

## Validation and Performance Check

---

### Background

STRmix was internally validated and implemented at SDPD Crime lab in 2015, and subsequently upgraded in 2017. The software regularly gets updated by the developers, sometimes incorporating minor changes, and other times adding or changing substantial elements. Depending on the extent of the changes, a performance check or validation is required before the newer version can be used in casework. At SDPD, STRmix v2.4.06 was being used in casework at the time of v2.6 release. This validation and performance check covers STRmix v2.6.2. Below is a summary of important changes made to the STRmix software since the implementation of v2.4.06. Comprehensive lists of changes can be found in the Release and Testing Reports compiled by ESR for each software version update.

#### Summary changes from v2.4 to v2.6 (changes particularly affecting SDPD)

- Changes to pre-burnin logic resulting in speed and memory improvements, as well as speed and memory savings when performing database searching and LR calculations.
- Addition of **generalized stutter** to the biological model (specifically allowing multiple stutter types, including n-2 and n-0.5 repeat stutter to be modeled).
- Independent variance parameters for each type of stutter modeled during Model Maker analysis. Each type of stutter now has its own variance parameters.
- Change in the peak variability model to incorporate the **Taylor Quantum Effect**.
- Addition of minimum expected stutter ratio of 0.001.
- Expanded functionality of **batch mode** (e.g., Mx Priors can be run in batch mode, MCMCs can be batched with LRs and any other analysis, and batch will run in user defined order).
- Ability to drag and drop files for input setup.
- Reference input files no longer require nucleotide size of the alleles.
- LR (likelihood ratio) batcher – the ability to set up comparisons of multiple reference files to an MCMC, and/or to multiple MCMCs in one batch.
- **H<sub>d</sub>-True Tester (HdTT)** – a separate tool incorporated into the STRmix user interface to generate numerous random profiles to compare to MCMCs.
- Nomenclature change for LRs on report: “LR Total (point LR)” is now referred to as the “sub-sub-source LR” and “factor of N!” is now referred to as the “sub-source LR” (HPD LR remains the same).
- Ability to incorporate F<sub>ST</sub> (theta) into database search LRs.
- Database (DB) search includes AMEL genotypes.
- The ability to model a **variable number of contributors (VarNOC)** – two deconvolutions, one each performed under both values of N, and a calculation of weights between the two models can be made.

- PDF report formatting changed for all analyses, PDF reports now made for database searches, and PDF reports can be made at any time after analysis is done.

This list includes several things affecting the biological model (which are considered more substantial changes), such as the Taylor Quantum Effect and modeling additional types of stutter. STRmix v2.4.06 models one repeat unit reverse stutter ( $n-1$ ) and one repeat unit forward stutter ( $n+1$ ). DNA results generated with the GlobalFiler (GF) kit frequently have both of these types of stutter at almost all loci, but also exhibit half repeat reverse stutter ( $n-\frac{1}{2}$ ) and double reverse stutter ( $n-2$ ) occasionally. Until this version of STRmix, these additional stutter peaks were filtered, or edited, in GeneMapper ID-X (GMID-X) prior to exporting, if the peaks are not being considered allelic. In STRmix v2.6, the half repeat reverse stutter ( $n-\frac{1}{2}$ ) is termed two base pair back stutter (2bp Back Stutter) and is observed routinely at the SE33 and D1S1656 loci in the GF kit. Two repeat unit reverse stutter ( $n-2$ ) have not been filtered in GMID-X when using STRmix v2.4.06; however, if a peak in an  $n-2$  location fell within  $n-2$  stutter ratio expectations, it could be edited out manually before exporting the data for STRmix v2.4.06 analysis. Both of these types of stutter can now be modeled using STRmix v2.6. Figure 1 is a graphical depiction of the STRmix v2.6 model. One change implemented between v2.4.06 and v2.6 is the ability to combine results from multiple kits. This feature functions similarly to replicate analyses within a single kit, but expands the capabilities to data obtained from amplifications of the same extract with different amplification kits. The multiple kit (Kit 2 in Figure 1) analysis is not currently being considered as part of the workflow for this lab, so the right half of the diagram can be disregarded for SDPD analyses.

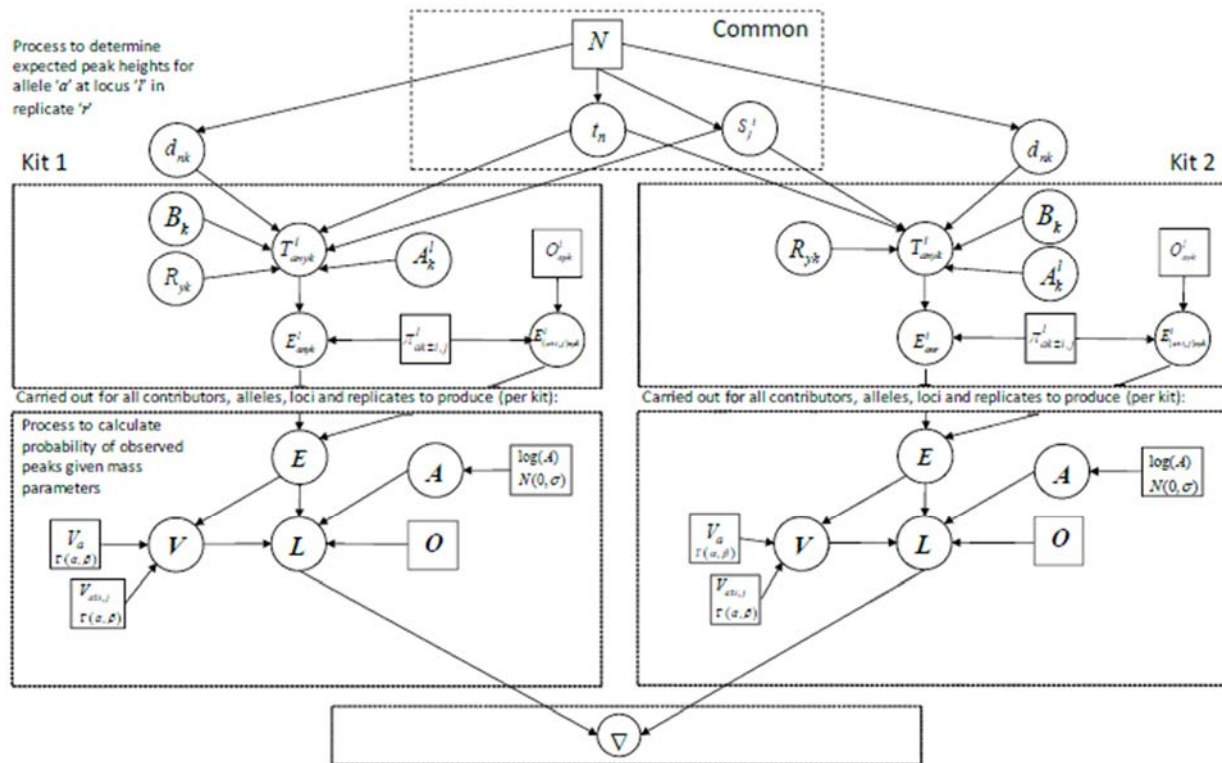


Figure 1 - From the STRmix v2.6 user's manual; a graphical depiction of the STRmix model.

In addition to biological model changes, the STRmix v2.6 user interface is very different from v2.4; however, the general process is very similar – STRmix is still applying the biological model to build profiles using proposed mass variables and comparing them to the observed data over thousands of iterations (MCMC process). Comparisons can still be made to the data, resulting in likelihood ratios. The STRmix v2.6 MCMC and LR reports also look very different from a report generated with v2.4; however, they still contain the same information as the previous version, namely mixture proportions, allele and stutter variances, average log(likelihood), genotype set weights, LRs and best fit contributor order (in LR reports).

## Scope of the upgrade to STRmix v2.6

SDPD currently has the ability to employ STRmix to deconvolute data typed with the Identifiler, Identifiler Plus, MiniFiler (analyzed on the 3500), and GlobalFiler kits. The upgrade to STRmix v2.6 requires different degrees of change for each of these kits. These are outlined in the Model Maker summary; but are summarized here as well:

### *MiniFiler*

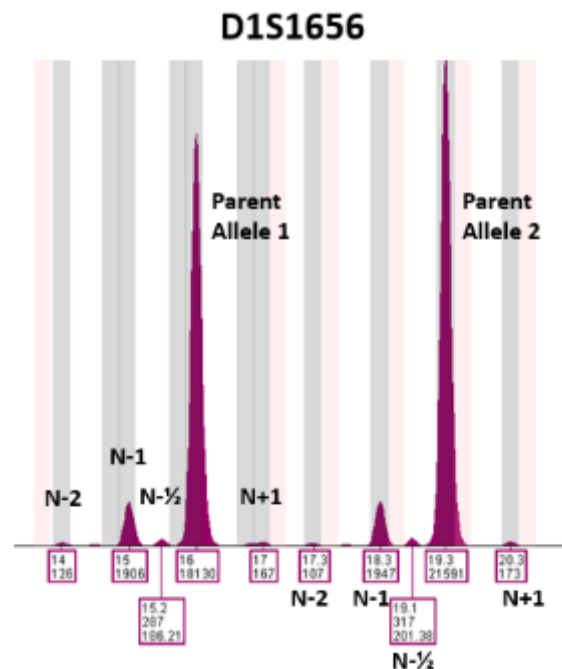
The STRmix kit for MiniFiler does not require any changes in parameters for deconvolutions in STRmix v2.6 because no additional stutter is being modeled. As such, the previous parameters were used to create a new kit (for compatibility purposes) within v2.6 to deconvolute MiniFiler data. The scope of the upgrade for MiniFiler constitutes a performance check to determine whether the STRmix v2.6 deconvolutions produce equivalent results to v2.4.06.

### *Identifiler/Identifiler Plus*

The STRmix kit for Identifiler Plus does require an additional variance parameter for  $n+1$  stutter. Based on that need, Model Maker was re-run using Identifiler Plus samples from the original validation (v2.4) and new variance parameters were obtained. The scope of the upgrade for Identifiler Plus constitutes a validation because of the new set of variance parameters generated by a new Model Maker analysis.

### *GlobalFiler*

The STRmix kits for GlobalFiler and GlobalFiler 24 second data also required Model Maker analysis to generate independent variance parameters for  $n+1$  stutter, as well as  $n-\frac{1}{2}$  and  $n-2$  stutter, which the lab opted to model. The scope of the upgrade for both the GlobalFiler kits constitutes a validation because of the new, expanded set of variance parameters



*Figure 2* - A single source sample with high parent allele peaks to show all types of stutter that may be observed in GlobalFiler amplifications.

generated by a new Model Maker analysis.

For analysis in STRmix v2.6, GlobalFiler kit results require re-analysis using a new combination of GMID-X analysis method/panel so that the n-½ and n-2 repeat stutter peaks are no longer filtered, see Figure 2 for an example where all possible stutter peaks are labeled. Note that in this example, all types of stutter are labeled for both heterozygote peaks, but also that the signal of the parent alleles is unusually high (~20,000 RFU) and results of this intensity are rarely observed in casework samples unless they are over-amplified.

## Descriptions of features added since STRmix v2.4.06 affecting SDPD analyses

### *Generalized Stutter*

The generalized stutter incorporated into STRmix v2.6 allows STRmix to model expected stutter peaks in addition to n-1 and n+1 repeat stutter. STRmix can now model n-2 repeat and n-½ repeat stutter. The model for total allelic product calculations is identical to the model used in STRmix v2.4.06; however, the calculations for the *expected height of an allele* now includes reductions in height due to the additional stutter peaks. The formula for the expected peak

heights is as follows:  $E_{anyk}^l = \frac{T_{anyk}^l}{1 + \sum \pi_{ak \pm i, j}^l}$ , where the sum of all expected stutter ratios are

added are considered, which has the effect of removing height from the total allelic product. For peaks in loci that have n-½ repeat stutter modeled, the expected peak height is calculated as follows:

$$E_{anyk}^l = \frac{T_{anyk}^l}{1 + SR_{a-2}^l + SR_{a-1}^l + SR_{a-1/2}^l + SR_{a+1}^l}.$$

With generalized stutter, the STRmix model also allows an option for determining the likelihood of peaks by having them be inversely proportional to either the height of the observed allelic peak or the height of the expected stutter peak (see the v2.6 model maker write-up for additional information). The likelihood of peaks is another area of the STRmix model that has been slightly altered by the addition of the Taylor Quantum Effect.

### *Taylor Quantum Effect*

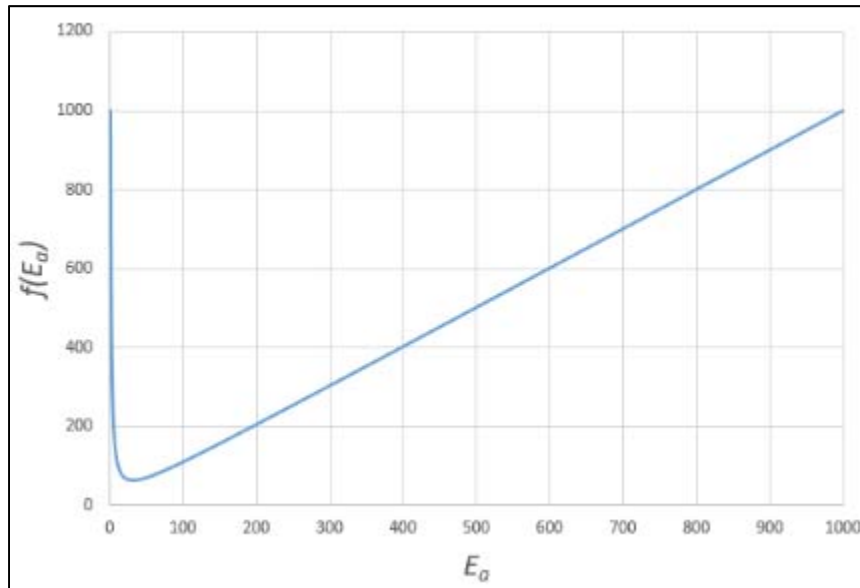
The likelihood of peaks within STRmix has been modeled in past versions by:

$\log\left(\frac{O_a}{E_a}\right) \sim N\left(0, \frac{c^2}{E_a}\right)$ , for allelic peaks, and  $\log\left(\frac{O_a}{E_a}\right) \sim N\left(0, \frac{c^2}{O_a}\right)$ , for stutter peaks. This

essentially detailed that as expected peak heights decreased, the more variability was expected. A limit on this variability was instituted in previous versions of STRmix such that the variability would not be infinite. This limit was incorporated such that the expected peak height for a peak could never be less than half the analytical threshold. The Taylor Quantum Effect has been implemented to improve on this variance model. The Taylor Quantum Effect limits the variance model such that as expected peak heights become increasingly small, the allowable variability decreases. This adjustment only affects peaks that are modeled based on expected

heights. By definition, if a peak is observed it is above the analytical threshold; however, STRmix v2.6 incorporates the same correction for both for consistency in calculations, but it will have no effect on observed peaks.

The Taylor Quantum Effect now models variance by:  $\log\left(\frac{O_a}{E_a}\right) \sim N\left(0, \frac{c^2}{f(E_a)}\right)$ , where  $f(E_a) = \frac{b}{O_a} + O_a$ , for back stutter, and  $f(E_a) = \frac{b}{E_a - x} + E_a - x$ , for all other stutter peaks, where  $b = 1000$  (a constant variable determined by developmental validation). This effect will allow for variability to increase (i.e., will allow for more peak height variability) until the expected peak height decreases to 32 RFU, at which point the allowable variability will decrease. The effect of



this is to limit unreasonable peak height differences between observed and expected for low level peaks. This is accomplished because the variance is divided by  $f(E_a)$ , and as the peak heights of the observed or expected peak heights diminish the standard deviation of the normal distribution will broaden, until the peak heights reach the inflection point, as demonstrated in Figure 3.

*Figure 3 - Graphical representation of the variance models in v2.6 with Taylor Quantum Effect applied*

### Batch Mode

In v2.4.06, batch Mode could be used to batch multiple MCMCs into sequential independent MCMC runs, but other functions (e.g., LRs) could not be included in a batch, and the order was determined alphanumerically. Batch mode in v2.6 has improved substantially. First, the minor glitch that did not allow for running MCMCs using informed priors was fixed. Second, an LR batcher function was added. This allows the setup of a comparison of one or more reference samples to one or more MCMCs in using only one screen. Third, batch mode can now accommodate analyses of any kind, including LR batcher setups, in a user defined order. A batch timer and counter also included on the run screen, so batch progress can be better monitored. LR batcher and batch mode are user interface changes and typically don't require in-depth performance check testing; however, these elements will significantly change the workflow at SDPD, so both LR batcher and batch mode were tested using multiple kits, multiple reference files, DB searches, and H<sub>d</sub>-True Tester analyses in different combinations and with different number of random profiles in DB search and H<sub>d</sub>-True Tester as part of this performance check.

*H<sub>d</sub>-True Tester and Database Search*

An H<sub>d</sub>-true test is a test calculating an LR for a profile that is known to not be a contributing component of the sample. At SDPD with STRmix v2.4, each casework MCMC is accompanied by a database search consisting of over 10,100 samples. A small portion of the database are elimination samples used to identify potential contamination events. The vast majority of the samples in this database are randomly generated profiles (i.e., profiles that theoretically could not be contributing to any evidence profiles). The reason to compare all deconvoluted samples against the 10,000 profiles generated randomly from the NIST combined allele frequencies was to provide context for comparisons to DNA results for persons of interest. In the SDPD process, when a sample is deconvoluted, a database search is performed against the 10,000 random profiles as H<sub>d</sub>-true tests, and an LR is calculated for each random profile in the database. In a simple, very discriminating DNA result (such as a robust, single source sample), LRs for each of the random profiles is likely to be 0. If the randomly generated profile happens to have the same DNA types as the ones found in the robust evidence (an extremely unlikely event), though, the LR will be very high. Not all DNA results are discriminating single source samples, however. At SDPD, the majority of DNA results are mixtures, and it is not uncommon for one or more components of these mixtures to have ambiguous contributing genotype possibilities, so the results of a 10,000 random database search can be quite different, with perhaps several LRs for random profiles (known non-contributor profiles) above 1. This is illustrated in the specificity tests of a wide range of validation mixtures (see STRmix original validation and 5 person addendum specificity results). The results of these H<sub>d</sub>-true tests are dependent on the discrimination of the evidence sample and the number of random profiles being compared to the mixture.

The goal of performing H<sub>d</sub>-true testing is to get an estimate of the discrimination power of the DNA profile. When there are random profiles that have LRs >1, it provides information that there are ambiguous component(s) to an evidence sample. This may be because of the complexity of the mixture, or one or more contributors having low level contribution with drop-out being possible at many loci. In assessing the context for LRs from persons of interest, H<sub>d</sub>-true testing using 10,000 randomly generated profiles is a limited tool. The limitations arise from the fact that when profiles are randomly generated for non-contributor testing, the vast majority of them will produce LRs of zero against the evidence profile, even profiles with ambiguous components. This is observed routinely in casework, where there are no random samples with LRs above 1 in complex mixtures with low level component(s) and high probability of drop-out from one or more components. As stated previously, the results of an H<sub>d</sub>-True test are dependent on the qualities of the sample used in the test and the number of random profiles tested. More contextual information could be obtained if more random profiles were tested against the evidence sample; however, the more samples that are created and compared against the evidence require more computing power to perform the computations; therefore using this approach is limited by computing capacity.

The H<sub>d</sub>-True Tester (H<sub>d</sub>TT) tool was first released as part of the changes to STRmix v2.5. This tool that compares randomly generated profiles to evidence deconvolutions. There are two options that can be employed when using this tool. The first option of H<sub>d</sub>TT is called random sampling,

which generates results that are almost identical to using a database of 10,000 (or another chosen value) random samples. The only difference is that a new set of random profiles is randomly generated, using the allele frequencies from a specified population, each time the tool is run. An LR for each of the random samples is generated using this tool, and the range of LRs can give information regarding the discrimination potential of the deconvolution and provide context for LRs of POIs. Just like the static DB searching the SDPD currently employs, the number of random samples can be increased at the expense of computational time, but the same limitations discussed above apply. For a robust single source sample, the number of random profiles generated before producing a profile that give an  $LR \geq$  the LR of the true contributor is estimated to be approximately  $1/LR$ . Based on the profile probabilities, random profiles favoring inclusion to the test sample are generated only a very small proportion of the time; therefore, the information obtained about the profile from random profile sampling is limited.

The second option of the  $H_d$ TT tool is importance sampling, which helps to address the limitation of  $H_d$ -True testing described in the previous paragraph. To obtain random profiles that would, by chance alone, fit into the mixture deconvolutions when doing  $H_d$ -true tests would often require generating billions to quadrillions of random profiles, which is computationally impractical. Importance sampling uses only the accepted genotype sets from the deconvolution to generate random profiles used for comparison. To compensate for selecting genotypes at a higher rate than random, the process of importance sampling mathematically corrects for the selective oversampling by factoring in: 1) the probability of selecting a particular contributor (generally  $1/\#$  unknowns in the profile); 2) the profile probability of the selected profile (with additional considerations when that profile contains a Q allele), and 3) the product of the weights across all loci. Importance sampling provides better context for LRs by providing an estimate of probability of obtaining any LR above 1, given the deconvolution. Importance sampling also provides an indication on the performance of the model by providing the probability of samples obtaining an LR higher than the LR of a POI. As stated in the STRmix user's manual, "this probability value is presented as a frequentist sounding interpretation to the LR is actually a statement that follows from the laws of probability." The output from an  $H_d$ TT using importance sampling includes a chart and table stating the probability (1 in x) of obtaining a range of LRs. The output also provides the number of effective iterations, average LR for the profiles compared in the  $H_d$ TT, max LR, and min LR (these will be non-zero LRs due to the importance sampling from within the accepted genotypes) for each deconvolution. These are described in detail in the results section.

There are several ways to incorporate this feature and the proposed incorporation of this tool will affect the casework workflow. The  $H_d$ TT tool would allow for random sample comparisons separately from searching the elimination database. This would enhance the efficiency of our analytical process because samples with no references for comparison would only be searched against the elimination database (of approximately 200 samples) at the time of deconvolution and would not have the  $H_d$ TT test performed until reference samples were submitted for comparison. This efficiency combined with the efficiencies of batch mode are benefits of upgrading STRmix to v2.6.

*Variable Number of Contributors (VarNOC)*

An important and sometimes challenging element of forensic DNA interpretation is assigning the number of contributors to a DNA result. In casework this number is ALWAYS unknown. Even when a result appears to be single source, there is always the possibility of masking and/or drop-out. We can create mixtures from known DNA extracts in the lab for validation, but perceived number of contributors to a mixture is not always consistent with the known number of contributors. This could be a result of masking of alleles, low contributor template amount, elevated stutter, or drop-in. Inaccurate contributor number interpretations may be diagnosed through STRmix diagnostic values outside of expected ranges, or could be diagnosed when comparisons made to reference DNA profiles indicate an additional contributor is required to explain the data. In these instances, two separate deconvolutions are generally performed and all diagnostics are evaluated to inform the determination of the number of contributors. This potential uncertainty with the number of contributors to a mixture is addressed with a new feature in STRmix v2.6 called the “Variable Number of Contributors function” (VarNOC).

With VarNOC, samples are deconvoluted using a contributor range, with the maximum difference between number of contributors (N) being 1. Due to the computer processing requirements of running two separate deconvolutions on the same evidence profile, and the information it provides, VarNOC is not recommended to be used routinely, and should be considered after other potential avenues to address the number of contributors to the sample have been explored (e.g., additional amplification). VarNOC is a tool that can be used to compare two separate deconvolutions when the number of contributors is ambiguous or there is a need to have a likelihood ratio reported that has a different number of contributors in the numerator than the denominator. VarNOC produces two separate independent analyses, each with separate genotype weights, mixture proportions, and diagnostics, and provides information on the probability of the number of contributors given the observed profile ( $\Pr(N|O)$ ).

As with single MCMC deconvolutions of samples, comparisons may be performed after VarNOC analysis. There are two options for performing an LR against a VarNOC analysis— stratified and Maximum Likelihood Estimate (MLE). The **stratified** likelihood ratio uses both deconvolutions, treating N as a variable and integrates N out by stratifying the LR across the variable contributor numbers range assigned in the deconvolution. This stratification of the LR essentially calculates the LR across both number of contributor analyses by incorporating in the adjusted contributor number probabilities into the LR equation. The **MLE** method of performing an LR assigns the number of contributors as the value that produces the maximum posterior probability of the profile. Ultimately, this means that the final LR may have the same number of contributors under both  $H_p$  and  $H_d$ , or could potentially have different number of contributors between the numerator and denominator of the LR. The differences between the LR calculations do not substantially affect the LR (compared to not using VarNOC, or to each other), but there are certain scenarios in which would benefit from one option over another so both options are explored in this performance check.



With the many new elements incorporated into STRmix since the release of v2.4.06 there are many steps to take and samples to analyze in order to ensure a robust performance check of this software upgrade. The all-encompassing purpose of this performance check is to evaluate the deconvolution capabilities and LR calculations of STRmix v2.6 for all kits (GF, ID/ID+, and MF), as well as assess the additional tools described above to make informed decisions about how to incorporate STRmix v2.6 into the casework process at SDPD.

## Methods

The samples used for this study were primarily taken from the respective amplification kit validations and addendums; see previous validation or modification studies for information about how each of the mixtures and DNA results were generated. A few additional samples were incorporated into the validation to test specific features, which are detailed below.

- For MF samples, the same input files and kit settings were used as the original validation.
- For ID and ID+ samples, the same input files were used, but the new STRmix v2.6 kit settings were used to deconvolute the input files.
- GlobalFiler evidence type samples were re-analyzed in GMID-X with Analysis Methods and Panel that no longer filtered any stutter; see Model Maker write-up for list of analysis methods and panels. Despite re-analysis the input file may not have changed from the original analysis; some input files remained the same as the original validation, and others had additional peaks detected that were no longer filtered as stutter. Some input files were different because of the GMID-X peak detection and window modification study that happened after the original validation of the input files. The new GF evidence input files were then exported for STRmix analysis.
- Reference samples were not re-analyzed or exported for this validation/performance check, as the reference genotypes did not change. Although size is no longer a required component of reference input files for STRmix v2.6, it can still be a part of that input file without any errors.
- All STRmix kit settings (variance parameters, stutter files, etc.) are detailed in the STRmix v2.6 Model Maker write up.

This write-up contains results from studies conducted on 4 different SDPD-specific STRmix kits: SDPD MiniFiler, SDPD Identifiler Plus, SDPD GlobalFiler, and SDPD 24s GlobalFiler. Studies were generally performed as follows:

- Deconvolutions were performed on all or a subset of mixtures from the original validation.
- Results from the deconvolution were compiled using the Interpretation Results to Excel tool (v1.2).
- Contributor proportions, DNA amounts, diagnostics such as allele variance, stutter variances, log(likelihood) and genotype weights were qualitatively assessed.

- Comparisons were made to each deconvolution using one or a combination of methods such as database searching (with and without  $F_{ST}$ ), LR from previous (sub-sub source and sub-source LRs), and LR batcher.
- Database search results were compiled using the Database Search Results to Excel tool (v1.2), and were sorted and plotted, where relevant.
- LR from previous results were compiled using Interpretation Results to Excel tool (v1.2) for assessment and plots, where relevant.
- NIST1036 allele frequencies (updated in July 2017) were used for LR calculations using the Caucasian, African American, Asian, Hispanic, or combined populations, depending on the type of comparison being done.
- In STRmix v2.6, the allele frequency files for each population require all loci, even markers not modeled by STRmix (i.e. Yindel and DYS391), to have total locus frequencies sum to 1; therefore, the previous files were updated to avoid an error message when performing database searches.

Over the course of this validation and performance check, three different evaluation versions of STRmix v2.6 were tested: v2.6.0.29 beta, v2.6.0.37 beta, and v2.6.2. Between v2.6.0 and v2.6.2, there was a change to the way VarNOC LRs are calculated when  $H_p$  and  $H_d$  contain different numbers of contributors. Although this was a minor change (more details can be found in that release and testing report), all VarNOC extended output calculations were recalculated using v2.6.2, and the majority of the other analyses were recalculated after this version was installed. All conclusions about VarNOC were made after reviewing results from the latest version of the software. Additionally, all evaluation of STRmix v2.6 was done on a 64 bit SiForce workstation computer (128 GB RAM with a dual Intel®Xeon® 2.10 GHz processor) using a zipped version of the software.

#### MiniFiler kit MCMCs and LRs

STRmix input files generated for the original STRmix validation were used for this performance check; samples were not re-analyzed with GMID-X because only n-1 repeat stutter is modeled with the STRmix MF kit; the other types of stutter are generally not observed. Using STRmix v2.6, two sets of mixtures were analyzed.

In the first set, a subset of 10 two- to four-person mixtures was chosen for direct comparison to results from v2.4.06. The seed was set to be identical to the seed used in the STRmix v2.4.06 performance check. Using that set of deconvolutions, a DB search was performed using a file that contained the true contributors;  $F_{ST}$  was set to 0, and the Caucasian allele frequencies were used so that a direct comparison could be made to true contributor LRs generated with STRmix v2.4.06.

In the second set, a random seed was used for analysis of 21 two- to five-person mixtures. Against this set of MCMCs, a DB search was run against a 76-profile file containing the true contributors as well non-contributors. The minimum LR was set to zero, the NIST combined allele frequencies were used, and the  $F_{ST}$  used for the search was 0, and sub-source LR was not

assigned.  $H_d$ TT (with importance sampling) results were also compiled (graphs and average LR).

Six of the mixtures that were analyzed in the second set were the same as the mixtures in the first set. This provided an opportunity to examine MCMC variability between two different deconvolutions with a different seed, analyzed with STRmix v2.6. Contributor DNA amounts, allele and stutter variance values and average log(likelihood) values were compared from one run to the next using v2.6.

#### Database search with $F_{ST}$

$F_{ST}$  beta distributions can now be incorporated into the database search LR calculations. Originally, database searching only calculated LR using the product rule, which doesn't incorporate NRCII recommendations to account for substructure. At SDPD, DB searching is used to provide information about potential associations to an elimination database, and then calculate an "LR from previous" for any potential associations returned above the LR cutoff of 1000. Incorporating  $F_{ST}$  into the database search LR calculations could potentially eliminate the necessity to calculate the "LR from Previous" thus streamlining the process. LR calculations performed this way use only a single population's allele frequencies.

MCMCs from three different kits (MF, ID+, and GF) were utilized for this part of the study. An LR calculation from a previous interpretation was done using "LR from Previous" or "LR batch" using each of the known contributors in the mixtures. This included low-level contributors whose contribution to the mixture were around the level of the analytical threshold, meaning that not all LR are above 1. Then a DB search was performed against a 76-profile file containing the true contributors and non-contributors. Minimum LR was set to zero, the Caucasian allele frequencies were used, and the  $F_{ST}$  used for the search was 0.01b(1.0,1.0) to indicate  $\theta = 0.01$ , with a flat prior distribution. Sub-sub-source LR (i.e., point LR) were calculated using database search and also using LR from previous calculations. Sub-sub source LR (point LR) from the known contributors using Caucasian allele frequencies were plotted against the corresponding DB search LR.

In STRmix v2.6, the DB search file no longer needs to be kit specific; the same DB search file can be used for MiniFiler, Identifiler Plus and GlobalFiler deconvolutions. The output is also different. DB searching now results in a slightly different collection of files, one of which is a PDF report for each DB search, reporting all LR values above the user defined threshold. In the v2.6 report settings, there is an option to sort the results in order of LR, so that the highest LR are now at the top of the list, should there be multiple associations to the MCMC. The DB search report will also include results for Amelogenin.

#### LR calculations from v2.3 and v2.4 deconvolutions

In casework, it may be necessary to calculate an LR from a deconvolution done with a previous version of STRmix. At SDPD, both v2.3 and v2.4 have both been used to analyze evidence. A small study was performed to test the functionality of calculating an LR from previous using deconvolutions from v2.3 and v2.4. A 2 person MCMC and 3 person MCMC each from both v2.3

and v2.4 were used to do multiple LR from Previous calculations, including one combined LR calculation.

#### Identifiler Plus MCMCs and LRs

A subset of Identifiler and Identifiler Plus mixtures from were analyzed with STRmix v2.6. They were not re-analyzed in GMID-X because forward stutter in these evidence samples had been unfiltered prior to the v2.4 validation. The 14 two-person, 15 three-person, and 14 four-person Identifiler and 14 two-person, 16 three-person, and 18 four-person Identifiler Plus evidence samples were interpreted with v2.6. Mixtures were first analyzed using the known number of contributors they were prepared with, but apparent NOC was also noted. Where these differed, the mixtures were analyzed using VarNOC, and those results will be discussed further in the VarNOC section below.

To compare versions, a DB search was run to obtain LRs of the true contributors. In order to compare it to v2.4, NIST combined allele frequencies were used, and the  $F_{ST}$  used for the search was 0, and sub-source LR was not assigned.

In order to evaluate sensitivity and specificity for v2.6 analyses, a DB search was run against a ~110-profile file containing the true contributors and non-contributors. The minimum LR was set to zero, the NIST combined allele frequencies were used to calculate a sub-sub source LR using an  $F_{ST}$  set to 0.01b(1.0,1.0). HdTT (with importance sampling) results were also compiled for a subset of these results (graphs and average LRs). The same types of graphs as the original validation were created with the results from v2.6 deconvolutions, showing true contributor and non-contributor LR values plotted for 2-, 3-, and 4-person Identifiler and Identifiler Plus mixtures

#### GlobalFiler and GlobalFiler 24s MCMCs and LRs

Twenty six GlobalFiler mixtures were used for this performance check/validation. This set was made up of all the original validation mixtures amplified with the reformulated master mix (prior to 2015) including five 2-person, six 3-person, three 4-person, and twelve 5-person mixtures. A subset of these mixtures that had been previously injected for 24 seconds (four 2p, four 3p, one 4p, and one single source sample) were used for testing the GlobalFiler 24 second STRmix kit. One aim of this validation was to test the addition of generalized stutter ( $n+ \frac{1}{2}$  and  $n-2$ ) to the biological model, so all of these samples were re-analyzed in GMID-X to label peaks that were previously filtered. In addition to the above samples, a set of sixteen 2-person low-level mixtures were also evaluated with STRmix v2.6. These samples were not necessarily samples that had additional peaks called after GMID-X analysis, but some had ambiguity with NOC, so were analyzed twice with two different NOCs, which made them great candidates for the validation of the VarNOC feature. All GlobalFiler mixtures were first analyzed with the number of contributors it was designed to have (i.e., ground truth NOC), but apparent NOC was also noted, and those samples will be discussed further in the VarNOC section below.

For the comparison of analyses between STRmix software versions, five GlobalFiler and five GlobalFiler 24 second samples were selected because they were used in the evaluation of v2.4.

After analysis in v2.6, a DB search was run for each of these to obtain LR of the true contributors. In order to compare results to v2.4, NIST combined allele frequencies were used, and the  $F_{ST}$  used for the search was 0, and sub-sub-source LR was calculated.

Similar to the MiniFiler and Identifiler Plus study, the sensitivity and specificity were evaluated for v2.6 analyses. A 76-person DB search was run against the true contributors and non-contributors in this database file. The minimum LR was set to zero, the NIST combined allele frequencies were used, and the  $F_{ST}$  used for the search was 0.01b(1.0,1.0), and sub-source LR was not assigned.  $H_d$ TT (with importance sampling) results were also compiled for a subset of these results (graphs and average LR). The same types of graphs as the original validation were created with the results from v2.6 deconvolutions, showing true contributors and non-contributors plotted for 2-, 3-, and 4-person mixtures.

One sample that had to be modified prior to STRmix v2.4 analysis because of a problem modeling forward stutter was also evaluated in v2.6 to test a change to peak modeling. Using STRmix v2.4, there was a sample with a homozygous genotype that had n+1 stutter detected at TH01. The n+1 stutter file specifies 0% stutter at TH01, thus when the profile containing the n+1 peak at TH01 was analyzed in v2.4, the result was a heterozygous genotype that included the n+1 peak weighted with 100% genotype weight. Qualitative expectations would dictate that this sample should have been a homozygote. To address this, the stutter peak was edited out for the STRmix analysis. In STRmix v2.6, a minimum expected stutter ratio of 0.001 was incorporated, so this sample was re-analyzed with STRmix v2.6 to observe the effect of the change.

Analysis time is included in the PDF reports for the MCMCs. Total analysis time was compiled for GlobalFiler results. For these analyses, three different versions of v2.6 software (v2.6.0.29 beta, v2.6.0.37 beta, and v2.6.2) were used, sometimes low memory mode was utilized and sometimes the program was limited to the default of only 4 GB of RAM, so these results can only be used as general estimates of analysis time of v2.6 deconvolutions.

#### LR from Previous and combined LR

Likelihood ratios were evaluated for propositions containing single individual as well as compound propositions containing multiple individuals to observe the assignment of contributor order as well as the effect of combining individuals in propositions on the magnitude of the LR.

LR calculations with v2.6 were also compared to LR calculations with v2.4. Three mixtures from the v2.4.06 performance check were used for LR comparisons. An LR was calculated for each of the contributors to these mixtures using v2.4.06. LRs were then calculated to the same MCMCs using v2.6.2 (setting the same seed for the LR calculation that was used in v2.4.06). The “LR total” value from the v2.4 calculation was compared to the “sub-sub-source LR” from the v2.6 calculation for each of the contributors.

#### Variable NOC

Past validations of STRmix have demonstrated the effect changing the NOC has on both known contributor and non-contributor LR. For the first part of the VarNOC study, samples were analyzed with different numbers of contributors (not using VarNOC) to replicate this effect on known contributor LR with these validation samples. They were analyzed twice, once with ground truth N, and another time with either N+1 or N-1 contributors. This was the same approach taken in previous STRmix validation work. All five 2-person, all five 3-person, and all three 4-person GF mixtures were analyzed with N+1 contributors (despite having no basis using current interpretation guidelines). The only two samples that COULD be interpreted as having N-1 contributors were two of the 4 person mixtures. All other samples had too many peaks detected, and STRmix would not allow the interpretation as fewer contributors. A 76-person DB search was run against the true contributors and non-contributors in this database file. The minimum LR was set to zero, the NIST combined allele frequencies were used, and the  $F_{ST}$  used for the search was 0.01b(1.0,1.0), and sub-source LR was not assigned. These LR were compared to LR from the same known contributors compared to MCMCs with N (ground truth) contributors.

The second part of the VarNOC study utilized the subset of samples that had uncertainty in the NOC, using current interpretation guidelines. Most of the samples used to test the VarNOC feature were amplified with GlobalFiler. Not every sample contained characteristics that would warrant an analysis using the VarNOC feature of STRmix v2.6, so each mixture was scrutinized for any ambiguity in determining the NOC. If any ambiguity was present, the mixture was included in this portion of the validation. In order to have a larger subset of samples to work with, the Identifiler/Identifiler Plus mixtures were also assessed with for ambiguity in the determination of NOC. A list of each sample used including a description of the justification for using VarNOC on each one is included below.

1. *Single source sample 152* has two drop-in peaks at one locus AND both true contributor peaks dropped out at the same locus. This was a validation sample with low target input (~12.5pg) injected for 24 seconds that was used for testing VarNOC MLE LR, because it was an apparent single source sample. Not until comparison to the known DNA profile from the donor did the possibility arise of it being a 2 person mixture (VarNOC 1/2).
2. *This same single source sample* input file was edited to have one of the aforementioned drop-in peaks changed to a known contributor allele to create a partial drop-put, partial drop-in scenario (VarNOC 1/2).
3. *Low-level Mixture 10* is a ground truth 2-person mixture with elevated stutter at one locus, a sub-threshold peak at another locus and several instances of extreme peak imbalance (VarNOC 2/3).
4. *Low-level Mixture 11* is a ground truth 2-person mixture with sub-threshold peaks at one locus (VarNOC 2/3).
5. *Low-level Mixture 15* is a ground truth 2-person mixture with slight peak imbalances (VarNOC 2/3).
6. *Low-level Mixture 16* is a ground truth 2-person mixture with peak imbalance at one locus (VarNOC 2/3).

7. *GF Mix 2-34 24-sec* is a ground truth 2-person mixture injected for 24 seconds in which there are two instances of peak balance (VarNOC 2/3).
8. *GF Mix 3-44* is a ground truth 3 person mixture with two instances of elevated (~20%) stutter at one locus (VarNOC 3/4).
9. *GF Mix 3-44 24-sec* is the same mixture injected for 24 seconds. One additional peak is called resulting in 3 instances of elevated stutter VarNOC 3/4).
10. *GF Mix 5-1* is an apparent 4-person mixture, but was designed as a 5-person mixture. This is a more robust sample with minimal dropout, but high allele overlap (VarNOC 4/5).
11. *GF Mix 5-5* is an apparent 4-person mixture, but sub-threshold peaks could justify a 5-person interpretation (VarNOC 4/5).
12. *GF Mix 5-10* is an apparent 4-person mixture, but sub-threshold peaks could justify a 5-person interpretation. The difference between this one and 5-5 is that one of the contributors to this mixture is dropping out almost completely (VarNOC 4/5).
13. *GF Mix 5-12* is designed as a very low template 5-person mixture. The total DNA input was so low that multiple contributors are dropping out to the point that this is an apparent 3-person mixture, but sub-threshold peaks could justify a 4-person interpretation (VarNOC 3/4).
14. *ID Mix 2\_9* is a 3-person mixture amplified with Identifiler. There is elevated stutter at one locus (VarNOC 3/4).
15. *ID Mix 2\_24* is a 3-person mixture amplified with Identifiler. There is slightly elevated stutter at one locus (VarNOC 3/4).
16. *ID Mix 2\_30* was designed as a 4-person mixture and amplified with Identifiler. Because of fewer discriminating loci, it is an apparent 3-person mixture with a sub-threshold peak (VarNOC 3/4).
17. *ID Mix 2\_35* was designed as a 4-person mixture and amplified with Identifiler. Because of fewer discriminating loci, it is an apparent 3-person mixture with sub-threshold peaks, but only in stutter positions (VarNOC 3/4).
18. *ID Mix 13 Study\_Case 5* was a sample designed as a 4-person mixture to have purposely high allele sharing. By allele count this sample could be assessed as a 3-person mixture, but taking into account some peak imbalance, it could be interpreted as a 4-person mixture (VarNOC 3/4).

Two other 5-person mixtures could have been assessed as having either 4 or 5 contributors, but the computer ran out of memory when attempting the 5-person MCMC.

After analyzing these samples using the VarNOC feature, information (probability of N given O, diagnostics, run-time, contributor %, etc.) was compiled for comparison purposes. Comparisons were then run for each of the known contributors, twice each. One comparison specified using a stratified VarNOC LR, and the other comparison specified using the Maximum Likelihood Estimate for the VarNOC LR. When using MLE, the propositions for each sample were recorded to determine which samples had a different NOC in the numerator and denominator.

In each LR from previous analysis, three sets of LRs are calculated: the first set consists of LRs compared to MCMC 1 with N contributors; the second set consists of LRs compared to MCMC 2 with N+1 contributors, and the third set of LRs are the VarNOC LRs (either stratified across contributors, or MLE). The first two sets of LRs are identical to the usual output when doing a comparison to a single deconvolution. If the settings do not indicate an HPD calculation, then there will only be three sub-sub-source LRs reported (for each population); one for each individual MCMC (N and N+1) and one for the VarNOC LR. When the HPD LR is calculated, then both individual MCMC LRs (N and N+1) will have the sub-sub-source LRs and the HPD LRs. The VarNOC LR will only have the HPD LR reported. As always there are multiple LR options, including sub-source LR (formerly factor of N! LR) which may be calculated if desired. No sub-source LRs was not compiled at this time.

LRs with all four sets of allele frequencies were used, and HPDs were calculated. In order to best compare LRs, the LOWEST HPD was selected from each of the four populations and used in the graphs for VarNOC LR analysis. Several graphs were prepared with this information. VarNOC LRs were compared to LRs of ground truth, N+1 and N-1 number of contributors. Stratified vs MLE LRs were also compared.

DB searches with known contributors, and also 10,000 random samples (generated using the NIST combined allele frequencies) were run after these mixture deconvolutions. The  $H_d$ TT tool is not compatible with VarNOC analyses. The minimum LR for the DB searches was set to zero, the NIST combined allele frequencies were used, and the  $F_{ST}$  used for the search was 0.01b(1.0,1.0) for some mixtures, and 0 for others.

In addition, the extended output of a VarNOC analysis was examined to determine what the process was and to determine whether the VarNOC maths could be reproduced. Both a SDPD validation sample as well as an ESR provided example of VarNOC were used to examine the VarNOC process and to recreate the maths.

#### Exploring the MCMC quality: $H_d$ -True Tester and database search of random profiles

DB searches with 10,000 random profiles,  $H_d$ TT (with random sampling), and  $H_d$ TT (with importance sampling) were run on the deconvolutions described above. As mentioned previously, the  $H_d$ TT tool cannot be used with VarNOC analyses, so only DB searching was available. To set up an  $H_d$ TT, the previous interpretation is selected. The number of random profiles can be specified, the population for sampling and LR is selected, and importance sampling is selected through a check box (if left unchecked random sampling will be performed). Unless otherwise specified, 10,000 random profiles were used for these tests. Since  $H_d$ TT using random sampling is similar to using a random profile database search, the only difference being that it generates a different set of 10,000 random profiles for each analysis, the need for in-depth of analysis was not as great as for  $H_d$ TT (with importance sampling). Therefore, in the interest of time,  $H_d$ TT (with random sampling) was only run on a small subset of mixtures to test functionality of the tool and get a general time estimate for completion of  $H_d$ TT.



In order to evaluate  $H_d$ TT (with importance sampling), this tool was used on 40 GlobalFiler analyses. The NIST combined allele frequencies were used in order to better compare results (in a qualitative way) to the 10,000 random profile DB searches that were done (both time and output). Charts, tables, results and analysis time were compiled for all those analyses. In two mixtures,  $H_d$ TT (with importance sampling) was run again to investigate the effect of increasing the number of random profiles to 100,000 (on both run time and results).

### Compatibility with COSTaR

The indexing in STRmix v2.6 is done using locus names instead of locus numbers (v2.4.06). STRmix data output still contains a results file in v2.6. COSTaR (COSTaR 4.2-G SDPD 020119.xls) was tested for compatibility with STRmix v2.6 data by importing results from a 2 person mixture (2-34) and a 3 person mixture (3-16). Every genotype and weight for each contributor from the STRmix results (PDF) was compared to the genotypes and weights for each contributor in COSTaR after the respective mixture was imported. COSTaR for Suspects (COSTaR for Suspects-G 071518.xls) was also checked to ensure indexing was working properly.

## **Results**

The changes to GMID-X required by switching STRmix v2.6 will only affect GlobalFiler profiles. In GlobalFiler analyses, n-2 repeat and n-0.5 repeat stutter peaks are no longer being filtered by the panel. Detection of n-2 stutter peaks is largely dependent on the peak height of the parent allele, and have the possibility of being detected at any locus. In the validation of GlobalFiler, n-0.5 repeat stutter peaks were predominantly observed at the SE33 and D1S1656 loci; however, there were a few observations of n-0.5 repeat stutter at D2S441 as well. SE33 and D1S1656 were the only loci that n-0.5 stutter was filtered for STRmix v2.4 analyses, and the only loci at which STRmix v2.6 will model that type of stutter. Note that n-0.5 and n-2 are not frequently observed occurrences. While n-2 repeat stutter was analyzed by linear regression for the n-2 stutter file, the guideline of n-2 stutter peaks being up to 1.5x the n-1 stutter ratio may still be used for the purpose of assessing the number of contributors. Analysts should keep in mind that these types of stutter are generally only observed when peak heights within the profile are robust, and the more instances that occur in a profile, the more likely it is that there is an additional contributor.

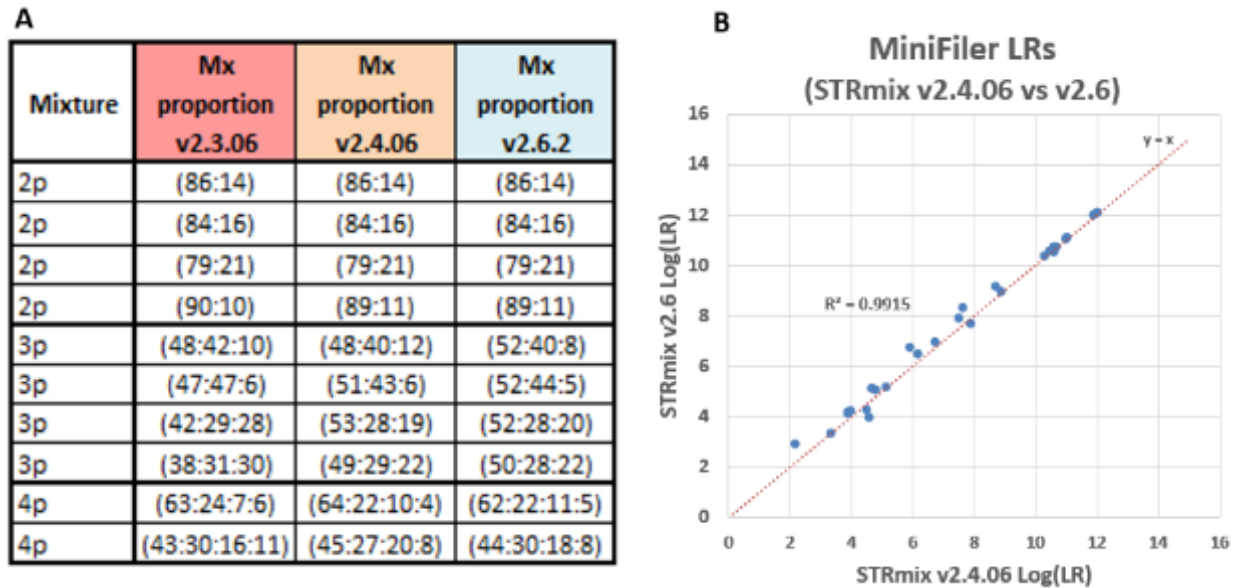
In higher order mixtures, this also means that the number of peaks detected at a locus could more easily exceed 16 (the limit for exporting for v2.4 and before analysis). STRmix v2.6 no longer requires the user to specify the number of alleles per locus, so the GMID-X table setting used for exporting electropherogram data to STRmix will be defaulted to contain 22 alleles per locus, but can be adjusted as needed without affecting STRmix analysis.

### MiniFiler kit results

None of the STRmix kit settings changed in the SDPD MiniFiler kit in going from v2.4 to v2.6. This provided an opportunity for testing the effects of biological model changes (i.e., Taylor Quantum Effect) the MCMC deconvolutions in the absence of any kit setting changes. The exact same input files, seed values, stutter files and variances, etc., as used in previous STRmix

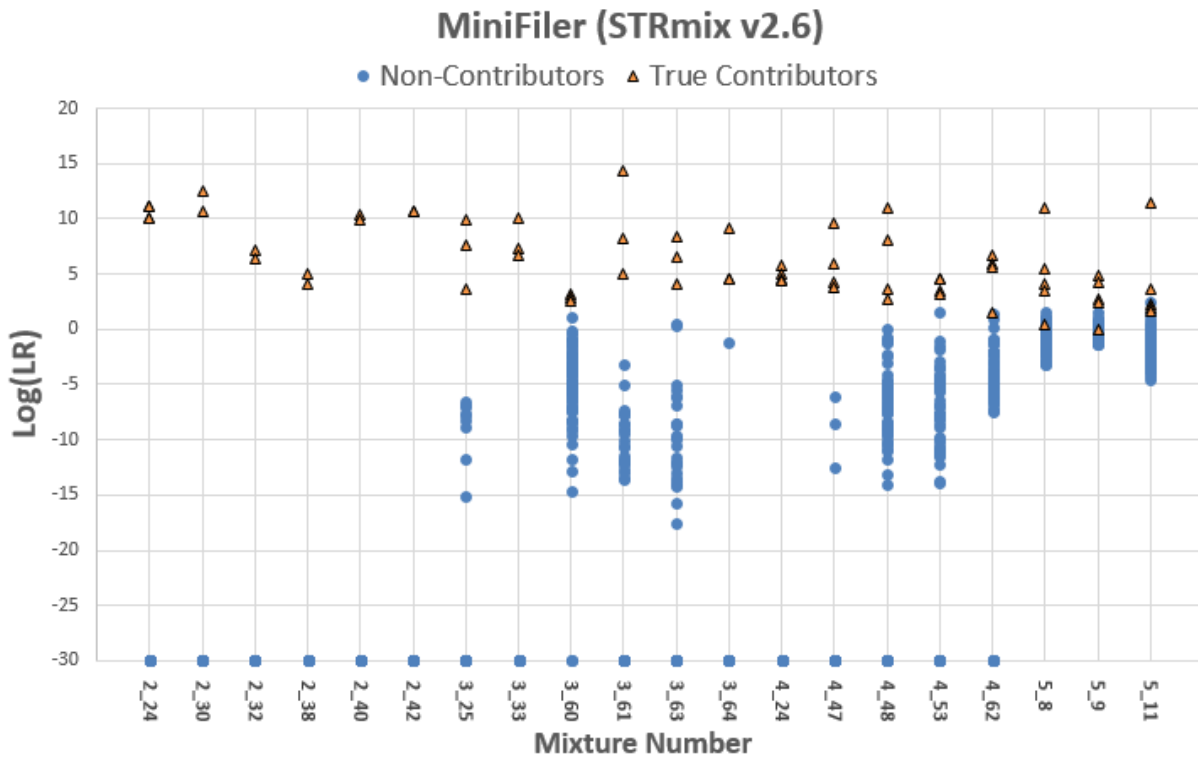
analyses of the MiniFiler samples were used for v2.6 deconvolutions. In order to compare the results from the different versions, known contributor LR<sub>s</sub> were calculated using the database search function ( $F_{ST} = 0$ ; NIST Caucasian allele frequencies). For the purposes of comparing to LR<sub>s</sub> from past versions of STRmix, the allele frequencies date back prior to July of 2017, which could account for small differences in some instances.

Although the same seed was set and the same settings were used, there were very slight differences in the mixture deconvolution. It is noted in the STRmix v2.6.0 release and testing report that due to changes in the how the seed is implemented, differences between weights for profiles with ambiguous genotypes are to be expected. Figure 4A shows the mixture proportions for the same samples analyzed with three different versions of STRmix. Figure 4B demonstrates the small magnitude of the variability in LR<sub>s</sub> for every contributor to these mixtures;  $R^2 = 0.9915$ . Although there are slight differences in the mixture proportion from version to version, the results are very similar.



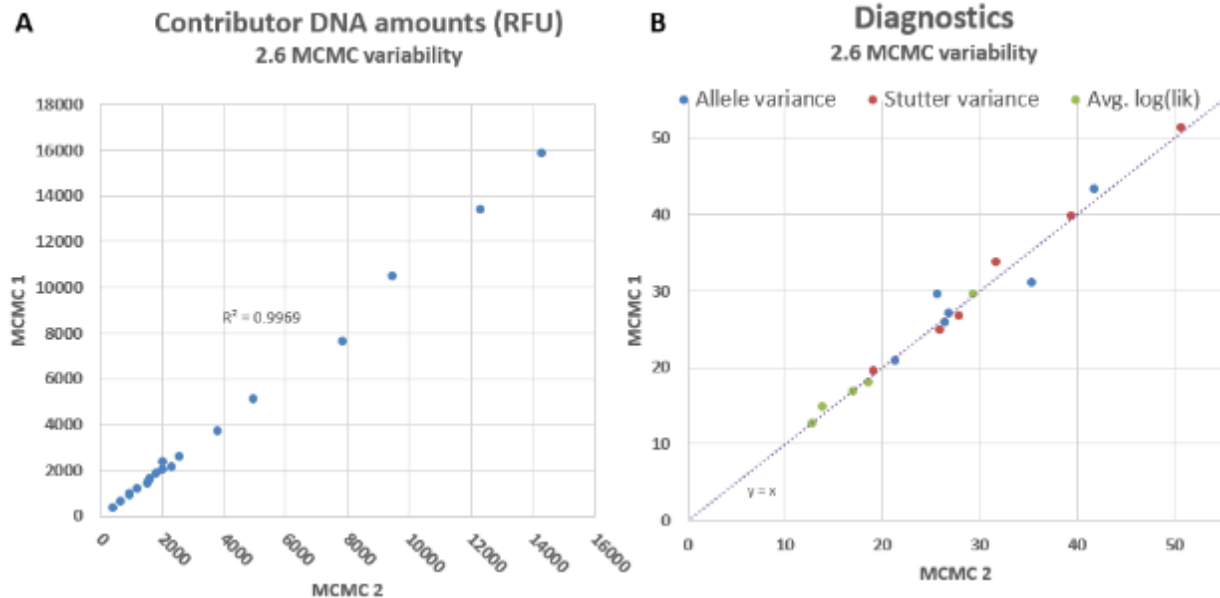
*Figure 4 - A subset of 2, 3 and 4 person MiniFiler mixtures was analyzed with three versions of STRmix. Mixture proportions change slightly from version to version (A). Likelihood ratios are highly correlated between the two versions; v2.4 Log(LR) is plotted against the v2.6 Log(LR) (using Caucasian allele frequencies) for each contributor to these mixtures (B).*

A larger subset of mixtures were analyzed for the purpose of investigating sensitivity and specificity. Figure 5 plots LR<sub>s</sub> of true contributors and non-contributors for 2, 3, 4, and 5 person mixtures. Although MiniFiler is rarely used in casework, and will only be used for legacy analysis, the wide range of samples was available, so was thoroughly tested. The results essentially show the same general pattern as analysis with previous versions STRmix. As the complexity of the mixture increases, and template decreases, LR<sub>s</sub> for true contributors tend to decrease, and LR<sub>s</sub> from non-contributors tend to increase. In some mixtures, there are non-contributors with limited support for inclusion, and one with moderate support for inclusion (highest LR for a non-contributor = 316).



*Figure 5 - True contributor and non-contributor Log(LR)s from MiniFiler mixtures analyzed with STRmix v.2.6 for 2, 3, 4, and 5 person mixtures (using combined allele frequencies).*

MCMC variability has been evaluated in previous versions of STRmix, and was also evaluated in this performance check. Six mixtures were analyzed twice (with different seeds). The DNA (template) amounts per contributor were compared to each other from run to run (Figure 6A), and showed minimal variability;  $R^2 = 0.9969$ . Diagnostic values for allele and stutter variance, and average log(likelihood) were also compiled for evaluation to determine the variability in these values are from run-to-run (Figure 6B). The average log(likelihood) and stutter variances were minimally variable with not much change from run-to-run. One value was outside expectations (a negative value), but was reproducible. Allele variance was a little more variable, but in general, when the value was high in one deconvolution, it was high in the paired deconvolution.



*Figure 6 – Six of the same MiniFiler mixtures were analyzed twice with v2.6. DNA amounts (per contributor) were reproducible from one deconvolution to the next (A) across a wide range of template amounts. Allele variance, stutter variance and log(likelihood) are all secondary diagnostic values, and were also consistent from run to run (B).*

#### Database search with $F_{ST}$

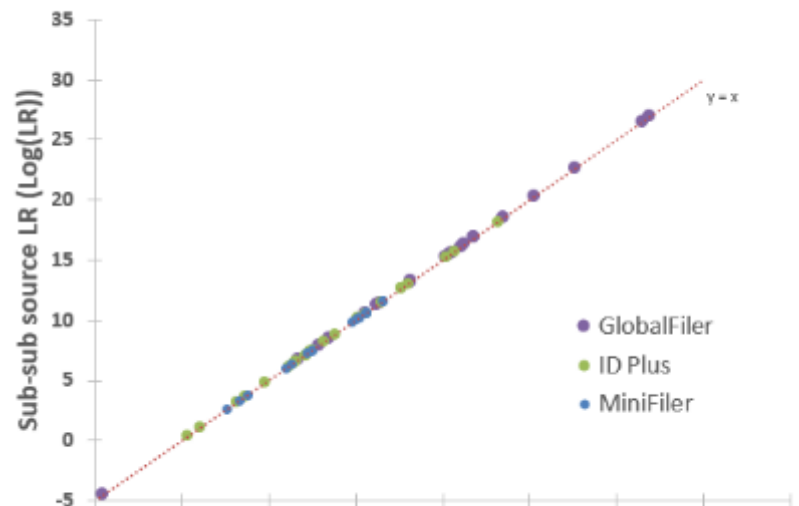
When theta is incorporated into the DB search it results in the same LR value as when calculating sub-sub source LRs if the same allele frequency files (Figure 7) are used for the calculations. This was observed regardless of the contributor, the kit, or the magnitude of the LR. With STRmix v2.6, a DB search against elimination samples will be done on every casework deconvolution.

#### R calculations from v2.3 and v2.4 deconvolutions

STRmix v2.6 was able to perform LR from Previous calculations using both LR batch and LR from Previous tools. When the same seed was set as used in the previous version, an identical sub-sub source LR was obtained. The values were the same regardless of how the calculation was done (LR from previous, or LR batcher).

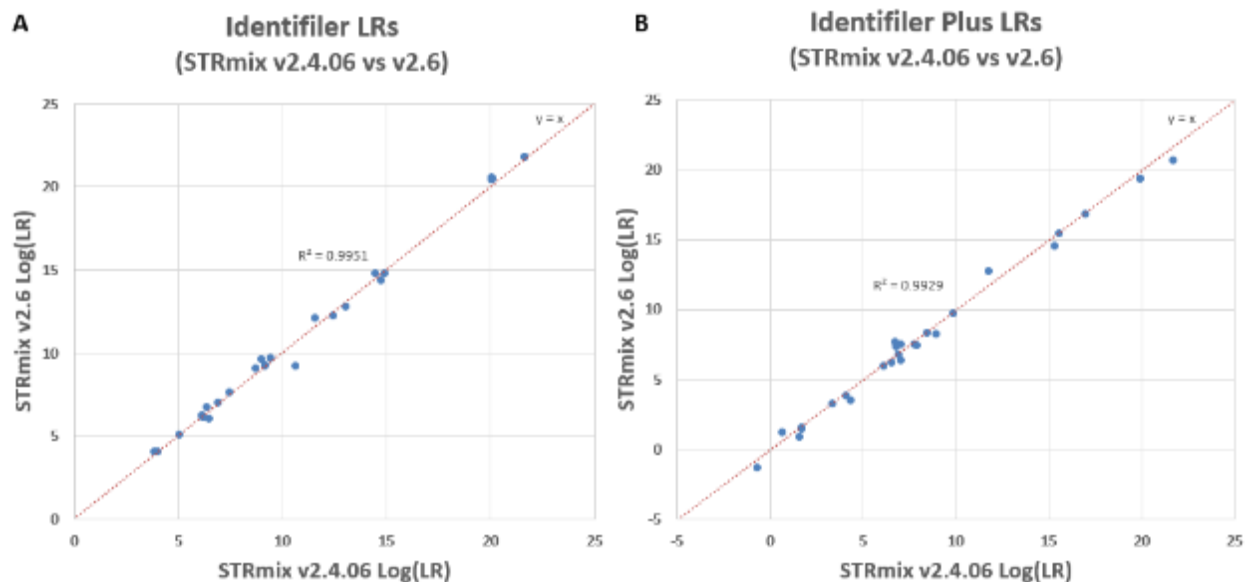
#### Identifiler Plus results

#### **DB search LRs with Theta vs. Point LR**



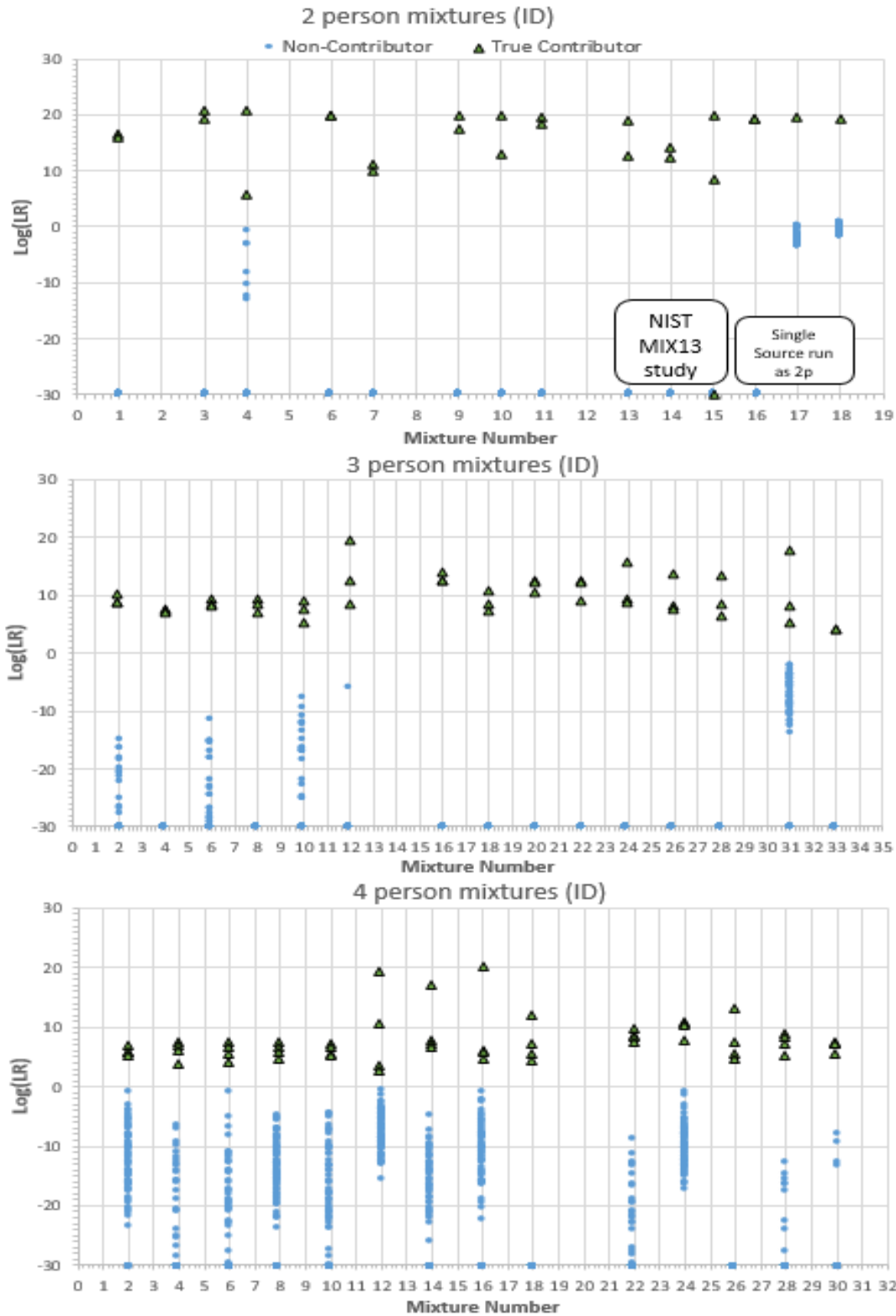
*Figure 7 – STRmix v2.6 incorporates the ability to include an  $F_{ST}$  distribution in DB search calculations. Using the same allele frequencies and  $F_{ST} = 0.01(1.0,1.0)$ , this value is now identical to the sub-sub source LR (previously point LR or LR total) from a direct reference comparison.*

In the original Identifiler/Identifiler Plus legacy kit validation, it was determined that only one set of STRmix settings was sufficient to analyze mixtures that were amplified with either of these two multiplex kits. In order to compare results between STRmix versions, known contributor LR<sub>s</sub> were calculated using the database search function ( $F_{ST} = 0$ ; NIST combined allele frequencies). Since model maker was re-run to incorporate a separate forward stutter variance, and by extension a new back stutter variance, the kit settings have changed from v2.4. Based on the new variance parameters, there was no expectation that the MCMC would be identical between versions even if the same seed was set; however, there was an expectation that the results in v2.6 would be very similar to the deconvolutions performed using the previous version. Despite the kit and deconvolution being slightly different, Figure 8 demonstrates the small magnitude of the variability in the LR<sub>s</sub> for every contributor to these Identifiler mixtures;  $R^2 = 0.9951$ , and Identifiler Plus mixtures;  $R^2 = 0.9929$ .

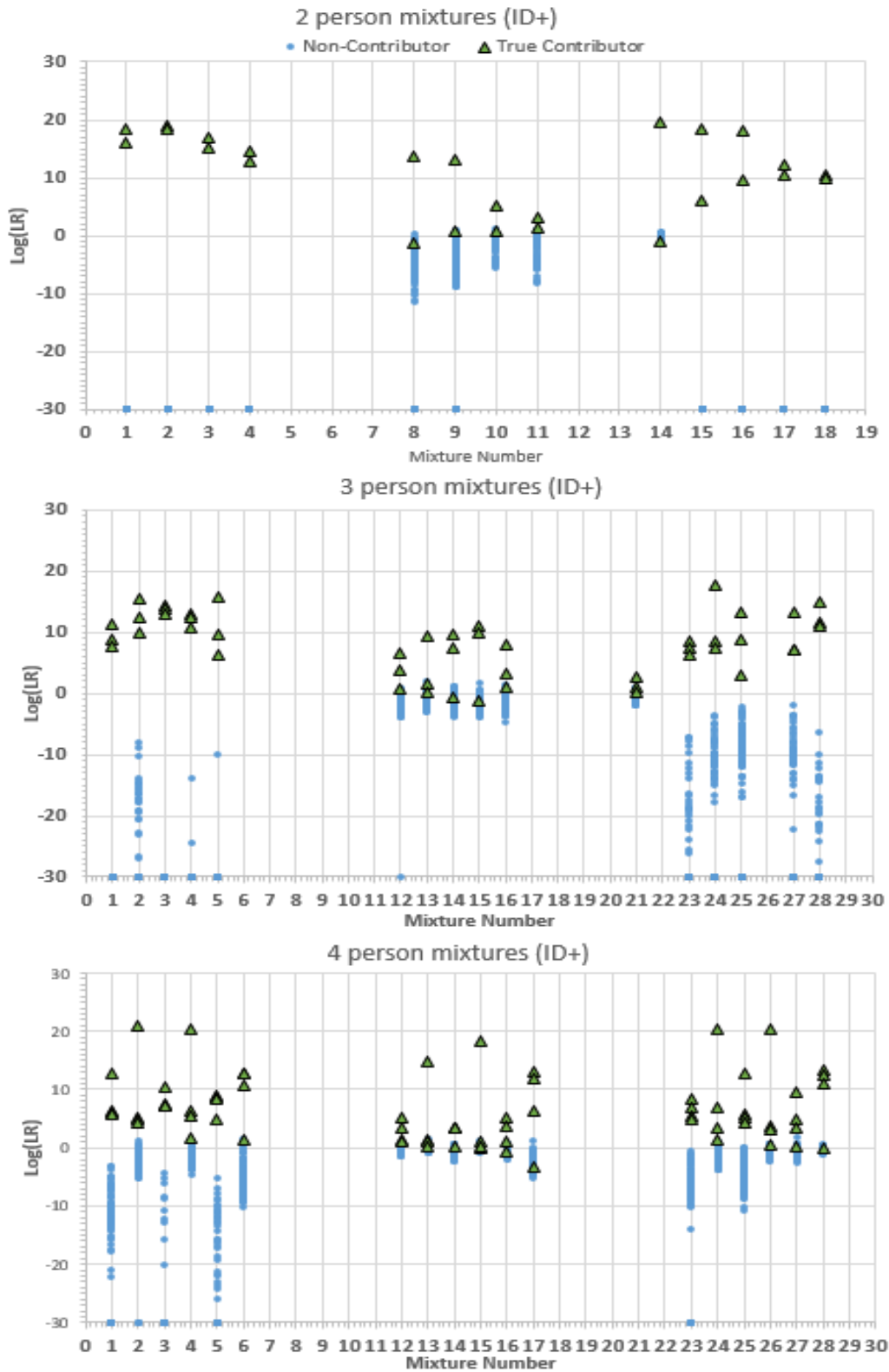


**Figure 8**– Likelihood ratios are highly correlated between the two versions for samples amplified with Identifiler (A) and Identifiler Plus (B); v2.4 Log(LR) is plotted against the v2.6 Log(LR) (using combined allele frequencies and  $\theta = 0$ ) for each contributor to these mixtures. Both sets of mixtures were deconvoluted using the Identifiler Plus STRmix kit.

Similar to previous validations and with MiniFiler results above, true contributors and non-contributors were plotted for 2-, 3-, and 4-person mixtures in both kits (in the same layout with the same mixture numbers as the original validation with STRmix v2.4). Identifiler mixtures analyzed with the Identifiler plus kit are shown in Figure 9, and Identifiler Plus mixtures are shown in Figure 10. High template, high quality mixtures are much more discriminating than low template (with dropout), low quality mixtures.



*Figure 9— True contributor (triangles) and non-contributor (circles) Log<sub>10</sub>(LR)s from Identifiler mixtures analyzed with STRmix v.2.6 for 2, 3, and 4 person mixtures (using combined allele frequencies and  $F_{ST} = 0.01(1.0,1.0)$ ).*

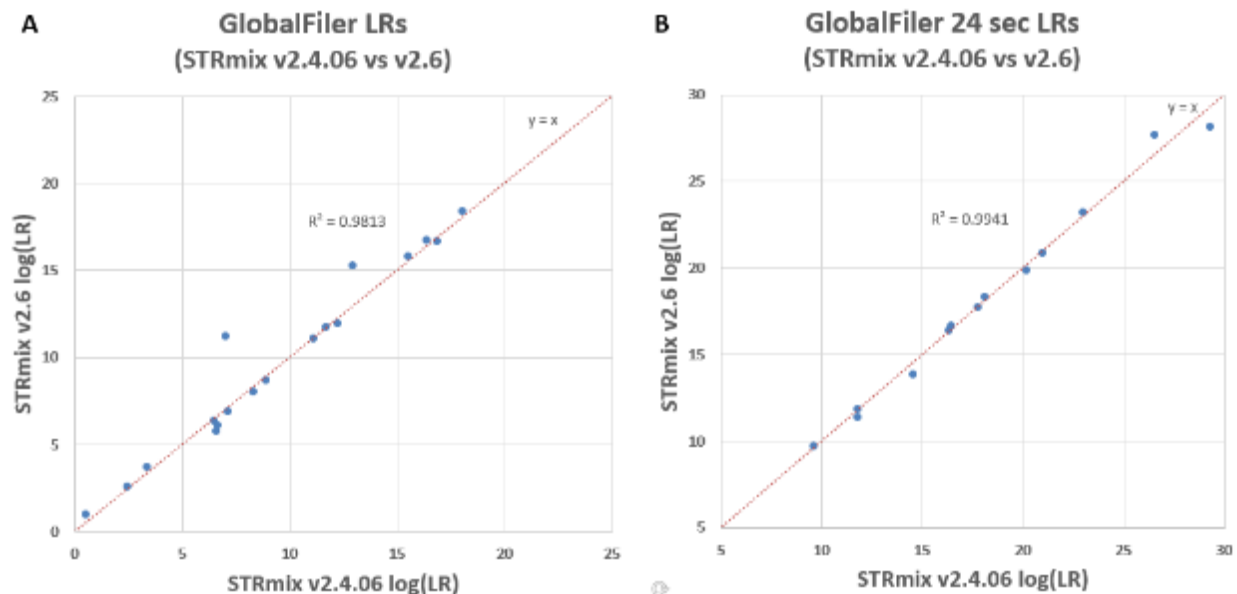


*Figure 10– True contributor (triangles) and non-contributor (circles) Log<sub>10</sub>(LR)s from Identifiler Plus mixtures analyzed with STRmix v.2.6 for 2, 3, and 4 person mixtures (using combined allele frequencies and  $F_{ST} = 0.01(1.0, 1.0)$ ).*

Consistent with the results in the original validation, there is a clear separation between true contributors and non-contributors in more robust samples, and as the template amount drops, the magnitude of the LR of the true contributors drop while the magnitude of the LR of non-contributors increase. The template details for each mixture can be found in the original validation. In some mixtures, there are non-contributors with LR that have limited support for *inclusion* (highest LR for a non-contributor = 86.5). In MiniFiler, the highest non-contributor LR is 316, which is higher than the 86.5 seen with this kit. Identifiler and Identifiler Plus kits amplify more loci, so the discrimination power of those kits is higher than that of the MiniFiler kit.

#### GlobalFiler and GlobalFiler 24s results

For the comparison between STRmix versions, known contributor LR were calculated using the database search function ( $F_{ST} = 0$ ; NIST combined allele frequencies). Not only did kit settings change, but two additional types of stutter modeling were incorporated into the biological model. Despite the many changes to the GlobalFiler and GlobalFiler 24 sec STRmix kits, there was still an expectation (based on the results from the developmental validation) that the results would, again, be very similar, but not identical to results obtained with the previous version. Figure 11 demonstrates the variability in LR for every contributor to five GlobalFiler mixtures;  $R^2 = 0.9813$ , and five GlobalFiler 24 sec mixtures;  $R^2 = 0.9941$ . In the GlobalFiler kit (Figure 11A), the correlation between v2.4 and v2.6 dropped slightly because of one 4 person mixture.

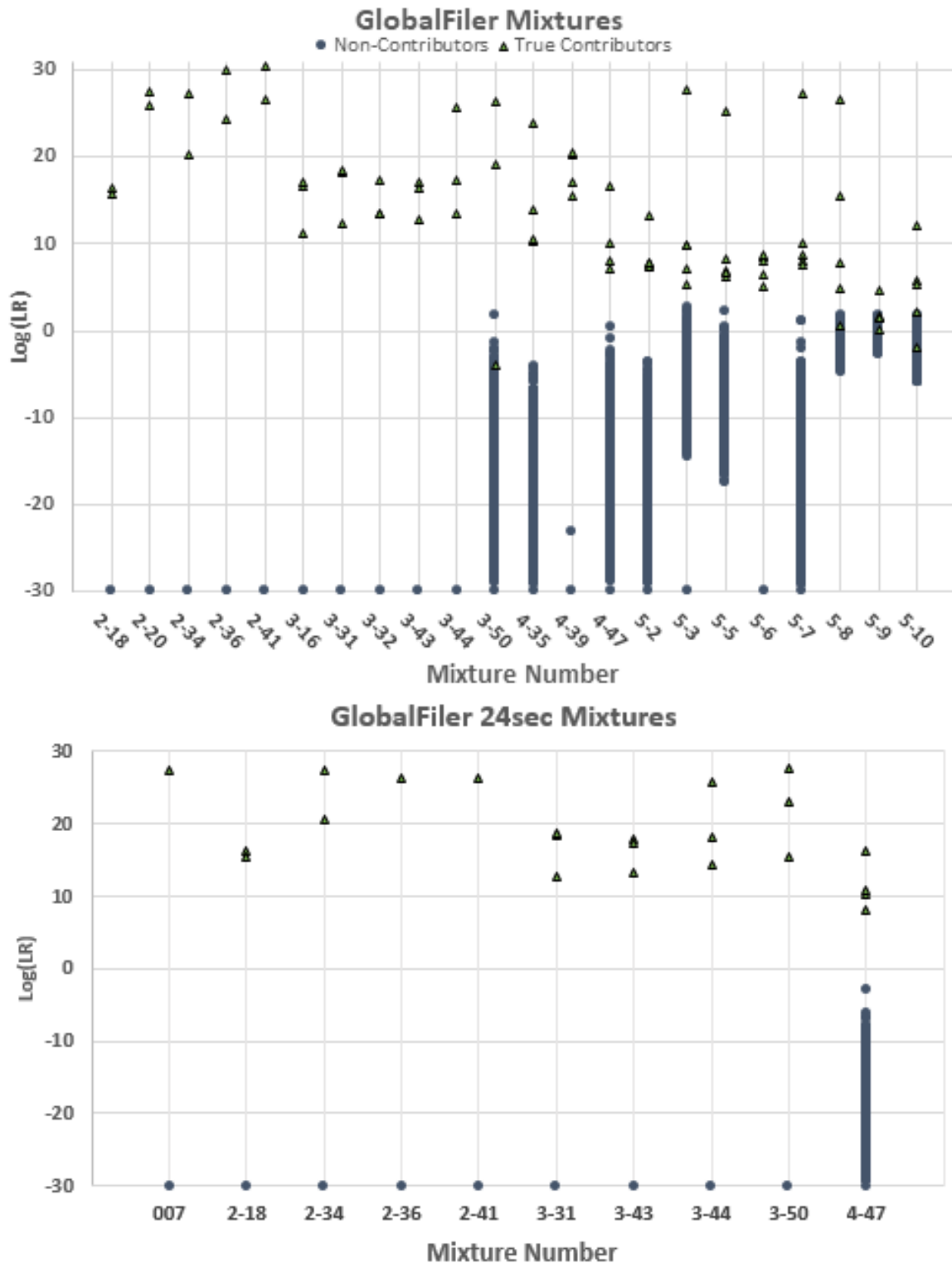


**Figure 11** – Likelihood ratios are highly correlated between the two versions for samples amplified with GlobalFiler and injected for 15 seconds (A) and 24 seconds (B), and analyzed with the respective STRmix kits; v2.4 Log(LR) is plotted against the v2.6 Log(LR) (using combined allele frequencies and  $\theta = 0$ ) for each contributor to these mixtures.



Two contributors to that mixture had higher LR<sub>s</sub> when analyzed with STRmix v2.6 than v2.4. With v2.6, this was described best as a 55:23:13:9 mixture with an avg log(likelihood) of 62. With v2.4, this was described as a 49:23:16:12 with an avg log(likelihood) of 43. The higher avg log(likelihood) in the v2.6 analysis suggests that this mixture was able to be better modeled than in the v2.4 deconvolution. It is likely that since there are more independent stutter variances in v2.6 that the software was able to reduce the differences between expected and observed data. The better modeling likely resulted in genotype set weights that were slightly higher (i.e., more probable) than they were in the v2.4 analysis. Upon closer inspection of those two contributor LR<sub>s</sub>, one fits best with contributor 2, and the other fits best with contributor 4 (in both analyses). The stutter variance is higher in the v2.4 analysis. It is worth noting that the input files are different between the v2.6 and v2.4 analyses. The differences between the files was an increase in the number of detected n-1 stutter peaks because of a GMID-X modification study that adjusted peak detection settings since the original analysis of the samples. The presence of the increased number of stutter peaks alone could have resulted in less required stutter variance in the v2.6 analysis and an associated increase in the average log(likelihood)s obtained throughout the MCMC process.

Similar to previous validations and the Identifiler, Identifiler Plus, and MiniFiler results above, true contributors and non-contributors were plotted for 2-, 3-, 4-, and 5-person mixtures in for GlobalFiler analyses. A subset of these were injected for 24 seconds and analyzed (in the same layout with the same mixture numbers), see Figure 12. High template, high quality mixtures are much more discriminating than low template (with dropout), low quality mixtures, and there is much more ambiguity with the low level contributors in 5-person mixtures. Similar to the results in the original validation, there is a clear separation between true contributors and non-contributors in more robust samples, and as the template amount drops and mixture complexity increases, the magnitude of the LR<sub>s</sub> of the true contributors drop while the magnitude of the LR<sub>s</sub> of non-contributors increase. The template details for each mixture can be found in the original validation. In some mixtures, there are non-contributors with LR values that indicate moderate support for inclusion (highest LR for a non-contributor = 600, which happens to be a 5-person mixture with one major contributor and 4 low-level minor contributors of roughly equivalent proportions). In 2- through 4-person mixtures, the highest LR for a non-contributor = 56.9, which falls in the limited support for inclusion category. This is consistent with the level of limited support seen by non-contributors in the Identifiler Plus analysis, which also analyzed 2- through 4-person mixtures. The more complex the mixture and the more similar contributor levels, the more ambiguity in contributing genotypes and therefore the more likely someone may be associated by chance alone.



*Figure 12 – True contributor (triangles) and non-contributor (circles)  $\text{Log}_{10}(\text{LR})$ s from GlobalFiler mixtures injected for 15 (top) and 24 seconds (bottom) analyzed with the respective STRmix v.2.6 kits for 2, 3, and 4 person mixtures (using combined allele frequencies and  $F_{ST} = 0.01(1.0, 1.0)$ ).*

The single source sample used to test the minimum stutter feature was analyzed with both STRmix v2.4 and STRmix v2.6. At TH01, the peaks (heights in RFU) are 5 (197), 6 (13,085), and 7 (104). When analyzed with STRmix v2.4, the genotype weights are 100% for 6, 7. The average log(likelihood) indicated that there might be a problem with this sample with a value of -3.39. In addition, allele and stutter variances were both over 40. When this sample was analyzed with v2.6, the genotype weight at TH01 was 100% for 6 homozygote (the expected genotype). In addition, the average log(likelihood) increased considerably. The allele variance, forward stutter variance, double back stutter variance and 2bp back stutter variance were all ~15. Interestingly, the back stutter variance was still high in this sample at ~70. All other genotype weights of this sample are as expected based on parent allele heights, however further analysis could be done to isolate the reason for the high variance.

As with v2.4, MCMC analysis time was highly mixture dependent. The results described in this paragraph do not include VarNOC analyses, but do include a few instances of the same mixture being analyzed with two different NOCs outside of the VarNOC feature. For seventeen 2-person MCMCs, the average analysis time was 4 min (ranged from 2-7 min). For seven 3-person MCMCs, the average analysis time was 25 min (ranged from 8 minutes to 70 min). For seven 4-person MCMCs, the average analysis time was 2 hrs, 45 min (ranged from 3 min to 5 ½ hrs). For nine 5-person MCMCs, the average analysis time was 11 hrs, 45 min (ranged from 5 min to 27 ½ hrs). One of the 5-person mixtures reached an “out of memory” error, and results could not be obtained. Another 5-person mixture, which presented as an apparent 4-person mixture had an MCMC completed analyzed as a 4-person mixture, but resulted in an “out of memory” error when analyzed as having 5 contributors. The analysis mixture also could not be completed when it was attempted with 4/5 VarNOC analysis. The range was very large on these analyses, but it is important to keep in mind that these weren’t all run in low memory mode, and some were limited to 4GB of RAM before the settings file allowed for more.

#### LR from Previous and combined LRs

In general, LR from Previous analyses function as expected for MCMCs performed within STRmix v2.6, as well as from analyses conducted using previous versions of the software.

The “sub-sub-source” LRs calculated with v2.6.2 were identical to the “LR Total” values from v2.4.06 calculations to the 3 mixtures analyzed with v2.4.06, when the same allele frequencies were used.

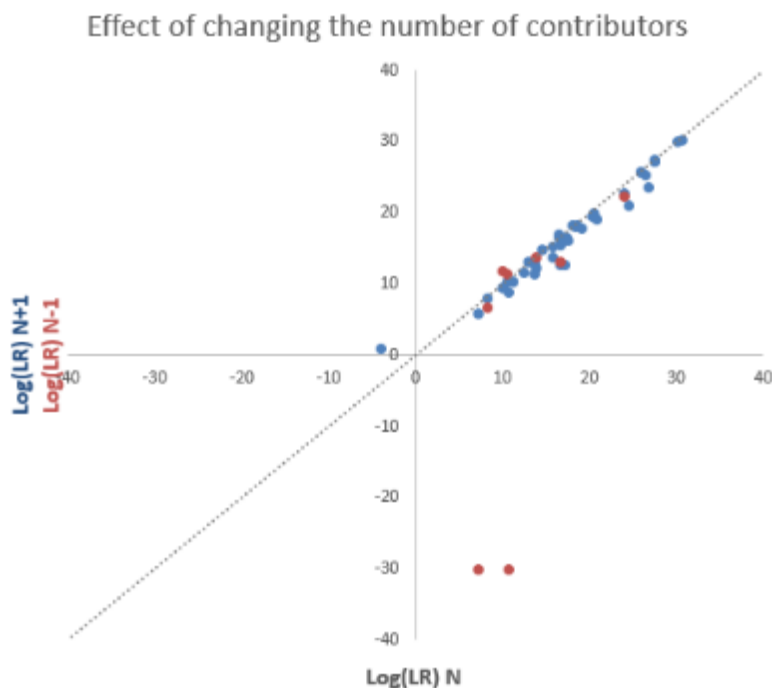
When calculating sub-sub-source LRs, STRmix associates the compared individual(s) with the contributor order producing the highest LR. In most situations, this is consistent between all populations; however, it was observed that on some LR calculations the compared profile was associated with different contributor orders in different populations. Occasionally when contribution proportions between contributors are similar there are small differences in genotype set weights between contributors with similar genotypes among them. When LRs are calculated it can occur that a particular allele (or combination of alleles) present in a different contributor order will cause the highest LR between the population groups to be different

contributor orders. This may happen when a rare allele is present within the genotype sets for particular population group.

In addition, some compound LR's calculated in v2.6 resulted in the contributor order for the compound LR being different than the contributor order reported when the LR's were calculated individually. The change in the reported contributor order for compound LR's can be explained by the different propositions that are involved between the independent LR's. It is somewhat expected that when you have a single individual compared to a mixture that one contributor order may appear to be the best fit, but with compound propositions the contributor orders of best fit may adjust to accommodate the additional individuals added into the equation.

### Variable NOC

It has always been an option to use STRmix to analyze a single sample under different assumptions regarding the NOC. In general, the effect on contributor LR's known. When the magnitude of the LR is large, there is a small effect on the LR when analyzed as N+1, or N-1 (when feasible). When the LR magnitude is low, the effects can be somewhat more dramatic. For instance where it is possible to analyze a sample as N-1 contributors (without the number of peaks detected prohibiting this), the LR for contributors can change from supporting inclusion to exclusionary. Conversely, when evaluating a mixture as N+1, LR's can change from supporting exclusion to being inclusionary.



*Figure 13 – Prior to VarNOC analysis, samples were analyzed several times with different number of contributors. This analysis was using allele count only, analyzing as N-1 (red) only if the input file allowed, and running a larger subset as N+1 (blue) regardless of any indication of an additional contributor; LogLR is plotted (using combined allele frequencies and  $\theta = 0.01$ ).*

This was demonstrated again by running a subset of validation samples through STRmix v2.6 (Figure 13). When ground truth 2-, 3-, and 4-person mixtures are analyzed as having N+1 contributors (in blue), the effect on the LR is very small compared to an N contributor deconvolution. Only two 4-person mixtures could be analyzed as having three contributors. The results demonstrated exactly what has been shown before, and what is intuitive for this scenario: one contributor from each of these mixtures was excluded when analyzed as having 3 contributors, but there was very strong support for inclusion when analyzed as

having 4 contributors (two red dots at the bottom of Figure 13). In casework, these results start to demonstrate the value of having the ability to assess the evidence samples two different ways, and comparing them mathematically.

Table 1 lists all the samples analyzed with the VarNOC feature in STRmix v2.6. The NOCs column lists the two number of contributors used in the VarNOC analysis of the samples. The results provided after a VarNOC analysis are identical to what would be provided after running two separate MCMCs, but with some additional VarNOC specific information.

With VarNOC, two independent MCMCs are run sequentially using the two different number of contributor assumptions. Mixture proportions, diagnostics, and genotype weights are provided for both MCMCs, in a single report, with additional information on the probability of the number of contributors, given the observed data  $\Pr(N|O)$ .  $\Pr(N|O)$  are the unadjusted contributor number probabilities, which can be used to determine which model is best supported.

In Table 1, NOC1 is the smaller NOC that each sample was analyzed with. NOC2 is the larger NOC (i.e., NOC1 + 1) that the sample was analyzed with. Each of these

Sample	NOCs	Pr N O for NOC1	Pr N O for NOC2
Single Source Sample Full dropout, full dropin	1;2	0.996	3.89E-03
Single Source Sample partial dropout, partial dropin	1;2	0.999	7.52E-04
Low Level Mix 10	2;3	1	3.67E-07
Low Level Mix 11	2;3	1.000	1.06E-04
Low Level Mix 15	2;3	1.000	4.00E-04
Low Level Mix 16	2;3	0.999	1.17E-03
Mix 3-44	3;4	0.993	7.26E-03
24 sec Mix 2-34	2;3	0.997	2.81E-03
24 sec Mix 3-44	3;4	1.000	8.03E-07
Mix 5-1	4;5	1.000	5.53E-07
Mix 5-5	4;5	1.000	7.13E-05
Mix 5-10	4;5	1	1.07E-16
Mix 5-12	3;4	0.753	0.247
ID_Mix_2_9	3;4	0.997	3.28E-03
ID_Mix_2_24	3;4	0.998	1.84E-03
ID_Mix_2_30	3;4	0.008	0.992
ID_Mix_2_35	3;4	0.837	0.163
ID_NIST Mix13_Case 5	3;4	0.999	1.41E-03

Table 1 – A list of all the samples included in the VarNOC analysis. The top 13 samples were amplified with GF, and the bottom 5 were amplified with ID. The NOCs column lists the two different ways the sample was analyzed with a VarNOC MCMC. The probability of the number of contributors (N) given the observed data (O) for the lower NOC is in column 4, and for the higher NOC is in column 5. Most of these samples are better described as having the lower number of contributors, according to STRmix.

samples is different, and there is a different reason for ambiguity in each one (see methods section). With this subset of samples, most of the time, the lower number of contributors is favored, mathematically. The reason for why STRmix would tend to prefer the option with fewer contributors is that minimizing the number of contributors in a given hypothesis increases the overall probability of the observed data. Each contributor added to the hypothesis requires an additional genotype probability in the calculation, which reduces the overall probability of the observed data given: 1) the different hypotheses, 2) the different mass parameters for each contributor, 3)

the different genotype set probabilities, and 4) the number of contributors. The counter weight to this drive to minimize the number

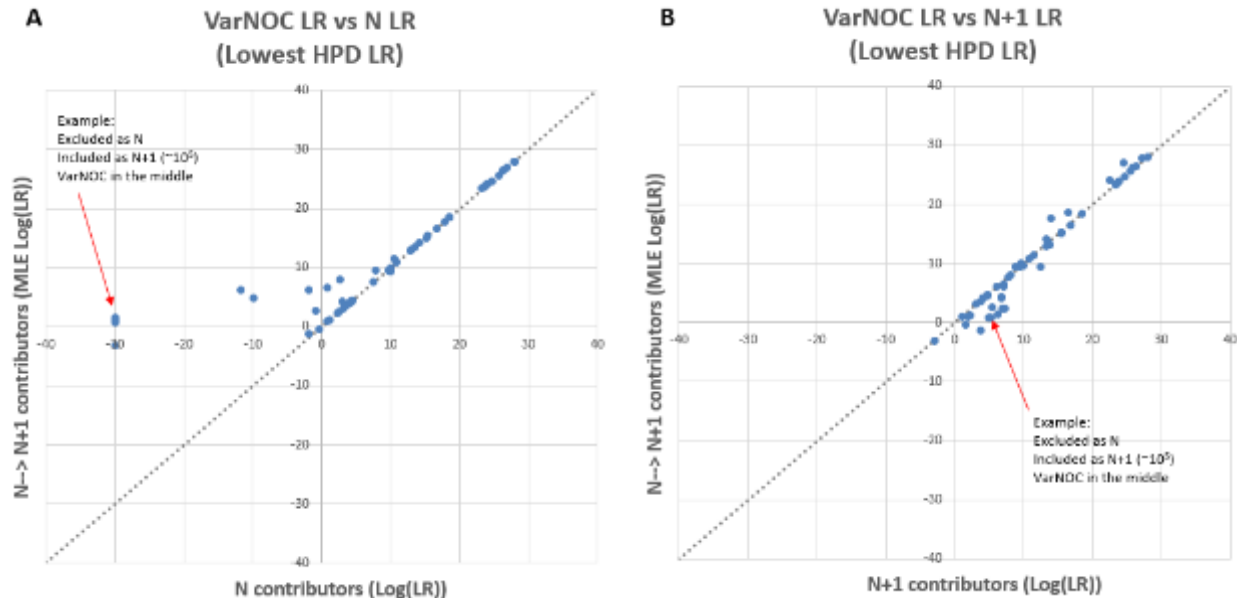
of contributors is that with each additional contributor, any stochastic events or imbalance within the profile can be compensated for by the contribution of the additional contributor(s). The addition of a contributor within STRmix would mostly likely be favored by providing a better description of the observed data. However, if this additional contributor has a low template contribution, the near zero mass of the additional contributor would tend to favor the analysis requiring the fewer number of contributors. This is almost always the case when the number of contributors is in question. In addition, as with any MCMC analysis in STRmix, only the peaks detected above the analytical threshold are input into the software: therefore, any potential DNA peaks present below the that could potentially inform the number of contributors determination are not considered for  $\Pr(N|O)$ .

After the VarNOC analysis is complete, likelihood ratio calculations can be performed against the VarNOC deconvolutions. There are two options for the LR calculations: the stratified LR, and the maximum likelihood estimate, which both use between model weights ( $Z_n$ ) for the calculation of the LRs.

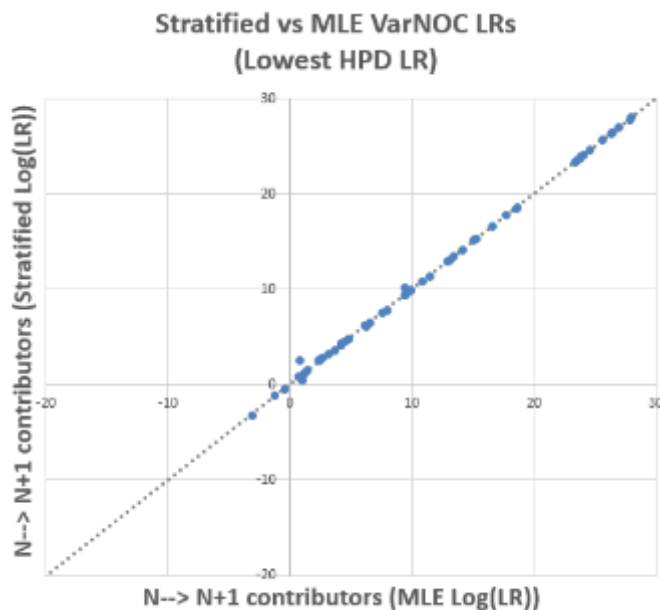
The stratified LR is an LR that essentially uses both NOC analyses in the LR calculation and stratifies the LR across a number of contributors. The stratification of the LR sums across the between model weights (i.e., the adjusted contributor number probabilities) in both the numerator and denominator. Since the unadjusted contributor number probabilities,  $\Pr(N|O)$ , are used to obtain the adjusted contributor number probabilities, their effect can be observed on the VarNOC LRs. For example, if the  $\Pr(N_N|O)$  is .99 for N contributors, and  $\Pr(N_{N+1}|O)$  is 0.01 for N+1 contributors, the VarNOC LR will be very close to the LR for N contributors.

In calculating the MLE VarNOC LR, the profile is evaluated, and STRmix identifies the highest posterior probability of the profile for each hypothesis. The analysis with the maximum posterior probability under each hypothesis is selected for the LR. An MLE VarNOC LR has only one NOC in the numerator and denominator – sometimes they are the same, and sometimes they are different. For the MLE VarNOC LR, the NOC for each proposition is determined by the evaluating the posterior probabilities of the profile (i.e., the sum of the log(likelihood) values for each iteration using the accepted genotype sets).

In Figure 14, VarNOC LRs are compared to the LRs from MCMC 1 (NOC = N, or the lower number in Table 1) in panel A. And these same VarNOC LRs are also compared to the LRs from MCMC 2 (NOC = N+1, or the higher number in Table 1) in panel B. In looking at panels A and B together, these results are similar to what was demonstrated in Figure 12 – that changing the NOC has a negligible effect on some LRs, but it can have a large effect on others, specifically in a situation where one reference is excluded after comparing to a deconvolution with one NOC, but included after a deconvolution with a different NOC. There are arrows pointing to one example of this in Figure 14. This figure shows how the VarNOC LR integrates results from two different deconvolutions using two different NOCs.



**Figure 14** –With every comparison to a VarNOC MCMC, three sets of LRs are provided: the LRs for MCMC 1 (N contributors), the LRs for MCMC 2 (N+1 contributors), and the VarNOC LRs (N  $\rightarrow$  N+1 contributors). VarNOC LRs can either be stratified LRs or MLE LRs. In panel A, the VarNOC MLE LogLRs on the left are plotted against the LogLRs from MCMC 1 with N contributors (on the horizontal axis). In panel B, the same VarNOC MLE LogLRs on the left are now plotted against the LogLRs from MCMC 2 with N+1 contributors (on the horizontal axis).



**Figure 15** – VarNOC stratified LogLRs are plotted against the VarNOC MLE LogLRs (horizontal axis). The LR difference between MCMC 1 and MCMC 2 can sometimes be very large, and the VarNOC LR takes both into account. There is a very small difference in the stratified vs MLE VarNOC LRs.

The MLE VarNOC LR is plotted in Figure 14, but the results would be similar if the stratified VarNOC LR was used. Since every LR comparison was run twice, once calculating a stratified LR, and once calculating the MLE LR, they could be directly compared. Figure 15 plots the two different VarNOC LRs against each other, and there is very little difference between the two.

Selecting whether to perform a MLE VarNOC LR or a stratified VarNOC LR will be dictated by the scenario and the evidence profile obtained. Although both LRs could be used when there is ambiguity in the NOC assigned, the stratified might be used when a profile could be analyzed as either one or another NOC, but there is little indication which NOC would be preferred by the prosecution or defense.

In this situation, the resulting LR would be stratified across both NOC analyses. In contrast, the MLE VarNOC LR might be selected if the profile could be analyzed under two different NOCs, but there was a clear indication that one analysis may be preferred by either party. For example, if there was a sample under which one NOC determination would lead to an exclusion of the person of interest from the mixture; but another NOC analysis would support the person's inclusion, it would be reasonable to assume that the defense proposition would favor the NOC analysis that tended to exclude, whereas the prosecution would propose the NOC that favored inclusion.

*Table 2 – A summary of LRs that can be calculated after VarNOC analysis. The numbers compiled here reflect multiple “LR from previous” calculations and two different DB searches.*

	VarNOC	
	Case 5	
	SDPD Identifiler Plus	
NOC	3	4
Mx Prop.	51:26:23	50:27:23:0
Pr N O	0.996691	0.0033088
POI A Cauc Point LR	37020.60	101125.76
POI A Cauc HPD	13196.69	37429.71
DB search Cauc (Strat=N)	37020.60	
DB search Cauc (Strat=Y)	37407.13	
VarNOC LR: POI A Cauc Stratified - Pt. Estimate	37407.13	
VarNOC LR: POI A Cauc Stratified - Unrelated (HPD)	13767.00	
VarNOC LR: POI A Cauc MLE - Pt. Estimate	37020.60	
VarNOC LR: POI A Cauc MLE (3/3) - Unrelated (HPD)	13196.69	

In the LR settings in STRmix, stratification (across contributor order) can be turned on or off. This also applies to DB searches on VarNOC analyses. When stratification is turned off, the DB search LR matches the point estimate of the number of contributors with the highest  $\Pr(N|O)$ , if FST is used in DB search and allele frequency populations are comparable. When stratification is turned on, the DB search LR matches the point estimate (which it is still called in VarNOC LR calculations) of the stratified VarNOC LR. Table 2 shows one example of the multiple LR calculations that can be made using a VarNOC analysis.

The VarNOC process was examined through both an internal SDPD validation sample as well as an ESR provided VarNOC example. Below follows the VarNOC process determined through the examination of the extended output as well as the ESR example. The VarNOC maths were recreated for the ESR-derived VarNOC example.

- 1) STRmix runs two independent MCMCs using both NOC1 and NOC2
- 2) STRmix will then identify the iteration with the highest posterior probability within each MCMC. This is done for every 10<sup>th</sup> accept by using the  $t$ ,  $d$ ,  $A$ , and  $\lambda$  and the posterior distribution of genotype sets. The  $\log(\text{profile probability})$  [including penalties] is summed across all genotype sets and that is the posterior probability for that iteration. The maximum  $\log(\text{profile probability})$  across all genotype sets in each MCMC is  $\theta_{\text{peak}}$  for each MCMC.



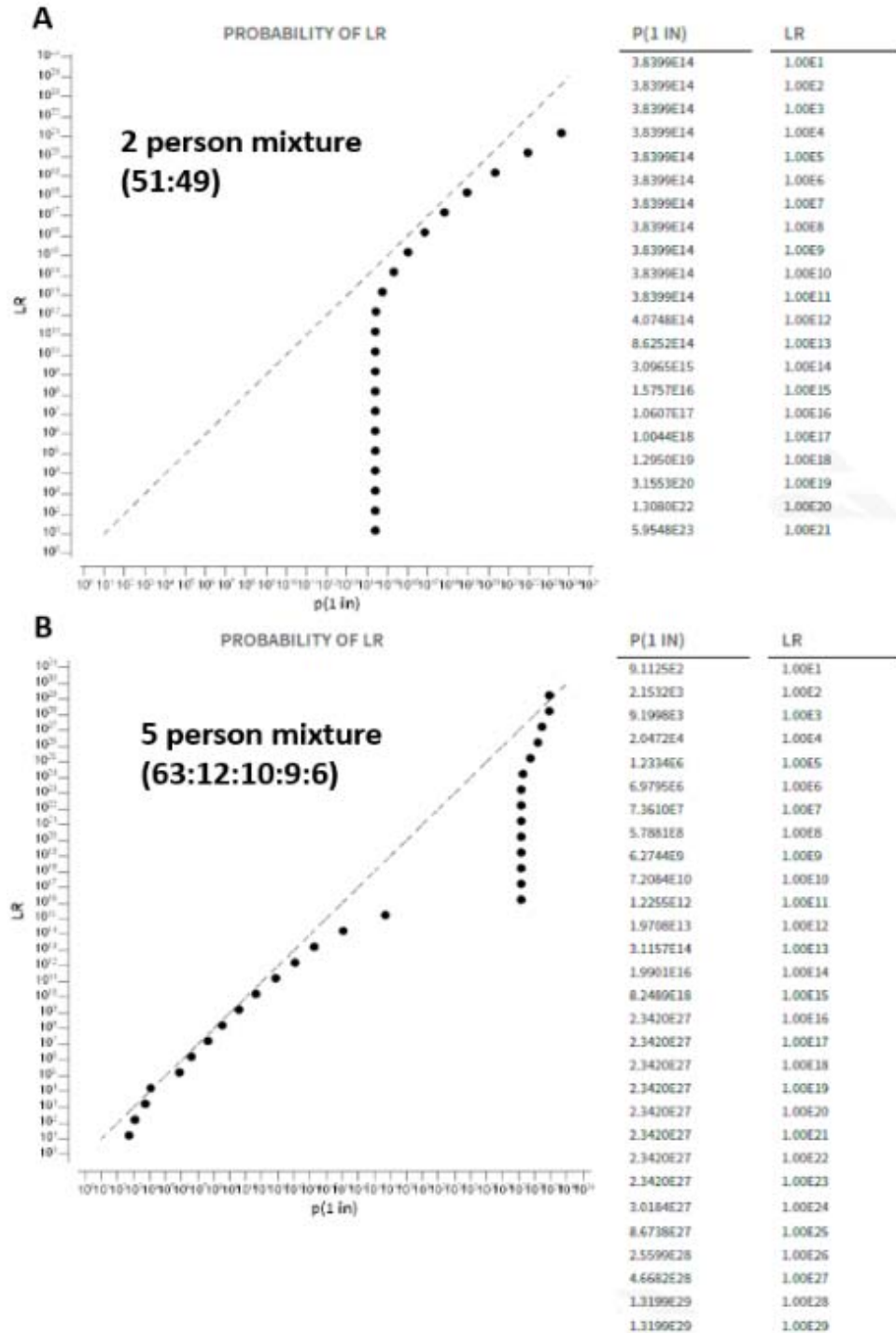
- 3) Each mass parameter (*in each MCMC – we'll just assume all the rest of these steps apply to both MCMCs from here*) in the  $\theta_{\text{peak}}$  iteration is used and the squared distance (squared to remove negative values), Q0 distances, of each mass parameter in all other iterations is determined. The distances are sorted from closest to farthest from the corresponding  $\theta_{\text{peak}}$  parameter. The closet 2.5% of all parameters are selected as one subset of the mass variables in the cosmic-rectangle.
- 4) Go back and calculate the root mean variance of each mass variable ( $t$ ,  $d$ ,  $A$ , and  $\lambda$ s) in all iterations (in each MCMC) and these are the Q1 distances. Sort again from closest to farthest and the top 2.5% of these become another set of parameters that are used in the cosmic-rectangle.
- 5) Now that we have these subsets of  $t$ ,  $d$ ,  $A$ , and  $\lambda$  parameters from the original MCMCs, STRmix will conduct two naïve MC analyses from which the marginal likelihoods will be calculated. The parameters defined in step 4 are the “prior” distributions of these parameters for the naïve MC.
- 6) The naïve MCs will conduct 10,000 iterations randomly selecting mass parameters from the prior distributions. The mass parameters selected will be used to calculate profile probabilities summed across all genotype sets and penalties are assigned for all  $A$ , and  $\lambda$ s, as well as the  $t$  and  $d$ 's. This is done for all iterations within the naïve MC.
- 7) Once all the log(profile probabilities), including penalties, are calculated the average profile probability across the entire naïve MC is determined. This average is then multiplied by the volume of the original MCMC divided by the volume of the hyper-rectangle which is the **adjusted marginal likelihood**. Incorporating the proportion of the hyper-rectangles allows for comparisons between sample spaces that have different dimensionality.
- 8) The log(profile probabilities) are also multiplied by the genotype set probabilities (calculated without  $F_{\text{ST}}$  and summed across multiple possible genotypes at each marker). These are the unadjusted profile probabilities. Once this is done the average unadjusted profile probabilities across the entire naïve MC is determined. This average is then multiplied by the volume of the original MCMC divided by the volume of the hyper-rectangle which is the **unadjusted marginal likelihood**.

The normalized weights for both the adjusted and unadjusted marginal likelihoods are then calculated by dividing the marginal likelihoods (either unadjusted or adjusted – separately) by the sum of the marginal likelihoods of both NOC analyses. The unadjusted is  $p(N|O)$  and the adjusted is  $Z_n$ , used in the LR calculations.

### Exploring the MCMC quality: H<sub>d</sub>-True Tester and database search of random profiles

Efficiency and flexibility are both increased with the addition of the H<sub>d</sub>TT tool. Much smaller DB searches with the elimination profiles can be separated from DB searches of 10,000 random profiles that assist the analyst in providing context for low level LR's. The output from an H<sub>d</sub>TT with importance sampling includes a PDF report with a chart and table stating the probability (1 in x) of obtaining a range of LR's, and also number of effective iterations, average LR, max LR, and min LR (non-zero) for each deconvolution. Figure 16 shows two examples of H<sub>d</sub>TT (with importance sampling) results for two very different mixtures. Figure 16A displays the H<sub>d</sub>TT results for a robust, balanced 2-person mixture (2-18). The results indicate that the probability of obtaining a low magnitude LR is low. This illustrates the high power of discrimination this sample has, and how STRmix was able to deconvolute the contributing genotypes clearly. To pull an example from the chart and relate it to Turing's principle, the probability of obtaining an LR > 10<sup>10</sup> (LR value in the right hand column of Figure 16A) should be approximately 1 in 10<sup>10</sup>. Since 1 in 3.84 x 10<sup>14</sup> (value in the left column of Figure 16A) is smaller than 1 in 10<sup>10</sup>, Turing's rule is demonstrated to hold for this sample. The average H<sub>d</sub>TT LR for this sample is 1.651. The average LR from an H<sub>d</sub>-TT would be anticipated to be around 1 for any sample. Obtaining a value more or less than one for the H<sub>d</sub>-TT indicates that there were slightly more samples in the test that had LR's above or below 1 than would be expected. This does not indicate a problem with the test, but is function of two components inherent in the H<sub>d</sub>TT, namely, 1) the number of random profiles tested and 2) the ambiguity in the component interpretation.

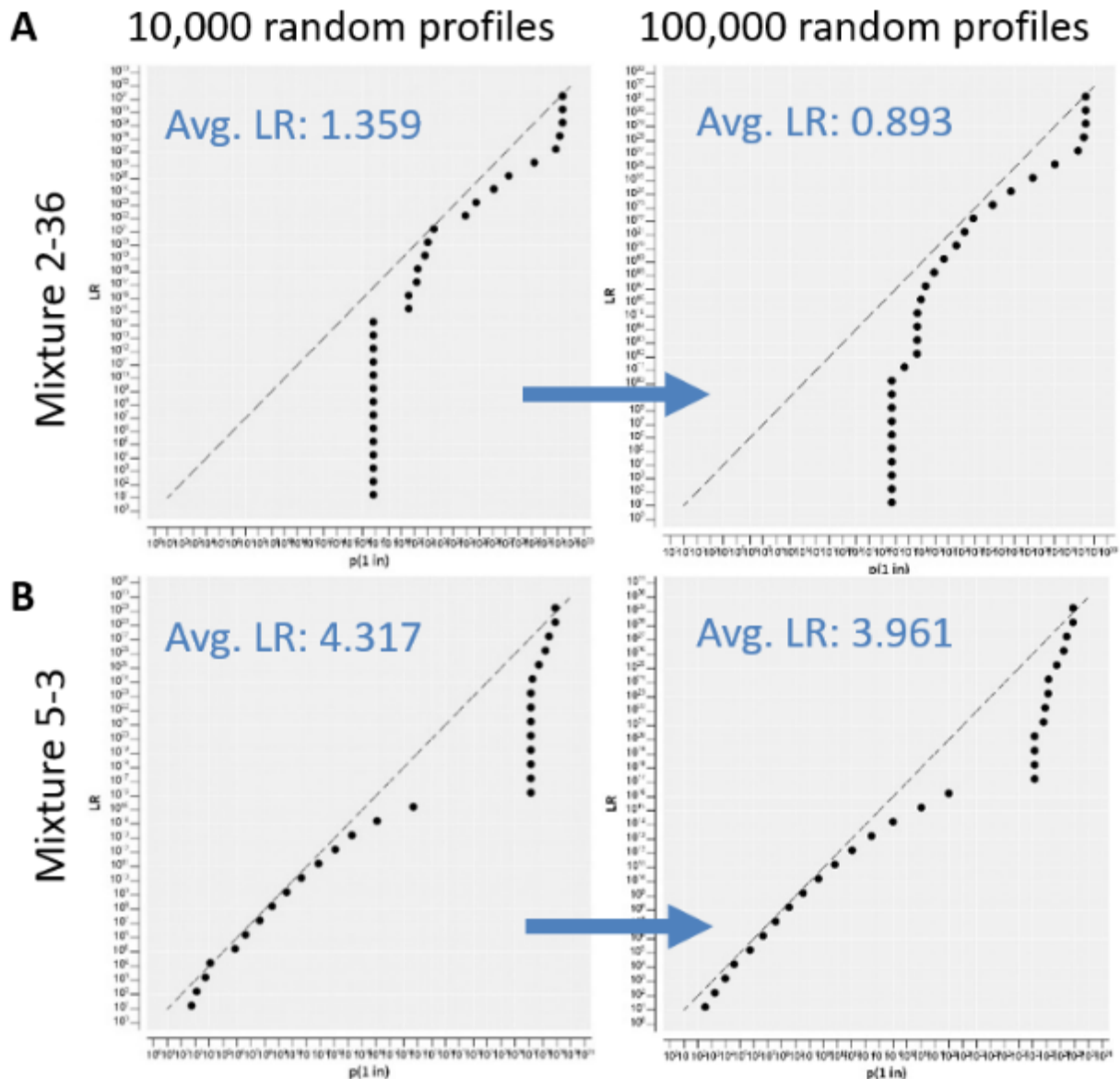
With importance sampling, the H<sub>d</sub>TT is biasing the random samples in favor of genotypes that are possible contributors to the mixture, then extrapolating based on allele frequencies as to how many random profiles that sampling equates with in the population as a whole. If more random profiles were tested in the H<sub>d</sub>TT, then the average LR would begin to move towards 1. In Figure 16B is a 5 person mixture (5-3) with one major component, and 4 low level contributors that all have dropout to some level. The results for this sample are very different – The probability of obtaining a large LR is very small, and stays small to a certain point to ~10<sup>15</sup>, but then the probability more closely follows the Turing principle in that the probability of obtaining an LR of ~10<sup>10</sup> is close to 1 in 10<sup>10</sup>. The average H<sub>d</sub>TT LR for this sample is 4.317. The H<sub>d</sub>TT results from this sample again provide us some context for the sample. The results indicate that there is a significant distinction between one component of the mixture (i.e., the major component) and the remaining components. Based on the deconvolution it would be unlikely to match to the major component based on chance alone. For the lower level components that are more closely aligned in contributor proportion, the deconvolution still adheres to the Turing principal, but there is more ambiguity in the contributing genotypes. In that regard, the average LR indicates that there are more random profiles yielding LR's above 1 than would be expected; however, the more random profiles that are tested would tend to skew the average LR towards 1.



*Figure 16 – Probability of LR results for two different mixtures after running  $H_d$  True Tester with importance sampling: mixture 2-18 (A), and mixture 5-3 (B). The table to the left of the graph contains the same information, just in a different format. Looking at this data graphically indicates several things: that the discrimination of a balanced 2 person mixture is different than a 5 person mixture with a major component, the 5 person mixture has several components, one with less ambiguity than the others (break in the line of dots in B), and that STRmix is robust in deconvolution (demonstrated with the dots below the diagonal line).*

Further illustration of this can be observed in the  $H_d$ -True tests for other samples in the data set. There, the average LR can also provide information about the individual deconvolutions. In this study, the higher order mixtures had higher average  $H_d$ TT LRs. Two of these mixtures (2-36 and 5-3) were selected to increase the number of random profiles used for  $H_d$ TT with importance sampling to see how this value changed (Figure 17), and also to examine the effect on run time. Mixture 2-36 has an estimated mixture ratio of 88:12 where the 12% contributor has a low DNA amount (280 RFU) and is dropping out at some loci. The average LR was 1.359 when using 10,000 random profiles in  $H_d$ TT (with importance sampling). When this was increased to 100,000 random profiles, the average LR changed to 0.893. The average LR in this instance moved closer to 1 when more random profiles were used for the  $H_d$ TT comparison. Adding the additional profile to the  $H_d$ TT had only a small effect on the probability of LR curve (see Figure 16A). Increasing the number of random profiles has a slightly greater effect on a more complex sample. The average LR for the 5-person mixture mentioned above (5-3) changed from 4.317 to 3.961 when 100,000 random profiles were used in the  $H_d$ TT rather than 10,000, but very little effect on the probability of the LR (Figure 17B). The additional random profiles added to the  $H_d$ TT again resulted in moving the average LR towards 1, but also highlights the inherent ambiguity in the profile. This result demonstrates the ambiguity in the low level components.

Because  $H_d$ TT is a tool meant to give information about the mixture as it relates to reference comparisons, this tool can be set up at the same time as a comparison, and isn't necessarily needed until that time.  $H_d$ TT (with importance sampling) can be a time consuming analysis, but since not every deconvolution will need this, there is an ultimate time savings in casework workflow. In addition to only running an  $H_d$ TT test when comparisons are performed, the computer processing times in v2.6 are generally faster than v2.4 due to coding changes, but the time it takes for  $H_d$ TT to run will be dependent on the individual mixture tested. With the caveat mentioned in the methods section about computer and software settings in mind,  $H_d$ TT (random sampling) is similar to random profile database searching, and  $H_d$ TT (importance sampling) is the most time consuming of the  $H_d$ TTs that can be performed with v2.6. Only a few  $H_d$ TT (with importance sampling) analyses were duplicated (not using the same seed) and the analysis time did vary. So, ultimately, just like v2.4, not only are random sample comparisons highly variable mixture-to-mixture, but they are also variable from run-to-run, so these results should only be used as a loose approximation of the time it takes to run an  $H_d$ TT. Table 3 gives an average run time to compare database searching to  $H_d$ TT (importance sampling) analysis for validation samples.  $H_d$ TT (importance sampling) analysis for seven 3-person MCMCs was always under 5 minutes. For seven 4-person MCMCs, it ranged from 20 minutes to an hour, and for nine 5-person MCMCs, it ranged from 3.5 hours to 9 hours.



*Figure 17 – Increasing the number of random profiles in  $H_d$  True Tester with importance sampling has a small effect on the average LR. Although the probability of LR plots have smoothed out, the general shape of the curves is about the same, which indicates that an  $H_d$  True Test is reproducible and reflects the level of ambiguity in the DNA result.*

	2 person	3 person	4 person	5 person
DB search	< 1 min	< 1 min	6 min	1.75 hrs
HdTT IS	< 1 min	2.5 min	40 min	6 hrs

*Table 3 – The average time it takes to evaluate comparisons to random profiles with two different tools: database (DB) search and HdTT importance sampling (IS). All comparisons are to 10,000 random samples, and are rounded averages of run time for each order mixture.*

The time it takes for H<sub>d</sub>TT (importance sampling) to run is also highly dependent on the number of random samples. For the two mixtures that were increased to use 100,000 random samples, the analysis time increased substantially. H<sub>d</sub>TT (importance sampling) analysis went from taking ~30 seconds to ~30 minutes for mixture 2-36, and went from ~5 hrs to ~48 hrs for mixture 5-3; increasing the number of samples 10 fold had a ~10 fold increase on analysis time.

### COSTaR

Data generated with STRmix v2.6 imported properly into COSTaR (COSTaR 4.2-G SDPD 020119.xls). There are more significant digits compared with data generated with v2.4, and each genotype and weight possibility for each contributor in the STRmix v2.6 PDF matched the COSTaR import file for the two mixtures tested. Reference samples from STRmix comparison .txt files were also imported correctly with COSTaR for Suspects (COSTaR for Suspects-G 071518.xls).

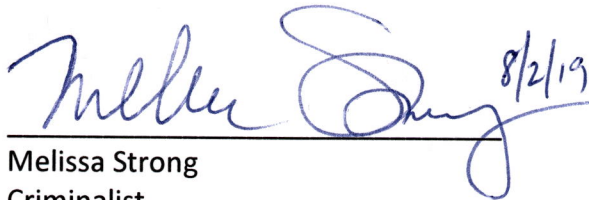
### **Conclusions**

Upgrading STRmix from v2.4.06 to v2.6.2 incorporates a long list of changes, some minor user interface upgrades, some incorporation of tools to increase efficiency (H<sub>d</sub>TT, batch mode and LR batcher), and some major changes, such as incorporating additional types of stutter into the biological model and VarNOC analysis options.

One of the benefits of upgrading to STRmix v2.6 would be an increase in efficiency. A pitfall of our current workflow include LR calculations that have to be set up individually and 10,000 random samples being compared to MCMCs that never have reference samples submitted for comparison. Additionally, not having a theta value in DB search LRs sometimes necessitates an additional LR calculation to be made before a conclusion can be made about the level of support for an elimination sample or a rough guess about how a random sample LR compares to an HPD from a POI. The SDPD casework workflow would change considerably if H<sub>d</sub>TT, DB search with theta (for the elimination database samples), and the newest batch mode were used. A database of only a few hundred samples (the size of the SDPD elimination database) takes considerably less time to run than a >10,000 sample database, so using the “start and search” button would give reportable results faster. If reference samples are being compared to a deconvolution, both the LR from previous and the H<sub>d</sub>TT could be set up at the same time in batch mode so that mixture discrimination is only evaluated when needed (because it is still a time consuming step. Furthermore, multiple references can be compared to a mixture or multiple mixtures (LR batcher) in one batch, getting rid of the night or weekend long lag time that is a hindrance in our workflow with v2.4.

VarNOC analysis increases the time it will take to analyze a sample, but the number of samples that will need VarNOC analysis is likely such that it will have little impact on workflow considerations. For the few cases it will be needed on, it increases our ability to provide information about ambiguous samples, and provides a way to calculate an LR considering two different MCMC.

Overall, there are some big changes in the way we will use STRmix in our casework workflow, but this validation/performance check demonstrates that it is still a robust, discriminating and consistent way of interpreting DNA results and providing conclusions regarding reference sample comparisons. Based on the data obtained from the STRmix v2.6 performance check and the reasons stated above, STRmix v2.6.2 should be implemented in casework at the SDPD.

 8/2/19  
 Melissa Strong  
 Criminalist

 08-02-2019  
 Shawn Montpetit  
 DNA Technical Manager

### References

1. STRmix v2.5.11 Release and Testing Report; ESR; July 3, 2017
2. STRmix v2.6.0 Release and Testing Report; ESR; August 15, 2018
3. STRmix v2.6.0.29 beta to v2.6.0.37 beta to v2.6.0 Summary of Changes; ESR; September 7, 2018
4. STRmix v2.6.2 Release and Testing Report; ESR; April 2, 2019 (updated July 5, 2019)
5. STRmix v2.6 Implementation and Validation Guide; ESR; August 1, 2018
6. STRmix v2.6 Installation Manual; ESR; August 15, 2018
7. STRmix v2.6 Operation Manual; ESR; June 26, 2018
8. STRmix v2.6 User's Manual; ESR; August 1, 2018
9. Recommendations of the SWGDAM Ad Hoc Working Group on Genotyping Results Reported as Likelihood Ratios; Approved July 12, 2018
10. Taylor, D.; Bright J.-A.; and Buckleton J. Interpreting forensic DNA profiling evidence without specifying the number of contributors. (2014) Forensic Science International: Genetics 13:269-280. [VarNOC]
11. Taylor, D; Curran, J; and Buckleton, J. Importance sampling allows Hd true tests of highly discriminating DNA profiles. (2017) Forensic Science International: Genetics 27:74-81. [Importance sampling.]
12. Kruijver, M.; Bright, J.-A.; Kelly, H.; and Buckleton, J. Exploring the probative value of mixed DNA profiles. (2019) Forensic Science International: Genetics 41:1-10. [Importance sampling.]
13. Buckleton et. al. NIST interlaboratory studies involving DNA mixtures (MIX13): A modern analysis. (2018) FSIG 37:172-179. [For comparison to LR for references using the MIX13 samples in this validation.]