SCIENTIFIC WORKING GROUP ON DNA ANALYSIS METHODS[1]

Internal Validation of Fully Continuous Probabilistic Genotyping Systems Module

Short Title: ***Internal PGS Validation Module***

Effective XXXXXXX, XX, XXXX

Scope

The SWGDAM Internal Validation of Fully Continuous Probabilistic Genotyping Systems Module contains minimum requirements and best practice guidelines to assist laboratories in designing internal validation experiments as required by the *FBI's Quality Assurance Standards for Forensic DNA Testing Laboratories (Forensic QAS)* Standard 8.8. A probabilistic genotyping system (PGS) is comprised of software, or software and hardware, with analytical and statistical functions that utilize formulae, models, and algorithms to analyze DNA single source and mixture profiles. The probabilistic genotyping process consists of inferring genotype weights using algorithms, such as the Markov chain Monte Carlo (MCMC) sampling method and using those weighted genotypes to calculate likelihood ratios (LRs) assuming one or more pairs of propositions regarding the donors to the forensic profile. Internal validation shall be conducted using samples of varying quantity, quality, and type (single source and mixtures) encountered in forensic casework to assess the reliability and determine the potential limitations of the PGS. Laboratories may have varied validation experimental approaches that differ from those listed in

[1] The Scientific Working Group on DNA Analysis (SWGDAM; see SWGDAM.org) is comprised of forensic science practitioners and other experts who represent government laboratories within the U.S and Canada, as well as intra- and international professional groups and academia. SWGDAM recommends to the FBI Director revisions to the *Quality Assurance Standards for Forensic DNA Testing Laboratories* and the *Quality Assurance Standards for DNA Databasing Laboratories (QAS)*. SWGDAM provides a forum for its members and invited guests to discuss research, technologies, techniques, and training; and conduct or recommend studies to develop, test, and validate methods for use by forensic laboratories. SWGDAM's Guidelines and Recommendations represent best practices within the discipline. The term "should" is used herein to indicate good practices identified by SWGDAM. "Must" distinguishes mandatory elements, which may be specified in the Quality Assurance Standards for Forensic DNA Testing Laboratories and/or Quality Assurance Standards for DNA Databasing Laboratories.

24 this document; such approaches can be utilized if they generate enough empirical data to
25 determine the capabilities and limitations of the system and support the laboratory's standard
26 operating procedures.
27
28

Table of Contents

Introduction 3
General Considerations 4
Defining System Parameters 7
Functional Testing 12
Reliability Testing 15
Accuracy Testing 19
Sensitivity 20
Specificity 22
Precision 23
Additional Testing 25
Additional Features 26
Additional Comments 27
Appendix A: Example Validation Studies 28
References 41
Overview Document (www.swgdam.org)
[Insert link to document before publication]

61

DRAFT

62 **Key Concepts:**

63 ❖ Probabilistic approaches can provide statistical weighting to the potential genotype
64 combinations unlike legacy approaches which consider all combinations equally
65 probable.
66 ❖ Each laboratory evaluating a PGS must determine which validation studies are relevant to
67 demonstrate the fitness for their use and identify potential limitations of the software.
68 ❖ Validation studies cannot account for all scenarios that may arise during casework
69 examinations; however, laboratories should attempt to cover the range of variation
70 expected to be encountered with forensic samples.
71

72 **Introduction**

A probabilistic genotyping system (PGS) is commonly employed as a software tool to assist the DNA analyst in the interpretation of forensic DNA typing results. In accordance with the *FBI's Quality Assurance Standards for Forensic DNA Testing Laboratories* (QAS), it is not intended to replace the human evaluation of the results or the human review of the results prior to reporting like an Expert System does. The analyst will need to employ manual analysis, as necessitated by the software, before employing the PGS. For example, the analyst may be required to estimate and use a specific number of contributors (NOC) to run a PGS analysis (including deconvolution and statistical calculations), or to assess whether typing results should be interpreted or not based on the quality of the data.

Probabilistic genotyping reduces subjectivity in the analysis of DNA typing results and is particularly useful for low-level DNA samples (i.e., those in which the quantity of DNA for one or more contributors is such that stochastic effects may be observed) and complex mixtures (i.e., mixture profiles that may exhibit evidence of three or more contributors, degradation, and/or stochastic variation). Many legacy statistical approaches applied to mixtures, such as a combined probability of inclusion, may consider all interpreted genotype combinations to be equally probable, whereas a probabilistic approach can provide a statistical weighting to the potential genotype combinations (e.g., based on allele/peak intensity). Probabilistic genotyping does not utilize a stochastic threshold but instead accounts for the possibility that an allele may have dropped out or conversely that an allele may have dropped in. In making greater use of the DNA profile information and eventually comparing potential DNA contributors, probabilistic genotyping improves the ability to distinguish true contributors and non-contributors compared to legacy (binary) interpretation methods.

To ensure optimal PGS performance, the laboratory should verify all hardware and software specifications have been met prior to beginning validation studies. Laboratories should also be aware of the features and limitations of the PGS they are implementing, and their impact on the validation process. Depending on the models applied by the software, prerequisite studies may be required to, for example, establish parameters for allele drop-out, drop-in, or stutter expectations. Each laboratory evaluating a PGS must determine which validation studies are relevant to demonstrate the fitness for their use and identify potential limitations of the software.

Laboratories should use samples amplified and subjected to capillary electrophoresis using all internally validated methods and conditions expected to be applied during casework. This will allow laboratories to assess how the validated amplification chemistries and instruments, as well as the genetic analyzers, will inform the PGS. It is recommended that laboratories use samples whose quantity and quality cover the range of variability encountered by the laboratory in casework and have been recently quantified using current laboratory operating procedures.

There are two main approaches to probabilistic genotyping and calculating LRs: the semi-continuous method and fully continuous method.

- The semi-continuous method focuses on the alleles present or absent in the profile and considers all possible genotype combinations of the observed alleles equally, in conjunction with probabilities of drop-out and drop-in. Analysis parameters such as peak height variation, mixture ratios and stutter percentages are not typically utilized by semi-

DRAFT

continuous software systems, although these elements may be considered during the initial manual evaluation of the data.

- The fully continuous method generally utilizes information such as peak heights, stutter percentages and peak height ratios as well as probabilities of drop-in and drop-out. The weighting of genotype combinations as more or less probable may be inferred from the data through methods such as MCMC samplings from probability distributions.

PGSs model data in varying ways. One PGS may require laboratory-specific calibration to create more informative prior data distributions while another may use generic prior distributions, making such calibration unnecessary. As a result, some studies may be necessary for one PGS but not for another.

These guidelines address studies required for parameter setting and the validation of a fully continuous PGS in conjunction with a multiplex autosomal STR typing kit and may be suitable for evaluating modifications to existing laboratory operating procedures. Additionally, some studies described herein may also be suitable for evaluating a semi-continuous PGS.

Validation studies cannot account for all scenarios that may arise during casework examinations; however, laboratories should attempt to cover the range of variation expected to be encountered with forensic samples. Following implementation, laboratories should review results and if necessary, conduct supplemental studies to improve workflow, thresholds and/or interpretations.

This module should be used in conjunction with the SWGDAM Validation Guidelines for DNA Analysis Methods: Overview Document (https://www.swgdam.org/publications). The studies herein are not synchronized to the Forensic QAS; instead, they are presented in a suggested order to streamline testing and conserve resources such as time, reagents, samples, and consumables. Both documents can be referred to for general background information regarding validation and definition of terms. Materials provided by the PGS developer can also be used as a resource for validation, training, and application. Example validation studies are provided in Appendix A.

GENERAL CONSIDERATIONS

1. Overarching PGS Validation Concepts

1.1. To identify aspects of the PGS that should be evaluated through validation studies, the laboratory should document, or have access to documentation, that explains how the software performs its operations, to include:

1.1.1. the methods of analysis and statistical formulae

1.1.2. the data to be entered in the system

1.1.3. the operations performed by each portion of the user interface

1.1.4. the workflow of the system; and

DRAFT

1.1.5. the system reports, diagnostic values, or other results.

1.2. The samples selected for validation (both single source and mixtures) should cover a range of characteristics that are representative of those typically encountered by the testing laboratory. Data should be selected to challenge the system's capabilities and identify any limitations. The same data (or a subset thereof) may be used for multiple studies to test the different applications of the PGS; however, software developer instructions may be more specific about data sharing during validation/customization.

2. System control – Installation and setup

2.1. The laboratory should verify that the computers running the software meet or exceed the recommended specifications (e.g., sufficient RAM, at least the minimum number and type of processors, and appropriate operating system), that the PGS has been properly installed, and that the initial software configurations are correct.

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

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.

2.2.2. Security protection should be used to ensure only authorized users can access the software and data.

2.2.3. Audit trails are available to track changes to system data and/or verification of system settings applied each time an analysis or comparison is run.

2.2.4. User-level security should be employed to ensure that system users only perform authorized actions. For instance, access to alter validated analytical parameters should be restricted to approved personnel, if possible.

3. Internal validation

3.1. Internal validation of a PGS is the accumulation and assessment 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. In a multi-laboratory system, however, functional and reliability testing should be conducted at each site to ensure the software operates as expected. If conducted within the same laboratory, developmental validation studies may satisfy some of the elements of the internal validation guidelines.

3.2. To identify data features (e.g., minimum quality requirements, NOC) that render a profile suitable or unsuitable for probabilistic genotyping, the laboratory should test data

DRAFT

across a range of characteristics that are representative of those typically encountered by the testing laboratory. Data should be selected to challenge the system's capabilities and identify limitations.

3.2.1. It is critical to include challenging single source samples and complex mixtures with low-level contributors during internal validation, as the data from such samples should be evaluated thoroughly and generally help to define the software's limitations and inform laboratory protocol. Sample and/or data types which may not be suitable for PGS analysis should also be included.

3.2.2. While some manual comparisons resulting in exclusion of a reference sample may be evident without the use of a PGS, these types of comparisons should still be included in validation testing to verify that the software performs as expected.

DEFINING SYSTEM PARAMETERS (PREREQUISITE STUDIES AS NECESSARY)

4. System Parameters – Prerequisite Studies May be Necessary

4.1 Depending on software requirements, prerequisite studies may be required to inform the modeling parameters of the PGS software. If a laboratory uses more than one genetic analyzer, consideration should be given to sensitivity variation between instruments prior to defining PGS parameters. Modeling parameters may need to be established for each amplification kit, platform, and/or set of variable DNA typing conditions utilized by the laboratory. For example, any variations in the amplification, post-amplification, and/or electrophoresis procedures used by the laboratory to increase or decrease the detection of alleles and/or artifacts must be evaluated. These parameters may include:

4.1.1. Analytical threshold (AT)

4.1.2. Allele drop-in expectations

4.1.3. Stutter expectations

4.1.4. Saturation limit (dynamic range) of the genetic analyzer; or

4.1.5. Allelic peak height variation (including allelic drop-out).

4.1.5.1. Intra-locus peak height

4.1.5.2. Inter-locus peak height

4.2 For PGSs that require laboratory-specific modeling parameters to be established, internal validation studies should be performed using the established parameters. *The dataset used to establish the lab-specific parameters must be different from the dataset used to validate the software.* The samples used to define the lab-specific parameters and to conduct validation testing must be generated using identical amplification and capillary

DRAFT

electrophoresis conditions but may be comprised of previously generated laboratory data (e.g., data used for a multiplex kit validation study).

5. *Analytical Threshold*

5.1. The Analytical Threshold (AT) is defined as the value where a true DNA signal can be reliably distinguished from instrument noise. This parameter is typically determined during validation of the associated amplification kit and genetic analyzer. Depending on the PGS used, this value(s) or another value(s) recommended by the software developer is used as the lower bound below which a DNA signal is not considered.

6. *Allele drop-in*

6.1. *Study purpose*

6.1.1. Allele drop-in is typically defined as a non-reproducible observance of a peak within a profile. Most often it is observed as the presence of one or two low-level peaks within a profile, where it is not possible to establish a source for the alleles observed.

6.1.2. Depending on the PGS being used, the presence of allele drop-in within an evidence sample can be assigned a probability as a lab-defined parameter. This probability is factored into the modeling of the PGS when considering potential genotypes of the contributor(s) to a sample. A drop-in peak height (RFU) maximum can also be determined, and if implemented, peaks detected above this threshold would not be considered as potential drop-in to preserve modeling and software run times.

6.2. *Study considerations*

6.2.1. The laboratory should monitor the detection of drop-in peaks with validation and previous casework negative controls (including extraction reagent blanks). Drop-in can also be assessed with positive controls to supplement the number of samples although drop-in may be masked by allelic peaks.

6.2.2. Depending on the PGS and the amount of drop-in observed, the laboratory may choose to model the probability of drop-in as a frequency or as a distribution correlated to peak height.

6.3. *Study outcome*

6.3.1. A drop-in frequency is calculated using the number of drop-in peaks observed divided by the total number of autosomal loci (number of autosomal loci multiplied by the number of samples) evaluated.

DRAFT

6.3.2. Based on the number of drop-in peaks observed, and if the PGS allows for it, a drop-in frequency may alternatively be modeled as a distribution correlated with peak height. A drop-in peak height (RFU) maximum can also be determined based on the data.

7. *Stutter expectations*

7.1. *Study purpose*

7.1.1. Expected stutter ratio values should be determined for each type of stutter that will be modeled by the chosen PGS for each locus within the laboratory's amplification kit. Stutter ratios may need to be determined per locus and/or per allele, depending on the PGS software. Similar to drop-in, a maximum stutter ratio can also be estimated to assist with modeling and software run times.

7.2. *Study considerations*

7.2.1. If customizing expected stutter ratios for the PGS, it should be determined which types of stutters are most often observed within the multiplex kit. Stutter may include one repeat shorter ("back") or longer than the parent allele ("forward"), two repeats shorter than the parent allele ("double-back"), or other variants such as two base pairs shorter than the parent allele ("half-back"). This may vary from locus to locus.

7.2.2. It may be necessary to lower the analytical threshold to increase detection of typical stutter types within a multiplex kit. Donors should be sought out for this study to maximize coverage of alleles and stutter types.

7.2.3. When there is ambiguity in determining the presence of a stutter peak at a particular locus such as alleles augmented by stutter (ex. a heterozygote allele that is in the back stutter position of the sister allele, ex. 16, 17) and stutter bracketed by two alleles (e.g., stutter peak 16, for a heterozygote pair of 15, 17) the data at that locus should be removed from the study.

7.3. *Study outcome*

7.3.1. Once typical stutter ratio values are captured empirically, the data is examined to determine the best fit regression for per-allele estimates of stutter.

7.3.2. Linear stutter ratio regressions created for each locus (by allele and/or the LUS) are evaluated to determine the most appropriate model for each locus for use within the PGS.

7.3.2.1. Outliers from the regression lines should be investigated to determine if they should remain within the dataset or if additional amplifications and/or donors are needed.

DRAFT

7.3.3. If neither regression by allele nor LUS appears to be a good fit for modeling the stutter at a particular locus, average observed values for stutter ratios per allele can be used where data (e.g., at least five observations) are obtained.

7.3.4. A stutter ratio maximum can also be determined based on the data.

7.3.4.1. For example, the laboratory can apply the maximum stutter values observed within the dataset or a stutter value that includes a selected percentage of the values in the dataset.

8. Saturation limit

8.1. ***Study purpose***

8.1.1. Depending on the PGS, the laboratory may decide to evaluate and set a saturation level for PGS lower than that of the genetic analyzer.

8.1.1.1. If interpreting a profile with off-scale data, then the saturation threshold may be an important parameter within the PGS to determine when the height of a peak, and any relative amount of corresponding stutter product, is accurately represented or when it reaches a height at which it is no longer accurately measured.

8.2. ***Study considerations***

8.2.1. The saturation limit is typically instrument/platform-specific determined by the manufacturer and the instrument hardware and software.

8.2.2. For a PGS that uses peak heights to model the genotypes and/or stutter ratios of the contributor(s) to the sample, it is important to assess the point at which the instrument is no longer able to accurately assign peak heights due to saturation. The actual saturation point for an instrument should be assessed to determine if it is lower than what is automatically flagged by the instrument software or conversely, to confirm that the instrument software is properly flagging the off-scale data.

8.3. ***Study outcome***

8.3.1. The data collected from this study can be used to define an approximate saturation point for the genetic analyzer by direct comparison of the observed allelic peak heights to the expected allelic peak heights. Expected allelic peak heights can be determined based on the observed back stutter peak heights in combination with expected average back stutter ratios.

DRAFT

8.3.2. The peak height value (RFU) at which the observed peak heights of the allelic peaks begin to diverge from a linear correlation with the expected allelic peak heights represents the approximate point at which quantitation becomes inaccurate.

8.3.3. This value can also be used to provide guidance for when saturated peak(s) are encountered within casework data.

8.3.3.1. For example, the laboratory may determine that if a DNA profile contains saturated peaks, it may need to be reanalyzed using a dilution or alternate typing condition before interpretation using the PGS. Alternatively, the PGS may incorporate the saturation value into the modeling of potential stutter peaks.

9. *Allelic and stutter peak height variation (including intra-locus peak height and inter-locus peak height)*

9.1. *Study purpose*

9.1.1. Peak heights are inherently variable, both at the intra-locus and inter-locus levels.

9.1.1.1. Depending on the PGS, a range of profiles can be used to model the allelic and stutter peak height variability observed within the laboratory. This variability parameter is then applied as prior expectations within the PGS during the deconvolution process.

9.2. *Study considerations*

9.2.1. Single source profiles from known donors can be used to determine the peak height variability, allelic drop-out, and heterozygote peak height ratio variation observed within a laboratory.

9.2.1.1. The study should include amplification of samples from multiple donors with varying DNA template amounts. The DNA profiles should encompass the range of profiles likely to be encountered in casework, from low-level partial profiles to full profiles approaching the previously determined saturation point. Profiles generated from DNA template amounts below the laboratory's minimum input amount, if applicable, should be included within the dataset, since data from this study will be applied to mixed DNA profiles which may contain contributors at these levels.

9.3. *Study outcome*

9.3.1. The PGS parameter setting module will generate allele and stutter peak variance distributions which will inform the modeling of the peak height variation observed in a sample.

DRAFT

9.3.2. A check for the reasonableness of the intra-locus peak height variance constant can be undertaken by comparison with the heterozygote balance values from the data and evaluating whether sufficient coverage is obtained over a range of expected allelic peak heights.

9.3.3. The variance parameters can be further assessed by determining whether there is any correlation between the observed or expected peak heights of the high and low molecular weight alleles within the heterozygote loci included in these studies.

9.3.4. These variance parameters are applied during the validation experiments and should be monitored throughout the studies to determine if they are appropriately covering the range of data that will be tested by the laboratory.

INTERNAL VALIDATION TESTING

10. Functional Testing

10.1. *Study purpose*

10.1.1. The functional testing of software with DNA profile interpretation and statistical calculation capabilities demonstrates that the software performs the intended tasks and functions as expected. PGSs often require:

10.1.1.1. the import or transcription of data from other software programs

10.1.1.2. the application of customized parameter values that inform the modeling used and calculations produced (PGS specific)

10.1.1.3. separate DNA profile modeling and the statistical calculation as two distinct functions and

10.1.1.4. production of a report detailing diagnostic indicators of the model's success and results.

10.1.2. When performing an initial validation of a PGS or evaluating a major software upgrade, each of the functions should be assessed to gain general familiarity with the software, its workflow, and any software-specific characteristics or behaviors prior to proceeding with further testing.

10.1.2.1. If a previously validated software has undergone a minor revision(s), the same functions should be assessed to demonstrate and ensure that the modification has not unintentionally affected the software's ability to perform the intended tasks.

10.2. *Study considerations*

DRAFT

10.2.1. DNA profiles used in functional testing should not be overly complex in nature to allow the formation of well-founded, manually derived expectations for comparison.

10.2.2. DNA profiles used during functional testing should be single source and mixtures of known origin, with a corresponding known reference profile, or profiles, for comparison.

10.2.3. A functional testing study may include an assessment of multiple software elements. Some aspects of functional testing which should be explored include:

10.2.3.1. Data import:

10.2.3.1.1. The format and file type required by the PGS must be determined and used for data import. Special considerations regarding the formatting of allele calls should be taken into consideration, if applicable.

10.2.3.1.2. A PGS may require the import of stutter data which may need to be replaced after being filtered out during traditional data analysis.

10.2.3.1.3. Some PGSs cannot model for a tri-allelic locus, microvariants not represented in the allelic ladder or alleles without a numerical designation (e.g., alleles labeled with “>” or “<”) which cannot be assigned a numerical designation. Such loci must be omitted from the analysis.

10.2.3.1.3.1. The means by which data are omitted from an analysis should be explored to determine whether the modeling is affected by removal of data prior to import into the system, or whether the system provides a mechanism for the removal of data post-import.

10.2.3.2. Maintenance and application of selected parameters and settings:

10.2.3.2.1. Functional testing of a PGS should determine which parameters and settings are automatically retained from one analysis to the next and which must be selected with each analysis.

10.2.3.2.1.1. This provides an understanding of which parameters and settings are global in nature and will be applied to each analysis conducted by the laboratory and which will vary from analysis to analysis as determined by validation and policy.

10.2.3.3. General profile deconvolution:

DRAFT

10.2.3.3.1. The deconvolution of DNA profiles of known origin and make-up, further characterized by manually derived expectations, serves as an initial evaluation of the parameters and settings established to inform the PGS and as a preliminary evaluation of the modeling.

10.2.3.3.1.1. The results of the software's modeling should be assessed to determine how closely the modeling fits the previously established expectations from the manual interpretation and known make-up of the DNA profile.

10.2.3.4. Assignment of LRs:

10.2.3.4.1. At least one known contributor and one known non-contributor should be compared to each single source and mixture profile analyzed during functional testing.

10.2.3.4.1.1. In general, the LR of the known contributor should be reflective of their contribution to the sample. Each non-contributor should result in an exclusion (LR of 0) or favor the alternate hypothesis (LR less than1).

10.2.3.4.1.2. A PGS may produce more than one LR (sub-sub-source, sub-source, etc.); if so, each LR produced should be reflective of the contribution, or lack thereof, of a known reference sample to the profile.

10.2.3.5. Evaluation of system report/results:

10.2.3.5.1. Result reports from the PGS (e.g., whether printed, electronic, or a combination of both) should be generated to ensure they are able to be successfully produced in a format that can be maintained in a case file.

10.2.4. Reports should also be evaluated to ensure their completeness to include all expected sections, diagnostic indicators, and results of the modeling. These elements do not require separate studies unless an element is not performing as expected and necessitates detailed exploration.

10.3. ***Study outcome***

10.3.1. Overall, functional testing:

10.3.1.1. provides a foundational assessment of a PGS's utility

10.3.1.2. explores procedures for the basic operation of the software

DRAFT

10.3.1.3. establishes familiarity with the operation of the software; and

10.3.1.4. may identify areas of potential further evaluation.

10.3.2. Functional testing does not rigorously challenge the software or demonstrate its limits, but rather serves to demonstrate that the PGS is operational and capable of further testing.

10.3.2.1. For laboratories validating a new version of a previously validated PGS, functional testing may be conducted simultaneously with other studies.

10.3.2.1.1. The way functional testing is evaluated, and the results of such testing must be documented.

10.3.3. Functional testing should aid the laboratory in defining file types and the format in which the data must be imported.

10.3.3.1. Software-specific data formatting requirements should be documented for inclusion in the standard operating procedure.

10.3.4. Maintenance of parameters and settings should also be documented during functional testing to ensure ongoing consistency of application during validation testing and casework use.

10.3.4.1. Subsequent validation studies may necessitate that a parameter be purposefully changed to study the effect.

10.3.5. The results of deconvolutions conducted during functional testing should mirror the expectations formed by manual interpretation prior to analysis using the software.

10.3.5.1. Any analyses demonstrating divergence from reasonable expectation should be investigated, with causes and resolutions thoroughly documented.

10.3.6. LRs produced from comparisons to known references during functional testing should, as with deconvolutions, be reflective of the expected contributions of the known individuals present in the profile.

10.3.6.1. Any LRs demonstrating divergence from reasonable expectations should also be investigated, with causes and resolutions documented.

10.3.7. Familiarity with the organization and content of the report or other system results should be gained during the functional testing.

*11. **Reliability Testing*** (“physical” reliability and “usability limits”)

DRAFT

DRAFT

11.1. *Study purpose*

11.1.1. Reliability testing is the process of testing a software program beyond its functional aspects to ensure it works appropriately in the laboratory environment and is specific to security, software communication and data transfer, stability of settings and load testing. Reliability testing of a software program or system such as a PGS requires that the software/system be technically evaluated to ensure it operates according to expectations in the virtual and physical environment (network) in which it resides and is used.

11.1.1.1. At a minimum, this may include physically testing multi-user or multi-site scenarios, direct-access and network/server-access scenarios, and interaction with other software programs. It may be useful to have a network administrator available to assist with this testing.

11.1.1.2. For example, a PGS is installed on a computer network; however, due to firewall settings on the individual user computer, the user is unable to access the software. As a result, permission and settings modification was needed from the network administrator to bypass the firewall to be able to run the software.

11.1.2. Reliability testing should also assess the usability limits of the PGS's functions. In this context, "usability limits", or operational limits, are considered defining conditions which cause a failure of the software to function for the user and may be indicated by instances such as the receipt of an error, failure of the analysis to proceed, the inadvertent loss of data on import, or the unexpected closure of the software during analysis.

11.1.2.1. Assessing the usability limits of a PGS may be achieved by both targeted approaches as well as through conducting other studies.

11.1.2.1.1. Some potential usability limits of PGS include, but are not limited to, the NOC which can be analyzed, DNA profiles exhibiting excessive dropout, or a contributor assessment where the number of alleles cannot be explained by the NOC input into the software (e.g., NOC set at 2 with 5+ alleles at a locus). This should not be confused with the limits of DNA profile interpretation or modeling indicators but are rather limits where the software will no longer operate.

11.1.2.1.2. For instance, failure to complete the analysis of a six-contributor mixture due to insufficient computer memory represents a usability limit of the software.

11.1.2.1.3. In addition, some PGSs will not proceed if an off-ladder (OL) allele call has not been assigned a numerical value in the import file.

665 11.1.3. Some reliability testing reveals physical limitations of the PGS, whereas other
666 studies may support a laboratory's decision to limit analyses in the software
667 before reaching the usability limit.
668
669 11.2. ***Study considerations***
670
671 11.2.1. When assessing the physical (and virtual) reliability and usability limits of a PGS,
672 the following should be considered:
673
674 11.2.1.1. The virtual environment in which the software resides should be
675 evaluated.
676
677 11.2.1.1.1. A PGS may be equipped to reside on a shared network or may not be
678 compatible with a shared network due to security, access, or system
679 specification considerations.
680
681 11.2.1.1.2. The operating system or server on which the PGS will reside must be
682 compatible with the functionality of the software.
683
684 11.2.1.2. When a PGS is intended for use in multiple locations and/or by multiple
685 users, the effect of utilizing the software in these conditions should be
686 assessed to identify potential issues such as overwriting, limits to access,
687 and licensing requirements and needs.
688
689 11.2.1.3. The security of parameters and settings, and their potential for alteration,
690 should also be assessed.
691
692 11.2.1.3.1. Developer-recommended and/or validation-derived settings should be
693 evaluated to understand the level of access required to change them
694 and ensure they are maintained from one analysis to another.
695
696 11.2.1.3.1.1. Testing should assess how to identify whether a parameter or
697 setting has been changed from a previously defined value.
698
699 11.2.1.3.2. A PGS should also be evaluated for its interactions and/or dependence
700 on other software programs or frameworks.
701
702 11.2.1.3.3. The programming language and updates to the programming language
703 as well as any software that either imports information to or exports
704 information from the PGS (e.g., genotyping software, CODIS entry
705 software, etc.) should be assessed for compatibility and completeness.
706
707 11.2.1.4 Usability limits are often observed through conducting internal
708 validation studies and may not require a defined reliability testing
709 study or studies. Even if reliability testing does not constitute a formal
710 study, how reliability testing was assessed must be documented.

DRAFT

11.3. *Study outcome*

11.3.1. A PGS must be operational in the physical (and virtual) environment in which it resides, whether that be on a shared network or isolated server or workstation.

11.3.1.1. The level of security needed should be determined based on how many analysts require access, the structure and requirements of licensing, and the security requirements of the laboratory or laboratory system.

11.3.1.2. If the laboratory is part of a laboratory system, a PGS may also be used in multiple locations and/or by multiple users concurrently, where potential issues such as overwriting, limits to access, and licensing requirements can be resolved through reliability testing.

11.3.2. Reliability testing further demonstrates the stability and/or accessibility of parameters and settings, and their potential for alteration.

11.3.2.1. Developer-recommended and/or validation-derived settings should be maintained from one analysis to another.

11.3.2.2. If a parameter or setting is purposefully or accidentally changed from a previously defined value, the results/report of the PGS should include the change and the laboratory should have a policy to define a check of the parameters used.

11.3.3. Communication and proper data transfer should be demonstrated between the PGS and data analysis software that imports data into the PGS, between the PGS and CODIS entry formatting software (if applicable), and any other software that imports information or extracts information from the PGS.

11.3.3.1. Barriers to communication and data transfer can be identified and should be resolved during this phase of testing.

11.3.4. Assessing the usability limits of a PGS may define elements or bounds which could result in a failure of the software, such as:

11.3.4.1. A maximum number of contributors (NOC) that can be analyzed using the software

11.3.4.2. Formatting or data importing limits

11.3.4.3. A minimum number of DNA alleles observed or number of loci with genetic data required for analysis, in combination with the estimated NOC present

11.3.4.3.1. The effects of contributor assessments of fewer contributors than the data (allele count) indicates could provide support for resolving casework NOC assessments which result in similar errors and be included in troubleshooting procedures.

11.3.5. For any elements evaluated during reliability testing, observations must be recorded, and the documentation retained with other validation materials.

11.3.5.1. Some of these elements may be appropriate for inclusion in the validation summary itself.

11.3.5.1.1. For instance, a limit to the NOC the software can analyze would be appropriate for inclusion in a validation summary.

11.3.5.1.2. Laboratory administrative IT security requirements may be defined by policy and may not need explicit reiteration in the validation summary.

12. Accuracy

12.1. *Study purpose*

12.1.1. The objective of accuracy studies is to demonstrate the ability of a measurement to give results close to a true value.

12.1.1.1. Within a PGS internal validation, accuracy studies may include demonstrating:

12.1.1.1.1. the ability of the PGS to accurately assign mixture proportions

12.1.1.1.2. genotype assignments conform to qualitative expectations

12.1.1.1.3. the LR is accurately calculated; or

12.1.1.1.4. for some systems, that raw data files are correctly analyzed.

12.2. *Study considerations*

12.2.1. A laboratory should ensure that the samples used in accuracy studies are within the range of sample complexity (e.g., NOC, mixture ratios, and template amounts) that a laboratory expects to interpret in casework analysis.

12.2.1.1. The mixtures selected for these studies may be used for multiple studies, as applicable, to cover the appropriate range of sample complexity.

DRAFT

12.2.1.2. The focus of this study should be on the known composition of mixtures and how that is reflected in mixture proportions and genotype weights.

12.2.2. Single source samples should be used to verify that a basic LR calculation is being done correctly.

12.2.2.1. The LR generated by the PGS should be similar to that calculated by hand (inverse of a random match probability (RMP)) or another validated software package (e.g., Popstats).

12.2.3. The LR assignments from sensitivity and specificity studies should support the accuracy of the PGS calculations for both true contributors and non-contributors and should be reflective of the data and informed expectations.

12.2.4. The genotype weight and mixture proportion estimates made by the PGS can be compared to the genotypes and intended mixture composition of the samples used to construct the mixtures. Mixture proportion estimation can also be compared to manually estimated proportions at loci with no allele sharing.

12.2.5. The accuracy of the allele call determination can be accomplished by various methods if the import for the PGS is raw data files.

12.2.5.1. Previously analyzed proficiency test samples and/or NIST-traceable samples can be used and verified against their known profiles. In addition, sample allele calls and approximate RFUs can be assessed by comparing them against a similar validated allele-calling program.

12.3. *Study outcome*

12.3.1. Accuracy studies should confirm that the LR obtained from a true contributor comparison to a single source profile is consistent with a calculation of 1/RMP.

12.3.2. Accuracy studies using mixture profiles should demonstrate consistency in:

12.3.2.1. the PGS-estimated mixture proportions compared to the known mixture proportions,

12.3.2.2. the PGS-estimated genotype weight assignments compared to the manual assessment of the profile and the known contributor genotypes,

12.3.2.3. the direct comparison of allele calls and peak heights observed using analysis software to those reported by the PGS.

12.3.3. Any analyses demonstrating divergence from reasonable expectations and/or known values should be investigated, with causes and resolutions documented. The lab should determine if the discordant data will remain within the dataset or if

DRAFT

additional data are needed to gain an understanding of the circumstances which may affect the analysis.

13. Sensitivity Testing

13.1. *Study purpose*

13.1.1. A sensitivity study will demonstrate and challenge the ability of a PGS to reliably provide support for the presence of a known contributor's DNA over a broad range of mixture proportions, template concentrations, and NOC observed with evidentiary typing results.

13.1.1.1. For each DNA profile being tested for sensitivity, the true contributor(s) is compared to the deconvoluted DNA profile.

13.1.1.2. The occurrence and range of LR values greater than one for true contributors (true positive) provide the laboratory with the sensitivity of the system and context for providing a verbal equivalent (if chosen to do so) of the calculated LR.

13.1.1.3. A sensitivity study will also identify the proportion of the profiles described above for which true contributors yield a LR value less than one, as well as the range of these values. The occurrence and range of LR values less than one for true contributors should be evaluated to determine whether they are due to sample quality or software failure.

13.2. *Study considerations*

13.2.1. The sensitivity studies must include, and may exceed, the range of sample complexity encountered by the laboratory, representing the bounds of the validation study which will be used to inform the future protocol.

13.2.2. Laboratories should incorporate DNA profiles into the study that are expected to show decreased sensitivity. The DNA profiles represented in this study should include, but are not limited to, the following complexities:

13.2.2.1. Profiles that exhibit a high NOC

13.2.2.2. Profiles with low template amounts

13.2.2.3. Profiles with disparate contributor ratios/proportions

13.2.2.4. Profiles with equal contributor ratios/proportions

13.2.2.5. Profiles with one or more degraded contributors

DRAFT

13.2.2.6. Profiles with inhibition patterns and

13.2.2.7. Profiles with a high degree of allele sharing between contributors

13.2.3. The laboratory should examine the sensitivity for true contributors across the range of single-source and mixture profiles included in these studies.

13.2.4. A laboratory should test the ground truth NOC, vary the assigned NOC (e.g., ±1 or 2), and/or use the apparent NOC based on the electropherogram to assess the impact on sensitivity.

13.3. *Study outcome*

13.3.1. The PGS's sensitivity demonstrates the system's ability to detect true contributors in a mixture profile, as well as the magnitude of the LRs as a reflection of each individual's contribution.

13.3.2. Sensitivity studies demonstrate the range of LR values that can be expected for known contributors based on the quality of the data.

13.3.2.1. For samples with low template or high contributor number, the mixture proportions become more ambiguous, and a decrease in the LR is expected to be observed due to the uncertainty in assigning alleles to each contributor in the mixture.

13.3.3. The sensitivity study identifies and reinforces general trends in the types (and characteristics) of mixture profiles where a true contributor's LR approaches, or is less than, one (e.g., a highly degraded sample, minor contributor, low template contributor, or high NOC).

13.3.4. A mixture component exhibiting limited data results in an increasing level of uncertainty using any DNA interpretation method and this should not be viewed as a failure of PGS.

13.3.4.1. True contributor comparisons to limited and undetected data should trend toward a LR equal to one, support for H2, or exclusion.

14. Specificity Testing

14.1. *Study purpose*

14.1.1. Specificity testing is intended to demonstrate the ability of a PGS to reliably differentiate between non-contributors and true contributors in a DNA profile and will be dependent on the data evaluated as well as the robustness of the settings/parameters within the PGS.

DRAFT

939 14.1.2. Evaluating specificity requires observing the proportion of non-contributors that
940 have a LR value of less than one (true negatives), as well as the range of LR
941 values observed.
942
943 14.1.2.1. For each DNA profile being tested for specificity, the non- contributor(s)
944 is compared to the deconvoluted DNA profile.
945 14.1.2.2. The occurrence and range of the LR values greater than one for non-
946 contributors (adventitious support) will provide the laboratory with an
947 estimate of specificity performance and the context of the reported LR
948 calculated by the PGS in relation to the quality of data.
949
950 14.1.3. This should be demonstrated using a set of profiles that challenge the PGS and
951 reflect the range of variation and quality of data observed by the laboratory.
952
953 14.2. ***Study considerations***
954
955 14.2.1. A laboratory should examine a range of sample types when assessing specificity.
956 The same profiles used for sensitivity testing can also be utilized for specificity
957 testing. The specificity of the PGS is affected by the quality of the data.
958 Increasing the NOC, including two or more contributors with approximately equal
959 proportions in the same mixture, decreasing the overall peak heights of the data,
960 increasing the degree of allele sharing, and introducing inhibition or degradation
961 will affect the specificity of the PGS in relation to a given DNA profile.
962
963 14.2.1.1. The specificity studies must include, and may exceed, the range of sample
964 complexity encountered by the laboratory, representing the bounds of the
965 validation study which will be used to inform the future protocol.
966
967 14.2.2. Laboratories should incorporate DNA profiles into this study that are expected to
968 demonstrate decreased specificity. Refer to 13.2.2 for a list of example sample
969 types where sensitivity is expected to decrease.
970
971 14.2.3. The number of non-contributors tested should be sufficient to result in a
972 spectrum of LRs from uninformative to exclusion. Any analyses resulting in LRs
973 greater than one should be scrutinized for the degree of allele sharing and quality
974 of the profile.
975
976 14.2.4. Known non-contributor profiles may be obtained from multiple sources,
977 including the following options:
978
979 14.2.4.1. DNA profiles derived from laboratory research samples
980
981 14.2.4.2. DNA profiles generated in silico using allele frequencies from the
982 appropriate population database(s)
983

DRAFT

14.2.4.3. DNA profiles obtained from a publicly available source

14.2.5. A laboratory should test the ground truth NOC, vary the assigned NOC (e.g., ± 1 or 2), and/or use the apparent NOC based on the electropherogram to assess the impact on specificity.

14.3. ***Study outcome***

14.3.1. The PGS's specificity, as characterized through this study, demonstrates the system's ability to distinguish true non-contributors in a profile, as well as the magnitude of the LRs as a reflection of quality and quantity of the data.

14.3.1.1. The plotted LRs should trend upwards to 1 (and possibly >1) for known non-contributors as less information is available within the profile.

14.3.2. A mixture component exhibiting limited data results in an increasing level of uncertainty using any DNA interpretation method and this should not be viewed as a failure of PGS.

14.3.3. The specificity study, in conjunction with the information obtained from the sensitivity study, will provide a demonstration that the PGS is providing expected levels of discrimination given the quality of the data analyzed.

14.3.4. Sensitivity and specificity studies should inform the development of training and policies regarding the proper characterization (e.g., strength) of a likelihood ratio statistic.

15. Precision

15.1. ***Study purpose***

15.1.1. The objective of precision studies is to characterize the degree of mutual agreement among a series of individual measurements, values, and/or results.

15.1.1.1. During a PGS validation, precision studies may include the following assessments, as applicable:

15.1.1.1.1. Variation in the deconvolution results (e.g., genotype weights and contributor ratios)

15.1.1.1.2. Variance parameters (e.g., allele RFU and stutter variance)

15.1.1.1.3. Diagnostic values (e.g., log (likelihood) and Gelman-Rubin diagnostic) or

15.1.1.1.4. LR values following multiple analyses of the same data

DRAFT

1030 15.1.1.2. The information acquired from precision studies can assist a laboratory in
1031 optimizing the PGS analysis parameters as well as understanding how the
1032 completeness and quality of the data influence variation in the modeling
1033 and analysis results.
1034
1035 15.1.1.3. Laboratories can also use the data from precision studies to define specific
1036 quality indicators for a PGS analysis, for example, the range of expected
1037 diagnostic values and variance parameters.
1038
1039 15.1.1.4. Defining these values and/or parameters in a standard operating procedure
1040 can assist a laboratory in ensuring the quality of PGS analyses.
1041
1042 15.2. ***Study considerations***
1043
1044 15.2.1. A laboratory should ensure that the mixtures used in precision studies are within
1045 the range of sample complexity (e.g., NOC, mixture ratios, and template amounts)
1046 that a laboratory expects to interpret in casework analysis.
1047
1048 15.2.1.1. The mixtures selected for this study should be, at a minimum, the sample
1049 types (complexity) that are expected to yield many genotype possibilities
1050 and exhibit a distribution of contributor weights following deconvolution
1051 with the PGS.
1052
1053 15.2.1.1.1. A likelihood should be calculated for each contributor to the profiles.
1054
1055 15.2.2. A laboratory may assess the impact of varying PGS analysis parameters (e.g., the
1056 number of MCMC accepts or the number of chains) on precision.
1057
1058 15.2.3. A laboratory may utilize data from other relevant internal studies, as applicable, to
1059 cover the appropriate range of sample complexity necessary when assessing
1060 precision.
1061
1062 15.3. ***Study outcome***
1063
1064 15.3.1. The information acquired from the precision studies will enable a laboratory to
1065 optimize the appropriate PGS analysis parameters and characterize the variance
1066 observed in the PGS parameters and diagnostics in relation to DNA profile
1067 completeness and quality.
1068
1069 15.3.2. For example, the primary result of a deconvolution is genotype weights, and the
1070 PGS utilizes the genotype weights to assign a LR.
1071
1072 15.3.2.1. If a PGS employs an MCMC approach, then the genotype weights and LR
1073 will vary if a PGS analysis is repeated.
1074

DRAFT

15.3.2.2. The PGS analysis parameters that affect the precision of the genotype weight estimation (e.g., the number of MCMC accepts) can be assessed, as described above, with the resultant data used to define the PGS analysis parameters that provide the desired level of precision.

15.3.3. The information acquired from the precision studies may also be used to define laboratory specific quality indicators (e.g., variance parameters, log (likelihood), and the Gelman-Rubin diagnostic) for a PGS in addition to defining the number of repeat analyses allowed (or not allowed) for a sample.

15.3.4. The precision studies described above will allow a laboratory to identify an expected range for these quality indicators for the sample complexity that the laboratory expects to interpret in casework analyses and describe trends in these indicators based on the complexity of the DNA mixture and/or relative strength of a LR.

15.3.4.1. Defining acceptable ranges for laboratory specific quality indicators will provide a mechanism to assess the quality of the PGS analyses performed within a laboratory.

15.3.5. These studies, as well as data from other validation studies, should aid the laboratory in understanding how profile completeness, complexity, and quality affect the precision of the genotype weights and calculated LR.

16. Recommended Additional Testing

16.1. While some of the following samples may have already been included in the validation studies as samples typically encountered by the laboratory, there are additional samples for laboratories to consider for inclusion during validation testing for a thorough understanding of their PGS.

16.1.1. Even though not all these sample types are encountered in casework, having tested them as part of a validation may be beneficial to inform PGS standard operating procedures and troubleshooting.

16.2. Many of these samples can be constructed in the laboratory by modifying the .txt file of the known references (or mixture profile, depending on the PGS) if actual samples do not exist.

16.3. These additional samples can include the following:

16.3.1. Challenged samples: profiles demonstrating inhibition and degradation.

16.3.2. Profiles with genetic anomalies: tri-alleles, heterozygote imbalance (e.g., primer site mutation), somatic mutations

DRAFT

16.3.3. Profiles with allele drop-in

16.3.4. Profiles demonstrating excessive allele sharing resulting in mixtures that appear to be fewer contributors than they are by allele count (e.g., biological relatives)

16.3.5. Profiles with artifacts: profiles with unedited spikes, unresolved single base pair resolution

16.3.6. Profiles where stutter peaks have been removed or artifacts have been left in (PGS dependent)

16.3.7. Profiles with the incorrect NOC (± 1 or 2)

16.3.8. Profiles deconvoluted using unsupported or incorrect assumed known references

16.4. It is strongly recommended that the laboratory tests any parameter that can be modified by the user. An example of this is the ability to modify the MCMC settings in the PGS. As described above, since MCMC is a sampling method, increasing or decreasing the number of accepts can influence the discrimination (and uncertainty) of the deconvoluted profiles.

16.5. A graduated approach allows a laboratory to focus additional testing on factors that can affect the PGS and ultimately LR results.

16.5.1. For example, performing deconvolutions on a set of mixtures, with and without an assumed contributor, will provide valuable information regarding the effect of conditioning during a deconvolution.

16.5.1.1. Subsequently, each true contributor can be tested in various combinations with other contributors and assumed contributors.

16.5.2. Testing with non-contributors or incorrect NOCs (± 1 and 2) can demonstrate how the PGS responds to what may be an incorrect assumption or simply a typo during data entry.

17. Additional features/options included in PGS

17.1. Fully continuous PGSs may include additional functions which may or may not be used by a laboratory. If the intent is to utilize such function(s), it must be validated prior to implementation and the results used to inform standard operating procedures and troubleshooting.

17.2. Not all functions have been described above; therefore, the laboratory must formulate its own study plan keeping in mind the intended use of the function and QAS requirements.

DRAFT

17.3. The studies included and the numbers and types of samples should be selected to demonstrate and challenge the validity of the feature with respect to its intended use by the laboratory.

18. Additional Comments

18.1. The above studies describe internal validation testing for a fully continuous PGS. When software updates or modifications are made by the PGS developer, the laboratory must decide whether the changes result in major or minor revisions to the software.

18.1.1. Minor updates such as additional data display, print functions or other cosmetic features will require a functional test prior to implementation.

18.1.2. Major revisions, such as updates or improvements to program code, calculations, or modeling, for example, will require functional, reliability and regression testing to ensure the PGS still functions as expected; precision, accuracy, sensitivity, and specificity studies may be needed, as applicable.

18.2. During validation, laboratories may find it helpful to create a subset of validation samples for the purpose of subsequent validation testing when the need arises.

18.2.1. These samples may include those that are at the limits of the current validation, in addition to those samples covering the range of variation.

18.2.1.1. For example, those profiles eliciting adventitious support for non-contributors or demonstrating a lack of support for true contributors could be identified and retained for sensitivity and specificity testing if a software developer makes a modification to the modeling, LR calculations or algorithms.

DRAFT

APPENDIX A: EXAMPLE PGS VALIDATION STUDY

The following study examples are informational and are not intended to dictate the types and numbers of samples every laboratory must use to satisfy each study. Validation studies cannot account for all scenarios that may arise during casework examinations; however, laboratories should attempt to cover the range of variation expected to be encountered with forensic samples. Each laboratory seeking to evaluate a new method must determine which validation studies are relevant to the methodology, in the context of its application, and determine the experiments required to satisfy each study. Following implementation, laboratories should review results and if necessary, conduct supplemental studies to improve workflow, thresholds and/or interpretations.

System Parameters (430 samples total)

Analytical Threshold
Drop-in
Stutter expectations
Saturation Limit
Allelic peak height variation (inter- and intra-locus) and stutter peak heigh variation

NOTE: This data set must be different than what is used to validate the software

Analytical Threshold:

- Determined during the validation of the genetic analyzer.

Drop-in: **(150 samples)**

- A total of 150 negative controls are evaluated, these include extraction reagent blanks and amplification negative controls
- If the amplification and CE analysis has not changed, previous extraction reagent blanks and amplification negative controls from casework analysis can be evaluated
- In addition, extra extraction reagent blanks can be extracted along with the samples used for the PGS validation and extra amplification negative controls can be amplified along with the samples used for the PGS validation
- The peak heights and number of drop-in instances were recorded.

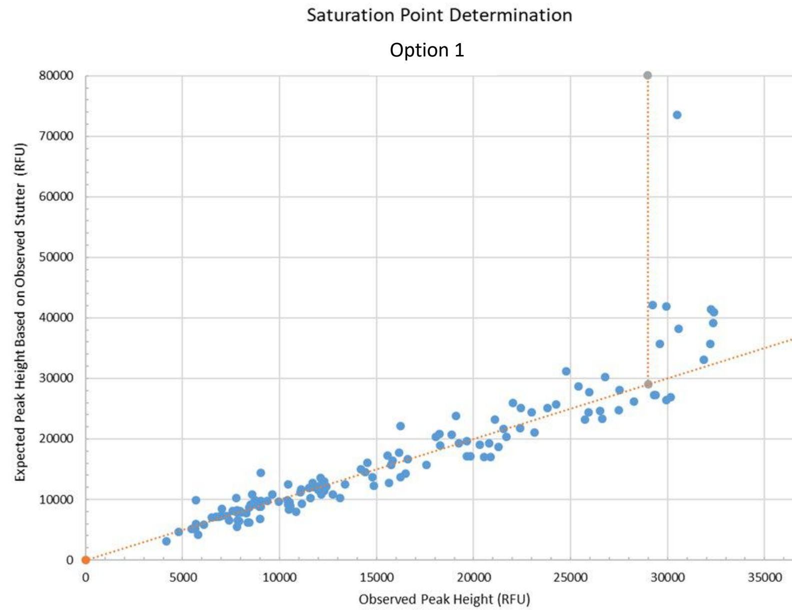
Stutter Expectations: **(100 samples)**

- Samples from 100 different donors and/or known reference samples are amplified for the multiplex kit as previously determined by the laboratory. A variety of donor DNA profiles are included to maximize coverage of alleles and stutter types. The samples are run on the genetic analyzer and analyzed without stutter filters using a lowered analytical threshold to maximize stutter observations.
- Instances of stutter being modeled (e.g., back, forward, double-back and half-back stutter) are recorded.
- Stutter ratio values are calculated for each allele at each locus.

DRAFT

1247 - Stutter ratios are graphed by allele and/or by longest uninterrupted sequence (LUS) of
1248 repeat units (where relevant) versus percent stutter calculated. Results are evaluated to
1249 determine best fit regressions for each locus and each stutter type.
1250 - The maximum expected stutter ratio can also be determined from this data set.
1251
1252
1253 *Saturation Limit:* **(48 samples)**
1254 - Option 1: A series of six known samples at varying template amounts that are expected
1255 to result in saturation are used. The samples are amplified in duplicate and run with
1256 standard injection times on each genetic analyzer in the laboratory.
1257 - Option 2: A series of six known samples of standard input template amount are run using
1258 standard injection times to determine stutter ratios. The samples are then run with
1259 increased injection times until the peaks demonstrate saturation.
1260 - The data is analyzed without stutter filters. It may be necessary to lower the AT to
1261 capture more stutter data.
1262 - Samples are interpreted, and peak heights of parent and back stutter peaks are recorded.
1263 Peak heights augmented by stutter are excluded from the data.
1264 • Option 1: Using the average back stutter ratio expectation values captured within
1265 the stutter study, the observed parent peak heights are compared to the expected
1266 parent peak heights.
1267 ▪ Plot the observed parent allelic peak height against the expected parent
1268 peak height. The point at which the expected parent peak height diverges
1269 represents the saturation level for that instrument.
1270 • Option 2: The observed back stutter ratio from the standard injection are
1271 compared to the stutter ratios obtained from the increased injection times to
1272 identify when the data becomes saturated.
1273 ▪ Plot the observed back stutter ratio from the standard injection against the
1274 stutter ratios from the increased injection times. The point at which the
1275 observed stutter ratios diverge represents the saturation level for that
1276 instrument.
1277 - For laboratories with multiple instruments, the lowest saturation limit may be used for all,
1278 or the saturation limit can be instrument specific. Alternatively, some data analysis
1279 software will identify peaks that have saturated the genetic analyzer's detector and result
1280 in off-scale data. Peak heights of alleles identified as off-scale by the data analysis
1281 software are recorded.
1282 - Alternatively, some data analysis software will identify peaks that have saturated the
1283 genetic analyzer's detector and result in off-scale data. Peak heights of alleles identified
1284 as off-scale by the data analysis software are recorded.
1285

DRAFT



1286
1287
1288
1289 *Allelic and Stutter Peak Height Variation:* **(132 samples)**

1290 - For a laboratory that targets samples with 50pg total DNA to an optimum target of 1 ng,
1291 mixture components will contribute only a fraction of the total DNA; therefore, samples
1292 from six (6) different donors are amplified at multiple target amounts ranging from 2.0 ng
1293 to 8 pg (refer to Example Dilution Series below) for a total of 132 amplifications.
1294 - Samples are run on both genetic analyzers used by the laboratory. Profiles are analyzed
1295 without stutter filters and using a lowered analytical threshold to capture allelic and
1296 stutter peaks below threshold.
1297 - Labels for allelic peaks and their associated stutter, along with apparent drop-in peaks,
1298 are retained while labels for other artifacts are removed.
1299 - The data from these samples can be used by the PGS parameter setting module to
1300 generate variance parameters and distributions for allelic peak height, intra-locus and
1301 inter-locus peak height variability and stutter peak height variability.

1302
1303
1304 **Internal Validation Testing**

1305
1306 *Functional Testing:*

1307 - Eight DNA profiles consisting of three single source and five mixture profiles were
1308 utilized for functional testing (see table below)
1309 - The mixture profiles were designed to contain optimal quantities of DNA and readily
1310 discernable mixture proportions when assessed manually
1311 - The DNA profiles were imported into the software using a designated file type and
1312 format
1313 • Import file format(s) documented
1314 • One instance of an off-ladder allele (OL) that is not assigned an allele call is
1315 included in the import file

- One instance of an allele value labeled as “<6” was included in the import file
- After import and analysis, the data (allele, peak height, and size) for the evidence and reference profiles analyzed were confirmed in the report ensure it was imported properly

- Parameters and settings
  - A single source DNA profile was analyzed, and the parameters and settings listed in the resulting report were compared to the expected values to confirm they were entered correctly
  - Each of the adjustable parameters (e.g., MCMC accepts, seed number, HPD calculations, population databases) were changed for an analysis
  - The analysis was subsequently repeated, and the parameters and settings were checked to verify that those that were adjusted reverted to the default settings
- General profile deconvolution
  - The eight DNA profiles were analyzed
    - The results of the deconvolution were compared to the experimental set-up of the DNA samples and the manual interpretation of the DNA profiles
  - LRs were assigned for the known contributors and one known unrelated non-contributor
    - The assigned LRs of the known contributors were evaluated to determine if the magnitude of the LR was reflective of the observed contribution to the DNA profile (e.g., assigned LRs for minor component contributors to a mixed DNA profile were less than the assigned LRs for the major component contributors)
    - The assigned LRs for the known non-contributors were all zero or less than 1

- Reports
  - The reports for each analysis were reviewed to ensure completeness:
    - All expected sections
    - Diagnostics
    - Results of modeling
    - Deconvolution results
    - LRs for each selected population
    - Settings and parameters, including an indicator for adjusted settings and parameters

Example: Functional Testing Summary Chart

	Data Import	**Parameters**	**Deconvolution**	**LR**	**Report**
Single Source 1 w/ “<6”	Error noted at import, correction made and completed analysis	All maintained	Consistent with expectations	K1: 1.4 E30 (H_1) K2: Excluded	All sections/info present

DRAFT

Single Source 2 w/ “OL”	Unassigned OL- failure of all subsequent data to import, correction made and analysis completed, all types	All maintained	Consistent with expectations	K3: 8.0 E27 (H_1) K4: Excluded K5: Excluded	All sections/info present
Single Source 3 – adjusted parameters and settings	All types	Adjusted	Consistent with expectations	K6: Excluded K7: 2.2E24 (H_1)	All sections present/info present, adjusted settings and parameters indicated
Single Source 3	All types	All maintained, back to default	Consistent with expectations	K6: Excluded K7: 2.2 E24 (H_1)	All sections/info present
1:1 Mixture	All types	All maintained	Mixture proportions are not consistent with a 1:1 mixture	K8: Excluded K9: 3.7 E20 (H_1) K10: 4.1 E21 (H_1)	All sections/info present
5:1 Mixture	All types	All maintained	Consistent with expectations	K11: 5.5 E26 (H_1) K12: 8.1 E8 (H_1) K13: Excluded	All sections/info present
2:1 Mixture	All types	All maintained	Consistent with expectations	K14: 7.2 E17 (H_1) K15: 1.6 E10 (H_1) K16: Excluded	All sections/info present
5:3:1 Mixture	All types	All maintained; Analyze a second time with additional accepts	Consistent with expectations	K17: 9.7 E23 (H_1) K18: 6.8 E15 (H_1) K19: 3.9 E7 (H_1) K20: 5.4 E-14	All sections/info present
5:5:1 Mixture	All types	Additional accepts setting mistakenly retained	Consistent with expectations	K21: 4.8 E19 (H_1) K22: 3.6 E20 (H_1) K23: 2.2 E6 (H_1) K24: Excluded	All sections/info present

Reliability Testing:

- The physical reliability of the software was evaluated during the *Functional Testing* (see above):
    - Software developer specifications were considered when installing the PGS in the testing environment (RAM, processor, etc.)

DRAFT

    - For each analysis run, the parameters were reviewed to ensure they did not change between analyses.
    - Any changes that were made to parameters returned to the default value whenever a subsequent analysis was conducted.
    - Changes to parameters were confirmed as readily displayed in the software results/report.
    - Input files containing "OL" or other artifact labels were identified at analysis set-up.
- NOC limitations
    - Equal two-person, three-person, four-person, five-person, and six-person mixtures (1:1, 1:1:1, etc.) were prepared, analyzed, and the inputs created for the PGS.
    - The mixtures were analyzed sequentially until the PGS was not able to complete the analysis.
    - Optional parameters such as low memory mode were employed to determine if the analysis could be successfully completed.
- Incorrect NOC
    - A three-person mixture with six alleles at multiple loci was analyzed using a NOC = 2
    - The analysis reported an error and could not be completed

Accuracy:

- Accurately assign mixture proportions
    - The two- and three-person mixtures from the *Functional Testing* (1:1, 2:1, 5:1, 5:3:1, and 5:5:1) were visually evaluated to estimate the mixture proportions by examining loci containing the maximum number of expected alleles.
    - The visual estimations were then compared to the PGS-generated mixture proportions which were consistent.
- Genotype assignments conform to qualitative expectations
    - The two- and three-person mixtures from the *Functional Testing* (1:1, 2:1, 5:1, 5:3:1, and 5:5:1) were visually evaluated to develop approximate expectations for the weightings for possible genotype sets.
    - The visual estimations were then compared to the PGS-generated weights which were generally consistent. Any inconsistencies were noted, and possible causes listed.
- LR is accurately calculated (single source)
    - Five high quality single source DNA profiles (i.e., not exhibiting drop-in, drop-out, or alleles in stochastic RFU range) were compared to the known contributor's reference sample and an LR was assigned by the PGS.
    - A random match probability (RMP) was calculated for each comparison using another statistical program validated by the laboratory. The RMP was calculated and recorded using the same allele frequencies, treatment of rare alleles, and theta coefficient.
    - A comparison was made between the LR and 1/RMP.
- LR is accurately calculated (mixture)
    - Evaluated in the *Functional Testing* above.

DRAFT

- LRs assigned to the known contributors to the two- and three-person mixtures were evaluated to determine if they conformed to the qualitative expectations based on the LR reflective of informed expectations based on the quantity and quality of the data should be obtained.

Sensitivity:

- This study is based on the laboratory's policy that mixtures containing up to four individuals will be interpreted and used for comparison purposes.
- A set of mixtures using known contributors was prepared as described in the table below. The mixtures covered the following:
  - NOC up to 5
  - Varying template amounts (total template from 0.01 ng to 2.0 ng)
  - Varying mixture proportions
  - Varying levels of degradation
  - Varying levels of allele sharing
  - All samples prepared, amplified, and analyzed in duplicate

Contributor Ratio	Total DNA Template (ng) (Template amplified for each set, for a total of 3 different donor sets)	Number of Amplifications
Single Source Profiles:		
1	2.0, 1.0, 0.75, 0.5, 0.25, 0.1, 0.05, 0.025, 0.01	18
Total		18
Two-Person Mixtures:		
95:1	2.0, 1.0, 0.5	6
50:1	2.0, 1.0, 0.5, 0.25	8
20:1	2.0, 1.0, 0.5, 0.15	8
10:1	2.0, 1.0, 0.5, 0.25, 0.1	10
5:1	2.0, 1.0, 0.5, 0.25, 0.1, 0.05	12
3:1	2.0, 1.0, 0.5, 0.25, 0.1, 0.01	12
2:1	2.0, 1.0, 0.5, 0.25, 0.1, 0.01	12
1:1	2.0, 1.0, 0.5, 0.25, 0.1, 0.01	12
Total		80

DRAFT

Three-Person Mixtures:		
96:1:1	2.0, 1.0, 0.5	6
64:32:1	2.0, 1.0, 0.5	6
32:1:1	2.0, 1.0, 0.5, 0.25	8
16:1:1	2.0, 1.0, 0.5, 0.25, 0.1	10
8:1:1	2.0, 1.0, 0.5, 0.2, 0.075	10
4:2:1	2.0, 1.0, 0.5, 0.15, 0.05	10
4:1:1	2.0, 1.0, 0.5, 0.15, 0.05	10
2:1:1	2.0, 1.0, 0.5, 0.15, 0.05	10
1:1:1	2.0, 1.0, 0.5, 0.15, 0.05	10
Total		80
Four-Person Mixtures:		
96:4:1:1	2.0, 1.0, 0.5	6
64:8:1:1	2.0, 1.0, 0.5	6
32:4:1:1	2.0, 1.0, 0.5, 0.25	8
16:1:1:1	2.0, 1.0, 0.5, 0.25, 0.1	10
8:1:1:1	2.0, 1.0, 0.5, 0.25, 0.075	10
4:3:2:1	2.0, 1.0, 0.75, 0.5, 0.25	10
4:2:1:1	2.0, 1.0, 0.5, 0.25, 0.1	10
4:1:1:1	2.0, 1.0, 0.5, 0.25, 0.1	10
2:1:1:1	2.0, 1.0, 0.5, 0.15, 0.05	10
1:1:1:1	2.0, 1.0, 0.5, 0.15, 0.05	10
Total		90
Five-Person Mixtures:		
50:4:2:1:1	2.0, 1.0, 0.5	6
10:2:2:1:1	2.0, 1.0, 0.5, 0.25, 0.1	10

DRAFT

5:2:2:1:1	2.0, 1.0, 0.25, 0.075	8
4:1:1:1:1	2.0, 1.0, 0.25, 0.05	8
Total		32
Degraded Profiles:		
Single source – moderately degraded	2.0, 1.0, 0.75, 0.5, 0.25, 0.1, 0.05	14
Single source – heavily degraded	2.0, 1.0, 0.75, 0.5, 0.25, 0.1, 0.05	14
Two-person mixture – major degraded (2:1, 5:1, 10:1, 20:1)	1.0, 0.5, 0.25	24
Two-person mixture – minor degraded (2:1, 5:1, 10:1, 20:1)	1.0, 0.5, 0.25	24
Three-person mixture – major degraded (4:2:1, 2:2:1, 8:1:1, 10:5:1)	1.0, 0.5, 0.25	24
Three-person mixture – mid degraded (4:2:1, 2:2:1, 8:1:1, 10:5:1)	1.0, 0.5, 0.25	24
Three-person mixture – minor degraded (4:2:1, 2:2:1, 8:1:1, 10:5:1)	1.0, 0.5, 0.25	24
Four-person mixture – major degraded (4:3:2:1, 4:2:1:1, 8:4:1:1, 20:10:5:1)	1.0, 0.5, 0.25	24
Four-person mixture – mid degraded (4:3:2:1, 4:2:1:1, 8:4:1:1, 20:10:5:1)	1.0, 0.5, 0.25	24
Four-person mixture – minor degraded (4:3:2:1, 4:2:1:1, 8:4:1:1, 20:10:5:1)	1.0, 0.5, 0.25	24
Total		220

DRAFT

Allele Sharing		
Parent / Child (5:1, 3:1, 1:1)	2.0, 1.0, 0.5, 0.25, 0.1	45
Parent/Child/Unrelated (5:3:1, 3:3:1, 1:1:1)	2.0, 1.0, 0.5, 0.25, 0.1	45
Parent/Parent/Child (5:3:1, 3:3:1, 1:1:1)	2.0, 1.0, 0.5, 0.25, 0.1	45
Parent/Parent/Child/ Unrelated (10:5:3:1, 5:5:3:1, 1:1:1:1)	2.0, 1.0, 0.5, 0.25, 0.1	45
Total		180
Grand Total		**640**

1425

1426 - Each of these profiles are analyzed using the apparent NOC and the apparent NOC+1 and
1427 the apparent NOC-1.
1428 - Comparisons to each of the knowns are performed for all the deconvoluted contributors
1429 (NOC = N, N+1, and N-1).
1430 - Plots of calculated LR values versus template amount, average peak height (APH),
1431 maximum number of obligate alleles, contributor proportion (or percent contribution),
1432 effect of replicates, etc. can inform the sensitivity of the PGS and the context of the
1433 magnitude of the LR. The following examples are not all-inclusive.
1434 • Plot the log (LR) values against the APH per contributor. The APH per known
1435 contributor is determined from the unmasked and unshared alleles. Where no
1436 DNA from the individual was detected within the profile, the APH may be set to
1437 half the AT. The APH per contributor value is used since this is the most
1438 comparable to the information an analyst will have with forensic casework and is
1439 therefore the most relevant explanatory variable to plot.
1440 • Data from all true contributors from mixtures with the same NOC are plotted with
1441 the log +(LR) on the y-axis and the APH on the x-axis. This plot can provide
1442 context of the LR magnitude as compared to the APH of each contributor.
1443 • Tabulate the fraction of true-contributor LR values that fall within ‘bins’ on either
1444 side of LR=1 and compare to specificity study results (see below). Instances
1445 where a true contributor exhibits a LR less than one should be addressed through
1446 a review of the data to determine if the results are as expected.
1447 • Assess the impact of the NOC on sensitivity by plotting LRs resulting from N+1
1448 and N-1 analyses, as compared to the true NOC.

DRAFT

1449 - Sensitivity can be expressed globally across the entire study and across each condition in
1450 the validation study. The laboratory should assess the global sensitivity of the system as
1451 well as the sensitivity of each condition or component part of the system.
1452 - Any analyses demonstrating divergence from reasonable expectations and/or known
1453 values should be investigated, with causes and resolutions documented. The lab should
1454 determine if the discordant data will remain within the dataset or if additional data is
1455 needed to gain an understanding of the circumstances which may affect the analysis.
1456
1457 *Specificity:*
1458 - For each validation DNA profile (created in the *sensitivity study*) tested for the *Specificity*
1459 study, a set of known non-contributors was compared to the deconvoluted DNA profile.
1460 • A database of 10,000 in silico generated non-contributor DNA profiles is used to
1461 evaluate specificity.
1462 • Using the same software settings as for a known contributor (see sensitivity
1463 study), an LR value is calculated and recorded for each non-contributor being
1464 evaluated against the validation profile.
1465 • The LR values per non-contributor and template amount of the lowest contributor
1466 for each DNA profile are tabulated and/or plotted. These are combined with the
1467 sensitivity plots to include both contributors and non-contributors within the same
1468 plot.
1469 • Instances where the non-contributor comparison results in a LR greater than one
1470 are recorded.
1471 • The range of LRs from non-contributors (e.g., minimum and maximum) is
1472 assessed.
1473 • The exact LR values are recorded to provide further context as to the LR
1474 magnitude in relation to profile type and validation conditions (e.g., NOC,
1475 contributor proportion, contributor DNA template, level of degradation, etc.).
1476 • Specificity is calculated as the proportion of non-contributors with a LR less than
1477 one (those favoring H2 and exclusions) divided by the total number of
1478 comparisons.
1479 ▪ Specificity can be expressed globally across the entire study or for a
1480 condition(s) in the validation study.
1481 o Note that the specificity calculation does not define a proportion of
1482 non-contributors expected to produce A LR>1 for any specific
1483 sample. The calculation is simply a metric for the validation to
1484 determine whether the software can effectively discern true
1485 contributors from non-contributors.
1486 • Further methods of data analysis to consider:
1487 ▪ Plots of LR values versus template amount, average peak height,
1488 contributor proportion (or percent contribution), and effect of replicates
1489 can be considered.
1490 o For instance, plot the log (LR) values against the average peak
1491 height (APH) per contributor. The APH is determined from the
1492 unmasked and unshared alleles.
1493 o Where no DNA from the individual was detected within the
1494 profile, the APH may be set to half the AT.

DRAFT

  - The APH per contributor value is used since this is the most comparable to the information an analyst will have with forensic casework and is therefore the most relevant explanatory variable to plot.
  - Tabulate the fraction of non-contributor log (LR) values that fall within ‘bins’ on either side of LR=1. APH for non-contributors to a given profile is taken as the minimum APH among the known donors to the profile. Instances where a non-contributor exhibits an LR greater than 1 should be addressed through a review of the data to determine if the results are as expected.
  - Assess the impact of the NOC on specificity by plotting LRs resulting from N+1 and N-1 analyses as compared to the true NOC.

Precision:

- Eight DNA profiles that span the laboratory’s intended application were selected for the precision studies and are listed below. These samples were amplified as part of the laboratory’s sensitivity and specificity studies.

Mixture Ratio	Template (ng)
1:1:1:1	1
4:3:2:1	0.75
3:2:1	0.5
1:1	1
1:1	0.5
1:1	0.25
Single source	0.50
Single source	0.10

  - Deconvolution was performed on each of the above listed DNA profiles using 5,000, 50,000, and 100,000 accepts per chain.
  - LRs were assigned to each of the known contributors to the DNA profiles using each of the deconvolutions.
  - The LRs and diagnostic values including the template amounts, mixture proportions, and genotype weights were evaluated for each deconvolution.
  - Based on this data and the manufacturer’s recommendation, 50,000 accepts per chain were selected as the default value.
- The precision of the deconvolutions and LR assignment was then evaluated by performing the deconvolutions and LR calculations for each contributor as described in the table below:

Mixture Ratio	Total Template (ng)	Number of Post Burn-in MCMC Accepts per Chain	Number of Analyses
1:1	1	50,000	5
1:1	0.5	50,000	5
1:1	0.25	50,000	5

DRAFT

3:2:1	0.5	50,000	5
4:3:2:1	0.75	50,000	5

- The range of LRs and diagnostic values including the template amounts, and mixture proportions for each set of replicate analyses was evaluated.

Additional Evaluations:

- Evaluate the range of diagnostic values for deconvolutions using the correct and incorrect NOC
- Evaluate the effectiveness of the use of a confidence interval calculation or its equivalent if available
- Evaluate the use of informed priors if available
- Evaluate the use of other user-selected parameters such as the maximum degradation value

References and Suggested Readings

An assessment of the performance of the probabilistic genotyping software EuroForMix: Trends in likelihood ratios and analysis of Type I & II errors. Benschop CCG, Nijveld A, Duijs FE, Sijen T. Forensic Sci Int Genet. 2019 Sep; 42:31-38. doi: 10.1016/j.fsigen.2019.06.005. Epub 2019 Jun 10. PMID: 31212207

Application of Random Match Probability Calculations to Mixed STR Profiles, Bille T, Bright J-A, Buckleton J. Journal of Forensic Sciences. 2013; 58; 474-85

Establishing the limits of TrueAllele® Casework: a validation study. Greenspoon, S.A., Schiermeier-Wood, L., and Jenkins, B.C. Journal of Forensic Sciences, 60(5):1263-1276, 2015

Internal validation of STRmix™ - A multi laboratory response to PCAST. Bright JA, Richards R, Kruijver M, Kelly H, McGovern C, Magee A, McWhorter A, Ciecko A, Peck B, Baumgartner C, Buettner C, McWilliams S, McKenna C, Gallacher C, Mallinder B, Wright D, Johnson D, Catella D, Lien E, O'Connor C, Duncan G, Bundy J, Echard J, Lowe J, Stewart J, Corrado K, Gentile S, Kaplan M, Hassler M, McDonald N, Hulme P, Oefelein RH, Montpetit S, Strong M, Noël S, Malsom S, Myers S, Welti S, Moretti T, McMahon T, Grill T, Kalafut T, Greer-Ritzheimer M, Beamer V, Taylor DA, Buckleton JS. Forensic Sci Int Genet. 2018 May; 34:11-24. doi: 10.1016/j.fsigen.2018.01.003. Epub 2018 Jan 8. PMID: 29367014

Internal validation of STRmix™ for the interpretation of single source and mixed DNA profiles Tamyra R Moretti, Rebecca S Just, Susannah C Kehl, Leah E Willis, John S Buckleton, Jo-Anne Bright, Duncan A Taylor, Anthony J Onorato; Forensic Sci Int Genetics 2017 Jul;29:126-144

Quality Assurance Standards for Forensic DNA Testing Laboratories; Federal Bureau of Investigation. (2020) available at https://www.fbi.gov/about-us/lab/codis/qas-standards-for-forensic-dna-testing-laboratories-effective-07-01-2020.

DRAFT

STRmix™ put to the test: 300 000 non-contributor profiles compared to four-contributor DNA mixtures and the impact of replicates. Noël S, Noël J, Granger D, Lefebvre JF, Séguin D. Forensic Sci Int Genet. 2019 Jul; 41:24-31. doi: 10.1016/j.fsigen.2019.03.017. Epub 2019 Mar 21. PMID: 30947115

Validating TrueAllele®DNA Mixture Interpretation. Mark W. Perlin,M.D., Ph.D.; Matthew M. Legler,B.S.; Cara E. Spencer,M.S.; Jessica L. Smith,M.S.; William P. Allan,M.S.; Jamie L. Belrose,M.S.; and Barry W. Duceman,Ph.D. Journal of Forensic Sciences. November 2011. 56:6

Informational Web Site: Additional information may be obtained from the following web site: https://strbase.nist.gov/

DRAFT

1586

Document Version	Revision History
XXXXX, 2024	Original (Published on SWGDAM website). Replaces the 2015 SWGDAM Guidelines for the Validation of Probabilistic Genotyping Systems

1587
1588
1589

DRAFT