Forensic Science International: Genetics 11 (2014) 56–63

Contents lists available at ScienceDirect

Forensic Science International: Genetics

journal homepage: www.elsevier.com/locate/fsig

An illustration of the effect of various sources of uncertainty on DNA likelihood ratio calculations

D. Taylor [a,*], J-A. Bright [b], J. Buckleton [b], J. Curran [c]

[a] Forensic Science South Australia, 21 Divett Place, SA 5000, Australia
[b] ESR Ltd, Private Bag 92021, Auckland 1142, New Zealand
[c] Department of Statistics, University of Auckland, Private Bag 92019, Auckland 1142, New Zealand

ARTICLE INFO

Article history:
Received 29 August 2013
Received in revised form 28 January 2014
Accepted 3 February 2014

Keywords:
DNA interpretation
Sampling uncertainty
MCMC
HPD
Continuous methods
Relatives

ABSTRACT

A typical assessment of the strength of forensic DNA evidence is based on a population genetic model and estimated allele frequencies determined from a population database. Some experts provide a confidence or credible interval which takes into account the sampling variation inherent in deriving these estimates from only a sample of a total population. This interval is given in conjunction with the statistic of interest, be it a likelihood ratio (LR), match probability, or cumulative probability of inclusion. Bayesian methods of addressing database sampling variation produce a distribution for the statistic from which the bound(s) of the desired interval can be determined.

Population database sampling uncertainty represents only one of the sources of uncertainty that affects estimation of the strength of DNA evidence. There are other uncertainties which can potentially have a much larger effect on the statistic such as, those inherent in the value of F_{st}, the weights given to genotype combinations in a continuous interpretation model, and the composition of the relevant population. In this paper we model the effect of each of these sources of uncertainty on a likelihood ratio (LR) calculation and demonstrate how changes in the distribution of these parameters affect the reported value. In addition, we illustrate the impact the different approaches of accounting for sampling uncertainties has on the LR for a four person mixture.

1. Introduction

The likelihood ratio is the recommended statistic for the calculation of the weight of forensic DNA evidence [1]. It is the ratio of the probability of the observed evidence DNA profile (O) given each of two competing hypotheses, H_1 and H_2.

Traditional methods of DNA profile interpretation are described as binary. In a binary interpretation system weights of zero or one are used to either exclude or include genotype sets respectively. The weights represent a relative assignment of the probability density of the observed profile if it is from the proposed genotype combination. Hereafter we will refer to a probability density as a probability for simplicity and because the difference, although important, is not required here.

This assignment of relative probability is often guided by a set of heuristics that may include heterozygous balance, dropout, and mixture proportion [2]. A continuous interpretation model uses the quantitative information from an electropherogram such as peak heights, to calculate the probability of the peak heights given all possible genotype set combinations, S_j. A weight, w_j, can be defined as the normalised probability density of the observed evidence data (O) given the proposed genotype set combination, $\Pr(O|S_j)$.

Weight is a relatively new term for a concept that has been in use in DNA profile interpretation for some time. The variation in these weights assigned using a binary method of interpretation is difficult to quantify, as any variation will be from the differences arising in the interpretations between two, or more, analysts. Advances in research along with access to increased computer resource, have given practicing forensic scientists access to software which generate and apply continuous (as opposed to binary) weights. These weights are often estimated by Markov chain Monte Carlo (MCMC) methods [3–5] utilising peak height information and models of DNA profile behaviours. Continuous methods allow weights to be assigned any value between zero and one. In the frequentist paradigm, the weights are regarded as having a fixed, but unknown value. A reasonable frequentist procedure looks to use the data to provide an estimate of the weights along with the associated uncertainty in the estimates. In the Bayesian school of thought, the weights are regarded as random, with their behaviour described by a statistical distribution. Under either framework it is typical to only use the average weights when calculating the LR. However, in reality the weights

* Corresponding author. Tel.: +61 8 8226 7700; fax: +61 8 8226 7777.
E-mail address: Duncan.Taylor@sa.gov.au (D. Taylor).

1872-4973/$ – see front matter.
http://dx.doi.org/10.1016/j.fsigen.2014.02.003

are unknown, and it is useful to consider the effect of this uncertainty on the resulting LR calculations.

Some forensic practitioners calculate a credible interval (CI) or confidence interval that accounts for the sampling variation inherent in allele frequency estimation from a sample of a population of interest; namely a population database. Not all commentators believe that an assessment of sampling error is necessary. Brenner [6] makes explicit his doubts about the usefulness of assessing sampling uncertainty with the following challenge:

> "Will someone tell me, please, what rational difference it ever can make to know the confidence limits in addition to knowing the best point estimate? Specifically, can you give premises under which, for a fixed point estimate, the decision to convict or not to convict would depend on the size of the confidence interval?"

There is a lot of substance to Brenner's challenge. However these comments may not have taken full account of the cross examination process where any uncertainty or doubt is often explored at length. An analyst who has prepared for such a cross examination will likely be of more assistance to the court than one who chooses to answer "would it make any difference?" Furthermore, and perhaps more importantly, in our experience it is an accepted practice in adversarial systems that all reasonable uncertainty is conceded to the defendant.

Commenting on statistical evidence in general, rather than DNA in particular, Good stated [7]:

> "The court expects us to provide both an average based on our sample and some measure of the accuracy of our average."

Almost any measurement in science has an associated measure of uncertainty. Well prepared lawyers correctly investigate this avenue of questioning. In our experience, this typically has been approached by asking a question along the lines: "Is your database of 180 individuals big enough?"

Is there any reason why DNA evidence should be exempt from this line of questioning? The position advocating a consideration of sampling uncertainty is also taken by many authors [8–14]. In most cases, even with the inclusion of an estimate of sampling uncertainty, the final answer is not vastly different to the point or 'best' estimate.

One method for calculating a CI is the highest posterior density (HPD) [14–16]. The HPD method allows the calculation of interval bounds and uses a Dirichlet distribution to describe the variation in allele frequencies estimated from a population database. The HPD interval bounds are evaluated using a Monte Carlo method. In this method a large random sample is taken from the posterior distribution of the LR and the empirical sample quantiles are used as estimates of the bounds.

Whilst the HPD method has been applied to forensic LR calculations with respect to the variability inherent in allele frequency estimates, it can also readily account for the variance of other parameters impacting on the LR, namely F_{st} (commonly called theta, or the coancestry coefficient), genotype set weights and population composition. We will investigate the effect of each of these factors of uncertainty on the LR distribution. There are other sources of uncertainty which we do not investigate in this work but should still be recognised, such as the number of contributors, the biological models underpinning the statistic and the potential for errors in the generation of the observed data.

The uncertainty in selecting an appropriate value of F_{st} has long been recognised. It is the authors' experience that a commonly used approach to this has been to assign a value believed to be at the top end of the plausible range.

Another matter of significant importance is the presence of relatives as potential alternative donor(s) of the DNA. [17] Traditionally, this matter is subsumed in the formation of the propositions being considered e.g.: *The DNA profile has originated from an unrelated individual from a certain population.* A less common approach is to produce two (or more) LRs; one considering the proposition that the donor is unrelated to the person of interest (POI) and one for the proposition that the donor is a relative of the POI.

A potential solution has been known for some time and is termed the 'unifying formula' [18–20]:

$$LR = \frac{Pr(O|H_1)}{\sum_i Pr(O|H_i)Pr(H_i|H_d)} \quad (1)$$

where typically i is the ith person (related or unrelated) under consideration. H_i is then the proposition that the ith person is the source of the DNA and H_1 is the proposition that the POI is the source of the DNA. In practice, this unifying formula cannot be implemented as when taken to its extreme a different hypothesis is generated for everyone on Earth. Hence i ranges from zero up to the size of the global population, i.e., encompassing every individual in every population. A plausible simplification is to change the meaning of i, firstly to be considered within one population at a time and secondly to be a relationship group, for example individuals whose relationship to the POI is unrelated, parent, child, sibling, cousin, etc. We term this method the 'unified method'. Making the described simplification, the prior may be assigned as the probability of someone in the population being related to the POI with relationship type i. These proportions can be reasonably estimated by making assumptions about population and family structure (see appendix A). Alternatively there may be additional evidence that informs the prior that, say, a brother is the donor of the profile.

A second method that could be used to account for population composition is to generate LRs for each relationship in the proportions that those relationships exist within the population, as estimated from available census data. We term this second method the 'picking method'. The picking method will likely produce an LR distribution over a wider range of values than the unifying method, with a heavy left tail attributed to relatives.

Whilst much is known about the effect of allele frequency variation on the LR distribution the combined effect of these additional sources of uncertainty has not previously been investigated. In this paper we report the effect of each of these sources of uncertainty on a likelihood ratio (LR) calculation and demonstrate how these sources would affect a reported value. The incorporation of these factors into the LR allows the scientist to report a CI using statements such as "*I am 99% sure that the true LR is greater than X*", or "*the LR is above X with 99% probability*". It also removes the need to stipulate that the alternative donor is unrelated when forming the propositions.

Statisticians, and other scientists, commonly use sensitivity analyses to understand how the behaviour of a system, or a model, changes with respect to changes in the inputs. In the forensic DNA interpretation context, parameters are varied over a plausible range and a number of LRs are produced. A statement is then made based on these different numbers. The similarity of this approach to the one described, we hope, will become obvious.

2. Method

A single source AmpF*l*STR® ProfilerPlus™ profile with an unambiguous genotype ($w_i = 1$) at all loci except one was artificially created to minimise the number of variables. The one ambiguous locus had a single peak at a height where dropout, although unlikely, was permitted. At this the locus the designation was *a, Q* where the *Q* allele could be any allele other than *a*.

Empirical distributions of the logarithm base 10 of the LR, $\log_{10}$(LR), distributions were generated from 10,000 HPD iterations, drawing from prior distributions of the parameter of interest in each analysis. The $\log_{10}$(LR) distributions are used for plotting Figures. LR values are calculated using the sub-population model of [21] using propositions:

H_1: The POI is the source of the DNA.

H_2: Someone other than the POI is the source of the DNA.

2.1. Considering only one factor at a time

Throughout this paper we explore the effects of each factor of variation on the LR, holding all other factors constant. To hold allele frequencies constant (in the assessment of all factors of uncertainty other than allele frequency) the database size was inflated to 1 billion (again artificially in the calculation) to effectively remove allele frequency variation. To hold F_{st} constant we use point values in LR calculations that correspond to the mean of the Beta distribution being used to model F_{st} distribution. To hold weights constant we use point values of 0.995 and 0.005 for the locus which is not completely resolved. In the case of population composition, when we wish to remove the effect of population composition on the LR we consider all members of the population as unrelated.

3. Results

3.1. Allele frequencies

Allele probabilities were modelled with a Dirichlet distribution [14] and evaluated by resampling allele counts from gamma distributions and renormalising at each iteration as described in Taylor et al. [3]. Allele frequency estimates from a pan-Australian Caucasian database [22] were used for all simulations. The size of the database was artificially changed to the values of 50, 100, 500, 1000 and 10,000 people and the LR distribution plotted for each database size Fig. 1.

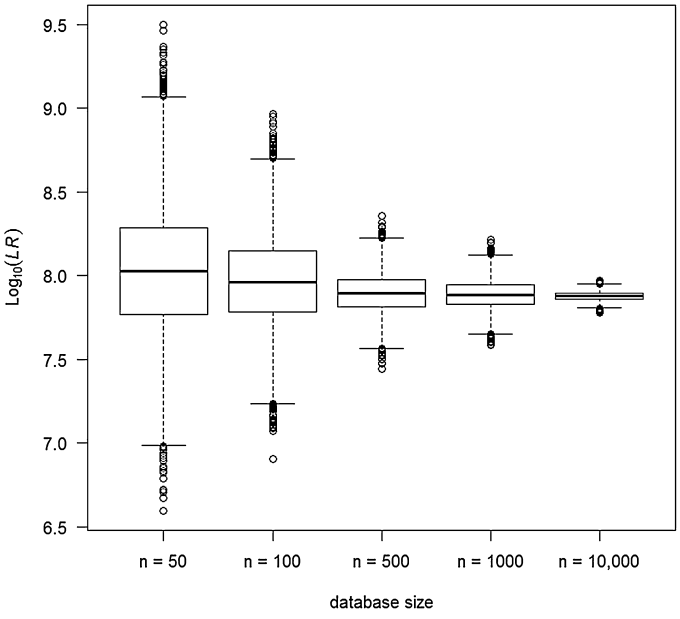

Fig. 1. Log_{10}(LR) distributions calculated using differing database sizes.

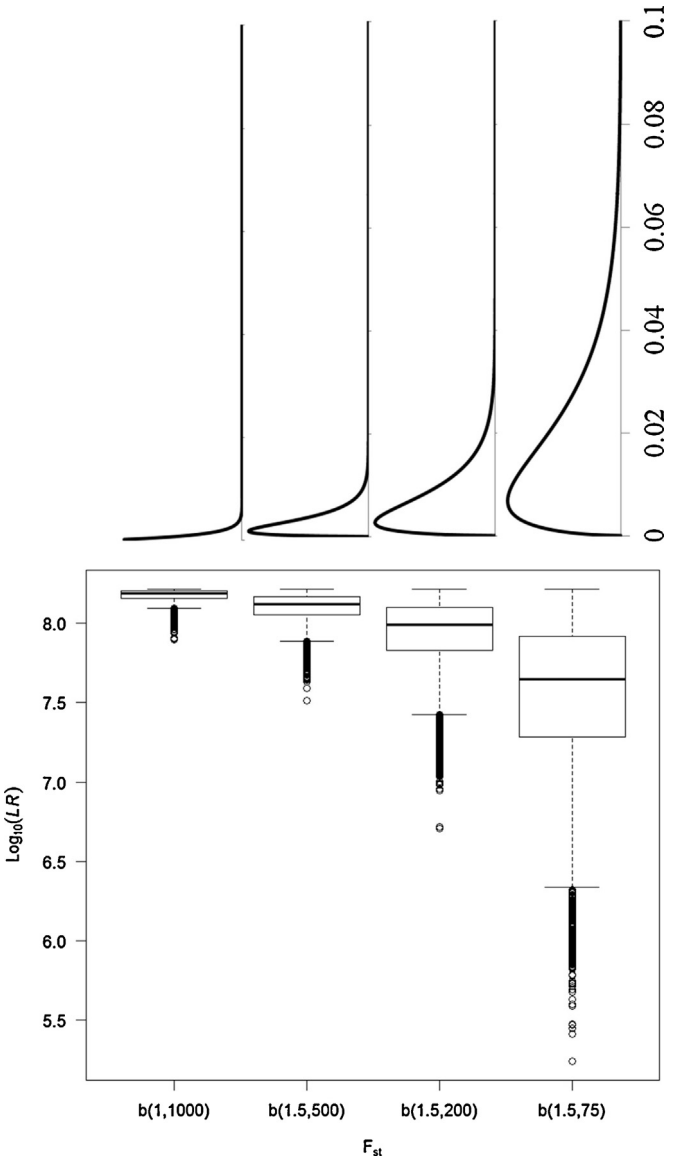

Fig. 2. Log_{10}(LR) distributions using four different F_{st} distributions (shown above the boxplot categories).

3.2. F_{st}

The probability that a pair of alleles one taken from each of two people in the population are identical by descent, IBD, will be different for each pair of two individuals. Hence F_{st} within a population may be usefully thought of as having a distribution. This distributions is likely to be asymmetrical, positively skewed, and is constrained to lie between zero and one. The Beta family of distributions have these properties, and so were selected to model the behaviour of F_{st}. Typical values of F_{st} used for forensic LR calculations tend to range between 0 and 5% [23]. We test a number of F_{st} distributions representing differing levels of co-ancestry. Fig. 2 shows the distribution of LRs and the F_{st} distribution used in the calculation.

3.3. Genotype set weights

The weights, w_i, for the genotype set combinations, S_j, were calculated using MCMC as described in Taylor et al. [3]. These w_i values correspond to counts of MCMC iterations where the specified genotype set has been the focus of the MCMC. If the

analysis was repeated, then these counts would be slightly different but would make up approximately the same proportion of total iterations. The variability in counts that make up the weights can be modelled using a gamma distribution as they are likely to be asymmetrical and the counts exist over the range (O, X) where X is the number of iterations in the MCMC. A low weight indicates that the specified genotype was the focus of the MCMC less often (attracted a low proportion of the total iterations) and therefore the relative run-to-run variation is expected to be greater.

It is recognised that successive MCMC iterations are not independent samples from the posterior. The effective sample size (corresponding to the number of independent MCMC iterations) was calculated using the package coda [24] in R [25]. The effective sample size (ESS) was then used to generate effective counts (EC) from the genotype set weights. The EC values were drawn from a gamma distribution $\Gamma(EC, 1)$ and normalised back into weights for use in the LR in each Monte Carlo iteration of the HPD calculation.

The variation in w_i will not only depend on the size of w_i but also on the total number of iterations the MCMC has been allowed to run. An obvious solution to difficulties caused by weight variation is to run each MCMC analysis for an extended number of iterations, making the EC so large that the MCMC run-to-run variation becomes insignificant. This approach may be impractical within a case-working forensic laboratory, where limited computation power may be expected and rapid case turnaround times are required. We test a number of weights with a fixed ESS of 100,000 (Fig. 3) and a number of ESS values with a fixed weight of 0.01 (Fig. 4).

3.4. Fraction of population related to the person of interest

The fraction of population related to the POI was taken into account using two different methods. We refer to the first method as the 'picking method', and the second as the 'unified method' [26]. The fraction of total population related to the POI was assigned using a number of simplifying assumptions about population behaviour that allow numbers of relatives of the POI in a population of size P to be directly related to the average number of children (n) had by couples within the population. Two populations were created, one with a high fraction of relatives and one with a low fraction. We place the details of this approximation

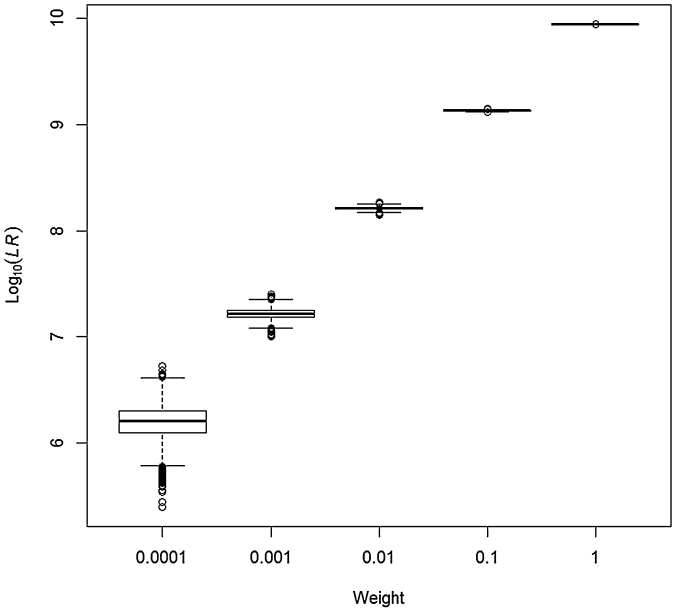

Fig. 3. $\text{Log}_{10}(\text{LR})$ distributions using ESS of 100,000 and different weights.

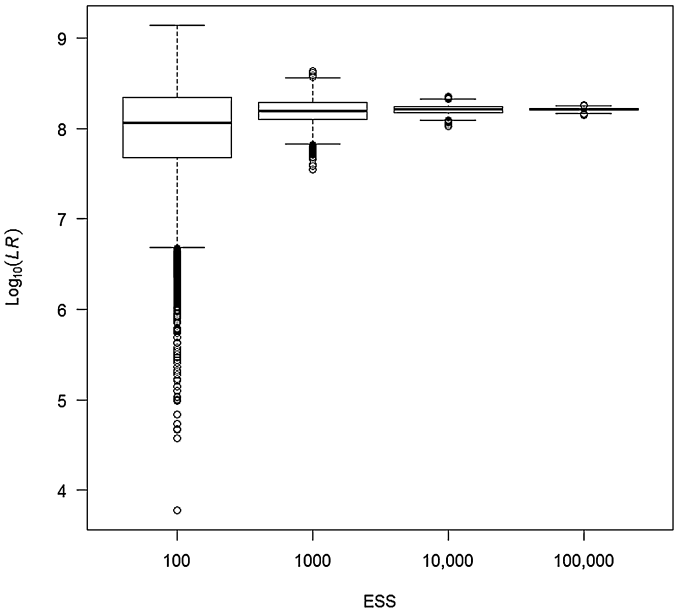

Fig. 4. $\text{Log}_{10}(\text{LR})$ distributions using a weight of 0.01 and differing ESS values.

in Appendix A. These fractions are used as the priors in Eq. (1) which models the situation of no additional information pointing towards or away from any particular relative.

3.5. Picking method

Using the proportion of the total population that the specified relationship comprises, an LR was calculated which considers either a related or unrelated person to the POI in H_i i.e:

$$\text{LR}_i = \frac{\Pr(O|H_1)}{\Pr(O|H_i)} \tag{2}$$

For I Monte Carlo iterations there will be $I \times r_i$ LRs considering a person as the source of DNA in H_i who has a relationship i to the POI. Where r_i is the proportion of the population with relationship i. Note that in Eq. (2) the LR is considering only one type of relationship of the true donor to the POI in the denominator, whereas in Eq. (1) the LR stratified the denominator across all possible relationships.

The picking method represents a discreet equivalent of the continuous methods of accounting for uncertainty outlined in the paper, i.e. a relationship type is drawn from a discreet prior distribution of relationship types and then the LR is calculated given that chosen relationship. To the authors' knowledge this method of accounting for a population's relatedness to a POI has not previously been explored.

3.6. Unified method

Again using the proportion of the total population that the specified relationship comprises, an LR was calculated weighted by the relationship types.

$$\text{LR} = \left(\sum_i \frac{r_i}{\text{LR}_i}\right)^{-1}$$

Where LR_i is the LR considering a relative of type i as the source of DNA in H_2.

Fig. 5 shows two LR distributions for two different population structures. The first represents a large population size (1 million) where the number of children per family is low (2) and the second

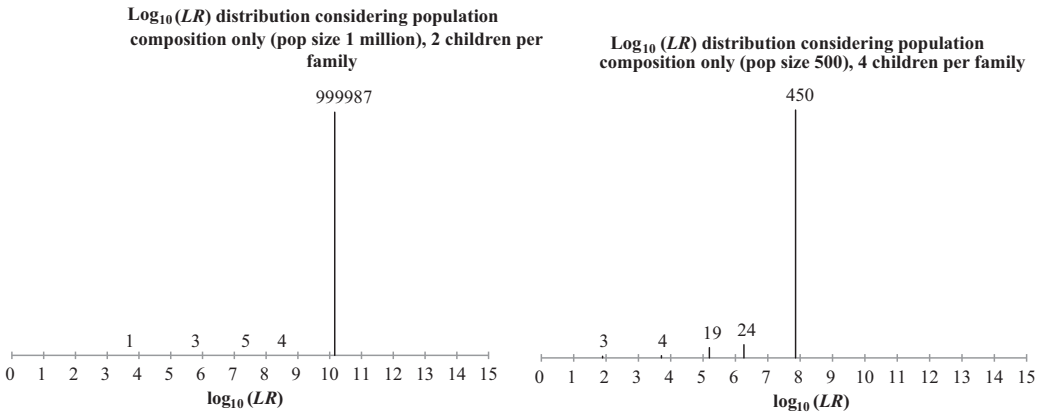

Fig. 5. Log_{10}(LR) distributions produced considering related or unrelated individuals (to the POI) as the source of DNA for two different population compositions.

represents a small population size (500) where the number of children per family is much higher (4). In fig. 5 the position of the vertical lines on the *x*-axis represents several LR values that can be obtained by considering someone other than the POI as the source of the DNA (related or unrelated). The height of the vertical lines represents the proportion of the population that would generate that LR, with the exact number given.

3.7. All uncertainty

Two extremes were considered for each parameter. These extremes represented a low and a high variance for that parameter, so that the range of effects on the LR distribution could be observed. Two ethnic groups were chosen to use as examples: Australian Caucasian [22] and Australian Aboriginal [27]. Assigned F_{st} values for these populations are 0.01 and 0.05 respectively [28–31], representing their markedly different levels of co-ancestry. A Beta distribution was fitted to the F_{st} values using maximum likelihood estimation. The estimated parameters for the Beta distributions are: Beta(0.98, 567.2) for the Caucasian population (low variation scenario), and Beta(0.3, 32.7) for the Aboriginal population (high variation scenario). Using these values, the associated distributions have respective means at 0.0017 and 0.0090, which are markedly lower than the conservative values of 0.01 and 0.05 typically in use.

The low variation scenario had the following properties:

- A large database size (N = 18,116 individuals [22]).
- A low level of co-ancestry where the range of plausible F_{st} values is low (equivalent to an Australian Caucasian population).
- A POI reference that matched at all unambiguous loci and was *homozygous* at the ambiguous locus.
- A large population size with few relatives in it.

The high variation scenario had the following properties:

- A small database size (N = 50, using frequencies from [22] but with N artificially changed in the calculation).
- A higher level of co-ancestry where the range of plausible F_{st} values is high (equivalent to an Australian Aboriginal population).
- A POI reference that matched at all unambiguous loci and was *heterozygous* at the ambiguous locus.
- A small population size with many relatives in it.

The distribution of LRs for each scenario is shown in Fig. 6.

Fig. 7 shows the same scenarios as seen in Fig. 6, but using the unifying method for familial relationships.

3.8. Example

To demonstrate the effect of accounting for multiple sources of uncertainty within the LR, an artificially constructed four person mixture was analysed under two scenarios; firstly with only sampling variation taken into account and then with all variation sources included. The artificial profile had one major and three roughly equal minor contributors. In this example, a POI reference profile corresponding to one of the minor contributors was compared to the mixed profile to generate a LR distribution. The LR distribution generated under the first scenario (as per current

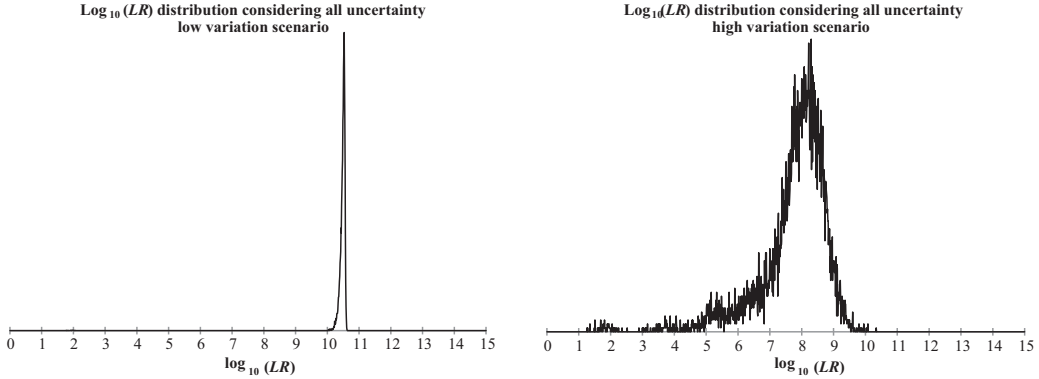

Fig. 6. Log_{10}(LR) distributions of the low and high variation scenarios using the picking method for familial relationships.

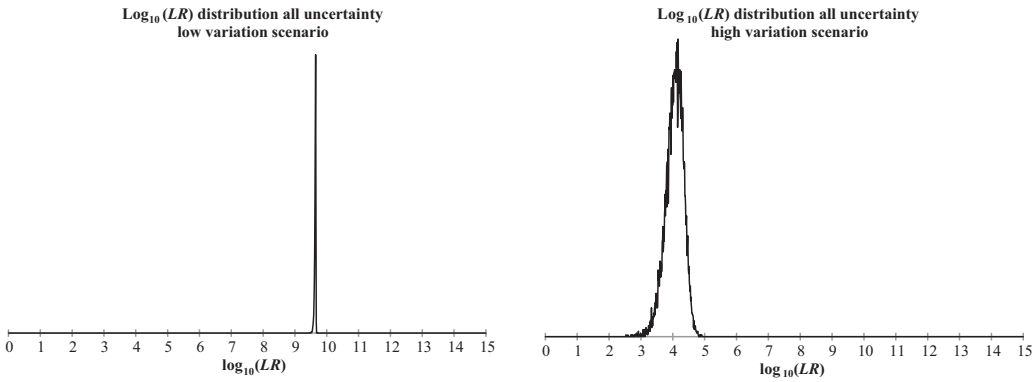

Fig. 7. Log_{10}(LR) distributions of the low and high variation scenarios using the unifying method for familial relationships.

forensic practice) was compared to the distribution created under the second scenario. The following properties were chosen to reflect a real casework situation:

- A South Australian Aboriginal allele frequency database, $N = 325$ individuals [27]
- An F_{st} distribution of Beta(0.3, 32.7) to correspond to an Australian Aboriginal population
- A population size of $P = 26{,}000$ [32] with $n = 3$ children per couple
- The unifying method to account for relatives of the POI in the population
- H_1: The mixture represents DNA from the POI and three unknown individuals, H_2: The mixture represents DNA from four unknown individuals
- An MCMC effective sample size of 772

The comparison of the two analyses can be seen in Fig. 8.

Inspection of Fig. 8B shows that including all uncertainty in the estimation of the LR pushes the lower end of the distribution below zero, into the range where the alternate hypothesis is supported. Note that the effect of the F_{st} on the LR distribution is more prominent given the accumulative effect that F_{st} has across genotype probabilities that contain three or four unknowns. The result is that whilst the LR distribution in Fig. 8B is more negatively skewed, its mode is pushed to the right of where it was in 8 A. Note that a graph similar to Fig. 8B has not been produced using the picking method to account for population relatedness to the POI. The reason for this is that given the population size and average number of children per family, the number relatives of the POI expected in the population is small, and there would be no visual (and very little numerical) difference to the Figure already produced. The reported LR taking into account all uncertainty parameters (Fig. 8B) will be either above or below 1 depending on whether the point estimate or a CI is chosen. The exact reportable values for the results seen in Figs. 6, 7 and 8 can be seen below in Table 1.

4. Discussion and conclusions

Our work shows that any of the sources of variation outlined in this study could dominate the total variation in the LR distribution under the right circumstances. For example, uncertainty in genotype weight distributions can dominate the LR variation if one or more weights for relevant genotype sets are low, or are in combination with inadequate MCMC iterations and the calculation performed uses a large population database. However if the weights are high for relevant loci and the MCMC is run for many iterations, but a small population database is used then the allele frequency distribution will dominate as the LR variation. In most cases some combination within the ranges tested in this study will be present. In combination, all factors of uncertainty will interact to produce an LR distribution that is wider than if any individual component is considered in isolation.

Although many sources of variation will exhibit interactions, some interactions of interest are worth mentioning with regards to F_{st}. Consideration of the Balding and Nichols [21] sub-population formula suggests the F_{st} variation will have a more pronounced effect on the LR distribution as more unknowns are considered in the genotype probabilities. It is also expected that F_{st} distribution will have a larger effect on *LR* distribution when considering rare genotypes due to the application of F_{st} in the sub-population

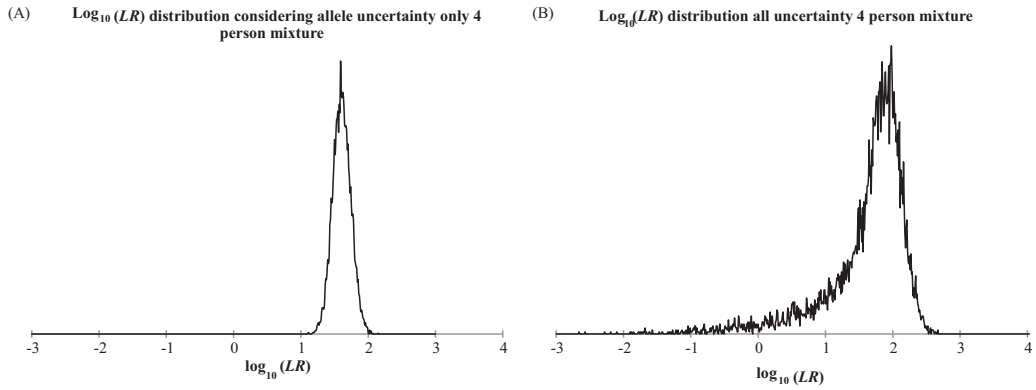

Fig. 8. Log_{10}(LR) distributions accounting for (A) allele frequency variation only (using mean F_{st} value) and (B) all variation parameters.

Table 1
Reportable LR values showing point estimate, 99% 1-sided credible interval (CI) considering only allele frequency variation (using mean F_{st} values) and the same CI considering all variation parameters.

Single source profile	Low variation scenario	High variation scenario
Point estimate	3.03×10^{10}	7.79×10^{7}
CI allele frequency variation only	2.70×10^{10}	1.47×10^{7}
CI all uncertainty (picking method)	1.70×10^{10}	5.13×10^{3}
CI all uncertainty (unifying method)	3.69×10^{9}	2.00×10^{3}
Four person mixture		Artificial crime sample
Point estimate		28.69
CI allele frequency variation only		19.95
CI all uncertainty (unifying method)		0.19

formula. This has the consequence that the LR distribution resulting from small database size can be partially masked by a wide F_{st} distribution.

Either method of accounting for population composition gives comparable results when considering a 99%, one-sided CI. However the picking method has several disadvantages. When population size is large in comparison to the number of relatives expected (so that the chance of picking a relative is less than the inverse of the number of HPD iterations) then often the picking method will not ever choose a relative and so the population composition has no effect on the LR distribution. In addition, there is no easy way to generate an LR point estimate using the picking method as the distributions can be multimodal.

It is typical in forensic laboratories to report an LR point estimate and a credible interval that takes allele frequency variation into account. Carrying out such a calculation can give the false impression that all uncertainty has been taken account of, and this may invite false statements or conclusions. In reality allele frequency variation is only one source of variation, and may not even be the greatest source of total variation within the LR. A reported LR probability interval which takes all sources of uncertainty into account can be similar in magnitude to the allele-frequency-only-LR if the other sources of variation are low (an order of magnitude as seen in the low variation scenario in Table 1). However, if other sources of variation are high, then the difference between the reported LR when taking only allele frequency and all sources of variation into account can be significant. In the high variation scenario this difference was four orders of magnitude. As seen in the four person example this difference amounted to a reduction in the reported LR of over two orders of magnitude, but perhaps more interestingly in this example, the 99% CI and the point estimate were on opposite sides of the neutrality line, LR = 1.

The use of MCMC methods to determine weightings, and Monte Carlo resampling to carry out the HPD calculations provide a powerful and flexible tool for assessing sources of uncertainty. The risk of using such a system is that the mathematics and general concepts are less transparent to an average user. Detailed training in the concepts of MCMC and HPD are required to avoid a 'black box' system. In addition, the transition to an MCMC based system requires a more complex plan for validation than simpler methods, which previously could be assessed with comparison to relatively simple hand calculations in many cases. Despite these complexities, we feel that the system described within this paper, coupled with appropriate training, would provide a powerful addition to a forensic laboratory's ability to interrogate EPG data.

The question remains, given that we can evaluate all these uncertainties within the *LR* calculations, what value should be reported to a courtroom? There are advocates that argue the most relevant figure to provide is the point estimate [6]. However, in our experience the court often concerns itself with exploring potential sources of uncertainty. If the decision is made to report a CI (or at least to calculate one for reference if required) then it would be difficult to justify only considering one aspect of total variation (i.e. allele frequency estimate, as is currently prevalent practice) and not others if an available method exists to reasonably do so. It is the personal view of the authors that a statistical evaluation of evidence should concede reasonable doubt and uncertainty to the defendant and so a reported value should be some lower quantile (e.g. 95 or 99%) of the LR distribution in criminal matters, but we respect alternative decisions. For civil matters there is often no side to which it is obvious to concede doubt and so a point estimate (mode or 50% quantile) may be the more appropriate figure to report.

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. We would also like to thank two anonymous reviewers for their helpful comments that improved this paper.

Appendix A

Proportions used for fraction of population related to POI.

The simplifying assumptions used to generate the relatives proportions are:

- There are three generations persisting in the population present in equal proportions.
- Generation 1 is the youngest and has no children.
- Generation 2 is the middle and has both children and parents surviving.
- Generation 3 is the oldest and has no surviving parents.
- There is only one union between two families i.e., there are no instances of two siblings from one family bearing children with two siblings from another family.
- Couples form from within their own generation.
- There are no instances of inbreeding closer than or equal to the first cousin level.
- Each family has the same structure and number of children.
- There are no instances of early death.

Table 2
Fraction of individuals in a population with known relationship.

Relationship type, i		The fraction of the population with relationship i, r_i	
		Low relatedness population	High relatedness population
Unrelated	1	0.999987	0.90000
Siblings	2	0.000001	0.00600
Children	3	1.33E-06	0.00600
Parents	4	1.33E-06	0.00533
Uncle/auntie	5	1.33E-06	0.00267
Niece/nephew	6	1.33E-06	0.00800
Grandparents	7	1.33E-06	0.01600
Grandchildren	8	1.33E-06	0.00267
First cousins	9	0.000004	0.04800

Using these assumptions the number of individuals of each relationship type to the POI were calculated for two populations; a large population with a low number of children per family and a small population with high numbers of children per family as seen in Table 2.

References

[1] 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 Sci. Int. 160 (2006) 90–101.
[2] T.M. Clayton, J.S. Buckleton, Mixtures. Forensic DNA Evidence Interpretation, CRC Press, Boca Raton, 2004, pp. 217–227.
[3] D. Taylor, J.-A. Bright, J.S. Buckleton, The interpretation of single source and mixed DNA profiles, Forensic Sci.e Int.: Gen. 7 (2013) 516–528.
[4] 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, J. Forensic Sci. 56 (2011) 1430–1447.
[5] J.M. Curran, A MCMC method for resolving two person mixtures, Sci. Justice 48 (2008) 168–177.
[6] C.H. Brenner, DNA frequency uncertainty – why bother (1997)
[7] P.I. Good, Applying Statistics in the Courtroom, Chapman&Hall/CRC, London, 2001.
[8] I.W. Evett, Weir BS. Interpreting DNA Evidence – Statistical Genetics for Forensic Scientists, Sinauer Associates Inc., Sunderland, 1998.
[9] R. Chakraborty, M.R. Srinivasan, S.F. Daiger, Evaluation of standard errors and confidence intervals of estimated multilocus genotype probabilities and their implications in DNA, Am. J. Hum. Genet. 52 (1993) 60–70.
[10] NRCII, National Research Council Committee on DNA Forensic Science, The Evaluation of Forensic DNA Evidence, National Academy Press, Washington, D.C, 1996.
[11] R. Chakraborty, Sample size requirements for addressing the population genetic issues of forensic use of DNA typing, Hum. Biol. 64 (1992) 141–160.
[12] B. Budowle, K.L. Monson, R. Chakraborty, Estimating minimum allele frequencies for DNA profile frequency estimates for PCR-based loci, Int. J. Legal Med. 108 (1996) 173–176.
[13] D.J. Balding, Estimating products in forensic identification using DNA profiles, J. Am. Stat. Assoc. 90 (1995) 839–844.
[14] J.M. Curran, J.S. Buckleton, C.M. Triggs, B.S. Weir, Assessing uncertainty in DNA evidence caused by sampling effects, Sci. Justice 42 (2002) 29–37.
[15] C.M. Triggs, J.M. Curran, The sensitivity of the Bayesian HPD method to the choice of prior, Sci. Justice 46 (2006) 169–178.
[16] J.M. Curran, J.S. Buckleton, An investigation into the performance of methods for adjusting for sampling uncertainty in DNA likelihood ratio calculations, Forensic Sci. Int.: Gen. 5 (2011) 512–516.
[17] J. Buckleton, Triggs C. Relatedness and DNA: are we taking it seriously enough? Forensic Sci. Int. 152 (2005) 115–119.
[18] D.J. Balding, P. Donnelly, Inference in forensic identification, JRSS A 158 (1995) 21–53.
[19] D.J. Balding, P. Donnelly, Inferring identity from DNA profile evidence, in: Proceedings of the 92 National Academy of Science, USA, (1995), pp. 11741–11745.
[20] J.S. Buckleton, A framework for interpreting evidence, in: J.S. Buckleton, C.M. Triggs, S.J. Walsh (Eds.), Forensic Dna Evidence Interpretation, CRC Press, Boca Raton, Florida, 2005.
[21] D.J. Balding, R.A. Nichols, DNA profile match probability calculation: how to allow for population stratification, relatedness, database selection and single bands, Forensic Sci. Int. 64 (1994) 125–140.
[22] S.J. Walsh, J.S. Buckleton, Autosomal microsatellite allele frequencies for a nationwide dataset from the Australian Caucasian sub-population, Forensic Sci. Int 168 (2007) e47–e50.
[23] J.S. Buckleton, C.M. Triggs, S.J. Walsh, DNA Evidence, CRC Press, Boca Raton, Florida, 2004.
[24] M. Plummer, N. Best, K. Cowles, K. Vines, CODA: convergence diagnosis and output analysis for MCMC, R News 6 (2006) 7–11.
[25] M. Plummer, Bayesian graphical models using MCMC, RJAGS (2012).
[26] D.J. Balding, Weight-of-evidence for Forensic DNA Profiles, John Wiley and Sons, Chichester, 2005.
[27] D.A. Taylor, J.M. Henry, Walsh SJ. South Australian Aboriginal sub-population data for the nine AMPFlSTR® Profiler Plus™ short tandem repeat (STR) loci, Forensic Sci. Int.: Gen. 2 (2008) e27–e30.
[28] S.J. Walsh, R.J. Mitchell, J.M. Curran, J.S. Buckleton, The extent of substructure in the indigenous Australian population and its impact on DNA evidence interpretation, Int. Congr. Ser. 1288 (2006) 382–384.
[29] J. Buckleton, S. Walsh, J. Mitchell, Autosomal microsatellite diversity within the Australian population, Report of the National Institute of Forensic Sciences Standing Committee on Sup-Population Data (2007).
[30] S.J. Walsh, J. Buckleton, Autosomal microsatellite allele frequencies for 15 regionally defined Aboriginal Australian population datasets, Forensic Sci. Int. 168 (2007) e29–e42.
[31] S.J. Walsh, R.J. Mitchell, N. Watson, J.S. Buckleton, A comprehensive analysis of microsatellite diversity in Aboriginal Australia, J. Hum. Genet. 52 (2007) 712–728.
[32] Australian Bureau of Statistics. 2011 Census QuickStats: South Australia 2011