# STRmix$^{TM}$ Glossary

**STRmix$^{TM}$**: a fully continuous probabilistic genotyping forensic software which combines biological modeling with mathematical processes in order to (1) interpret 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 creates 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$^{TM}$). STRmix$^{TM}$ 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 **DART**:

**D**egradation rate for each contributor to the DNA profile
**A**mplification efficiencies for each locus within the profile
**R**eplicate amplification strength for each PCR replicate
**T**emplate for each contributor (measured in relative fluorescence units (RFU)

**Model Maker –** estimation of the STRmix$^{TM}$ 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$^{TM}$.

**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 a mathematical method that 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$^{TM}$, the positions define certain combinations of genotypes and parameters that are being tried as explanations for the profile. At each iteration, the algorithm may either step or stay put. The chain is the path of steps and stay puts.

- **Markov property**: the property of the MCMC analysis where the values in each step are independent of the values in the previous step.

- **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$^{TM}$, this range of possibilities consists of series of proposed genotypes for each given locus.

**Iteration** – a proposed genotype combination that is either accepted or rejected during the MCMC process. These proposed genotype combinations are not the final result.

**Accepts or Moves –** an iteration that is accepted and 'moved' towards the next step in the MCMC process. This is defaulted in STRmix$^{TM}$ to 500,000 accepts

- **Burn-in**—During the MCMC process, the first 100,000 accepts are discarded. This is done to allow the chains time to reach a desired or 'good space'.

**Metropolis-Hastings algorithm:** The Metropolis-Hastings algorithm is an equation used to determine whether or not to accept or reject a new profile combination.

**Summary of Contributors**

**DNA Amounts (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 Proportions (%) -** approximate percentage of each contributor to the sample.

**Degradation –** estimated for each contributor, the decreasing trend of peak heights with increasing molecular weight – a higher number indicates a steeper slope of degradation.

- **Degradation starting at (rfu/bp) -** the lowest RFU peak at the smallest molecular weight locus in the sample.

**Run Information**

**Total Iterations -** total number of proposed genotype combinations made during the MCMC process.

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

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

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

**Average (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.

**Allele Variance ($c^2$) -** Parameter that describes how variable allele peak heights are within the run

**Stutter Variance ($k^2$) -** Parameter that describes how variable stutter peak heights are within the run

**Seed value -** 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 and should only be used to assist with validation and performance checks.

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

**Locus efficiencies (also referred to as Locus Specific Amplification Efficiencies, or LSAE) -** 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

**<u>Parameters</u>**

**Minimum allowed Variance from the mode –** 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.

**Locus amplification variance (LSAE variance, $\sigma^2$) -** A variable that penalizes loci which have locus amplification efficiencies that differ from the mean of the other loci values. 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 toward s the whole profile expected heights.

**Maximum stutter** – the maximum allowable back stutter proportion permitted, i.e. 0.30 = 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.

**Forward stutter max** – the maximum allowable forward stutter proportion permitted, i.e. 0.05 = 5%. Setting forward stutter max = 0 turns this parameter off. The forward stutter max parameter should be determined by the laboratory by internal validation studies and is expected to be affected by laboratory specific variables.

**Drop-in cap** – maximum summed heights of drop-in allele(s) (in RFU) permitted per locus. STRmix$^{TM}$ will not model an allele as drop-in if it is above the drop-in cap.

**Drop-in frequency** – 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$^{TM}$; described as a gamma distribution ($\alpha_3$, $\beta_3$).

**Random Walk Standard Deviation-** 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$^{TM}$ takes smaller steps which leads to more accepts and is quicker.
- When RWSD is large – STRmix$^{TM}$ takes larger steps which leads to less accepts and is slower, but will more likely step across valleys.

**Effective Sample Size thinning** – the number of values STRmix$^{TM}$ 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$^{TM}$ uses all of them. The default value is 1,000,000.

**MCMC Accepts-** The number of MCMC acceptances required (including burn-in) before the MCMC finishes. The default value is 500,000.

**Maximum degradation** – the maximum allowable degradation for any one contributor during burn-in.

**Saturation** – All data above this level is only considered qualitatively. For example, if Saturation = 8000 RFU and the evidence input file contains some peaks with heights >8000 RFU, STRmix$^{TM}$ will convert these peak heights to 8000 RFU and run accordingly. The value for saturation is determined during internal validation studies and is expected to be specific to the model of electrophoresis instrument used.

**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. The default setting is 1000.
- **HPD Significance value –** the percentile used within the HPD calculation for the probability interval. The default setting is 99.0.
- **HPD Sides –** The number of sides used within the HPD calculation for the probability interval (1 or 2).

**Factor of N! –** When this value is set to "yes", the DNA profile of a comparison sample (i.e. victim or suspect) is compared to the mixture as a whole. When this value is set to "no", the DNA profile of a comparison sample is compared to each component of the mixture.

**MCMC Uncertainty –** When set to "yes", STRmix$^{TM}$ considers genotype set weights, allele frequencies and theta as distributions and re-samples from these distributions during the LR HPD calculation. This acts as an additional layer of conservatism.

**MCMC Chains-** a Markov chain is a process that steps from one position to another. The positions define certain combinations of genotypes and parameters that are being tried as explanations for the profile. At each decision the algorithm may either step to a new position (accept) or stay (reject the new position) at the same position. The chain is the path of steps and stays.

**Summary of LRs (All LRs in this table are 99.0% 1-sided lower HPD)**

**Total LR-** total 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 comparison sample in H1 ($H_p$).

**Unified LR –** LR that takes into account that the unknown contributors within H1 ($H_p$) and H2 ($H_d$) are made of up of both relatives and unrelated people.

**Stratified LR –** when multiple populations are selected to calculate an LR, STRmix$^{TM}$ will calculate LRs for each population individually and then provide a single LR that samples across all populations.

## Per Locus Likelihood Ratios

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

**LR total -** combined LR for all loci.

**Factor of N! –** LR which compares the POI to the total mixture, regardless of contributor order.

**99.0% 1-sided lower HPD -** The combined LR calculated by STRmix™ is referred to as a point estimate. Because the true answer is not known, a confidence interval is then applied around the point estimate known as a 1-sided 99.0% highest HPD credible interval. This interval accounts for the uncertainty associated with the point estimate LR. This interval, commonly applied in Bayesian statistical calculations, gives a range (i.e. with 99% confidence) of where the true allele probabilities actually lie. The lower end of the HPD interval is reported from STRmix™ to be the most conservative to the person of interest.

## Miscellaneous

**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$^{TM}$ output rather than intuitive knowledge such as allele variance, stutter variance, average log likelihood, and Gelman-Rubin diagnostic.