Science for Communities

Article:

Russell, L., Cooper, S., Wivell, R., Kerr, Z., Taylor, D., Buckleton, J., & Bright, J. A. (2019). **A guide to results and diagnostics within a STRmix™ report**. *Wiley Interdisciplinary Reviews: Forensic Science*, 1(6), e1354.

This is an **Accepted Manuscript** (final version of the article which included reviewers' comments) of the above article published by **Wiley-Blackwell** at https://doi.org/10.1002/wfs2.1354

A guide to results and diagnostics within a STRmix™ report

Laura Russell[1*], Stuart Cooper[1], Richard Wivell[1], Zane Kerr[1], Duncan Talyor[2,3], John S. Buckleton[1,4], Jo-Anne Bright[1]

[1] *Institute of Environmental Science and Research Limited, Auckland, New Zealand*

[2] *Forensic Science South Australia, Adelaide, South Australia, Australia*

[3] *School of Biological Sciences, Flinders University, Adelaide, South Australia, Australia*

[4] *Department of Statistics, University of Auckland, Auckland, New Zealand*

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

Until recently, forensic DNA profile interpretation was predominantly a manual, time consuming process undertaken by analysts using heuristics to determine those genotype combinations that could reasonably explain a recovered profile. Probabilistic genotyping (PG) has now become commonplace in the interpretation of DNA profiling evidence. As the complexity of PG necessitates the use of algorithms and modern computing power it has been dubbed by some critics as a 'black box' approach. Here we discuss the wealth of information that is provided within the output of STRmix™, one example of a continuous PG system. We discuss how this information can be evaluated by analysts either to give confidence in the results or to indicate that further interpretation may be warranted. Specifically, we discuss the 'primary' and 'secondary' diagnostics output by STRmix™ and give some context to the values that may be observed.

Keywords: diagnostic, probabilistic genotyping, STRmix™

1.0 Introduction

The development of more sensitive forensic DNA techniques means that more DNA profiles are obtained from samples collected from crime scenes. Improvements to both chemistry and detection technology have led to the generation of more mixtures and profiles exhibiting both allelic dropout and drop-in (Bright, Taylor, Gittelson, & Buckleton, 2017; Coble & Bright, 2019). A DNA profile is represented graphically as an electropherogram (epg). An individual possesses two copies of an allele (one inherited from each parent) which are sections of DNA. Forensic DNA profiling targets numerous polymorphic sites in the genome, with either one or two alleles detected at each site depending on whether the donor is a homozygote or heterozygote, respectively. These alleles are represented as coloured peaks within the epg and are separated on the horizontal axis according to their size, measured in molecular weight. The heights of the alleles are approximately proportional to the amount of DNA template (Bill et al., 2005; Edwards, Civitello, Hammond, & Caskey, 1991) however this relationship is affected by a number of factors including the size of the allele and whether or not the sample is degraded.

Forensic DNA profiles may also contain a number of artefact peaks and it is imperative that analysts recognise these during profile analysis. The most commonly encountered artefact is stutter. Stutter forms as a by-product during amplification via polymerase chain reaction (PCR) of the short tandem repeat (STR) loci typically used within forensic DNA testing kits (Brookes, Bright, Harbison, & Buckleton, 2012; Hauge & Litt, 1993; Walsh, Fildes, & Reynolds, 1996). Stutter is generally one repeat unit smaller than the target allele, where it is termed back stutter. Less often, stuttering can result in artefact peaks two repeats smaller (double back stutter) or one repeat larger (forward stutter) than the parent allele (Bright, Buckleton, Taylor, Fernando, & Curran, 2014; Bright, Huizing, Melia, & Buckleton, 2011; Gibb, Huell, Simmons, & Brown, 2009; Krenke et al., 2005). Some loci, notably SE33 and D1S1656, also produce stutter products two base pairs smaller than the parent allele. The degree of stutter formation is related to the type of repeat; STRs with trinucleotide repeat structures (such as D22S1045) are known to stutter more than tetra- and pentanucleotide repeats (Butler, 2012). Stutter appears allelic in all aspects and can confound profile interpretation, particularly in cases where allelic peaks from a minor contributor are present at similar levels to stutter.

The interpretation of forensic DNA profiles involves the determination of those genotype combinations that could reasonably explain the DNA profile. The difficulty in interpreting complex DNA profiles means forensic laboratories are increasingly adopting probabilistic genotyping methods. Probabilistic genotyping (PG) "refers to the use of biological modelling, statistical theory, computer algorithms, and probability distributions to calculate likelihood ratios and/or infer genotypes for the DNA typing results of forensic samples" (Scientific Working Group on DNA Analysis Methods (SWGDAM), 2015). PG methods are now accepted and are in widespread use by the forensic community (Coble & Bright, 2019). Broadly, there are two types of PG methods: fully continuous and semi-continuous. Semi-continuous methods do not use peak heights within the interpretation and do not model artefacts generated during DNA profile generation such as stutter. Fully continuous methods

use more information from within the profile, such as peak heights, to evaluate the probability of a set of peak heights and some have the ability to model artefacts such as stutter.

One example of a fully continuous PG system is STRmix™ (Taylor, Bright, & Buckleton, 2013). The increased use of STRmix™ in courts around the world has meant that there is interest from legal participants and especially the defence, in being able to read and interpret a STRmix™ output. To meet this interest, we publish here an interpretation of the diagnostics output by STRmix™.

2.0 Profile analysis in STRmix™

2.1 Assigning a number of contributors

The interpretation of a DNA profile using STRmix™ starts with the assignment of the number of contributors, N, to the profile. We advocate that this is done in the absence of profiling information from any persons of interest (POI) in a case. However, in circumstances where an individual's DNA is expected to be present (for example, when considering DNA results produced from an intimate swab in a sexual assault case), knowledge of their DNA profile could help to better inform N. When considering crime scene samples, the "true" number of contributors is always unknown and unknowable. It therefore falls to the analyst to utilise their knowledge, experience, and expertise to provide their best estimate of N. Several methods have been proposed to assign the number of contributors (Biedermann, Bozza, Konis, & Taroni, 2012; Haned, Pène, Lobry, Dufour, & Pontier, 2011; Swaminathan, Grgicak, Medard, & Lun, 2015; Taylor, Bright, & Buckleton, 2014). Studies evaluating the performance of some of these have also been published (Biedermann et al., 2012; Coble, Bright, Buckleton, & Curran, 2015). In our experience, analysts most commonly use the Maximum Allele Count method in conjunction with peak height information.

2.2 Deconvolution

Provided that the analyst has been able to assign the number of contributors and decided to progress an interpretation, the first step is to deduce all possible genotype sets that might explain the data, including the consideration of dropout and drop-in. Depending on the number of peaks detected at a locus and the number of assigned contributors, there could be millions or even billions of genotype combinations at a locus. Within STRmix™, settings for the maximum allowable stutter ratios and maximum allowable peak heights for drop-in peaks are used to eliminate unreasonable genotype combinations from further consideration, thereby improving run-time.

The next step is the profile deconvolution. STRmix™ assigns a relative weight to the probability of the epg given each possible genotype combination at a locus. It does this using a statistical method called Markov Chain Monte Carlo (MCMC) which is an iterative re-sampling process. For each iteration, genotype combinations and biological parameters are proposed to describe the profile (Bright, Taylor, Curran, & Buckleton, 2013b). For the

simplest profile (for example, one amplification resulting in one epg) these biological parameters are:

1. DNA amount (template) for each contributor to the profile,

2. The level of degradation for each contributor to the profile, and

3. Amplification efficiencies for each locus within the profile.

Collectively, these are referred to as the *mass parameters*. The per-locus amplification efficiency recognises that not every locus within a profile amplifies equally well resulting in some loci with peak heights that are either higher or lower than average. The general process during each iteration is that a genotype set and a set of values for the mass parameters described above are proposed. These proposed values are used to generate an expected DNA profile (*E*). Biological models are employed for the modelling of expected stutter heights, expected allele heights, and the variance in peak heights, with peak height variability dependent on the kit, number of PCR cycles, and capillary electrophoresis (CE) instrumentation used (Bright, Taylor, Curran, & Buckleton, 2013a; Bright et al., 2013b; Taylor, Bright, Buckleton, & Curran, 2014; Taylor, Buckleton, & Bright, 2016). The expected profile is then compared with the observed profile (*O*), that is the epg, to determine how well the proposed values explain the data.

A probability density is calculated for each peak, *O*, compared with *E*, across the profile. The product across all peaks in the profile is a measure of the model 'fit', i.e. how well the proposed genotype set and parameter values describe the epg. The higher the probability density the better the fit of the parameter values to *O*. The results are then compared with those of the previous iteration. The proposed values for the genotypes and mass parameters are either accepted or rejected depending on the probability density. A new set of genotypes and mass parameters is then proposed and new probabilities are assigned given the new expected profile. Each of these processes is called an iteration and the process to accept or reject a proposed iteration is called Metropolis-Hastings (Hastings, 1970; Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, 1953) acceptance rejection sampling.

The MCMC progresses in two parts: burn-in and post burn-in. Burn-in is a preliminary MCMC that is run to ensure that the post burn-in MCMC begins in an area of high probability space and can be likened to a 'warm-up'. In STRmix™, burn-in is typically undertaken on a number of independent chains (eight by default), with each chain running until it reaches a set number of accepted iterations (100,000 by default). Burn-in begins with a randomly chosen genotype set and fixed mass parameters. The progression of the MCMC is influenced by a 'seed' which is set using a random number generator. Each run has a different seed unless specifically fixed either to conduct validation studies or to repeat an analysis for a justifiable reason. Post burn-in starts at the completion of burn-in. Post burn-in is undertaken on the same number of chains until they each reach a set number of accepted iterations (50,000 accepts per chain by default).

The genotypes sets accepted during post burn-in are tallied. At completion of the MCMC, these values are normalised at each locus so that they range from zero (indicating that the observed data cannot be explained by the proposed genotype set) to one (indicating that this is the only genotype set that explains the DNA profile). Although mathematically unnecessary, the counts are normalised so the values provide an intuitively helpful diagnostic for analysts. These are described as 'weights' and are the primary output of STRmix™.

2.3 Likelihood ratio calculation

Following deconvolution, a likelihood ratio (*LR*) may be assigned for any POI for which profiling data are available. Within the STRmix™ output, per-locus *LR*s are also displayed; as discussed in Section 3.1.3 below, these are one of the primary diagnostics of STRmix™ and should be reviewed by the analyst. The *LR* differs depending on the propositions considered, the allele frequency data used in the calculation, and whether or not a theta correction is applied, and its value. Other settings within STRmix™ will also affect the *LR* calculated, such as whether to take account of certain sources of uncertainty, whether to calculate a stratified *LR*, and whether to consider relatives of the POI as possible sources of the DNA.

A report is generated at the completion of a STRmix™ interpretation. This PDF format report is a record of all the settings used within the interpretation and *LR* calculation. The sections of the report are configurable and may be turned on or off and/or reordered depending on laboratory policy.

In Appendix S1, the epg of a two person mixed GlobalFiler™ DNA profile with approximately equal DNA amounts from the PROVEDIt dataset (Alfonse, Garrett, Lun, Duffy, & Grgicak, 2018) is given. The profile has been interpreted conditioning (or assuming the presence of) one of the known donors using STRmix™. Conditioning on an individual in the deconvolution and subsequent *LR* calculation may be undertaken when their DNA is expected to be present under both the prosecution and defence propositions. For example, DNA from the complainant may be assumed on an intimate swab collected from them. After deconvolution, the resolved profile was compared to one POI. The following propositions were used in the assignment of the *LR*:

H_p: The DNA originated from the complainant and the POI

H_d: The DNA originated from the complainant and one unknown individual.

The STRmix™ report generated following deconvolution of this profile is provided in Appendix S2.

3.0 STRmix™ diagnostics

A number of diagnostics have been included in the STRmix™ interpretation and are written to the report. These diagnostics can help the analyst determine how the interpretation has progressed. They can be used to assess the results to ensure they are suitable for reporting. Diagnostics can help to inform the user on two different aspects of the analysis. Firstly,

diagnostics may inform the user on how well the interpretation has performed in accordance with the underlying models and theory. Diagnostics that fall into this type may indicate whether the analysis has achieved a stable distribution for its parameter values, or how well the observed data can be explained with the biological models used by STRmix™. The other type of diagnostic is the mean posterior values of model parameters. Many of these parameter values (such as genotype set weights, DNA amount, or degradation) can be intuitively related to observable patterns in the epg by the analyst, using their knowledge of DNA profile behaviour. While this second type of diagnostic does not directly inform the user as to how well the analysis has performed, it achieves this same outcome through the alignment of the diagnostic values with the analyst's intuitive expectation.

STRmix™ developers have classified these diagnostics into 'primary' and 'secondary' categories. The primary diagnostics are the weights assigned to genotype sets, the mixture proportions assigned to individual contributors, and, where an *LR* calculation is undertaken, the per-locus *LR* values. Secondary diagnostics include the average log(likelihood) which is the average of the probability density across the post burn-in accepts, the Gelman-Rubin convergence diagnostic to assess the convergence of the independent MCMC chains, and descriptors of the variability of allelic and stutter peak heights in the profile. Each of these diagnostics is discussed in turn within this paper. The primary diagnostics have been classified together as they are thought to be of key importance. They are also areas that are familiar to analysts who have experience with conventional DNA interpretation techniques and can be cross-referenced or reconciled against the observed data. The secondary diagnostics are new concepts to most analysts and are more difficult to check for intuitiveness. There is no 'right' value for each of the secondary diagnostics although there may be a range of expected values dependent on the profile presentation and complexity.

3.1 STRmix™ primary diagnostics

3.1.1 Weights

Weights are the primary output of STRmix™. Genotype sets that best explain the observed data should be given the highest relative weight at a locus. In contrast, genotype sets that do not explain the profile well should be assigned a relatively low weight (or no weight at all). The weights are used in the assignment of any subsequent *LR*.

The weights produced through the MCMC process should make intuitive sense to experienced analysts. After interpretation, it is recommended that the analyst reviews the weights in conjunction with the epg to verify that the accepted genotype combinations and their respective weights are intuitive.

For example, consider locus D8S1179 in the profile displayed in Appendix S1. There are four alleles present with similar peak heights: 11, 12, 13, and 14. The smaller peak visible above the analytical threshold in the 10 position is likely stutter.

As an example, where there are two assumed contributors donating DNA in approximately equal proportions with four distinct alleles detected, there are six possible genotype combinations that could explain the observed profile. These six combinations are listed in Table 1.

Table 1: Genotype combinations and weights following STRmix™ interpretation of the D8S1179 locus of the profile displayed in Appendix S1 assuming two contributors and no conditioning on the complainant's profile.

Contributor 1	Contributor 2	Weight
11,14	12,13	0.344
13,14	11,12	0.176
11,13	12,14	0.15
12,14	11,13	0.137
11,12	13,14	0.116
12,13	11,14	0.077

Given the peak heights of the four alleles at this locus all six genotype sets would be expected to have reasonably similar weights rather than high weight being attributed to one particular genotype set. This can be seen in the genotype weights displayed in Table 1. In this instance, the [11,14] [12,13] genotype set has been assigned a little more weight than the other genotype combinations as it provides a somewhat better explanation of the observed profile. Note however that the other genotype combinations have still been assigned some weight by STRmix™, indicating that it accepted them as possible explanations for the observed profile.

Consider that the case circumstances indicate that one of the two contributor's DNA could reasonably be assumed to be present in the mixed DNA profile, for example if the DNA originates from an intimate swab of which they are the donor. It would be reasonable to proceed with an interpretation conditioned on the genotype of this individual. This information is used within the interpretation and will affect the genotype sets considered during the profile interpretation and the subsequent generation of weights. If, in this example, DNA is assumed to be present from a contributor with genotype 11,13 and their genotype at each locus is held in the contributor 1 position then there is only one possible genotype for the second contributor (12,14) under the assumption of two contributors. We would expect STRmix™ to only accept this genotype combination and assign it a weight of one.

From the weights section of the report in Appendix S2, it can be seen that there are several loci (D16S539, D8S1179, D18S51, D19S433, D5S818, D1S1656 and D12S391) that have one genotype set that has been assigned a weight of one. Given the assignment of two contributors (with DNA in approximately equal proportions) and the assumption of a known contributor under both H_p and H_d, the single genotype sets at these loci should be intuitive to analysts.

For more complex profiles, higher order mixtures, or profiles that include contributors at low template where allelic dropout is considered, the number of possible genotype combinations can be very large. Whilst estimating both the complete list of accepted genotype combinations and their relative weights becomes intractable for an analyst, review of some of the elements of the weights section of the report should still be undertaken. For example, one would expect a large number of genotype combinations to be accepted for an unresolvable four-person mixture, with weights diffused across these possibilities. However, if only one or a few genotype combinations were accepted, we would advocate that the results are closely scrutinised as this would appear to be a counterintuitive result.

3.1.2 Mixture proportions

Mixture proportions (M_x) recorded in the STRmix™ output are an approximation of the proportion of each contributor's DNA in the sample based on template values that are also displayed in the STRmix™ report. The template values are the average of the modes of the accepted iterations of each chain of the post burn-in phase of the interpretation. Mixture proportions (or ratios) have traditionally been used to assist in the assessment of the number of contributors to a mixed DNA profile and to help determine putative genotype combinations to a profile. It is advised that this conventional interpretation approach be maintained as it is an element of the interpretation that should be relatively easy for an analyst to assess intuitively.

As an example, assuming the mixture proportions observed at the D8S1179 locus in the profile provided in Appendix S1 extend across the profile, and assuming the mixture originates from two contributors, an analyst may conclude that the individuals who contributed DNA did so in approximately equal proportions. Following STRmix™ interpretation, the mixture proportions calculated should also reflect this and therefore would be in keeping with the analyst's expectation. This is the case in the example given in Appendix S2 where it can be seen mixture proportions have been determined to be 0.56 and 0.44. This would be a useful indicator that the STRmix™ run has progressed as expected. However, if the resulting mixture proportions in a STRmix™ output indicated highly divergent proportions then this could indicate that the data input into STRmix™ or the weights and other diagnostics require further review. Extreme examples (where M_x is patently unintuitive) can be the result of user error where, for example, the incorrect input file was interpreted, or due to the incorrect assignment of N.

There are instances where STRmix™ gives no or very little 'mass' to a contributor. This may mean that STRmix™ does not require the additional contributor to explain the observed result. This may either be due to the contributor being present at trace levels (peaks just above the analytical threshold), a minor contributor being masked by peaks of other higher level contributors, due to an over assignment in the number of contributors, or due to the presence of close relatives where there is a high degree of allele sharing between contributors and hence masking or allele sharing.

By default, mixture proportions in STRmix™ are uninformed and the per-contributor template values are optimised within the MCMC. However, there is the option of informing the template parameter values within the MCMC using the M_x prior function within STRmix™. This function is useful when a contributor is heavily masked by another or if the number of peaks above analytical threshold from a minor contributor is low and hence provides very little information to drive parameter values within the model, as discussed in in Taylor et al. (Taylor, Buckleton, & Bright, 2017).

3.1.3 Per-locus LRs

The *LR* is the method employed by all current PG software to assign a weight of the evidence for a POI (Coble & Bright, 2019). The *LR* is a ratio of two probabilities that evaluates the evidence given two mutually exclusive propositions:

$$LR = \frac{\Pr\left(O \mid H_p\right)}{\Pr\left(O \mid H_d\right)}$$

Where O is the observed DNA profile, H_p is the prosecution proposition, and H_d is the defence proposition. Typically, H_p is inclusionary with respect to the POI whilst H_d is exclusionary. The profile *LR* and the per-locus *LR*s are displayed in Appendix S2 under 'Summary of LR' and 'Per Locus Likelihood Ratio', respectively.

For a given POI, an *LR* is calculated for each contributor order position. There are N! contributor orders for a profile originating from N individuals, when there are no conditioned contributors. For example, there are six possible contributor orders for a three-person mixture (3!=6). Within the report, the POI is displayed in the position giving the largest *LR*. Consider a mixture with fully-resolved major and minor contributors. If the POI corresponds with the major component, we would expect STRmix™ to assign a strong inclusionary *LR* in the major contributor position and, very likely, an exclusionary *LR* (or *LR* strongly favouring exclusion) in the minor contributor position.

The prosecution hypothesis considers the POI to be a contributor to the DNA profile. A value of one for $\Pr\left(O \mid H_p\right)$ in the first column of the Per Locus Likelihood Ratios table is obtained when the contributors specified under the prosecution scenario fully explain the recovered profile. For example, this could be a locus in a single-source profile where only one genotype combination which corresponds to the POI reference genotype is considered and assigned a weight of one or with reference to the example in Appendix S2, all of the evidence is explained given the prosecution proposition of the assumed contributor and the 'POI'. The defence proposition typically considers the probability of the evidence if the DNA was not from the POI but rather from some other individual. An *LR* greater than one provides support for the prosecution proposition, whereas an *LR* less than one supports the alternative proposition. An *LR* of one is 'neutral' or 'uninformative'. The verbal equivalent for *LR*=1 is variously neutral or uninformative depending on the scale used (Buckleton, Bright, & Taylor, 2016; Scientific Working Group on DNA Analysis Methods, 2018). An *LR* of 0 indicates that

the evidence cannot be explained by, or is extremely unlikely, given the prosecution proposition.

The *LR*s in the 'Per Locus Likelihood Ratio' section of the report in Appendix S2 are all above 1 and hence support inclusion of this POI. However, in a different sample or with a different POI there may be instances where one or more individual locus *LR*s are closer to one or favour exclusion (<1) and this will affect the overall *LR* accordingly.

If most loci favour inclusion (*LR*>1), but one locus displays a very low *LR* or an *LR* of 0, as is the case in the example shown in Figure 1, this indicates that further review of the epg and the interpretation is warranted. An exclusion or low *LR* at one locus may be the correct result or could be due to an error within the input file (for example, retention of an artefact peak). This incorrect information can lead to incorrect genotype combinations being assigned weight and hence causes a false exclusion at this locus.

PER LOCUS LIKELIHOOD RATIOS

	FBI_EXTENDED_CAUC 0.01b(1.0, 1.0)		
LOCUS	Pr(E|Hp)	Pr(E|Hd)	LR
D3S1358	9.8909E-2	6.5292E-3	1.5149E1
vWA	7.0375E-2	2.5484E-3	2.7616E1
D16S539	3.2225E-2	2.6089E-3	1.2352E1
CSF1PO	5.7573E-2	1.2647E-2	4.5523E0
TPOX	2.8867E-1	3.3915E-3	8.5117E1
Yindel			
D8S1179	4.0287E-3	1.9606E-4	2.0548E1
D21S11	1.1023E-5	4.5721E-7	2.4110E1
D18S51	3.0300E-2	1.0461E-3	2.8966E1
DYS391			
D2S441	7.3027E-2	7.4445E-3	9.8094E0
D19S433	7.2828E-3	5.9207E-4	1.2300E1
TH01	4.5686E-2	1.0416E-2	4.3860E0
FGA	1.5251E-2	9.7418E-4	1.5656E1
D22S1045	2.0798E-2	2.2380E-3	9.2929E0
D5S818	1.8797E-1	7.1539E-2	2.6274E0
D13S317	2.9475E-2	1.9223E-3	1.5334E1
D7S820	4.0045E-2	2.6943E-3	1.4863E1
SE33	2.8293E-4	2.7409E-6	1.0322E2
D10S1248	7.4243E-2	6.6427E-3	1.1177E1
D1S1656	0	4.5010E-4	0
D12S391	5.3955E-3	1.6154E-4	3.3401E1
D2S1338	1.8884E-2	7.3032E-4	2.5858E1
SUB-SUB-SOURCE LR			0
SUB-SOURCE LR			0
99% 1-SIDED LOWER HPD INTERVAL			0

Figure 1: Per locus *LR*s assigned to a POI evaluated with a two person mixture displaying a false exclusion at D1S1656

The *LR* column in Figure 1 suggests all loci provided an *LR* favouring inclusion, except D1S1656, which has been assigned an *LR* of 0. A review of the epg indicates that the exclusion is likely due to a one base pair CE resolution issue. An allele appears to be present but falls in the shoulder of a stronger peak and has not been detected by the profile analysis software (Bright et al., 2018; Moretti et al., 2017). Rework options that attempt to resolve the two peaks could be considered. Alternatively, settings within the profile analysis software can be adjusted to try and improve resolution. If these approaches are unsuccessful, one option available within STRmix™ is to ignore the locus during deconvolution. Prior to ignoring any locus from the deconvolution we advocate a full review of the results to ensure justifiable concordance between the questioned and known samples.

One low or zero *LR* at a single locus has also been observed when profiles with peaks above the saturation threshold have been interpreted. This results in higher than expected stutter peaks which are then modelled as being allelic in origin. In addition, pull up peaks that are retained in the STRmix™ input file can also cause false exclusions at a single locus.

3.2 STRmix™ secondary diagnostics

The secondary diagnostics are summarised in the 'Post burn-in summary' section of the STRmix™ report as displayed in Appendix S2. The key secondary diagnostics are discussed in turn.

3.2.1 Log(likelihood)

The log(likelihood) value displayed in the report is the average log(likelihood) across all chains for all iterations of the post burn-in phase of the MCMC. The log(likelihood) is a summary of fit of the modelled or expected profile compared with the observed profile. In general, the larger this value, the better STRmix™ has been able to describe the observed data. A low or negative value suggests that STRmix™ has not been able to describe the data very well given the information it has been provided. There are a number of reasons why this may be:

1. The profile is very low-level (partial or trace)
    a. If there are few peaks in the profile, there would be only a limited number of values used within the calculation of the log(likelihood) at each iteration.
    b. If the peaks are low, the probability density may be low due to the increased variability associated with low peak heights.
2. There are large stochastic events in the DNA profile (e.g. large unexpected heterozygote peak imbalances or variation in mixture proportions across the profile). These may simply be due to stochastic events during amplification or could be forced by under-assigning the number of contributors.
3. Data have been removed that was real, particularly stutter peaks, and must be explained as dropout.

4. Artefact peaks have been left labelled and must be explained as drop-in.

A low or negative average log(likelihood) diagnostic may indicate to the user that the interpretation requires additional scrutiny. However, a low or negative log(likelihood) alone does not necessarily negate the use of the results, referring to points 1 or 2 above as examples.

3.2.2 Gelman-Rubin

During the MCMC process STRmix™ uses multiple independent chains within the MCMC (eight chains by default (Bright et al., 2016)) to efficiently explore the probability space for genotype sets, mass parameters, and allele and stutter peak height variance parameters. At the end of a STRmix™ deconvolution, the within-chain variance and between-chain variances of the log(likelihood) values are determined. These variances can be used to calculate the Gelman-Rubin (GR) convergence diagnostic, $\hat{R}$. This is a common mathematical indicator used in MCMC processes (Gelman & Rubin, 1992) to determine if all the chains have sampled from the same probability space and therefore have *converged*. If chains have diverged to different spaces the variation *between* the chains is likely to be larger than the variation *within* the chains and this can result in an $\hat{R}$ greater than 1.2. This can indicate that:

1. The interpretation has not run for a sufficient number of MCMC accepts and the values determined during the MCMC have not reached equilibrium, or
2. One or more chains has become stuck in different parts of the probability space

If each chain is sampling from the same space (i.e. they have converged) then the intra- and inter-chain variances should be approximately equal, and $\hat{R}$ should be close to 1.

A GR less than 1.2 does not guarantee that all chains have converged and conversely a GR greater than 1.2 does not indicate that the results are invalid. As with the other secondary diagnostics, the GR should be interpreted in the context of all of the results.
A slightly elevated GR may be the expected outcome for a complex, higher order mixture. The inherent complex nature of such a profile means that there are many possibilities for the chains to explore and the default number of accepts may not be sufficient to allow STRmix™ to explore all of these. In this scenario, an analyst may opt, prior to or following STRmix™ interpretation, to increase the number of accepts, either manually in the STRmix™ 'Run Settings' or use the 'Auto-extend' feature of the software. With the latter option, STRmix™ runs a check at the end of the post-burn-in phase. If the GR is greater than a user-specified value (for example, 1.2) then a set number of additional accepts will be performed in an attempt to allow all the chains to converge.

An unexpected and infrequently observed outcome of a STRmix™ run can be a GR value grossly in excess of 1.2. Rare instances of single-source samples giving GR values in excess of 3 have been reported by STRmix™ users. Given the relatively simple nature of a single-source, high template profile, a GR greater than 1.2 would not be the expected result. In such instances, one or more chains have likely taken a very different route through the probability

space compared with the other chains. This can lead to counter intuitive genotype combinations and weights relative to the observed DNA profile. In these circumstances, re-running the input file again, using the default number of accepts but with a different seed will likely lead to the chains converging on a similar high probability space and will lead to results that are more intuitive with respect to the observed data.

3.2.3 Peak height variance parameter

STRmix™ models assume that, as expected peak height decreases, peak height variability increases. This relationship should be intuitive for most forensic scientists and is why a stochastic threshold is generally used in traditional interpretation methods. Typically, for profiles with low-template or degraded DNA there is increased variation in peak heights within a profile and between replicate amplifications due to stochastic effects. Variation in peak heights can also be affected by the platform and conditions used within a laboratory, including profiling kit, amplification volume, PCR cycle number, CE instrument, and injection protocol. The range of expected variation in peak heights is determined as part of the STRmix™ implementation process. This prior distribution is determined for both alleles and all modelled stutter types using a function within STRmix™ called Model Maker (Taylor, Buckleton, et al., 2016). Model Maker models the variability in peak heights of a range of single-source samples with known genotypes that encompass a wide range in profile quality.

During a profile deconvolution using STRmix™, peak height variability parameters are sampled within the MCMC. The average of the allele and the average of the stutter variance constants across all post burn-in iterations are displayed numerically in the Post Burn-in Summary within the STRmix™ report. They are also graphically overlaid (as a black dot) on a plot of their respective prior distribution, where c^2 is the allele variance parameter, and k^2 the stutter variance parameters for each stutter type modelled. These values can be used as a diagnostic to determine the amount of peak height variability, and hence profile quality, STRmix™ has settled on in its analysis of the profile. Examples of these plots are given on the second page of the report provided in Appendix S2.

If the one or more variance parameters for a deconvolution are significantly larger than the mode of the respective prior distribution, then this may indicate the profile is not being explained well. A more thorough review of the profile and the STRmix™ deconvolution is recommended. Elevated values for these variance constants may be due to the DNA profile showing high stochastic effects or the assigned number of contributors may be incorrect.

Used in conjunction with the average log(likelihood), elevated allele or stutter variance constants can indicate poor PCR. If the sample is simply low-level this should result in a low average log(likelihood) and variance constants close to the mode of the relevant prior distribution. If some data have been omitted or artefact labels retained in the input file this may result in a low or negative average log(likelihood) and high variance constants. Within the plots output to the STRmix™ report, the variance constants will likely lie in the right hand tail of the prior distributions.

An example of an excessively large back stutter variance (k^2) and negative log (likelihood) from a single-source profile is displayed in the Post Burn-in Summary excerpt of a STRmix™ report shown in Figure 2**Error! Reference source not found.**. The elevated posterior mean stutter variance value relative to the mode can be observed in the plots provided.

POST BURN-IN SUMMARY

Total iterations	533,820	Acceptance rate	1 in 1.33
Effective sample size	2,292.43	log(likelihood)	-58.89
Gelman-Rubin convergence diagnostic	1.18		
Allele variance (mode = 11.985)	18.679	Back Stutter variance (mode = 12.003)	262.087
Forward Stutter variance (mode = 11.985)	22.752		

VARIANCE CHARTS

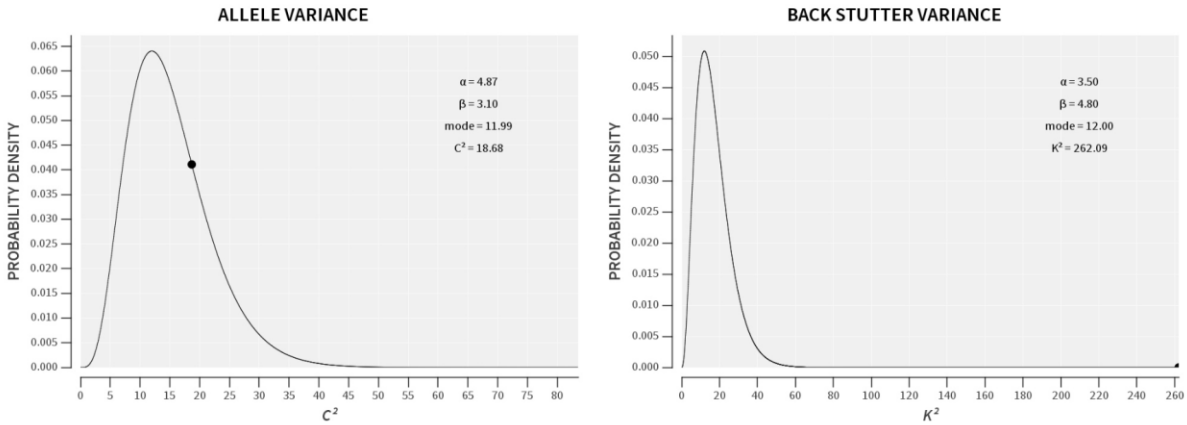

Figure 2: Excerpt of a STRmix™ report displaying the post burn-in summary and variance distributions

Re-examination of the input file that was analysed in STRmix™ and led to the values displayed in Figure 2 revealed that a stutter filter had been applied at profile analysis prior to STRmix™ interpretation resulting in no stutter peaks being present in the STRmix™ input file. During deconvolution, STRmix™ had to explain the absence of expected stutter peaks by invoking dropout, leading to the elevated back stutter variance and negative log(likelihood) diagnostics observed.

Other aspects of a profile that can result in larger than expected variance constants include:

1. the inability to resolve an allelic or stutter peak that is 1 base pair from an adjacent allelic peak during CE,
2. peaks within a profile above the CE instrument saturation threshold, or
3. due to differences in expected and observed stutter peak heights given different sequence variants of particular alleles (for example, the 14 allele at vWA).

Conclusion

As discussed in (Taylor, Bright, et al., 2016), despite many labs moving to PG, analysts still need to be trained and experienced in manual interpretation methods in order to properly assess whether the results are intuitive.

The nature of the MCMC process within STRmix™ allows the user to interrogate several elements of the deconvolution enabling analysts to have a degree of confidence in the deconvolution and the output. Several primary and secondary diagnostics have been written into the STRmix™ output.

The primary diagnostics include genotype weights, mixture proportions, and per-locus *LR*s. The results of these primary diagnostics can be reviewed and checked by experienced DNA analysts.

The secondary diagnostics inform how well the MCMC has progressed within a STRmix™ interpretation. We advocate that an interpretation is closely scrutinised if the diagnostics are not intuitive or are out of range, especially if any of the primary diagnostics are involved. It is advisable to evaluate all resulting diagnostic values in the context of the profile being interpreted and consider further options subsequent to this.

If one or many diagnostics do not align with expectation given the observed profile, there are a number of options to interrogate the results further. These include:

- Checking the interpretation setup, for example the input files, the kit selected and any conditioning profiles
- Checking whether any of the other diagnostics are not intuitive or are out of range
- Checking that the number of contributors, N, reflects the best estimate for the profile
- Checking the input file does not contain any artefact peaks that should have been removed at analysis. In addition, checking that all peaks are included in the input file, including all stutter variants that are to be modelled and peaks that are separated by 1 base pair.
- Checking whether the interpretation requires an increased number of MCMC accepts to allow the MCMC process to reach equilibrium.

It is important that diagnostic values are not used in isolation to authenticate or discount results as invalid. Rather, these should be considered in combination with the other indicators discussed in this article.

In this article we have outlined, in part, how to read a STRmix™ output. Scientists are trained to read these outputs and rapidly become familiar with them. We intend this article to assist the legal participants in any court proceeding to be able to assess the information in these reports.

Acknowledgements

This work was supported in part by grant 2017-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 their organizations.

Appendix S1 & S2: Supplementary Information

Supplementary information related to this article can be found, in the online version, at https://doi.org/10.1002/wfs2.1354.

References

Alfonse, L. E., Garrett, A. D., Lun, D. S., Duffy, K. R., & Grgicak, C. M. (2018). A large-scale dataset of single and mixed-source short tandem repeat profiles to inform human identification strategies: PROVEDIt. *Forensic Science International: Genetics, 32*, 62-70. doi:10.1016/j.fsigen.2017.10.006

Biedermann, A., Bozza, S., Konis, K., & Taroni, F. (2012). Inference about the number of contributors to a DNA mixture: Comparative analyses of a Bayesian network approach and the maximum allele count method. *Forensic Science International: Genetics, 6*(6), 689-696. doi:10.1016/j.fsigen.2012.03.006

Bill, M., Gill, P., Curran, J., Clayton, T., Pinchin, R., Healy, M., & Buckleton, J. (2005). PENDULUM—a guideline-based approach to the interpretation of STR mixtures. *Forensic Science International, 148*(2–3), 181-189. doi:http://dx.doi.org/10.1016/j.forsciint.2004.06.037

Bright, J.-A., Buckleton, J. S., Taylor, D., Fernando, M. A. C. S. S., & Curran, J. M. (2014). Modelling forward stutter: towards increased objectivity in forensic DNA interpretation. *ELECTROPHORESIS, 35*(21-22), 3152-3157. doi:10.1002/elps.201400044

Bright, J.-A., Huizing, E., Melia, L., & Buckleton, J. (2011). Determination of the variables affecting mixed MiniFiler DNA profiles. *Forensic Science International: Genetics, 5*(5), 381-385. doi:10.1016/j.fsigen.2010.08.006

Bright, J.-A., Richards, R., Kruijver, M., Kelly, H., McGovern, C., Magee, A., . . . Buckleton, J. S. (2018). Internal validation of STRmix™ – A multi laboratory response to PCAST. *Forensic Science International: Genetics, 34*, 11-24. doi:https://doi.org/10.1016/j.fsigen.2018.01.003

Bright, J.-A., Taylor, D., Curran, J. M., & Buckleton, J. S. (2013a). Degradation of forensic DNA profiles. *Australian Journal of Forensic Sciences, 45*(4), 445-449.

Bright, J.-A., Taylor, D., Curran, J. M., & Buckleton, J. S. (2013b). Developing allelic and stutter peak height models for a continuous method of DNA interpretation. *Forensic Science International: Genetics, 7*(2), 296-304. doi:http://dx.doi.org/10.1016/j.fsigen.2012.11.013

Bright, J.-A., Taylor, D., Gittelson, S., & Buckleton, J. (2017). The paradigm shift in DNA profile interpretation. *Forensic Science International: Genetics, 31*, e24-e32. doi:10.1016/j.fsigen.2017.08.005

Bright, J.-A., Taylor, D., McGovern, C. E., Cooper, S., Russell, L., Abarno, D., & Buckleton, J. S. (2016). Developmental validation of STRmix™, expert software for the interpretation of forensic DNA profiles. *Forensic Science International: Genetics, 23*, 226-239.

Brookes, C., Bright, J.-A., Harbison, S., & Buckleton, J. (2012). Characterising stutter in forensic STR multiplexes. *Forensic Science International: Genetics, 6*(1), 58-63. doi:10.1016/j.fsigen.2011.02.001

Buckleton, J. S., Bright, J.-A., & Taylor, D. (2016). *Forensic DNA Evidence Interpretation* (2nd ed.). Boca Raton: CRC Press.

Butler, J. M. (2012). Chapter 5 - Short Tandem Repeat (STR) Loci and Kits. In J. M. Butler (Ed.), *Advanced Topics in Forensic DNA Typing* (pp. 99-139). San Diego: Academic Press.

Coble, M. D., & Bright, J.-A. (2019). Probabilistic genotyping software: An overview. *Forensic Science International: Genetics, 38*, 219-224. doi:https://doi.org/10.1016/j.fsigen.2018.11.009

Coble, M. D., Bright, J.-A., Buckleton, J. S., & Curran, J. M. (2015). Uncertainty in the number of contributors in the proposed new CODIS set. *Forensic Science International: Genetics, 19*, 207-211. doi:10.1016/j.fsigen.2015.07.005

Edwards, A., Civitello, A., Hammond, H. A., & Caskey, C. T. (1991). DNA typing and genetic mapping with trimeric and tetrameric tandem repeats. *American Journal of Human Genetics, 49*(4), 746-756.

Gelman, A., & Rubin, D. B. (1992). Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science, 7*(4), 457-511.

Gibb, A. J., Huell, A.-L., Simmons, M. C., & Brown, R. M. (2009). Characterisation of forward stutter in the AmpFlSTR® SGM Plus® PCR. *Science & Justice, 49*(1), 24-31. doi:http://dx.doi.org/10.1016/j.scijus.2008.05.002

Haned, H., Pène, L., Lobry, J. R., Dufour, A. B., & Pontier, D. (2011). Estimating the Number of Contributors to Forensic DNA Mixtures: Does Maximum Likelihood Perform Better Than Maximum Allele Count? *J Forensic Sci, 56*(1), 23-28. doi:10.1111/j.1556-4029.2010.01550.x

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika, 57*, 97--109.

Hauge, X. Y., & Litt, M. (1993). A study of the origin of 'shadow bands' seen when typing dinucleotide repeat polmoorphisms by the PCR. *Human Molecular Genetics, 2*(4), 411-415.

Krenke, B. E., Viculis, L., Richard, M. L., Prinz, M., Milne, S. C., Ladd, C., . . . Budowle, B. (2005). Validation of a male-specific, 12-locus fluorescent short tandem repeat (STR) multiplex. *Forensic Science International, 148*(1), 1-14.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics, 21*, 1087-1091.

Moretti, T. R., Just, R. S., Kehl, S. C., Willis, L. E., Buckleton, J. S., Bright, J.-A., . . . Onorato, A. J. (2017). Internal validation of STRmix for the interpretation of single source and mixed DNA profiles. *Forensic Science International: Genetics, 29*, 126-144. doi:10.1016/j.fsigen.2017.04.004

Scientific Working Group on DNA Analysis Methods. (2018). Recommendations of the SWGDAM Ad Hoc Working Group on Genotyping Results Reported as Likelihood Ratios. Retrieved from https://docs.wixstatic.com/ugd/4344b0_dd5221694d1448588dcd0937738c9e46.pdf

Scientific Working Group on DNA Analysis Methods (SWGDAM). (2015). Guidelines for the Validation of Probabilistic Genotyping Systems. Retrieved from http://media.wix.com/ugd/4344b0_22776006b67c4a32a5ffc04fe3b56515.pdf

Swaminathan, H., Grgicak, C. M., Medard, M., & Lun, D. S. (2015). NOCIt: A computational method to infer the number of contributors to DNA samples analyzed by STR genotyping. *Forensic Science International: Genetics, 16*, 172-180. doi:http://dx.doi.org/10.1016/j.fsigen.2014.11.010

Taylor, D., Bright, J.-A., & Buckleton, J. (2013). The interpretation of single source and mixed DNA profiles. *Forensic Science International: Genetics, 7*(5), 516-528. doi:http://dx.doi.org/10.1016/j.fsigen.2013.05.011

Taylor, D., Bright, J.-A., & Buckleton, J. (2014). Interpreting forensic DNA profiling evidence without specifying the number of contributors. *Forensic Science International: Genetics, 13*, 269-280. doi:http://dx.doi.org/10.1016/j.fsigen.2014.08.014

Taylor, D., Bright, J.-A., Buckleton, J., & Curran, J. (2014). An illustration of the effect of various sources of uncertainty on DNA likelihood ratio calculations. *Forensic Science International: Genetics, 11*, 56-63. doi:http://dx.doi.org/10.1016/j.fsigen.2014.02.003

Taylor, D., Bright, J.-A., McGovern, C., Hefford, C., Kalafut, T., & Buckleton, J. (2016). Validating multiplexes for use in conjunction with modern interpretation strategies. *Forensic Science International: Genetics, 20*, 6-19. doi:http://dx.doi.org/10.1016/j.fsigen.2015.09.011

Taylor, D., Buckleton, J., & Bright, J.-A. (2016). Factors affecting peak height variability for short tandem repeat data. *Forensic Science International: Genetics, 21*, 126-133. doi:http://dx.doi.org/10.1016/j.fsigen.2015.12.009

Taylor, D., Buckleton, J., & Bright, J.-A. (2017). Does the use of probabilistic genotyping change the way we should view sub-threshold data? *Australian Journal of Forensic Sciences, 49*(1), 78-92. doi:10.1080/00450618.2015.1122082

Walsh, P. S., Fildes, N. J., & Reynolds, R. (1996). Sequence analysis and characterization of stutter products at the tetranucleotide repeat locus vWA. *Nucleic Acids Research, 24*, 2807-2812.