STRmix V2.0.6 BFS Casework Internal Validation Summaries

Summaries prepared by Steven Myers: SPM 1/5/16 and Jeanette Wallin: 1/5/16 and reviewed and approved for use in casework by Gary Sims: 1/5/16

Internal Validation Summaries for the Use of Probabilistic Genotyping Software STRmix in Analysis and Interpretation of DNA Case Results

This document summarizes studies conducted by the BFS Richmond Laboratory to assess and internally validate the probabilistic genotyping software program STRmix (ESR/Niche Vision) for the CA Dept. of Justice Bureau of Forensic Services (BFS). STRmix V2.0.6 is a fully continuous probabilistic genotyping program for the interpretation of autosomal STR profiles. While STRmix V2.0.6 was designed to interpret evidence profiles ranging from one to four contributors, the BFS protocol has been validated only for one to three contributors. The validation was specific to Identifiler Plus PCR Amplification Kit results from both the 3130/3130*xl* and 3500/3500xL Genetic Analyzers (Thermo-Fisher/Life Technologies), following current DNA Technical Procedures. FBI population databases [African American, Caucasian and Southwest Hispanic databases - JFS 1999 44(6); FSC 1999 1(2); FSC 2001 3(3)] were used throughout the validation, including a change to the amended data [JFS 2015 erratum 60(4)] once available. The validation studies were performed in accordance with the SWGDAM Guidelines for the Validation of Probabilistic Genotyping Systems (June 2015) and satisfy Standard 8.7 of the FBI Quality Assurance Standards for Forensic DNA Testing Laboratories (September 1, 2011).

In addition to assessing STRmix, this validation included the development and testing of two internally-developed Excel programs: Phantom and CAL DOJ STRmix Report. The first was developed to address forward (N+4) stutter as version 2.0.6 of STRmix does not have the capability to model forward stutter. Phantom creates a profile for STRmix input as an assumed additional "contributor," enabling the software to consider forward stutter peaks as possible alleles from the other contributors. The Phantom also performs a second key function, which is to verify and correct, as needed, locus order in GMID-X exported tables; the STRmix required locus order in exported data from the GMID-X Genotypes tab is not always maintained.

The CAL DOJ STRmix Report serves two main functions. First off, this program must be used to calculate likelihood ratios from STRmix interpreted data that includes a phantom contributor, *instead* of the likelihood ratio function in STRmix. Secondly, this program is used to assess and generate potential profiles for upload into CODIS. CAL DOJ STRmix Report may also alternatively be used to calculate likelihood ratios when there is one person of interest (POI) to be evaluated; if there are two POI's, the STRmix likelihood ratio function must be used instead at this time.

Documentation

SWGDAM probabilistic genotyping validation Guideline 1.3 states, "*The laboratory should document or have access to documentation that explains how the software performs its* *operations and activities, to include the methods of analysis and statistical formulae, the data to be entered in the system, the operations performed by each portion of the user interface, the workflow of the system, and the system reports or other outputs. This information enables the laboratory to identify aspects of the system that should be evaluated through validation studies.*"

There are numerous peer-reviewed publications from the creators of the STRmix program (Duncan Taylor, Jo-Anne Bright, and John Buckleton) describing the function of the components of STRmix, which are additionally addressed in STRmix User Manuals. In addition, the BFS Richmond Laboratory documented validation included verification of the methods and statistical formulae used through calculation reproduction in Microsoft Excel, including genotype probability distribution weights and the different likelihood ratio values. As part of this work, it was determined, in communication with the STRmix creators, that some formulae in the manual are in fact not accurate (see validation corresponding to Guideline 3.2.6 for more information). The workflow of the system and accuracy verification of system reports and output files were inherent to these evaluations.

System Control

System control is addressed in the SWGDAM probabilistic genotyping validation guidelines as Guideline 2.1, which states, "*The laboratory should verify that the software is installed on computers suited to run the software, that the system has been properly installed, and that the configurations are correct.*"

STRmix uses Markov Chain Monte Carlo (MCMC), a random process. For routine use, no seed is set, and so no two runs of a mixture would be identical. You can, however, set a seed to the randomization. By starting the randomization process from that seed, the computer will always return the same set of random values. Thus, results using those random values will be identical as long as all settings and conditions are identical. Four tests were run to examine whether STRmix is running as expected in relation to the SetSeed function:

- Test 1: Demonstrated that SetSeed function leads to identical runs under identical conditions.
- Test 2: Examined whether putting the computer to sleep during a run affects results.
- Test 3: Testing was done on a four processor computer. Test 3 examined whether the same results are obtained when using 4 chains vs. 8 chains.
- Test 4: Confirmed that, after returning the settings to their previous state (no seed; 8 chains), different results were obtained from all runs that used a seed.

Using genotype weight distributions, it was determined that SetSeed function is working as expected, giving identical results when using identical settings. Pausing the computer using the Sleep mode did not affect these results. A different set of weight distributions was obtained when using different numbers of chains, all else being equal. This indicates that all settings, including the number of chains, must match between runs when using SetSeed. Test 4 gave results that differed from the results observed for Tests 1, 2, and 3. Since Test 4 used no seed, this was the expected result. In conclusion, SetSeed will allow you to obtain identical results from identical data and settings. This is being used for quality assurance purposes such as testing to make sure that the settings applied to STRmix on a particular computer are correct.

Guideline 2.2. "*The laboratory should, where possible, ensure the following system control measures are in effect.*"

See below.

Guideline 2.2.1. "*Every software release should have a unique version number. This version number should be referenced in any validation documentation or published results.*"

While the STRmix validation performed at BFS Richmond Laboratory began with version 1.0.7.49, the version ultimately validated is 2.0.6. This is clearly addressed in the BFS Richmond Laboratory validation.

Guideline 2.2.2. "*Appropriate security protection to ensure only authorized users can access the software and data.*"

The BFS Richmond Laboratory is a secure facility.

Guideline 2.2.3. "*Audit trails to track changes to system data and/or verification of system settings in place each time a calculation is run.*"

Per the DNA Technical Procedures involving STRmix, such verifications are required with each case analysis.

Guideline 2.2.4. "*User-level security to ensure that system users only perform authorized actions.*"

N/A.

Developmental Validation

SWGDAM probabilistic genotyping validation Guideline 3. states, "*Developmental validation of a probabilistic genotyping system is the acquisition of test data to verify the functionality of the system, the accuracy of statistical calculations and other results, the appropriateness of analytical and statistical parameters, and the determination of limitations. Developmental validation may be conducted by the manufacturer/developer of the application or the testing laboratory. Developmental validation should also demonstrate any known or potential limitations of the system.*"

Although the developer performed the developmental validation, several studies performed at the BFS Richmond Laboratory may be considered developmental in nature.

Guideline 3.1. states, "*The underlying scientific principle(s) of the probabilistic genotyping methods and characteristics of the software should be published in a peer-reviewed scientific journal. The underlying scientific principles of probabilistic genotyping include, but are not limited to, modeling of stutter, allelic drop-in and drop-out, Bayesian prior assumptions such as allele probabilities, and statistical formulae used in the calculation and algorithms.*"

There are numerous peer-reviewed publications that include but are not limited to these stated topics. See the reference list at the end of this document for key and related publications.

3.2. Developmental validation should address, where applicable, the following:

3.2.1. Sensitivity – Studies should assess the ability of the system to reliably determine the presence of a contributor's(s') DNA over a broad variety of evidentiary typing results (to include mixtures and low-level DNA quantities). This should be evaluated using various sample types (e.g., different numbers of contributors, mixture proportions, and template quantities).

3.2.1.1. Sensitivity studies should demonstrate the potential for Type I errors (i.e., incorrect rejection of a true hypothesis), in which, for example, a contributor fails to yield a LR greater than 1 and thus his/her presence in the mixture is not supported.

3.2.1.2. Sensitivity studies should demonstrate the range of LR values that can be expected for contributors.

See Internal Validation

3.2.2. Specificity – Studies should evaluate the ability of the system to provide reliable results for non-contributors over a broad variety of evidentiary typing results (to include mixtures and low-level

DNA quantities). This should be evaluated using various sample types (e.g., different numbers of contributors, mixture proportions, and template quantities).

3.2.2.1. Specificity studies should demonstrate the potential for Type II errors (i.e., failure to reject a false hypothesis), in which, for example, a non-contributor yields a LR greater than 1 and thus his/her presence in the mixture is supported.

3.2.2.2.Specificity studies should demonstrate the range of LR values that can be expected for non-contributors.

In considering the specificity of the system, the question is asked, "What is the chance a randomly selected person who is not a contributor would be included as a possible contributor to the evidence profile." Using simulated non-contributors, both related and unrelated to the true contributors, specificity tests were performed to address this question using one STRmix interpretation from each of the following:

- Two sets of 1 ng (19:1, 9:1, 4:1, 2:1, 1:1, 1:2, 1:4, 1:9, and 1:19) 2-person mixtures
- One set of 0.5 ng (9:1, 4:1, 1:1, 1:4, and 1:9) 2-person mixtures
- Two sets of 3-person mixtures (1:1:1, 4.5:4.5:1, 6:3:1, and 8:1:1) amplified at varying template inputs (1.5, 0.75, and 0.375 ng)
- A second set of 3-person mixtures (1:1:1, 4.5:4.5:1, 6:3:1, and 8:1:1) amplified at varying template inputs (1, 0.5, and 0.25 ng)
- Two- and 3-person differentially degraded mixtures (1:1 and 6:3:1, both 1 ng input)

For each relationship category, 10,000 profiles were modeled in relation to the tested contributor's reference profile. The following relationships were simulated: parent/child, full-sibling, half-sibling (which has the same degree of autosomal genetic relatedness as uncle/aunt, niece/nephew, grandparent, and grandchild), first cousin, second cousin, and unrelated. Alleles shared by descent between the reference and the modeled non-contributor were in proportion to the kinship coefficients for each relationship:

	Both Ref Alleles	Ref Allele 1	Ref Allele 2	Neither Ref Allele
Relationship	k_2	$k_{1,allele\ 1}$	$k_{1,allele\ 2}$	k_0
Parent-Child	0	0.5	0.5	0
Full-Siblings	0.25	0.25	0.25	0.25
Half-Siblings	0	0.25	0.25	0.5
1st Cousins	0	0.125	0.125	0.75
2nd Cousins	0	0.0625	0.0625	0.875
Unrelated	0	0	0	1

All alleles not shared by descent were selected at random. A population group was randomly selected from four databases in proportion to 2013 California Census data:

Population	Proportion
African American	0.066
Caucasian	0.39
Hispanic	0.384
Other (Han)	0.16

All alleles were sampled from the selected population. The alleles selected for the Unrelated category served as the alleles that weren't identical by descent for the relatives. Substructure was not incorporated into the modeling process.

With full-siblings, parents, and children, there is a clear positive correlation in the reference log(LR) vs. the maximum log(LR) comparisons, and this correlation is seen to decrease as the relationship becomes more distant. As the POI's LR increases, all of the categories tested showed increased dispersion of the proportion of non-contributors, with a downward trend and an increase in LR = 0. This is likely due to the POIs with high LRs being from more discriminating mixture interpretations (*e.g.,* higher template amounts). Of course, the POI could be a major contributor, and the non-contributor could best fit the same or another contributor in the mixture, yielding a high or low LR.

Non-contributors with LRs greater than that of the POI were mostly limited to the low end of the LR range. However, even an unrelated non-contributor gave an LR ≈ 290,000 for comparison when using challenging 3-person mixture data. The same mixture, but in a different capillary electrophoresis (CE) run and with a different reference comparison, gave an LR ≈ 12,000 for an unrelated non-contributor, which was higher than the LR = SPM 2/11/16 ~~8.7~~ 8.27 of the known contributor.

Overall, for unrelated non-contributors to a mixture, the worst case sample had over ~~97%~~ 74%* SPM 2/11/16 of the likelihood ratios less than 1.0. While LRs were observed in the 10^5 range, most LRs were 0. The instances where unrelated non-contributors had LRs greater than the LRs of the comparison reference were limited to reference LRs < 10^3. As predicted by basic biology, LRs of relatives of the reference tended to more often be > 0 and generally showed a positive correlation with the reference LRs.

The Laboratory's proposed verbal predicates were also examined for how they align with the LRs observed for the unrelated non-contributors. None of the 20,000 to 30,000 unrelated non-contributors had an LR > 10 million. This supports that an LR of at least 10,000,000 is a conservative value for the verbal equivalent "strong support." Less than 0.1% of the 20,000 to 30,000 unrelated non-contributors had an LR in the "moderate support" range and there was a progressive decrease in the proportions of non-contributors when going from the "weak evidence" range to the "moderate evidence" range. Most of the comparisons were LR = 0 for all unrelated non-contributors. These data support that the verbal predicates are reasonably conservative.

*SPM 2/11/16 THE WORST CASE SAMPLE HAD OVER 96% OF THE UNRELATED NON-CONTRIBUTOR LRS < 10.

Edits reviewed ✓ 2/11/2016 2/11/16

3.2.3. Precision – Studies should evaluate the variation in Likelihood Ratios calculated from repeated software analyses of the same input data. This should be evaluated using various sample types (e.g., different numbers of contributors, mixture proportions, and template quantities).

3.2.3.1. Some probabilistic genotyping approaches may not produce the same LR from repeat analyses. Where applicable, these studies should therefore demonstrate the range of LR values that can be expected from multiple analyses of the same data and are the basis for establishing an acceptable amount of variation in LRs.

3.2.3.2. Any parameter settings (e.g., iterations of the MCMC) that can reduce variability should be evaluated. For example, for some complex mixtures (e.g., partial profiles with more than three contributors), increasing the number of MCMC iterations can reduce variation in the likelihood ratio.

See Internal Validation

3.2.4. Case-type Samples – Studies should assess a range of data types exhibiting features that are representative of those typically encountered by testing laboratories. These features include those derived from mixtures and single-source samples, such as stutter, masked/shared alleles, differential and preferential amplification, degradation and inhibition.

3.2.4.1. These studies should demonstrate sample and/or data types that can be reliably evaluated using the probabilistic genotyping system.

See Internal Validation

3.2.5. Control Samples – If the software is designed to assess controls, studies should evaluate whether correct results are obtained with control samples.

N/A

3.2.6. Accuracy – *Studies should assess the accuracy of the calculations performed by the system, as well as allele designation functions, where applicable.*

3.2.6.1. These studies should include the comparison of the results produced by the probabilistic genotyping software to manual calculations, or results produced with an alternate software program or application, to aid in assessing accuracy of results generated by the probabilistic *genotyping system. Calculations of some profiles (e.g., complex mixtures), however, may not be replicable outside of the probabilistic genotyping system.*

MCMC Process

STRmix uses the Metropolis-Hastings algorithm to decide when to "take a step" in the MCMC chain. When used routinely, the individual steps along that chain are not memorialized in the STRmix output, but a user-defined option will allow such "Extended Output" to be saved. This output saves every accepted step, omitting rejected steps that weren't taken. To better understand the approach and math behind the STRmix application of MCMC, sets of "accepts" from the extended output were examined and their associated Metropolis-Hastings values were recreated. Files listing the tested genotypes and drop-out probabilities were also recreated in full or in part. In addition, the creation of parameters such as the template amount, mixture proportions, and degradation, were explored. This exploration and reproduction involved identification of coding nuances, such as rounding versus truncation, and coding rules, such as X is always 1 for drop-out (-1) alleles and no stutter is subtracted from these -1 alleles. A 2-person differentially degraded mixture was used in this evaluation, initially using a single amplification STRmix interpretation and later as a joint interpretation.

The parameters and calculations explored included the following:

- The list of possible genotypes for several loci were recreated.
- The lists were compared to the genotypes listed in the Extended Output file for the 20,000 burn-in accepts.
- A maximum of one locus change in genotypes per accept was verified.
- The application of the probability of drop-out was verified to the 13th decimal place; note that it was discovered in collaboration with the developer that the formulae used to calculate the cumulative probability are not correct in the V2.0 manual.
- The change per iteration to mass parameters *t* (template), *d* (degradation), A^l (LSAE – locus-specific amplification efficiency), and *R* (replicate amplification) was verified. This included the application of zygosity, molecular weights, and expected stutter.
- The expected peak heights were compared to the "accepts" listed below and matched to the 3rd decimal place*:
  - 0 to 8, representing the initial phase of burn-in
  - 9,996 to 10,004, representing the half-way point in burn-in, during which a variance setting transition occurs
  - 19,991 to 19,999, representing the end of burn-in
- Log(p) values for the comparison of the expected and observed peak heights at the accepts noted above were recreated to the 4th decimal place*; these values along with the LSAE penalties (see next bullet) contribute to the Metropolis-Hastings (M-H) comparison values.
- LSAE penalty values were recreated for these accepts, matching to 7 significant figures.

- M-H values were recreated for these accepts, matching to 7 significant figures.
- The reported average output mixture proportion, degradation, and LSAE values were verified.
- For the joint interpretation, the effect of the replicate on MW (for assessing expected stutter and -1 alleles) and on M-H values was verified for the same accepts as described above.
    - Expected stutter values at all loci matched when rounded to the 3rd decimal place*.
    - All allele and/or stutter peak $\log(p)$ values matched at all loci when rounded to the 4th decimal place*.
    - The LSAE penalties matched at all loci to 7 significant figures; note the same LSAE values are applied to both amplifications' profiles.
    - M-H values, which are the sum of all $\log(p)$ for alleles, stutter, and LSAE, matched at all loci to 8 significant figures.

In summary, the information present in the extended output run folder is logical and can be replicated with a high degree of accuracy despite the different programs used for the calculation. The small differences may be due to rounding and differences in the number of decimal places allowed by Excel and Java.

In the joint interpretation, the allele and/or stutter peaks from each profile were treated as independent amplifications of the same mixture; the same genotype combinations and d, A, and t values were applied, with only R changing. This reinforces that only replicate amplifications, and not replicate injections or separate extractions/stains/swabs, should be jointly interpreted.

* The STRmix extended output lists expected peak heights to three decimal places and log(p) to four decimal places, so these set the limits for those comparisons.

Likelihood Ratio

The likelihood ratio calculations performed by STRmix were evaluated by manual recalculation and through an alternate, in-house, macro-based Excel spreadsheet (see below) to ensure the calculations matched the descriptions and stated formulae in the relevant literature (*e.g.*, User Manual). The following observations were noted:

- It was discovered that STRmix V1.0.7.49 had a coding error related to the Balding and Nichols (1994) application of θ when there were unknown individuals in the Hp. This was corrected in V2.0.
- There is no requirement for the assumed contributor order to remain constant across multiple population groups (*e.g.*, the POI might be assumed to be Contributor 3 for African American but Contributor 2 for Caucasian). Instead, the LR reported for a particular population group appears to be the highest population-specific LR across all contributor orders. This can lead to the inconsistent selection of contributor order across

the range of population groups tested. However, this avoids the inconsistent selection of contributor order when performing multiple LR calculations but with different sets of populations (*e.g.,* ID_FBI_C calculated in isolation vs. ID_FBI_C calculated concurrently with ID_FBI_ AA and ID_FBI_H). By always using the contributor order that gives the highest LR for the individual population, the LR never changes as you add or subtract concurrently tested population groups.

- The stratified LR calculation is not accurately represented in the User's Manual where, in communication with the STRmix creators, an earlier approach is described. The formula STRmix actually applies appears to be appropriate, and the value reported is predictable. This was observed in both V1.0.7.49 and V2.0. When more than one contributor order is used to report a range of population-specific LR values (second bullet above), the stratified LR included in the report is the one calculated for the contributor order of the first population entered.
- The run time text is transient, disappearing once the results are viewed. Without that information, it is not possible to fully evaluate the issues noted in the second and third bullets above. Therefore, our procedure will recommend saving that text.

CAL DOJ STRmix Report V1.0.xltm

An Excel spreadsheet that was initially created as a means to recreate and test the STRmix LR calculation (see the Likelihood Ratio section above) was subsequently updated with additional features. The spreadsheet has two primary functions:

1. The calculation of an LR, summarized in a 1-page report; and
2. The development of CODIS-ready search profiles, one for each non-assumed contributor to the evidence profile, summarized in a 1-page report.

The final version, "CAL DOJ STRmix Report V1.0.xltm", has the following functionality:

"Report worksheet"

- STRmix interpretation results are imported from the STRmix run folder summary file.
- A comparison reference profile is imported either from GMID-X export tables created using the same reference-sample format as used by STRmix, or from CSV files in the STRmix reference sample file format.
- Run information (e.g., case and item numbers; file names for the evidence profile(s), assumed donor(s), and person of interest; comments; and iterations settings), interpretation quality assessment values (e.g., estimates of mixture proportions; template amounts and degradation estimates are graphed), and quality control measures (*e.g.,* confirmation of correct settings; and evaluations of iterations in relation to the Java cap – see Guideline 4.1.6.3) are captured and displayed in the printed report.
- Likelihood Ratios for a single person of interest, calculated for three FBI amended population groups (African American, Caucasian, and Southwest Hispanic) at $\theta = 0.01$

according to the approach of Balding and Nichols [FSI 64 (1994)]. The individual locus LRs and multilocus profile LRs are displayed, and log(LR) values are graphed.

  - When a phantom profile (see Guideline 4.1.9) has been incorporated into the interpretation, the default LR calculation setting is to remove that profile from the results. This serves to eliminate a slightly anti-conservative shift in the LR due to the integration of the phantom alleles into the Balding and Nichols approach.

"CODIS Search" worksheet

- Run information (e.g., case and item numbers; file names for the evidence profile(s), assumed donor(s), and person of interest; comments; and iterations settings) and quality control measures (e.g., confirmation of correct settings; and evaluations of iterations in relation to the Java cap) are captured and displayed in the printed report.
- CODIS profiles are created for each contributor using the following rules:
  - At a locus, the search profile is based upon the profiles contained in the top [#] subset of genotype combinations, where [#] is a user-defined proportion (default = 0.95, or 95% of the total weight) and the genotype combinations have been ranked by order of descending weight.
  - The list of genotypes for a contributor is reduced to the corresponding set of alleles:
    - Alleles found in every row of the subset are assigned as obligate.
    - If there are two obligate alleles, those alleles are assigned as a genotype. Provisions are also made for homozygous genotypes.
    - If a -1 allele (drop-out) is within the list, the locus will only be used if there is an obligate allele.
    - If a locus has more than the user defined number of alleles (4 is the CODIS default), the locus is not used.
- Random Match Probabilities at $\theta = 0$ are calculated at each locus in a manner consistent with CODIS moderate stringency rules. For example the msRMP for a single allele P would include the probability of P,P homozygotes as well as P,NotP heterozygotes, while the msRMP for alleles P and Q would include the probabilities for P,P and Q,Q homozygotes as well as P,Q heterozygotes.
- Combined msRMP are calculated for each contributor at each of the three databases listed above. These profile msRMP values may or may not include the loci D2S1338 and D19S433 per a user-defined setting. The inclusion of these two loci may be advantageous for California SDIS searches, but they are not currently compared for NDIS searches.

Accuracy of the LR calculations was checked through comparison to STRmix V2.0.6 using a number of different sample types with different features, including single-source, two-person, and three-person, profiles with and without drop-out, profiles with and without a phantom contributor, interpretations with and without an assumed contributor, and single amplification vs.

joint amplification interpretations. With the partial exception of the phantom interpretation, STRmix V2.0 and CAL DOJ STRmix Report V1.0.xltm gave identical LRs. With the single-source phantom interpretation, the default setting in the Excel spreadsheet removed the phantom, and thus gave a number different than STRmix but identical to the STRmix interpretation of the same donor's profile without a phantom. An option in the spreadsheet to include the phantom did, however, give an LR identical to STRmix for the same data.

CODIS profiles were created manually for a number of different STRmix interpretations with different features, including single-source, two-person, and three-person, profiles with and without drop-out, profiles with and without a phantom contributor, and interpretations with and without an assumed contributor. msRMP were also recreated for three comparisons. All profiles and msRMPs were identical to those created by CAL DOJ STRmix Report V1.0.xltm.

3.2.6.2. If the software uses raw data files from a genetic analyzer as input data, the peak calling, sizing and allele designation functions should be compared to the results of another software system to assess accuracy. Allele designations should also be compared to known genotypes where available.

N/A

Internal Validation

SWGDAM probabilistic genotyping validation Guideline 4 states, "*Internal validation of a probabilistic genotyping software system is the accumulation of test data within the laboratory to demonstrate that the established parameters, software settings, formulae, algorithms and functions perform as expected. In accordance with the QAS, internal validation data may be shared by all locations in a multi-laboratory system. Depending on the features and capabilities of the probabilistic genotyping system, some DNA typing results may or may not be determined to be suitable for such analysis. To identify data features (e.g., minimum quality requirements, number of contributors) that render a profile appropriate or inappropriate for probabilistic genotyping, the laboratory should test data across a range of characteristics that are representative of those typically encountered by the testing laboratory. Data should be selected to test the system's capabilities and to identify its limitations. In particular, complex mixtures and low-level contributors should be evaluated thoroughly during internal validation, as the data from such samples generally help to define the software's limitations, as well as sample and/or data types which may potentially not be suitable for computer analysis. In addition, some exclusions may be evident without the aid of probabilistic software. If conducted within the same laboratory, developmental validation studies may satisfy some of the elements of the internal validation guidelines.*"

Studies involving all such criteria are described below.

Guideline 4.1. "*The laboratory should test the system using representative data generated in-house with the amplification kit, detection instrumentation and analysis software used for casework. Additionally, some studies may be conducted by using artificially created or altered input files to further assess the capabilities and limitations of the software.*"

All data tested was generated from the following components currently used for DNA casework in BFS: Identifiler Plus PCR Amplification Kit, either or both the 3130/3130*xl* and 3500/3500xL Genetic Analyzers, and GeneMapper ID-X, version 1.4. Artificially created data was used to evaluate saturated peaks (see Guideline 4.1.4) and the Phantom program (see Guidelines 4.1.9 and 5).

Internal validation should address, where applicable to the software being evaluated:

4.1.1. Specimens with known contributors, as well as case-type specimens that may include unknown contributors."

Known Contributors, With and Without Assumed Donors

When performing mixture interpretation using STRmix, the hypotheses may include an assumption of the presence of one or more contributors to the mixture. For example, in a typical sexual assault case, the sperm fraction extract may contain some incompletely separated non-sperm fraction DNA, typically from the victim's vaginal epithelial cells. In such a case, it would be reasonable to assume for the sperm fraction mixture interpretation that one of the contributors is the victim (or her profile as observed in the non-sperm fraction).

This study examined the impact of assuming contributors using a variety of mixed profiles. Two and three-person mixtures of various ratios were used, as was a limited set of samples with differential degradation. In each mixture, each contributor was run as an assumed contributor. Each pair of donors was also run as two assumed contributors for the three-person mixtures.

For 2-person mixtures, a mixture study was tested that included 1ng (19:1, 9:1, 4:1, 2:1, 1:1, 1:2, 1:4, 1:9, and 1:19) and 0.5 ng (9:1, 4:1, 1:1, 1:4, and 1:9) input mixtures, all amplified in duplicate. For 3-person mixtures, three samples were amplified at varying ratios (1:1:1, 4.5:4.5:1, 6:3:1, and 8:1:1) in duplicate using 1.5, 0.75, and 0.375 ng input template. Each amplification was interpreted twice with STRmix, including joint interpretations for the duplicate amplifications and each interpretation involving an assumed contributor(s).

The 3-person mixtures were challenging at the 375 pg level because of dropped-out alleles leaving insufficient evidence of donor proportions. In three of the four ratios, one or both amplifications didn't have more than 4 alleles at any one locus. In such situations, STRmix appears to divide the mixture proportions evenly across the donors. The current study in part assessed whether assuming contributors would aid the software in the interpretation of such low-level results.

2-Person Mixture Results

All LRs were > 1 regardless of whether a contributor was assumed. For 4:1, 1:4, 9:1, 1:9, 19:1, and 1:19 mixtures, assuming a contributor had almost no effect on the LR, suggesting that there was little gained from the increased knowledge with these interpretations. This is likely due to the observation that the major contributor's genotypes are already well defined at these ratios without the assistance of an assumed contributor. Knowing the minor donor's genotypes for these mixture ratios didn't appear to improve the genotype calls for the major donor. Similarly, narrowing down an already well-defined major donor profile didn't appear to meaningfully reduce the ambiguity in the minor donor's genotype calls. On the other hand, the 1:1, 2:1 and 1:2 mixtures demonstrated obvious benefits from assuming a contributor. This was true for both the major and minor contributors to the 2:1 and 1:2 mixtures.

Duplicate interpretations were examined for consistency in the log(LR). Almost all duplicate LR results were within a factor of 2, regardless of whether or not one contributor was assumed. Precision generally improved with higher quantities of template DNA. The addition of an assumed contributor generally improved precision for contributors who contributed ~0.25 ng or more to the mixture's total template DNA. Below this level, the addition of an assumed contributor had little/no practical effect on precision.

3-Person Mixture Results

In the great majority of comparisons, and for every average of the differences between multiple interpretations, the addition of one or more assumed contributors led to greater sensitivity [higher log(LR)], and none of the contributors gave a complete false negative (LR = 0). In other words, assuming one or two of the contributors generally led to increases in the LR for the remaining comparisons. As with interpretations using no assumed contributors, assuming one contributor occasionally led to negative log(LR) values for non-assumed donors. Where LRs went down with assumed contributors, reductions were generally small (< 1 log unit). Only one sample (a 6:3:1 0.75 ng mixture) displayed decreases of more than one log unit. The percentages of comparisons with LRs < 1 are listed in the table below.

	All contributors		10% contributors	
Assumed	Comparisons	LR < 1	Comparisons	LR < 1
0	108	7.4%	36	17%
1F	72	1.4%	36	2.9%
2F	72	2.8%	27	3.7%
3M	72	2.8%	9	0.0%
1F and 2F	36	0.0%	27	0.0%
1F and 3M	36	0.0%	9	0.0%
2F and 3M	36	0.0%	n/a	n/a

When two contributors were assumed, all interpretations gave positive log(LR) values for the non-assumed donors.

To assess precision, duplicate interpretations were examined for consistency in the log(LR). Precision generally improved with higher quantities of template DNA. Interpreting with one assumed contributor did not appear to have a demonstrable effect on precision with this set. Using two assumed contributors had a more visible effect. The minimum and maximum log(LR) shifts, as well as the proportion with shifts > 1 log unit, are summarized in the table below.

			Δlog(LR)	
Assumed	Comparisons	Min	Max	% > 1
0	108	3.9E-04	2.5E+00	1.85%
1F	72	9.3E-04	4.3E-01	0.00%
2F	72	4.4E-05	8.8E-01	0.00%
3M	72	2.8E-04	1.2E+00	2.78%
1F and 2F	36	0.0E+00	8.2E-01	0.00%
1F and 3M	36	0.0E+00	2.8E-01	0.00%
2F and 3M	36	0.0E+00	6.1E-01	0.00%

Overall, donor assumptions generally have a positive effect on the ability of STRmix to assess a mixture. With assumed donors, sensitivity tended to be equivalent or higher, and precision tended to be tighter. Where the assumption does not add meaningful information (*e.g.,* assuming the major or minor contributor to a 9:1 mixture), the effect may be neutral. Only one instance was observed where the assumption was detrimental to the point of causing a complete false negative. For one amplification of a three-person, 6:3:1 mixture with differential degradation, when the minor (intact) contributor was assumed alone or in concert with the major (intact) donor, the middle (degraded) donor gave LR = 0 for half of the interpretations performed.

Case-Type Specimens

Two- and 3-person differentially degraded mixtures were assessed for sensitivity and precision comparing interpretations performed using no assumed donor and one assumed donor. The 2-person mixture was a 1:1 mixture with 1ng input DNA and one of the two donors degraded. The 3-person mixture was 6:3:1 with 1ng input DNA and the 3-part donor degraded. These samples showed the predominant trend of equivalent or better sensitivity and precision with assumed donors. One amplification of the 3-person differential degradation sample did, however, give sporadic false negatives for the degraded donor when the minor contributor was assumed alone or in concert with the major contributor.

Specifically, for the 2-person mixture, all interpretations had log(LR) > 0, but the results assuming a contributor were at a magnitude observed for single-source profiles. While each profile or combination of profiles had only two interpretations for each assumed donor condition, the precision was patently better for the first of two amplifications and the joint interpretation than when no contributor was assumed.

For the 3-person differentially degraded mixtures, all comparisons were log(LR) > 0 except for the two (see above) in which the degraded donor was excluded. All replicate interpretations were within 1 log unit with the few sporadic exceptions noted above.

An additional 3-person challenging mixture without degradation was assessed through STRmix. This was a 4.5:4.5:1 mixture with 0.375 ng input DNA. Assuming donors gave equivalent or better sensitivity for individual and joint interpretations when compared to the equivalent without an assumed donor(s). All log(LR) values from assumed-donor interpretations were > 0 (*i.e.,* LR > 1) with none of the false negatives that were observed when no donors were assumed. Relative to no assumed donor(s), precision was equivalent or better with assumed donors, such that all but one Δlog(LR) value was < 1; the Δlog(LR) > 1 was 1.58 log units, which would be inconsequential since the lower LR was over 100 quadrillion. When no donors were assumed, precision was good with the vast majority Δlog(LR) < 1.

Additional Known Contributor Studies

Additional 3-person mixtures (1:1:1, 4.5:4.5:1, 6:3:1, and 8:1:1) were each amplified at 1, 0.5, and 0.25 ng and run on both a 3130*xl* and a 3500xL Genetic Analyzer. The performance of STRmix using 3500xL data appears to be equal to or better than when using 3130*xl* data. Consistent with previous findings, sensitivity again appeared to suffer when STRmix failed to correctly identify the mixture proportions of the contributors. The only results that gave log(LR) < –1 were 3130*xl* 6:3:1 and 4.5:4.5:1 interpretations that had STRmix-estimated mixture proportions of ~1:1:1 related to low template input. However, all results gave LRs > 0 with one exception involving a Java cap limitation (see Guideline 4.1.6.3).

Also, see Guideline 4.1.13.

4.1.2. Hypothesis testing with contributors and non-contributors

4.1.2.1. The laboratory should evaluate more than one set of hypotheses for individual evidentiary profiles to aid in the development of policies regarding the formulation of hypotheses. For example, if there are two persons of interest, they may be evaluated as co-contributors and, alternatively, as each contributing with an unknown individual. The hypotheses used for evaluation of casework profiles can have a significant impact on the results obtained.

The studies performed with and without assumed donors considered varying hypotheses, such as all permutations of:

- Hp = assumed + known *versus* Hd = assumed + unknown
- Hp = assumed + known + unknown *versus* Hd = assumed + 2 unknowns
- Hp = assumed 1 + assumed 2 + known *versus* Hd = assumed 1 + assumed 2 + unknown

See the summaries for Guidelines 4.1.1 and 4.1.13 for more information.

Additionally, one of the training sets run by each of the five trainees encompassed varying likelihood ratios that included the following:

- Hp = known 1 + known 2 *versus* Hd = 2 unknowns

The varying likelihood ratios followed predictable trends based on varying assumptions, the number of unknown contributors, and varying weights related to template level and donor number.

4.1.3. Variable DNA typing conditions (e.g., any variations in the amplification and/or electrophoresis parameters used by the laboratory to increase or decrease the detection of alleles and/or artifacts).

3130/3130*xl* and 3500/3500xL Genetic Analyzers

Both Genetic Analyzer models – 3130/3130*xl* and 3500/3500xL – are currently in use for DNA casework across BFS, therefore sensitivity and precision were compared between the models. Two 3-person mixture sets were included in this study. The mixtures of each set (1:1:1, 4.5:4.5:1, 6:3:1, and 8:1:1) were amplified in duplicate using either 1.5, 0.75, and 0.375 ng DNA input or 1.0, 0.5, and 0.25 ng DNA input. All amplifications were interpreted at least twice, and when duplicate amplifications were available, jointly.

With one exception, the 3130 and 3500 results for these mixture series all gave LRs > 0. LRs tended to diverge when STRmix assessments of mixture proportion diverged. For example, the 3130 profiles for the two 0.375 ng amplifications of mixture 4.5:4.5:1 displayed no more than four alleles at any one locus. When performing joint interpretations of these profiles, six of ten interpretations misjudged the true mixture proportions (each donor was interpreted by STRmix to have contributed ~1/3 of the template, a trend also observed in the single-amplification interpretations of this mixture's profiles). This led to reduced LRs for the two majority contributors, and LR = 0 for the minor contributor. With the four of ten interpretations that interpreted a more accurate set of mixture proportions, all contributors had higher LRs with all LRs > 0.

In considering precision, only one replicate from the higher template quantity set had a difference greater than 1 log unit (LR difference of 10X). This was in the two 3130 interpretations of the second 4.5:4.5:1 0.375 ng amplification and was related to differences in the mixture proportions between the interpretations. The second set had 4 and 5 differences greater than 1 log unit, respectively, for 3130*xl* and 3500xL. These were generally contributors with low template quantities, and this set was already uniformly lower template than the first set.

Under the current set of Bureau laboratory procedures and analytical settings thresholds, the performance of STRmix using 3500 data appeared to be equal to or better than when using 3130 data. This is largely related to the difference in dynamic ranges which is a result of disproportionate analytical thresholds (50 RFU versus 150 RFU). Regardless, it is acceptable to use STRmix, with the appropriate corresponding parameters (*e.g.,* allele variance), to interpret Identifiler Plus results from either the 3130/3130*xl* or 3500/3500xL Genetic Analyzer. Furthermore, it is expected that 3500/3500xL data will provide higher sensitivity with comparable precision.

Injection Time

The effect of reduced injection time was examined using three-person mixtures (1:1:1, 4.5:4.5:1, 6:3:1, and 8:1:1) that were each amplified at 1.0 ng. The mixtures were injected both for the standard time and at a reduced injection time on a 3130*xl* (5 and 3 seconds) and a 3500xL (24 and 12 seconds) Genetic Analyzer. Unsurprisingly, reducing the injection time often led to similar or lower LRs compared to the full injection times. This is to be expected since reduced injection time could lead to the reduced detection of alleles and the associated issue of poorer estimates of the mixture proportions. Additionally, the larger allele variances for reduced peak height would allow for more possible combinations of genotypes, contributing to the reduced LRs and an increased proportion of replicate pairs that differ by > 1 log unit. Regardless, all results gave LRs > 0 with one exception involving a Java cap limitation (see Guideline 4.1.6.3).

4.1.4. Allelic peak height, to include off-scale peaks

An off-scale peak study was conducted to test the ability of STRmix to detect true genotype combinations and donor proportions in mixtures with saturated peaks. The study used two artificial 4:1 two-person Identifiler Plus mixtures scaled to various template levels correlating to

RFU heights above and below the STRmix saturation RFU setting. In the test profiles, RFUs were capped at the saturation point to simulate what can occur when the PCR product's fluorescent signal exceeds the 3130/3130*xl* CCD camera's ability to quantify. Additionally, one interpretation was performed using a high RFU (scaled above the saturation setting) profile that had no RFU cap applied. By setting STRmix to collect the extended output, this allowed for an examination of the MCMC Metropolis-Hastings calculation in the face of saturation.

It was determined that although saturation can have an effect on the probability assigned to genotype combinations and the overall likelihood ratios, STRmix demonstrated a high level of success at using the information from non-saturated peaks to adequately determine contributors' relative template amounts as well as correct genotype combinations. Even faced with a major contributor that is fully saturated, template amounts maintained a highly accurate ratio, and the overall likelihood ratios were within an order of magnitude of those obtained without any saturated peaks. Ratios other than the tested 4:1 might be expected to show greater ambiguity in genotype assignment if the saturation masks critical information regarding allele sharing, although the stutter peaks appear to be acting as a surrogate for this lost RFU. Increased numbers of contributors will also, as usual, lead to greater ambiguity since the "elevated stutter" could also reasonably be accounted for as minor donor alleles.

Overall, it appears that there is no significant detriment in using STRmix with data that has some saturated peaks, especially in lower-order mixtures and where multiple contributors are not saturated. It should be noted, though, that saturated data may be associated with spectral and amplification artifacts that could prove challenging to both the examiner and STRmix.

4.1.5. Single-source specimens

Identifiler Plus sensitivity studies from five qualified analysts, run on both 3130/3130*xl* and 3500/3500xL Genetic Analyzers, were used to evaluate the interpretation of single-source samples by STRmix. The performance of STRmix using 3500/3500xL data appears to be equal to or better than when using 3130/3130*xl* data. As noted in previous studies, much of this may be attributed to the greater overall proportion of peaks detected under the 3500/3500xL procedure due to differences in the analytical thresholds. Each study included single amplifications at the following quantities: 2, 1, 0.5, 0.25, 0.125, 0.062, and 0.031 ng, and most studies included 0.016 ng. Each amplification was interpreted twice in STRmix.

Regardless of the CE used for single-source samples, STRmix sensitivity was 100%, no population-specific LRs were < 1.0, and STRmix behaved in a logical manner as template quantity increased. For all template quantities at or above 0.25 ng, STRmix assigned 100% of the weight to a single genotype at all loci. Therefore, those template quantities gave identical LRs within a sensitivity study.

When considering template quantities below 250 pg, data from both Genetic Analyzer models displayed a logical increase in log(LR) values as the template amount increased from the lowest quantity. The 3500/3500xL Genetic Analyzer data trend lines in this range showed a higher log(LR) trend than for the 3130/3130*xl*, likely due to the increased proportion of alleles being detected for 3500/3500xL profiles at the same template level.

Duplicate interpretations were examined for consistency in the log(LR). Precision was well within a factor of 2.

See also Single-Source under Guideline 4.2.

4.1.6. Mixed specimens

4.1.6.1. Various contributor ratios (e.g., 1:1 through 1:20, 2:2:1, 4:2:1, 3:1:1, etc.)

See Guidelines 4.1.1, 4.1.6.4, and 4.1.13.

4.1.6.2. Various total DNA template quantities

See Guidelines 4.1.1, 4.1.6.4, and 4.1.13.

4.1.6.3. Various numbers of contributors. The number of contributors evaluated should be based on the laboratory's intended use of the software. A range of contributor numbers should be evaluated in order to define the limitations of the software.

Samples containing DNA from one, two, and three donors were tested. While STRmix V2.0.6 can accommodate up to four donors, the use of the fourth donor is reserved for the Phantom assumed contributor, when needed. All such combinations were tested as described within this document. In particular, see Guidelines 4.1.5, 4.1.6, and 4.1.13.

Limitation – Java Cap

When jointly analyzing replicate amplifications of an extract, a limitation of the software was identified when sample entry order was changed. Variation in estimated template quantity, inter-PCR efficiency, and log(LR) values and discussions with STRmix creators led to the discovery of the following:

- The list of candidate genotypes is created from the first evidence profile imported into STRmix. This list can affect the rate of accepts.

- In our case involving a joint interpretation of two amplifications, importing the second amplification prior to the first appeared to slow the acceptance rate relative to the first amplification being imported before the second.
- Java has a limitation on the number of iterations: 2,147,483,647. Once that number of iterations has been exceeded, STRmix may start deleting genotype combinations, possibly leading to false exclusions as occurred in our case.

As a result, as part of the DNA Technical Procedure, the analyst is required to verify the number of iterations has not exceeded the Java maximum; as an additional verification, the CAL DOJ STRmix Report was programmed to have an automatic check of this value. If a STRmix run exceeds the Java "cap," the analyst must re-run the sample to proceed with mixture interpretation. The analyst may either change the sample order or increase the proportion of accepts assigned to the burn-in phase using the same overall number of accepts (as a way of preserving convergence time). These changes were recommended by the programmer of STRmix and evaluated during validation.

4.1.6.4. If the number of contributors is input by the analyst, both correct and incorrect values (i.e., over- and under-estimating) should be tested.

The number of contributors assumed to be present in a mixture is one of the required entries for STRmix version 2.0/2.0.6. When limiting interpretations to profiles assumed to be from 1 to 3 individuals, a combination of allele counts and inter-allelic peak height observations should minimize incorrect assumptions. However, the true number of contributors to an unknown mixture may not always be self-evident. Part I of evaluating a correct (True) vs. incorrect (True ± 1) donor number assumption in STRmix included single-source samples and mixtures where the donor number would be easily identified by simple allele counts. For the single-source samples, a sensitivity study was used that included a dilution series from 2.0 ng to 0.016 ng. For 2-person mixtures, a mixture study was tested that included 1ng (19:1, 9:1, 4:1, 2:1, 1:1, 1:2, 1:4, 1:9, and 1:19) and 0.5 ng (9:1, 4:1, 1:1, 1:4, and 1:9) input mixtures, all amplified in duplicate. For 3-person mixtures, three samples were amplified at varying ratios (1:1:1, 4.5:4.5:1, 6:3:1, and 8:1:1) in duplicate using 1.5, 0.75, and 0.375 ng input template.

Part II of this study used a 3-person mixture that was designed to have significant genotype overlap so as to prevent simple allele counts from being an accurate evaluation of donor number. At any one locus, anywhere from 2 to 4 separate alleles are detected. While inter-allelic peak height observations should bring into question incorrect assumptions, this study examined the effect on interpretation by STRmix of assuming correct (True) and incorrect (True – 1) numbers of donors. The 3-person mixtures were at varying ratios (1:1:1, 4.5:4.5:1, 6:3:1, and 8:1:1), amplified in duplicate using 1.0, 0.5, and 0.25 ng input template.

An additional element to Part II is an examination of the effect from increasing burn-in time (20K to 80K) while also increasing the total number of accepts (100K to 130K). These numbers were selected to allow for an extended period of MCMC exploration at the looser allele variance used during the first half of burn-in (*40K* accepts vs. *10K* under our standard settings) while maintaining an identical number of accepts (across the second half of burn-in plus readout) performed at the Laboratory's allele variance (*40K + 50K = 90K* vs. *10K + 80K = 90K* under our standard settings).

Part I

In single-source samples, the software appeared to maintain the true contributor as a clear major contributor across much of the tested template range. With template as low as the 62.5 pg sample, the LR under the assumption of two contributors was largely unchanged in comparison to an assumption of a single contributor. Below that level of template, STRmix interpreted enough ambiguity that the LR reduced to 1.0. False exclusion rates were still 0% regardless of donor number assumption, because no LRs < 1.0 were observed. Inter-interpretation precision was within a factor of 2 regardless of the donor number assumption.

None of the 2-person mixtures tested had a low enough template level for both contributors to pass as single-source. The software always detected an excess of alleles for this assumption. When interpreted as a 3-person mixture, the major contributor's LRs were minimally affected, whereas the minor contributor's LRs were generally shifted downward. This shift was most notable for the more disparate ratios (*i.e.,* 1 ng amplifications at 1:9, 9:1, 1:19, and 19:1) and with lower template amounts (*i.e.,* 0.5 ng amplifications at 1:4, 4:1, 1:9, and 9:1). These samples had a median shift of 1.2 log units between correct and incorrect donor number assumptions.

The 2:1 and 1:2 mixtures could be very sensitive to donor number assessments. When interpreted as three-person mixtures, the nature of the ratio led to the software sometimes settling upon a 1:1:1 ratio and other times a more correct 14:7:1 ratio. When both interpretations were "correct," the median shift was less than 0.05 log units. When one or both interpretations were "incorrect," shifts were as high as 14 log units. Inter-interpretation precision within a donor-number assumption was worse for donors present at 0.05 ng of template DNA (*i.e.,* the minor contributor to a 19:1 or 1:19 1 ng amplification or a 9:1 or 1:9 0.5 ng amplification) when interpreted as a three-person mixture than as a two-person mixture. Even then however, the two interpretations were within 10X.

Three-person assumed interpretations of both 1:2 1 ng amplifications gave LRs that differed by a factor of >10,000 (major contributor) and >1 trillion (minor contributor) due to the different mixture proportions determined for each interpretation. Other than these, only one pair of interpretations differed by >10X, and that was the minor contributor to a 1:4 mixture amplified at 0.5 ng (~~the~~ LRs differed by ~36X.) Overall, 100% of the 2-person mixtures were within 10X when interpreted assuming two contributors, while 96.4% were within 10X when interpreted incorrectly assuming three contributors.

The three-person mixture set used for this study was selected specifically because it had a number of low-template amplifications where no more than 4 alleles were detected at any one locus above the analytical threshold. [Note that upon inspection of results below the analytical threshold, it is unlikely that a trained casework analyst would actually interpret these mixtures as resulting from just two contributors.] Having no more than 4 alleles detected allowed an interpretation to proceed with an incorrect assumption of two contributors, and this resulted in full false negatives (LR = 0) for minor contributors in 42% of the comparisons because the comparison reference genotype could not be created using any combination of those alleles. This increased to 92% when focusing on just the contributors comprising 10% of the total template. When interpreted correctly as three contributors, there were no false negative results, although LRs were < 1 in ~6.5% of total comparisons. For those comparisons from the 2-person assumption that were not LR = 0, an incorrect 2-person assumption on average increased the likelihood ratio compared to a correct 3-person assumption by ~2.6X (*i.e.,* 0.41 log units) overall and ~3.9X (*i.e.,* 0.59 log units) for the 10% contributor. Aside from an LR=0, another potential diagnostic of an incorrect donor number assumption for this scenario (3-person mixture with no more than 4 alleles detected at any locus) is a STRmix interpretation indicating evenly distributed mixture proportions across the donors.

Interpreting this mixture set as four contributors had the general effect of flattening the LR with log(LR) trending toward 0 whether the three-contributor interpretation had a log(LR) > 0 or < 0. Comparing average log(LR) values across two STRmix interpretations for each assumption, the median difference between correct and incorrect donor number was less than half a log unit. Inter-interpretation precision was lower when an incorrect donor number assumption was made, but across all four-contributor assumptions at least ~94% of the comparisons had duplicate interpretations in which the LRs were within a factor of 10 (1 log unit).

Part II

The profiles for the three-person mixture used in this study were selected specifically because no more than 4 alleles would be detected at any one locus. Thus, a combination of these profiles could be mistaken for a 2-person mixture if you were to make that determination based solely upon the number of alleles detected. As it turned out, having fewer alleles than would readily indicate the number of contributors was not a challenge for STRmix when interpretations were performed assuming the correct number of donors. A key difference here: the reason for there being "too few" alleles was not that alleles from a donor had fallen below the analytical threshold. Instead, it was simply due to allele stacking/overlap.

As demonstrated in the previous donor number study, donor number assumptions do have an effect on the ability of STRmix to assess a mixture. In this study, 54% of the reference comparisons involved false negatives (LR = 0) when interpreted as a 2-person mixture. This is an increase from the 42% false negative rate in Part I. On the contrary, no LRs were below 1.0 when the mixture was interpreted correctly using a 3-person assumption. In the previous study, there were no false negatives at LR = 0 when interpreted with the correct donor number, but LRs < 1 were observed in ~6.5% of the total comparisons. That difference could in part be due to Part I having used 3130 data and the Part II study using 3500 data. Another possibility is that the additional allele stacking could mean that fewer alleles fell below the analytical threshold. Both the Genetic Analyzer model and the profile combinations could affect whether peaks are detected above or below the analytical threshold. When an assessment of allele count and peak heights leaves ambiguity as to the number of donors, it may be prudent to run the mixture under both the minimum number of contributors (by allele count) and as the minimum + 1. In Part I, adding an extra contributor to a true 2-person mixture led to a median downward log(LR) shift of 1.2 log units, but did not lead to LRs < 1.

Likelihood ratios from replicate analyses were generally within 10X of each other. For this mix of profiles, the proportion outside 10X was slightly higher than in previous studies (~5%, up from ~2% to ~3%). Lastly, no benefit or detriment was observed by using the altered run conditions: 130K total accepts with 80K burn-in accepts.

4.1.6.5. Sharing of alleles among contributors

All mixture studies evaluated included loci with overlapping alleles. In particular, see Part II of Guideline 4.1.6.4 for results on mixtures in which excessive allele sharing was evaluated in 3-person mixtures.

4.1.7. Partial profiles, to include the following:

4.1.7.1. Allele and locus drop-out

Varying template levels of single-source and 2- and 3-person mixture samples were evaluated throughout the validation, including levels that led to allele and locus drop-out.

4.1.7.2. DNA degradation

DNA degradation studies were performed to examine the sensitivity and reproducibility of STRmix when testing especially challenging differential degradation samples. A 2-person 1:1 mixture series was created with one of the two contributors degraded at varying levels. Additionally, a 3-person 6:3:1 mixture was created with the 3-part donor degraded. Duplicate amplifications of these mixtures were performed.

Sensitivity was measured as the proportion of comparisons (including each mixture being compared to each contributor) that gave positive log(LRs). STRmix sensitivity was 100% for the 2-person mixtures, and 96.3% for the 3-person mixture. Overall, STRmix had positive log(LRs) in 99.47% of the comparisons. The remaining 0.53% was attributed to a single false exclusion (LR = 0) for the degraded DNA contributor to the 3-person mixture. In nine total comparisons to this person, the false exclusion occurred in one interpretation of the first of two amplifications. The other two interpretations of the first amplification, all three interpretations of the second amplification, and all three joint interpretations gave positive log(LR) values.

Precision was measured as the proportion of pairwise comparisons, within like-interpretations, that were within one log(LR) unit. STRmix precision was 100% for the 2-person mixtures, and 88.89% for the 3-person mixture. Overall, STRmix had 98.41% of the pairwise comparisons within one log unit. The remaining 1.59% was attributed to the single false exclusion (LR = 0) for the degraded DNA contributor to the 3-person mixture. This one false exclusion was compared to two other interpretations that have positive log(LR) values.

4.1.7.3. Inhibition

The effect of inhibitors on PCR amplification typically manifests as either a degradation-like pattern of decreasing RFU inversely proportional to molecular weight, or as locus-specific decreases in RFU that may not correlate to molecular weight. The degradation-like pattern can be corrected for by the STRmix degradation variable. To correct for locus-specific patterns, STRmix does allow for separate amplification efficiencies for each locus. However, STRmix also limits the extent to which a locus can be above or below expectations. This limitation is done via the LSAE variance and Metropolis-Hastings penalty.

This study examined whether the LSAE adjustment and penalty in STRmix can adequately allow for locus-specific inhibition patterns. Sensitivity and precision studies compared STRmix interpretations of a two-person mixture series amplified with no inhibitor to the same mixtures amplified with an inhibitor (hematin and humic acid were both tested).

STRmix adequately dealt with the inhibition-induced increases in locus amplification variation. The differences between inhibited and non-inhibited results came down to the completeness of the profiles. When inhibition led to fewer alleles detected for a minor contributor, the LR was reduced. This is no different than any other situation where fewer alleles were detected for a contributor.

4.1.8. Allele drop-in

The BFS-validated STRmix V2.0.6 procedure utilized an analytical threshold of 50 RFU for 3130/3130*xl* and 150 RFU for 3500/3500xL, therefore drop-in was not implemented and validated. Per the DNA Technical Procedure, the analyst will enter 0 for the probability of drop-in STRmix parameter.

4.1.9. Forward and reverse stutter

Reverse (N-4) stutter was accounted for throughout the validation in all studies as any detected N-4 peaks for the "evidence" are always imported into STRmix for interpretation.

STRmix V2.0.6 does not consider forward (N+4) stutter, and therefore a workaround was developed. While different approaches were evaluated, such as assuming an extra contributor or implementing drop-in, the best approach was determined to be the use of an assumed "phantom" (artificial) contributor that is assigned the uncertain allelic and/or forward stutter peaks. This approach appeared to best replicate the LRs of profiles comprised of just alleles and reverse stutter. As expected, the LRs for loci with forward stutter peaks did decline when applying the assumed contributor. Those allele positions would be available for possible minor contributor genotype, and the increase in possible genotypes would generally reduce the weight assigned to the true genotypes.

As a result, the Phantom Excel program was developed and validated. This program performs the following functions:

- Confirms that the locus order in the GeneMapper ID-X (GMID-X) exported table is correct for Identifiler Plus, correcting the order where required; the STRmix required locus order in exported data from the GMID-X Genotypes tab is not always maintained.
- Examines the profiles for possible forward stutter using rules adapted for this purpose from the laboratory's current tiered thresholds for 3500/3500xL and another set developed for 3130/3130*xl*.
- Allows the user to select from the list of candidates those samples for which "Phantom" assumed donor profiles will be created. These Phantom profiles will include the possible forward stutter peaks, all other alleles being placeholders ("1" alleles).

The program was used during the validation, which demonstrated its ability to accurately perform these functions.

4.1.10. Intra-locus peak height variation

Intra-locus peak height variation was evaluated throughout the validation by using varying template levels of single-source and 2- and 3-person mixtures.

4.1.11. Inter-locus peak height variation

During the Markov Chain Monte Carlo (MCMC) process, STRmix utilizes a variable ("mass parameter") referred to as Locus-Specific Amplification Efficiency ("LSAE") and represented as A^l. This is a multiplier that will increase (value >1) or decrease (value <1) the heights of all expected peaks at a locus. It could also be described as an offset or scaling factor to allow for inter-locus RFU imbalances due to kit chemistry or locus-specific inhibition. This variable is integral to STRmix functionality and was thus effectively assessed throughout the validation.

See also Guideline 4.1.7.3 addressing data which specifically tested PCR-inhibited samples.

4.1.12. For probabilistic genotyping systems that require in-house parameters to be established, the internal validation tests should be performed using those same parameters. The data set used to establish the parameters should be different from the data set used to validate the software using those parameters.

Laboratory-specific parameters incorporated into STRmix include reverse (N–4) stutter, the detection threshold (*i.e.,* the Genetic Analyzer analytical threshold), drop-in probabilities, and saturation values.

- The detection thresholds were set to the analytical thresholds in the current laboratory technical procedures (50 RFU for the 3130/3130*xl* and 150 RFU for the 3500/3500xL).
- At the current analytical thresholds, drop-in is not expected, and that probability was set to 0.0.
- N–4 stutter was based upon an assessment of 367 profiles. Using only loci that are either homozygous or heterozygous with alleles that are 2 or >4 bases apart, stutter percentages were graphed by repeat number (adjusted to a decimal scale). The linear regression slope and y-intercepts for each locus were entered into STRmix. The same stutter file was applied to both 3130/3130*xl* and 3500/3500xL data.
- Saturation was determined for each instrument through a comparison of the observed height of an allele to the expected height given the height of the N–4 stutter peak and the expected stutter values calculated above. The inflection point at which observed RFUs tended to be lower than expected RFUs was the saturation setting for STRmix. Using the same 367 profiles included in the stutter study, the 3130/3130*xl* saturation value was 7500 RFU. Based upon 232 profiles, no clear inflection point was observed for 3500/3500xL data. A saturation value of 30,000 RFU was selected for further validation and ultimately adopted. This value is noted by ESR as typical for the 3500 instrument.

The Model Maker module in STRmix takes single-source profiles created by the laboratory using a specific amplification kit-instrument model combination and assesses the data for peak variance and locus amplification variance. This process allows the software to later use data relevant to the laboratory's process in the interpretation of single-source and mixed profiles. By establishing these variables in advance, it also speeds the computer analysis time. All validation beyond the initial studies used samples different from the data sets used to establish parameters in Model Maker.

Since Model Maker uses MCMC to establish the variance settings, this study evaluated the following using 172 profiles from the 3130/3130*xl* and 216 profiles from the 3500/3500xL:

- Run-to-run variation in the resulting variance values when using the same data set.
- Variation when using multiple subsets of 100 profiles randomly drawn from one larger data set (3130/3130*xl* only).
- Comparison to the allele variance calculated by ESR from the same data set (3130/3130*xl* only).
- Values for the LSAE variance.

Because Model Maker, in all of its various forms, employs MCMC, some amount of run-to-run variation is to be expected. Using the full data sets, Model Maker in V2.0 gave a range of allele variance values from 3.248 to 3.827 for 3130/3130*xl* data over ten runs and 9.162 to 10.430 for 3500/3500xL data over eleven runs. Using the 100-profile 3130/3130*xl* subsets, ten Model Maker runs gave a range of allele variance values from 3.239 to 3.922, which contains the range for the full set of profiles. All variances captured approximately the same percentage of data points in the "Sanity Check" (97% to 98%).

The allele variance (3.392) created by ESR and used for DOJ's STRmix 3130/3130*xl* validation fits within the range of observed DOJ values within this set. The allele variance values obtained by DOJ V2.0 and ESR V2 are similar (even identical for one DOJ run). The median of the allele variances from the 3500/3500xL runs (9.767) was selected as the default setting for further validation.

Model Maker for STRmix V1.0.7.49 uses a different assessment for LSAE variance than Model Maker for STRmix V2.0. In V1.0.7.49 the variance setting is calculated as the mean of the individual sample variances (from the STRmix results file), while in V2.0 it is the mode of a gamma distribution fitted to the results from the individual sample variances. Although DOJ originally calculated the locus amplification variance using Model Maker V1.0.7.49, it is equivalent to the mean LSAE variance calculated from the ESR V2 Model Maker results (the source of the allele variance setting) based upon the same set of profiles. Using an LSAE variance calculated as the mean of the individual sample variances is acceptable and gives a variance that is always somewhat higher (and therefore more tolerant of variation) than the gamma mode variance. The effect of setting the LSAE variance to the sample mean or the gamma mode was tested using challenging 2-person and 3-person mixtures that included differential degradation. STRmix V2.0 interpretation results were consistent whether using the gamma mode or the sample mean. In an effort to reduce the number of changes between

STRmix versions, it was decided to base the LSAE variances for Identifiler Plus on the sample mean.

Using the full data sets, Model Maker V2.0 gave a range of LSAE mean variance values from 0.01908 to 0.02386 for 3130/3130*xl* data over 10 runs and 0.02268 to 0.0298 for 3500/3500xL data over eleven runs. Using the 100-profile 3130/3130*xl* subsets, ten Model Maker runs gave a range of LSAE variance values from 0.0192 to 0.02674, which contains the range for the full set of profiles. The LSAE variances selected as the default values for the validation were from the Model Maker runs selected for the allele variance: 0.022 for 3130/3130*xl* and 0.0229 for 3500/3500xL.

4.1.13. Sensitivity, specificity and precision, as described for Developmental Validation

2-Person Mixtures

STRmix interpretations from two separate 2-person mixture studies were examined for sensitivity and precision. Both studies included 1 ng template input for the following ratios: 19:1, 9:1, 4:1, 2:1, 1:1, 1:2, 1:4, 1:9, and 1:19. One of these studies was performed with duplicate amplifications, and also included, in duplicate, the following ratios with a 0.5 ng template input: 9:1, 4:1, 1:1, 1:4, and 1:9. All mixtures were interpreted twice in STRmix and those mixtures with duplicate amplifications were also interpreted jointly in STRmix, in duplicate. Sensitivity was measured as the proportion of comparisons that gave positive log(LR) values for true contributors. Not surprisingly, sensitivity and LRs tend to go down with lower amounts of template DNA, when comparing to a minor contributor. Precision was measured as the proportion of pairwise comparisons that were within 2X and 10X of each other.

For sensitivity (and specificity), 100% of data points had positive log(LR) values, and therefore, there were no LRs < 1.0. For a given contributor, the LR was lower when the mixture was 1:1 than when the mixture was 4:1, 2:1, 1:2, or 1:4. In other words, within the range spanning 4:1 to 1:4, STRmix was better able to define a contributor's genotype when the person was a major or minor contributor than when they were an equal-parts contributor. This is because mixtures of equal proportions will result in each contributor having an increased number of possible genotypes, and therefore an increased number of genotype combinations. With regard to minor contributors, in general, unless two minor contributor alleles are detected at a locus, there will be greater ambiguity about the minor donor genotype due to possible overlap with major donor alleles or drop-out. Joint interpretations may improve likelihood ratios as they tended to give higher LRs than interpretations of a single amplification in this study.

In addressing precision, 100% of the replicate interpretations were within a factor of 10 (*i.e.,* 1.0 log unit). In fact, all but two replicate interpretations were inside a factor of 2 (~0.3 log units).

3-Person Mixtures

Sensitivity and precision of STRmix was assessed using 3-person mixtures prepared among four BFS laboratories: Fresno, Riverside, Sacramento, and Richmond. Mixtures tested included ratios 1:1:1, 4.5:4.5:1, 6:3:1, and 8:1:1. Each mixture was amplified in duplicate at three separate template quantities: 1.5, 0.75, and 0.375 ng. STRmix interpretations were performed either in duplicate or triplicate for both each amplification separately and as a joint interpretation. Sensitivity was measured as the proportion of comparisons that gave positive log(LR) values for true contributors. Not surprisingly, sensitivity and LRs tend to go down with lower amounts of template DNA, when comparing to a minor contributor, and when interpreting more even mixtures (*e.g.,* 1:1:1). Precision was measured as the proportion of pairwise comparisons that were in the ranges of 0 – 0.3 and 0 – 1.0 log(LR) units. This corresponds to LRs within 2X and 10X of each other, respectively.

Overall, STRmix had a high degree of sensitivity with positive log(LR) values in 96.59% of the comparisons. Of the 3.41% of negative log(LR) values, 0.55% of comparisons were the result of complete false exclusions (LR = 0). These false exclusions, however, were solely the result of the software exceeding the Java cap on iterations (see Guideline 4.1.6.3). When rerun in a manner that kept the iterations below the cap, all of the LR = 0 comparisons became LR >> 0 comparisons.

Of the remaining 2.86% of comparisons with 0 < LR < 1.0, all occurred with 0.375 ng amplifications, and seemed to be the result poor estimates by STRmix of the mixture proportions likely due to few or no loci where at least 5 alleles (for these 3-person mixtures) were detected, increased stochastic variation (*esp.* for 6:3:1), and/or multiple donor alleles that fell below the analytical threshold. Care should be taken when interpreting such mixtures, especially if most/all of the indicators that they consist of 3 people fall below the analytical threshold. In such cases, jointly interpreting replicate amplifications could prove helpful to correct for this, especially in regard to mixture proportion estimates.

With regard to precision, STRmix had 83.85% within 0.3 log units, and 96.87% of the pairwise comparisons within one log unit. When the pairs with at least one LR = 0 result are removed from consideration, the maximum difference was 2.87 log units, which corresponds to a factor of ~740. The largest deviations occurred in pairs with a minimum log(LR) > 7 (LR > 10 million). LRs of 10 million and 10 billion are likely to lead to the same conclusions about the strength of the evidence. Below this level, deviations ranged up to ~100X, which could possibly lead to moderately different conclusions. The precision results led to the procedural recommendation to perform two interpretations: if the LRs for a sample fall within a factor of 10, the lower LRs will be reported; if they diverge by more than a factor of 10, a third interpretation should be performed.

It should be noted that this study was based upon 3130/3130*xl* data. As discussed in section 4.1.3, the performance of STRmix using 3500/3500xL data appeared to be equal to or better than when using 3130/3130*xl* data, especially in regard to sensitivity.

CODIS Profile Worksheet Functionality in CAL DOJ STRmix Report

The samples tested for specificity under Guideline 3.2.2 were tested here for sensitivity and specificity using a cumulative genotype probability of 0.95. Overall, the CODIS profiles developed by the CAL DOJ STRmix Report were generally efficient at capturing true contributors under a default cumulative genotype probability setting of 0.95. Except for one 2-person mixture, only the differentially degraded samples had any false negative CODIS profiles. However, increasing the cumulative genotype probability to 1.0 allowed the missing profiles to be included. To maximize the inclusion of true contributors, some profiles will need to be set to 1.0.

Moderate stringency random match probabilities (msRMP) were calculated for each CODIS profile per the moderate stringency match rules using the standard African American, U.S. Caucasian, and Southwest Hispanic databases; note that D2S1338 and D19S433 are not included in the statistic under default NDIS settings. When applying the msRMP in a way similar to the CODIS Match Estimator, as expected, many of the CODIS profiles were not searchable due to a high moderate stringency random match probability. This was caused, in part, by loci that could not be searched due to excessive allele counts (above the CODIS per-locus limit) and/or drop-out. The msRMP increase from the addition of the D2 and D19 results would help, but those are currently only useful for California's SDIS.

4.1.14. Additional challenge testing (e.g., the inclusion of non-allelic peaks such as bleed-through and spikes in the typing results)

Non-allelic peaks, such as bleed-through ("pull-up") and spikes must be filtered out in GeneMapper ID-X prior to exporting the data for STRmix. On the occasions that a peak was inadvertently not filtered out, the particular locus resulted in an LR = 0 for at least one contributor to the mixture. The curious result prompted inspection of the data, identification of the inclusion of the non-allelic peak, and re-analysis.

STRmix also requires that all peaks have only a number designation with no non-numerical text such as >, <, or OL. The interpretation of profiles that included an unedited pull-up peak labeled OL or an N–4 stutter peak listed as <8 proceeded to completion but did not include any results for the particular locus or any loci after it.

The import function of CAL DOJ STRmix Report V1.0.xltm was also tested for incorrect file types and entry. Attempts were made to import the following file types: txt-formatted

[…]_Results files from STRmix LR run folders, which don't include genotype probability distributions but do have the same name format as the necessary results files; csv-formatted population frequency files; txt-formatted […]_GenotypePDF files; and non-reference samples with two or more peaks per locus from GMID-X Genotypes Table files. With the exception of evidence peaks that had no more than two alleles per locus, all of the erroneous imports failed. When a GMID-X table includes both evidence and reference profiles, extra caution should be made to ensure that a low-level evidence profile is not entered in place of a reference. True reference files, whether in GMID-X txt format or STRmix csv format, were correctly imported.

4.2. Laboratories with existing interpretation procedures should compare the results of probabilistic genotyping and of manual interpretation of the same data, notwithstanding the fact that probabilistic genotyping is inherently different from and not directly comparable to binary interpretation. The weights of evidence that are generated by these two approaches are based on different assumptions, thresholds and formulae. However, such a comparison should be conducted and evaluated for general consistency.

4.2.1. The laboratory should determine whether the results produced by the probabilistic genotyping software are intuitive and consistent with expectations based on non-probabilistic mixture analysis methods.

4.2.1.1. Generally, known specimens that are included based on non-probabilistic analyses would be expected to also be included based on probabilistic genotyping.

4.2.1.2. For single-source specimens with high quality results, genotypes derived from non-probabilistic analyses of profiles above the stochastic threshold should be in complete concordance with the results of probabilistic methods.

4.2.1.3. Generally, as the analyst's ability to deconvolute a complex mixture decreases, so do the weightings of individual genotypes within a set determined by the software.

Single-source

Eight previously generated sensitivity studies (4 from 3130/3130*xl* and 4 from 3500/3500xL Genetic Analyzers) were run through STRmix. Each study included single amplifications of 2 ng, 1 ng, 500 pg, 250 pg, 125 pg, 62 pg, 31 pg, and 16 pg; note one 3130/3130*xl* study included 10 replicates at both 31 pg and 16 pg.

The summations below illustrate that STRmix gives full profiles to at least the same level of template DNA, and sometimes a dilution below, as compared to the current interpretation procedure. There are a few factors that contribute to this increased sensitivity of the same data. One factor is that during the STRmix procedure development, it was decided to include low-level alleles that have been detected in only one injection if the same allele was visible below the analytical threshold in the duplicate injection. In the current procedure, peaks detected in only one injection are removed from further interpretation as "LLI" (low-level inconclusive).

Sensitivity study	CURRENT BINARY PROCEDURE	STRMIX
	3130/3130*xl* data	
1	250 pg	125 pg
2	0.5 ng	250 pg
3	250 pg	250 pg
4	250 pg	250 pg
	3500/3500xL data	
5	125 pg	125 pg
6	125 pg	125 pg
7	0.5 ng	250 pg
8	250 pg	125 pg

Another factor contributing to this increased sensitivity is the use of a stochastic threshold, or lack thereof. In the current procedure, a peak below the set threshold (365 RFU for 3130/3130*xl* and 1075 RFU for 3500/3500xL) is treated as an allele call rather than a genotype. STRmix, on the other hand, assigns weight to the different genotype possibilities it considers during MCMC and was shown to often assign 100% weight to homozygous genotypes for single peaks just below the current stochastic threshold for each platform. This is not surprising and suggests the current stochastic thresholds are generally conservative.

In comparison to the current manual approach, in no instance did STRmix incorrectly assign 100% weight as a homozygote to a single peak with a dropped-out partner allele nor did it incorrectly genotype a true homozygote as a heterozygote. STRmix rather appeared to appropriately consider the different possibilities, tending toward higher weights assigned to homozygous genotypes for low-level single peaks due to the inclusion of a probability of drop-out; the probability of drop-out is based on the allele variance (from Model Maker) and the analytical threshold.

Of the eight studies, four peaks were detected as off-scale, all being homozygous peaks from a 2 ng DNA template input. These four peaks were genotyped correctly by STRmix.

Sample files were run through STRmix in duplicate. The duplicate LRs were all within 10x for a given template quantity and population. For results in which a complete 15-locus STR profile was obtained (*i.e.,* one genotype per locus, with 100% weight assigned to the genotype), the LRs for a given population were identical. Identical LRs were obtained for the higher quantities tested (*i.e.,* 2 ng through the quantity listed in the STRmix data column above, either 125 or 250 pg).

In comparing the statistical weight calculation between the current manual procedure (1/RMP) and the STRmix procedure (likelihood ratios), most of the higher template quantities were within 10x of one another; one study was within 19x. At the lower quantities (≤ 125 pg), more variation, though not dramatic, was seen. These results were expected due to differences between the two approaches. Overall, the data sets showed a trend of 1/RMP values being slightly larger than LRs for the higher template quantities tested (until drop-out was considered) and slightly lower than LRs for the lower template quantities tested (generally ≤ 125 pg). The differences for the higher templates are likely attributable to the fact that CAL DOJ STRmix Report uses the full Balding and Nichols method while STR10-ID Profile uses theta only; the full Balding and Nichols method will drive down LR values. At the lower templates, the opposite trend is observed due to the introduction of the factors described previously (the difference in handling of low-level peaks between the procedures and the difference in the binary/stochastic threshold versus continuous/probability of drop-out approaches).

2-Person Mixtures

A comparison of five different 2-person mixture studies was done to evaluate the consistency of the interstudy statistics: LRs and 1/random match probability (RMP). These 2-person mixtures were amplified at both 1 ng and 0.5 ng of total DNA (made up of entirely different male and female contributors), run on a 3500 or 3500xL, and consisted of nine male:female mixture ratios (19:1, 9:1, 4:1, 2:1, 1:1, 1:2, 1:4, 1:9, and 1:19). Several of these mixtures were additionally evaluated with an assumed contributor.

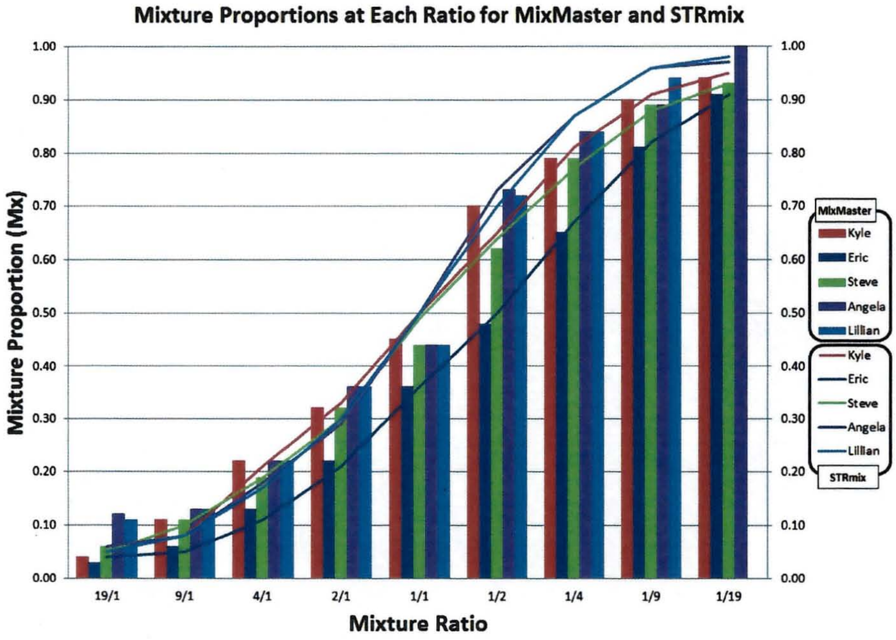
The statistics for both methods demonstrated concordance with expectations. Across all the mixtures, the calculated RMP values rose and fell in tandem with the STRmix-generated LRs. The STRmix LRs were higher than the RMPs when the interpreted profiles were not single-source. However, when the interpreted profiles were single-source, the RMP exceeded the LR. This "improved" RMP statistic for single-source profiles is due to the fact that the current BFS technical procedures for RMP calculations do not fully incorporate theta using the subpopulation correction model of Balding and Nichols as do the new methodologies for LR calculations.

As the female contributor being compared rose from very minor (in the 19:1 mixtures) to very major (in the 1:19 mixtures), the pattern of statistical weight from STRmix mirrored that of MixMaster, the 2-person mixture deconvolution Excel program currently in use in BFS for DNA casework. This is a consequence of the increasing peak heights of the female contributor. These increased peak heights allowed STRmix, as well as the manual method, to reduce the number of included genotypes, thereby increasing the weight assigned to each. Conversely, when the manual deconvolution resulted in more ambiguity and reduced RMP, the STRmix weightings and subsequent LRs were reduced as well.

Across each of the studies at all of the mixture ratios, all known contributors that were included in the MixMaster results were also included in the STRmix results. This is the case for the mixtures run without an assumed contributor (both 1 ng and 0.5 ng amplifications) as well as the 0.5 ng mixtures run with an assumed contributor.

Additionally, for all instances in which a single-source profile was interpreted using MixMaster (with or without an assumed contributor), STRmix was also able to interpret a single-source profile. In all five mixture studies STRmix was able to deduce more single-source profiles than was MixMaster.

As shown in the figure below, comparisons of these studies showed mixture proportions tracked closely between the existing mixture interpretation method (MixMaster) and the probabilistic system (STRmix). Note that the statistics and mixture proportions were evaluated using only the African American statistics for the female contributor as the standard basis for comparison.

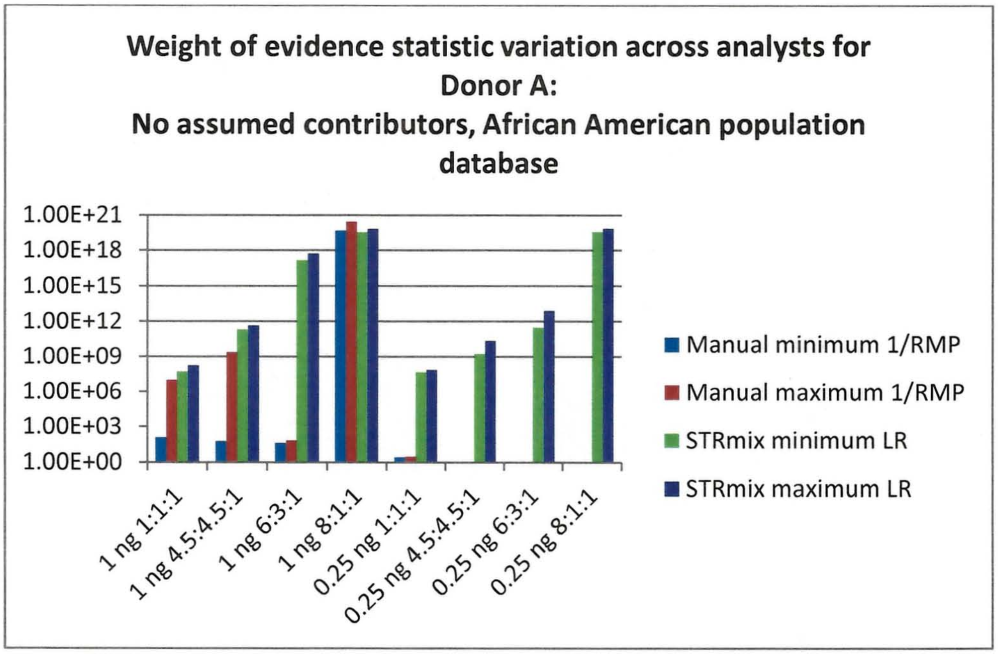


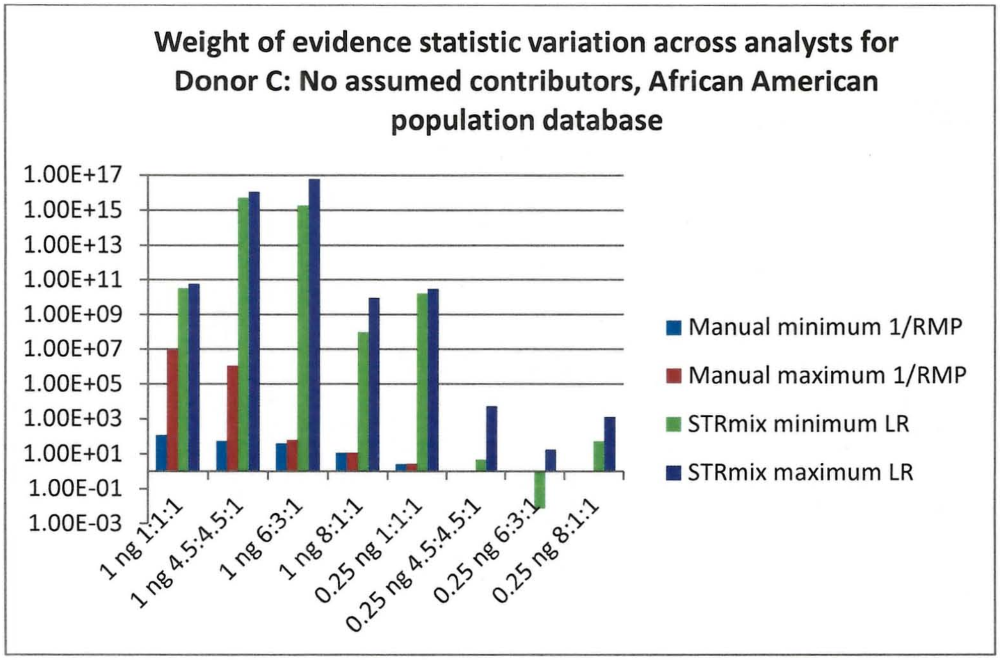
Additional 2-Person Mixture Comparison Data

STRmix results of the 2-person differentially degraded mixtures described under Guideline 4.1.7.2 were additionally compared to MixMaster results. Between the two programs, the known contributors were correctly included with one partial exception. While STRmix sensitivity results were 100%, MixMaster sensitivity was 97.22% when using the standard 4-allele approach to estimating Mx. The remaining 2.78% was attributed to a single false exclusion (LR = 0) at a single locus for the comparison of one reference to one amplification. When MixMaster was rerun using the MxCalculator 3 and 4-allele Mx estimator (also currently in use for DNA casework), this comparison was no longer LR = 0.

3-Person Mixtures

A comparison study was conducted in which five qualified analysts performed parallel interpretations, using the current technical procedures for manual interpretation of three-person mixtures and STRmix, on the same 3-person mixture data. The mixtures were created at varying ratios (1:1:1, 4.5:4.5:1, 6:3:1, and 8:1:1) and amplified in duplicate using 1.0 and 0.25 ng input template. Comparisons were not possible between interpretations using the current technical procedures and STRmix for all ratios and input amounts because interpretation was not possible for some of the contributor ratios at the 0.25 ng input level using the current procedures. At the contributor ratios and input amounts where an interpretation was possible, the determined genotypes for all interpretations using the current procedures included the known (non-assumed) contributors, allowing for calculation of a weight of evidence statistic (1/RMP). However, differences of several orders of magnitude were observed in calculated 1/RMP statistics across the five analysts for some of the interpretations using the current procedures, whereas the STRmix weight of evidence statistics (LRs) did not exhibit the same degree of variation among the analysts. Given an A:B:C donor order, the figures below show the LRs for the A (usually major) and C (usually minor) donors.





Wide variation in 1/RMP was observed in the current method for the 1 ng 1:1:1 and 1 ng 4.5:4.5:1 amplifications due to the analysts' assumptions concerning allele dropout; specifically, an assumption that all alleles had been detected for all three contributors resulted in higher 1/RMP statistics for the 1 ng 1:1:1 amplification, and an assumption that all alleles had been detected for the two higher level contributors of the 1 ng 4.5:4.5:1 amplification resulted in higher 1/RMP statistics. Inspection of the STRmix genotype probability distributions for these mixtures suggests that these assumptions are well-founded, since no genotype combination including a dropped allele (referred to as a "Q" allele in the STRmix report) was assigned any probability at any locus for the contributors whose alleles were assumed to be fully detected. However, the imposition of these assumptions is, to some extent, dependent on analyst judgment instead of quantifiable information. STRmix analysis, by contrast, is minimally dependent on analyst discretion and uses the same "random walk" algorithm to explore genotype possibilities every time it is run, which led to much more consistent, precise results between interpretations performed by the different analysts.

STRmix was capable of performing mixture interpretations for all contributor ratios and input amounts, as well as provide intuitively reasonable LRs for all known contributors. STRmix also generated estimates for useful descriptive parameters, such as mixture proportion, that were in accord with the known properties of the mixtures and exhibited increased variance with decreased input amount in a predictable manner. All STRmix LRs for the known contributors were positive, and were comparable or higher than the corresponding 1/RMP calculated using the current technical procedures, with the exception of the 6:3:1 0.25 ng mixture.

At this 6:3:1 contributor ratio and 0.25 ng input level, with no known contributors assumed, the STRmix LRs for the lowest-level contributor to the mixture were slightly negative for some

STRmix interpretations. However, the corresponding random match probability for the interpretation of this mixture using the current procedures was just over 1 for all ethnic groups, which means that most of the population would be included as a possible contributor to the mixture; a negative LR, which provides support for the proposition that a random person contributed to the mixture, is reasonable in this instance.

Apart from an assessment of STRmix LR precision within a single set of data is a determination of whether STRmix LRs vary from one interpretation to another in a way that is internally consistent and comports with the expectations of an experienced analyst. Within this comparison study, STRmix LRs for the known contributors tracked with separation between the input amount of the contributor in the numerator of the LR and the input amounts of the remaining contributors, unless the overall input levels for the numerator contributor were low and genotypes including dropped alleles predominated the results for that contributor. LRs for known, high-level contributors gradually decreased as contributor input levels became increasingly similar; this relationship is intuitive and in agreement with the concept that STR mixture data involving more ambiguity in the possible genotypes should be given less weight.

The 1/RMP value was comparable to the STRmix LR at the contributor ratio with the greatest input level separation (*i.e.,* 8:1:1) but quickly fell off (in the absence of additional assumptions about allele detection), primarily because mixture proportions are not used to restrict genotypes for 3-person mixtures under the current technical procedures.

In the same way that a valid analysis method should give less weight to more ambiguous evidence, the same method should give more weight to less ambiguous evidence. The 0.25 ng input level comparison between interpretations with no assumed contributor and one assumed contributor demonstrated this principle for both the current technical procedure interpretation method and the STRmix interpretation method. In most cases where interpretation was possible with the current procedures so that a comparison could be made, the addition of information about an assumed contributor increased the weight of evidence statistic for the unassumed contributors, using both methods. The magnitude of the increase tended to be less for the interpretations using the current procedures, mainly because of the same mixture proportion intractability issues that caused the 1/RMP value to fall off quickly as the input level separation decreased with no assumed contributors. Some instances of lower STRmix LRs for an interpretation involving an assumed contributor were observed. In these instances, the assumed contributor and one non-assumed contributor were both at low levels, and the addition of the assumed contributor information limited the possibilities for the low level non-assumed contributor such that possibilities involving dropped "Q" alleles were assigned higher probability than in the interpretation without an assumed contributor.

Modification to Software

SWGDAM probabilistic genotyping validation Guideline 5 states, *"Modification to probabilistic genotyping software shall be addressed in accordance with the QAS."*

5.1. Modification to the system such as a hardware or software upgrade that does not impact interpretation or analysis of the typing results or the statistical analysis shall require a performance check prior to implementation.

5.2. A significant change(s) to the software, defined as that which may impact interpretation or the analytical process, shall require validation prior to implementation.

5.3. Data used during the initial validation may be re-evaluated as a performance check or for subsequent validation assessment. The laboratory must determine the number and type of samples required to establish acceptable performance in consideration of the software modification.

STRmix Version 2.0 versus 2.0.6

A performance check assessing MCMC via the SetSeed Function and likelihood ratio calculations was performed to assess minor programming corrections in STRmix V2.0.6 from V2.0. STRmix V2.0.6 was determined to be identical to V2.0 when performing the MCMC interpretation process and when creating LRs with no assumed contributors.

STRmix Version 1.0.7.49 versus 2.0/ 2.0.6

The initial assessment of STRmix at the California Department of Justice began with V1.0.7.49, progressing through V2.0 to V2.0.6. V2.0.6 is the version intended for our initial casework use. An Excel spreadsheet that was initially created as a means to recreate and test the STRmix LR calculation was subsequently updated with additional features (see Guideline 3.2.6.1 regarding the Likelihood Ratio calculations) and used to calculate the LRs for previous studies, so any LR calculation differences between V2.0 and V2.0.6 are of no consequence here. Direct comparisons of the LRs from V2.0.6 to the LRs from the final version of the Excel spreadsheet ("CAL DOJ STRmix Report V1.0.xltm") are described in Guideline 3.2.6.1.

V2 and V2.0.6 have identical MCMC programming (see the V2.0.6 performance check), though both differ from V1.0.7.49. While key studies have been reanalyzed in full or in part with either V2.0 or V2.0.6, this study looked at the applicability of results from V1.0.7.49 studies to the current V2.0.6. The following sample types were run through STRmix in duplicate, including joint interpretations:

- A 2-person mixture series with no degradation (1 and 0.5 ng inputs);
- A 3-person mixture series with no degradation (1.5, 0.75, and 0.375 ng inputs); and
- Both 2-person and 3-person mixtures with differential degradation.

Comparisons between these two software versions using challenging samples with differential degradation (2-person and 3-person) have also been documented in the studies described under Guideline 4.1.12.

Most of the STRmix validation studies in this lab have been performed on V2.0/2.0.6. For those studies that were performed only on V1.0.7.49, the current study suggests that the information gained from them can be generally applied to V2.0.6. There do appear to be some systematic differences, most clearly detected in the log(LR) shifts in the 3-person differential degradation samples, but those shifts are not unidirectional across all contributors or even within a contributor. Looking across all sets in this study, the versions mostly gave nearly identical results, and neither version consistently gave better (higher LR) or worse (lower LR) results. Therefore, assessments from one version regarding sensitivity, precision, and how well the system determines mixture proportions will generally apply to the other version.

Phantom Versions 1.0 versus 1.2

During the greater STRmix validation, the criteria/settings that the Phantom spreadsheet uses to create Phantom profiles were modified incrementally. (See the Internal Validation introductory information and Guideline 4.1.9 for the purpose of the program and additional details.) As a result, the spreadsheet "Phantom 1.2.xltm" was created to be the final version approved for casework. This study used the same artificial profiles designed to test each of the spreadsheet's functions as were used in the initial validation of Phantom 1.0.0 BETA.xltm.

The following options/changes were added since Phantom 1.0.0 BETA.xltm:

- Peaks that fall in both forward and reverse stutter positions will be identified as possible forward stutter if they satisfy the following conditions:
    - The peak's percentage of the parental allele is > the reverse stutter percentage calculated as *mx + b + 0.035.* (Based upon the data used to create the STRmix stutter file, 3.5% is a supplement above the expected stutter percentage. When combined, the expected plus the supplement resulted in a percentage that exceeded 99% of observed stutter percentages.)
    - The remaining RFU is below the instrument-specific maximum forward stutter RFU (3130/3130*xl* 125 RFU; 3500/3500xL 250 RFU).

  NOTE: Phantom 1.0.0 BETA.xltm evaluated these peaks differently. Previously, peaks that fell in both forward and reverse stutter positions were only identified as possible forward stutter if the reverse stutter was calculated to be >30%. This changed for version 1.1.0 and later.
- A tolerance for multiple profiles having the same sample name was added. However, it's still the case that only one phantom profile can be made per sample name without the previous phantom profile being overwritten. This changed for version 1.1.1 and later.
- Microvariant allele names are rounded to the first decimal place. Without this, the Excel macro's comparison of two peaks might not recognize that one might be stutter. This can

happen because of unforeseen Excel-related micro-differences at higher numbers of decimal places. This modification is included in version 1.1.2 and later. This was the last truly functional change to the spreadsheet. This version or later would have been used for almost all of the training sets performed by the five analysts training in the use of STRmix V2.0.6, including mixture studies from those sets that are included in the validation. (See the Internal Validation introductory information and Guideline 4.2.1.)

- Alerts or text were added to indicate that the genotypes table had been checked/corrected for locus order, and that there were no samples with possible forward stutter if that was the case.
- The "Het Only or Hom/Het?" drop-down menu on the "multiplex" worksheet was locked as "Hom/Het", the previous default value, and the cell color changed to gray.
- The default selection for the Genetic Analyzer option was changed from "31XX" to "35XX."
- The spreadsheet name on the "run it" worksheet was updated and an email address was removed.
- The "multiplex" worksheet was protected.

Visual inspection of the txt files created by the Phantom 1.2 spreadsheet showed locus order was both corrected properly for all samples with out-of-order loci and properly maintained for the others. The pop-up window for Phantom selection worked correctly. The proper file names were listed on the main worksheet of the Phantom spreadsheet following a run. All Phantom CSV files had profiles that matched the expected results; this included the creation of a second Phantom for profiles with >2 possible forward stutter peaks. Therefore, the Phantom 1.2 spreadsheet was found to be working as expected.

FBI Quality Assurance Standard 9.5.5 was addressed by processing NIST standard reference material (SRM) 2391c sample D (a 2-person mixture) through STRmix, Phantom, and CAL DOJ STRmix Report. The correct results were obtained.

STRmix and Probabilistic Genotyping Reference List

Principal references

Curran, J.M., and J.S. Buckleton (2011) "An investigation into the performance of methods for adjusting for sampling uncertainty in DNA likelihood ratio calculations." Forensic Science International: Genetics 5:512-516.

Bright, J., Taylor, D., Curran, J.M., and J.S. Buckleton (2013) "Degradation of forensic DNA profiles." Australian Journal of Forensic Sciences http://dx.doi.org/10.1080/00450618.2013.772235

Gittelson, S., Kalafut, T., Myers, S., Taylor, D., Hicks, T., Taroni, F., Evett, I. W., Bright, J., and

J. Buckleton (2015) "A practical guide for the formulation of propositions in the Bayesian approach to DNA evidence interpretation in an adversarial environment." J Forensic Sci. 2015 Aug 6

Taylor, D., Bright, J., and Buckleton, J. (2013) "The interpretation of single source and mixed DNA profiles." Forensic Science International: Genetics 7:516-528.

Taylor, D. (2014) "Using continuous DNA interpretation methods to revisit likelihood ratio behavior." Forensic Science International: Genetics 11:144-153.

Taylor, D., Bright, J-A., Buckleton, J., and J. Curran (2014) "An illustration of the effect of various sources of uncertainty on DNA likelihood ratio calculations." Forensic Science International: Genetics 11:56-63.

Scientific Working Group on DNA Analysis Methods (June 2015), "SWGDAM Guidelines for the Validation of Probabilistic Genotyping Systems." http://www.swgdam.org/

Institute of Environmental Science and Research (December 2014) STRmix User's Manual, v2.0.

Additional literature - general

Bright, Taylor, Curran, Buckleton. (2014) "Searching mixed DNA profiles directly against profile databases." Forensic Sci Int Genet 9:102-110.

Coble, Bright, Buckleton, Curran. (2015) "Uncertainty in the number of contributors in the proposed new CODIS set." Forensic Sci Int Genet 19:207-211.

Taylor D, Bright JA, Buckleton J. (2014) “The 'factor of two' issue in mixed DNA profiles.” J Theor Biol. Dec 21;363:300-6.

Curran JM, Buckleton J. (2014) “Uncertainty in the number of contributors for the European Standard Set of loci.” Forensic Sci Int Genet. Jul;11:205-6.

Bille T, Bright JA, Buckleton J. (2013) “Application of random match probability calculations to mixed STR profiles.” J Forensic Sci. Mar;58(2):474-85.

Curran JM, Buckleton JS. (2011) “An investigation into the performance of methods for adjusting for sampling uncertainty in DNA likelihood ratio calculations.” Forensic Sci Int Genet. Nov;5(5):512-6.

Walsh SJ, Curran JM, Buckleton JS. (2010) “Modeling forensic DNA database performance.” J Forensic Sci. Sep;55(5):1174-83.

Curran JM, Buckleton J. (2010) “Inclusion probabilities and dropout.” J Forensic Sci. Sep;55(5):1171-3.

Gill P, Buckleton J. (2010) “Commentary on: Budowle B, Onorato AJ, Callaghan TF, Della Manna A, Gross AM, Guerrieri RA, Luttman JC, McClure DL. Mixture interpretation: defining the relevant features for guidelines for the assessment of mixed DNA profiles in forensic casework. J Forensic Sci 2009;54(4):810-21.” J Forensic Sci. Jan;55(1):265-8.

Bille TW, Weitz SM, Coble MD, Buckleton J, Bright JA. (2014) “Comparison of the performance of different models for the interpretation of low level mixed DNA profiles.” Electrophoresis. Nov;35(21-22):3125-33.

Lauc G, Dzijan S, Marjanovic D, Walsh S, Curran J, Buckleton J. (2008) “Empirical support for the reliability of DNA interpretation in Croatia.” Forensic Sci Int Genet. Dec;3(1):50-3.

Walsh SJ, Buckleton JS. (2007) “Autosomal microsatellite allele frequencies for a nationwide dataset from the Australian Caucasian sub-population.” Forensic Sci Int. May 24;168(2-3).

Buckleton JS, Curran JM, Gill P. (2007) “Towards understanding the effect of uncertainty in the number of contributors to DNA stains.” Forensic Sci Int Genet. Mar;1(1):20-8.

Buckleton JS, Curran JM, Walsh SJ. (2006) “How reliable is the sub-population model in DNA testimony?” Forensic Sci Int. Mar 10;157(2-3):144-8.

Curran JM, Buckleton JS, Triggs CM. (2003) “What is the magnitude of the subpopulation effect?” Forensic Sci Int. Jul 29;135(1):1-8.

Additional literature - TrueAllele

Perlin MW, Hornyak JM, Sugimoto G, Miller KW. (2015) "TrueAllele(®) Genotype Identification on DNA Mixtures Containing up to Five Unknown Contributors." J Forensic Sci. Jul;60(4):857-68.

Perlin MW, Dormer K, Hornyak J, Schiermeier-Wood L, Greenspoon S. (2014) "TrueAllele casework on Virginia DNA mixture evidence: computer and manual interpretation in 72 reported criminal cases." PLoS One. Mar 25;9(3).

Perlin MW, Belrose JL, Duceman BW. (2013) "New York State TrueAllele ® casework validation study." J Forensic Sci. Nov;58(6):1458-66.

Perlin MW, Legler MM, Spencer CE, Smith JL, Allan WP, Belrose JL, Duceman BW. (2011) "Validating TrueAllele® DNA mixture interpretation." J Forensic Sci. Nov;56(6):1430-47.

Pálsson B, Pálsson F, Perlin M, Gudbjartsson H, Stefánsson K, Gulcher J. (1999) "Using quality measures to facilitate allele calling in high-throughput genotyping." Genome Res. Oct;9(10):1002-12.