STRmix™ Glossary

STRmix™ – a fully continuous probabilistic genotyping forensic software which combines biological modeling with mathematical processes in order to (1) assist with interpretation and attempt to deconvolute DNA profiles in the presence or absence of conditioned samples, and (2) compare suspect/informative reference samples (comparison samples) to evidence samples and provide statistical weight in the form of a likelihood ratio (LR).

- The deconvolution is performed using a Markov Chain Monte Carlo process which considers possible genotype combination(s). Each combination is assigned a weight which reflects how well it explains the evidence profile.

- LRs are calculated by comparing the probabilities of two hypotheses, H1 and H2 (Hp and Hd, in STRmix™). STRmix™ incorporates the assigned weights and sub-population models (Balding and Nichols, 1994, also known as NRC II recommendation 4.2) to calculate the LR.

Mass parameters – variable parameters used to generate expected peak heights during a deconvolution, collectively referred to as **DR BAT**:

Degradation rate for each contributor to the DNA profile
Replicate amplification strength for each PCR replicate
Balance between amplification kits
Amplification efficiencies for each locus within the profile
Template for each contributor (measured in relative fluorescence units (RFU))

Model Maker – a module of the software that provides an estimation of the STRmix™ parameters for an STR amplification kit using empirical data. The parameters within Model Maker must be determined before casework samples can be analyzed in STRmix™.

Weighting or Weight – a probability that reflects how well a particular genotype combination explains the evidence profile. For example, if a proposed combination of genotypes is unlikely to lead to the observed evidence profile, then that combination will be given a low weighting (close to zero).

MCMC – Markov Chain Monte Carlo is an algorithm based on standard mathematical principles that assigns a likelihood for each genotype combination. This method uses a random re-sampling process in order to give a best explanation for an observed set of data.

- **Markov chain –** a process that steps from one position to another. In STRmix™, the positions are defined by a combination of genotypes and parameters that are proposed as explanations for the data. At each iteration, the algorithm will compare the likelihood of the

current iteration to the previous iteration and then determine if it will step to the new iteration or stay at the current iteration. The chain is the path of steps and stays.

- **Markov property** – the property of the Markov chain where the values of each iteration are independent of the values in the previous iteration.

- **Monte Carlo –** a random re-sampling method used to address complex problems. It is used to systematically search the entire range of possibilities that are being considered to ensure that all likely combinations are considered. In STRmix™, this range of possibilities consists of a series of proposed genotypes for each given locus, along with mass parameters.

Iteration – a proposed genotype combination and set of mass parameters that are either accepted or rejected during the MCMC process.

Accepts or Moves – an iteration that is accepted and 'moved' towards the next step in the MCMC process.

- **Burn-in** – during the initial phase of the MCMC process, a set number of accepts are discarded. This is done to allow each chain the time to reach a desired or 'good space'.

Metropolis-Hastings algorithm – an equation used to determine whether or not to accept or reject a new iteration based on the likelihood of evidence data given the proposed genotype combination and mass parameters.

SUMMARY OF CONTRIBUTORS

Template (rfu) – the best estimate of the "amount" (proportion) of DNA, expressed in rfu, determined by the mean of all post burn-in accepts of the calculated template (t) for each contributor.

Mixture Proportion (%) – approximate percentage of each contributor to the sample based on template (rfu) amount.

Degradation – modelled as a negative exponential curve for each contributor, the decreasing trend of peak heights with increasing molecular weight – a higher number indicates a steeper slope of degradation.

- **Degradation start point –** the point in bp where degradation is first applied at the smallest molecular weight peak observed in the sample.

- **Linear approximation (rfu/bp) –** linear model of how many rfu each contributor degrades per base pair increase.

- **Degradation curve** – plot of the degradation modelled for each contributor

POST BURN-IN SUMMARY

Total iterations – total number of proposed genotype and parameter combinations (steps and stays) made during the MCMC process.

Acceptance rate – total number of accepts during the MCMC process, divided by total iterations.

Effective sample size (ESS) - the number of independent samples that have been taken from the posterior distribution of the MCMC likelihood.

- **Effective Sample Size thinning –** the number of values STRmix™ uses in the ESS calculation i.e. if there were 2 million iterations and the ESS thinning setting is 1 million then the ESS calculation will be performed by thinning out every second value. This assists with run time. If the number of iterations is less than the ESS thinning value then STRmix™ uses all of them. The default value is 1,000,000.

log(likelihood) - this value is the average $\log_{10}$(likelihood) for the entire post burn-in MCMC.

Gelman-Rubin convergence diagnostic - this value informs the user whether the MCMC chains have converged. This is calculated by comparing the within-chain and between-chain variances of the MCMC chains.

LSAE variance (σ^2) – (Locus specific amplification efficiency) variance constant that describes how variable the different loci amplified during the run. The likelihood is penalized for loci which have locus amplification efficiencies that differ from the mean of the other loci values.

Allele Variance (c^2) – variance constant that gives an indication of the level of stochastic variation amongst allelic peak heights within a sample.

Stutter Variance (k^2) – variance constant that gives an indication as to the level of stochastic variation amongst stutter peak heights within a sample.

Inter replicate efficiency – the comparison of the variation in sampling between replicates**.** This value is reported as a percentage.

LOCUS EFFICIENCIES

LSAE (Locus Specific Amplification Efficiencies) – how well each locus in the sample amplified in comparison to other loci within the sample; this is modeled as one of the variable parameters within the MCMC process.

MCMC SETTINGS

Mx Priors – setting that allows for the approximate mixture proportion (%) and variance of each contributor to be set prior to sample interpretation.

Number of chains –set number of chains used for the MCMC process.

MCMC accepts – The number of MCMC accepts required before the MCMC finishes.

Random walk standard deviation (RWSD) – sets the step size distributions for the random Gaussian walks. During the MCMC, the next iteration will be close but not too close to the previous iteration.

- When RWSD is small – STRmix™ takes smaller steps which leads to more accepts and is quicker.
- When RWSD is large – STRmix™ takes larger steps which leads to less accepts and is slower but will more likely step across valleys.

Post burn-in shortlist – The log(likelihood) used to remove genotypes sets from the post burn-in list.

KIT SETTINGS

Detection threshold(s) – lab-specific analytical threshold (AT) determined during internal validation studies.

Saturation – all data above this level is only considered qualitatively. The value for saturation is determined during internal validation studies and is expected to be specific to the model of electrophoresis instrument used.

Degradation starts at – the point in base pairs where degradation is first applied to the profile.

Degradation max – the maximum allowable degradation for any one contributor during the entire MCMC process.

Drop-in cap – maximum height of a drop-in allele (in rfu) permitted per locus. STRmix™ will not model an allele as drop-in if it is above the drop-in cap.

Drop-in rate parameter – laboratory observed rate of drop-in observed during internal validation.

Drop-in parameters – parameters such as analytical threshold, peak height, observed drop-in rate, and probability of drop-in that are used to model drop-in in STRmix™; described as a gamma distribution (α, β).

Minimum variance factor – the minimum allowable value the allele and stutter variance constants can take in relation to the mode of their prior distributions. For example, if the mode is 4.2 and the minimum allowed variance from the mode setting is 0.5, then the smallest value the variance constant can take is 4.2 x 0.5 = 2.1.

Variance minimization parameter – this value is used within the calculation of the log(likelihood).

LSAE variance parameter – the locus amplification variance parameter was determined through the internal validation using Model Maker and is expected to be affected by laboratory specific variables. It is used to correct for variation in amplification efficiencies in order to pull locus expected peak heights back towards the whole profile expected heights.

Allelic variance parameters – used to describe how variable allele peak heights are during the STRmix™ run. These were determined through the internal validation using Model Maker and are expected to be affected by laboratory specific variables.

Maximum stutter ratio – the maximum allowable stutter proportion permitted for a specific stutter type, i.e., 0.3 = 30%. Setting stutter max = 0 turns this parameter off. The maximum stutter parameter should be determined by the laboratory through internal validation studies and is expected to be affected by laboratory specific variables.

Stutter variance parameters – Used to describe how variable stutter peak heights are during he STRmix™ run. These were determined through the internal validation using Model Maker and are expected to be affected by laboratory specific variables.

LR SETTINGS

Assign sub-source LR (formerly Factor of N!) – When this value is set to "yes", the sub-source LR is provided, which provides statistical weight to the comparison of the reference sample (i.e., victim or suspect) compared to a mixture as a whole, as opposed to a specific contributor order.

Highest Posterior Density (HPD) – defines the interval most likely to contain the true value; used when calculating a likelihood ratio.

- **HPD iterations –** the number of iterations used within the Highest Posterior Density calculation to create the probability interval.
- **HPD quantile –** the quantile used within the HPD calculation for the probability interval.
- **HPD sides –** The number of sides used within the HPD calculation for the probability interval (1 or 2).

Use MCMC uncertainty – When set to "Y", STRmix™ considers genotype set weights as distributions and re-samples from these distributions during the LR HPD calculation.

Use Allele Frequency uncertainty – When set to "Y", STRmix™ considers allele frequencies as distributions and re-samples from the distributions during the LR HPD calculation.

SUMMARY OF LR

(All LRs in this table are sub-source LRs: 99.0% 1-sided lower HPD calculated from 1000 iterations, MCMC uncertainty on, Allele frequency uncertainty on.)

Unrelated LR (formerly Total LR) – LR for each population. Contributors within H1 (H_p) and H2 (H_d) are assumed to be unrelated individuals.

Relationship LRs- LR which considers the evidence being explained in H2 (H_d) by a relative of the POI (person of interest) in H1 (H_p).

Unified LR – LR that takes into account that the unknown contributors in H2 (H_d) are made up of both relatives of the POI and unrelated people.

Stratified LR – when multiple populations are selected to calculate an LR, STRmix™ will calculate LRs for each population individually and then provide a single LR that is a weighted average across all populations.

PER LOCUS LIKELIHOOD RATIOS

Per Locus Likelihood Ratios - For each locus, the probability of the evidence given H1 (H_p) and H2 (H_d) are individually listed, as well as the ratio of the two probabilities (LR).

SUB-SUB-SOURCE LR (formerly LR total) - combined LR for all loci for the contributor order with the highest calculated LR.

SUB-SOURCE LR (formerly Factor of N!) – LR which provides statistical weight to the comparison of the POI to the mixture as a whole, regardless of contributor order.

Values listed as the 99.0% 1-SIDED LOWER HPD INTERVAL - The combined LR calculated by STRmix™ is referred to as a point estimate. Because the true answer is not known, a distribution is then applied around the point estimate known as a highest posterior density (HPD) credible interval. This interval accounts for the uncertainty (allele frequency and MCMC) associated with the point estimate LR. This interval, commonly applied in Bayesian statistical calculations, gives a range of where the true LRs

actually lie. The lower end of the calculated HPD interval is reported from STRmix™ to be conservative to the person of interest.

Miscellaneous

Seed – starting number used within the random number generator. Setting the seed to the same number will return the same results for a sample from run to run. It is used as a QA measure and to assist with validation and performance checks.

Q allele – signifies a genotype possibility that includes a dropped allele which is not observed in the evidence profile.

Primary diagnostics- diagnostics that can be intuitively approximated by an experienced analyst such as weights, LRs, and mix ratio/proportion.

Secondary diagnostics- diagnostics that need to be evaluated based on STRmix™ output rather than intuitive knowledge such as allele variance, stutter variance, log(likelihood), effective sample size, and Gelman-Rubin convergence diagnostic.