Developmental validation of STRmix™, expert software for the interpretation of forensic DNA profiles

Jo-Anne Bright[1*], Duncan Taylor[2,3], Catherine McGovern[1], Stuart Cooper[1], Laura Russell[1], Damien Abarno[2], John Buckleton[1]

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

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

[3] *School of Biological Sciences, Flinders University, GPO Box 2100, Adelaide, SA 5001, 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

Abstract

In 2015 the Scientific Working Group on DNA Analysis Methods published the SWGDAM Guidelines for the Validation of Probabilistic Genotyping Systems [1]. STRmix™ is probabilistic genotyping software that employs a continuous model of DNA profile interpretation. This paper describes the developmental validation activities of STRmix™ following the SWGDAM guidelines. It addresses the underlying scientific principles, and the performance of the models with respect to sensitivity, specificity and precision and results of interpretation of casework type samples. This work demonstrates that STRmix™ is suitable for its intended use for the interpretation of single source and mixed DNA profiles.

Keywords

DNA mixtures; probabilistic genotyping; continuous method; validation; STRmix.

Introduction

The dominant method for forensic DNA analysis involves the amplification of short tandem repeats using PCR. Amplified products are separated via capillary electrophoresis (CE). Fluorescently labelled tags are used to colour code the markers or loci. A laser excites the primer tags as the different lengths of DNA travel through the capillaries of the electrophoresis instrument, which emit a signal that is recorded. The signals are visualised as peaks in a graph of fluorescence versus time, known as an electropherogram (epg). The height of the peaks is approximately proportional to the initial amount of DNA template and is measured in relative fluorescent units (rfu). In this way height can be used as an approximation of DNA quantity or template.

Manual techniques for DNA profile interpretation are heuristically based and may be difficult to apply consistently between laboratories, individual scientists and even a single scientist. Variable decisions often occur early in the manual interpretation process and can even occur at allele assignment. Divergence in these choices can have significant downstream consequences [2, 3]. Phenomena such as stutter (artifactual amplicons produced as a consequence of the PCR process), allelic drop-in (the presence of low amounts of extraneous DNA) and dropout (which is a consequence of low template and/or degraded DNA and results in partial DNA profiles) [4] are all considered at profile analysis and interpretation. Interpretation of DNA profiles is also complicated by mixed samples (the presence of DNA from more than one individual).

The interpretation of an epg or evidentiary DNA profile should initially be undertaken 'blind'; in isolation of the person of interest's (POI) reference DNA profile, and where possible avoiding contextual effects [5, 6]. Comparison with reference profiles of any POI or other relevant evidentiary profiles is undertaken after profile interpretation. Traditionally there are three primary conclusions that can be drawn: *cannot exclude* (or *inclusion*), *can exclude*, or *inconclusive* which is sometimes also called *uninterpretable* [7]. It is desirable when an association is reported (cannot exclude or inclusion) to present the evidence with the associated statistical weight [7]. When the evidence profile originates from a single individual, the weight of evidence can be presented as a match probability. This is an assignment of the probability that a random person might match the crime scene stain given the observation of that crime stain profile. A favoured alternative to the match probability, which can be extended to use for mixed DNA profiles, is the likelihood ratio (*LR*). The *LR* considers the probability of obtaining the evidence profile(s) given two competing propositions, usually aligned with the prosecution case and defence case. The *LR* is used throughout Australasia and the UK and is used in some laboratories within the US and Europe for criminal forensic work to express the weight of evidence. The *LR* is accepted to be the most relevant and powerful statistic to calculate the weight of the evidence and is the only method recommended by the International Society for Forensic Genetics (ISFG) for ambiguous profiles [8]. Ambiguous profiles include all mixtures and single source profiles where dropout and drop-in are a consideration.

Known shortcomings of traditional methods of DNA profile interpretation have led to the development of improved models that factor in the probability of dropout [9-13]. The drop model (also known as the semi-continuous method) can optionally incorporate a probability for dropout, Pr(D), and/or a probability for drop-in, Pr(C). Semi-continuous methods do not use peak heights when generating possible genotype sets and do not model artifacts such as stutter. Continuous methods make assumptions about the underlying behaviour of peak heights across all profiles to evaluate the probability of a set of peak heights in a given profile. These methods are designed to be used in expert systems and reduce the requirement for the manual assignment of peaks as allelic within evidence profiles, and hence reduce the opportunity for inconsistency in interpretation to occur. The calculations are sufficiently complex that software is needed. STRmix™ is one such continuous method that employs a fully continuous approach for DNA profile interpretation (http://strmix.esr.cri.nz/ [14]).

In 2015 the Scientific Working Group on DNA Analysis Methods published the SWGDAM Guidelines for the Validation of Probabilistic Genotyping Systems [1]. The developmental validation of a probabilistic genotyping system has been described by SWGDAM as "the acquisition of test data to verify the functionality of the system, the accuracy of statistical calculations and other results, the appropriateness of analytical and statistical parameters, and the determination of limitations" [1].

The developmental validation of STRmix™ was initially undertaken in 2012 following the requirements outlined within the FBI Quality Assurance Standards [15] by analysts at Forensic Science South Australia (FSSA) and the Institute of Environmental Science and Research Limited (ESR; http://www.esr.cri.nz/). FSSA is the South Australian State Forensic Science Laboratory and is accredited by the National Association of Testing Authorities, Australia. ESR is the New Zealand Government Crown Research Institute that undertakes forensic services for the NZ Police. ESR forensic DNA laboratories are accredited by the Laboratory Accreditation Board of the American Society of Crime Laboratory Directors (ASCLD/LAB) under the International Testing Program (ISO 17025).

Within this paper we describe the developmental validation activities undertaken for STRmix™ following the SWGDAM recommendations [1]. Each of the guidelines is discussed in turn under their recommendation number.

Guideline 3.1 Publication of underlying scientific principles

All significant portions of the statistical algorithms and underlying scientific principles behind STRmix™ have been published in peer reviewed scientific literature. Within Table 1 we provide a summary of these models and algorithms and their references aligned with the software version in which they were introduced.

STRmix™ uses the quantitative information from an electropherogram (epg) such as peak heights (O), to calculate the probability of the profile given all possible genotype combinations (S_j). A value, or weight (w_i), is assigned to the normalised probability density $p(O|S_j)$. STRmix™ assigns a relative weight to the probability of the epg given each possible genotype combination at a locus. The weights across all combinations at that locus are normalised so that they sum to one. Therefore, a single unambiguous genotype combination at any locus would be assigned a weight of one.

STRmix™ describes the fluorescence observed in one or more epgs using a number of models that describe various properties of DNA profile behaviour. These are described as mass parameters and include a template for each contributor, a locus specific amplification efficiency for each locus, a replication efficiency for each PCR replicate, and a degradation for each contributor. This biological model is described in Bright et al. [16]. Profile degradation is modelled as exponential [17, 18]. Drop-in is optionally modelled as a gamma distribution following Puch-Solis [19]. In addition, STRmix™ employs a per allele stutter model, the parameters of which are based on empirical data [16, 20, 21].

Posterior distributions of mass parameters are sampled from using Markov chain Monte Carlo (MCMC). In general, MCMC is a numerical method used, in this case, to approximate an integral (typically multi-dimensional) of the observed data across all parameters. MCMC methods sample from the posterior distribution of the desired integral. It does so by using Markov chains that have the posterior distribution as their equilibrium distribution. These chains 'walk' around in a memoryless fashion using an acceptance-rejection criterion to determine whether to take a step or not. At each step that the chain accepts the integrand value, it is counted towards the integral. At each step that the chain rejects the integrand value at that proposed point, the current point is counted towards the integral. The rejection-acceptance rule used within STRmix™ is called the Metropolis-Hastings algorithm [22, 23]. The chain will then propose new steps in its search for a state that provides a reasonably high contribution to the integral until it finds a state which it will accept and move to. The statistical algorithms within STRmix™ are described in Taylor et al. [14].

STRmix™ does not use the reference profiles during profile deconvolution unless a reference from a known contributor is available (for example the complainant's DNA on their intimate samples collected as part of an investigation into a sexual assault). Where a reference profile is available from a person of interest (POI) a likelihood ratio may be calculated. It is the ratio of the probability of the observed crime stain (O) given each of two competing hypotheses, H_p and H_d, and given all the available information, I. Mathematically, we express this as:

$$LR = \frac{\Pr(O \mid H_p, I)}{\Pr(O \mid H_d, I)}$$

The likelihood ratio is calculated in STRmix™ incorporating values for F_{ST} (theta) using the subpopulation model of Balding and Nichols in 1994 [24], referred to as recommendation 4.2 in the 1996 National Research Council report (NRCII) [25]. As a continuous extension to the classic incorportation of a theta value (which is typically a fixed value) STRmix™ can consider a distribution for theta. Propositions within STRmix™ are flexible. The defence proposition aligns with exclusion of the person of interest and typically considers an unknown, unrelated individual within a selected population. Where appropriate, alternate propositions are calculated under the defence propositions such as a sibling, parent, child or cousin of the person of interest [26]. Additionally STRmix™ can provide an *LR* based on the unifying theory. This is where rather than specifying either an unrelated individual or a nominated relative (sibling, parent etc.) under the defence propositon, all members of the population, including possible relatives of the POI can be considered by taking into account their prior probabilities based on population properties.

If one or more contributors is known to be present (i.e. conceded by both parties) then this information can be provided to STRmix™ at the deconvolution stage in order to assist in the deconvolution of the remaining questioned contributors. This assumption of a known contributor is then carried forward to the *LR* calculation. If a reference profile is not available from a person of interest, the profile may be compared directly with a database of known individuals [28] to identify investigative leads.

STRmix™ uses the highest posterior density (HPD) method for calculating an *LR* distribution, from which a quantile can then be chosen in order to report a bound of the probability density distribution [29, 30]. Within STRmix™ versions 2.3 onwards, the variability due to MCMC, the sampling variation inherent in generating allele frequency databases and the variability in F_{ST} (theta) can be estimated.

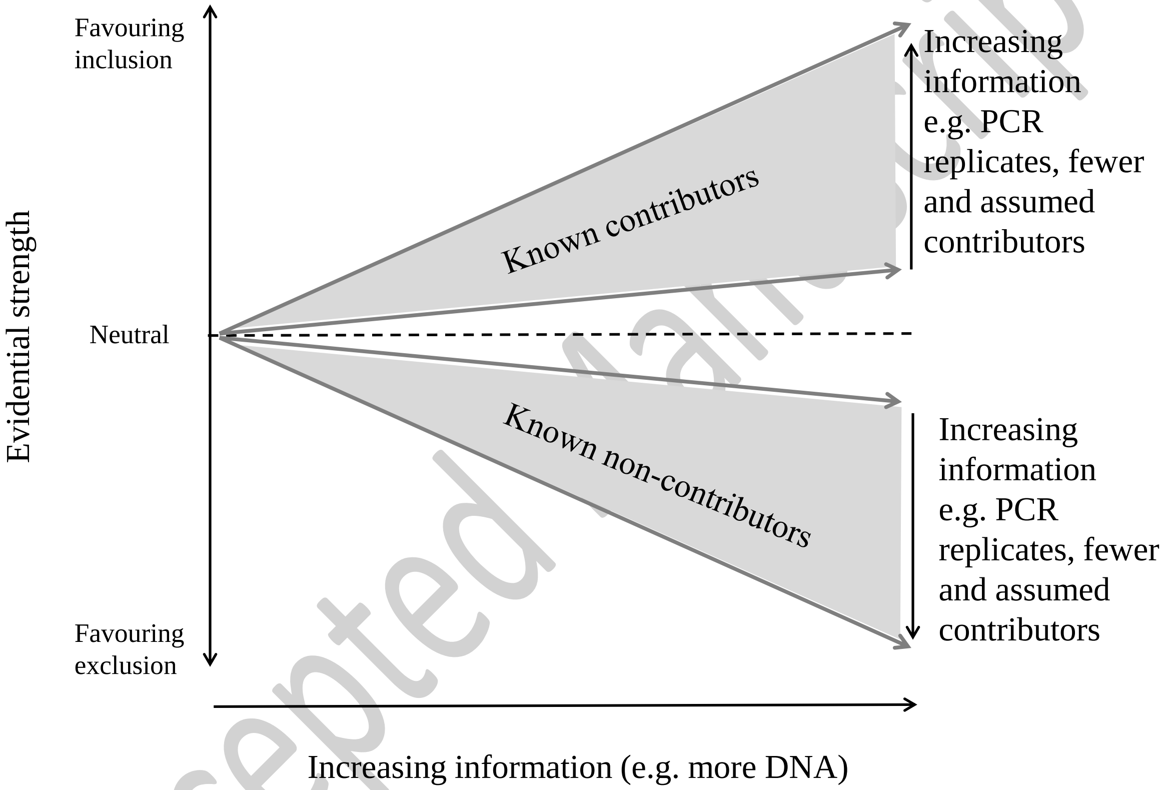
Table 1: A summary of the scientific principles, the STRmix™ version in which they were introduced and their publications

Algorithms, scientific principles and methods	Version introduced	Reference
Allele and stutter peak height variability as separate constants within the MCMC	V2.0	[14][14]
Peak height variability as random variables within the MCMC	V2.3	[31]
Model for calibrating laboratory peak height variability	V2.0	[31]
Application of a Gaussian random walk to the MCMC process	V2.3	Described within this paper
Modelling of back stutter by regressing stutter ratio against allelic designation	V2.0	[16, 20, 32, 33]
Modelling of back stutter by regressing stutter ratio against *LUS*	V2.3	[16, 20, 21, 33]
Modelling of forward stutter	V2.4	[34]
Modelling of allelic drop-in using a simple exponential or uniform distribution	V2.0	[14][14]
Modelling of allelic drop-in using a Gamma distribution	V2.3	[19]
Modelling of degradation and dropout	V2.0	[17]
Modelling of the uncertainties in the allele frequencies using the HPD	V2.0	[30]
Modelling of the uncertainties in the MCMC	V2.3	[29, 30, 35]
Database searching of mixed DNA profiles	V2.0	[28]
Familial searching of mixed DNA profiles	V2.3	[26]
Relatives as alternate contributors under the defence proposition	V2.3	[26]
Modelling expected stutter peak heights in saturated data	V2.3	[34]
Taking into account the 'factor of two' in *LR* calculations	V2.3	[36]
Model for incorporating prior beliefs in mixture proportions	V2.3	[37]

Guideline 3.2 Sensitivity and specificity studies

With respect to interpretation methods, sensitivity is defined as the ability of the software to reliably resolve the DNA profile of known contributors within a mixed DNA profile for a range of starting DNA template. The log(LR) for known contributors (H_p true) should be high and should trend to 0 as less information is present within the profile. Information includes the amount of DNA from the contributor of interest, conditioning profiles (for example the victim's profile on intimate samples), PCR replicates and decreasing numbers of contributors. Specificity is defined as ability of the software to reliably exclude known non contributors (H_d true) within a mixed DNA profile for a range of starting DNA template. The LR should trend upwards to neutral as less information is present within the profile. This is shown diagrammatically in Figure 1.

Figure 1: A diagram showing the desired performance of a method of mixture interpretation.


Favouring inclusion
Evidential strength
Neutral
Favouring exclusion
Known contributors
Known non-contributors
Increasing information e.g. PCR replicates, fewer and assumed contributors
Increasing information e.g. PCR replicates, fewer and assumed contributors
Increasing information (e.g. more DNA)

Specificity and sensitivity within STRmix™ were tested by calculating the LR for a number of GlobalFiler™ mixtures for both known contributors and known non-contributors [38]. Two, three and four contributor mixtures were constructed in varying proportions and amplified with varying amounts of template DNA as described in Table 2.

Table 2: A summary of the experimental set up

Sample	Mixture proportions for contributor				Total DNA added to PCR (pg)
	One	Two	Three	Four	
1-3	0.50	0.50	-	-	400,200,50
4-6	0.33	0.67	-	-	
7-9	0.20	0.80	-	-	
10-11	0.17	0.83	-	-	
13-15	0.09	0.91	-	-	
16-18	0.33	0.33	0.33	-	
19-21	0.50	0.33	0.17	-	
22-26	0.25	0.25	0.25	0.25	400,200,50,20,10
27-31	0.40	0.30	0.20	0.10	

Each sample was amplified in triplicate giving a total of 93 samples. Profiles were interpreted using STRmix™ v1.08 and *LR*s calculated for the known contributors and 186 non contributors. The propositions considered were:

H_p: The DNA originated from the person of interest and *N*-1 unknown contributors

H_d: The DNA originated from *N* unknown individuals

Where *N* was the number of contributors within the profile.

The plots of $\log_{10}(LR)$ versus DNA in the PCR (pg) produced for these comparisons are reproduced in Figures 2 through 6. The *LR*s produced from comparisons to known contributors (sensitivity tests) are signified by a blue point and those produced from comparisons to known non-contributors (specificity tests) are signified by a red point. A minimum value for $\log_{10}(LR)$ of -30 was used, and any *LR*s obtained that fell below this were given the value of -30. The lines on figures are given only as a visual indication of trends in the scattered results. The polygons seen give a visual indication of the spread of the *LR*s.

The plots in Figures 2 through 6[1] clearly demonstrate the sensitivity of STRmix™ for these mixtures by inspection of the spread of blue points. They show the range of expected *LR* values for contributors given the amount of input DNA (guideline 3.2.1.2). Type I errors (incorrect rejection of a true hypothesis) are clearly identified as blue points *below* the horizontal line of $\log_{10}(LR) = 0$. As expected, this is dependent on the amount of DNA per contributor and the number of contributors to a profile (guideline 3.2.1.1).

The plots also demonstrate the specificity of STRmix™ by inspection of the red points. The per contributor amount for H_d true contributors was taken as the average of the known contributors (guideline 3.2.2.2). Type II errors (failure to reject a false hypothesis) are clearly identified as reds points *above* the horizontal line of $\log_{10}(LR) = 0$. As for sensitivity tests, this depends on the amount of DNA within the profile and number of contributors (guideline 3.2.2.1). A series of much larger simulations (over 100 million *LR*s in total) exploring the specificity of STRmix™ and comparing it to theoretical expectations was carried out in [39]. This work found close alignment with expected and observed specificity from STRmix™ results.

[1] Reprinted from Forensic Science International: Genetics, Volume 11, Duncan Taylor, Using continuous DNA interpretation methods to revisit likelihood ratio behaviour. Forensic Science International: Genetics, Pages 144-53, Copyright 2014, with permission from Elsevier.

The *LR* distributions for H_p true and H_d true are very well separated at high template for two contributor mixtures. As the number of contributors increased and the template lowered the two distributions converged on $\log_{10}(LR) = 0$. At high template STRmix™ correctly and reliably gave a high *LR* for true contributors and a low *LR* for false contributors. At low template or high contributor number STRmix™ correctly and reliably reported that the analysis of the sample tends towards uninformative or inconclusive.

There are some arguments [1-3] that a single point estimate of the *LR* as given in Figures 2 through 6 is actually the best and most theoretically sound estimate to give if the goal was an even handed and probabilistic treatment of uncertainty. In DNA profile interpretation we typically deliberately give an underestimate. In our own casework we predicate this with the words "at least" by which we mean that the number reported is either below or very near the bottom of the plausible range. Our experience suggests that this is done because of the desire by the courts and forensic scientists to avoid overstating the evidence. Over time the avoidance of overstatement has changed into what is probably a very considerable and deliberate understatement. This has been facilitated, we believe, because DNA can afford this understatement given the magnitude of our likelihood ratios.

Sensitivity and specificity studies however have a scientific component to them and it may be desirable to use the best estimate available for these. If these studies are used to formulate decisions such as assigning terms to a verbal scale then it should be noted that they refer to the point estimate and not the lower bound. This has an additional and possibly undesirable consequence that if the verbal scale is calibrated from the sensitivity and specificity plots and then this scale is applied to the lower bound, the scale itself now possesses an element of conservativeness.

Figure 2: LRs produced for two person mixtures

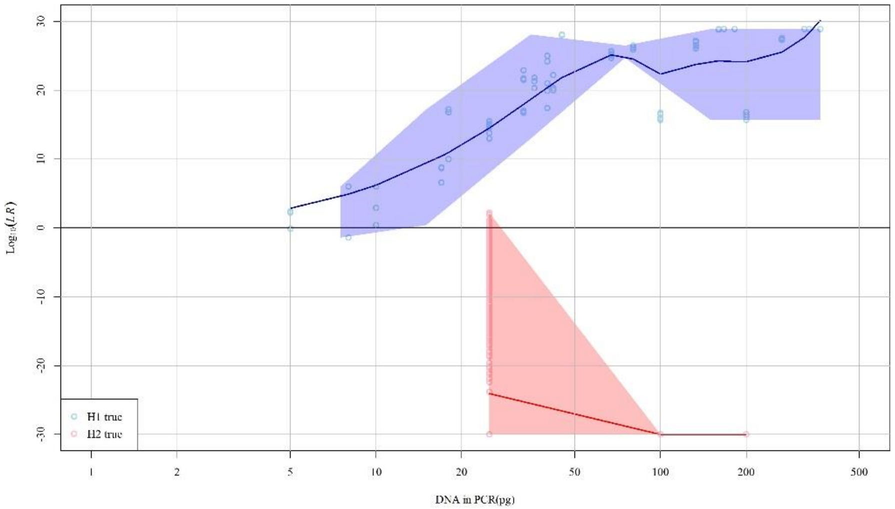


Figure 3: *LR*s produced for three person mixtures

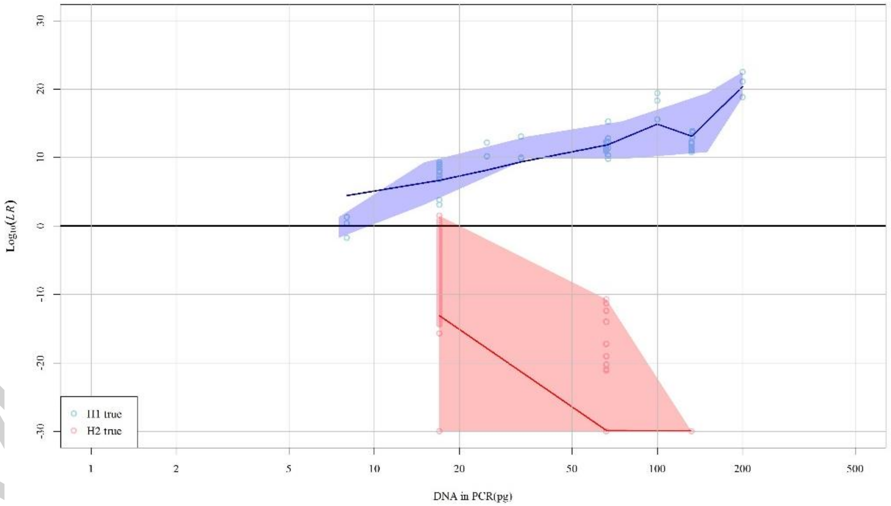


Figure 4: *LR*s produced for four person mixtures

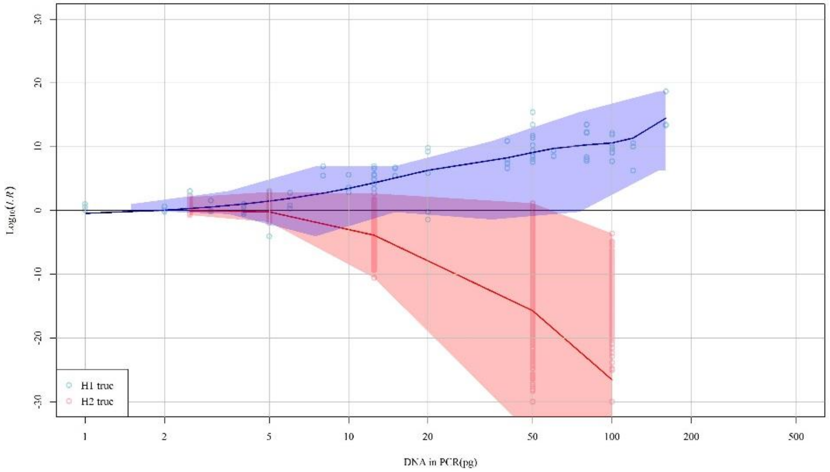


Figure 5: *LR*s produced for four person mixtures using three replicate amplifications

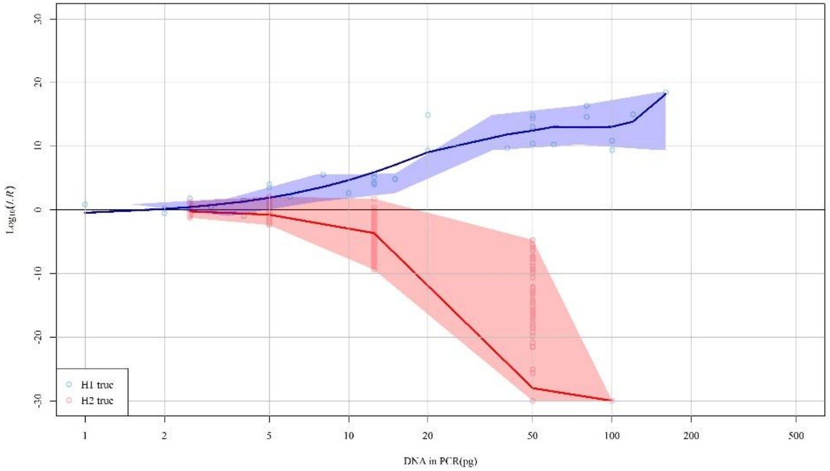
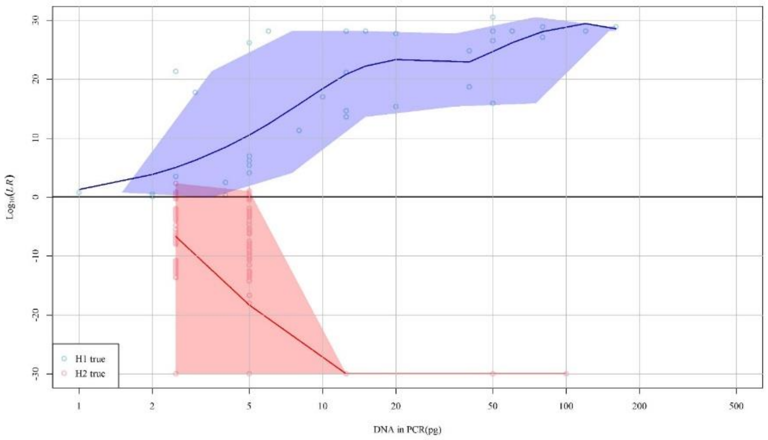


Figure 6: ***LR*s produced for four person mixtures using three replicate amplifications and assuming three out of the four known contributors in each analysis**



There is no specific SWGDAM guideline regarding error rate but it is one of the Daubert standards regarding the admissibility of expert evidence in the US [40], with acknowledgement that these guiding factors are neither exclusionary nor mandatory [41]. With respect to forensic DNA evidence, the concept of error rates and false inclusions[2] are similar and often confused. False inclusions would come under the specificity guideline of SWGDAM (guideline 3.2).

Our preferred procedure when using STRmix™ is that the analyst assesses whether a person of interest is excluded prior to either their assessment of the results of software calculations or interpretation of the profile using the software at all. Following this procedure, STRmix™ is being continually checked against human expectations and hence is being continually validated.

The number of *LR*s >1 is largely determined by the sample. Factors include the number of contributors and template. Considerable research has been undertaken that allows informed statements to be made about the false inclusion rate for any given sample [14, 28, 38, 42].

The *LR* is an assessment of the weight of evidence. It is developed by considering two propositions: one aligned with the prosecution and an alternative. *LR*s >1 support the prosecution proposition and those lower than one support the alternative.

To highlight the matter, consider that we make up a DNA mixture and hence we know the donors. Consider that this mixture is made from Smith and Brown. If we test the proposition that it contains Smith we expect a high *LR*. Suppose the *LR* is a billion. Is this correct? It is larger than one and as such that part is correct, but is a billion too large or too small or just right? The problem is that we do not have the 'true answer' and this cannot be obtained by any method.

False exclusions or false inclusions need to be interpreted in an *LR* framework. A false exclusion most nearly corresponds with an *LR* markedly less than one when H_p is true. A false inclusion most nearly corresponds with an *LR* markedly greater than one when H_d is true. *LR*s near one are best described as uninformative and this may be the correct indication of the value of the profile even for comparisons with true or false donors if the information present in the profile is limited.

When we consider a possible error rate for STRmix™ this must be balanced against the error rate for the entire DNA analysis process which can cause false inclusions and exclusions independent of the program. A false inclusion occurs when:

- A non-donor has the correct alleles by chance, in total or in large part, to explain the mixture.

It is very improbable that operator error (such as the inclusion of artifacts) or false information about a known contributor would cause a false inclusion.

The rate of false inclusion is increased in situations where the true DNA donor is a close relative of the POI[3]. Higher order mixtures, say four contributors, increase the chance of false inclusions. Depending on the type of profile and proportion of DNA corresponding to the POI, replication and the correct use of known contributors can reduce the chance of false inclusions

[2] Note that the terms 'false inclusion' and 'false exclusion', whilst commonly used, imply an error has occurred, when in reality the probability has been assigned as expected in accordance with theory. A better term would be 'support for a false proposition'; however we retain the terms 'false inclusion/exclusion' for general understanding.

[3] Exploratory experimental work (ongoing) undertaken in conjunction with USACIL and the FBI suggests that STRmix™ can handle most of these situations.

(refer Figures 5 and 6). In addition, more loci used in the analysis will reduce the chance of false inclusion.

A false exclusion occurs when:

- The PCR reaction runs sufficiently poorly that the peak or stutter heights give misleading information, or
- A non-contributor is assumed to be present, or
- There is an operator error, notably inclusion of an artifact in the peak information used by STRmix™ at interpretation. An artifactual peak that has been retained within the input file will become part of the information used by STRmix™ to build genotype combinations. This will result in genotype combinations containing the artefact which will not align with the "true" genotypes of contributors to the profile. If the POI aligns with one of these altered (false) genotypes, this might result in a false exclusion.

There are a number of factors within STRmix™ under the control of the operator or the laboratory that affect errors. Most significantly are the two variance terms. If these are set too low they increase false exclusions. Set too high they increase false inclusions. These variances are set during a laboratory's internal validation by modelling the observed variation in allelic and stutter peak heights within a set of single source profiles of varying quality [31]. There are a number of diagnostics output by STRmix™ that allow a human check of the results including the genotypic weights ($p(O|S_j)$), the posterior mean of the variance terms and summary statistics of the MCMC (discussed later).

False inclusions and false exclusions may occur as a result of a combination of specific software, multiplex and operator factors. These are measurable. The most significant factors affecting them are the number of contributors, the number of known contributors, template levels, and the multiplex used. These factors are wrapped up in the *LR* in a way that the chance of producing an *LR* equal to or larger than the one in any particular case (LR_{case}) is less than $1/LR_{\text{case}}$. This relationship has been tested in trials of over 120 million cases of simulated false contributors and has always held [39].

The fraction of false donors exceeding LR_{case} has been termed the *p*-value [43-45] and it has been convincingly argued that they do not replace the *LR* [46]. Nor is the *p*-value a direct measure of the false inclusion rate since an *LR* for a false donor less than LR_{case} but still much larger than one would be considered a false inclusion.

We have no realistic way of measuring the false exclusion rate except to say that we have no undiagnosed instances of false exclusion.

The pink data within Figures 2 through 6 are the $\log_{10}(LR)$ values for non-donors. Any red data points above the line support H_p and may therefore be considered false inclusions. These data, which are towards the low template end, are slightly above the $\log_{10}(LR) = 0$ line, and are usually likelihood ratios between 1 and 1,000 ($\log_{10}(LR)$ 0 to 3). We term these low grade false inclusions since the *LR*s are low and near neutrality or only slightly to the inclusionary side. They occur when the false donor has the correct alleles for inclusion and hence they are a property of DNA rather than a consequence of the software not performing. There are no modelling improvements that could ever be made which will eliminate all *LR*s that falsely favour inclusion. This is because the phenomenon causing these results is not a modelling phenomenon, but is due to the available biological data. With any interpretation method there is a modelling component (including probability of dropout and drop-in) that will affect the magnitude of the *LR*, and this could mean the difference between a false inclusion and correct exclusion for a particular non-donor.

Uncertainty in the number of contributors

The determination of the effect of incorrectly assigning the number of contributors to a profile on the interpretation is not explicitly a requirement of developmental validation within the SWGDAM guidelines however this is something the STRmix™ development team has explored. The true number of contributors to a profile is always unknown. Analysts are likely to add contributors in the presence of an artifact, high stutter, or forward stutter peak. The assumption of one fewer contributor than that actually present may be made when contributors are at very low levels, are affected by peak masking and are dropping out (or visible below the analytical threshold), and in profiles where DNA is from individuals with similar profiles at the same concentrations.

The effect of the uncertainty in the number of contributors within STRmix™ has previously been reported for a number of profiles with *N* and *N*+1 assumed contributors, where *N* is the known number of contributors [28, 42]. The inclusion of an additional contributor beyond that present in the profile had the effect of lowering the *LR* for trace contributors within the profile. STRmix™ adds the additional (unseen) profile at trace levels which interacts with any known trace contribution, diffusing the genotype weights and lowering the *LR*. There was no significant effect on the *LR* of the major or minor contributor within the profiles.

Separately, the effect of underestimating the number of contributors to a profile (*N* versus *N*-1) has been investigated. Assigning the number of contributors as *N*-1 (where *N* is the known number of contributors) may result in an exclusionary *LR* for a known contributor. This occurs as STRmix™ is more likely to favour an incorrect genotype as it had to account for profiling information that does not explain the data accurately.

Guideline 3.2.3. Precision

STRmix™ assigns a relative weight to the probability of the epg given each possible genotype combination at a locus. These weights are determined by Markov chain Monte Carlo (MCMC) methods. The results of no two analyses will be completely the same using a stochastic system like MCMC. This is a phenomenon that is relatively new to forensic DNA interpretation, which up until this point has always had the luxury of, at least theoretically, completely reproducible interpretation results. The reproducibility of *LR*s calculated using STRmix™ has previously been explored by Bright et al. [35, 48].

The main cause of high variability within STRmix™ is non-convergence with the MCMC. Strictly, Markov chains do not converge. They explore the sampling space forever until they are told to stop. What we mean when we say Markov chains have reached convergence is that all chains are sampling from, and remain in, the ‘same’ high probability space.

Non-convergence is caused by the MCMC chains not being run for a sufficient number of accepts. The MCMC process starts with a number of iterations termed the ‘burn-in’. Accepted genotypes from the burn-in process are not counted as they are likely to start at a low probability location. At the completion of burn-in the MCMC progresses to post burn-in. STRmix™ is set to run for a user defined number of burn-in and post burn-in accepts. STRmix™ uses accepts as a method of controlling how long the MCMC runs rather than total iterations. The reason for this is that by ensuring a defined number of accepts is obtained there is some degree of automatic scaling, whereby more complicated problems (with lower acceptance rates) will automatically run for more iterations, without the need for user intervention.

Non-convergence can be diagnosed using the Gelman-Rubin statistic [49, 50]. A high Gelman-Rubin statistic in conjunction with other diagnostics may be an indication of non-convergence. The solution to non-convergence is to run the problem for longer, i.e. for more MCMC accepts. We typically multiply the number of burn-in and post burn-in accepts by 10.

Putting aside non-convergence, there will always exist a level of MCMC run to run variability. This is simply due to the fact that the analysis is based on random number generation to function, which as the name suggests, is random. Ideally this variability in some output value is small in comparison to the size of the value itself and hence its impact on interpretation is minimal, and in some instances can be taken into account. Variation in *LR*s produced from STRmix™ analyses will depend on both the sample and the run parameters. Sample specific factors that affect precision include:

1. Number of contributors to a DNA profile
2. Quality/intensity of the DNA profile
3. Number of replicates available for analysis
4. The probability of the observed data given the genotype of the POI as a contributor (commonly referred to as the 'fit' of the POI)
5. The amount of STR information available in the profile.

STRmix™ run specific parameters that affect precision include

1. Number of iterations the MCMC has run
2. The number of Markov chains used
3. The step size of the Markov chain (termed the random walk standard deviation, RWSD).

The RWSD is a metaparameter that describes the standard deviation of the normal distributions from which the step size for each continuous parameter is drawn. We describe this metaparameter in more detail below. The effect of these run specific parameters on the variability of the *LR* is discussed in detail below.

Number of MCMC chains and accepts

Increasing the number of either accepts or moves and adjusting the step size (the RWSD) can reduce but not totally remove the variation. There is, however, an associated runtime cost. Hence a trade-off between reproducibility and runtime must be struck.

The variation in the calculated *LR* due to sample factors and run specific parameters in STRmix™ has been explored for a number of different profiles with varying numbers of contributors and quality. Eight profiles were generated 'in silico'. These included one, two, three and four contributor profiles, in various template (high and low level) and proportions, in the GlobalFiler™ kit configuration. Each profile was interpreted in STRmix™ v2.3.07 ten times giving 80 runs in a batch. For each batch, a different combination of number of chains, burn-in and post burn-in accepts were trialled. In total, sixteen different chain/iteration combinations were tested generating data for over 1200 profile deconvolutions. The data was analysed to determine which chain/iteration combination resulted in the best reproducibility whilst also considering the impact on run time. A summary of the number of chain and accepts combinations considered is provided in Table 3.

Table 3: Summary of run parameters (chains, burn-in and post burn-in accepts) undertaken to interpret the sixteen profiles in order to explore the precision of STRmix™.

Set	Chains	Burn-in accepts	Post burn-in accepts
1	4	50,000	150,000
2	4	500,000	200,000
3	4	50,000	2,000,000
4	4	500,000	2,000,000
5	8	50,000	400,000
6	8	500,000	400,000
7	8	50,000	4,000,000
8	8	500,000	4,000,000
9	16	50,000	800,000
10	16	500,000	800,000
11	16	50,000	8,000,000
12	16	500,000	8,000,000
13	20	50,000	1,000,000
14	20	500,000	1,000,000
15	20	50,000	10,000,000
16	20	500,000	10,000,000

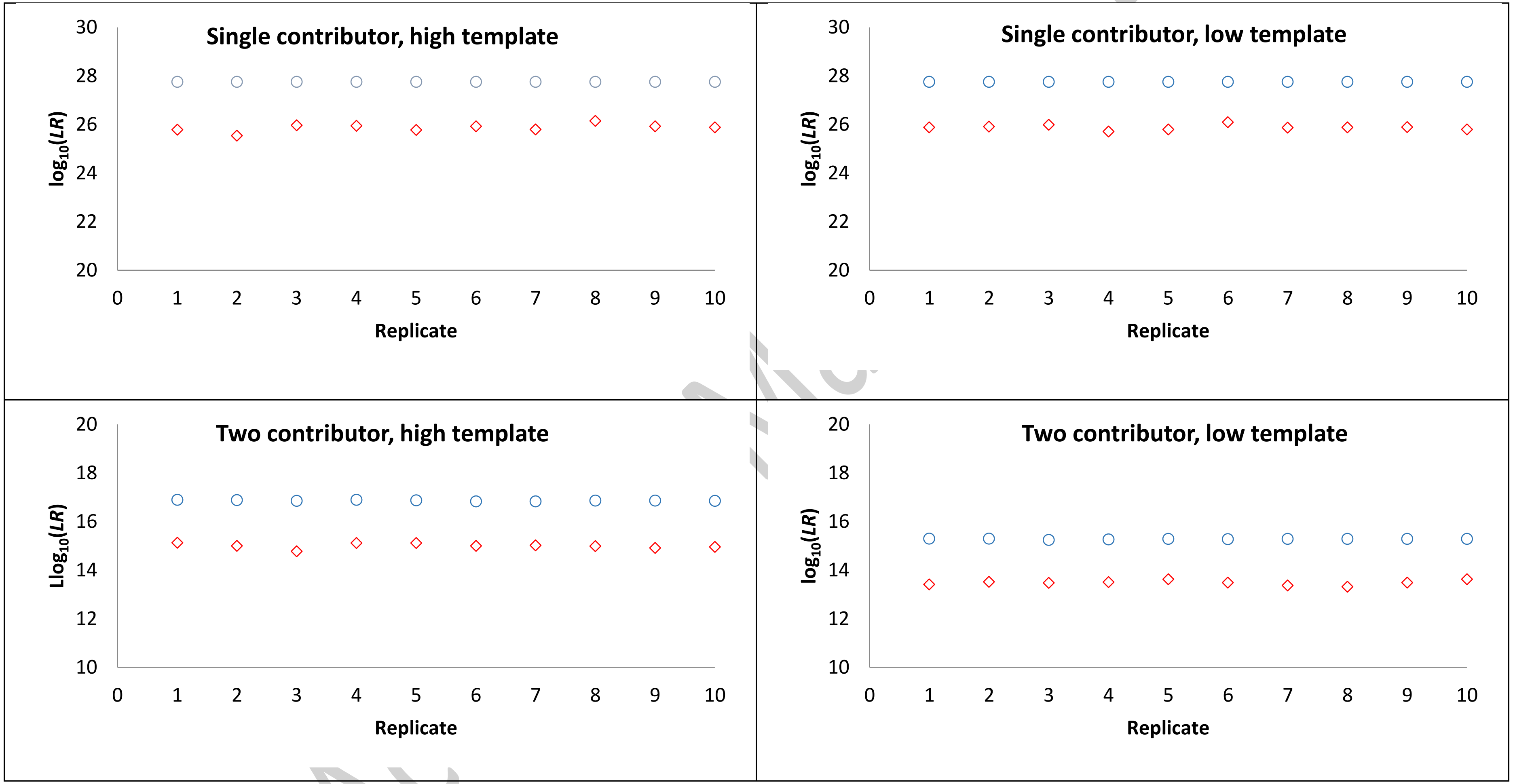
A summary of the point estimate and 1st percentile (taking into account sampling variation in allele proportions and weights) of the distribution of $\log_{10}(LR)$ value (called the $\log_{10}(LR)$ and $\log_{10}$(HPD) respectively) for each of the ten replicates is provided in Appendix 1 (ordered by run parameter set) and Appendix 2 (ordered by profile). In addition summary statistics including the Gelman-Rubin diagnostic and posterior means of the allele and stutter variance constants are provided.

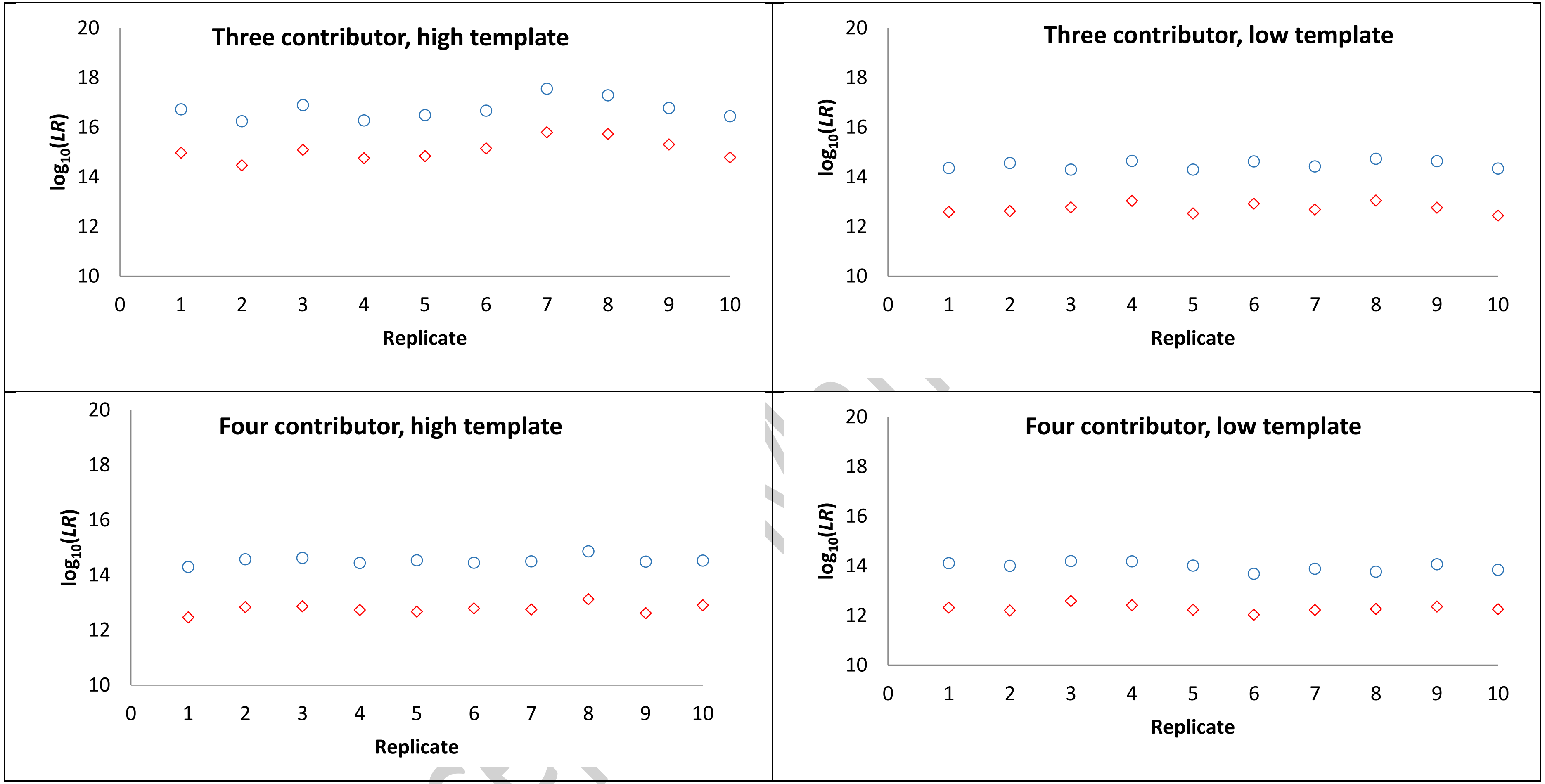
Inspection of the results in Appendix 1 and 2 show that as the profile is interpreted using more Markov chains and higher numbers of accepts, STRmix™ analyses are more likely to converge to the same parameter values, resulting in more reproducible *LR*s. The number of chains, total number of burn-in and post burn-in accepts and number of contributors all had an effect on run times. Consequently some interpretations were not completed after reviewing the wider results.

The *LR* for the two GlobalFiler™ single source profiles under all run configuration was identical. Due to the peak heights of these profiles dropout was not considered, resulting in a single genotype combination at each locus with weights equalling one. This was the expected result. The two person mixtures all gave *LR*s within one order of magnitude across all run configurations. There was an increase in observed *LR* variability within the three and four person mixtures with lower numbers of chains and lower total iterations.

A summary of the distribution of the $\log_{10}(LR)$ and $\log_{10}$(HPD) for ten replicates of the eight GlobalFiler™ profiles using eight chains with 50,000 burn-in accepts and 400,000 post burn-in MCMC accepts is provided in Figure 7.

Figure 7: Log$_{10}$(*LR*) (○) and log$_{10}$(HPD) (◊) of ten replicate interpretations of different GlobalFiler™ profiles, interpreted using eight chains with 50,000 burn-in accepts and 400,000 post burn-in MCMC accepts

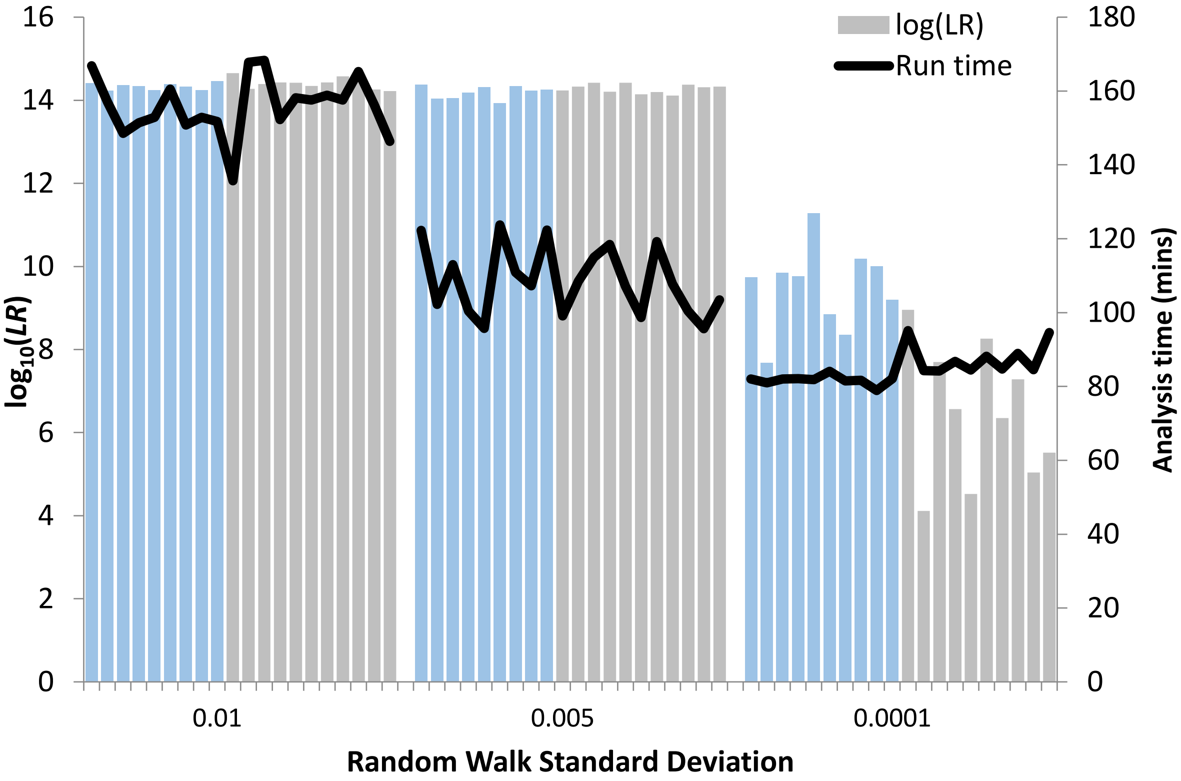



Three contributor, high template
log10(LR)
20
18
16
14
12
10
0
1
2
3
4
5
6
7
8
9
10
Replicate
Three contributor, low template
log10(LR)
20
18
16
14
12
10
0
1
2
3
4
5
6
7
8
9
10
Replicate
Four contributor, high template
log10(LR)
20
18
16
14
12
10
0
1
2
3
4
5
6
7
8
9
10
Replicate
Four contributor, low template
log10(LR)
20
18
16
14
12
10
0
1
2
3
4
5
6
7
8
9
10
Replicate

Random walk standard deviation

At each iteration, the MCMC will have a particular set of values stored that describe the profile. When proposing new values for the next MCMC iteration the values will be chosen close to the current set of values. The distance of the step-size is based on a normal distribution with a mean set to the current value and a variance that dictates step-size. This is known as a Gaussian random walk. In a Gaussian walk the size of the step for any given variable is sampled randomly from $\sim N(0, sd^2)$. The size of sd^2 is dependent on the parameter and is tuned by the RWSD. Setting the RWSD too high will result in the values for the mass parameters that are used to describe the profile differing significantly between steps. This will allow the Markov chain to explore much more posterior topography but will result in many rejected iterations, where parameters have been chosen that do not describe the profile adequately, resulting in longer run times. It may also have the effect of requiring additional iterations to ensure fine scale posterior topography is adequately explored. A RWSD that is set too small will mean the larger scale topography may not be explored sufficiently resulting in a decrease in precision and, potentially, accuracy. While this suggests that values for RWSD which are either too high or too low can have determimental outcomes, in practise the MCMC can accommodate a broad range of values with little negative effect, but some potentially positive. A demonstration of the effect of varying the RWSD on the $\log_{10}(LR)$ for the four contributor high and low template GloabFiler™ profile is given in Figure 8. The profile was interpreted ten times each using three different values for the RWSD: 0.01, 0.005 and 0.0001. Interpretations were undertaken using eight chains with 50,000 burn-in accepts and 400,000 post burn-in MCMC accepts within STRmix™ version 2.4.02.

Figure 8: $\text{Log}_{10}(LR)$ of ten replicate interpretations of the high template (blue bars) and low template (grey bars) four person GlobalFiler™ profiles, interpreted using eight chains with 50,000 burn-in accepts and 400,000 post burn-in accepts and varying RWSD. Runtime (in minutes) is indicated by the black lines, which correspond to the right hand vertical axis.



Inspection of Figure 8 shows that reproducible *LR*s (within one order of magnitude) were generated using both 0.01 and 0.005. The run times using a RWSD of 0.005 were significantly less however than when using 0.01. The *LRs* assigned using a RWSD of 0.0001 were highly variable indicating STRmix™ had not likely explored the probability space sufficiently. On balance the RWSD value of 0.005 afforded a reproducible LR with a low run time.

We have demonstrated that at least 50,000 burn-in and 400,000 post burn-in accepts across eight chains and a RWSD of 0.005 are suitable MCMC run parameters leading to reproducible *LR*s (within one order of magnitude) for many different types of profile. These settings are likely to be excessive for many one, two and some three person profiles. They will be sufficient for the remaining three and most four person profiles. Decreasing the number of accepts may mean that STRmix™ has not converged and, even with convergence, more variability is expected. Increasing the number of accepts has been shown to help with reproducibility for more complex profiles and will certainly mean higher run times. A summary of the approximate run times for different profiles interpreted using STRmix™ v2.4.02 on a laptop(Windows 7 64 bit, Intel Core i7-5600U CPU, 2.6 GHz, 16 GB RAM) are given in Table 4.

Table 4: Approximate time taken to complete interpretation of various GlobalFiler™ profile types within STRmix™ (hours:minutes:seconds), 8 chains with 50,000 burn-in and 400,000 post burn-in MCMC accepts, and RWSD of 0.005.

Number of contributors	High template profile	Low template profile
Single contributor	0:00:12	0:00:12
Two contributors	0:00:34	0:01:13
Three contributors	0:16:52	0:16:42
Four contributors	1:53:37	1:42:50

In calculating the *LR*, the numerator is the weighted sum of the probability of fewer genotype sets than the denominator. In many cases the numerator may have only one term. Since the denominator is the weighted sum across the probability of many genotype sets it has a stability to variation in the *LR*. However the numerator of the *LR* is more sensitive and this effect is at its greatest when the weight for the numerator genotype set(s) is low. This is most obvious for profiles where the inclusion of a POI requires an improbable peak height variability (observed as large heterozygote balances or dropout) i.e. where the fit of the POI to the profile is poor, or when the inclusion of the POI requires one or more drop-in events to have occurred (which will also increase *LR* variability due to allele proportion uncertainty).

We have demonstrated that higher order mixtures and profiles with low template and/or poor quality lead to a decrease in precision (replication in *LR* across replicates runs). As a general guide, we have observed that if the overall *LR* is greater than 1 and one or more of the locus *LR* values are less than or equal to 1, the POI is likely to have a poor fitting genotype to the observed data at these loci. In these cases the MCMC can be run at 10 or more times the default number of accepts and/or by increasing the RWSD in order to ensure improved precision.

In general, using the default settings as described above, when comparing a POI who is a good fit to the observed profile the difference between the smallest and largest *LR* is small in relation to the size of the *LR*. For profiles where an unlikely stochastic effect has occurred, or the POI is a poor fit to the profile then the difference between the largest and smallest *LR* may be higher but again small in relation to the *LR*. In the 1200 dataset described above (Appendix 2) the largest differences between the smallest and largest log(*LR*) using the recommended run settings was 1.3 fold. For profiles where an unlikely stochastic effect occurred, or the POI was a poor fit to the profile then the difference in log(*LR*) values can be above one. These situations can be minimised or eliminated via policies that suggest increasing iterations based on the profile data.

Reproducibility

Reproducibility is often stated as one of the main principles of the scientific method. A value is reproducible if there is a high degree of agreement between *LR*s run on the same input in different locations by different people. Reproducibility is one component of the precision of a measurement or test method. The other component is repeatability which is the degree of agreement of *LR*s on the same input by the same observer in the same laboratory.

Reproducibility is not intended to mean "exactly the same". Reproducibility means that the results are very similar within the limits of measurement or they lead to the same conclusion. In any real world application we must accept measurements to a degree of resolution and models of a limited level of complexity, or we must accept that the property we are measuring has a degree of variability. A level of uncertainty can exist in a measurement (or model) and yet the measurement can still be informative. In fact science and statistics rely on this fact.

If the same or a different operator interpreted the same input file using STRmix™ with the same random number seed[4] they would obtain exactly the same answer. So why then do we not set the seed and obtain exactly the same answer each time? Strangely this is dishonest repeatability. It would give a false impression of perfect precision. We prefer to give a true measure of our precision.

For very simple situations we can manually calculate the value of the *LR* from the mixture deconvolution part of the software. For the remaining situations, which comprise the vast majority of situations, we can predict limits and patterns but not exact values (for example by referring to plots such as those in Figures 2 through 6). If we retain the concept of a correct, but unknowable, answer, and we plot the output from STRmix™ against these limits the patterns can be assessed to draw conclusion about the function of the STRmix™ models.

Guideline 3.2.4. Case-type Samples

The mixtures described in section 3.2.3 above (Precision) include a range of profile types typically encountered in casework. These profiles include single source and mixed DNA profiles containing up to four contributors generated for both Identifiler™ and GlobalFiler™ profiles. In addition, the developmental validation of STRmix™ involved the testing of a number of profiles generated using other kits and different capillary electrophoresis instruments (3130 and 3500) including ProfilerPlus®, PowerPlex® 21, Fusion, MiniFiler™, SGMPlus™ (profiles amplified at 34 cycles) and NGM Select™ (data not shown). Back stutter is explicitly modelled in all versions of STRmix™ and version 2.4 introduces to modelling of forward stutter. The profiles included contributors with shared alleles. STRmix™ models the variability of single peaks. The variance of this model is determined by directly modelling laboratory data. This is undertaken within STRmix™ using the Model Maker function.

[4] No computer code can actually produce a truly random number. When you tell a computer to generate a sequence of random numbers it draws upon an algorithm that generates what looks like (to humans) as being random, but will eventually start repeating itself. If a computer was told to generate a set of 1000 random numbers twice then it would generate two lists of 1000 seeming random looking numbers, but the lists would be identical. The way to get around this is by providing the algorithm with a random starting value (or 'seed').

Mock samples versus casework

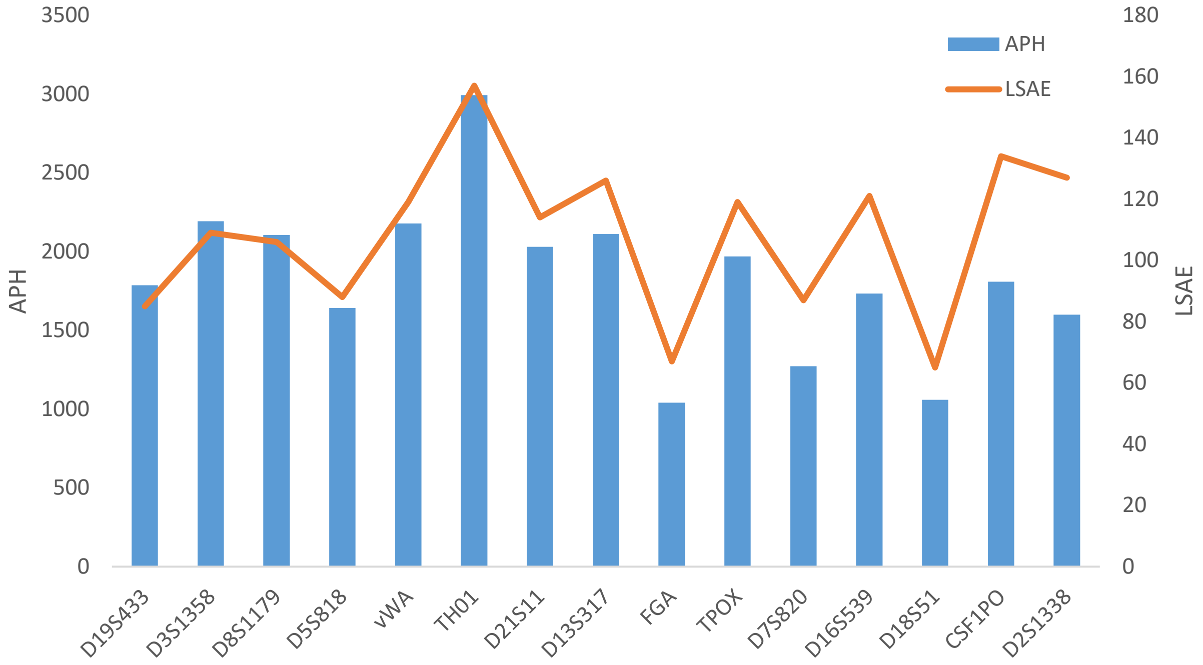
Three experiments have previously been reported comparing the use of mock case samples and casework samples, or single source and mixed DNA profiles, to form interpretation policies [31, 51, 52]. None of the studies found any obvious difference between these sets. This may be the expectation from theory. Peak height is approximately linearly proportional to the number of template molecules sampled. The standard deviation in that peak height is proportional to the square root of the number of template molecules [53, 54].

If we posit that casework has degradation and inhibition effects not modelled with mock samples then we need to see how that would affect the peak heights and their variability. Degradation effectively reduces template from the starting extract but whatever number of quality template molecules survive this number is still the primary explanatory variable for peak height and relative variation. Therefore if 50% of the template was degraded we would expect this to behave similarly to a mock sample with half the template.

The effect of inhibition is more difficult to predict. Inhibitors may bind to the single stranded DNA or to the polymerases or any other co-factor. If they are simply removing template from the reaction then they would act the same as degradation. In any case what we tend to observe is that a whole locus or sets of loci amplify poorly and all peaks are lowered [55]. We could easily see how the relative variability might remain the same. STRmix™ explicitly models locus specific amplification efficiencies (LSAE). The LSAE model reflects the observation that even after template DNA amount, degradation and variation in peak height within loci are modelled, the peak heights between loci are still more variable than predicted, resulting in poorer amplification of some loci possibly due to inhibition. The variance of the LSAE model is determined by directly modelling laboratory data (see [31]). LSAE values for each STRmix™ interpretation appear within the results. We can demonstrate the relationship of LSAE values to average peak heights (APH) via a simple plot. The LSAE values should mimic the average peak heights of the locus if degradation is minimal, otherwise you will see a trend across sets of loci within dye colours according to molecular weight. This is demonstrated for one single source Identifiler™ profile in Figure 9. The differences in APH and LSAE in this figure are due to overall profile degradation which is modelled separately.

We have described above the theoretical expectations from the interpretation of inhibited and degraded profiles using STRmix™. Separately, we have interpreted a number of DNA profiles derived from various mock crime samples such as cigarette butts, bloodstains on wood, touched items and worn clothing. Inspection of the diagnostics from these STRmix™ interpretations including degradation and LSAE values align with our expectations (data not shown).

Figure 9: Plot of APH (bars) and LSAE value (line) for each locus ordered by molecular weight for a single source Identifiler™ profile



Guideline 3.2.6. Accuracy

There is a subset of profiles where the expected answer may be replicated relatively easily by hand. By comparing the software output with the expected answer, the performance and limitations of the software may be examined. An understanding of the models behind the methods is essential for this process. Examples of where we can predict the answer include single source profiles, mixtures where the profile of a major contributor is unambiguous (major/minor) and mixtures of two contributors in equal proportions (balanced). STRmix™ has been shown to give the expected result in each case [48].

Functionality has been installed within STRmix™ to facilitate validation and performance checks. This includes the extended output and set seed functions. The extended output contains all of the parameters and calculated probabilities for each iteration within a run. The 'set seed' function turns off the random processes within STRmix™ and allows direct comparison of runs within and between different versions of the software. STRmix™ is built in two separate parts that communicate via a text file. The first part runs the MCMC, the second the *LR* calculation. Hence, in some version releases it is possible to test one part using an old output from the other part variously using the set seed or checks of the extended output to allow the direct comparison of outputs and lessen the validation load.

The following functionality and outputs from STRmix™ were verified by hand as part of the developmental validation tasks for each commercial version:

1. Expected allele and stutter heights given mass parameters
2. Expected peak heights of drop or 'Q' alleles given mass parameters
3. Probabilities given expected and observed peak heights and varying analytical thresholds
4. Locus specific amplification efficiency calculations
5. Summation of probabilities for each allele in a locus and across a profile
6. Summation of probabilities across multiple replicate profiles
7. Informed priors on mixture proportion

8. *LR* values where there are no assumed contributors
9. *LR* values for propositions with assumed contributors
10. *LR* values with varying theta values
11. Relatives calculations (where a relative is considered as an alternate contributor under H_d)
12. Sampling from the Beta distributions for theta
13. *LR* stratified point estimates
14. *LR* highest posterior density (HPD) interval values
15. Gaussian walk
16. Gelman-Rubin statistic, ESS, weight resampling
17. Drop-in function
18. Database search functionalities
19. Model maker.

The comparison of expected heights, probability and *LR* values was conducted in MS Excel or by comparison to results generated in the r_{HPD} package written by Professor James Curran in R [56].

The likelihood ratios calculated using STRmix™ have been compared to two probabilistic genotyping methods employing semi-continuous models and two binary methods of profile interpretation [48, 57]. Where a profile was able to be fully resolved or for single source profiles where dropout was not a consideration (weight, w_i, equals one at each locus) the *LR* between STRmix™ and the semi-continuous methods were comparable where they were using the same population genetics model. For mixed DNA profiles, generally STRmix™ resulted in higher *LR*s for ground truth known trials as continuous models use more of the profiling information (for example peak height information) compared with semi-continuous and binary interpretation methods.

Conclusion

Within this paper we describe the exercises undertaken as part of STRmix™ developmental validation following the SWGDAM guidelines for the validation of probabilistic genotyping software [1]. This work demonstrates that STRmix™ is suitable for its intended use for the interpretation of single source and mixed DNA profiles including profiles of a complex and low level nature.

A number of different parameters within STRmix™ that are known to affect *LR* reproducibility were investigated. We have interpreted over 1200 profiles and conclude that at least 50,000 burn-in and 400,000 post burn-in accepts across eight Markov chains and a RWSD of 0.005 are suitable STRmix™ run parameters leading to reproducible *LR*s (within one order of magnitude) for many different types of profile.

Having undertaken both internal and developmental validations following the SWGDAM guidelines we find them a good template within which to work. Recommendations 3.2.5 (control samples) and 3.2.6.2 (analysis of raw data files) are not applicable to STRmix™.

Acknowledgements

The authors would like to thank Johanna Veth (ESR), John Simich (Erie County, NY), Shawn Monpetit (San Diego, CA), Bjorn Sutherland (ESR) and two anonymous reviewers for their helpful comments that greatly improved this paper.

References

[1] Scientific Working Group on DNA Analysis Methods (SWGDAM). Guidelines for the Validation of Probabilistic Genotyping Systems. 2015.

[2] Budowle B, Onorato AJ, Callagham TF, Manna AD, Gross AM, Guerreri RA, et al. Mixture Interpretation: defining the relevant features for guidelines for the assessment of mixed DNA profiles in forensic casework. Journal of Forensic Sciences. 2009;54:810-21.

[3] Gill P, Buckleton J. Commentary on: Budowle B, Onorato AJ, Callaghan TF, della Manna A, Gross AM, Guerrieri RA, Luttman JC, McClure DL. Mixture interpretation: Defining the relevant features for guidelines for the assessment of mixed DNA profiles in forensic casework. J Forensic Sci 2009;54(4):810-21. Journal of Forensic Sciences. 2010;55:265-8.

[4] Buckleton JS, Triggs CM, Walsh SJ. DNA Evidence. Boca Raton, Florida: CRC Press; 2004.

[5] Dror IE, Charlton D, Peron AE. Contextual information renders experts vulnerable to making erroneous identifications. Forensic Science International. 2006;156:74-8.

[6] Thompson WC. Painting the target around the matching profile: the Texas sharpshooter fallacy in forensic DNA interpretation. Law, Probability and Risk. 2009;8:257-76.

[7] Scientific Working Group on DNA Analysis Methods (SWGDAM). SWGDAM Interpretation Guidelines for Autosomal STR Typing by Forensic DNA Testing Laboratories. 2010.

[8] Gill P, Brenner CH, Buckleton JS, Carracedo A, Krawczak M, Mayr WR, et al. DNA commission of the International Society of Forensic Genetics: Recommendations on the interpretation of mixtures. Forensic Science International. 2006;160:90-101.

[9] Haned H. Forensim: An open-source initiative for the evaluation of statistical methods in forensic genetics. Forensic Science International: Genetics. 2011;5:265-8.

[10] Haned H, Gill P. Analysis of complex DNA mixtures using the Forensim package. Forensic Science International: Genetics Supplement Series. 2011;3:e79-e80.

[11] Balding DJ, Buckleton J. Interpreting low template DNA profiles. Forensic Science International: Genetics. 2009;4:1-10.

[12] Lohmueller K, Rudin N. Calculating the weight of evidence in low-template forensic DNA casework. Journal of Forensic Sciences. 2013;58(s1):s234-59.

[13] Mitchell AA, Tamariz J, O'Connell K, Ducasse N, Budimlija Z, Prinz M, et al. Validation of a DNA mixture statistics tool incorporating allelic drop-out and drop-in. Forensic Science International: Genetics. 2012;6:749-61.

[14] Taylor D, Bright J-A, Buckleton J. The interpretation of single source and mixed DNA profiles. Forensic Science International: Genetics. 2013;7:516-28.

[15] FBI Quality Assurance Standards for Forensic DNA Testing Laboratories. 2011.

[16] Bright J-A, Taylor D, Curran JM, Buckleton JS. Developing allelic and stutter peak height models for a continuous method of DNA interpretation. Forensic Science International: Genetics. 2013;7:296-304.

[17] Bright J-A, Taylor D, J.M. C, Buckleton JS. Degradation of forensic DNA profiles. Australian Journal of Forensic Sciences. 2013;45:445-9.

[18] Buckleton J, Kelly H, Bright J-A, Taylor D, Tvedebrink T, Curran JM. Utilising allelic dropout probabilities estimated by logistic regression in casework. Forensic Science International: Genetics. 2014;9:9-11.

[19] Puch-Solis R. A dropin peak height model. Forensic Science International: Genetics. 2014;11:80-4.

[20] Brookes C, Bright J-A, Harbison S, Buckleton J. Characterising stutter in forensic STR multiplexes. Forensic Science International: Genetics. 2012;6:58-63.

[21] Bright J-A, Curran JM, Buckleton JS. Investigation into the performance of different models for predicting stutter. Forensic Science International: Genetics. 2013;7:422-7.

[22] Hastings WK. Monte Carlo sampling methods using Markov chains and their applications. Biometrika. 1970;57:97--109.
[23] Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E. Equations of state calculations by fast computing machines. Journal of Chemical Physics. 1953;21:1087-91.
[24] Balding DJ, Nichols RA. DNA profile match probability calculation: how to allow for population stratification, relatedness, database selection and single bands. Forensic Science International. 1994;64:125-40.
[25] National Research Council Report: The evaluation of forensic DNA evidence. Washington DC: National Academy Press; 1996.
[26] Taylor D, Bright J-A, Buckleton J. Considering relatives when assessing the evidential strength of mixed DNA profiles. Forensic Science International: Genetics. 2014;13:259-63.
[27] Taylor D, Bright J-A, Buckleton J, Curran J. An illustration of the effect of various sources of uncertainty on DNA likelihood ratio calculations. Forensic Science International: Genetics. 2014;11:56-63.
[28] Bright J-A, Taylor D, Curran J, Buckleton J. Searching mixed DNA profiles directly against profile databases. Forensic Science International: Genetics. 2014;9:102-10.
[29] Taylor D, Bright JA, Buckleton J, Curran J. An illustration of the effect of various sources of uncertainty on DNA likelihood ratio calculations. Forensic Science International: Genetics. 2014;11:56-63.
[30] Triggs CM, Curran JM. The sensitivity of the Bayesian HPD method to the choice of prior. Science & Justice. 2006;46:169-78.
[31] Taylor D, Buckleton J, Bright J-A. Factors affecting peak height variability for short tandem repeat data. Forensic Science International: Genetics. 2016;21:126-33.
[32] Bright J-A, Curran JM. Investigation into stutter ratio variability between different laboratories. Forensic Science International: Genetics. 2014;13:79-81.
[33] Kelly H, Bright J-A, Buckleton JS, Curran JM. Identifying and modelling the drivers of stutter in forensic DNA profiles. Australian Journal of Forensic Sciences. 2013;46:194-203.
[34] Taylor D, Bright J-A, McGoven C, Hefford C, Kalafut T, Buckleton J. Validating multiplexes for use in conjunction with modern interpretation strategies. Forensic Science International: Genetics. 2016;20:6-19.
[35] Bright J-A, Stevenson KE, Curran JM, Buckleton JS. The variability in likelihood ratios due to different mechanisms. Forensic Science International: Genetics. 2015;14:187-90.
[36] Taylor D, Bright JA, Buckleton J. The 'factor of two' issue in mixed DNA profiles. Journal of Theoretical Biology. 2014;363:300-6.
[37] Taylor D, Buckleton J, Bright J-A. Does the use of probabilistic genotyping change the way we should view sub-threshold data? Australian Journal of Forensic Sciences. 2015:DOI:10.1080/00450618.2015.1122082.
[38] Taylor D. Using continuous DNA interpretation methods to revisit likelihood ratio behaviour. Forensic Science International: Genetics. 2014;11:144-53.
[39] Taylor D, Buckleton J, Evett I. Testing likelihood ratios produced from complex DNA profiles. Forensic Science International: Genetics. 2015;16:165-71.
[40] Daubert et al v Merrell Dow Pharmaceuticals Inc., 509 US 579 (1993). 1993.
[41] Kumho Tire Co. Ltd et al. v. Carmichael et al. In: Court USS, editor. 526 US 1371999.
[42] Bright J-A, Curran JM, Buckleton JS. The effect of the uncertainty in the number of contributors to mixed DNA profiles on profile interpretation. Forensic Science International: Genetics. 2014;12:208-14.
[43] Dørum G, Bleka Ø, Gill P, Haned H, Snipen L, Sæbø S, et al. Exact computation of the distribution of likelihood ratios with forensic applications. Forensic Science International: Genetics. 2014;9:93-101.

[44] Gill P, Haned H. A new methodological framework to interpret complex DNA profiles using likelihood ratios. Forensic Science International: Genetics. 2013;7:251-63.
[45] Haned H, Dorum G, Egeland E, Gill P. On the meaning of the likelihood ratio: is a large number always an indication of strength of evidence? 25th Congress of the International Society for Forensic Genetics. Melbourne, Australia2013.
[46] Kruijver M, Meester R, Slooten K. <em>p</em>-Values should not be used for evaluating the strength of DNA evidence. Forensic Science International: Genetics.16:226-31.
[47] Taylor D, Buckleton J. Do low template DNA profiles have useful quantitative data? Forensic Science International: Genetics. 2015;16:13-6.
[48] Bright J-A, Evett IW, Taylor D, Curran JM, Buckleton J. A series of recommended tests when validating probabilistic DNA profile interpretation software. Forensic Science International: Genetics. 2015;14:125-31.
[49] Gelman A, Rubin DB. Inference from Iterative Simulation Using Multiple Sequences. Statistical Science. 1992;7:457-511.
[50] Gelman A, Carlin JB, Stem HS, Rubin DB. Bayesian data analysis. New York: Chapman & Hall; 1995.
[51] Bright J-A, McManus K, Harbison S, Gill P, Buckleton J. A comparison of stochastic variation in mixed and unmixed casework and synthetic samples. Forensic Science International: Genetics. 2012;6:180-4.
[52] Bright J-A, Turkington J, Buckleton J. Examination of the variability in mixed DNA profile parameters for the Identifiler(TM) multiplex. Forensic Science International: Genetics. 2009;4:111-4.
[53] Gill P, Curran J, Elliot K. A graphical simulation model of the entire DNA process associated with the analysis of short tandem repeat loci. . Nucleic Acids Research. 2005;33:632-43.
[54] Weusten J, Herbergs J. A stochastic model of the processes in PCR based amplification of STR DNA in forensic applications. Forensic Science International: Genetics. 2012;6:17-25.
[55] Bright J-A, Cockerton S, Harbison S, Russell A, Samson O, Stevenson K. The effect of cleaning agents on the ability to obtain DNA profiles using the Identifiler™ and PowerPlex® Y multiplex kits. Journal of Forensic Sciences. 2011;56:181-5.
[56] R Development Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2004.
[57] Bille TW, Weitz SM, Coble MD, Buckleton JS, Bright J-A. Comparison of the performance of different models for the interpretation of low level mixed DNA profiles. ELECTROPHORESIS. 2014;35:3125-33.