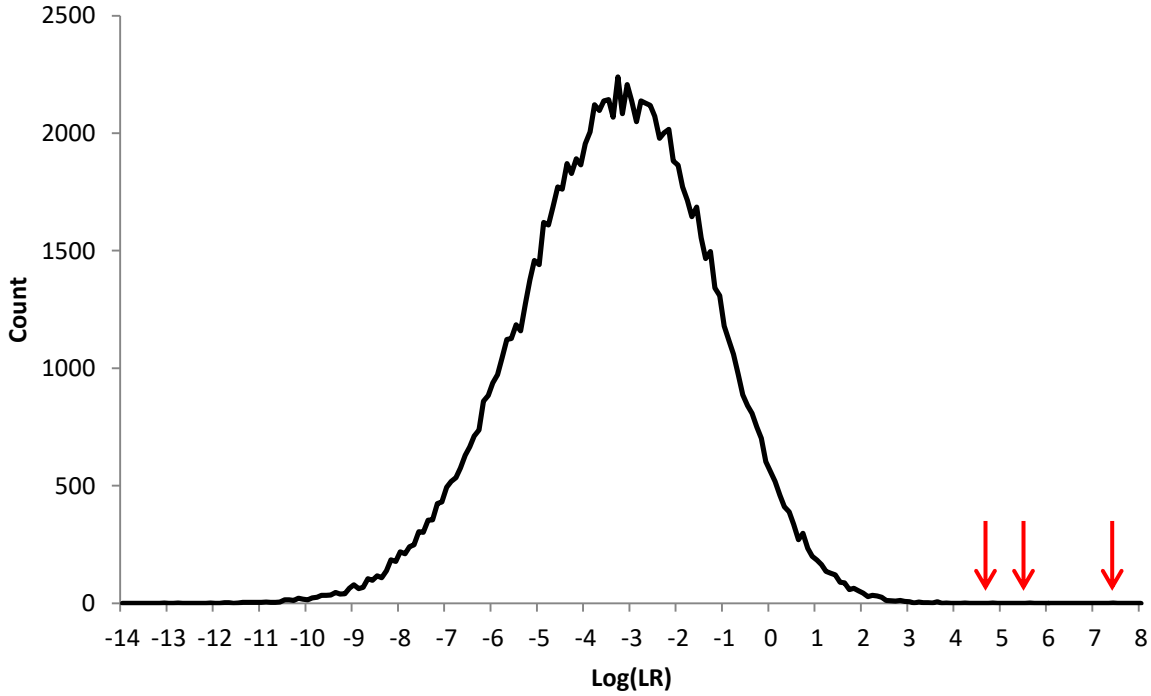
Figure 1: Summary of counts of log(LR_C) values for all adventitious matches for the two person profile with drop (panel A), three person profile without drop (panel B) and three person profile with drop (panel C). The log(LR_C) of the known contributors is indicated by arrows.



Panel A



Panel B



Panel C

Experiment 2

A representative epg from one of the eight artificial mixed DNA profiles (Profile 1) is given within the supplementary material. The LR_C values for all adventitious matches are summarised in Table 1. The incorrect assignment of a fourth contributor to the interpretation generates many more possible genotype combinations and results in a large increase in the number of low grade adventitious links (LR_C < 1,000). There was no trend observed which could be attributed to the number of contributors for adventitious links with LR_C > 1,000. The highest observed LR_C for an adventitious match with either N = 3 (correct) or N = 4 (incorrect) was 114,000 (Profile 3, interpreted incorrectly as a four person mixture).

The LR_C for each of the known contributors considered individually as potential contributors to the artificial mixtures under H_1 is given in Table 2. Interpretation of the profile incorrectly assuming four contributors had little impact on the LR_C where the known individual was a major contributor. For minor contributors however, interpretation of the profile assuming four contributors had the effect of reducing the LR_C, in some cases up to three orders of magnitude, when compared to the LR_C calculated using the true number of contributors. As expected, comparison to the interpretation assuming three contributors resulted in the highest *LR*s for all profiles.

Table 1: Count of adventitious links per profile for experiment 2, a true three person mixture interpreted assuming either three or four contributors

Profile		1		2		3		4		5		6		7		8		Total counts	
Assumed no. contributors		3	4	3	4	3	4	3	4	3	4	3	4	3	4	3	4	3	4
Ranges of LR_C	$1 - 10^1$	3,076	31,464	1,209	4,057	45	16,956	1	23,582	22	24,433	330	26,303	254	24,781	203	29,685	5,140	181,261
	10^1-10^2	960	3,036	497	164	32	2,717	31	3,319	105	3,678	287	2,850	152	2,777	826	3,845	2,890	22,386
	10^2-10^3	168	125	123	2	10	196	43	137	120	102	85	123	36	191	301	139	886	1,015
	10^3-10^4	17	2	22	0	3	18	24	1	31	0	15	2	15	22	15	4	142	49
	10^4-10^5	2	1	2	0	1	1	0	0	0	0	0	0	3	4	3	0	11	6
	10^5-10^6	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1
Total as % of database size		2.9%	23.8%	1.3%	2.9%	0.1%	13.7%	0.1%	18.6%	0.2%	19.4%	0.5%	20.1%	0.3%	19.1%	0.9%	23.1%	9,069	204,718

Table 2: LR_C for known contributors to each of the artificial mixtures, experiment 2

Sample	1		2		3		4		5		6		7		8	
Assumed no. contributors	3	4	3	4	3	4	3	4	3	4	3	4	3	4	3	4
*LR*s for the three known contributors	1.4×10^{12}	2.0×10^{12}	3.2×10^{6}	9.7×10^{3}	4.8×10^{12}	3.0×10^{12}	1.7×10^{17}	1.5×10^{17}	1.3×10^{18}	1.3×10^{18}	1.4×10^{16}	7.7×10^{16}	1.4×10^{9}	5.8×10^{7}	1.4×10^{17}	1.6×10^{17}
	4.1×10^{5}	5.8×10^{3}	2.3×10^{16}	2.3×10^{16}	1.7×10^{8}	1.2×10^{5}	2.3×10^{15}	2.4×10^{15}	1.7×10^{16}	1.8×10^{16}	7.0×10^{13}	6.4×10^{14}	2.3×10^{16}	2.3×10^{16}	1.2×10^{15}	1.3×10^{15}
	6.4×10^{13}	8.2×10^{13}	8.0×10^{15}	5.9×10^{15}	1.1×10^{14}	1.0×10^{14}	3.3×10^{5}	4.6×10^{3}	1.0×10^{5}	1.6×10^{3}	2.1×10^{6}	3.5×10^{2}	1.9×10^{16}	5.2×10^{16}	1.4×10^{5}	3.7×10^{3}

Table 3: Count of adventitious links per profile for the extreme low level profiles for experiment 3

Profile		1		2		3		4		5		6		7		8		Total counts	
Assumed no. contributors		2	3	2	3	2	3	2	3	2	3	2	3	2	3	2	3	2	3
Ranges of LR_C	>1	7	1,039	28	551	8	1,181	7	755	3	864	1	930	8	596	1	749	63	6,665
	10^1-10^2	25	143	20	216	9	40	6	111	3	64	1	0	0	328	1	122	65	1,024
	10^2-10^3	6	18	4	34	2	9	5	15	2	13	3	5	2	54	3	9	27	157
	10^3-10^4	1	5	0	1	2	2	1	5	0	6	0	1	3	7	1	5	8	32
	10^4-10^5	1	1	0	0	0	1	0	0	0	0	0	1	0	0	0	1	1	4
	10^5-10^6	1	1	0	0	0	0	0	1	0	1	1	1	0	0	1	1	3	5
Total as % of database size		2.9%	84.5%	3.6%	56.1%	1.5%	86.3%	1.3%	62.1%	0.6%	66.3%	0.4%	65.6%	0.9%	68.9%	0.5%	62.1%	167	7,887

Table 4: LR_C for known contributors to each of the extreme low level artificial mixtures, experiment 3 (trace contributor highlighted)

Sample	1		2		3		4		5		6		7		8	
Assumed no. contributors	2	3	2	3	2	3	2	3	2	3	2	3	2	3	2	3
LRs for the three known contributors	1.6×10^7	3.8×10^6	1.1	1.4	1.6×10^9	4.5×10^8	7.8×10^8	2.2×10^8	1.5×10^{13}	5.5×10^{10}	1.6×10^{11}	1.8×10^{11}	0	2.8	7.6×10^{10}	1.3×10^{11}
	0	4.6	1.6×10^{15}	1.5×10^{15}	0	2.3	9.5×10^{11}	2.1×10^{12}	4.1×10^{14}	3.4×10^{13}	1.1×10^8	4.1×10^7	1.8×10^{16}	1.7×10^{16}	9.7×10^7	4.0×10^7
	3.8×10^{11}	8.7×10^{10}	1.7×10^8	2.3×10^5	7.4×10^{11}	6.4×10^{11}	0	3.1	0	2.0	0	1.2	3.5×10^8	9.7×10^7	0	4.1

Experiment 3

A representative epg from one of the eight artificial mixed DNA profiles (Profile 1) reduced in scale by 90% is given in Figure 2, supplementary material. The LR_C values for all adventitious matches are summarised as counts in Table 3. As in experiment 2, the assumption of more contributors to the profile results in many more possible genotype combinations. Fewer adventitious matches were observed when an assumption of two contributors was made. The adventitious match with the highest LR_C (730,000) occurred when Profile 8 was interpreted as a three person mixture. More adventitious matches with high LR_C values (in the order of 10^5) were obtained for the extreme low level profiles compared to experiment 2.

The LR_C for each of the known contributors to the 90% scaled mixtures is given in Table 4. The scaling of the profiles downwards by 90% resulted in the complete dropout of the trace contributor to each profile (highlighted in Table 4). As expected, the *LR* values for the known contributors are lower than the original comparison (Table 2) because of the increased uncertainty in the profile interpretation. As in experiment 2, comparison to the known contributors resulted in the highest LR_C values for all profiles.

Experiment 4

The artificially constructed four person mixture interpreted incorrectly as a three contributor profile linked to the three known 'major' contributors with *LR*s of 1.5×10^{11}, 1.1×10^5 and 4.9×10^{13}. These *LR*s are in within one order of magnitude of the original profile interpretation results in Table 2 indicating that the introduction of a trace contributor to a profile has little effect on the interpretation of the major profiles. The LR_C for the highest adventitious match was 4,260.

The profile, when interpreted correctly as a four contributor profile, linked to the three known 'major' contributors with *LR*s of 5.5×10^{11}, 6.5×10^3 and 1.3×10^{14}. The additional fourth contributor matched to the corresponding database profile with LR_C of 11.6.

Case 1. A mixed DNA profile was obtained from a semen stain on a carpet at the scene of an alleged sexual assault involving two male offenders. DNA from most likely two contributors was detected, present in approximately equal proportions. The EPG is shown in Figure 2. The profile was interpreted assuming two contributors and searched against the DPD using a *LR* threshold of one million (10^6). The threshold was determined by rounding upwards the *LR* from the highest observed adventitious match in experiment 3. Each of the individuals on the database was considered as a potential contributor in turn under the following two hypotheses:

H_1: Database individual under consideration and one unknown contributor

H_2: Two unknown contributors

The crime profile was linked to two individuals. The profiles of the two individuals are in Table 5. Direct comparison of the individual profiles to the crime profile reveals a potential non-concordance with Contributor 2 at D18S51. On close inspection of the epg in Figure 2 a peak in the 17 allele bin is visible below the analytical threshold. Despite this non-concordance and the large imbalance at D21S11, we note that these two contributors fully explain the complete profile.

Figure 2: Epg of mixed DNA profile from a semen stain from Case 1

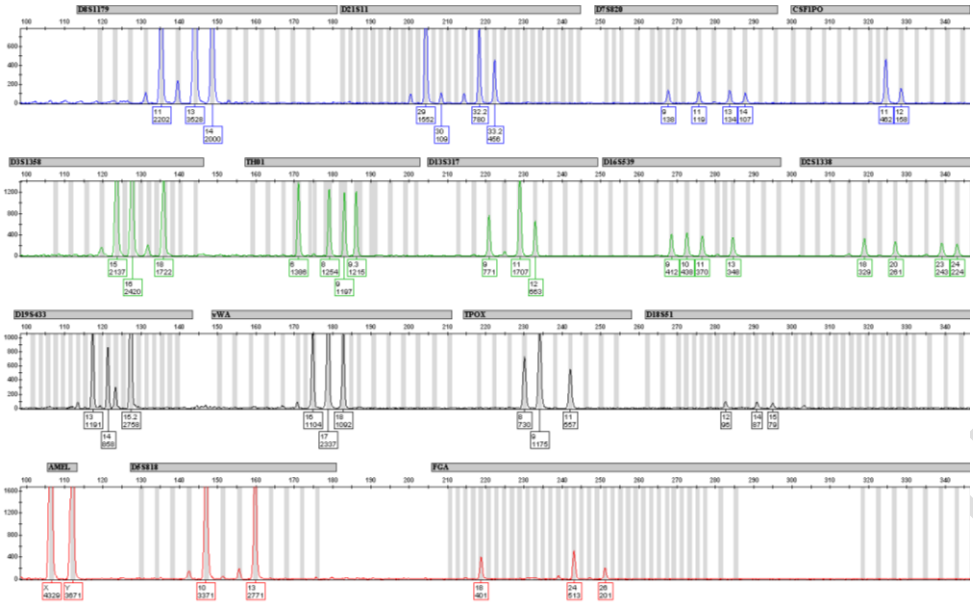


Table 5: Profiles of the two matching individuals and LR_C for Case 1

Locus	Contributor 1	Contributor 2
D8	11,14	13,13
D21	30,32.2	29,33.2
D7	13,14	9,11
CSF	11,12	11,11
D3	15,18	16,16
TH01	8,9.3	6,9
D13	11,11	9,12
D16	10,13	9,11
D2	18,24	20,23
D19	15.2,15.2	13,14
vWA	17,18	16,17
TPOX	9,11	8,9
D18	14,15	12,17
D5	10,13	10,13
FGA	24,26	18,24
LR_C	4.9×10^{13}	9.3×10^{9}

Case 2. A low level mixed DNA profile was obtained from cellular material recovered from a shoe that was located in car at the scene of an aggravated burglary. The epg of one of the replicate amplifications is shown in Figure 3. The profile was interpreted assuming three contributors based on the minimum peak count, and supported by sub-threshold peak information. Each of the individuals on the database was considered as a potential contributor in turn under the following two hypotheses:

H_1: Database individual under consideration and two unknown contributors

H_2: Three unknown contributors

The profile linked to one individual profiled using the SGMPlus multiplex on the DPD. The DNA profile of that individual and the corresponding LR_C is in Table 6.

Figure 3: Epg of the mixed DNA profile from a shoe insole from Case 2

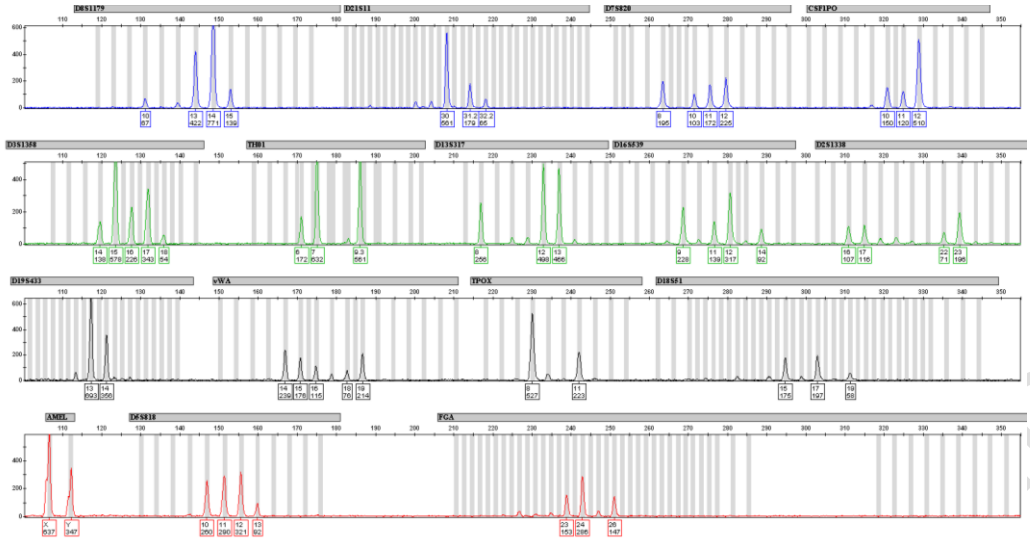


Table 6: Profile of the matching individual and LR_C for Case 2

Locus	Contributor 1
D8	14,14
D21	30,30
D7	-
CSF	-
D3	15,17
TH01	7,9.3
D13	-
D16	9,12
D2	16,23
D19	13,14
vWA	14,19
TPOX	-
D18	15,17
D5	-
FGA	24,26
LR_C	9.1×10^8

Conclusion

Direct searching of unresolved mixtures against databases of known individuals has been shown to be feasible as an investigative technique with the use of a suitable *LR* threshold to filter out low grade adventitious links. For this dataset, an appropriate LR_C threshold of approximately 1 million would ensure the risk of reporting an adventitious match is mitigated when interpreting extreme low level profiles (the majority of peaks less than 400 rfu and all peaks below 800 rfu). Complex DNA profiles with different contributors in the same proportions resulted in the highest LR_C values when the known contributors were considered individually as potential contributors under H_1, even when dropout was a consideration. The choice of a threshold is undertaken as part of a risk assessment. Setting the threshold too low risks increasing the chance of obtaining an adventitious match whereas setting the threshold too high risks missing true, legitimate matches. Table 7 shows the rate of adventitious matches (false positives) and incorrect non-matches (false negatives) that arise from using different *LR* cut off values using data from Tables 1 and 2.

Table 7: Numbers of false negative and false positive results obtained in experiment 2 using different *LR* cut off values

LR cut off	Considered a 3 person mix		Considered a 4 person mix	
	Number of false positives	Number of false negatives	Number of false positives	Number of false negatives
10^6	0	4	0	7
10^5	0	0	1	6
10^4	11	0	7	6
10^3	153	0	56	1
10^2	1,039	0	1,071	0
10^1	3,929	0	23,457	0
1	9,069	0	204,718	0

Regardless of where the search threshold is set there will always be the possibility of false positive and false negative results. There is a limited capability to identify a true contributor if they are a trace contributor to a complex mixture without also flagging a large number of false positive links. This is also true if there is substantial dropout of an individual's alleles or if a minor contributor's alleles are masked by a major contributor's alleles within a mixed DNA profile. This information can be used to help form guidelines in order to limit the numbers of mixed DNA profiles searched against a database to those that have the greatest potential to provide strong investigative leads.

The assumption of additional contributors to a profile beyond those suggested by allele count alone tended to lower the LR_C for the true minor and major contributors and increase the number of low grade adventitious links, where $1 < LR_C < 1{,}000$. A match against the database is unlikely for a trace contributor that has very few alleles either present above the analytical threshold and present in non-masked allele positions. This is the expected outcome.

The assumption of additional contributors also resulted in significantly increased computational effort.

The multiplex used to determine the genotype for the known database profile did not appear to have an effect on whether an adventitious link was made. This was evident from the make-up of the high risk profile database where the profile multiplexes were in the same appropriate proportions of the original database.

Two case examples are described where profiles that were considered previously unsuitable for database comparison were interpreted and searched against the NZ DPD with a LR_C threshold of 1 million. Both cases resulted in links to individuals with high LR_C values. It is worth cautioning the reader that, as with any links resulting from a database search, their primary purpose in investigative only and further scrutiny is warranted.

This work reinforces the power of DNA databases as investigative tools and demonstrates the ability to directly search mixed DNA profiles using a *LR* framework without the need to identify a single contributor profile. Even using fully continuous methods of interpretation an individual's profile may not be able to be reasonably inferred from a complex DNA mixture. This would make the profile unsuitable for entry to a database using traditional database search methods. The searching method proposed in this paper allows real time searching of complex mixed DNA profiles. Using an *LR* strategy is a more powerful method than counting matching alleles, for example, and allows phenomena such as drop in and dropout to be taken into account. The functionality is available on the NZ DPD where

average search times are 10 minutes against a database of over 145,000 profiles.

Acknowledgements

This work was supported in part by grant 2011-DN-BX-K541 from the US National Institute of Justice. Points of view in this document are those of the authors and do not necessarily represent the official position or policies of the U.S. Department of Justice. The authors thank Catherine McGovern and Johanna Veth (ESR) and two anonymous reviewers whose helpful comments greatly improved this paper.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.fsigen.2013.12.001.

References

[1] S. J. Walsh, J. S. Buckleton. DNA Intelligence Databases. In: Buckleton JS, Triggs MC, Walsh SJ, editors. Forensic DNA evidence interpretation. Boca Raton: CRC Press; 2005.

[2] L. Mueller, Can simple population genetic models reconcile partial match frequencies observed in large forensic databases?, Journal of Genetics. 87(2) (2008).

[3] J. M. Butler. DNA databases: uses and issues. In: Butler JM, editor. Advanced topics in forensic DNA typing: methodology: Elsevier; 2012.

[4] P. Gill, C. H. Brenner, J. S. Buckleton, A. Carracedo, M. Krawczak, W. R. Mayr, et al., DNA commission of the International Society of Forensic Genetics: Recommendations on the interpretation of mixtures, Forensic Science International. 160(2–3) (2006) 90-101.

[5] B. Caddy, G. Taylor, A. Linacre. A review of the science of low template DNA analysis.; 2008 [cited 14th April 2008]. Available from: http://police.homeoffice.gov.uk/news-and-publications/publication/operational-policing/Review_of_Low_Template_DNA_1.pdf?view=Binary

[6] The Forensic Science Regulator. Response to Professor Brian Caddy's Review of the Science of Low Template DNA Analysis; 2008 7 May 2008 [cited 12 May 2008]. Available from: http://police.homeoffice.gov.uk/news-and-publications/publication/operational-policing/Review_of_Low_Template_DNA_1.pdf?view=Binary

[7] J.-A. Bright, D. Taylor, J. M. Curran, J. S. Buckleton, Developing allelic and stutter peak height models for a continuous method of DNA interpretation, Forensic Science International: Genetics. 7(2) (2013) 296-304.

[8] H. Haned, L. Pène, J. R. Lobry, A. B. Dufour, D. Pontier, Estimating the number of contributors to forensic DNA mixtures: Does maximum likelihood perform better than maximum allele count?, Journal of Forensic Sciences. 56(1) (2011) 23-28.

[9] A. Biedermann, S. Bozza, K. Konis, F. Taroni, Inference about the number of contributors to a DNA mixture: Comparative analyses of a Bayesian network approach and the maximum allele count method, Forensic Science International: Genetics. 6(6) (2012) 689-696.

[10] B. Budowle, A. J. Onorato, T. F. Callaghan, A. D. Manna, A. M. Gross, R. A. Guerrieri, et al., Mixture interpretation: defining the relevant features for guidelines for the assessment of mixed dna profiles in forensic casework, Journal of Forensic Sciences. 54(4) (2009) 810-821.

[11] J.-A. Bright, P. Gill, J. Buckleton, Composite profiles in DNA analysis, Forensic Science International: Genetics. 6(3) (2012) 317-321.

[12] H. Kelly, J.-A. Bright, J. Curran, J. Buckleton, The interpretation of low level DNA mixtures, Forensic Science International: Genetics. 6(2) (2012) 191-197.

[13] T. M. Clayton, J. S. Buckleton. Mixtures. In: Buckleton JS, Triggs CM, Walsh SJ, editors. Forensic DNA Evidence Interpretation. Boca Raton, Florida: CRC Press; 2004. p. 217-274.

[14] H. Kelly, J. A. Bright, J. S. Buckleton, J. M. Curran, A comparison of statistical models for the analysis of complex forensic DNA profiles, Science and Justice. http://www.scopus.com/inward/record.url?eid=2-s2.0-84881142115&partnerID=40&md5=48ba7495cff95503c481dde967ae91d3 (2013).

[15] D. Taylor, J. A. Bright, J. S. Buckleton, The interpretation of single source and mixed DNA profiles, Forensic Science International: Genetics. 7 (2013) 516-528.

[16] J. A. Bright, D. Taylor, J. M. Curran, J. S. Buckleton, Degradation of forensic DNA profiles, Australian Journal of Forensic Sciences. DOI: 1080/00450618.2013.772235 (2013).

[17] M. W. Perlin, M. M. Legler, C. E. Spencer, J. L. Smith, W. P. Allan, J. L. Belrose, et al., Validating TrueAllele® DNA Mixture Interpretation, Journal of Forensic Sciences. 56(6) (2011) 1430-1447.

[18] R. Puch-Solis, L. Rodgers, A. Mazumder, S. Pope, I. Evett, J. Curran, et al., Evaluating forensic DNA profiles using peak heights, allowing for multiple donors, allelic dropout and stutters, Forensic Science International: Genetics. 7(5) (2013) 555-563.

[19] J. S. Buckleton, J. M. Curran, P. Gill, Towards understanding the effect of uncertainty in the number of contributors to DNA stains, Forensic Science International: Genetics. 1(1) (2007) 20-28.

[20] S. A. Harbison, J. F. Hamilton, S. J. Walsh, The New Zealand DNA databank: its development and significance as a crime solving tool, Science and Justice. 41 (2001) 33-37.

[21] J.-A. Bright, J. S. Buckleton, C. E. McGovern, Allele frequencies for the four major sub-populations of New Zealand for the 15 Identifiler loci, Forensic Science International: Genetics. 4(2) (2010) e65-e66.

[22] Applied Biosystems, AmpFlSTR Identifiler Amplification Kit User Guide. Foster City, CA: Applied Biosystems; 2009.

[23] D. Taylor, J.-A. Bright, J. S. Buckleton, The interpretation of single source and mixed DNA profiles Forensic Science International: Genetics. 7(5) (2013) 516-528.