Review of PowerPlex® 21 & STRmix™ v1.05 Validations

20 November 2022

PREPARED FOR: **The Commission of Inquiry into Forensic DNA Testing in Queensland**

AUTHORS: **Dr Bruce Budowle: PowerPlex® 21 validation**

Ms Johanna Veth (ESR Ltd): STRmix™ v1.05 validation

DISCLAIMER AND DECLARATION OF CONFLICTS OF INTEREST

The Institute of Environmental Science and Research Limited (ESR) owns STRmix Limited (a fully owned subsidiary company) which licenses STRmix™ software to laboratories around the world, including the forensic DNA Laboratory at Queensland Health Forensic and Scientific Services.

The authors have used all reasonable endeavours to ensure that the information contained in this client report is accurate. However, the authors do not give any express or implied warranty as to the completeness of the information contained in this client report or that it will be suitable for any purposes other than those specifically contemplated during the Project or agreed in writing. The conclusions in this report are opinions based on the information provided to the authors at a point in time and may change if additional information is provided.

Contents

1. EXECUTIVE SUMMARY

1. The Australian laboratories were required to concurrently validate a new profiling kit, such as PowerPlex® 21, and STRmix™ software which was an enormous challenge.

2. The QHFSS laboratory should have taken more time to complete their validations, which was needed by the other Australian laboratories.

3. In general, the design of the PowerPlex® 21 validation was consistent with best practice.

4. In general, the STRmix™ v1.05 validation appears to have been competently undertaken and demonstrated a good understanding of the software. As there were no 'best practice' guidelines available for probabilistic genotyping software or formalised validation guidelines from the STRmix™ developers, the guidance provided by STRmix™ was perhaps haphazard.

5. Flawed interpretation of data was evident in both validations.

6. It seems that the 132pg threshold was set *prior* to the completion of the two validations. Each validation demonstrated that data from DNA from samples containing less than 132pg could be reliably obtained and interpreted.

7. Implementing a threshold that was unsupported by the data created a great risk of not detecting potentially probative, exculpatory or otherwise informative profiling results.

8. The laboratory was remiss in not conducting a proper 6-month review after implementing PowerPlex® 21.

2. POWERPLEX® 21 VALIDATION

9. The Commission of Inquiry into Forensic DNA Testing in Queensland requested review of the validation of the PowerPlex® 21 kit by Queensland Health Forensic and Scientific Services (QHFSS). The question specifically posed is "Whether the validation of the PowerPlex® 21 kit was performed in accordance with best practice and, if not, whether any failures had, or could have had, an effect upon the results of DNA testing in the Blackburn case."

10. The opinion in this Report is based on information in various files and communications provided by the Commission and is contained within the folders titled: Experiments and data; Other; PowerPlex 21; Standard Operating Procedures; and Validation documents. While all files contributed to the review, the specific files cited in this report are:

- PowerPlex 21 – Direct Amplification of Reference FTA samples validation - 26.09.2012
- PowerPlex 21- Amplification of Extracted DNA Validation – 14.12.2012
- 16. Experiments PP21 2012
- PowerPlex 21 Direct Amp of Ref FTA samples 26_09_2012
- 11. Project Report 131 PP21 Post-implementation review – K.Scott, L.Ryan, C. Allen 01_2016
- 47. Summary Half and Full Volume PP21 15_02_2013

2.1 INTRODUCTION

11. Validation of a method is an important aspect of a quality system. The studies performed inform a laboratory of the limitations of a methodology or technology and guide that laboratory in developing analytical and interpretation protocols that can be fit for purpose. It is imperative that a laboratory allows the results of a validation study to dictate the performance parameters of the methodology it seeks to test and subsequently implement.

12. A review of the studies indicates that QHFSS was actively engaged in the design and validation of PowerPlex® 21 kit (herein PP21). The study design of the samples that were analyzed included sensitivity studies for assessing stochastic effects, limits of detection and limits of reporting; analyses of population samples for determination of stutter and assessment of concordance; determination of a baseline to address differentiation between noise and signal; mixture studies for potential deconvolution and/or detection of minor contributors; samples analyzed on two different capillary electrophoresis instruments for reproducibility (see 16. Experiments PP21 2012); determination of probability of drop in; an evaluation of amplification reaction volumes (i.e., designated half (i.e., 12.5 µl) and full (i.e., 25 µl) volumes); and an assessment of potential contamination (see 2. PowerPlex 21- Amplification of Extracted DNA Validation – 14.12.2012). These tests are

the primary studies performed during validation[1] and are internationally accepted aspects to assess during validation of a STR typing kit.

13. Additional studies, that were not performed, include testing with mock case material or nonprobative evidence and exposure to PCR inhibitors. These studies while interesting do not add *per se* to protocol development for implementation of PP21, but could help with sample preparation, and thus should not be considered as critical for the validation of PP21. Also, there did not appear to be an analysis of precision. However, the data obtained from the validation studies performed by QHFSS are consistent with the authors' personal experience and a publication on a developmental study of PP21 by Ensenberger *et al*[2] titled "Developmental validation of the PowerPlex 21 System". Thus, the data generated were consistent with expectations and can be relied on, to a degree, to assist in developing effective standard operating protocols.

14. There were some issues with the validation processes such as determining probability of drop in, use of an incorrect injection time for one of the capillary electrophoresis (CE) instruments, baseline determination, and carryover. These issues already have been addressed in the report titled "Review of DNA analysis undertaken in the Blackburn case" by Johanna Veth (17 November 2022) and need not be repeated herein. One issue noted, although, was the explanation by QHFSS for the observed differences in performance between the CE instruments – designated A and B, which was subsequently identified to be due to a lower injection time used for samples analyzed on instrument B. On page 28 in file "2. PowerPlex 21- Amplification of Extracted DNA Validation – 14.12.2012" the laboratory suggests that the observed performance differences

> "*could be due to a prolonged period between spectral calibrations, aging reagents and arrays and was taken into consideration when setting thresholds.*"

When performing a validation study that will set the conditions of performance for a method, it would be better practice to use reagents, materials and calibrations that are relatively fresh or up to date especially if they may impact setting of the thresholds as well as comparison of performance between two instruments. It would be preferable to use the highest quality materials at the onset so a laboratory could establish high performance data as its operational baseline. Then, during operation a diagnostic could be instituted to assess if output signal begins to decay due to aging of consumables and drift of system performance. If there were such differences as suggested by QHFSS, it would raise some additional concerns with the design of the validation studies and the data generated.

2.2 THE 132 PG THRESHOLD

15. Regardless, of the above issues, the general design of the analytical part of the PP21 validation study was consistent with best practices; however, the

[1] for example see swgdam_validation_guidelines_approved_dec_2012.pdf (bloodstainevidence.com)
[2] in Forens. Sci. Genet (2014) 9:169-178

interpretation of the data and subsequent implementation of the method (and concomitant interpretation guidelines) did not meet the level of best practices. It is evident that QHFSS set a threshold for input DNA to analyze which impacts the data generated (i.e., the 0.132 ng (or 132 pg) minimum bound concentration for full amplification reactions) prior to assessing the sensitivity and performance of PP21. As an example, in "2. PowerPlex 21- Amplification of Extracted DNA Validation – 14.12.2012" (on page 7) it is stated:

> *"Generally samples with lower templates (reaching the often termed 'low copy number' level of 100-150pg) tend to exhibit enhanced stochastic effects as one would expect. Therefore, it should be considered whether samples around this input template level should be amplified given that interpretation of the results could be unwieldy."*

16. Clearly, the laboratory is focused on input amounts driven solely by its preferred threshold of 132 pg. Note that this concern of unwieldy interpretation, though, is inconsistent with the initial adoption of a half volume for the amplification reaction. The half volume data had notably greater demands on interpretation, and this issue is discussed later in this report. Very soon after implementation QHFSS ceased using half reactions in its operational workflow.

17. QHFSS appears to have adopted the 132 pg input threshold before analyzing any of the validation data even though the results supported that a lower amount of input DNA can yield substantial informative data. The laboratory was well aware of the risk of setting a threshold too high or too low. For example, in file "1. PowerPlex 21 – Direct Amplification of Reference FTA samples validation -26.09.2012" (on page 37) the laboratory, referencing other authors, states:

> *"The use of thresholds for reporting is essentially a risk assessment (23), if the thresholds are set too low then labelling of artefacts and noise may occur, if set too high then real peaks will not be labelled and information will be lost (20, 24)."*

18. We have not been provided with data that indicate that such a risk assessment was performed. Instead, QHFSS accepted the 132 pg input threshold, even if the validation and the laboratory's findings support that it is too high. For example, in "2. PowerPlex 21- Amplification of Extracted DNA Validation – 14.12.2012" (on page 7) it is stated:

> *"The sensitivity of this next generation STR kit has greatly increased, however the increased sensitivity does not necessarily result in increased information. The results of this validation indicates that Promega's PowerPlex® 21 system is a very sensitive STR amplification kit, but to reduce the risk of type 2 errors (calling a heterozygous locus homozygous[1]) consideration needs to be given to restricting the range of DNA template added."*

19. This position, in itself, is not supportable. One should not consider "a type 2 error" as stated by QHFSS as an error. It has been known for more than 20

years[3] and well before the validation studies were undertaken by QHFSS that allele drop out can occur, and there are two possible genotype categories – a single peak may be a homozygote or part of a heterozygote. Both genotype states would be entertained when interpreting DNA profiles. Manual methods[4] have been in place since the inception of STR typing to address the issue and effectively interpret some data. Moreover, with the advent of probabilistic genotyping (and in particular STRmix™) methods for profile interpretation are available that are more effective than manual methods at addressing missing data. Thus, restricting data solely on the potential of allele drop out is inconsistent with the interpretation methodologies at the timeframe of 2012-2013.

20. Even if allele drop out in which a heterozygote would present only one of the alleles was set as a main criterion for selecting a threshold, the data generated during the validation studies support that a much lower input amount than 132 pg should have been implemented. In file "2. PowerPlex 21- Amplification of Extracted DNA Validation – 14.12.2012" (on pg 32) Figure 1 shows that at 50 pg input DNA near to full profiles were obtained and even at 10 pg half the alleles were detected from samples at full and half amplification reactions. The laboratory concludes on page 34 that:

> *"The average peak height ratio decreased as the DNA template decreased to 50pg. Below a DNA template of 50pg less heterozygote pairs were observed (as expected) which resulted in the peak height ratio becoming more variable and drop out being observed."*

and on page 40 determined that:

> *"The µPHR_25 at 25pg input was 0.736 and at 0.5ng input was 0.851 compared with the µPHR_12.5, at 25pg input was 0.598 and at 0.5ng was 0.832."*

21. µPHR is the overall average peak height ratio (PHR). The closer the value approaches 1.0 the more balanced are the peaks suggesting less impact or severity due to stochastic effects. A PHR in the range of 0.7 and above is considered relatively robust.

22. Thus, these results and findings by QHFSS stated on pages 34 and 40 support that the data from 50 pg of input DNA and indeed 25 pg of input DNA did not exhibit substantial allele drop out. Thus, the validation data do not support a "Type 2 Error" threshold being set at 132 pg input DNA. More appropriately a much lower amount of input DNA could be tolerated and in turn obtain interpretable data.

23. Indeed, on pages 51 and 52, Figures 23 and 24, respectively, show only one drop out event was observed at 50 pg input for a full amplification reaction

[3] for example, see Moretti et al. Validation of short tandem repeats (STRs) for forensic usage: Performance testing of fluorescent multiplex STR systems and analysis of authentic and simulated forensic samples. J. Forens. Sci. 46(3):647-660, 2001; Budowle et al. Mixture interpretation: defining the relevant features for guidelines for the assessment of mixed DNA profiles in forensic casework. J. Forens. Sci. 54:810-821, 2009.
[4] for a revisit of a manual approach see Bieber et al. Evaluation of forensic DNA mixture evidence: protocol for evaluation, interpretation, and statistical calculations using the combined probability of inclusion. BMC Genetics 17(1):125, 2016

between the two samples tested. This observation indicates a very low rate of drop out rate for the amount and types of samples tested.

24. In contrast with these data QHFSS concludes on page 63:

> *"The PowerPlex®21 system is a very sensitive amplification kit when used at either the standard amplification volume (25µL) or reduced volume amplification (12.5µL); however the increased sensitivity does not necessarily result in more reliable information."*

Also on page 63 QHFSS concludes that:

> *"The two sensitivity experiments explored the range on DNA template inputs from very large inputs (4ng) to very small inputs (0.00059ng). Within this validation complete PowerPlex® 21 DNA profiles were obtained with as little as 0.01875ng of template DNA. However, the PHR data indicate that as the amount of template DNA decreases the µPHR decreases and σPHR increases. The risk of type 2 errors is greatly increased from template DNA amounts of less than 0.132ng for both 25µL and 12.5µL total PCR volumes, which is supported by the experimental drop out data."*

25. The experimental data do not support either of these conclusions. Likely, the interpretation by QHFSS is driven by the bias of maintaining the 132 pg threshold. The latter statement on page 63 does not hold for 50 pg (even lower) and amounts above. There was little evidence of allele drop out at 50 pg of input DNA (or the Type 2 Error as defined by QHFSS).

26. Even mixture data generated by QHFSS support that a lower template amount could be applied. For example, in file "2. PowerPlex 21- Amplification of Extracted DNA Validation – 14.12.2012" (on page 7) QHFSS observed that:

> *"At a total DNA input template of 0.5ng, for 25µL and 12.5µL total PCR volumes, all alleles were detected for the mixtures with ratios of 1:1, 2:1 and 5:1."*

27. With an input amount of 0.5 ng, the minor contributor of a 5:1 mixture is at an input of 83.3 pg[5], which is less than the 132 pg threshold.

28. It is important to note that partial profiles are expected to be generated with forensic evidence and have been interpreted since the inception of forensic DNA typing. So, a criterion of no drop out would be inconsistent with international practices and would in effect toss out useful genetic information. Moreover, driving a threshold to a value that is excessively conservative and not in line with manual methods or with the capabilities of STRmix, does not make good use of genetic data, especially when some of the data are quite interpretable. Oddly enough in file "2. PowerPlex 21- Amplification of Extracted DNA Validation – 14.12.2012" (on page 48) QHFSS recognizes that:

[5] 0.5 ng = 500 pg. 500pg/6=83.3 pg.

> *"The homozygote threshold calculated in this validation will be used for extracted reference samples as case work samples do not require a homozygote threshold for STRmix™ analysis."*

29. It would seem that QHFSS did recognize that the "Type 2 Error" approach would not necessarily apply to casework when using STRmix™. Yet, the laboratory maintained the 132 pg threshold for its case workflow. Based on the overall validation studies performed by QHFSS, processing samples only if they have a minimum input of 132 pg is not best practice. QHFSS should reassess the data with all due speed and put into place a threshold (only as a guideline) that is concomitant with its validation data.

2.3 POST-IMPLEMENTATION REVIEW

30. QHFSS implemented a procedure for analysis of casework evidence using PP21 with half volume amplification reactions. To ensure that the process was performing adequately in file "2. PowerPlex 21- Amplification of Extracted DNA Validation – 14.12.2012" (on pg 65) QHFSS recommended:

> *"A post implementation review should be performed to review the appropriateness of points 3 – 8. The review will at minimum examine the outcomes of samples amplified within 0.0176ng/µL and 0.0244ng/µL concentration range. Similarly, all of the extracted reference samples will be reviewed with regards to the AITH and homozygote threshold."*

31. A similar recommendation is found in file "PowerPlex 21 Direct Amp of Ref FTA samples 26_09_2012" (on page 40) which stated:

> *"Six months after implementation of the PowerPlex® 21 system for direct amplification of reference samples; a review of the thresholds must be carried out to determine appropriateness of the thresholds set."*

32. Monitoring and following up on performance of a newly implemented procedure (a post-implementation review) to assess if it is performing as expected or if modifications are necessary is laudatory and part of a good quality system. However, while QHFSS recommended the subsequent reviews, it did not follow up in a timely fashion. It should be noted that a failure to follow up with such reviews is not unique to the validation of PP21. The same failure was observed with a recommended follow up for samples and data results based on the process changes associated with the Option Paper[6].

33. It was not until January 2016 (approximately 3 years after implementation) that QHFSS performed its post-implementation review[7]. On page 5 of the review document, QHFSS stated:

> *"The introduction of the PowerPlex® 21 (PP21) amplification kit into Forensic DNA Analysis in September 2012 (reference) and December 2012 (casework and extracted reference) was a significant change for the laboratory, as it involved changes to methodologies, sample processing*

[6] see report titled "Assessment of the Options Paper and Update Paper Prepared by Queensland Health Forensic and Scientific Services (QHFSS)" by Bruce Budowle, dated 19 September 2022.
[7] 11. Project Report 131 PP21 Post-implementation review – K.Scott, L.Ryan, C. Allen 01_2016

> *workflows and to aspects of reporting. Forensic DNA Analysis simultaneously introduced STRmix™ as the profile analysis software, this co-introduction made the changes for the laboratory complex. Due to the extensive nature of these changes, the Forensic DNA Analysis Management team decided that a post-implementation review 6 months after the implementation date would be of benefit. The PP21 post implementation review was to be completed under change management project number #131, however due to staff movements, and low staffing numbers in the Quality and Project team - the PP21 review was not completed at that time (in mid-2013).”*

and on page 6 stated:

> *“The value of a full review was reduced as many supplementary projects had been undertaken in Forensic DNA Analysis in the period September 2012 to December 2015, all of which had already assessed/reviewed aspects of PP21 use/application.”*

34. While it is appreciatied that staffing can impact some projects and deadlines, performing such reviews three years later poses a substantial risk and should not be tolerated. Moreover, justification that supplementary projects reduced the need for a comprehensive review also poses a substantial risk for identifying weaknesses, corrective actions, and proactively addressing limitations or errors that may arise. As noted by QHFSS the demand to implement PP21 and STRmix™ was a “significant change.”

35. Indeed, in its 2016 post-implementation review QHFSS notes on page 12 the value of a timely review:

> *“Where there is the introduction of new technology or where it is a large project - a process of review of data, workflows and reporting is completed. It is suggested that this should occur within 6-12 months post—implementation.”*

36. It should be expected that modifications are needed with most implemented procedures. Indeed, a notable change was the removal of using half volume amplification reactions within a few months of implementation of PP21 (implementation for casework was in December 2012). By February 2013[8] clearly there were concerns raised:

> “*Half volume reactions showed a high incidence of artefacts that had the appearance of allelic peaks above the reporting threshold. In samples with a low template contributor it was not possible to reliably assess whether such peaks were true alleles or artefacts”*

37. The document also notes:

> *“Post-implementation feedback suggested full-volume reactions should be revisited due to the difficulties in case management of half-volume samples”*

[8] see 47. Summary Half and Full Volume PP21 15_02_2013

38. It is surprising that QHFSS did not observe the difficulties with interpretation of half volume amplification reactions during its validation studies. Such variation would be predicted based on basic molecular biology, sampling, and enhanced sensitivity of detection. Additionally, the determination of the half and full volume variance performed by Jo-Anne Bright and John Buckleton (requested by QHFSS for STRmix™) indicated greater variance with half volume amplification results[9].

39. It may have been the potential cost savings that drove in part the acceptance of half volume amplification reactions[10] as opposed to consideration of interpretation challenges. Another factor may have been not including casework samples (or mock samples) in the validation study (see page 13 of 2016 post-implementation review) and not making more use of analysts in the validation studies, as analysts would have relevant experience on the challenges of data interpretation.

40. While implementation of half volume reactions is an example of failure to adequately assess the data generated during validation, it also demonstrates the importance of post-implementation reviews. This particular issue of half volume reactions became evident almost immediately and thus was addressed in part by cessation of the practice (what would have been well before a six-month review).

41. More critical, however, the failure to carry out a timely review did not contribute to laboratory improvement; a comprehensive review may have identified basic quality functions of the laboratory system, issues with data analyses of validation studies, human errors, etc. Recognizing the need for timely reviews post-implementation may have been identified three years earlier as a risk as opposed to the time of the 2016 post-implementation review. Indeed, the 3-second injection issue with CE instrument B, the error with stutter values (OQI 40636 identified 18/11/2015), and optimization of PP21 for mixture interpretation, are but a few issues that may have been identified earlier if a timely review was undertaken. Clearly, a root cause of the failures with properly validating PP21 should have been evident (as such a demand by the Australian and New Zealand Police Advisory Agency (ANZPAA) to validate PP21 simultaneously with STRmix™ in the short time frame required should have been questioned) and was eventually recognized by QHFSS in its 2016 post-implementation review on page 13:

> *"Simultaneous implementation of multiple new processes. In the case of PP21 it was implemented simultaneously with STRmix™. Both PP21 and STRmix™ were substantial projects in terms of impact on data analysis and interpretation within the laboratory. This co-implementation made initial troubleshooting more difficult, as it was challenging to determine which issues related to the new kits characteristics, and which issues related to the change in data analysis (with STRmix™). While there may be some instances where co-implementation is necessary, the laboratory should have a sound justification for this approach, as co-implementation*

[9] See Table 1, page 8 of file 1. Verification of the DNA Profile Analysis module of STRmix using the Promega PowerPlex 21 system - 12.2012

[10] see pages 29 and 63 in file 2. PowerPlex 21- Amplification of Extracted DNA Validation – 14.12.2012

does result in additional complications in the early adoption and implementations of technologies."

42. The conclusions reached in this report are based solely on the documentation made available to the author at the time of this report. If further information is made available, the conclusions may need to be revised.

3. STRMIX™ V1.05 VALIDATION

43. This review of key documents has been conducted as part of the Commission of Inquiry into Forensic DNA testing undertaken by the Queensland Health Forensic and Scientific Services (QHFSS). This portion of the report is a review of the validation of STRmix™ v1.05 undertaken by the QHFSS DNA laboratory in 2012.

44. Several documents related to this validation were provided. These included different iterations of the validation report, emails and meeting minutes related to the validation, and the STRmix™ v1.05 user manual.

45. Specifically, the Commission has requested the following:

a. Whether, and why, the performance of the validation was scientifically sound and consistent with international best practice (including consideration of issues including the experimental design employed, the experiments conducted, the analysis of those experiments, statistical methods employed, the report and dealing with any feedback received, as well as particular issues raised below); and

b. To what extent, if any, any deficiency in the above two characteristics could have or did have an impact on:

i. Whether the methods, systems and processes for forensic DNA testing and analysis in place at the laboratory were or are reliable;

ii. Whether the methods, systems and processes for forensic DNA testing and analysis in place at the laboratory were or are in accordance with international best practice;

iii. Whether the methods, systems and processes for forensic DNA testing and analysis in place at the laboratory would or have resulted in accurate reporting of results and accurate matching.

46. The conclusions reached in this report are based solely on the documentation made available to the author at the time of this report. If further information is made available, the conclusions may need to be revised.

3.1 BACKGROUND

47. At the direction of the Australian and New Zealand Police Advisory Agency (ANZPAA) all Australian jurisdictions were expected to implement by the end of 2012 one of the larger DNA profiling kits, such as PowerPlex® 21 (PP21) or Globalfiler®, and the probabilistic genotyping software STRmix™.

48. It should be acknowledged at each of these tasks - validating a new profiling kit or validating software that completely changes how forensic DNA profiling results were interpreted and reported – involves a considerable amount of work. Ideally, the changes would be implemented sequentially allowing laboratory staff time to fully understand and find ways of dealing with the consequences of the first change before embarking on the second. That both these changes had to be accomplished concurrently placed an enormous burden on the Australian laboratories that should not be underestimated.

49. STRmix™, developed by researchers from the Institute of Environmental Science and Research (ESR) and Forensic Science South Australia (FSSA), is a software tool to enable forensic DNA analysts to interpret DNA evidence that was previously thought to be too complex. This software became available during 2012 and while the Australian laboratories had access to earlier versions it was v1.05 that was validated for Australian casework[11].

50. In addition to having access to earlier versions of the software, at least one person from each of the Australian laboratories, including QHFSS, undertook 'Train the Trainer' instruction in Melbourne in July 2012.

51. Version 1.05 of the software was rolled out to the laboratories during the first two weeks of November 2012.

52. Unless stated otherwise, this review focusses on the first iteration of the QHFSS STRmix™ validation report which was signed off in mid-December 2012[12]. The first validation report contained and compared data from both half-volume and full-volume amplifications.

53. Addendums to the original validation report were issued in early 2013. An addendum dated February 2013 was specifically related to low template samples. In March 2013 two additional addendums were issued, one related to full-volume amplifications and another related to half-volume amplifications. The low template and full-volume addendums specifically incorporate re-interpretation of mixed DNA profiles after a process change altered how peaks in double-back and forward stutter positions were dealt with during the initial analysis of profiles. These analysis changes had little bearing on the STRmix™ validation as a whole. The half-volume addendum appears to be the same as the original validation document but with only data from half-volume amplifications[13] included in the report.

54. Over the past few years, the STRmix™ team have developed extensive documentation and tools to support laboratories in their validation activities. Furthermore, in 2015 the Scientific Working Group on DNA Analysis Methods (SWGDAM) published guidelines[14] for the validation of probabilistic genotyping systems such as STRmix™. These documents now provide a framework for the types of experiments and data required to validate the software. Guidance such as this helps establish internationally accepted best practice.

55. However, in 2012 the Australian laboratories were relying on guidance provided directly by the STRmix™ developers. Rather than centralised in one document, this guidance seemed to have been dispersed in emails, often addressing particular questions. The STRmix™ v1.05 User Manual includes some information related to the running of Model Maker, a critical step in validating STRmix™, however this information in no way constitutes as providing instructions on how to carry out a complete validation. Furthermore, how to validate STRmix™ does not appear to have been discussed in any great depth in the July 2012 Train the Trainer workshop.

56. Questions raised in an email[15] dated 22 August 2012, from QHFSS to representatives of laboratories across Australia and New Zealand indicated that there may have been

[11] Although FSSA had validated an earlier version of STRmix™ in August 2012.

[12] Document 1. Verification of the DNA Profile Analysis Module of STRmix using the Promega PowerPlex 21 system – 12.2012.pdf.

[13] Perhaps this document was created as half-volume amplifications were ceased in February 2013, allowing the full-volume addendum to stand alone.

[14] https://www.swgdam.org/_files/ugd/4344b0_22776006b67c4a32a5ffc04fe3b56515.pdf

[15] Email 23082012.pdf

plans for each laboratory to contribute to a national validation document. In this email chain a QHFSS staff member asked what each lab was planning to do to validate STRmix™, indicating that there was, at least at this point in time, no clear guidance on this matter.

57. In an effort to gain consistency across the Australian laboratories, members of the Biological Specialist Advisory Group[16] (BSAG) were discussing and directing some aspects relating to the use of STRmix™. It is difficult to determine from the documentation provided to what extent, if at all, the BSAG discussions shaped the validation activities in the Australian laboratories.

58. Another email[17] dated 3 September 2012 indicates that a STRmix™ project steering committee was in place to coordinate the validation and implementation of STRmix™ into the Australian laboratories. This email outlined the deployment of STRmix™ v1.05, future developments and some information about the release of the user manual. The email also noted that it would be up to BSAG "…to determine what additional testing/verification may be required."

59. As there was no formal validation guide, this review of the QHFSS STRmix™ v1.05 validation has been undertaken based on the details recorded in the validation document, the STRmix™ v1.05 user manual, emails which include those from the STRmix™ developers, and the author's own memory, knowledge and experience of various versions of STRmix™, including v1.05. Given that it is almost a decade later, it is possible that not all guidance provided by the STRmix™ developers or BSAG was included in the information made available for this review.

60. The author has also informally asked members of the STRmix™ team specific questions related to STRmix™ v1.05 functionality and validation requirements, and advice that was provided in 2012, particularly in relation to drop-in modelling.

3.2 INTRODUCTION

61. Unless stated otherwise, the QHFSS STRmix™ validation document referred to in this review is called 'Verification of the DNA Profiling Analysis module of STRmix using the Promega PowerPlex 21 system – 12.2012.pdf'.

62. In general, the QHFSS STRmix™ v1.05 validation appears to have been competently undertaken and demonstrated a good understanding of the software.

63. It is clear from the documentation provided that STRmix™ team members assisted with some critical aspects of the validation based on the data provided by the laboratory. These included:

    a. Choosing an appropriate peak variance constant to be used in STRmix™

    b. Checking Model Maker results

    c. Providing drop-in parameters

    d. Providing advice on how to calculate the saturation limit

    e. Providing a stutter ratio file based on inter laboratory data

[16] BSAG members include representatives from each of the Australian and New Zealand laboratories.
[17] Email 04092012.pdf

64. How QHFSS determined some of the key tasks required in a STRmix™ validation is discussed in the following paragraphs.

3.3 PEAK VARIANCE AND LSAE[18] VARIANCE

65. These two parameters are modelled by STRmix™ using its Model Maker functionality. Model Maker is run as part of the STRmix™ validation and the results are subsequently incorporated into the software and used in the interpretation of DNA profiles.

66. Based on the information in the validation report and communication between QHFSS staff and STRmix™ developers, it appears that this part of the validation was done correctly, and the values seem appropriate for the profiling kit and CE instrument. Variances were calculated separately for data from full-volume and half-volume amplifications. The variance for the half-volume reactions were larger which was expected given increased variability in peak heights.

67. The performance of the variances was tested using six 2-person and 3-person mixed DNA profiles. This seems to be a small set, compared to what would be tested now for a new version of STRmix™. However, the testing undertaken by the QHFSS laboratory was likely based on advice from the developers which was

> *"We are suggesting you run a few profiles through to test your variance value before launching into casework*."[19]

68. The peak variance decisions made based on the mixture testing that was undertaken were appropriate.

3.4 BACK STUTTER

69. Back stutter is a common artefact of the polymerase chain reaction (PCR) that can be modelled in STRmix™. A back stutter peak is an artefact associated with an allele[20] and there is usually a predictable relationship between the position and height of an allele and the position and height of its associated back stutter peak.

70. There are other forms of stutter, including double back and forward stutter, however these occur at a lower frequency than back stutter. Back stutter was the only form of stutter that early versions of STRmix™, including v1.05, modelled.

71. The STRmix™ developers collated and compared the back stutter ratio data generated by the Australian laboratories using PP21. The developers' analysis[21] of the inter laboratory data, which included data from QHFSS, determined that as there was very little difference between the laboratories' datasets, a common set of stutter ratios could be used by all the laboratories provided they were using the same DNA profiling kit (PP21) and 30 PCR cycles. Given the findings of the inter laboratory study, it is appropriate that the QHFSS laboratory adopted this common stutter ratio file into its STRmix™ v1.05 configuration.

[18] LSAE=locus specific amplification effect
[19] Email 13112012.pdf
[20] An allele is a DNA profiling result.
[21] This work was published: J-A. Bright, J.M. Curran, Investigation into stutter ratio variability between different laboratories, Forensic Science International: Genetics, 13, (2014), pp. 79-81.

3.5 SATURATION

72. Saturation is a term used when too much DNA is present causing the camera in the capillary electrophoresis (CE) instrument to be overloaded. When the height of an allele exceeds the saturation limit the normally predictable relationship between the height of the DNA profiling result and the associated stutter peak breaks down. The saturation value is important as STRmix™ relies on the predictable relationship between allele height and associated stutter peak height for profile interpretation.

73. While the method used by QHFSS to determine the saturation value was correct, the data used in this experiment appear to not have included profiles that approached and exceeded the saturation limit. In order to determine the saturation limit of an instrument the experimental data set should include some samples overloaded with DNA, and this requirement certainly could be inferred from the advice provided by one of the STRmix™ developers[22]. Section 4.1 of the QHFSS STRmix™ validation document, which outlines the experimental design for saturation, states:

> "*The observed peak heights and observed stutter heights of between approximately 100 and 450rfu (dependant on locus data) were recorded.*"

If these values are correct, then the samples used simply do not contain enough DNA to accurately determine the saturation limit. This lack of appropriate data was confirmed in section 5.1 which noted that there were no data that deviated from the expected relationship between stutter and DNA profiling result peak heights. The data presented in Table 6 (page 12) appear to present lines of best fit in table form between observed and expected stutter heights at each locus when they ought to have assessed at what point there is a deviation from this linearity.

74. The laboratory ultimately settled on a saturation value of 7000rfu[23] because

> "At the 7000-8000rfu heights, the DNA profiles had a tendency to demonstrate the effects of excess template and often possessed poor baseline integrity."

While this reasoning has nothing to do with the effect of DNA template on the relationship between allele and stutter peak heights, the 7000rfu value is a saturation value commonly determined by laboratories using 3130 CE instruments.

3.6 DROP-IN CAP AND PARAMETERS

75. The drop-in cap and parameter were based on data from the PP21 validation. Drop-in refers to fragments of DNA present in the laboratory environment that are inadvertently introduced into samples. Drop-in peaks may be detectable, especially if sensitive profiling techniques are used. Unlike contamination, drop-in[24] is not reproducible and is characterised by the presence of just 1 or 2 peaks. It should also be a rare event and laboratories that employ sensitive profiling techniques monitor drop-in rates as a general laboratory 'health check'[25].

76. The STRmix™ software is able to model the drop-in events that may be observed in a laboratory using a combination of parameters that are dependent on setting a drop-in cap and estimating the rate of drop-in. In version 1.05 of STRmix™ the cap is a peak

[22] Email 12112012

[23] rfu=relative fluorescent units, the unit of measure of peak heights in electropherograms.

[24] It can be difficult to determine what is contamination and what is drop-in. However, drop-in is usually determined from negative controls where there are no more than 2 or 3 peaks. More than this is normally classed as gross contamination.

[25] Gill P, Whitaker J, Flaxman C, Brown N, Buckleton J. An investigation of the rigor of interpretation rules for STRs derived from less than 100 pg of DNA. Forensic Science International 2000;112;17-40.

height whereby a peak up to and including that height in an evidence input file[26] may be proposed as possible drop-in during a STRmix™ interpretation. The cap can be calculated in different ways. For example, if there are several occurrences of drop-in, a laboratory could use the average peak height of the drop-in peaks plus 3 (or more) standard deviations as the cap. If very few drop-in events are observed, then a laboratory may choose some value that exceeds the tallest peak height of the observed drop-in peaks.

77. STRmix™ v1.05 could model drop-in in two ways. The first was using a constant probability where the probability of being drop-in of any peak equal or below the height of the drop-in cap was the same. The constant probability model could be used when very few drop-in peaks were observed in the data. The second model used an exponential probability distribution where the probability of drop-in decreases as the height of the peak approaches the drop-in cap. This model could be used when more drop-in data were available.

78. One way to collate drop-in data is to use experimental data associated with the validation of a new DNA profiling kit. Initially, this may be the only way to collect these data. Once a DNA profiling kit is implemented, drop-in data can continue to be collected and reviewed from time to time. If increased rates of drop-in are observed, then a laboratory can amend the drop-in parameters. If the increase is significant, then a laboratory should consider addressing its anti-contamination procedures.

79. QHFSS used data from its PP21 validation to inform the STRmix™ drop-in parameters. The experimental design was captured in section 5.6 of the PP21 validation document[27] and was appropriate for this study. There is one very small error in this section as it stipulates that only peaks above 20rfu were considered as possible drop-in, but the threshold must have been lower – perhaps 15rfu. Section 6.5 details the results of the drop-in study. Three drop-in peaks were observed, the tallest of which was 21rfu and the remaining peaks were both 19rfu.

80. As all the observed instances of drop-in were below the laboratory's analytical threshold, the drop-in cap was set to 40rfu, which is the same as the laboratory's analytical threshold. While this decision is not unreasonable given the data, it did mean that only peaks in an evidence input file that were exactly 40rfu in height could be proposed as drop-in by the STRmix™ software.

81. Drop-in parameters using the exponential probability distribution[28] method were calculated by the STRmix™ developers for each of the Australian laboratories that observed drop-in[29].

82. According to the STRmix™ validation document, QHFSS implemented a 40rfu cap and the drop-in parameters provided by the developers.

[26] This is the data file STRmix uses to perform its interpretations. It includes the heights of all the alleles and associated stutter peaks observed in a DNA profile.

[27] Document 2.PowerPlex 21 – Amplification of Extracted DNA Validation – 14.12.2012.pdf

[28] The author asked one of the developers about this choice given the constant probability method would be more commonly used when drop-in data is scarce, however the developer could not recall why they chose the exponential probability method. (Pers Comms J-A Bright, 9 November 2022.)

[29] The STRmix™ developers summarised the parameters for the laboratories in the document 'Drop-in modelling II_final.pdf'

83. However, in the standard operating procedure[30] (SOP) the drop-in cap was 100rfu and parameters differed from those provided by the STRmix™ developers. It is possible that these were transcription errors. A later version[31] of the SOP is correct.

84. During the validations of PP21 and STRmix™ v1.05 there was an error in the run conditions of one of the QHFSS CE instruments. This error was discovered some months after PP21 and STRmix™ v1.05 were implemented and the subsequent investigation was documented in OQI34817[32]. As some of the validation drop-in data were obtained from results generated by the CE instrument with the incorrect setting, the samples were run again on the instrument with the correct run setting.

85. The findings of that reanalysis are discussed at length in section 5.5 of the 'Review of DNA analysis undertaken in the Blackburn case' report and will not be repeated in this document. There is evidence in the documentation provided that drop-in was occurring more frequently than determined by the initial PP21 validation data and at peak heights greater than the implemented cap of 40rfu. The documentation also suggests that another confounding issue, carryover, was also causing difficulties with analysing drop-in data. Carryover is discussed further in sections 5.5 and 5.6 of the 'Review of DNA analysis undertaken in the Blackburn case' report.

86. The increased amount of drop-in detected in the reanalysis of the original validation data, in conjunction with increased drop-in events recorded through ongoing monitoring after PP21 was implemented, are cause for concern.

87. Opportunities to adjust the drop-in settings in STRmix™ v1.05 were not taken. Furthermore, drop-in was not mentioned at all in either of the STRmix™ v2.0.1 validation reports[33]. An email dated 9 December 2014[34] suggests that a formal re-evaluation of the drop-in cap and parameters had still not been undertaken.

88. While the effect of not reassessing the drop-in cap and parameters is likely minor for the majority of samples and cases, there are ramifications for the interpretation of low-level peaks. This is especially true given the laboratory's practice of calculating and reporting likelihood ratios for very trace results. In effect, STRmix™ would have been underestimating the probability of drop-in for low-level peaks.

3.7 MIXTURE INTERPRETATION

89. As part of the validation, the profiling results obtained from DNA samples containing different ratios of DNA from known contributors were interpreted (or 'deconvoluted') by the STRmix v1.05 software, and likelihood ratios (LRs) were generated. The interpretations and LRs were then reviewed to determine if they produced expected results and intuitively made sense. Reproducibility of the results was also tested.

[30] SOP 31523V1 – Interpretation and Statistical Analysis of DNA profiles Using the STRmix Expert System – 14.12.2012.pdf at page 20
[31] SOP 31523V2 – Use of STRmix Software - 10.11.2014.pdf
[32] OQI stands for 'opportunity for quality improvement'.
[33] Verification and Implementation of STRmix Version 2.0.1 – 06.2014, V0.1 & V0.2.
[34] Document '30. Corro – Drop in parameter – 09.12.2014.pdf'

90. The samples used in the mixture study[35] contained total DNA templates of 0.06ng, 0.125ng, 0.250ng and 0.500ng, from 2, 3 or 4 contributors in a variety of contributor ratios.

91. The 4-person mixture interpretations either failed because of computer runtime issues or the analyst was not able to competently review the results. For these reasons the validation recommended that no 4 person mixtures be interpreted using STRmix™. Setting the issue of computer runtime limitations aside, the decision to not include 4 person mixtures was perhaps premature. Not all 4 person mixtures are inherently complex, depending on the contributor proportions. That the results from 4 person mixtures were difficult for the analyst to competently review perhaps reflects the limited training provided by the STRmix™ developers during the Train the Trainer workshop.

92. This study found that for the 2 and 3 person mixtures where the total DNA template was 0.500ng, the interpretations and LRs were intuitively correct and reproducible.

93. For the samples with total template of 0.250ng or 0.125ng, variability was sometimes seen in the interpretations and LRs for the donor of the smallest amount of DNA in the mixture. For instance, in the reproducibility study at page 17 the laboratory identified that when a two-person mixture with a contributor ratio of 50:1 was repeatedly deconvoluted and compared to the DNA profiles of the 2 known contributors, the LRs generated for the known trace minor contributor were close to 1, either weakly supporting inclusion or exclusion. In this mixture, the trace minor donor contributed a theoretical 4pg of DNA to the mixture. The report correctly explained that this variability was not unexpected and was largely due to the loss of allelic information from this contributor.

94. Appendix 1 of the STRmix™ validation document presents the results of each of the mixture interpretations. For some of the mixtures with total template of 0.125, the LRs for the donor contributing the least amount of DNA are sometimes close to 1, and sometimes in support of exclusion. Again, this is to be expected and it is important to note that this is not an example of STRmix™ getting the wrong answer. The program is simply using the DNA profiling information that is available to inform the interpretation. When an LR close to 1 is generated, it quite correctly reflects the weakness or insufficiency of the DNA results being interpreted. For the major contributors to these mixtures, STRmix™ produced intuitively correct and reproducible effects.

95. The validation report noted[36] that in at least one instance an exclusion occurred because the data that had been input into STRmix™ was incorrect. It is unclear if the interpretation was repeated with the correct profiling data.

96. The validation report at page 17 concluded that the low and variable LRs generated for the trace donors provided evidence that when DNA template was as low as 0.1ng – 0.150ng profiling results may not be reliably interpreted. Although this conclusion is actually erroneous, it was used to provide support for recommendation 3 from the PP21 validation that any sample containing a DNA concentration of less than 0.01ng/µL[37] would not be routinely processed because profiling results would likely

[35] Somewhat confusingly, these are described in the PP21 validation document, not the STRmix™ validation document. See table 8, pg 24 of document '2. PowerPlex 21 – Amplification of Extracted DNA Validation – 14.12.2012.pdf'

[36] At section 5.5, page 16.

[37] For a full-volume amplification, this equates to approximately 0.150ng total DNA template.

exhibit stochastic effects[38]. That samples with less than 0.01ng/µL DNA concentration would not be processed was also given as a reason to not undertake further exploration of low level DNA mixture interpretation during the validation.

97. The presence of stochastic effects can complicate profile interpretation, but low-level DNA profiles may still be interpretable. STRmix™ is designed to manage the variability caused by stochastic effects in low level DNA profiles. The mixture interpretation and reproducibility studies conducted by QHFSS demonstrated that the DNA profiles of the major DNA contributors, and sometimes those of the minor contributors, could be reliably interpreted from samples with total DNA of 0.06ng to 0.125ng thus demonstrating that the 0.01ng/µL (or 0.150ng total DNA) threshold was too conservative.

98. Whether or not the profile of a minor contributor can be interpreted is largely based on the DNA template of that contributor in comparison to the remaining contributors, rather than the DNA template of all the contributors combined. By focussing on the interpretability of only the DNA profiles of the very trace DNA donors to a mixture, the laboratory failed to recognise that valuable information may be derived from the profiling results of the remaining DNA contributors.

99. It was remiss of the STRmix™ v1.05 validation process to use the PP21 0.01ng/µL threshold recommendation as a reason to not explore the interpretation of mixed DNA profiles at lower total DNA templates. This additional exploration may have assisted in determining a more reasonable threshold[39] or reassess the appropriateness of employing such a threshold, particularly for major crime samples.

100. Implementing a process that did not routinely profile any case sample with a DNA concentration below 0.01ng/µL created the great risk of not detecting potentially probative, exculpatory or otherwise informative profiling results.

3.8 FURTHER COMMENTS

101. The Train the Trainer workshop may not have been entirely adequate. It is understood that only one person from QHFSS attended and this person was subsequently responsible for validation, documentation, implementation and training. This demand was a considerable responsibility for one person, especially for an entirely new and complex software package for which formalised validation guidelines were not available. This comment is not intended as a criticism of the work undertaken by this QHFSS staff member, but rather a recognition of the burden carried by that staff member.

102. QHFSS had access to earlier versions of STRmix™ and had undertaken some preparatory validation work during the second half of 2012. However, some key tasks, such as the determination of the peak variance and the testing of mixture interpretations, could not be undertaken until STRmix™ v1.05 was deployed in early November 2012. The validation was signed off in mid-December which seems rather rushed, although perhaps possible if a staff-member was devoted to the validation full-time.

[38] Increased stochastic effects are common in low-level DNA profiles and include phenomena such as peak imbalance and drop-out (absence) of some or many DNA profiling results. Determining the minimum number of contributors can also be difficult.

[39] The validity of the 0.01ng/µL threshold and whether or not implementing such a threshold is appropriate for all crime sample types is discussed in section 3.5 of the 'Review of the Blackburn case' document.

103. An email to QPS dated 06 March 2013[40] described the significant burden that implementing both PP21 and STRmix™ v1.05 had placed on the laboratory during the previous year. This email noted, perhaps with some pride, that, “To date, Queensland is the only jurisdiction to implement these huge changes within the allocated timeframe (all other jurisdictions expect to implement by the end of March 2013).” In retrospect, the rush to implement both PP21 and STRmix™ by the end of 2012 meant that opportunities to gain a better understanding of the capabilities of STRmix™ were lost.

104. Had more time been taken, further mixture interpretations, including lower template profiles, may have been undertaken. Extra time may have allowed those staff members involved in the validation to consider more fully the results obtained from the STRmix™ interpretations and in turn better inform the setting of thresholds and other policy decisions. Further training should have been sought, both for the personnel responsible for the validation and implementation of STRmix™ and for reporting scientists who had to use STRmix™ and understand and communicate its results.

105. The conclusions reached in this report are based solely on the documentation made available to the author at the time of this report. If further information is made available, the conclusions may need to be revised.

[40] Document 25. Corro – Update – 06.03.2013