Review of the validation material from the Queensland Health Forensic and Scientific Services (QH)

Requested by Commission of Inquiry

into Forensic DNA Testing in Queensland

Prepared by Duncan Taylor

07 October 2022

1 Contents

1.0 Purpose:

I have been asked by the Commission of Inquiry into Forensic DNA Testing in Queensland (hereafter 'the Commission') to review a number of validation reports produced by the Queensland Health Forensic and Scientific Services (QH). The specific tasks are to comment on whether the validations were scientifically sound, whether appropriate experimental design and data analysis were used, whether the validations met best practice, and ultimately whether the findings may have impacted on the reliability of the methods as implemented in QH. In particular for two validations (Quant Trio and Quant Studio) I have been asked to provide commentary on the statistical techniques used in the data analysis. In addition, the Commission have provided me with a number of comments that raise issues with the validations that I am to take into consideration. To assist me in this task the Commission have also provided a number of supporting documents and communications (listed in Appendix I). I provide my Curriculum Vitae in Appendix II.

Two questions are of importance with this review:

1. Has the validation been conducted using scientifically sound and appropriate statistical principles and according to best practise

2. If not, has there possibly been an adverse outcome, such as unreliable results being produced, or opinions being given

I will address these for the individual validations below.

2.0 Executive Summary:

The following provides a general summary of the findings in relation to the validation reports that have been reviewed. For more detail see the individual sections in the main body of the report

- Validation of Quantifiler® Trio
    - Scientific soundness of validation

        Many instances the statistical techniques used, and the ways the data were presented, are not the best methods that could have been chosen. There were instances where Student's t-tests were performed inappropriately (suggesting a misunderstanding of the limitations and assumptions of the technique). There were also a number of instances where published examples of equivalent data analysis could have been followed that would have aligned the validation with accepted practise.

- Reliability of results

  In general, the appropriate laboratory work was carried out, and the main issue lies with the way the data was presented or analysed. In many instances the experiment only called for a graphical display of the data rather than a (misapplied) statistical technique. Looking past the analysis and more at the data itself, I believe that the decisions made throughout the validation and the recommendations would have been the same had appropriate statistical analysis been carried out. I do not believe there is any evidence to suggest the manner in which Quantifiler was implemented into casework is unreliable or would lead to unreliable quantification values being produced.

  For some experiments, such as repeatability and reproducibility, the experimental deign was sub-optimal and this led to an assessment of variability that falls below best practice. An understanding on the variability of results can be important when describing results to stakeholders. At worst, if the lack of knowledge of variability is misunderstood as meaning there is a lack of variability itself then this could lead to inaccurate advice or information being given.

  Another issue with the experimental design is that the validation was not adequate to appropriately calculate a limit of detection (LOD). Some additional laboratory work is required to calculate this LOD according to best practice. This is an important aspect of the validation to complete if the LOD is being used as a decision point either for laboratory processes or reporting. It is also important to make sure staff understand the meaning of a LOD. If either the LOD is not appropriately calculated, or its meaning is misunderstood then this has the potential to lead to unreliable decisions being made or information being given.

  If a LOD value is going to be used as a decision threshold, then until its value has been appropriately calculated all quantified DNA samples should be treated (with respect to decision making or laboratory processes) as though they have exceeded the LOD.

- Validation of Quant Studio 5
  - Scientific soundness of validation

    Many of the same experiments were carried out in this validation as they were for the Quantifiler Trio validation. Therefore, again there were instances where inappropriate statistical tests carried out in all sections.

One exception is that there was a lack of appropriate experimental design for the repeatability and reproducibility components of the validation.

- Reliability of results

  As with Quantifiler Trio, I do not think this validation provides any evidence to suggest quantification values being produced by the QS5 are unreliable.

  As the use of QuantiFiler Trio and Quant Studio 5 are linked (in that they both must work in tandem to produce quantification results) the calculation of the LOD value will necessarily include Quant Studio 5, but there is nothing further to add here.

  The sub-optimal repeatability and reproducibility experiments result in a lack of knowledge about the variability in quantification values produced. Such a lack of knowledge can lead to an inability to answer questions of importance and has the potential to be misunderstood as a lack of variability itself. This could lead to incorrect advice being provided, or decisions being made.

  Finally, in experiment 3b of the validation an outlying quantification result was obtained. It would have been beneficial to carry out some additional troubleshooting or to analyse some additional samples to ensure this was not an ongoing, albeit sporadic, issue. As the instrument has been in use for some time, it would probably be possible at this stage to retrieve data from already run control samples to show that the cause of the outlying result is not an ongoing issue.

- Verification of 3500xL B
  - Scientific soundness of validation

    As with previous validations, there are instances of statistical tests being used that are not (the most) appropriate for the data. A point to consider for follow-up is whether the homozygous threshold chosen in the validation has the intended level of coverage, and this should be checked by the laboratory.

  - Reliability of results

    As with previous validation, given the work that has been carried out (even if not analysed according to best practice) there is no evidence to suggest DNA profiles produced by this instrument are unreliable or that they would lead to unreliable opinions. The main point to follow-up on is the actual coverage of the homozygous threshold chosen. This is particularly important to the reliability of decisions if they are being

made based on the probability of peaks reaching this threshold, or on the intended coverage.

- Validation of 3500xL Analysis of Casework PowerPlex®21 WEN
  - Scientific soundness of validation

    As with previous validations, there are components that have not used formally appropriate statistical methods. Also, in parts I believe the validation has not been large enough to show the range of reproducibility of results. There is also the same comment with regards to the setting of the homozygous threshold as for the previous validation, and again this should be checked by the laboratory.

  - Reliability of results

    As before, there is no evidence to suggest unreliability of DNA profiles being produced, but the homozygous threshold should be followed up on to ensure reliable decisions are being made and reliable information given.

- Assessment of 3500xl A Genetic Analyzer for Processing Casework Powerplex® Samples
  - Scientific soundness of validation

    This validation was carried out quite thoroughly and demonstrated good performance over a range of samples. It also included the most comprehensive statistical analysis of any of the validation reports I reviewed (of the baseline fluorescence level).

  - Reliability of results

    There is no evidence that unreliable DNA profiles are being produced, and good indication that the performance of the instrument / process has been demonstrated.

- 3500xL Genetic Analyzer Validation for Reference samples Amplified with Powerplex®21 using Direct Amplification
  - Scientific soundness of validation

    There are a number of aspects of this validation that have not been carried out in an appropriate manner. In particular the assessment of sensitivity, repeatability and reproducibility have not been carried out appropriately.

- Reliability of results

  There is no evidence that unreliable DNA profiles are being produced, however there are aspects of this validation that warrant revisiting to ensure a proper understanding of the instrument performance is gained. As before, without an adequate knowledge of the performance of the instrument there is a risk that unreliable decisions can be made, or information provided.

  There is a recommendation in the report that certain aspects of the validation are revisited post implementation to ensure fitness of the thresholds chosen in the validation I would recommend a check is made that these post-implementation revisitations have occurred.

- 3500xL Genetic Analyzer Validation for Extracted Reference Samples Amplified with PowerPlex®21. Forensic DNA Analysis
  - Scientific soundness of validation

    As with previous 3500xL validations there are various aspects that have not been carried out according to best practice.

  - Reliability of results

    There is no evidence that unreliable DNA profiles are being produced.

    There is a recommendation in the report that certain aspects of the validation are revisited post implementation to ensure fitness of the thresholds chosen in the validation I would recommend a check is made that these post-implementation revisitations have occurred.

- 3500xL Genetic Analyzer Validation for Casework Samples Amplified with PowerPlex®21. Forensic DNA Analysis
  - Scientific soundness of validation

    As with previous 3500xL validations there are various aspects that have not been carried out according to best practice.

  - Reliability of results

    There is no evidence that unreliable DNA profiles are being produced.

    There is a recommendation in the report that certain aspects of the validation are revisited post implementation to ensure fitness of the thresholds chosen in the validation I would recommend a check is made that these post-implementation revisitations have occurred.

- Optimisation of the cleaning protocol for bone crusher vials

- Scientific soundness of validation

  This is quite a specialised and focussed validation and appears to be appropriately performed

- Reliability of results

  The conclusions are sound and there is no evidence that unreliable results are being produced

- Validation of Hamilton STARlet A for Quantification and amplification Assay Setup
  - Scientific soundness of validation

    In many experiments there was only a graphical display of results and no formal statistical testing, however I believe this is an appropriate assessment of the results. In my opinion, the validation of this instrument was appropriate.

  - Reliability of results

    There is no evidence that unreliable results are being produced

- Verification of Hamilton STARlet B for Quantification and Amplification Assay Setup
  - Scientific soundness of validation

    This is much the same as validation of the STARlet A. I believe this is an appropriate assessment of the results. In my opinion, the validation of this instrument was appropriate.

  - Reliability of results

    There is no evidence that unreliable results are being produced

- Validation of Hamilton® STARlet C for Capillary Electrophoresis Setup
  - Scientific soundness of validation

    I believe this is an appropriate assessment of the results. In my opinion, the validation of this instrument was appropriate.

  - Reliability of results

    There is no evidence that unreliable results are being produced

- Verification of ProFlex™96 Well PCR System using PowerPlex®21
  - Scientific soundness of validation

In modern DNA profiling the performance of the PCR process (including the performance of the thermocycler) are particularly important, as they can affect DNA profile quality in subtle ways that are important to probabilistic genotyping systems such as STRmix. There has been some evidence published that individual thermocyclers of the same model do not perform different with respect to the profile quality, and that there is some robustness to STRmix to small changes in DNA profile quality. In my opinion the ProFlex validation is inadequate to demonstrate that the performance of the instruments was similar enough to each other, and the previous 7500 that they can be considered in one group.

Some additional laboratory work to show the repeatability of the individual instruments, and the use of STRmix in this validation would have been useful. I note in the introduction of the validation it states it was "*not to assess the suitability of the current analysis and profile interpretation thresholds*", however STRmix would still have been a useful tool to assess and compare performance.

There are also some misunderstandings about the meaning of statistical test results (in this case R^2 values) that have led to a conclusion of similarity between instruments that may not be warranted.

- Reliability of results

While there is no evidence that the ProFlex systems are unreliable, there is a deficiency in the experimental design meaning there isn't sufficient data available to provide much support for the reliability of the results being produced by the ProFlex instruments either. I believe some additional laboratory work and data analysis would have been beneficial in this validation. A check should be conducted to ensure that the misunderstanding of R^2 values have not led to the conclusion that all instruments were performing similarly, and that additional testing has been carried out to confirm this fact.

There is a risk of unreliable results being produced and reported (ultimately being reflected in the likelihood ratio produced to QPS) if there is an undiagnosed divergence in performance between the ProFlex instruments.

However, I do not believe a suspension of laboratory functions are required whilst this additional validation work is being carried out. My opinion is based on three factors:

1) The current STRmix settings appear to be based on a combination of data from all ProFlex instrument and so will be somewhat

representative of their grouped average performance. This minimises the difference in modelled peak height variability that might be present in of any one instrument from that of the combined group

2) Published work by Kelly et al [1] show that STRmix is robust to small differences in profiling performance that might occur from instrument-to-instrument differences. Therefore, a small difference between instruments is likely to have a minimal effect on STRmix analyses in most cases.

3) I assume that at some point any STRmix analyses that have been carried out on DNA profiles produced using these ProFlex instruments are reviewed by an analyst (and possible secondarily as part of a case peer review). If there were dramatic issues with the performance of one ProFlex instrument, then I would expect this to be noticeable in the results of STRmix analyses. There are numerous diagnostic features in the STRmix output, which were neatly summarised by Russell et al [2], and also in STRmix user's and operation manuals. Implicit in this third point is that analysts are reviewing STRmix outputs, and that they have the appropriate training and understanding of STRmix functionality and DNA profile interpretation.

- Validation of the QIAsymphony® SP/AS Modules
  - Scientific soundness of validation

    Much of this validation examines factors where the purpose is to choose the best out of several options (or to demonstrate that a process is possible). These sorts of examinations often do not need extensive formal statistical evaluation. In general, many of the experiments would have benefitted with a more informative summarisation of the data (particularly graphically), and some would have benefitted from additional samples, however there is no evidence to suggest that unreliable results are being produced from this instrument.

  - Reliability of results

    There is no evidence to suggest that unreliable DNA extractions are being performed from this instrument, or that unreliable results would be provided in opinions derived from this instrument.

- Validation of QIAsymphony SP for bone extraction
  - Scientific soundness of validation

There are some instances where inappropriate use of statistical tests may have been carried out. However, I think all the information needed to make the decision could be obtained from just the graphical display of results that was given by the authors, and I do not believe any issues with the statistical tests have affected the conclusions.

- Reliability of results

In my opinion that the methods chosen were done so appropriately and there is no evidence to suggest that unreliable results would be produced from this instrument. The size of the experiments carried out was on the small size, but I think adequate to assess the DNA extraction process.

3.0 Introduction:

Validation is the process of providing objective evidence that a method, process or device is fit for the specific purpose intended [3]. For forensic laboratories, validation is carried out so that the laboratory can:

- have confidence that the method, process or equipment they wish to implement performs at a desired level,
- identify the limits of use or performance, and from those limits devise appropriate interpretation guidelines and implementation strategies, and
- demonstrate to external groups, such as Courts, that they are producing results of high quality and reliability, which then gives confidence in the opinions being provided that are based on those results

There are a number of bodies that provide guidance on validation of laboratory equipment. As the fields of study that utilise some type of laboratory equipment is very broad, so too are the aspects of equipment that can be validated, and the different requirements for individual disciplines. Within forensic biology specifically there are several sources of material that can be relied on when determining what constitutes best practice.

- Published recommendations: International forensic bodies publish recommendations on a range of forensically relevant topics, including aspects of validation. For example the USA based Scientific Working Group on DNA Analysis Methods (SWGDAM) publish guideline documents (see https://www.swgdam.org/publications) on a range of topics such as DNA profile interpretation, contamination minimisation, validation of probabilistic genotyping systems, and most relevant to this report Guidelines for Forensic DNA Analysis Methods [4] (which provides a general outline for the types of factors to include in a validation of forensic laboratory equipment). Also the European Network of Forensic Science Institutes (ENFSI) (https://enfsi.eu/) publish guidelines on a number of forensically relevant topics, including the validation of new methods or equipment [5]. In the UK the Forensic Science Regulator (FSR) produces recommendation documents on topics including validation requirements [3]. The FSR is particularly useful as it provides information on the documentation, formation of acceptance criteria, implementation, and a series of explanation of terms and concepts relevant to validation.
- Other validations: It is often the case that a laboratory which is an early adopter of a new technology or process (and / or who may be involved in the development of the technology / process) will carry out a validation and publish their findings in a peer reviewed journal. While these publications are not guidelines or recommendations,

they do provide good examples of the different aspects that can be validated and the types of data analysis that can be carried out in the various components of the validation. A relevant example of such a publication for this work is the Developmental Validation of the Quantifiler™ HP and Trio Kits for Human DNA Quantification in Forensic Samples by Holt et al [6].

- Developmental validation work: When developers of equipment or processes wish to demonstrate to the community the performance of their system, they will carry out a developmental validation. Often these are published in peer reviewed science journals when they constitute a new or novel component. However, if they do not contain a novel component then scientific journals will tend not to accept the validations for publication and the developers will provide that information in other ways. Most common in this situation is to provide the results of developmental validation on the company website and/or as part of a user manual. These developmental validations are not recommendation or guidelines but will often have been carried out in a way that addresses published validation guidelines and so provide an example for laboratories on how to carry out their own validation work. A relevant example of this is the User's manual for Quantifiler Trio and HP provided by ThermoFisher Scientific [7].

- Published standards: These are perhaps less specific for validation as they tend to talk to issues of having documented processes, systems in place for implementation and training, requirements of reporting results etc. However, they will also have information on the requirements of a laboratory to have a validation standard that is documented. For example, again using SWGDAM, section 8 from The Quality Assurance Standards for DNA Databasing Laboratories [8] focuses on validation, and has recommendations such as:

> *STANDARD 8.3 Except as provided in Standard 8.3.1.1, internal validation of all manual and robotic methods shall be conducted by each laboratory with the appropriate sample number and type to demonstrate the reliability and potential limitations of the method.*

> *8.3.1 Internal validation studies shall include as applicable: known database-type samples, precision and accuracy studies, sensitivity and stochastic studies, and contamination assessment studies.*

In Australia forensic laboratories tend to be accredited to International Standards Organisation (ISO) standards, and in particular ISO17025 [9]. Again, this document highlights the needs to have investigation of aspects such as precision, working range, selectivity, specificity, bias, robustness as part of a validation.

Some aspects of validation requirements can vary between published guidelines. There is not a requirement for a laboratory to meet every aspect of every guideline. The main purpose of a validation is for a laboratory to have enough understanding of the function and performance that they believe it is behaving reliably, and to use that understanding when developing implementation guidelines. Of course, in order for a laboratory to reasonably feel this way should require some minimum level of testing, and so by necessity usually the majority of points from any published guidance document will tend to have been addressed.

3.1 Developmental vs in-house validation

In many guidelines produced on validation there is a distinction between developmental validation and in-house validation (or sometimes referred to as verification). Developmental validations are those carried out by the developers of a new technique or process and carry out a broad set of testing that is designed to demonstrate to the community the performance of their system. Often these developmental validations will be large and will cover aspects that an individual laboratory would not cover in their own in-house validations. The in-house validations are carried out by individual laboratories in order to show the functioning of the system in their hands. It allows the laboratory to gain an understanding of the functioning and performance of the system in the way they are going to use it and from that develop implementation and interpretation guidelines specific to their own organisation. The aspects of developmental validation that are not covered in internal validations are often those aspects that do not need to be re-demonstrated and will not affect interpretation. For example, in the case of a human DNA quantification system part of the developmental validation will be to show that non-human DNA is not being quantified. Once shown by the developers, this aspect is usually accepted as true by individual laboratories validating the system and so not re-tested. On the other hand, developers may show the variability in measurements as part of their developmental validation, however it is important for individual laboratories to know the variability of the instruments in their hands and so this aspect of the developmental validation will usually be repeated in an in-house validation.

There is some guidance available on the factors to consider when validating instruments or processes at different stages of DNA profiling. The USA based SWGDAM have published a general validation guideline document “Scientific Working Group on DNA Analysis Methods Validation Guidelines for DNA Analysis Methods” in 2016 [4]. The factors that they outline for the internal validation are given in Table 1 from their document, which I reproduce in Figure 1.

TABLE 1 – Summary of Recommended Studies for Internal Validation

	Extraction System	Quantitation System	Amplification System/Reaction Conditions	Detection System
Known / Non-Probative Samples	X	X	X	X
Precision and Accuracy: Repeatability	X	X	X	X
Precision and Accuracy: Reproducibility	X	X	X	X
Sensitivity Studies	X	X	X	X
Stochastic Studies		X	X	X
Mixture Studies	X*	X*	X	X
Contamination Assessment	X	X	X	X

"System" includes methodology, chemistries, and instrumentation.

* Mixture studies will be required if the assay is intended to distinguish different contributors (male/female, major/minor, etc.).

Figure 1: Table 1 from [4] showing recommended studies for internal validation of laboratory systems

3.2 Level of statistical evaluation required

Statistical analysis of data can often be approached in a number of ways, each of which would be a valid assessment of the results depending on the specific questions being asked, and the way in which the data will be used to form guidelines. The important point is that if a statistical test is performed that it is used correctly, and the assumptions of that test type are met (or approximately met). For example, some statistical tests require an approximation of the distribution of the underlying data, otherwise they may not perform as intended and conclusions drawn from an analysis that inappropriately used that test could therefore be flawed.

Often a full formal statistical analysis of each aspect of a validation is not required, and a simplified analysis will suffice. Often this simpler analysis amounts to displaying the data in a specific way so that the performance can be visually assessed. For example, part of a validation may wish to show that as a DNA sample degrades some instrumental indicator of degradation also shifts. This type of relationship between degradation and the indicator may be easily displayed in a graph, with a line that shows the level of shift that the indicator makes. In this instance the relationship may be quite obvious without the need for a formal statistical analysis showing the slope of that line is significantly different from zero. However, if the point was to choose between two different potential indicators of degradation

(based on their performance) then it is more likely to require a formal statistical test comparing the two (unless one tests very obviously fails to perform).

The important point is to recognise what level of statistical analysis is fit for purpose, i.e., what analysis will provide the information required for the analyst to make their decision or draw a conclusion. I note that some of the sentiments expressed above are given in QH's own documentation, for example in the Appendix of the "Writing Guidelines for Validation and Change Management Reports." (document FSS.001.0012.0269.pdf) there is a section headed "Are statistics necessary given the experiment or analysis being considered", which states:

- *For strong statements "significant difference", "linear trend" etc. a statistic will be required to support the statement. For comparative statements it may not always be informative, or operationally appropriate to complete a statistic i.e. "differences were seen", or "appears to be a trend" statements do not require a statistic.*

- *Where a statistical test is not informative, and/or particularly where the difference between the experimental groups will not have an operational meaning - use of box plots are recommended. Box plots display the variation present in a system. Generally if the box plots overlap the difference between the groups is functionally non-significant.*

Another important factor is to recognise the difference between a finding of statistical significance and a one of practical significance. This is particularly important when the number of data points is large, providing power for a statistical test to identify very small differences with statistical significance. If these small differences are of a magnitude that they would not affect the decision being made, or an interpretation of the data then there is no practical difference and so the statistical difference can be noted, but not acted on. This concept is nicely described in a forensic context by Aitken et al in [10]. The decision on whether a statistically significant finings has a practical effect must be made in the context of how the data will be used to form implementation or interpretational guidelines within a laboratory.

3.3 Appropriate sample size

A point that commonly arises in the experimental design of validation work is the appropriate number of samples to test. The answer to this question is not straight-forward and will depend on the test being performed, the question being asked and the intended use of the results. There is a misconception that large numbers of samples are always required to investigate the comparative performance of two systems. As a general rule, as the expected difference in performance of those two systems decreases, and/or as the variability in the measurements increases so too would the number of samples needed to identify the difference increase. It is important to ask whether a small statistically significant difference will have any practical

significance to the outcome (as previously explained) and therefore whether large numbers of samples are required. In a paper by Butler [11] entitled "Debunking Some Urban Legends Surrounding Validation Within the Forensic DNA Community" he makes the point:

> *URBAN LEGEND #1: HUNDREDS OR THOUSANDS OF SAMPLES ARE REQUIRED TO FULLY VALIDATE AN INSTRUMENT OR METHOD*
>
> *[…some text removed…]*
>
> *The Student's t-test can be helpful in defining potential sample numbers for validation experiments. After running 5–10 replicate samples for a particular experiment, there are diminishing returns to adding additional results. The number five is already in use throughout the forensic DNA community. The 1996 National Research Council report The Evaluation of Forensic DNA Evidence requires at least five observations of an allele when establishing a minimum allele frequency (5), and the National DNA Index System (NDIS) expert system validation requirements involve the observation of at least five challenge events for each issue such as stutter, spikes, etc. (6). When conducting an internal validation, the SWGDAM Revised Validation Guidelines recommend running a total of at least 50 samples—not 50 samples per experiment. Typical internal validation studies include concordance testing with known and nonprobative evidence samples, examining precision, reproducibility and sensitivity, and assessing stochastic effects and the detectable range of mixtures and contamination.*

Some guidance on sample numbers for validation, specifically for liquid handling platforms, is given by BradShaw et al [12], who states:

> *2.3 Number of Replicates*
>
> *Guidelines for the number of replicate quantities dispensed per tip of an automated liquid handler are not as specific as those for handheld single-channel pipettes. For single-tip handheld pipettes, several regulatory agencies, including ASTM […] and the International Organization of Standardization (ISO) […], offer guidance and recommendations. For example, ASTM (American Society of Testing and Materials) E 1154 […] specifies taking four replicates monthly for a "quick check" and 10 replicates quarterly for a comprehensive volume verification. This recommendation is generalized and does not take into account the needs of individual laboratories. When a performance evaluation is conducted for an automated liquid handler, at least three replicates per tip should be collected. The authors recommend that four to 10 replicates be collected per tip on a monthly to quarterly basis until a system's baseline performance is established. At this point, the frequency of performance verification (volume checks) could be re-evaluated for shorter or longer durations, depending on the reliability of the liquid handler under test.*

I note that some of the sentiments expressed above are given in QH's own documentation, for example in the Appendix of the "Writing Guidelines for Validation and Change Management Reports." (document FSS.001.0012.0269.pdf) there is a section headed "Sample numbers", which states:

> *When deciding how many samples are require for an individual experiment the following should be evaluated:*
>
> - *Consider the amount of variation you are expecting to see. Where little variation is expected (e.g. number of alleles obtained from blood samples) small experimental sample numbers are needed. Where variation is higher (e.g. peak heights from low DNA quantification samples) sample numbers should be much higher. Where the amount of expected variation is unknown it is possible to run one set of samples, assess the results and then run additional samples if required.*
> - *The experimental design is always aiming to include enough samples to model the expected variation in the relevant experiment (given the experimental factors under consideration). Thereby producing sufficient information (via sample numbers) for the development of methods/thresholds to cover "most situations". It is not possible for a study/validation to cover all possible situations.*
> - *In cases were a project/validation is assessing locus amplification efficiency, and inter-locus peak height balance larger sample sizes may be required (suggest use of population samples ~200-250); this is particularly relevant for Y kits where a linear relationship may not be seen.*

There are formal statistical tests that can indicate the number of samples required in a study. These are referred to as power tests and by using the level of variability known to exist in the data, the expected difference between groups, the level of significance the difference is being examined under and the level of power desired to detect that difference, a sample number can be calculated. Power tests are commonly used in biomedical trials but rarely used by forensic laboratories to pre-determine sample sizes required for validation work. One of the limitations of power tests is that some expectation of measurement variability and the difference between groups is required before any work is carried out. In some instances, this information could be approximated using other published validation work, although that may not exist, and runs the risk of not mirroring the performance of the system in the laboratory carrying out the validation. Even in published forensic validation literature it is not common to see power analyses having been performed. The most common practice is for laboratories to start with a reasonable number of samples that are informed by other validation work that may have been carried out (either published or in-house) and after analysing the data produced, determine whether additional data is required.

3.4 Reliability of a technique

Formally in statistics, reliability is equated with a consistency of measure. Therefore, there is some component of the variability in a technique within reliability as well as the ability for the results / findings (even if they are variable) to be reproduced by another individual. However, variability cannot be the only measure of reliability. A measurement could be taken in a very reliable manner but not be valid for the question being asked.

Additionally, when a set of repeated measurements is taken and shows some level of variability, it is often the case that people will interpret this as the measure, or test, being unreliable. This is not necessarily the case and depends on the way the measure will be implemented. In forensic science the term reliability would more often be thought of in terms of the final outcome of the testing, i.e., the assuredness that when an opinion is being given, that the results it is based on have been interpreted in a way such that the opinion is well-founded and the strength of evidence is appropriately expressed. This means that a technique that produces some measurement, for example peak height variability, could show high levels of variability in that measurement. However, as long as that level of variability is known and taken into account when the peaks are being evaluated, then the implementation of the technique will be reliable, and by extension the technique can be considered reliable. Therefore, when assessing reliability, as well as a measure of the variability of any system (either in an absolute sense, or in comparison to an existing method) the implementation should be considered.

One final point to make is that there is a difference between showing that a method is producing reliable results and having no indication that suggests it is producing unreliable results. For a method to be found to be producing reliable results (i.e., information provided to stakeholders, or decisions being made), it requires consideration of the validation work, and the way in which is it implemented i.e., has the implementation considered the findings of the validation with an appropriate understanding of performance. Similarly, finding a method or instrument as unreliable would require validation work to be conducted and for there to be a misunderstanding of the findings, leading to a misapplication of that technique when implemented. It follows then that a finding of my review can be that there is no evidence of a technique being unreliable, which can be made even when validation work has not been optimal. Again, I iterate the difference of this opinion to stating that a technique is reliable. One final point towards this idea is that a validation can be carried out entirely appropriately, to best practice, and with appropriate statistical tests, and still the results of that validation can be misunderstood in a way that leads to unreliable information being provided, or decisions being made.

Given the level of developmental validation carried out by manufacturers on instruments used in forensic institutions, and the wide-spread use of these instruments, there is a general prior expectation that they will perform reliably, and any unreliability will generally come from the way a laboratory chooses to implement the instrument.

I have attempted to provide context around when a sub-optimal validation (leading in most instances to a lack of knowledge in the performance of that process or instrument) could lead to an unreliable decision being made, or unreliable information being provided. Note again

that I do not have evidence to the fact that this occurring (except perhaps apart from the DIFP statements that have already been noted by the Commission), but highlight the possibility when appropriate.

3.5 The QH Validation SOP

QH have a standard operating procedure (SOP) entitled "Forensic DNA Analysis Validation and Verification Guidelines" (FSS.0001.0024.1370) that covers the topic of when to carry out a validation compared to a verification, and what each constitutes. The SOP covers much of the same material I have discussed above and provides an appropriate framework under which validations can be carried out.

3.6 The DNA profile generation and evaluation process

When an exhibit is submitted to a forensic facility, there are a number of stages for the exhibit to progress through to ultimately have an opinion expressed in a report about potential DNA donors. Without going into detail for any one stage, these are:

1. Sampling – examining the exhibit, carrying out testing to identify stains or areas of interest, and using a sampling device (such as swab or a tapelift) or taking a cutting in an attempt to collect DNA from an area of interest
2. DNA extraction – submitting the sample (taken from an exhibit, person, or area) to a series of physical and/or chemical reactions so as to release cellular material from the sampling device, and then release DNA from their cells. The result is a solution called a DNA extract.
3. Quantification – A process in which a small aliquot of the DNA extract is tested in order to determine the concentration of DNA present in the extract
4. Polymerase Chain reaction (PCR) – A process where the regions of DNA that are of forensic interest are targeted and copied numerous times. Each copy of the DNA that is generated during PCR has a marker (called a fluorophore) added that can be detected during capillary electrophoresis. The information about DNA concentration is used to determine how much DNA extract to use in the PCR setup (or whether to proceed to PCR at all). The copied DNA fragments are referred to as amplicons, and the post-PCR tube containing the amplicons is referred to as the PCR products.
5. Capillary electrophoresis (CE) – In this stage the amplicons produced during PCR are passed through a capillary containing a gel matrix. This process sorts the fragments according to their molecular weight (the length of the DNA fragment). Once fragments have passed through the capillary, the fluorophores (added during PCR) are excited by a laser and emit light, that is detected by a camera. The result is a graph showing fluorescence over time for different colours of fluorophore, which is commonly referred to as a DNA profile. When the DNA extract contains a particular sequence of DNA (called an allele) at a region being targeted (called a locus, plural

loci) the PCR process will generate multiple copies, each appended with a fluorophore. The CE process will detect these copies as they leave the capillary at a particular time, which can be equated to their molecular weight. The intensity of the fluorescence (and hence the height of the resultant peak) depends on the number of copies of that amplicon, which in turn depends on the number of copies of that allele in the initial DNA extract. Hence the DNA profile is a collection of peaks, each with a molecular weight, and an intensity. Through the use of known controls, the molecular weight is translated to the allelic designation.

6. DNA profile reading – During the processes of PCR and CE the DNA profile that is generated will consist of a number of peaks that represent alleles in the DNA extract, and a number of artefactual peaks that are inherent in the generation of a DNA profile. The process of 'reading' a DNA profile is the identification and removal of the artefacts from a DNA profile, leaving only the peaks of interest for an evaluation. Typically, the process of reading is carried out by two people, who independently read the profile, compare results, and resolve any differences

7. DNA profile evaluation – The final stage is taking the generated and read DNA profile and interpreting the information. Typically, this will be carried out in an attempt to identify the DNA profile of a single DNA donor to the sample (so that the profile can be compared to a database of reference profiles), or to compare to reference profiles of persons of interest in a case (so that an opinion can be provided on their potential DNA donation to the sample). In the past 10 years, DNA profile interpretation and evaluation is assisted by computer software programs that are grouped under the heading 'probabilistic genotyping'. STRmix is one such tool in use in Australia.

The validations that I have reviewed play roles in various stages I have mentioned above. It may assist to identify their place within the DNA profiling process so that the connection between them becomes clearer. The technology that is the subject of these validation reports is:

- QIAsymphony® - This is an automated system used to carry out DNA extractions. This related to step 2 above.
- Quantifiler® Trio – This is quantification kit that is used in the quantification stage of DNA profiling. Quantifiler Trio targets three areas of human DNA, a short fragment, a long fragment (which together can provide an indication of the level of degradation in the sample), and a male DNA target (so that the male and female DNA amounts in a mixture can be determined). This relates to step 3 above.
- Quant Studio 5 – This is the laboratory equipment that is used to process and analyse the quantification reactions i.e., the Quantifiler Trio reactions will be processed on Quant Studio. Quant Studio is a type of quantification instrument called a real time PCR (or a qPCR) instrument, as the quantification reaction is a type of PCR reaction. This relates to step 3 above.

- PowerPlex®21 – This is a DNA profiling kit. It contains the reagents used to set up a PCR, targeting the loci of interest within the human DNA. PowePlex 21 target 21 loci (regions). This relates to step 4 above.
- ProFlex™96 – This is the laboratory instrument that carries out the PCR i.e., the PowerPlex21 reactions are run on the ProFlex. This relates to step 4 above.
- ABI 3500xL – This is a capillary electrophoresis instrument that is used to separate the fragments produced during PCR and generate the DNA profile. The PowerPlex21 reactions (after they have been run on the ProFlex) are run on the ABI 3500xL. This relates to step 5 above.
- Hamilton STARlet – This in an automated system of setting up forensic processed. It is referred to a liquid handling platform. Common forensic applications are to set up multiple PCR, quantification, or CE processes. This relates to steps 3, 4 and 5 above.
- STRmix™ - This is a software program known as a probabilistic genotyping tool. It is used to analyse and interpret DNA profile information. This relates to step 7 above.

4.0 Validation of Quantifiler® Trio - September 2015 (FSS.0001.0003.7266.pdf)

4.1 Experimental design

I note that the guidelines from which Figure 1 was taken were published after the compilation of the Queensland Validation of Quantifiler Trio and so could not have been used as a guide and are mentioned here just as a modern reference for validation expectation. It is not necessary that all aspects in Figure 1 are carried out in order for a laboratory to satisfy themselves that a technique, process or item of laboratory equipment is performances to a level that is suitable for implementation into active casework. Some additional trust in the performance of the Quantifiler Trio system can also be gained from the fact that the developmental validation [7] has addressed each of these points.

The Queensland validation study has not explicitly included an experiment focussed on contamination assessment, which would include an assessment of the level of extraneous human DNA in the reagents used in the Quantifiler Trio kit, or the results obtained from blank samples. However, this point could arguably be said to have been carried out implicitly by the presence (or absence) or DNA in reagent blank samples used in the study.

My opinion is that the main points addressed in the validation study satisfy the requirements of a validation of a Human DNA quantification system. Note this is not a comment on the quality of how each point has been addressed (that commentary is provided below), it is only a comment on the breadth of points being addressed.

4.2 Experiment 1: Assessment of Quantification Standards

Accuracy is used to describe the closeness of measurements to their true values. I am not aware of a standard technique that measures percentage inaccuracy. In the project proposal #152 document (FSS.0001.0019.7147.pdf) the acceptance criteria for experiment 1 refers to precision, which is different from accuracy when using statistical terminology, and refers to how closely clustered measurements are to each other (but not necessarily the true value). Precision and accuracy are often used in machine learning tasks that classify measurements into categories, so that any individual assignment can be considered correct or incorrect. The value being measured here (i.e., DNA concentration) is continuous, which means that it doesn't fall into distinct categories. For continuous variables usually performance measures which would be akin to precision and accuracy are based on results of a regression and come from the value taken by parameters in the regression and their standard error. For example, the development of a standard curve from Quantified results determines performance by the use of an R^2 value (or coefficient of determination), which is a measure produced from a linear regression that indicates how much of the variation in the data can be explained by the model (i.e., the independent variables being used to describe the data).

The method used appears to be to use the standard curves created by both the Life Technologies (LT) and Promega (PR) standards to quantify the DNA amount in a set of NIST DNA standards. The percentage of change has been calculated and referred to as the 'percentage inaccuracy'. There are a few points that arise with viewing data in this way:

1. Calculating the percentage change creates a skewed metric i.e., this metric can only take theoretic values from -1 to infinite. The value of equality in this range (i.e., when the two values being are the same) is 0, and therefore the most extreme underperformance can only differ from the value of equality by -100%, whereas the most extreme overperformance can differ from the value of equality by a large amount (open ended). This fact can be seen in the graphs in Figures 17 to 22 by the fact that the lowest value that the 'percentage inaccuracy' can take is -100% whereas there is no specific upper bound value. One solution to the skewness is when dealing with a ratio, to either transform (or display) the results on a log scale. However, percentage change is a common measure of similarity between two measures and in this respect the graphs in the validation report are valid.

   However, the skewness in this measure means an average of the 'percentage inaccuracies' (as is given in this validation) will visually appear to favour higher percentages, and when averaged will tend to give values above 100%. As a simple example, imagine two measurements of performance compared to a known standard, one measure which represents a halving of performance (hence -50%) and one which represents a doubling of the performance (hence +100%). While we might wish an overall statement on these two observations to be that there is no bias in the

performance an average of -50% and +100% leads to +25% and therefore suggests an overall higher level of performance.

2. The ultimate value that was used to decide between the two standards was calculated by averaging across percentage change for all quantification values. This data (as is common for many data types) displays a trend in the variability of results (or technically put a trend in the variance, which is called heteroscedasticity), whereby the smallest concentrations have the highest variance. Therefore, averaging across all values leads to a statistic that is dominated by the performance of the small concentrations i.e., the choice of which set of standards has performed the best will really be based on which has performed the best for the smallest concentrations.

A simple treatment of this data could have been just to graph the observed quantification values against the expected quantification value, showing some measure of variability in the data. This could have been achieved in a manner that is similar to that shown in Figure 17 of the Quantifiler HP and Trio Quantification Kit User Guide (4485354 Rev H), which I reproduce below in Figure 2. Note the DNA concentration scales in Figure 2 are spaced on a log10 scale (which is a common way of dealing with data that spans a large range of values, and/or data that is skewed).

Figure 17 Dynamic range of male samples using Quantifiler™ Trio DNA Quantification Kit and Quantifiler™ HP DNA Quantification Kit (DNA concentrations <10 ng/µL)

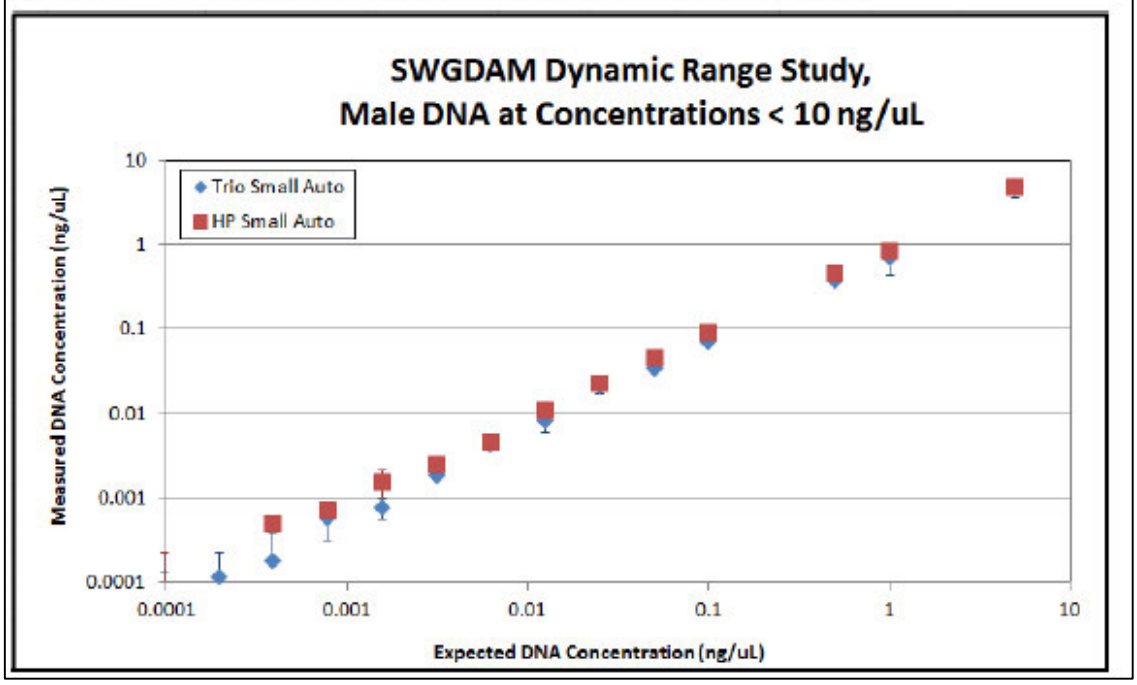

Figure 2: Showing Figure 17 of the Quantifiler HP and Trio Quantification Kit User Guide (4485354 Rev H)

Alternatively, the results of the observed/expected ratio could have been plotted (again on a log scale). The formal method for determining which method performed best would be regression, however some account of the heteroscedasticity would ideally be made. This does take the complexity of the analysis beyond most standard regressions, and for the purpose of choosing which standard set to use a choice made on a simple visual comparison of performance (as described above) would be adequate.

Also note that a common method used (the most common method used for research into the agreement of medical instruments according to a review by Zaki et al [13]) to compare the performance of two instruments over a range of input values (such as used in a dilution series) is described by Bland and Altman [14]. This method also allows a Bland-Altman plot (sometimes referred to as difference plots, or Tukey difference-mean plots) to be produces that graphically demonstrates the performance of the two systems compared to each other. The graphs in Figures 17 to 22 are similar to a derivate of a Bland-Altman plot, called a Krouwer plot [15] which shows change against a known amount. The main difference with the plots in Figures 17 to 22 and a Krouwer plot is that Krouwer plots will also provide bounds that shows the level of acceptable variability. As the level of variability for different concentrations is likely to be different (i.e., lower concentrations will have higher relative variability) then it may be that multiple Krouwer plots showing difference would be needed, or a single Krouwer plot showing percentage change.

Also, there is an apparent bias to positive levels of change, particularly at lower concentrations. I suspect the effect comes from not including the undetected results in the calculations. It is typical to include undetected results in analysis by assigning them a value that corresponds to some lower bound (e.g., half the detection threshold of the instrument), which then avoids the issues of apparent positive bias.

The method used to measure the performance has several issues that may confound the decision. However, the question now is whether a different decision would have been made had a formal statistical treatment been applied. In this case the choice between LT or PR standards was to determine which standards would be used in the remainder of the validation work. Given that both standards are produced by professional biotechnology companies and have both presumably been used in forensic laboratories it would be reasonable to expect a high level of performance from either system. From purely an ‘eyeballing’ of the spread of values the decision to proceed with the LT set is a reasonable choice and would likely be supported by a more formal analysis. Additionally, there is the fact that there was a higher pass rate for the LT standard sets (based on the R^2 value) compared to the LT kit, which further supports the decision.

There is a further note at the end of section 6.1 that compares the results of the current validation to those of Quantifiler HP (another modern DNA quantification kit produced by ThermoFisher) and draws the conclusion that Quantifiler Trio is more accurate than Quantifiler HP. In this instance the literature I can find that compares the performance of these two kits finds their performance to be similar (as shown in Figure 2, and also [6] although that study it doesn't directly compare performance between the kits). The Quantifiler Trio system has the advantage that it also quantifies male DNA as well as total DNA and so has an information advantage that would make it a more attractive choice. Therefore again, the decision made in the validation study is likely to reflect a decision that would be made with a formal analysis.

4.3 Experiment 2: Standard stability assessment

The proposal suggested that Student t-tests would be the criteria for determining the longest time period that would be accepted, but this does not appear to have been used and instead a criterion of the change in quantification value being less than 21% was used. I am not sure where this value came from, but I suspect it is more of a statement of observation rather than a specifically set criteria. Updating or modifying acceptance criteria is fine (sometimes a criterion that seemed reasonable, with additional knowledge on the working of a system, is realised to be not appropriate). As stated by the UK FSG [3] in section 7.8 on acceptance criteria:

> *7.8.2 If certain criteria were considered mandatory and prove to be unachievable, this assessment offers the opportunity to review the stated criteria with the proponent of that criteria (or rating thereof), or with the customer, to see what the next steps should be before finalising and writing up the report. As with all records, this review should be maintained for traceability.*

The important point is to document the initial non-adherence, and then the new criteria along with the rationale behind the change.

Regardless of the criteria, a simpler method of assessment would have been to determine when the intercepts, slopes and ultimately R^2 value obtained from the standards started to fail the acceptance criteria. This would then align with the manner in which the kit manufacturer tested the stability of the standard dilution series (see reproducibility study section in [7]).

The manufacturer's recommendations for the length of storage of standards is two weeks, whereas the recommendation given in the QH report is longer. This in itself is no issue, and often manufacturer recommendations are set conservatively. I note that in the recommendation the length of time accepted for storage given was five weeks, whereas in the text it is clear that four weeks is the intended timeframe. I suspect this is simply an oversight, but it should be checked to ensure that in practice the four-week timeframe is being used. I note that in the "Preparation & Testing of Quantification Standards, In-house DNA Controls,

Quantification Kits and Amplification Kits" standard operating procedure, under section 5.3, Registering New Standards, step 7 states "Enter the expiry date as 4 weeks from the date of preparation.", which suggests the appropriate timeframe determined by the validation is indeed being used.

4.4 Experiment 3a: Single source sensitivity (LOD)

In the context of a quantification system studies into the limit of detection are also sometimes referred to as sensitivity studies. Figure 3, shows a Figure from [5] (section 3.2.3) and shows graphically the concepts of a dynamic range over which measurements can reliably be taken.

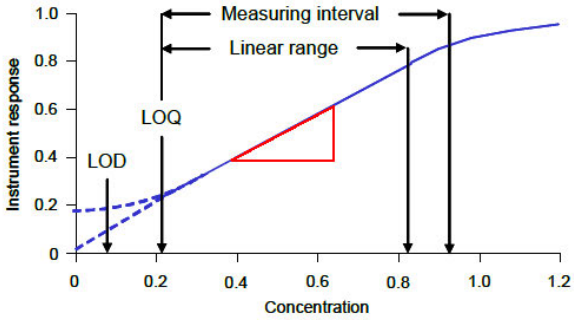

Figure 3 *A calibration curve showing instrument response (indication) versus concentration; the measurement interval, linear range, LOQ and LOD are identified. The triangle illustrates the calculation of the sensitivity or the slope of the calibration curve (=Δindication/Δconcentration). Figure reproduced from ref [3] by permission of Eurachem.*

Figure 3: Example of a dynamic range over which measurements can be used to predict some property of a sample

In general, there will usually be a range which some measurement will increase (or decrease) linearly with respect to a test property (such as concentration show in Figure 3). As the property lowers it will start to be confounded by the limits of the system (for example interaction with instrument 'noise') and so the linear trend will break down. Where the relationship starts to deviate from linear is considered the limit of quantitation, LOQ (i.e., where measurements can be used in a quantitative way). Below this level is the limit of detection, LOD, where the measurements may still be used to detect the presence of an item, but the ability to predict amounts is limited or absent. At the upper end of the property the linear trend can break down to saturation of the system (either saturation of the chemical, biological or physical processes). In the case of Quantifiler this upper limit is identified in section 6 of the user's manual [7]:

A linear relationship between expected quantity and actual concentration was observed for DNA dilutions within the supported quantification range of the assay, from 5 pg/µL to 100 ng/µL.

QH demonstrate this dynamic range by providing tables of average observed quantification values (based on two replicates) compared to the expected quantification values. They have also utilised results from experiment 1 and 2 in their assessment.

Formally a LOD is the lowest amount that an analyte in a sample can be detected with a stated probability. This can be set based on the strength of instrument signal produced when analysing blank samples compared to strength of instrument signal produced when analysing sample with an analyte. Alternatively, and commonly for quantitative PCR (qPCR) validation, the LOD can be set at the concentration where the probability of detecting DNA at that concentration is above a pre-determined threshold (typically 0.95). The reason for this slightly different definition for qPCR comes from the fact that it is not the raw fluorescent data being used in the validation (as it would in other instruments), but rather the quantification instrument's processing of that raw data into a DNA concentration value. Under this paradigm the variability in the detected signal does not play into the LOD assessment (although as variability increases then naturally so too will the instances of a non-detection). Examples of this type of analysis can be found in literature outside forensic science (for example [16, 17]). Often these published examples of determining LOD will employ approximately 20 replicated per dilution factor, and I note that the QH validation has only performed analysis in duplicate (i.e., 2 samples at each dilution) between the ranges of 0.09ng/µL and 0.001ng/µL. In the QH validation work all experiments lead to DNA being detected. Low numbers of replicates do not necessarily mean that a determination of LOD cannot be made, but it does mean the determination is more prone to stochastic events. Technically, given the results in this experiment, the validation has not found the LOD, but set it at a level above the LOD (but noting that additional replicates could in fact reveal a higher LOD). I recommend that the LOD experiment be redone according to standard methodology so that the LOD can be set with more rigour. As well as a dilution series of DNA concentrations the validation should also include quantification of a number of negative control samples to determine whether any false positive quantification values are obtained, and if so the rate at which this occurs.

When an LOD is based on the similarity between signal produced by a blank sample and a low-concentration sample then it represents a level at which the instrument cannot be used to distinguish between the two i.e., a signal at this level provides no evidence to support an analyte being present, but if it were present below the LOD then it wouldn't be distinguishable by the instrument output from background noise. The interpretation of the LOD for qPCR is slightly different, as it relies on a probability of detection. A common misconception about an LOD set in this manner is that it means any value quantification value below this level cannot be distinguished from a blank sample. In fact, the LOD set in this manner simply means that if an analyte is below the LOD level, there is less than 0.95 probability of detecting it. However, if the instrument has detected DNA below this level,

then it can still be said with some confidence that DNA is present at low levels, i.e., a DNA quantification value below the LOD should not be interpreted as no DNA being present.

In the QH determination of a dynamic range there is no upper bound specified, but in this instance that is expected given the way in which quantification results will be used in the construction of PCRs, i.e., there is unlikely to be an upper-bound quantification value beyond which samples are not progressed to PCR in the same way that is done for a lower-bound.

The LOD, set at 0.001ng/μL, has not been set in a manner with as much rigour as I have seen in other published examples (either by statistical methodology or number of replicates). Nevertheless, the value chosen for LOD aligns approximately with other cited values, for example the 0.005ng/μL limit given in the Quantifiler user's manual [7] and even the findings of Forootan et al [17] (despite using a different quantification kit with different DNA targets) of 3 molecules per reaction (which would equate to 0.0012ng/μL if converted to a value that was approximately equivalent to the Quantifiler work). There are numerous other examples that could be drawn upon.

The final point I will make here is that the LOD for quantification using Quantifiler Trio and the QuantStudio 5 is similar to the lower sensitivity of modern profiling kits used in conjunction with a 3500xL capillary electrophoresis instrument. There is also the unavoidable issue of sampling variation in the aliquoting of DNA taken to set up quantification or PCR reactions, which can lead to some DNA fragments in one and not the other. The consequence of these points is that samples which produce an undetected quantification value can still contain low levels of DNA that are abundant enough to generate (at least a partial) DNA profile. Whilst this fact doesn't affect the manner in which a LOD is calculate for a quantification system, again it should be born in mind when the meaning of that LOD is conveyed to other scientists and stakeholders.

Taking all of these factors into account, I don't find any evidence to suggest that there would be unreliable results obtained by the use of QuantiFiler Trio at low DNA concentrations. However, I believe the validation would benefit from some additional work to better define the LOD, and the meaning of the LOD should be made clear to analysts who need to interpret it in the context of forensic casework.

4.5 Experiment 3b: mixture studies

There is an email (document FSS.0001.0066.4907.pdf) relating to the mixture studies that comments on the varied performance of Quantifiler Trio on some of the range of male to female mixtures. The mixtures highlighted are those where the male DNA predominates and in the final report the M:F ratio is reported as n/a, even though a ratio could be calculated. No explanation is given for the lack of ratio in the validation report. While the email suggests that the performance is poor in these samples there are two points to consider:

1. As a general point of validation, when a value can be calculated then it should be calculated, even if those results do not show good performance (in fact it is especially these values where it is important to show the performance so that limitation can be developed).

2. The potential issue identified in the email (and presumably leading to the n/a values in the final report) are exacerbated by the way the data is shown in the email. The area of reported low performance is simply due to the fact that male DNA predominates the sample and there is some stochastic variability expected in both the total and male quant values. As there are no indications of the standard error in the validation report it is not possible to provide error bars or a confidence interval around the male DNA proportions, which would demonstrate whether there were any values falling outside an expected range. Regardless, if the results are graphed showing the male DNA proportion (and capping the upper observed value to 1) then it produces the graph seen in Figure 3, which I imagine would have been less concerning if viewed.

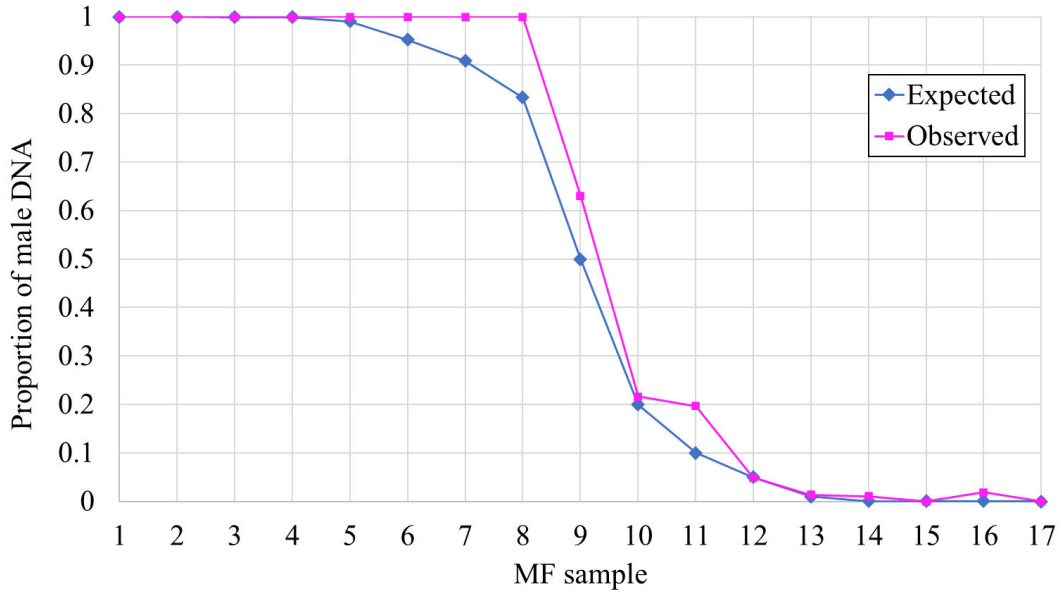

Figure 3: Ability of Quantifiler Trio to predict male DNA in male:female mixed DNA samples

4.6 Experiment 4a: repeatability

There is no indication of the variability in the data. I would expect a standard deviation, or a standard error, or a coefficient of variation to be provided so that the performance of the instrument could be assessed. To explain further, the experiments on reproducibility have carried out tests in an attempt to show whether there is a difference between the plates, but even if no difference is found we don't know whether both plates performed equally but with high variability or equally with low variability. More important than showing the plates were

statistically the same, is a modelling of the amount of variation in results produced by the instrument.

A point to make here is that carrying out comparative statistical tests such as the Student's t-test may give an indication that there is no support for a difference between two sets of measurements. This is slightly different to stating that two sets of measurements are 'the same' (or that two instruments are performing the same). This latter statement requires some definition of what is considered 'the same', which would normally be set based on bounds of biological or laboratory significance.

Individual Student t-tests are not the best test to use here as each has a chance of showing a false significance (commonly referred to as the 'multi-testing problem'). There are ways of correcting for multiple significance tests (such as the Bonferroni correction), but these too come with limitations and can lead to greater risks of failing to reject a false null hypothesis. There are a couple of ways in which this data could have better been analysed:

1. Draw a plot (such as a boxplot, or a scatterplot with error bars) for each concentration and plate to show the spread of values and draw an opinion by 'eyeballing' the level of overlap of the two plates

2. Carry out a multiple linear regression which seeks to describe the expected quantification value given both the observed quantification value (as a continuous variable) and the plate (as a categorical variable). The regression would certainly find the observed quantification value significant but would also provide information as to whether the plate was a significant factor in predicting the expected value. This would answer the question of main interest.

Given the main issue with the multi-testing problem is seeing significance where none exist, and in this instance the majority of the individual t-tests did not find any significant difference, it is unlikely that significant differences do indeed exist.

Having said this, the Quantifiler user's guide does compare plates in a similar manner as the QH validation, in that results of individual concentrations are compared between groups. If this type of testing is to be conducted, then the appropriate type of t-test would be a paired t-test. A paired t-test uses the information that the same sample is being compared under two conditions (in this case different runs of the plate).

Without the results of each repeated run being tabulated, or shown graphically, it is difficult to assess whether there is any real difference between repeated plates. Given the t-tests performed (although not ideal) it seems likely that there would be no difference if appropriately tested and so I do not believe there is evidence that the results would be unreliable when ultimately implemented into casework guidelines. However, there is a lack of information about the level of variability in the repeatability study.

4.7 Experiment 4b: reproducibility

Normally this test would explore the difference the repeatability of the instrument tested in 4a) and the variability of different setups of the plate (from the same samples) and potentially run on different days, and/or using different instruments, and/or by different operators. Therefore, the experimental setup would test multiple instances of repeatability against reproducibility. In the QH validation ideally the plate C would have also been run multiple times to obtain its repeatability, and then the combined repeats of plate C would be compared to the combined repeats of plates A and B. The absence of repeatability information about plate C is probably not an issue that would affect the ability to draw conclusions on reproducibility, but rather just not the best practice. I do not think there is any serious issue with the experimental design here to the point that it suggests unreliability of the outcome. I also note that the reproducibility only extended to one additional plate setup, where best practice would have been at least 3. In the Appendix of the "Writing Guidelines for Validation and Change Management Reports." (document FSS.001.0012.0269.pdf) it states for reproducibility:

- *"Run a plate over multiple days (as many as is practicable e.g. over 3-5 days), with different operators.*
- *The "plate" of samples used for reproducibility may include the same samples used for repeatability. It is suggested that ~12 samples (min 7 samples), plus controls are included in the reproducibility plate.*
- *Scatter plots or box plots can be a way to display the data to evaluate reproducibility within the system."*

In the QH validation report the measurements for each concentration were compared between plate C and plate A, and also for plate C and plate B, using a Student's t-test. Again, I note that there is the occurrence of the multi-testing problem here, in that there are multiple tests all with a probability of showing a significance just by chance (i.e., when one does not exist). Again, a potential solution to this would be to carry out a multiple regression with potentially plate and day as categorical factors, where day would then represent the significance related to reproducibility.

There is a comment (#6) in the 'Instructions to expert' document that a comment about t-test results given on page 47 being nonsensical due to the misapplication and misunderstanding of t-tests. I agree that the statement given in the report:

> "*…the low t-test score at 0.001ng/uL is due to the low accuracy and the high variability at that DNA concentration level, therefore the t-test score of 0.00787 ($p \geq 0.05$) is not unexpected*."

Is not correct as high variability in this instance would tend to cause no significant differences to be identified, rather than expecting to see them as suggested.

However, without carrying out a regression, and if validation work was carried out in a manner similar to the Quantifiler user's guide (which I believe would be fit for purpose) then the appropriate test would have been to compare plate C to plates A and B combined. Again, a graphical display of results would be highly informative in this instance (such as given in the reproducibility section of the Quantifiler user's manual). Again, I do not believe that there is evidence of unreliability of the implementation of Quantifiler resulting from the points I raise above. If anything, there is perhaps a lack of knowledge of the amount of variability that exists in the quantification result (although again I note that the general level of variability could be gained from other published validation work and is likely to be similar to that of QH). The impact of any lack of knowledge about variability is minimal in the implementation, as it is not common to take into account variability in the quantification value in the decision on how much sample to use during analysis (or whether to proceed to profiling). This is because the best estimate for DNA amount is still aways going to be the point estimate of the calculated contraction (regardless of variability). The main time the variability may be useful is when troubleshooting unexpected DNA profiling results.

However, an understanding on the variability of results can be important when describing results to stakeholders. For example, if a DNA concentration is returned as 0, or below the LOD, then it may be of interest to the stakeholder whether this means that there could be DNA in the sample above the LOD, and with what probability. Additionally, the variability in quantification may be important to understand when assessing whether to rework a sample. In the QH 'Procedure for Case Management' SOP (FSS.0001.0001.9477), section 19.2, 'Quantification issues', it states:

> *If the profile seems inconsistent with the quant value or if the quant value is unexpected given other results or testing (such as numerous spermatozoa present), consider a re-quant as the best option.*

But the question arises as to how an analyst identifies a profile that is inconsistent with the Quantification value. These sorts of questions can only be answered with knowledge of the variability of the quantification result. At worst, if the lack of knowledge of variability is misunderstood as meaning there is a lack of variability itself then this could lead to inaccurate decisions being made or information being given.

4.8 Experiment 5: inhibition

Nothing to comment on here as no formal statistical analysis is required.

4.9 Experiment 6a: degradation protocol

Nothing to comment on here as no formal statistical analysis is required.

4.10 Experiment 6b: degradation index proof of concept

Nothing to comment on here as no formal statistical analysis is required.

4.11 Experiment 6c: degradation index threshold

Nothing to comment on here as no formal statistical analysis is required. The decision on whether to proceed to profiling is not really a statistical one but rather a laboratory resource driven policy decision. I also not that in the implementation there was no DI threshold and so there is no impact of this degradation work.

4.12 Experiment 7: Quantifiler Trio kit new formulation (IPC modification)

As many of the previous tests were carried out again on data produced using a new plate of data the same issues of statistical test choice exist and have the same impact. Most of the decision on kit performance only requires the data to be displayed visually to see whether a significant difference has occurred as a result of the new formulation.

I make one additional note here that there is a paragraph (second to last in this section) that provides a p-value for a test comparing the ‘average quantification values’ between the original and modified kit. This reads as though one t-test was carried out, perhaps comparing entire dilution series. If this is the case then it is not a correct application of a t-test, as the underlying assumptions about the data being independent and identically normally distributed would be violated.

4.13 Correspondence

In general, the comments on the report are grammatical, language preference, or requests for more/clearer explanation that do not substantially affect any outcomes or decisions.

The advice given by Kaye Ballantyne (document FSS.0001.0011.2167.pdf) on statistical analyses is good advice and reflective of an appropriate balance between statistical analysis and practical validation.

There are some correspondences that refer to troubleshooting issues in the laboratory with quantification. For example, documents FSS.0001.0052.1771.pdf, FSS.0001.0052.1773.pdf, FSS.0001.0066.4916.pdf. These are not specific issues arising from any lack of rigour with the validation and appear to be operational issues that inevitably occur during the lifetime of a laboratory. This is particularly so in the course of bedding down new instruments, processes or techniques and is what validation seeks to minimise (but can never entirely avoid).

Some correspondence addresses logistic issues of cost, timing or relay of results and I have no comments about these.

In the 'Instructions to expert' document comment 8 states "The report used both 0.0088uL and 0.05uL for the DIFP process" however I am not sure what this refers to.

In the 'Quantification of Extracted DNA using the Quantifiler® Trio DNA Quantification Kit' document (FSS.0001.0019.8074.pdf) in Tables 7 and 8 there appears to be a threshold of 0.00241ng/µL below which DNA detected in a reagent blank and negative controls need only to be noted in a log (and is not considered contamination). This value is above the LOD, and is said to have been set "*as determined by in-house validations*" but I did not see where in the Quantfiler or QuantStudio 5 validations this value originated (it may be from another validation). In a later version of this document (FSS.0001.0001.6684) provided to me this value has been changed to 00001 ng/µL in line with the LOD as determined in the validation of Quantifiler Trio.

4.14 Overall conclusion on Quantifiler Trio validation

Many instances the statistical techniques used are not the best tests that could have been chosen. However, in most instances I believe that the decisions that were made and the recommendations mean that there is no evidence to suggest the manner in which Quantifiler was implemented into casework would adversely affect the results produced. In addition to the validation work carried out at QH there is a body of literature showing that the performance of Quantifiler Trio, and the manner in which QH are using Quantifiler Trio does not significantly deviate from the manufacturer's recommended methods of validation in any way that would impact on the results.

With this in mind the question turns as to whether the performance of the QuantiFiler Trio kit has been validated to best practice, and to provide analysts with an understanding of the performance of the kit.

Additional testing should be carried out to appropriately identify the LOD, if this value is going to be used as a decision threshold for analysis or reporting. If either the LOD is not appropriately calculated, or its meaning is misunderstood then this has the potential to lead to unreliable decisions being made or information being given.

Specifically, the ability to detect DNA over a range of concentrations should be tested, whereby each step in the dilution series has 10 to 20 replicates. The LOD should then be set at the concentration at which DNA is detected less that 0.95 of the time. This can be done by calculating the detection rate at each concentration, or by use of logistic regression. If none of the concentrations tested fall below this probability, then the concentration range should be extended down to lower levels until the LOD is reached. This limit maybe below, or above, the current LOD value of 0.001ng/µL being used by QH.

Also, additional repeatability and reproducibility should be carried out, at least 3, and preferably 5 runs of sample on the same day, by the same operator for repeatability. Then these same samples should be run by a different analyst on a different day for at least 3 to 5 days (as suggested in QH own SOP).

5.0 Validation of Quant Studio 5 (FSS.0001.0005.0767.pdf)

5.1 Experiment 1: Sensitivity, Limit of Detection and Inaccuracy

While a formal statistical test could be performed here the graphs do seem to show a similar performance of the three quant instruments, based on the percentage change. Even if there were some undiscovered statistical differences found here it seems unlikely that it would have a practical effect on the choice to move to Quant Studio 5 or the reliability of results produces in casework from using Quant Studio 5.

With regards to the LOD of 0.001ng/µL determined, again a formal statistical analysis would have investigated the level of variance in quantification across the range of input values, and would have determined when the linear relationship between quantification result and DNA amount broke down, or determined the point where the detection rate fell below a pre-defined value.

However, as with the Quantifiler validation, there is little to no issue in the values chosen that would affect the reliability of results that are ultimately produced by the laboratory. In explanation, PCR kits (such as Powerplex 21 used at QH) will normally have a target input DNA amount determined during validation. Often the target amount is in the range of 0.4ng to 1ng. When the target amount cannot be added (due to low levels of DNA in the DNA extract) then the maximum volume of DNA extract that the PCR kit allows will be added to the PCR (so as to maximise the amount of available DNA). Typically, the maximum volume of DNA extract that can be added to a PCR is around 15uL, which would occur for any DNA extract with less than 0.03ng/µL to 0.07ng/µL (depending on the target DNA amount for PCR). These values are well above the LOD determined by the QH validation, meaning that even if quantification values below this were highly variable, and often not representative of

the amount of DNA in the DNA extract, it would not affect the outcome of how the PCR was set up. For PCR setups where the maximum volume is added, the DNA profiling results does not depend on the variability in the quantification value.

A brief comment here is that the results of the quantification shown in Figure 5 to 7 of the Quant Studio 5 report for the 0.0001ng/μL sample are above 0. This does not necessarily mean that there is a bias in the instruments to detect more DNA than is present in the sample. I suspect the effect comes from not including the undetected results in the calculations. It is typical to include undetected (or negative) results in analysis by assigning them a value that corresponds to some lower bound (e.g., half the detection threshold of the instrument). This is a minor point and does not really affect the validation in a meaningful way.

5.2 Experiment 2: Comparison of QS5s and 7500

The p-values given in table 4 seem to pool all dilutions together to perform the paired Student's t-test. One of the assumptions of the t-test is independent and identically distributed samples in the underlying data and the dilution series does not fill this requirement. Additionally, there is a dependency between the three markers (SAT, LAT and Y-Target) which is not taken into account by independent datasets.

Again, a formal statistical test to carry out the comparison of a dilution series is relatively complex and the best simple way to show the relative performance of QS5 and the 7500-A would be simply to plot the results of running each dilution series on the two instrument and showing the amount of variability and overlap.

As these results are presumably just using the results of the previous experiment in a statistical test, the graphed results of the previous experiment go some way towards showing closeness of performance of the QS5 and 7500.

As a general point, it is preferable to set passing criteria based on independent criteria rather than by closeness of performance to an existing instrument. Passing criteria based on the closeness to performance of existing instrument assumes a high level of performance of that existing instrument.

5.3 Experiment 3a: Repeatability

The repeatability results are shows graphically and display the range of quantification values obtained by replication. In my opinion the graph shown in Figure 8 adequately demonstrates the results of repeatability and allows conclusions to be drawn. I note that the report highlights a higher variation in samples with more DNA. This would appear to be at odds

with their previous findings that suggest lower DNA concentrations have the highest degree of variability in results. In fact, this simply a consequence of the two different ways in which the data are being displayed and there is no contradiction. It is a common effect that occurs simply due to the fact that larger values will tend to have a larger range over which values can vary although relative to their size their variability is low. For example, the range between the smallest and largest quantification values for sample 5 are approximately 7 to 10, which represents an approximate 1.4-fold change. The quantification values for sample 1 are approximately 0.5 to 1, which has a smaller range, but a 2-fold change (i.e., therefore larger than the 1.4-fold change seen for sample 5). To overcome this issue a commonly calculated statistic, used to compare variability across groups of values of different magnitude, is the coefficient of variation.

In the acceptance criteria it notes that the repeatability across the replicates was comparable. It doesn't explain how this comparability is being defined and using the coefficient of variation would be an easy way to do this. It also notes that the results are comparable to those produced on the 7500 instrument (and cites the Quantifiler Trio validation) although again it doesn't explain what this statement is based on, doesn't provide any comparison of results (statistically or visually), in the Quantifiler Trio validation there is no equivalent graph shown to allow a comparison, nor is there any numerical indication of the variability in results in either validation to allow comparison. It is not clear what has been done to draw the conclusion that the results are comparable. Again, in this instance Bland-Altman plots could have been used to visually display the differences between the instruments and provide a mechanism for numerically defining 'comparable'.

With regards to the experimental design of the repeatability, there were 12 samples with 7 repeats on a plate and that plate was run twice. In this instance it is the repeatability of the QS5 instrument that is being tested and so the informative experiments will have multiple runs on the QS5 rather than multiple samples on the same run. As such, this repeatability experiment has only carried out two repeats, which is less than required by best practice.

Despite the lack of definitions or information provided (as discussed above), from the display of results that are present then I believe the repeatability experiments were enough to determine the QS5 repeatability and decide that it was fit for implementation. I do not think there was enough information to make the statement that repeatability the same as the 7500, however this doesn't mean that unreliable results are being produced, nor that it shouldn't be implemented (as the important point is to demonstrate the repeatability of the QS5 instrument, rather than specifically comparing it to the performance of the 7500). I believe there is no evidence to suggest unreliability from this experiment. There is a lack of information available that can describe the amount of variability in the performance of the QS5. As explained in section 4.7 of this report, such a lack of knowledge can lead to an inability to answer questions of importance and do have the potential to be misunderstood as a lack of variability itself.

5.4 Experiment 3b: Reproducibility

The results for the reproducibility study are displayed appropriately in Figure 9. Graphically this result shows the reproducibility of the system and draws the conclusion that reproduced samples give comparable quantification. As with the repeatability study a coefficient of variation would have assisted in having a meaningful way to define this comparability.

I note the presence of one outlying result for sample 7. This is noted by the author and possibly attributed to any one of Quantification PCR variations, pipetting variation or QS5 detection anomalies.

The presence of unexplained outliers like this can cause concern in validations and it is not clear to me what any of the three possible situations listed as causes for the outlier mean. Are these factors that can be investigated, or their occurrence minimised? This may have been an instance where some further replication to determine whether repeated instances of anomaly occurred.

Quantification results have a downstream quality check in the DNA profiles they produce and so the overestimation of quant may not be such an issue. A different issue would be if the anomaly can also occur in the other direction (i.e., anomalously lower than expected) which could lead to the decision not to analyse a profile that did in fact have abundant DNA for profiling to occur.

5.5 Experiment 4: Y-Intercept Thresholds

A quick not here about the statement form the validation report:

> *The newly calculated Y-Intercept ranges for QS5 are considerably narrower than the current ranges, which is in part due to the relatively small number of plates used to calculate them.*

The lower sample size would not necessarily be the cause of a narrower range as the range is not based on the maximum and minimum values but rather a standard deviation. Therefore, lower sample numbers could well lead to give a higher standard deviations and hence larger ranges.

A point raised in the acceptance criteria was the additional data would be compiled and used to reassess the ranges. A check should be made to determine whether this was carried out, and if so, how it affected the ranges.

5.6 Correspondence:

There is one email here to the ThermoFisher representative and I have no comment.

There are also a number of comments given about this validation in the 'Instructions to expert' document I received as part of my brief. I mention this explicitly here as these comments do not appear to relate to the validation document itself, but rather the "Project Proposal #185 Validation of two QuantStudio™ 5 Real-Time PCR Systems - July 2018" document.

Comment 13)

The comment made is that the comparison is carried out results from different populations, an example of which is seen in Table 4 where entire dilution series are compared in the t-tests. This comment is correct, and I have already addressed the issue with carrying out comparisons in this way in my discussion of experiment 2.

Comment 15)

The comment notes that in the repeatability results shown in Figure 4 there appears to be some outlying results for DNA concentrations 0.09ng/µL and 0.009ng/µL (note this may in fact just be one outlying result if the 0.09ng/µL was diluted to create the 0.009ng/µL sample). However, because the graph shows the percentage change and not the actual quantification result it is not directly comparable to Figure 9 of the QS5 validation report (FSS.0001.0005.0767.pdf). I suspect that if displayed in the manner of Figure 9 of the QS5 validation report (given its relatively low concentration) that the apparent outlying nature of the result would not be so marked and would sit within the expected level of reproducibility for a sample of that quantification.

Comment 16)

The comment made here is that the comparison of reproducibility should be between the QS5A and QS5B instruments, whereas in the document FSS.0001.0005.0538.pdf, table 5 shows the comparison of QS5A and the 7500. The correct comparison to carry out depends on the acceptance criteria, or the aspect of performance being examined. It seems in the case of the results shown in Table 5 there was a desire by the authors to compare the performance of the new QS5 instruments to their previous instrument, the 7500. In this instance then the instruments that were chosen to be compared is appropriate, even if the multiple Student's t-tests are not the best way to approach the question (for the reasons previously specified). If the 7500 was being used as the gold standard for comparison, then the comparison to the QS5 instruments could have been carried out with a Krouwer plot.

Comment 17)

The comment here is that Table 3, when calculating the average percentage inaccuracies, incorrectly sums positive and negative values rather than absolute values. I believe the intent

of the average % inaccuracy value (putting aside issues of whether this is the best method to measure instrument performance) was to show which instruments were generally giving more inaccurate (or better put imprecise) answers. If this is the case, then comment 17 is correct. If two concentrations had equal percentage inaccuracies but in opposite directions (for example +10% and -10%) then the average that best represents these values (as I believe is the author's intent) is the average of the absolute values (in this case of the example this would be 10%) rather than the average of the signed values (in this case of the example this would be 0%). Averaging across the signed values will not provide information the precision. I note that this table did not survive to the final validation report and so presumably was not used in the final acceptance criteria.

5.7 Overall conclusion:

Again, although there were instances where inappropriate statistical tests carried out in all sections, I do not think this validation provides any evidence to suggest quantification values from the QS5 would be unreliable.

Seeing the outlying result in experiment 3b it may have been beneficial to carry out some additional troubleshooting or analyse some additional samples to ensure this was not an ongoing albeit sporadic issue.

Also, additional repeatability and reproducibility should be carried out, at least 3, and preferably 5 runs of samples on the same day, on the same instrument (for each instrument), by the same operator for repeatability. Then these same samples should be run by a different analyst on a different day for at least 3 to 5 days (as suggested in QH own SOP).

6.0 3500 genetic analyser validation work

Note that the first three validation reports for 3500xL that I have reviewed are actually the latter three validation that have been carried out chronologically. These validations (in 6.1, 6.2 and 6.3) were the original reports I received, and they made reference to earlier validations of the 3500xL instruments. I later received the first 3500xL validation documents (in 6.4, 6.5 and 6.6) which I then append to the end of the 3500 genetic analyser validation work section.

6.1 Project Final Report #145 - Verification of 3500xL B (FSS.0001.0006.2778.pdf)

Whenever an instrument is validated, the fitness for purpose of that validation depends in large part on how the instrument will be used. In the case of CE instruments (such as the

3500xl) the main tasks (in a forensic context) with be either to analyse evidence profiles, or to analyse reference profiles. The types of profiles being analysed under these two groups will have different features, for example reference profiles will generally be of high quality, undegraded DNA from a single donor. There will be thresholds used that can quite aggressively remove unwanted peaks (such as stutter peaks) and the analytical threshold can be set relatively high without the fear of missing low-level information that cannot be recovered by retesting. For these reasons, the validation of 3500xl intended for use with reference samples can be more accepting of variability in results, and instrument noise, as the data being interpreted is relatively straightforward.

6.1.1 Experimental design

This validation does not cover all the points suggested by SWGDAM shown in Figure 1 for a detection system, however it is not a full validation of the use of 3500xl instruments in the laboratory. This is an assessment of the performance of an upgraded instrument by comparison to an existing instrument of the same model and so the components of the validation do not need to be as extensive as a full validation would be. Additionally, as the 3500xl B is only intended for use on reference samples then some components of validation (i.e., testing casework-type samples) may not relevant. This is somewhat indicated by the fact that the work has been referred to as a verification rather than validation. Importantly the instrument validation covered all variants of use the instrument was intending to be used for (i.e., profiles generated by direct amplification and profiles generated from DNA extract).

The only component of validation that is commonly carried out that was absent from this validation was an assessment of the peak resolution of the instrument. Typically, there is a limited ability for capillary electrophoresis (CE) instruments to resolve peaks that are one base pair apart, and when one peak is significantly lower this can result in the lower peak not being detected. This is arguably not as critical to test for an instrument that is only to be used on reference samples due to the fact that it is expected both peaks would be roughly equal in height. I do not believe the omission of this component of validation would cause any issue with reliability of reference profiles generated on this instrument.

6.1.2 Experiment 1: Baseline, LOD and LOR — Direct Amplification

The method of baseline determination used is a very commonly used method. An often-cited publication that demonstrates the use 3 and 10 standard deviations for determining a limit of detection (LOD) and limit of quantitation (LOQ, which has been called limit of reporting LOR, in the QH validation report) is by Gilder et al [18]. In my comments on the validation of Quantifier Trio I explained that there two ways that LOD is typically considered; the point

at which instrument signal cannot be distinguished from analyte signal, or the lowest concentration at which an analyte can reliable be detected (based on a probability). In the calculation of LOD in this validation it is the former definition being used, i.e., the lowest point at which fluorescence from an amplified DNA can be distinguished from fluorescence associated with instrument noise.

Some of the maximum RFU values of baseline noise seen in table 4 seem very high. Perhaps this is something unique to the instrument, or perhaps artefacts were included in the baseline calculations.

In the results in table 5 of the validation report is described as the average across all dyes. I assume this is meant to mean that all data from all the dye lanes were considered in the calculation of the profile-wide average and standard deviation (as the values in table 5 are not the average of the values in table 4). If my suspicions are correct, then the profile-wide values in table 5 are fine. If the values in table 5 were calculated as the average across the summary statistics other the other dyes, then they would be in error.

When choosing a single threshold to apply across all dye lanes it is more usual to calculate a threshold for each dye and then choose the most conservative of these for the whole profile. This is due to the fact that choosing a whole-profile average will lead to some dye lanes (those with values greater than the whole-profile average) to be labelling baseline noise. However, the only dye lane with an individual threshold that is greater than the threshold finally implemented (i.e., 100RFU and 300RFU for LOD and LOR) is the orange dye lane. The orange dye would be fine to exclude from baseline calculations as it is not a dye lane from which allelic information is interpreted (it is a dye for use with internal lane standards).

The other point to consider is the purpose of choosing a LOR. When a human reads a DNA profile their goal is to remove artefactual peaks and retain peaks of interest. For references, the peaks of interest are alleles, and all other peaks are artefactual. In order to prevent the need to remove an abundance of artefactual peaks, most DNA profile reading software has the ability to employ rules and thresholds that screen some peaks out automatically. One such threshold is commonly referred to as an analytical threshold (or baseline, or LOR), which is designed to minimise the labelling of raised fluorescence coming from instrument noise. The higher this threshold is set the less baseline noise will need to be manually removed by a reader, but this comes at a cost of failing to label low-level alleles. Therefore, the analytical threshold is set by balancing reading efficiency with information loss (which leads to sample reworking). There is no ‘right’ value at which to set this threshold, and different laboratories will have different preferences or needs (see Taylor et al [19] for a discussion on setting analytical thresholds for CE instruments).

Ultimately, I believe the values chosen are appropriate.

6.1.3 Experiment 2: Baseline, LOD and LOR — Extracted Reference

As per experiment 1 – values chosen are appropriate

6.1.4 Experiment 3: Baseline, LOD and LOR — Casework

As per experiment 1 – values chosen are appropriate

6.1.5 Experiment 4: Peak Height Ratio, Allelic Imbalance and homozygote thresholds — Direct Amplification

Peak height ratio:

Peak height ratio (PHR) has been calculated using a common method. A limitation with calculating the PHR in this way is that all values will always be between 0 and 1 and are not normally distributed. Calculating a standard deviation on a set of numbers does not rely on the data being normally distributed, but if it is then used as determining a level of coverage (i.e., plus or minus 3 standard deviations should give a 99.9% coverage of the data) then this does assume normality of the underlying data distribution. Although not directly stated, I believe this might be the intent of the validation report authors when using the standard deviation in the way they have to calculate the allelic imbalance threshold, and hence then the homozygous threshold.

Literature suggests that a more informative manner of calculating PHR is the second (high molecular weight) peak height divided by the first (lowest molecular weight) peak height. This gives an indication of the direction of imbalance, but also importantly sets up the data in a way where the distribution of PHR can be modelled as a set of normally distributed data (with a log transformation).

The effect of calculating the PHR as the lower height divided by the higher height (as done in the validation) is basically showing half a normal distribution, which is likely to underestimate the standard deviation and hence the peak variability.

Having said that, this is method of calculating PHR which has been used a number of times by laboratories and in publications and continues to be used. Work by Kelly et al in 2012 [20] nicely explained the difference:

There are two well-known definitions of heterozygous balance and both are typically referred to as *hb*. In this paper we will use the definition

$$h = \frac{\phi_{HMW}}{\phi_{LMW}} \quad (1)$$

where *HMW* and *LMW* refer to the higher and lower molecular weight allele, respectively, and where ϕ is peak height. We will discuss the alternative definition in Section 5.1.

1639

1640 and later:

5.1. Comparison of methods used to calculate heterozygous balance

As mentioned above there are two main methods that are used within the literature to calculate heterozygous balance. The calculation:

$$h^{(2)} = \frac{\phi_{smaller}}{\phi_{larger}} \quad (6)$$

where $\phi_{smaller}$ and ϕ_{larger} represent heights of the alleles with the smaller and larger peak heights respectively, is common [24].

We suggest, however, that calculating heterozygous balance by this method results in a misleading representation of the data and a loss of information content. We attempt to explain our reasoning below.

1641

1642 Allelic Imbalance:

1643 Allelic imbalance (AI) has been calculated using equation 4. As this relies on the PHR value
1644 and given the explanation in the PHR section above if the intention was to model 99.9% of
1645 the imbalance based on the mean and standard deviation of the PHR then the actual coverage
1646 may be lower than this. There is also a dependence of peak height variability with peak
1647 height. It is common for allele imbalance thresholds to only apply down to a particular
1648 height. Below (again from Kelly et al 2012 [20]) demonstrates this.

1649

4.1. Comparison to the conventional thresholds

Analysis was undertaken to determine where the conventional thresholds and the 95% credible intervals derived in this work intersect. It should be noted that the traditional bounds were developed for SGM+™, whereas this work uses Identifiler™. In the current model the traditional lower boundary of 0.60 is 'safe' when the average peak height is above 434 rfu. The upper boundary of 1.67 is 'safe' when the average peak height is above

1650

1651

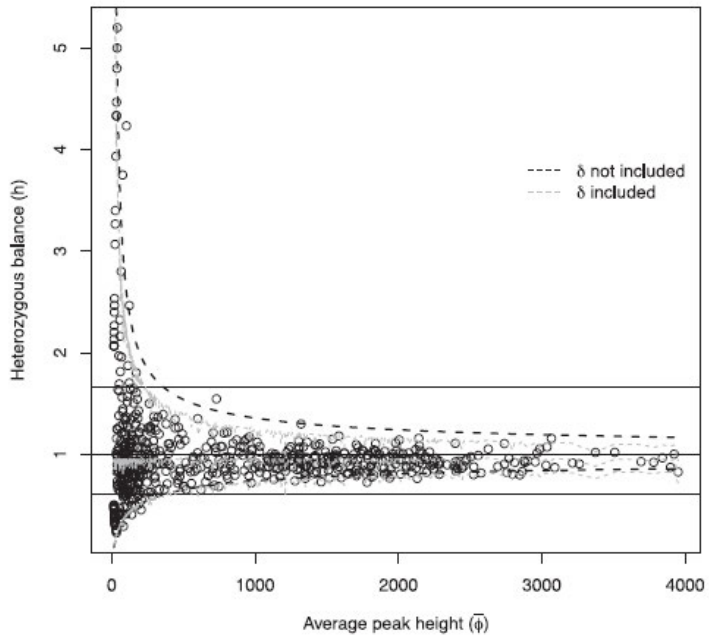

Fig. 2. Fitted values from the lognormal model with and without the effect of δ. The horizontal lines represent h thresholds (0.60, 1.67). The dashed lines indicate the 95% credible intervals determined in this work.

174 rfu. Fig. 2 shows the 95% credible intervals and the conventional h thresholds with respect to $\overline{\phi}$. The graph shows that the distribution of h is not even, and widens at low $\overline{\phi}$. The conventional thresholds can be supported when they are outside the 95% credible intervals but not once they intersect. The graph also shows that above approximately 1000 rfu, the 95% credible intervals become much narrower. These narrower guidelines could be useful in the interpretation of complex mixtures at higher levels as fewer genotype combinations would be considered possible.

Homozygous threshold:

The formula (given in equation 5) is not explained (or may be done so in previous validation reports). The homozygous threshold has been set as the LOR divided by the AI value (which would provide the height of a peak that is not expected to have imbalanced below the LOR at a particular probability). The multiplication by 2 is the part that is not explained and may just be an arbitrary conservativeness factor.

Note that any method for determining a homozygous threshold has a level of arbitrariness to it, and at its base is a decision based on probability that weighs up the risk of an incorrect homozygous designation against the need to carry out unnecessary laboratory work. In other words, a homozygous threshold determines when a single peak at a locus is deemed homozygous and is set at a level than minimises (but does not eliminate) the possibility that a locus designated as homozygous has instead come from a heterozygous individual where one

peak has not been detected. If the threshold is set too high, then too many homozygous loci will not reach the threshold and will cause unnecessary reworking of samples. Therefore, the threshold balances the effects of unnecessary reworking against incorrect homozygous designation. There is no standard level at which this balance is set (except for the broad agreement that it is heavily weighted towards reworking samples and not incorrectly designating a homozygous locus). In this respect the homozygous threshold set using the method in the QH validation will achieve this requirement, and is not incorrect, it just may not be designating homozygotes at the probability that is expected / intended.

Another commonly performed test when setting homozygous thresholds is to model the observed rate of drop-out against the expected rate of drop-out from the model. Then the homozygous threshold can be shown to align with its theoretically desired dropout probability. This has not been done in the QH validation; however, it is not a requirement to do so, simply a method that some laboratories find useful to demonstrate the performance of their chosen threshold. An example can be seen in Taylor et al [19].

Given the conservative nature in which the homozygous threshold has been set, and that there is no 'right' value to set such a threshold, I believe the threshold chosen is reasonable. There is no evidence to suggest interpretations based on this threshold are unreliable. At some point QH should model their observed dropout rates to demonstrate exactly how their threshold is performing.

6.1.6 Experiment 5: Peak Height Ratio, Allelic Imbalance and Homozygote Peak Thresholds — Extracted Reference

Same comments as for experiment 4

6.1.7 Experiment 6: Peak Height Ratio, Allelic Imbalance and Homozygote Peak Thresholds — Casework

Same comments as for experiment 4

6.1.8 Experiment 7: Concordance

This work is appropriate, and I have no comment.

6.1.9 Experiment 8: Sizing Precision

As a general comment it is best to have a pass criterion that is not dependant on any other instrument. Setting absolute criteria means there is some known level of performance that have an absolute meaning i.e., otherwise you could say that 3500xl B is at least as good as 3500xl A, but does that mean that they are both performing well or poorly. Presumably both instruments are performing well, and this information might be in the first validation report as there was no other instrument to compare back to at that point. A criterion that was set based on an amount of variability means that both instruments could independently be compared back to it and shown to be performing well, and then as a secondary point a comparison to machine A would be made.

Additionally, while the bar charts seen in Figures 4 to 23 are fine, boxplots would be more informative of the distribution of values obtained, rather than just the average (however this is a very minor point).

The results from this section have been appropriately interpreted. Looking at the standard deviation values, the instruments look to be performing in line with developmental validation carried out by the manufacturer.

6.1.10 Experiment 9: Repeatability and Reproducibility

Peak heights were used to compare samples from different instruments using Student's t-tests. I doubt they are normally distributed and so a non-parametric test might have been more appropriate. The non-parametric equivalent of a Student's t-test is a Mann-Whitney U test, the non-parametric equivalent of a paired t-test is the Wilcoxon Rank sum Test and a non-parametric equivalent of a ANOVA is a Kruskal-Wallis H test. In this instance the Wilcoxon Rank sum Test would have been appropriate.

Carrying out Student's t-tests (even on appropriate datasets) with a significance level of 0.05 have a 5% chance of showing a significant difference when none exist. As I understand there were 25 samples compared as part of this study, and so if none of them had any difference in peak height you would expect (by chance) one to have a p-value < 0.05. This is the classic occurrence of the multi-testing problem. There are ways to overcome this, but potentially pooling all samples and comparing using a Wilcoxon Rank sum Test could be done. Alternatively, a p-value correction can be carried to (such as a Bonferroni correction), however these are often criticised for being overly conservative i.e., fail to reject the null hypothesis more often than is reflected in reality.

In this case 8 samples did show a difference using the t-tests, which is more than expected by chance if there were no difference. However, it is well known that individual injections and different capillaries will exhibit small variations that affect the profiles produced. Bright et al [21] showed that a sample injected using one instrument in one injection, but different capillaries, showed profile to profile variation. They also showed that one sample injected

multiple times by the same capillary on subsequent injection exhibited profile to profile variation. The best way that the performance of the two instruments can be meaningfully compared is by statistics that pool across multiple samples (rather than looking on a sample-by-sample basis).

Perhaps the best way to show a non-significance has already been partially carried out (seen in Figures 24 to 26), which is to plot the distribution of peak height differences and show that the distribution is not significantly shifted from 0 (by drawing lines where the upper and lower 2.5% of the data lay, and that 0 is within these bounds). Another way could have been to produce Bland-Altman plots.

Therefore, I believe the conclusions drawn are correct (even if the use of Student's t-tests was not). There is no evidence to suggest unreliable DNA profiles would be being produced based on this analysis.

6.1.11 Cross-talk

This work is appropriate, and I have no comment.

6.1.12 Overall conclusion:

As with previous validations, there are instances of statistical tests being used that are not the most appropriate for the data. A point to consider for follow-up is whether the homozygous threshold chosen has the desired level of coverage. This can be carried out by analysing reference profiles that have already been analysed and determining a drop-out rate, and then comparing this to the expected rate at the homozygous threshold. It could also be calculated from the existing data, but using the appropriate measure of heterozygote balance. I suggest this as the method used to develop the threshold may have made an assumption about the underlying data that is not appropriate. However, I suspect that given the conservativisms built into the derivation of the homozygous threshold that this will not be an issue. A check of the performance of the existing threshold is likely to be available by auditing database matches (which are typically flagged when one mismatch is present, and so will identify the rate of misassigned homozygous peaks).

There is no evidence to suggest DNA profiles produced by this machine would lead to unreliable opinions.

6.2 Project #182 - Validation of 3500xL Analysis of Casework PowerPlex®21 WEN (FSS.0001.0006.4152.pdf)

6.2.1 Baseline, Limit of Detection and Limit of Reporting

These baseline levels appear much more in line with typical CE baseline levels (for example see [22]). Clearly new PP21 formulation has improved the performance of the kit, and it suggests that any issues that were identified in the previous validation study were likely kit issues, rather than 3500xl issues.

The acceptance criteria section seems not to be complete. Regardless, I assume the baseline values have been accepted, and this is appropriate given the data.

6.2.2 Stutter

In the acceptance criteria section, there are claims that the stutter thresholds are not affected by model of CE (e.g., 3130 vs 3500xl), but rather the profiling kit. This statement is backed up in by the findings in Bright et al [23] (which could be added to the validation report to provide some justification for the statement).

Implementation of the higher SR out of Promega and calculated values (as shown in table 5) is a reasonable practice to adopt. When reading references, the main task is to remove stutters, with the only counterbalance being the avoidance of removing an allele in an imbalanced heterozygous pair. Usually, the ratio of stutters to their parent peaks, and the ratio of peak heights from a heterozygous pair are quite well separated and so the stutter ratio filters can be set conservatively high to ensure stutter peaks are removed in references. Given the earlier calculated allele imbalance thresholds, interaction with stutter peaks will not be an issue.

The stutter ratios appear to have been set appropriately. There are also minimal risks to setting these ratios too high or too low. If they are set too low, then stutters will tend to be left labelled on reference DNA profiles and it would become obvious that the ratios were too low. If the ratio was set too high, then peaks in heterozygous pairs would be removed and additional reworks would be requested (which would then show that the original peak was an allele) and it would become obvious the ratios were too high. In either scenario, human assessment of the profiles would ensure correct profile information was being obtained.

6.2.3. Peak Height Ratio

The PHR calculated here has same issue as I mentioned for the previous validation (Verification of 3500xL B). In addition, I am not sure whether the 'All loci' results have been obtained by averaging the values for the individual loci. If this has occurred then it is not the best representation of the entire dataset for two reasons, 1) the number of observations for the loci are different hence the categories are not equally weighted and 2) the average of standard

deviations is not the standard deviation of the whole dataset. As with my earlier comment on setting a whole-profile baseline from individual-dye baselines, there is an argument that rather than using an average across the whole profile, a better approach might be to choose the most conservative individual value (in this case for an individual locus).

Despite any issues I have raised above, I do not believe there is any evidence to suggest unreliable DNA profiles are being produced.

6.2.4. Homozygote Peak Threshold

The version of the validation document I have in incomplete and doesn't have the specific value that the homozygous threshold was set. This is inconsequential, and no matter what value it is set, I mirror my comments from the homozygous threshold setting section in the verification of 3500xL B.

6.2.5. Repeatability and Reproducibility

I have the same comments on use of Student's t-test for repeatability and reproducibility as were given in the validation report for the 3500xl B.

A very minor point with the sentence:

> *Nine samples showed there was a significant difference ($p \geq 0.05$) and 16 samples showed there was no significant difference ($p \leq 0.05$) in peak heights between run 1 and run 2.*

Is that the inequality signs are both reversed from what they should be.

Note that Fig 4 suggests there was an overall difference in peak heights between runs as it appears there is a significant shift from 0. Given what looked like a significant shift in peak heights for repeatability it may have been worth running at least a third plate (and possibly fourth) to show the spread of peak height variability from repeated runs. The finding of a difference in peak heights between runs is not unexpected (again, see Bright et al [21]), and would not prevent implementation of the instrument, however just two runs hasn't given a clear indication of the level of repeatability of the instrument. I note that the instrument was not being implemented into casework and so the issue of reproducibility is less significant (as references are much clearer to interpret and usually have abundant opportunity for rework if required by any ambiguity in the profile).

6.2.6 Overall conclusion:

As with previous validations, there are components that have not used formally appropriate statistical methods. Also, in parts I believe the validation has not been large enough to show the range of reproducibility of results. There is no evidence to suggest unreliability of DNA profiles being produced.

However, additional repeatability and reproducibility should be carried out, at least 3 (and preferably 5) runs of sample on the same day (for each instrument), by the same operator for repeatability. Then these same samples should be run by a different analyst on a different day (for each instrument) for at least 3 to 5 days (as suggested in QH own SOP).

6.3 Project Report #186 - Assessment of 3500xl A Genetic Analyzer for Processing Casework Powerplex® Samples (FSS.0001.0006.4174.pdf)

There are various guidelines for the components of validation of probabilistic genotyping (PG) systems [24, 25]. It is always difficult to separate the performance of just the CE from the profiling system, due to the fact that the most common way to validate the CE is by visualising profiles, which have been produced by profiling kits. Therefore, the behaviour of the final results being evaluated are unavoidably a combination of the performance of both components. A good example of exactly this effect is the determination of baseline for the 3500xl B in a previous validation report, which was then subsequently greatly improved by a modification of the profiling kit, rather than any change in the CE itself.

A reasonable range of mixture proportions have been created for this validation. A quite in-depth analysis of the ability to interpret CE data produced on the instrument is given. From the summary of the findings, the CE would appear to be performing at a level that was reasonable for use in casework. I note that in the recommendations it is stated that the 3500xl-A should not be implemented in casework given some complicating factors with interpretation. However, I feel that the use in casework could have been recommended as long as analysts were provided with appropriate supporting documentation that allowed them to gain an understanding of the performance of the 3500xL instrument.

In the appendix an impressive, in-depth and considered analysis of baseline data in the development of LOD and LOR is given. The analysis takes into consideration of the effects of different dyes, and accounts for the skewness of the data.

I think the overall assessment of the 3500xl-A is appropriate.

6.4 3500xL Genetic Analyzer Validation for Reference samples Amplified with Powerplex®21 using Direct Amplification - February 2015 (FSS.0001.0025.9181)

This validation does not include mixture studies (suggested by SWGDAM, show in Figure 1 of my report, as an appropriate factor to test for detection systems) however the validation is for use on reference samples, and hence mixture studies are not required.

Many of the acceptance criteria were initially based on the comparability to the previous model of capillary electrophoresis instrument (the 3130xl). The level of similarity with the 3130xl is not defined and therefore the acceptance of the 3500xL is somewhat arbitrary. For example:

> *The 3500xL AI_{TH} for direct amplification was calculated as 54%. The current AI_{TH} for direct amplification on the 3130xl is 60%. The 3500xL passes this experiment as these thresholds are comparable.*

In many instances there are differences between the 3500xl and the 3130xl and this is not unexpected given the different generation of instrument and the different dynamic range over which the two instruments operate.

> *For +1 repeat stutter the 3500xL thresholds did vary from current 3130xl thresholds. This has resulted from […]. Given these analysis issues, the +1 repeat stutter thresholds must be accepted as calculated and cannot be used to assess whether the 3500xL has passed or failed this experiment.*

These types of issues raise the question of why even set an acceptance criterion. The issue here is not the performance of the 3500xL but rather one of setting appropriate criteria. As mentioned in comments to other validation reports it is best to set acceptance criteria based on absolute values of performance rather than as a comparison to an existing instrument (unless the fundamental passing criteria must be a similar performance to another instrument and not a particular level of performance).

6.4.1 Baseline, Limit of Detection and Limit of Reporting

As with other 3500xL validations, the determination of LOD and LOR have been carried out in a standard manner and appear to be appropriately set.

6.4.2 Stutter

Stutter thresholds have been calculated in a standard manner and appear to be set appropriately.

6.4.3 Peak Height Ratio

The peak height ratio calculations here have the same issues as described in comments on previous 3500xL validations.

6.4.4 Homozygote Peak Threshold

The homozygote peak threshold set here has the same issues as described in comments on previous 3500xL validations as it is based on the peak height ratio. I suspect this threshold is probably functioning appropriately but does not necessarily have the intended coverage. A check of the actual coverage should be performed so that any decisions based on this threshold can be made using correct information.

6.4.5 Concordance and Sensitivity

The concordance has been carried out appropriately. I am not sure why the sensitivity has been included in the title of this section. Typically, a sensitivity study for a 3500xL would involve testing the instrument with a range of input DNA to see at what concentration peaks are no longer distinguishable from baseline. I would consider the sensitivity component unvalidated.

6.4.6 Drop-in

No drop-in peaks were observed across the validation dataset, however only 11 negative amplification controls were analysed. Given this number of controls, and that a common drop-in rate for PCR negative controls on a 3500xL would mean drop-ins were seen once in every 30 profiles [26]. It is not uncommon that drop-in may not be seen during validation, and in such an instance a default level that aligns with literature finings might be used for casework. The important factor to follow-up is the regular reassessment of drop-in rates from negative controls in casework (i.e., every 6 months is common).

6.4.5 Artefacts

The assessment of artefacts is generally observational (i.e., reporting what was observed). This section seems to have been carried out appropriately.

6.4.6 Off Scale Allele Peaks

This section is generally observational (i.e., reporting what was observed). This section seems to have been carried out appropriately.

6.4.7 Repeatability and Reproducibility

My comments for this section are the same as comments for repeatability and reproducibility in other 3500xL validations (for example, section 6.1.10). Multiple student's t-tests carried out on a per-profile basis are not the appropriate way to assess the repeatability or reproducibility, and an appropriate plot could have been used to provide a means to assess the performance.

The acceptance criteria:

> *The 3500XL has passed this experiment because allele designations for repeatability and reproducibility were completely concordant.*

Is inappropriate as concordance of alleles is not a standard criterion for assessing repeatability or reproducibility of a 3500xL.

6.4.8 Carry Over and Cross Talk

There is little here to comment on. No observations of these phenomenon were observed, and this was reported, which is appropriate for the validation. Again, these phenomena usually continue to be monitored after implementation.

6.4.9 Overall conclusion

As with previous 3500xL validations there are aspects of this validation that have not been carried out to recognised standards. In particular the assessment of sensitivity, repeatability and reproducibility have not been carried out appropriately.

There is nothing in the validation that makes me think the results produced on the 3500xL are unreliable. Given this validation is being used for reference samples, and that this type of DNA profile is the easiest to interpret I do not believe there would be an issue with the implementation of this 3500xL into casework. However, the validation has not given the analysts a good understanding of the performance of this specific instrument given the lack of sensitivity assessment and the inappropriate measure of repeatability and reproducibility. This lack of knowledge can lead to sub-optimal decisions being made or information being provided, particularly when troubleshooting an unexpected, or unusual result.

Additional repeatability and reproducibility should be carried out, at least 3 (and preferably 5) runs of sample on the same day (for each instrument), by the same operator for repeatability. Then these same samples should be run by a different analyst on a different day (for each instrument) for at least 3 to 5 days (as suggested in QH own SOP).

I note in the recommendation that the authors suggest:

> *Conduct a post implementation review to determine if thresholds set in this report are appropriate when more 3500xL data is available. The post implementation review should also include determination of thresholds which were not possible in this project due to sample size (i.e. some -2 and +1 repeat stutter thresholds and N-2 and N+2 artefact thresholds).*

I suggest the Commission check that this post implementation review was carried out.

6.5 Genetic Analyzer Validation for Extracted Reference Samples Amplified with PowerPlex®21. Forensic DNA–Analysis - June 2015 (FSS.0001.0026.0656)

This validation has many of the same components as the validation outlined in section 6.4 (baseline, LOD, LOR, stutter, PHR, allelic imbalance, and homozygote threshold), that have the same type of data produced, the same statistical analyses carried out and the same issues as I have previously highlighted.

There were some components of the validation in section 6.4 that were not retested in this validation (concordance and sensitivity, drop-in, artefacts, off-scale peaks, repeatability and reproducibility, carryover and cross-talk). These omissions are reasonable given that they are generally more associated with the instrument, rather than the process.

There were two additional components in this validation that I detail below.

6.5.1 Sizing Precision Comparison

The sizing precision analysis appears to have been carried out appropriately and has given precision values similar to other published studies.

6.5.2 Comparison of Peak Heights between 3500xL and 3130xl

The 3500xL is known to yield higher peaks that a 3130xL. Stochastic effects associated with capillary electrophoresis will mean that there is some variability in the exact ratio, but from my experience the peak height increase is generally between 3 and 10-fold. A comparison of

peak heights between a 3500xL and 3130xL may be of interest to show in a validation, so that the range of peak height increase is known, however it is not likely to lead to the rejection of an instrument, and there is no specific criterion that such a comparison must meet. I am not sure that the claim in the acceptance criteria that:

> *There was a linear relationship (characteristic) between the peak heights of the 3500xL and 3130xL.*

Is well supported by the data, but showing this as a linear trend is not important. However, having produced the plots shown in Figure 46 and 47 there appears to be a small outlying cluster of points, which is likely to be from a single profile and would have been good to track down to see what has caused the slightly unusual performance.

6.5.3 overall conclusion

As with previous 3500xL validations there are aspects of this validation that have not been carried out to recognised standards.

There is nothing in the validation that makes me think the results produced on the 3500xL are unreliable. As with the previous validation, there is a recommendation given for post-implementation follow-up and I suggest the Commission check that this post implementation review was carried out.

6.6 3500xL Genetic Analyzer Validation for Casework Samples Amplified with PowerPlex®21. Forensic DNA Analysis - September 2015 (FSS.0001.0006.2855)

The baseline, LOD, LOR, stutter, peak height ratio and homozygous peak threshold, repeatability and reproducibility sections are very similar to those for the validations outlines in section 6.4 of my report and I have the same comments on the experiment and data analysis.

6.6.1 Drop In

In this validation there were 54 negative controls tested and two instances of drop-in detected above the LOD. This is an approximately expected level of drop-in for a 3500xL.

6.6.2 Saturation

There was no saturation level set and a note that:

"Further analysis of this data was required to determine the STRmix™ saturation setting."

Therefore, there is nothing to comment on here.

6.6.3 Mixture Study: Number of Contributors

This section of the validation was not carried out due to issues with a type of instrument artefact known as pull-up affecting the ability to interpret DNA profiles. There is nothing to comment on here.

6.6.4 Mixture Study: Two and Three Person Mixtures

This section of the validation was not carried out due to issues with a type of instrument artefact known as pull-up affecting the ability to interpret DNA profiles. There is nothing to comment on here.

6.6.5 Overall conclusion

The only contribution that this validation has provided to the existing validations of the 3500xL is the observation of drop-in. Other components that are specific to casework (such as mixture analysis) were deemed unsuitable due to the performance of the instrument. I note that this was later resolved with a reformulation of the profiling kit and presumably a reanalysis of the spectral calibration of the instrument (as indicated in the validation reviewed in section 6.3 of my report).

7.0 Project #148 – to optimise the cleaning protocol for bone crusher vials (FSS.0001.0056.9408.pdf)

The validation looked at the effectiveness of different techniques on the ability to remove DNA from bone crushing equipment. The investigation was relatively simple, but that is all that is warranted as the question is relatively simple one to answer.

In the investigation into whether there was a significant difference between the performance of the dishwasher and the Tergazyme, Student's t-tests were performed on heterozygote balance, stutter percentage and peak heights. The heterozygous balance levels were calculated in a sub-optimal, but commonly applied manner (as mentioned in previous comments for the 3500xl-B validation). Some tests for approximate normality could have been run to ensure appropriate use of t-tests, but I think in this instance they were valid.

Therefore, the findings of non-significance are appropriate (in fact the non-significance of peak height can be seen by the amount of overlap between the two groups in Figure 3).

There was a good comparison of results to those found in published literature.

The investigation into whether the different cleaning methods affected profile quality was small but would have shown if there were any dramatic effects. A larger study would have been required to detect mild effects on profile quality, but I do not necessarily think this is warranted.

I believe this validation was conducted appropriately, and the conclusion was supported by the data.

8.0 Project Proposal #173 Validation of Hamilton STARlet A for Quantification and amplification Assay Setup January 2017 (FSS.0001.0047.1056.pdf)

8.1 Experiment 1: Verification of STARlet A with ARTEL MVS

Validation has used a professional validation tool (Artel MVS) which appears to have specified passing criteria based on coefficient of variation, which is appropriate.

Again (as with Quant Trio and HS5 validations), I am not familiar with the term ‘percentage inaccuracy’, but I suspect this is just the term being used for percentage change and I am taking it as such. In that case the passing criteria also seems to be set by the Artel MVS and again seems an appropriate value.

The number of repeats seems appropriate according to published expectations [12].

8.2 Experiment 2: Preparation of DNA quantification standards

No comment – all seems appropriate

8.3 Experiment 3: Quantifiler® Trio Contamination Checks

It would have been informative to profile the one sample in the EXP 3a that gave a quantifiable DNA amount to see whether a profile was obtained (it may have been a very low-level spurious result in the quantification of the sample). If it did produce a profile a check could have been carried out to see whether it aligned with the DNA being used in the zebra pattern (as, if not, it may have been a drop-in or individual tube contamination from the

manufacturer, or staff). Regardless, if the zebra configuration is not being used for casework, then perhaps it is of limited consequence. I also note that the quantification value is very low (and below LOD according to the discussion) and so presumably would not have been sent to DNA profiling had they occurred in casework, however there is always the opportunity for contamination to occur where the contaminating amount is low but occurs into a sample with other DNA. This would mean that the sample would pass quantification LOD criteria and be profiles, and the contaminating profile could then be detected in the sample as an additional contributor. Therefore, the fact that the sample with the quantifiable level of DNA fell below the quant LOD does not completely safeguard against contaminated samples being profiled.

Despite the fact that some additional investigation may have been informative, I believe the passing of the instrument based on these results seems appropriate.

8.4 Experiment 4: Profiler® Plus Contamination Checks

No comment – seems appropriate

8.5 Experiment 5: PowerPlex® 21 Contamination Checks

No comment – seems appropriate.

8.6 Experiment 6a: Quantifiler® Trio Repeatability

The authors have identified the difficulty with measuring repeatability on just the Hamilton instrument (as total variability observed accumulates variability from the Hamilton, but also Quantifiler trio). Again, here Bland-Altman plots may have been useful in visually showing repeatability results. This would have also made it possible to set some meaningful acceptance criteria that took previously observed variability into account.

Despite this I believe this experiment has been appropriately passed.

8.7 Experiments 6b & 6c: Quantifiler® Trio Reproducibility and Performance Check

There could have been formal statistical tests carried out to assess reproducibility, however in this instance I agree with the authors that a visual representation of the results is all that is warranted to be satisfied with the performance of the STARlet A. As mentioned previously, a boxplot would have been more informative that a bar chart, but this is a minor point.

8.8 Experiments 7a & 7b: Profiler® Plus Repeatability and Reproducibility

There could have been formal statistical tests carried out to assess reproducibility, however in this instance I agree with the authors that a visual representation of the results is all that is warranted to be satisfied with the performance of the STARlet A.

8.9 Experiment 7c: Profiler® Plus Performance Check

There could have been formal statistical tests carried out to assess reproducibility, however in this instance I agree with the authors that a visual representation of the results is all that is warranted to be satisfied with the performance of the STARlet A.

8.10 Experiments 8a & 8b: PowerPlex® 21 Repeatability and Reproducibility

There could have been formal statistical tests carried out to assess reproducibility, however in this instance I agree with the authors that a visual representation of the results is all that is warranted to be satisfied with the performance of the STARlet A.

8.11 Experiment 8c: PowerPlex® 21 Performance Check

There could have been formal statistical tests carried out to assess reproducibility, however in this instance I agree with the authors that a visual representation of the results is all that is warranted to be satisfied with the performance of the STARlet A.

8.12 Experiment 9: Dilution Plate Performance Check — PowerPlex® 21 and Profiler® Plus

No comments – all seems fine

8.13 Experiment 10: Testing Kits Performance Check — Quantifiler® Trio, PowerPlex® 21 and Profiler® Plus

No comments – all seems fine

8.14 Overall conclusion

In my opinion the instrument was appropriately validated.

9.0 Project Proposal #173 Verification of Hamilton STARlet B for Quantification and Amplification Assay Setup July 2017 (FSS.0001.0047.1114.pdf)

Very much a repeat of the STARlet A validation experiments 1 to 5 and all carried out appropriately. Again, there was one quant >0 identified on the contamination check that would have been informative to profile. But even without the profile information I think the validation showed the instrument was fit for purpose.

10.0 Project Final Report #175 Validation of Hamilton® STARlet C for Capillary Electrophoresis Setup April 2019 (FSS.0001.0047.7168.pdf)

Experiment 1: Verification of STARlet C with ARTEL MVS

Similar to the previous STARlet validation tests for pipetting accuracy – all carried out appropriately

10.1 Experiment 2: Plate Seal Assessment

No comment – conclusions seem appropriate given the performance of each seal

10.2 Experiment 3: Whole Plate Transfer PowerPlex® 21

The comparison of outcomes of profile reads resulting from plate transfers is carried out only using high level summaries i.e., the percentage of reworks on the transferred plate compared to a manually prepared plate. The values in the tables do not suggest any issues with the plate transfer ability of the STARlet C.

There could have been some additional checks of the properties of the profiles to ensure that the summary results didn't obscure any underlying, unexpected trends. For example, it would be expected that the majority of samples requiring reworks on the transferred plates were the same as those that required reworks on the original plates, if this were not the case then some further investigation might be warranted. Similarly, the profiles where more or less alleles were detected in the transferred plate compared to the manual plate you may expect to be due to profiles that were low-level and had peaks affected by stochastic height variability so that by chance, they fell either side of the LOR.

However, the absence of this information from the validation does not suggest any unreliability in plate transfer.

10.3 Experiment 4: Re-CE PowerPlex®21

All seems appropriate

10.4 Experiment 5: Repeatability and Reproducibility

There could have been some summary statistic of the variability given, such as cov that would have been informative. However, this would be a combination of the variability due to pipetting and the performance of the CE and so would only be partially informative of just the pipetting aspect. Still, as the comparison is being made on the spread the points anyway, then the cov would at least give a simple criterion to accept on that was less subjective.

Despite this the analyses all seem appropriate.

10.5 Experiment 6: Plate Piercer Assessment

All seems appropriate

10.6 Overall conclusion

In my opinion the STARlet C was appropriately validated.

11.0 Project Report#199 – Verification of ProFlex™96 Well PCR System using PowerPlex®21 December 2021 Version 2.0 (FSS.0001.0048.6801.pdf)

When considering the validation of performance of a PCR thermocycler, it is common to find literature on the validation and calibration of the temperature profile produced by the instrument. This aspect is not of direct interest in validations carried out in a forensic setting (beyond the fact that the thermocycler must have passed these calibration tests in order to be used). Instead, the performance of a thermocycler in a forensic context is typically gauged by the quality of a DNA profile that has been produced. There is some evidence showing that individual thermocyclers of the same model provide approximately equivalent levels of peak height variability in the DNA profiles produced [27].

The complexity of validating a PCR thermocycler instrument is that in order to carry out such a validation some material must be amplified, and once amplified the material must be visualised (typically by the use of capillary electrophoresis). In other words, the quality of the

final product being assessed (i.e., a DNA profile) will depend not only on the performance of the thermocycler, but also the DNA profiling kit and the performance of the capillary electrophoresis instrument. Therefore, experiments must be set up in a way that would reveal trends in data that can be attributed specifically to the PCR thermocycler. These trends may be difficult to observe in amongst the effects of the profiling kit and capillary electrophoresis instrument and may be best revealed with the use of probabilistic genotyping software (such as STRmix) even if the probabilistic genotyping system is not the focus of the validation. It is common that in the absence of observing any effects in DNA profile quality that can be specifically attributed to the PCR thermocycler performance, that an opinion be reached that there is no evidence to suggest an underperformance of the thermocycler. This is an acceptable opinion to reach in this situation.

11.1 Average Peak Heights

The plots typical of those seen in Fig 1 to 4 show some variability in peak heights between PCR thermocycler instruments. The data is not sufficient to address the question of interest i.e., do the instruments perform to the same level as the 9700. What can be said from the data is that the range of performance from different Proflex instruments encompasses a typical 9700 results, which is a comment on the group of instruments rather than the performance of any specific individual instrument. To explain further, these graphs (and the similar graphs shown in the appendix) often show the red (Grumpy) instrument to produce peak heights near, or at, the lower range of all Proflex instruments and the blue (Bashful) instrument appears commonly at the top. There may be a difference in performance in these two instruments that has not been explored. Also, it would be informative to know the range of peak heights expected from the same samples at the same dilution amplified multiple times on the same instrument. It may be the that the amount of peak height variation across all the ProFlex instruments is within the amount of expected variation seen in peak heights produced from multiple amplification on the 9700 (and therefore of no practical significance, even if a statistical difference between instruments was found).

On page 5 the comment is made "*An R^2 value close to 1 indicates similarity between the ProFlex and 9700 average peak heights*". This is not correct, in fact an R^2 value close to 1 simply means the variability of the points around the trendline is small compared to the overall variability (put simply, the points will fall close to the trendline). It would only be if the slope of the trendline was close to 1 (and the intercept was 0) that similarity between instruments would be indicated. For example, in Figure 5 the line that would most closely indicate similarity (i.e., the line of equality, which passes through the points [500,500], [1000,1000], [1500,1500], etc) is for the green (Sleepy) Proflex, and it has an R^2 value of 0.9731. The trendline that is the furthest from the line of equality (i.e., indicating it is performing more differently from the 9700 than the sleepy instrument) is that of the teal (Bashful) Proflex, but it has a higher R^2 value of 0.9829.

11.2 Inter-locus balance

The graph shown in Figure 11 demonstrates inter-machine consistence of profile balance. Coupled with the peak height component to the study suggests that much of the variability in peak height seen in the previous section is due to absolute differences in the amount of DNA present (and not stochastic variability caused by PCR setup). Again, it might have been useful to know the level of expected variability for a single instrument before assessing all together as a group. There is also the ability to STRmix (which has an inbuilt indication of the amount of inter-locus peak height variability) here, but this is not a requirement, it just may have provided an easy alternative.

If the results shown in Figure 11 are the averages across all 53 profiles, then there will be a very strong stabilising effect of the averaging. Ideally the graph would show error bars on the plot to give an indication of individual profile variability.

11.3 Stutter

The analysis that has been carried out is valid, but only shows the upper bound and so some trends in stutter peaks would not be observable (i.e., if there was a tendency for stutter to be reduced in any one instrument). The upper bound is certainly more important from a reference point of view, but for evidence samples it would be helpful to demonstrate no great shift in expected ratios, or in variability in the ratio. Again, this could have been achieved in STRmix, particularly the use of Model Maker, which assesses the stutter peak height variability.

11.4 Concordance, Drop-In and Drop-Out

Drop-in assessment was reasonable. It would have been useful to calculate a drop-in rate and compare that to existing rate from the 9700. However the rate at which peaks were observed, and their height is quite typical of drop-in for kits of this generation [28].

There was no real assessment of drop-out as none of the profiles were in the dop-out range and so I would consider this aspect largely untested.

11.5 Allelic Imbalance

Again, it would be useful to have an indication of the spread of PHR values per instrument (e.g., in the form of error bars) shown on the graph, or as a graph type used that shows the data spread (such as a boxplot). There were no dramatic outlying results here and so I believe the assessment of the performance of the allelic imbalance in profiles generated on the

Proflex systems is performing adequately. As before, STRmix could have been used to provide a more robust measure of peak height variability.

11.6 Artefacts

This is a subjective assessment, but that seems fine for this aspect

11.7 Experiments and Results – Reference Protocols

This component appears to have been carried out appropriately

11.8 Summary – Model Maker results for Project #199

The performance of a thermocycler can affect PCR kit performance. Given the use of STRmix to analyse all profile data, in this instance a test of the effect on profile performance (and a check to ensure no major recalibration of STRmix) was warranted. It appears that the ProFlex data was analysed in Model Maker in January 2022. Model Maker is a component of software STRmix, which is used during the calibration stage of setting up STRmix for use in a laboratory. In essence Model Maker takes a set of profiles produced in the laboratory (for which the reference profiles of the DNA sources are known) and uses that to determine the level of peak height variability in profiles being produced for that laboratory. There is a note in the discussion on the difference between some of the peak height variability results produced on the ProFlex compared to the existing peak height variability, particularly for forwards stutters. I have the following comments:

- It appears from the correlation plots on page 4 that the ProFlex dataset had a much reduce number of datapoints for forward stutter (there appears to be only approximately 20 observations across the entire dataset, and these could be from only a handful of profiles) than the existing dataset and so the variability results will be more prone to stochastic variation. This is one possible cause for the difference in peak height variability values and could be investigated by lowering the analytical threshold to detect more forward stutters and carrying out a Model Maker analysis (this would not necessarily then lead to a practice of lower analytical threshold in casework but would identify the cause of the differences). Alternatively, some additional profiles with higher concentrations of DNA could be produced and added to the dataset, in which forward stutters would be more likely to be observed. I note that in the Model Maker check output graph on page 5 there are very few instances of peaks with heights above 4000rfu, when the dynamic range of 3500xl instruments tend to extend out to approximately 30000rfu. Ideally there would be profiles in the dataset that extended out to the higher rfu ranges in order to inform the peak height variability model. As an example, to demonstrate my point in Figure 4 I show a

side-by-side comparison of the Model Maker check output graph from the QH validation report (left) and the equivalent plot from a validation produce in a different forensic facility.

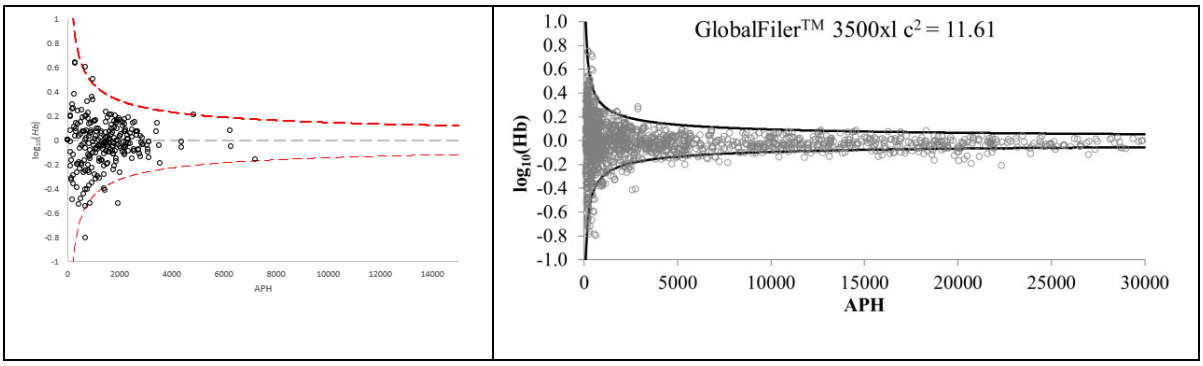

Figure 4: Model Maker heterozygous balance (Hb) plots for datasets produced in the QH ProFlex validation (left) and a different forensic facility (right)

This lack of a range in peak height data may also account for differences in other variance types that are noted in the discussion of the Model Maker results for Project #199 document.

- Alternatively, the noted differences in peak height variability may be due to a difference in the performance of the ProFlex instruments. I have already noted that there is some superficial evidence of ProFlex instruments that may be performing at different levels. One way this could be investigated would have been to run a Model Maker analysis on data produced by each instrument individuals (rather than the combination of data from all instruments that was carried out). It may be that any differences seen in peak height variability in the combined dataset is being driven by a single ProFlex instrument. Once identified this single instrument could then be investigated more deeply and potentially recalibrated if required. I note however that given the relatively low number of datapoints for some peak types in the combined dataset that to investigate the performance of individual instruments some additional laboratory work to produce additional samples may be required.

- If there is no individual instrument that appears to be performing differently (to a practical level) from the rest, then the appropriate action to take for ongoing profile analysis work would be to consider all the ProFlex instruments in a combined manner as was done in the Model Maker results for Project #199 report.

- It can be difficult to determine the practical impact of any differences between peak height variability parameter distributions. The STRmix models are quite robust to mild differences in peak height variability values (as shown by Kelly et al [1]). To assess the practical impacts of any differences it is common to have a standard set of test profiles to

which the DNA donors, and a number of non-donors are compared. Carrying out such an exercise may have found that any differences in peak height variability have very little practical consequence.

11.9 Overall conclusion

In my opinion the ProFlex validation has not been carried out according to best practise, due to its inadequate experimental design, and so has not given an adequate indication of instrument performance. Use of STRmix to a greater extend would have been more useful. I note in the introduction of the validation it states it was "*not to assess the suitability of the current analysis and profile interpretation thresholds*", however STRmix would still have been a useful tool to assess and compare performance (as mentioned in the preamble to this validation review).

This is not to say the ProFlex systems are unreliable (there is no evidence of unreliability), but equally there is limited ability to demonstrate reliability based on the results shown. I believe some additional laboratory work would have beneficial in this case. Specifically, a DNA dilution series leading to DNA profiles that cover the full dynamic range of the 3500xL instruments should be created and profiles generated on each of the ProFlex instruments. This dilution series should consist of at least 5 different references and total at least 50 samples. The generation of Model Maker parameters should occur for each instrument and experimentation comparing the LRs resulting from Hp and Hd true tests on constructed mixtures should be carried out. If performance of all instruments is similar (based on the alignment of Hp and Hd true LRs to some defined level) then the individual ProFlex datasets can be combined into a single dataset and analysed in Model Maker to obtain STRmix settings.

There is a risk of unreliable results being produced and reported (ultimately being reflected in the likelihood ratio produced to QPS) if there is an undiagnosed divergence in performance between the ProFlex instruments, which is currently assumed not to be occurring.

12.0 Validation of the QIAsymphony® SP/AS Modules November, 2016 (FSS.0001.0027.5378.pdf)

As with most other validation documents (excluding the components of the STARlet that uses the standard validation acceptance criteria that comes with the use of the Artel MVS), there is no mention of any specific guidance to which is being adhered or followed. This is not necessarily an issue, but when a guidance is followed then it generally ensures all aspects required for validation have been addressed. It is also easier for external bodies, such as auditors, to identify the breath of the validation as sufficient.

12.1 Experiment 1: Modification of Pre-Lysis Protocol

Graphically showing distribution for this data is appropriate. I am not sure that there would be significant differences between these methods if formal tests had been undertaken. Regardless I believe the conclusion drawn from the analysis done was appropriate.

12.2 Experiment 2: Modification of Extraction Protocol

It would have been useful to provide the value that the excluded outlier gave. Best practice for data analysis is only to remove outliers for specific reasons, e.g., there was a measurement error, or sampling issue. Just because an unusually large (or small) value is obtained, by itself is not a justification for removal. Having said that you could then choose to use a comparison metric that is more robust to outliers e.g., you could compare medians rather than means, and quantiles rather than standard deviations. In effect you may wish to use boxplots rather than bar-charts.

Also, again I am not sure whether formal statistical comparisons would have found much difference between the groups.

Regardless of this, I think the method chosen is appropriate.

12.3 Experiment 3: Lysate Storage

The experimental results shown here are quite limited (peak heights for two loci). In this instance some statistical tests may have been warranted to test for differences between storage conditions.

While there is nothing here to suggest the storage practice recommended is unreliable, some additional data analysis may be warranted.

12.4 Experiment 4: Sensitivity

Table 21 have extraction efficiency greater than 100%. Not unexpected given that not all DNA might come from cells (cell free DNA, see [29] for an example of this) and so an estimated DNA amount from cell count will not take this into account.

There are plenty of ways comparative sensitivity could have been formally tested (regression, or even a Bland-Altman plot), but the bar plot is sufficient.

In my opinion the decision was appropriate.

12.5 Experiment 5: Verification of Additional Substrates

The comparison of DNA yield was carried out qualitatively and involved a visual comparison of graphed results. In my opinion this is all that was needed to demonstrate a performance that was at least equivalent to that of the existing instrument (Maxwell 16).

12.6 Experiment 6: Inhibition

In my opinion the decision made was appropriate given the data. Again, testing could have been done, but was not needed to make choice.

12.7 Experiment 7: Degradation

The authors note that this experiment does not provide *"...conclusive evidence of the performance of the QIAsymphony® instrument to extract degraded DNA*.". This is due to the fact that the tests performed do not test the QIAsmyphony to retrieve small fragment of DNA.

With the limitations highlighted already by the authors my opinion is that the validation of this component is appropriate.

12.8 Experiment 8: Pipetting Accuracy

This component uses Artel MVS and standard assessment methodology and pass criteria. My opinion is that it has been carried out appropriately.

12.9 Experiment 9: Contamination Check

No comment. Has been carried out appropriately.

12.10 Experiment 10: Integrated Runs

No comment. Has been carried out appropriately.

12.11 Experiment11: Repeatability and Reproducibility

The validation would benefit from some criteria for passing (e.g., a %cov based on previous experience with DNA extractions or published findings on the topic). It would also be useful to see a plot of the spread of individual quantification amounts. The experimentation was quite minimal with only 10 samples that underwent repeated extractions, which covered two different groups of samples.

The sampling and extraction of DNA are the components of the DNA profiling process that are likely to add the most variability to the DNA amount obtained, and the final DNA profile quality. This means that there can be quite high acceptable level of variability in DNA amount from repeatability studies into DNA extraction. Therefore, the documentation of the variability is not likely to be critical for acceptance, but it is useful in understanding how much variability there is in the technique, particularly when undertaking troubleshooting.

From the work that was carried out there is no evidence that unreliable results are being produced.

12.12 Experiment 12: Sample Recovery

Again, it would be beneficial to the understanding of the performance of the instrument, and for the comparison of methods, to show some additional information. In this instance there is not even a standard deviation shown, meaning that there is no real sense of whether these results are performing differently in any practical sense, or whether there were any extreme/outlying results that could be skewing the averages.

It would also have been useful to show how the amount of DNA recovered compared with expected DNA amount (even if this is by comparison to a non-interrupted DNA extraction) to know whether there were any negative impacts of a run being interrupted. Again, this likely would not have changed the choice of method, nor the assessment of reliability, but would provide some information on the performance of the method that could be used when assessing results.

The choice of method seems logical given the results, and there is no evidence that unreliable results would be produced when implemented.

12.13 Experiment 13: Re-extraction of Substrates

In this instance the method is seeking to determine whether there is the possibility of re-extraction. All that is really required is showing that there is some possibility of obtaining DNA in such circumstances, and the study has shown this. While there could have been

additional data shown, in this instance it is not so important. The validation appears appropriate.

12.14 Overall conclusion

Much of this validation examines factors where the purpose is to choose the best out of several options (or to demonstrate that a process is possible). These sorts of examinations often do not need extensive formal statistical evaluation. In general, many of the factors would have benefitted with a more informative summarisation of the data (particularly graphically), however there is no evidence to suggest that unreliable results are being produced from this instrument.

13.0 Project Report #192 Validation of QIAsymphony SP for bone extraction April 2018 (FSS.0001.0025.5114.pdf)

13.1 Experiment 1: Current Organic Extraction

Contamination of the control and one sample was detected but dealt with appropriately and does not invalidate any part of the experiment.

I note that in the 'Instruction to Expert' document comment #20 comments on the quantification value of the positive control in Table 2 being lower than expected. I am not concerned with this result, or perhaps have not gleaned the same significance as the author of the comment. I have no further remarks.

13.2 Experiment 2: QIAGEN Pre-Lysis with Overnight Incubation and QIAsymphony® SP Extraction

All seem fine

13.3 Experiment 3: QIAGEN Pre-Lysis with 5 hour Incubation and QlAsymphony® SP Extraction

All seem fine

13.4 Experiment 4: Organic Pre-Lysis with Overnight Incubation and QlAsymphony® SP Extraction

All seem fine

13.5 Experiment 5: Organic Pre-Lysis with 5 hour Incubation and QIAsymphony® SP Extraction

All seem fine

13.6 Overall conclusion

In this case some Student's t-tests were performed to compare methodologies. In my opinion it would have been fine to just show the results graphically and draw conclusions from that, but there is no harm in statistical comparison.

It is not clear from the text what type of t-test was performed, but the statistical analysis that should be used is a paired test, as the interest is in the difference between samples and not the difference in the means of the group of samples. With multiple comparisons, if the question was whether the extraction method had any effect then a multi-test method should have been used (such as a repeated measure ANOVA). Also, depending on the distribution of the data a non-parametric test may have been the best to use. However, as I mentioned at the start, I think all the information needed to make the decision could be obtained from just the graphical display of results given, and so any issues with the statistical tests used have not affected my opinion that the methods chosen were done so appropriately.

14.0 Statement of Rhys Parry 28/09/2022

On 29/09/2022 I received the statement of Rhys Parry dated 28/09/2022 and updated instructions to expert that requested my comment on:

> *The specific concerns raised in reporting scientist, Rhys Parry's statement dated 28 September 2022.*

I restrict my comments to the points Raised by Rhys Parry on validations, or matters affected directly from validation outcomes, or matters of a statistical nature.

14.1 Microcon (paragraph 9 to 20)

Mr Parry has made comments that providing a success rate for profiles with DNA concentrations between 0.0011ng/µL and 0.0088ng/µL as a flat proportion of success (as appeared to have been done) was not appropriate. He based this opinion on the assumptions this made about the distribution of quantification values in the underlying data, and also from

a point of view that it simplified the opinion to a potentially unhelpful level. I agree with these general concerns.

He carried out his own analysis of the data (provided in the 'Rhys Parry 28 September 2022 – exhibits part 1' file) to provide probabilities for obtaining a result given low DNA concentrations. The analysis of Mr Parry was superior to providing a flat proportion of success. In his work Mr Parry broke the data into brackets of quantification and carried out an analysis that allowed a description of the probability of obtaining a result at any particular quantification level. While there was nothing significantly wrong with the way in which Mr Parry analysed the data, a more standard way to carry out such a task would have been using a logistic regression on the quantification data where each sample was assigned a binary outcome (i.e., result vs no result). Nevertheless, I expect the prediction of a probability of obtaining a result from the data produced by Mr Parry would give an appropriate value.

14.2 Validations: Quant Trio (project 152) paragraphs 54 to 65

At paragraph 55 I agree that there is a generally different skill set required for experimental designing and data analysis than that required for DNA profile interpretation and analysis. The latter is a common module of training provided to forensic analysts as part of their employment at a forensic institution. The former (experimental design and data analysis) is, in my experience, not usually provided as part of standard inhouse training in forensic institutions.

At paragraph 56 Mr Parry provides a number of dot points that highlight concerns he has with the data analysis in the Quant Trio validation. Many of these points are the same as those that I identified in my own assessment of the validation. The first two dot points highlight a general overestimation of quantification values shown in the Quantifier validation, which I had not initially noticed. Having re-examined the tables there is indeed a consistent over-estimate of quantification results compared to the level that was expected. It may be that the standards used for these experiments were not optimal (and hence tended to overestimate quantification values), however this point should be investigated to ensure that quantification over-estimates are not a persistent issue. This may be achievable by data mining from existing work, which would give an indication of the level of overestimation (or the absence of overestimation) over time.

At paragraph 57 Mr Parry comments on the limitation with the LOD analysis, which I have comment on in depth within my own review.

At paragraphs 58 to 61 Mr Parry discusses the issues of carrying out appropriate repeatability and reproducibility experimental design. I agree with his descriptions of these factors and that some of the validations have not carried out an ideal number of experiments in this regard (which I bring up in my own review of the various validations). I cannot comment on Mr Parry's assertion in paragraph 60. I note his specific mention in paragraph 61 of project 199

(for the ProFlex instruments) and agree with him assessment that the validation did not demonstrate repeatability and reproducibility. I have commented on this point at length in my own review.

14.3 Validation of Quant Studio 5 (Project 185) paragraphs 66 to 83

Paragraph 72 to 74 describe issues with using percentage change as a measure of difference. There are situations where the percentage change can be used (particularly when one measurement is being considered as the ground truth), and other situations where the percentage change can be misleading. I have covered this in my own review of Quant Trio. The best practice solution to this issue (as I mentioned in my Quant Trio review) is that a log base-10 transformation of a ratio is employed. In this case the order of values does not affect the magnitude of the measured change, i.e., 150 then 100 would lead to a log10(150/100) ~ 0.176 or and 100 then 150 would lead to log10(100/150) ~ -0.176 i.e., equal in size but opposite in sign to the first order.

In paragraph 75 Mr Parry points out that the repeatability study uses 7 replicates of 12 samples run twice. As correctly pointed out by Mr Parry, the experimental unit here is the instrument, which has only had 2 repeats. I had not picked this point in my initial review, but on reading of Mr Parry's statement I agree that this has not met the best practice for number of repeatability replicates.

At paragraphs 76 to 83 are various comments on the tests performed and the use of t-tests, on which I have commented throughout my own review.

14.4 NIST QRI – paragraphs 84 to 87

At paragraph 85 Mr Parry comments on the issue of comparing the performance of the QS5 to the previous 7500. While there is no evidence of the effect being described (i.e., that the 7500 was performing in a sub-standard way) this is indeed one of the types of risk that can arise when setting pass criteria based on other instruments rather than on independent performance metrics and is something I have commented on in my own review, and also suggested being amended in my recommendations.

Paragraph 84 I have no comment and paragraphs 86 and 87 are on matters I have already addressed.

14.5 Validation of QIAsymphony SP for Bone Extraction (project 192) – paragraphs 88 to 102

At paragraphs 90 to 92 Mr Parry comments on the variability of the results between extractions, but also in comparison to the originally obtained results. He suggests that some additional investigation could have been carried out to investigate the differences and some repeatability and reproducibility studies could be performed (paragraph 94). I agree in

general with these sentiments, but also acknowledge the fact that DNA extractions can be highly variable, and I do not have strong concerns with the results that were obtained. I am also aware that it is common for there to be a scarcity of availability of bone samples, which can sometimes dictate the size of the study that is possible.

15.0 Recommendations

Recommendation 1:

Standard Operating Procedures should exist that outline the requirements of validations and should include some indication of the type of statistical tests that can be performed and methods of graphically displaying results. For any tests mentioned in an SOP, explanations of these tests should be included to provide analysts some guidance as to when they should (and when they shouldn't) use the tests and the assumptions underlying each.

QH partially meet this recommendation as in the Appendix of the "Writing Guidelines for Validation and Change Management Reports." (document FSS.001.0012.0269.pdf), there is a section headed "Which statistics might be most appropriate" and gives the information:

- *ANOVA – to compare independent groups of samples*
- *Kruskal-Wallis – to compare independent groups of samples*
- *Paired T-test – to compare repeated samples i.e. same samples run through two different methods.*
- *Chi-square test – may be applied to demonstrate the average peak heights between loci (or dye layer) may differ*

Along with a reference for each that provides an example of application of these tests. This list should be expanded to include additional test types and methods of graphically displaying results. It also should include descriptions of tests and limitations as described above.

Recommendation 2:

Acceptance criteria should be set that are based on absolute values rather than being relative to the performance of previous instruments (unless it is a specific requirement that two instruments must be performing equivalently). This will ensure that all instruments that are accepted are being accepted to a level of high performance, and that performance is understood in absolute terms rather than relative to other instruments e.g., "*Quantifier Trio can quantify DNA with a cov less than 1.5*" is more informative with regards to performance than "*Quantifiler Trio has an equal or smaller cov than Quantifier Duo*".

Recommendation 3:

Acceptance criteria should be set in a manner that specifies the type of testing that needs to occur. For example, rather than an acceptance criterion such as:

> *Quantifiler® Trio will be assessed as acceptable if there is no significant difference in quantification results for the two plates run by Operator 1 on Day 1 (for SAT, LAT and Y Targets).*

the criterion could be set as:

> *Quantifiler® Trio will be assessed as acceptable if when comparing the two plates (run by Operator 1 on Day 1) each of the SAT, LAT and Y Targets provide a p-value less than 0.05 when a Wilcoxon Rank sum Test in conducted.*

Setting the criteria in this manner forces the analyst to carry out an analysis in a predetermined manner where there is no ambiguity about how to measure a concept such as 'significant difference'.

Recommendation 4:

For each validation carried out that requires a statistical analysis of results, an individual who has formal training or qualifications should be involved. This could be in several different ways:

a) A professional statistician from within QH (if one is employed) should be consulted in the design of validation experiments and should sign off the final validation report to indicate that data has been appropriately analysed and interpreted.
b) A professional statistician (but not an employee of QH) could be consulted
c) A member of QH could undergo formal training in statistical analysis to a level deemed appropriate (e.g., certificate, diploma, etc)

Recommendation 5:

The criteria set in a way suggested by recommendation 3 should be devised by a professional statistician

Recommendation 6:

An Australian and New Zealand group should be formed that develop guidelines for the appropriate minimum requirement of validation of laboratory instruments in a forensic context. The group should include members from forensic institutions, but also statisticians, and potentially members of other professions or groups such as those involved in instrument manufacture, members of standard-setting groups, and/or metrologists. These guidelines could then be used to inform all forensic laboratories on best practice and provide assurance to stakeholders and the public that all forensic laboratories are carrying out the same minimum accepted level of validation.

Recommendation 7:

Of the members that sign off validation reports, at least one should be external to the group who is carrying out the validation. This provides external feedback to ensure consistency between laboratories or groups.

Recommendation 8:

In the case of the ProFlex validation, some additional experimental laboratory work should be carried out to show the relative differences in performance of the instruments. Specifically, a DNA dilution series leading to DNA profiles that cover the full dynamic range of the 3500xL instruments should be created and profiles generated on each of the ProFlex instruments. This dilution series should consist of at least 5 different references and total at least 50 samples. The generation of Model Maker parameters should occur for each instrument and experimentation comparing the LRs resulting from Hp and Hd true tests on constructed mixtures should be carried out. If performance of all instruments is similar (based on the alignment of Hp and Hd true LRs to some defined level) then the individual ProFlex datasets can be combined into a single dataset and analysed in Model Maker to obtain STRmix settings.

Recommendation 9:

In the case of the QuantiFiler Trio validation, additional testing should be carried out to appropriately identify the LOD. Specifically, the ability to detect DNA over a range of concentrations should be tested, whereby each step in the dilution series has 10 to 20 replicates. The LOD should then be set at the concentration at which DNA is detected less that 0.95 of the time. This can be done by calculating the detection rate at each concentration, or by use of logistic regression. If none of the concentrations tested fall below this probability, then the concentration range should be extended down to lower levels until the LOD is reached. This limit maybe below, or above, the current LOD value of 0.001ng/µL being used by QH.

Recommendation 10:

If a LOD value for QuantiFiler Trio is going to be used as a decision threshold, then until its value has been appropriately calculated (as detailed in recommendation 9) all quantified DNA samples should be treated (with respect to decision making or laboratory processes) as though they have exceeded the LOD.

Recommendation 11:

In the case of the QH validation reports I have reviewed, apart those mentioned in recommendations 8 and 9, some additional statistical work should be conducted to address the issues raised within my report. In many instances this is simply a matter of displaying the results graphically in a more meaningful way or carrying out different statistical tests on existing data.

Recommendation 12:

Applicable validations should include, in the introduction, some reference to which general validation guideline is being followed. For example:

> *This validation for QuantiFiler Trio is following the SWGDAM guideline "Scientific Working Group on DNA Analysis Methods Validation Guidelines for DNA Analysis Methods" for which they list as experiment to conduct:*

- *Known / non-probative samples*
- *Repeatability*
- *Reproducibility*
- *Sensitivity studies*
- *Stochastic studies*
- *Mixture studies*
- *Contamination studies*

In this validation, experiment 1 addresses repeatability, experiment 2 addresses….etc

In this validation contamination studies are not being addressed due to …..

In addition to the SWGDAM guidelines this validation report will include experiments…

Providing this information will immediately identify to stakeholder the validation as aligning with best practice guidelines. It will also serve to ensure that all validations have a minimum breath to the experiments being carried out.

If no national guidelines are developed (as outlined in recommendation 6) then the validation introduction could also include references to methods used in published examples of similar validations. Again, this provides some assurance to use of recognised and accepted statistical methodology.

Recommendation 13:

Following the completion of validations, a presentation should be given to members of the forensic organisation, explaining the work that was done, the tests carried out and the meaning of the test results. This ensures that each member understands the statements being made in their own reports within the context of how they relate to the performance of laboratory instruments.

16.0 References:

[1] H. Kelly, J.A. Bright, M. Kruijver, S. Cooper, D. Taylor, K. Duke, M. Strong, V. Beamer, C. Buettner, J. Buckleton, A sensitivity analysis to determine the robustness of STRmix with respect to laboratory calibration, Forensic science international. Genetics 35 (2018) 113-122.

[2] L. Russell, S. Cooper, R. Wivell, Z. Kerr, D. Taylor, J. Buckleton, J.-A. Bright, A guide to results and diagnostics within a STRmix™ report, WIREs Forensic Science 1(6) (2019) e1354.

[3] FSR, The Forensic Science Regulator guidance on Validation FSR-G-201, The Forensic Science Regulator, 5 St Philip's Place, Colmore Row, Birmingham, 2020, pp. 1-63.

[4] SWGDAM, Scientific Working Group on DNA Analysis Methods Validation Guidelines for DNA Analysis Methods, (2016).

[5] ENFSI, Guidelines for the single laboratory Validation of Instrumental and Human Based Methods in Forensic Science, (2014).

[6] A. Holt, S. Wootton, J. Mulero, P. Brzoska, E. Langit, R. Green, Developmental Validation of the Quantifiler™ HP and Trio Kits for Human DNA Quantification in Forensic Samples, Forensic Science International: Genetics 21 (2015).

[7] ThermoFisher, Quantifiler™ HP and Trio DNA Quantification Kits USER GUIDE, 4485354:H (2018).

[8] SWGDAM, The Quality Assurance Standards for DNA Databasing Laboratories, (2020).

[9] ISO, ISO/IEC 17025 Testing and calibration laboratories, (2017).

[10] C.G.G. Aitken, A. Wilson, R. Sleeman, Statistical significance—meaningful or not, Law, Probability and Risk 17(2) (2018) 157-162.

[11] J.M. Butler, Debunking Some Urban Legends Surrounding Validation Within the Forensic DNA Community | NIST, 2006.

[12] J.T. Bradshaw, K.J. Albert, Instrument Qualification and Performance Verification for Automated Liquid-Handling Systems, Practical Approaches to Method Validation and Essential Instrument Qualification2010, pp. 347-375.

[13] R. Zaki, A. Bulgiba, R. Ismail, N.A. Ismail, Statistical Methods Used to Test for Agreement of Medical Instruments Measuring Continuous Variables in Method Comparison Studies: A Systematic Review, PLOS ONE 7(5) (2012) e37908.

[14] J. Martin Bland, D. Altman, STATISTICAL METHODS FOR ASSESSING AGREEMENT BETWEEN TWO METHODS OF CLINICAL MEASUREMENT, The Lancet 327(8476) (1986) 307-310.

[15] J.S. Krouwer, Why Bland-Altman plots should use X, not (Y+X)/2 when X is a reference method, Statistics in medicine 27(5) (2008) 778-80.

[16] M. Burns, H. Valdivia, Modelling the limit of detection in real-time quantitative PCR, European Food Research and Technology 226(6) (2008) 1513-1524.

[17] A. Forootan, R. Sjöback, J. Björkman, B. Sjögreen, L. Linz, M. Kubista, Methods to determine limit of detection and limit of quantification in quantitative real-time PCR (qPCR), Biomolecular detection and quantification 12 (2017) 1-6.

[18] J.R. Gilder, T.E. Doom, K. Inman, M. Crim, D.E. Krane, Run-Specific Limits of Detection and Quantitation for STR-Based DNA Typing. , Journal of the Forensic Sciences 52(1) (2006) 97-101.

[19] D. Taylor, J.-A. Bright, C. McGoven, C. Hefford, T. Kalafut, J. Buckleton, Validating multiplexes for use in conjunction with modern interpretation strategies, Forensic Science International: Genetics 20 (2016) 6-19.

[20] H. Kelly, J.-A. Bright, J.M. Curran, J. Buckleton, Modelling heterozygote balance in forensic DNA profiles, Forensic Science International: Genetics 6(6) (2012) 729-734.

[21] J.-A. Bright, K. Stevenson, J. Curran, J. Buckleton, The variability in likelihood ratios due to different mechanisms, Forensic Science International: Genetics 14 (2015) 187-190.

[22] J.W. Tay, J.A. Murakami, P.L. Cooper, M.S. Rye, Sensitivity and baseline noise of three new generation forensic autosomal STR kits: PowerPlex® Fusion, VeriFilerTM Plus and Investigator® 24plex QS, Forensic Science International: Reports 1 (2019) 100049.

[23] J.-A. Bright, J.M. Curran, Investigation into stutter ratio variability between different laboratories, Forensic Science International: Genetics 13 (2014) 79-81.

[24] SWGDAM, Guidelines for the Validation of Probabilistic Genotyping Systems. 2015).

[25] M.D. Coble, J. Buckleton, J.M. Butler, T. Egeland, R. Fimmers, P. Gill, L. Gusmão, B. Guttman, M. Krawczak, N. Morling, W. Parson, N. Pinto, P.M. Schneider, S.T. Sherry, S. Willuweit, M. Prinz, DNA Commission of the International Society for Forensic Genetics: Recommendations on the validation of software programs performing biostatistical calculations for forensic genetics applications, Forensic Science International: Genetics 25 (2016) 191-197.

[26] D. Moore, T. Clayton, J. Thomson, A comprehensive study of allele drop-in over an extended period of time, Forensic Science International: Genetics 48 (2020) 102332.

[27] D. Taylor, J. Buckleton, J.-A. Bright, Factors affecting peak height variability for short tandem repeat data, Forensic Science International: Genetics 21 (2016) 126-133.

[28] D. Moore, T. Clayton, J. Thomson, A comprehensive study of allele drop-in over an extended period of time, Forensic Science International: Genetics 48 (2020).

[29] P. Kanokwongnuwut, B. Martin, D. Taylor, K.P. Kirkbride, A. Linacre, How many cells are required for successful DNA profiling?, Forensic Science International: Genetics 51 (2021) 102453.

2854 **Appendix 1: material provided by the Commission**

2855

No	Document	Date	Link
1	Letter – Sofronoff to Dr Duncan Taylor		
2	Commission of Inquiry Terms of Reference	10/06/2022	Queensland Government Gazette
	Instructions to expert	05/09/2022	
3	Validations (Quant Studio 5 and Quant Trio)		
	Quant Studio 5		
	Data		Data
	Internal Feedback		
	Report for QIS OQI 56218 Use of NIST standard in Project #185	3/05/2022	FSS.0001.0002.3710.pdf
	RE: Project #185- Validation of QS5s A and B (Allan McNevin and Luke Ryan)	27/06/2018	FSS.0001.0005.0500.pdf
	Feedback	N/A	FSS.0001.0005.0502.pdf
	Feedback	N/A	FSS.0001.0005.0502.xls
	Technical Review of Proposal #185- Validation of QS5 (Updated Experiment 3)	1/01/2019	FSS.0001.0005.0791.pdf
	Technical Review of Proposal #185- Validation of QS5	1/05/2018	FSS.0001.0006.0799.pdf
	Correspondence between Justin Howes and Kaye Ballantyne	24/09/2018	FSS.0001.0011.2167.pdf
	Correspondence between Luke Ryan and Cathie Allen	6/07/2018	FSS.0001.0066.4890.pdf
	Correspondence between Luke Ryan and Cathie Allen	23/05/2019	FSS.0001.0066.4912.pdf
	Correspondence between Luke Ryan and Cathie Allen	15/02/2019	FSS.0001.0066.4914.pdf
	Correspondence between Luke Ryan and Kylie Rika	19/06/2018	FSS.0001.0066.4954.pdf
	Correspondence between Luke Ryan and Kylie Rika	9/06/2017	FSS.0001.0066.4956.pdf
	Project Files		
	QuantSTudio5 Real-Time PCR System for HID	N/A	FSS.0001.0005.0218.pdf
	HID RT 1.3 Software and Data Analysis	N/A	FSS.0001.0005.0267.pdf
	QuantStudio 5 Maintenance	N/A	FSS.0001.0005.0357.pdf
	Abstract	N/A	FSS.0001.0005.0515.pdf
	Project Report #185 Validation of two QuantStudio 5 Real-Time PCR Systems November 2018	1/11/2018	FSS.0001.0005.0516.pdf
	Project Proposal #185 Validation of two QuantStudio 5 Real-Time PCR Systems March 2018	1/03/2018	FSS.0001.0005.0538.pdf
	Technical Review of Proposal #185- Validation of QS5 (Updated Experiment 3)	1/01/2019	FSS.0001.0005.0791.pdf\

	Technical Review of Proposal #185- Validation of QS5	1/05/2018	FSS.0001.0005.0799.pdf
	Procedure for Change Management in Forensic DNA Analysis Appendix 2	6/10/2016	FSS.0001.0005.0803.pdf
	Project Proposal #185 Validation of two QuantStudio 5 Real-Time PCR Systems Repeatability and Reproducibility Experiment 3 November 2018	1/11/2018	FSS.0001.0005.0812.pdf
	Project Proposal #185 Validation of two QuantStudio 5 Real-Time PCR Systems June 2017	1/06/2017	FSS.0001.0005.0822.pdf
	Technical Review of Proposal #185- Validation of QS5	1/05/2018	FSS.0001.0005.0836.pdf
	Project Proposal #185 Validation of two QuantStudio 5 Real-Time PCR Systems May 2017	1/05/2017	FSS.0001.0020.7877.pdf
	Project Proposal #185 Validation of two QuantStudio 5 Real-Time PCR Systems May 2017	1/05/2017	FSS.0001.0020.7890.pdf
	Abstract	N/A	FSS.0001.0020.8427.pdf
	Validation of QuantStudio 5 Real-Time PCR Systems	N/A	FSS.0001.0020.8430.pdf
	Project Proposal #185 Validation of two QuantStudio 5 Real-Time PCR Systems Repeatability and Reproducibility Experiment 3 November 2018	1/11/2018	FSS.0001.0020.8631.pdf
	Project Proposal #185 Validation of two QuantStudio 5 Real-Time PCR Systems May 2017	1/05/2017	FSS.0001.0020.8681.pdf
	Project Proposal #185 Validation of two QuantStudio 5 Real-Time PCR Systems June 2017	1/06/2017	FSS.0001.0020.8694.pdf
	Project Proposal #185 Validation of two QuantStudio 5 Real-Time PCR Systems May 2017	1/05/2017	FSS.0001.0020.8707.pdf
	Project Proposal #185 Validation of two QuantStudio 5 Real-Time PCR Systems May 2017	1/05/2017	FSS.0001.0020.8720.pdf
	Project Report #185 Validation of two QuantStudio 5 Real-Time PCR Systems February 2019	1/02/2019	FSS.0001.0025.4982.pdf
	Project Proposal #185 Validation of two QuantStudio 5 Real-Time PCR Systems June 2017	1/06/2017	FSS.0001.0025.5016.pdf
	Validation Report		
	Project Report #185 Validation of two QuantStudio 5 Real-Time PCR Systems November 2018	1/11/2018	FSS.0001.0005.0516.pdf

	Project Report #185 Validation of two QuantStudio 5 Real-Time PCR Systems March 2018	1/03/2018	FSS.0001.0005.0560.pdf
	Project Report #185 Validation of two QuantStudio 5 Real-Time PCR Systems March 2018	1/03/2018	FSS.0001.0005.0584.pdf
	Project Report #185 Validation of two QuantStudio 5 Real-Time PCR Systems March 2018	1/03/2018	FSS.0001.0005.0609.pdf
	Project Report #185 Validation of two QuantStudio 5 Real-Time PCR Systems June 2018	1/06/2018	FSS.0001.0005.0634.pdf
	Project Report #185 Validation of two QuantStudio 5 Real-Time PCR Systems July 2018	1/07/2018	FSS.0001.0005.0658.pdf
	Project Report #185 Validation of two QuantStudio 5 Real-Time PCR Systems November 2018	1/11/2018	FSS.0001.0005.0682.pdf
	Project Report #185 Validation of two QuantStudio 5 Real-Time PCR Systems January 2019	1/01/2019	FSS.0001.0005.0706.pdf
	Project Report #185 Validation of two QuantStudio 5 Real-Time PCR Systems January 2019	1/01/2019	FSS.0001.0005.0726.pdf
	Project Report #185 Validation of two QuantStudio 5 Real-Time PCR Systems January 2019	1/01/2019	FSS.0001.0005.0747.pdf
	Project Report #185 Validation of two QuantStudio 5 Real-Time PCR Systems February 2019	1/02/2019	FSS.0001.0005.0767.pdf
	Correspondence between Beatric Goh and Adam Kaity Quant Studio 5 (moving and RNAse P plate)		FSS.0001.0020.8456.pdf
	Quantifiler Trio		
	Data		Data
	Internal Feedback		
	RE: Project #185- Validation of QS5s A and B	27/06/2018	FSS.0001.0005.0500.pdf
	Feedback	N/A	FSS.0001.0005.0502.pdf
	RE: Project #185- Validation of QS5s A and B	19/06/2018	FSS.0001.0005.0505.pdf
	Correspondence between Justin Howes and Kaye Ballantyne	24/09/2018	FSS.0001.0011.2167.pdf
	Correspondence between Sharon Johnstone and Kylie Rika	3/02/2016	FSS.0001.0052.1771.pdf
	Correspondence between Megan Mathieson and Kylie Rika	11/02/2016	FSS.0001.0052.1773.pdf
	Correspondence between Kirsten Scott, Pierre Acedo and Cathie Allen	27/04/2015	FSS.0001.0066.4882.pdf
	Correspondence between Luke Ryan and Cathie Allen	3/02/2015	FSS.0001.0066.4889.pdf

	Correspondence between Lucretia Angus and Cathie Allen	29/05/2014	FSS.0001.0066.4901.pdf
	Correspondence between Lucretia Angus and Cathie Allen	5/06/2014	FSS.0001.0066.4903.pdf
	Correspondence between Kirsten Scott and Cathie Allen	11/05/2015	FSS.0001.0066.4907.pdf
	Correspondence between Luke Ryan and Cathie Allen	23/11/2015	FSS.0001.0066.4916.pdf
	Correspondence between Luke Ryan and Cathie Allen	7/09/2015	FSS.0001.0066.4920.pdf
	Correspondence between Luke Ryan and Cathie Allen	17/08/2015	FSS.0001.0066.4921.pdf
	Correspondence between Megan Mathieson, Maria Aguilera, Pierre Acedo, Cecilia Lannuzzi and Cathie Allen	4/11/2015	FSS.0001.0066.4926.pdf
	Correspondence between Luke Ryan and Cathie Allen	5/06/2014	FSS.0001.0066.4927.pdf
	Project Files		
	Asset Management Acquisition Equipment	2016-2020	FSS.0001.0003.6109.pdf
	Project Proposal #152 Validation of Quantifiler Trio February 2015	1/02/2015	FSS.0001.0004.8724.pdf
	Quantifiler Trio DNA Quantification Kit Implementation	N/A	FSS.0001.0004.8770.pdf
	Minor Process Change Stage 2	5/07/2018	FSS.0001.0019.1252.pdf
	Customer Notification: Component Modification to the Quantifiler HP and Trio DNA Quantification Kits	1/06/2015	FSS.0001.0019.6844.pdf
	Quantifiler Trio DNA Quantification Kit Implementation	N/A	FSS.0001.0019.6873.pdf
	Experimental Design- Proposal #14 Validation of Quantifiler Trio	1/01/2015	FSS.0001.0019.6884.pdf
	Project Proposal #14 Validation of Quantifiler Trio	1/01/2015	FSS.0001.0019.6966.pdf
	Project Proposal #152 Validation of Quantifiler Trio January 2015	1/01/2015	FSS.0001.0019.7048.pdf
	Project Proposal #152 Validation of Quantifiler Trio January 2015	1/01/2015	FSS.0001.0019.7079.pdf
	Project Proposal #152 Validation of Quantifiler Trio February 2015	1/02/2015	FSS.0001.0019.7113.pdf
	Project Proposal #152 Validation of Quantifiler Trio February 2015	1/02/2015	FSS.0001.0019.7147.pdf
	Experiment Results Report	9/04/2015	FSS.0001.0019.7679.pdf
	WinPREP Quant Trio Kits Test	N/A	FSS.0001.0019.7686.pdf
	WinPREP Quant Trio Kits Test	N/A	FSS.0001.0019.7687.pdf
	WinPREP Quant Trio Kits Test	N/A	FSS.0001.0019.7688.pdf
	WinPREP Quant Trio Setup Ver 1.1	N/A	FSS.0001.0019.7689.pdf
	Quantification of Extracted DNA using the Quantilifer Trio DNA Quantification Kit	28/09/2015	FSS.0001.0019.8043.pdf

	Quantification of Extracted DNA using the Quantilifer Trio DNA Quantification Kit	8/02/2017	FSS.0001.0019.8074.pdf
	Next Revision Update	8/12/2015	FSS.0001.0019.8130.pdf
	Experiment Results Report	18/09/2015	FSS.0001.0019.8133.pdf
	HID Real-Time PCR Analysis Software	1/03/2014	FSS.0001.0019.8141.pdf
	Quantifiler HP and Trio DNA Quantification Kits User Guide	1/03/2014	FSS.0001.0019.8223.pdf
	Quantifiler Trio Bubbles Issue	21/12/2016	FSS.0001.0019.8333.pdf
	Quantifiler Trio SOP		FSS.0001.0019.8335.pdf
	Validation of Quantifiler Trio April 2015	1/04/2015	FSS.0001.0019.8338.pdf
	Validation of Quantifiler Trio April 2015	1/04/2015	FSS.0001.0019.8414.pdf
	Validation of Quantifiler Trio August 2015	1/08/2015	FSS.0001.0019.8495.pdf
	Validation of Quantifiler Trio August 2015	1/08/2015	FSS.0001.0019.8587.pdf
	Validation of Quantifiler Trio August 2015	1/08/2015	FSS.0001.0019.8679.pdf
	Validation of Quantifiler Trio August 2015	1/08/2015	FSS.0001.0019.8679.pdf
	Validation of Quantifiler Trio August 2015	1/08/2015	FSS.0001.0019.8876.pdf
	Validation of Quantifiler Trio April 2015	1/04/2015	FSS.0001.0019.8964.pdf
	Validation of Quantifiler Trio April 2015	1/04/2015	FSS.0001.0019.9052.pdf
	HID Real-Time PCR Analysis Software	1/03/2014	FSS.0001.0020.5891.pdf
	Quantifiler HP and Trio DNA Quantification Kits User Guide	1/03/2014	FSS.0001.0020.5973.pdf
	Validation of Quantifiler Trio September 2015	1/09/2015	FSS.0001.0020.6083.pdf
	Validation Report		
	Validation of Quantilifer Trio September 2015	1/09/2015	FSS.0001.0052.1773.pdf
	Validation of Quantilifer Trio September 2015	1/09/2015	FSS.0001.0052.1773.pdf
4	3500 Genetic Analyzer		
	3500A bline laser change 20191105		3500A bline laser change 20191105
	3500xl A Baseline (CW) 202009		3500xl A Baseline (CW) 202009
	Internal Feedback		Internal Feedback
	Project Files		Project Files
	Validation Report	2016; 2019	Validation Report
5	Bone Instrument Cleaning Methods		
	Data		Data
	Internal Feedback		Internal Feedback
	Project Files		Project Files
	Validation Report	2014;2015	Validation Report
6	Hamilton STARlet		
	Hamilton STARlet A		Hamilton STARlet A
	Hamilton STARlet B		Hamilton STARlet B
	Hamilton STARlet C		Hamilton STARlet C
	Internal Feedback		Internal Feedback
	Project Files		Project Files
7	ProFlex		
	Data		Data
	Internal Feedback		Internal Feedback
	Project Files		Project Files

	Validation Report	12/1/2021	Validation Report
8	QIAsymphony		
	Data		Data
	Internal Feedback		Internal Feedback
	Project Files		Project Files
	Validation Report	11/1/2016	Validation Report
9	General		
	Kirsten Scott Statement		FSS.0001.0011.5388.pdf
	Kirsten Scott Exhibits		Kirsten Scott Exhibits
	Current Standard Operating Procedures		
	17168 Basics of DNA profile interpretation	13/07/2020	FSS.0001.0012.0147.pdf
	22871 Procedure for Change Management in Forensic DNA Analysis	19/04/2022	FSS.0001.0012.0247.pdf
	23402 Writing Guidelines for Validation and Change Management Reports	18/08/2021	FSS.0001.0012.0269.pdf
	33315 Procedure for Verification and Maintenance of Equipment (Version 5, in force 5 June 2022)	17/08/2021	FSS.0001.0053.1266.pdf
	33315 Procedure for Verification and Maintenance of Equipment (Version 6, in force 30 June 2022)	14/07/2022	FSS.0001.0012.1074.pdf
	33756 Operation and Maintenance of the QIAsymphony SP and AS modules	10/03/2022	FSS.0001.0012.1224.pdf
	34034 Forensic DNA Analysis Workflow Procedure	28/04/2022	FSS.0001.0012.1591.pdf
	34045 Quantification of Extracted DNA using the Quantifiler Trio DNA Quantification Kit	03/05/2022	FSS.0001.0012.1753.pdf
	34050 Operation and Maintenance of the Microlab STARlet and LabElite Integrated I.D. Capper	17/01/2022	FSS.0001.0012.1788.pdf
	34054 Microlab STARlet and LabElite Integrated I.D. Capper Training Module	10/03/2022	FSS.0001.0012.1845.pdf
	34055 Training Delivery Plan Microlab STARlet and LabElite Integrated I.D. Capper	24/02/2022	FSS.0001.0012.1856.pdf
	34062 Capillary Electrophoresis Setup	15/02/2021	FSS.0001.0012.1861.pdf
	34063 Preparation & Testing of Extraction Quality Controls and Testing of Extraction Reagents	07/12/2021	FSS.0001.0012.1889.pdf
	34131 Capillary Electrophoresis Quality (CEQ) Check- Forensic Register	15/02/2021	FSS.0001.0012.2067.pdf
	34132 DNA Extraction and Quantification of Samples using the QIAsymphony SP and AS-FR	10/05/2021	FSS.0001.0012.2091.pdf
	34298 Validation of Examinations	30/05/2022	FSS.0001.0012.2504.pdf
	34312 Operation and Maintenance of the Applied Biosystems 3500xL Genetic Analyzer	31/05/2021	FSS.0001.0012.2560.pdf
	34514 Preparation & Testing of Quantification Standards, In-house Controls, Quantification Kits and Amplification Kits	07/06/2022	FSS.0001.0012.2630.pdf

	34610 Basic Programming for Microlab STARlet and LabElite Integrated I.D Capper using Venus Software	08/11/2021	FSS.0001.0012.2669.pdf
	35028 Operation and Maintenance of the AB Quant Studio RT-PCR Instrument	26/05/2022	FSS.0001.0012.2733.pdf
	36187 Maintenance of the ProFlex PCR System & ProFlex Server	10/01/2022	FSS.0001.0012.2847.pdf
	Standard Operating Procedure for Validations (QIS 23401)		
	23401V3 DNA Analysis Validation Guidelines	17/12/2012	FSS.0001.0001.3880.pdf
	23401V8 Forensic DNA Analysis Validation and Verification Guidelines	08/04/2022	FSS.0001.0024.1370.pdf
	Any guideline or other document about who performs validations		
10	Additional Materials received 23/09/2022		
	3500xL Genetic Analyzer Validation for Reference samples Amplified with Powerplex®21 using Direct Amplification	February 2015	FSS.0001.0025.9177
	3500xL Genetic Analyzer Validation for Extracted Reference Samples Amplified with Powerplex®21	June 2015	FSS.0001.0026.0606
	3500xL Genetic Analyzer Validation for Casework Samples Amplified with Powerplex®21	September 2015	FSS.0001.0006.2852
	Review and Assessment of the Use of "No DNA Detected" by Queensland Health Forensic and Scientific Services (QHFSS) - Prepared by Bruce Budowle, Ph.D.	05/09/2022	EXP.0001.0003.0001
	Emails between Emma Caunt and Justin Howes on Model maker results for Proflex validation	May – August 2022	FSS.0001.0057.2197
	Forensic DNA Analysis: Summary – Model Maker results for Project #199	25/01/2022	FSS.0001.0008.6790
	Project Report #185 Validation of two QuantStudio™ 5 Real-Time PCR Systems	July 2018	FSS.0001.0005.0659
	Emails between Rhys Parry and Kylie Rika	July 2018	FSS.0001.0060.1096
11	Additional Materials received 29/09/2022		
	Updated instructions to expert		
	Procedure for Case Management	08/03/2021	FSS.0001.0001.9446
	Quantification of Extracted DNA using the Quantifiler® Trio DNA Quantification Kit	03/05/2022	FSS.0001.0001.6663
	Statement of Rhys Parry	28/09/2022	
	Rhys Parry 28 September 2022 - exhibits part 1	28/09/2022	
	Rhys Parry 28 September 2022 - exhibits part 2	28/09/2022	
	Rhys Parry 28 September 2022 - exhibits part 3	28/09/2022	
	Statement of Professor Linzi Wilson-Wilde OAM PhD	25/08/2022	EXP.0002.0004.0001

2856

Appendix II: Curriculum Vitae

Education

2017-2018: Deep learning specialisation, Coursera

2016-2019: PhD (Discipline of statistics, College of Science and Engineering), Flinders University. Thesis title: Improving the statistical evaluation of forensic DNA evidence

2016: Machine Learning, Coursera

2015-2016: Certificate of Advanced Studies in "Statistics and the Evaluation of Forensic Evidence" offered through the Formation Continue UNIL-EPFL Lausanne - Suisse

2014: Lean Six Sigma – Yellow Belt (Advanced)

2008-2011 - Diploma in Biostatistics, Biostatistics Collaboration of Australia.

2001-2005: PhD (School of Biological Sciences), Flinders University and South Australian Museum. Thesis title: Genetics Using DNA markers for wildlife management and protection: a study of the population structure and systematics of the Australian carpet pythons utilising STRs, mitochondrial DNA sequence and allozymes.

2001: Honours Degree, Flinders University. Thesis title: Stable isotope ratio analysis of the human bone retrieved from St Mary's churchyard, and chemical analysis of the surrounding soil.

1998-2000: Undergraduate Degree, Flinders University - Forensic and Analytical Chemistry.

1997: Matriculation, Prince Alfred College.

Research history

Statistics:
Peer-reviewed publications: 123
i10 index: 72
h-index: 34
(as per google scholar):
https://scholar.google.com.au/citations?hl=en&pli=1&user=F38i2VMAAAAJ
Article review : publication ratio of 1:1

Peer reviewed publications:

2009:

1. Duncan A. Taylor, Julianne M. Henry, Simon J. Walsh. South Australian Aboriginal sub-population data for the nine AMPFlSTR® Profiler Plus™ short tandem repeat (STR) loci. Forensic Science International: Genetics, 2009 2(2):e27-e30.

2012:

2. Taylor DA, Henry JM Haplotype data for 16 Y-Chromosome STR loci in Aboriginal and Caucasian populations in South Australia Forensic Science International: Genetics. 2012; 6(6):e187-8.

3. Taylor DA, Nagle N, Ballantyne KN, van Oorschot RA, Wilcox S, Henry J, Turakulov R, Mitchell RJ. An investigation of admixture in an Australian Aboriginal Y-chromosome STR database. Forensic Science International: Genetics. 2012; 6(5):532-8.

2013:

4. Ottens R, Taylor D, Abarno D, Linacre A. Successful direct amplification of nuclear markers from a single hair follicle. Forensic Science Medicine and Pathology. 2013; 9(2):238-43

5. Jo-Anne Bright, Duncan Taylor, James M. Curran and John S. Buckleton. Degradation of forensic DNA profiles. Australian Journal of Forensic Sciences. 2013; 45(4):445-449

6. Jo-Anne Bright, Duncan Taylor, James M. Curran, and John S. Buckleton . Developing allelic and stutter peak height models for a continuous method of DNA interpretation Forensic Science International: Genetics. 2013; 7(2), 296-304.

7. Duncan Taylor, Jo-Anne Bright and John S. Buckleton. The interpretation of single source and mixed DNA profiles, Forensic Science International: Genetics. 2013;7(5): 516-528.

2014:

8. Buckleton, J., H. Kelly, J.-A. Bright, D. Taylor, T. Tvedebrink and J. M. Curran. Utilising allelic dropout probabilities estimated by logistic regression in casework. Forensic Science International. Genetics, 2014; 9: 9-11

9. Tegan E. Collins, Renée Ottens, Kaye N. Ballantyne, Nano Nagle, Julianne Henry, Duncan Taylor, Michael Gardner, Alison J. Fitch, Amanda Goodman, Roland A.H. van Oorschot, R. John Mitchell and Adrian Linacre Characterisation of novel and rare Y-chromosome short tandem repeat alleles in self-declared South Australian Aboriginal database, International Journal of Legal Medicine. 2014; 128(1): 27-31.

10. Jo-Anne Bright, Duncan Taylor, James Curran, John Buckleton. Searching mixed DNA profiles directly against profile databases. Forensic Science International: Genetics. 2014; 9: 102–110

11. Duncan Taylor. Using continuous DNA interpretation methods to revisit likelihood ratio behaviour. Forensic Science International: Genetics. 2014; 11: 144-153

12. D Taylor, JA Bright, J Buckleton, J Curran. An illustration of the effect of various sources of uncertainty on DNA likelihood ratio calculations. Forensic Science International: Genetics. 2014; 11: 56-63

13. J Buckleton, JA Bright, D Taylor, I Evett, T Hicks, G Jackson, JM Curran. Helping formulate propositions in forensic DNA analysis. Science & Justice 2014; 54(4): 258-261

14. Jo-Anne Bright, Cathie Allen, Shelley Fountain, Kerryn Gray, Denise Grover, Sharon Neville, Adam L Poy, Duncan Taylor, Gavin Turbett, Linzi Wilson-Wilde. Australian population data for the twenty Promega PowerPlex 21 short tandem repeat loci. Australian Journal of Forensic Sciences. 2014; 46 (4): 442-446

15. Ballantyne KN, Ralf A, Aboukhalid R, Achakzai NM, Anjos MJ, Ayub Q, Balažic J, Ballantyne J, Ballard DJ, Berger B, Bobillo C, Bouabdellah M, Burri H, Capal T, Caratti S, Cárdenas J, Cartault F, Carvalho EF, Carvalho M, Cheng B, Coble MD, Comas D, Corach D, D'Amato ME, Davison S, de Knijff P, De Ungria MC, Decorte R, Dobosz T, Dupuy BM, Elmrghni S, Gliwiński M, Gomes SC, Grol L, Haas C, Hanson E, Henke J, Henke L, Herrera-Rodríguez F, Hill CR, Holmlund G, Honda K, Immel UD, Inokuchi S, Jobling MA, Kaddura M, Kim JS, Kim SH, Kim W, King TE, Klausriegler E, Kling D, Kovačević L, Kovatsi L, Krajewski P, Kravchenko S, Larmuseau MH, Lee EY, Lessig R, Livshits LA, Marjanović D, Minarik M, Mizuno N, Moreira H, Morling N, Mukherjee M, Munier P, Nagaraju J, Neuhuber F, Nie S, Nilasitsataporn P, Nishi T, Oh HH, Olofsson J, Onofri V, Palo JU, Pamjav H, Parson W, Petlach M, Phillips C, Ploski R, Prasad SP, Primorac D, Purnomo GA, Purps J, Rangel-Villalobos H, Rębała K, Rerkamnuaychoke B, Gonzalez DR, Robino C, Roewer L, Rosa A, Sajantila A, Sala A, Salvador JM, Sanz P, Schmitt C, Sharma AK, Silva DA, Shin KJ, Sijen T, Sirker M, Siváková D, Skaro V, Solano-Matamoros C, Souto L, Stenzl V, Sudoyo H, Syndercombe-Court D, Tagliabracci A, Taylor D, Tillmar A, Tsybovsky IS, Tyler-Smith C, van der Gaag KJ, Vanek D, Völgyi A, Ward D, Willemse P, Yap EP, Yong RY, Pajnič IZ, Kayser M. Towards male individualization with rapidly mutating Y-chromosomal STRs . Human Mutation. 2014; 35(8): 1021-1032

16. Duncan Taylor, Jo-Anne Bright and John Buckleton. The ‘factor of two’ issue in mixed DNA profiles Journal of Theoretical Biology. 2014; 363: 300-306

17. Duncan Taylor, Jo-Anne Bright and John Buckleton. Considering relatives when assessing the evidential strength of mixed DNA profiles Forensic Science International: Genetics. 2014; 13: 259–263

18. Duncan Taylor, Jo-Anne Bright and John Buckleton. Interpreting forensic DNA profiling evidence without specifying the number of contributors Forensic Science International: Genetics. 2014; 13: 269-280

19. Jo-Anne Bright, John S. Buckleton, Duncan Taylor, M. Fernando and James M. Curran. Modelling forward stutter: towards increased objectivity in forensic DNA interpretation (2014) Electrophoresis: 35 (21-22) 3152-3157

20. R Ottens, D Taylor, A Linacre. DNA profiles from fingernails using direct PCR. (2014) Forensic science, medicine, and pathology, 1-5

2015:

21. S Cooper, C McGovern, JA Bright, D Taylor, J Buckleton. Investigating a common approach to DNA profile interpretation using probabilistic software. (2015) Forensic Science International: Genetics 16, 121-131

22. D Taylor, J Buckleton. Do low template DNA profiles have useful quantitative data? (2015) Forensic Science International: Genetics 16, 13-16

23. JA Bright, IW Evett, D Taylor, JM Curran, J Buckleton. A series of recommended tests when validating probabilistic DNA profile interpretation software. (2015) Forensic Science International: Genetics 14, 125-131

24. Duncan Taylor, John Buckleton, Ian Evett. Testing likelihood ratios produced from complex DNA profiles. Forensic Science International: Genetics 16 (2015) 165–171

25. Renée Blackie, Duncan Taylor, Adrian Linacre. Successful direct amplification of nuclear markers from single dog hairs using DogFiler multiplex. Electrophoresis (2015) 36(17), 2082-2085

26. JEL Templeton, D Taylor, O Handt, P Skuza, A Linacre. Direct PCR Improves the Recovery of DNA from Various Substrates. Journal of forensic sciences (2015) 60(6), 1558-1562.

27. Simone Gittelson, Tim Kalafut, Steven Myers, Duncan Taylor, Tacha Hicks, Franco Taroni, Ian W Evett, Jo-Anne Bright, John Buckleton. A Practical Guide for the Formulation of Propositions in the Bayesian Approach to DNA Evidence Interpretation in an Adversarial Environment. Journal of Forensic Sciences. (2015) 61(1), 186-195

28. D Taylor, J Buckleton, JA Bright. Does the use of probabilistic genotyping change the way we should view sub-threshold data? Australian Journal of Forensic Sciences (2015) DOI: 10.1080/00450618.2015.1122082

29. Jo-Anne Bright, John Buckleton and Duncan Taylor. A response to 'How to cross-examine forensic scientists: A guide for lawyers'. Australian Bar Review (2015) 41, 1-4

2016:

30. D Taylor, JA Bright, C McGoven, C Hefford, T Kalafut, J Buckleton. Validating multiplexes for use in conjunction with modern interpretation strategies. Forensic Science International: Genetics (2016) 20, 6-19

31. Nano Nagle, Kaye N Ballantyne, Mannis van Oven, Chris Tyler-Smith, Yali Xue, Duncan Taylor, Stephen Wilcox, Leah Wilcox, Rust Turkalov, Roland AH van Oorschot, Peter McAllister, Lesley Williams, Manfred Kayser, Robert J Mitchell. Antiquity and diversity of aboriginal Australian Y-chromosomes. American Journal of Physical Anthropology (2016) 159(3), 367-381

32. D Taylor, JA Bright, J Buckleton. Using probabilistic theory to develop interpretation guidelines for Y-STR profiles. Forensic Science International: Genetics (2016) 21, 22-34

33. D Taylor, J Buckleton, JA Bright. Factors affecting peak height variability for short tandem repeat data. Forensic Science International: Genetics (2016) 21, 126-133

34. D Taylor, D Abarno, T Hicks, C Champod. Evaluating forensic biology results given source level propositions. Forensic Science International: Genetics (2016) 21, 54-67

35. D Taylor, D Abarno, E Rowe, L Rask-Nielsen. Observations of DNA transfer within an operational Forensic Biology Laboratory. Forensic Science International: Genetics (2016) 23, 33-49

36. J Buckleton, J Curran, J Goudet, D Taylor, A Thiery, B Weir. Population-specific Fst values for forensic STR markers. Forensic Science International: Genetics (2016) 23, 91-100

37. JA Bright, D Taylor, C McGovern, S Cooper, L Russell, D Abarno, J Buckleton. Developmental validation of STRmix™, expert software for the interpretation of forensic DNA profiles. Forensic Science International: Genetics (2016) 23, 226-239

38. D Taylor. The evaluation of exclusionary DNA results: A discussion of issues in R v Drummond. Law, Probability and Risk. 2016; 15(3): 175-197

39. D Taylor. Is technology the death of expertise? Forensic Science International: Genetics (2016) 24, e1-e3

40. D Taylor. Probabilistically determining the cellular source of DNA derived from differential extractions in sexual assault scenarios. Forensic Science International: Genetics (2016) 24, 124-135

41. D Taylor, T Hicks, C Champod. Using sensitivity analyses in Bayesian Networks to highlight the impact of data paucity and direct future analyses: a contribution to the debate on measuring and reporting the precision of likelihood ratios. Science & Justice (2016) 56(5), 402-410

42. D Taylor, D Powers. Teaching artificial intelligence to read electropherograms. Forensic Science International: Genetics (2016) 25, 10-18

43. R Blackie, D Taylor, A Linacre. DNA profiles from clothing fibers using direct PCR. Forensic Science, Medicine, and Pathology (2016) 12 (3), 331-335

44. Geoffrey Stewart Morrison, David H Kaye, David J Balding, Duncan Taylor, Philip Dawid, Colin GG Aitken, Simone Gittelson, Grzegorz Zadora, Bernard Robertson, Sheila Willis, Susan Pope, Martin Neil, Kristy A Martire, Amanda Hepler, Richard D Gill, Allan Jamieson, Jacob de Zoete, R Brent Ostrum, Amke Caliebe. A comment on the PCAST report: Skip the “match”/“non-match” stage. Forensic Science International (2016) - 272: e7-e9

45. Alex Biedermann, Christophe Champod, Graham Jackson, Peter Gill, Duncan Taylor, John Butler, Niels Morling, Tacha Hicks Champod, Joelle Vuille, Franco Taroni. Evaluation of forensic DNA traces when propositions of interest relate to activities:

analysis and discussion of recurrent concerns. Frontiers in Genetics (2016) 7 (article 215)

2017:

46. Duncan Taylor, James Curran, John Buckleton. Importance sampling allows Hd true tests of highly discriminating DNA profiles. Forensic Science International: Genetics (2017) – 27, 74-81

47. D Taylor, A Biedermann, L Samie, KM Pun, T Hicks, C Champod. Helping to distinguish primary from secondary transfer events for trace DNA. Forensic Science International: Genetics (2017) – 28, 155-177

48. Duncan Taylor, Jo-Anne Bright, John Buckleton. Commentary: A "Source" of Error: Computer Code, Criminal Defendants, and the Constitution. Frontiers in Genetics (2017) – 8

49. Duncan Taylor, Jo-Anne Bright, Catherine McGovern, Sharon Neville, Denise Grover. Allele frequency databases for GlobalFiler™ STR loci in Australian and New Zealand populations. Forensic Science International: Genetics (2017) – 28, e38-e40

50. Duncan Taylor, Oliva Handt, Damien Abarno, John Buckleton. Likelihood ratio formulae for disputed parentage when the product of conception is trisomic. International Journal of Legal Medicine, (2017) 313(6); 1-9

51. Tamyra Moretti, Rebecca Just, Susannah Kehl, Leah Willis, John Buckleton, Jo-Anne Bright, Duncan Taylor, Anthony Onorato. Internal validation of STRmix™ for the interpretation of single source and mixed DNA profiles. Forensic Science International: Genetics (2017) 29, 126-144

52. Jennifer Templeton, Duncan Taylor, Oliva Handt, Adrian Linacre. Typing DNA profiles from previously enhanced fingerprints using direct PCR. Forensic Science International: Genetics (2017) 29, 276-282

53. Jo-Anne Bright, Duncan Taylor, Simone Gittelson, John Buckleton. The paradigm shift in DNA profile interpretation. Forensic Science International: Genetics (2017) 31: e24-e32

54. Duncan Taylor, Ash Harrison, David Powers. An artificial Neural Network system to identify alleles in reference electropherograms. Forensic Science International: Genetics (2017) 30, 114-126

55. Duncan Taylor, Jo-Anne Bright, Hannah Kelly, Meng-Han Lin, John Buckleton. A fully continuous system of DNA profile evidence evaluation that can utilise STR profile data produced under different conditions within a single analysis. Forensic Science International: Genetics (2017) – 31; 149-154

56. John Buckleton, Jo-Anne Bright and Duncan Taylor. Response to Lander's Response to the ANZFSS council statement on the President's Council Of Advisors On Science And Technology Report. Australian Journal of Forensic Sciences. 2017; 49:366-8

2018:

57. D Taylor, A Biedermann, T Hicks, C Champod. A template for constructing Bayesian networks in forensic biology cases when considering activity level propositions. (2018) Forensic Science International: Genetics 33, 136-146

58. Jo-Anne Bright, Rebecca Richards, Maarten Kruijver, Hannah Kelly, Catherine McGovern, Alan Magee, Andrew McWhorter, Anne Ciecko, Brian Peck, Chase Baumgartner, Christina Buettner, Scott McWilliams, Claire McKenna, Colin Gallacher, Ben Mallinder, Darren Wright, Deven Johnson, Dorothy Catella, Eugene Lien, Craig O'Connor, George Duncan, Jason Bundy, Jillian Echard, John Lowe, Joshua Stewart, Kathleen Corrado, Sheila Gentile, Marla Kaplan, Michelle Hassler, Naomi McDonald, Paul Hulme, Rachel H Oefelein, Shawn Montpetit, Melissa Strong, Sarah Noël, Simon Malsom, Steven Myers, Susan Welti, Tamyra Moretti, Teresa McMahon, Thomas Grill, Tim Kalafut, MaryMargaret Greer-Ritzheimer, Vickie Beamer, Duncan A Taylor, John S Buckleton. Internal validation of STRmix™–A multi laboratory response to PCAST. Forensic Science International: Genetics. 2018; 34, 11-24

59. Duncan Taylor, James Curran and John Buckleton. Likelihood ratio development for mixed Y-STR profiles. Forensic Science International: Genetics. 2018; 35, 82-96

60. Hannah Kelly, Jo-Anne Bright, Maarten Kruijver, Stuart Cooper, Duncan Taylor, Kyle Duke, Melissa Strong, Vickie Beamer, Christina Buettner, John Buckleton. A sensitivity analysis to determine the robustness of STRmix™ with respect to laboratory calibration. Forensic Science International: Genetics 2018; 35; 82-96

61. Simone Gittelson, Charles E.H. Berger, Graham Jackson, Ian W. Evett, Christophe Champod, Bernard Robertson, James M. Curran, Duncan Taylor, Bruce S. Weir, Michael D. Coble, John S. Buckleton. A response to "Likelihood ratio as weight of evidence: A closer look" by Lund and Iyer. Forensic Science International. 2018;288; e15-e19

62. Tim Kalafut, Curt Schuerman, Joel Sutton, Tom Faris, Luigi Armogida, Jo-Anne Bright, John Buckleton, Duncan Taylor. Implementation and validation of an improved allele specific stutter filtering method for electropherogram interpretation. Forensic Science International: Genetics 2018; 35; 50-56

63. Belinda Martin, Renee Blackie, Duncan Taylor, Adrian Linacre. DNA Profiles Generated from a Range of Touched Sample Types. Forensic Science International: Genetics 2018; 36; 13-19

64. Duncan Taylor, Bas Kokshoorn, Alex Biedermann. Evaluation of forensic genetics findings given activity level propositions: A review. Forensic Science International: Genetics (Special Issue on Trends and Perspectives in Forensic Genetics 2018) 2018; 36; 34-39

65. Peter Gill, Tacha Hicks, John M Butler, Ed Connolly, Leonor Gusmão, Bas Kokshoorn, Niels Morling, Roland AH van Oorschot, Walther Parson, Mechthild Prinz, Peter M Schneider, Titia Sijen, Duncan Taylor. DNA Commission of the International Society for Forensic Genetics: Assessing the value of forensic biological

evidence-guidelines highlighting the importance of propositions: Part I: Evaluation of DNA profiling comparisons given (sub-) source propositions. Forensic Science International: Genetics 2018; 36; 189-202

66. Mikkel Meyer Andersen, James Curran, Jacob de Zoete, Duncan Taylor, John Buckleton. Modelling the dependence structure of Y-STR haplotypes using graphical models. Forensic Science International: Genetics 2018; 37; 29-36

2019:

67. John S Buckleton, Jo-Anne Bright, Simone Gittelson, Tamyra R Moretti, Anthony J Onorato, Frederick R Bieber, Bruce Budowle, Duncan A Taylor. The Probabilistic Genotyping Software STRmix: Utility and Evidence for its Validity. Journal of Forensic Sciences 2019; 64(2) 393-405

68. D Taylor, M Kitselaar, D Powers. The generalisability of artificial neural networks used to classify electrophoretic data produced under different conditions. Forensic Science International: Genetics 2019; 38; 181-184

69. John S. Buckleton, Jo-Anne Bright, Kevin Cheng, Hannah Kelly, and Duncan A. Taylor. The effect of varying the number of contributors in the prosecution and alternate propositions. Forensic Science International: Genetics 2019; 38; 225-231

70. D Abarno, T Sobieraj, C Summers, D Taylor. The first Australian conviction resulting from a familial search. Australian Journal of Forensic Sciences Supplementary 1. 2019; S56-S59

71. D. Catoggio, J. Bunford, D. Taylor, G. Wevers, K. Ballantyne & R. Morgan. An introductory guide to evaluative reporting in forensic science. Australian Journal of Forensic Sciences Supplementary 1. 2019. S247-S251

72. Duncan Taylor, John Buckleton, Jo-Anne Bright. Comment on “DNA mixtures interpretation – a proof-of-concept multi-software comparison highlighting different probabilistic methods’ performances on challenging samples” by Alladio et al. Forensic Science International: Genetics: 2019; 40; e248-e251

73. Jo-Anne Bright, Duncan Taylor, Zane Kerr, John Buckleton, Maarten Kruijver. The efficacy of DNA mixture to mixture matching. Forensic Science International: Genetics: 2019; 41; 67-71

74. Duncan Taylor, Emily Rowe, Maarten Kruijver, Damien Abarno, Jo-Anne Bright and John Buckleton. Inter-sample contamination detection using mixture deconvolution comparison. Forensic Science International: Genetics 2019; 40; 160-167

75. Catherine Hopkins, Duncan Taylor, Kelly Hill, Julianne Henry. Analysis of the South Australian Aboriginal population using the Global AIMs Nano ancestry test. Forensic Science International: Genetics: 2019; 41; 34-41

76. Julianne Henry, Hoa Dao, Lenara Scandrett, Duncan Taylor. Population genetic analysis of Yfiler® Plus haplotype data for three South Australian populations. Forensic Science International: Genetics: 2019; 41; e23-e25

77. Jo-Anne Bright, Kevin Cheng, Zane Kerr, Catherine McGovern, Hannah Kelly, Tamyra R. Moretti, Michael A. Smith, Frederick R. Bieber, Bruce Budowle, Michael D. Coble, Rashed Alghafri, Paul Stafford Allen, Amy Barber, Vickie Beamer, Christina Buettner, Melanie Russell, Christian Gehrig, Tacha Hicks, Jessica Charak, Kate Cheong-Wing, Anne Ciecko, Christie T. Davis, Michael Donley, Natalie Pedersen, Bill Gartside, Dominic Granger, MaryMargaret Greer-Ritzheimer, Erick Reisinger, Jarrah Kennedy, Erin Grammer, Marla Kaplan, David Hansen, Hans J. Larsen, Alanna Laureano, Christina Li, Eugene Lien, Emilia Lindberg, Ciara Kelly, Ben Mallinder, Simon Malsom, Alyse Yacovone-Margetts, Andrew McWhorter, Sapana M. Prajapati, Tamar Powell, Gary Shutler, Kate Stevenson, April R. Stonehouse, Lindsey Smith, Julie Murakami, Eric Halsing, Darren Wright, Leigh Clark, Duncan A. Taylor, John Buckleton. STRmix™ collaborative exercise on DNA mixture interpretation. Forensic Science International: Genetics 2019; 40; 1-8

78. Duncan Taylor, Lydie Samie and Christophe Champod. Using Bayesian Networks to track DNA movement through complex transfer scenarios. Forensic Science International: Genetics 2019; 42; 69-80

79. Laura Russell, Stuart Cooper, Richard Wivell, Zane Kerr, Duncan Taylor, John Buckleton and Jo-Anne Bright. A guide to results and diagnostics within a STRmix™ report. WIREs Wiley Interdisciplinary Reviews (2019) e1354

80. John S. Buckleton, Kirk E. Lohmueller, Keith Inman, Kevin Cheng, James M. Curran, Simone N. Pugh, Jo-Anne Bright, Duncan A. Taylor. Testing whether stutter and low-level DNA peaks are additive. Forensic Science International: Genetics 2019; 43; 102166

81. JA Bright, H Kelly, Z Kerr, C McGovern, D Taylor, JS Buckleton. The interpretation of forensic DNA profiles: an historical perspective. Journal of the Royal Society of New Zealand, 2019; 50(2) 1-15

2020:

82. Peter Gill, Tacha Hicks, John M Butler, Ed Connolly, Leonor Gusmão, Bas Kokshoorn, Niels Morling, Roland AH van Oorschot, Walther Parson, Mechthild Prinz, Peter M Schneider, Titia Sijen, Duncan Taylor. DNA Commission of the International Society for Forensic Genetics: Assessing the value of forensic biological evidence-guidelines highlighting the importance of propositions: Part II: Evaluation of biological traces considering activity level propositions. Forensic Science International: Genetics 2020; 44; 102186

83. Duncan Taylor, Bas Kokshoorn, Tacha Hick. Structuring cases into propositions, assumptions, and undisputed case information. Forensic Science International: Genetics 2020; 44; 102199

84. John Buckleton, Miss Anne Ciecko, Maarten Kruijver, Benjamin Mallinder, Alan Magee, Simon Malsom, Tamyra Moretti, Steven Weitz, Todd Bille, Sarah Noel, Rachel Oefelein, Brian Peck, Tim Kalafut, Duncan Taylor. Response to: Commentary on: Bright et al. (2018) Internal validation of STRmix™ – a multi laboratory response

to PCAST, Forensic Science International: Genetics, 34: 11–24. Forensic Science International: Genetics 2020; 44;

85. K Cheng, JA Bright, Z Kerr, D Taylor, A Ciecko, J Curran, J Buckleton. Examining the additivity of peak heights in forensic DNA profiles. Australian Journal of Forensic Sciences, 2020 (IN PRESS)

86. Belinda Martin, Piyamas Kanokwongnuwut, Duncan Taylor, Paul Kirkbride, David Armitt, Adrian Linacre. Successful STR Amplification of Post-Blast IED Samples by Fluorescent Visualisation and Direct PCR. Forensic Science International: Genetics 2020; 46; 102256

87. P Ramos, O Handt, D Taylor. Investigating the position and level of DNA transfer to undergarments during digital sexual assault. Forensic Science International: Genetics 2020; 47; 102316

88. Lutz Roewer, Mikkel Meyer Andersen, Jack Ballantyne, John M Butler, Amke Caliebe, Daniel Corach, Maria Eugenia D'Amato, Leonor Gusmão, Yiping Hou, Peter de Knijff, Walther Parson, Mechthild Prinz, Peter M Schneider, Duncan Taylor, Marielle Vennemann, Sascha Willuweit. DNA Commission of the International Society of Forensic Genetics (ISFG): Recommendations on the Interpretation of Y-STR results in Forensic Analysis. Forensic Science International: Genetics 2020; 47; 102308

89. Lydie Samie, Christophe Champod, Duncan Taylor, Franco Taroni. The use of Bayesian Networks and simulation methods to identify the variables impacting the value of evidence assessed under activity level propositions in stabbing cases. Forensic Science International: Genetics 2020; 48; 102334

90. Catherine McGovern, Kevin Cheng, Hannah Kelly, Anne Ciecko, Duncan Taylor, John Buckleton, Jo-Anne Bright. Performance of a method for weighting a range in the number of contributors in probabilistic genotyping. Forensic Science International: Genetics 2020; 102352.

91. Duncan Taylor, David Balding. How can courts take into account the uncertainty in a likelihood ratio? Forensic Science International: Genetics 2020; 102361.

92. Duncan Taylor, Maarten Kruijver. Combining evidence across multiple mixed DNA profiles for improved resolution of a donor when a common contributor can be assumed. Forensic Science International: Genetics 2020; 49; 102375.

93. John Buckleton, Simone N Pugh, Jo-Anne Bright, Duncan Alexander Taylor, James Michael Curran, Maarten Kruijver, Peter David Gill, Bruce Budowle, Kevin Cheng. Are low LRs reliable? Forensic Science International: Genetics 2020; 49; 102350.

94. Paul Stafford Allen, Simone Nicole Pugh, Jo-Anne Bright, Duncan Alexander Taylor, John Simon Buckleton. Relaxing the assumption of unrelatedness in the numerator and denominator of likelihood ratios for DNA mixtures. Forensic Science International: Genetics. 2020; 51; 102434.

2021:

95. John Buckleton, Duncan Alexander Taylor, Jo-Anne Bright, Tacha Hicks, James Michael Curran. When evaluating evidence within a likelihood ratio framework, should the propositions be exhaustive? Forensic Science International: Genetics. 2021; 50; 102406.

96. Jo-Anne Bright, John Buckleton, Duncan Taylor. Probabilistic interpretation of the Amelogenin locus. Forensic Science International: Genetics. 2021; 52; 102462.

97. John Buckleton, James Curran, Duncan Taylor, and Jo-Anne Bright. What can forensic software developers learn from four significant software failures? 2021; Wiley Interdisciplinary Reviews: Forensic Science 3 (2), e1398.

98. Tacha Hicks, Zane Kerr, Simone Pugh, Jo-Anne Bright, James Curran, Duncan Taylor, John Buckleton. Comparing multiple POI to DNA mixtures. Forensic Science International: Genetics 2021; 52; 102481.

99. Piyamas Kanokwongnuwut, Belinda Martin, Duncan Taylor, K. Paul Kirkbride, Adrian Linacre. How many cells are required for successful DNA profiling? Forensic Science International: Genetics 2021; 51; 102453.

100. Duncan Taylor, Jo-Anne Bright, Lenara Scandrett, Damien Abarno, Shan-I Lee, Richard Wivell, Hannah Kelly, John Buckleton. Validation of a top-down DNA profile analysis for database searching using a fully continuous probabilistic genotyping model. Forensic Science International: Genetics 2021; 52; 102479.

101. Hannah Kelly, Jo-Anne Bright, Maarten Kruijver, Duncan Taylor, John Buckleton. The effect of user selected number of contributors within the LR assignment. Australian Journal of Forensic Science. 2021; IN PRESS.

102. Jess Champion, Piyamas Kanokwongnuwut, Roland van Oorschot, Duncan Taylor, and Adrian Linacre. Evaluation of a fluorescent dye to visualise touch DNA on various substrates. Journal of Forensic Science. 2021. IN PRESS

103. Duncan Taylor, Luke Volgin, Bas Kokshoorn, Christophe Champod. The importance of considering common sources of unknown DNA when evaluating findings given activity level propositions. Forensic Science International: Genetics 2021; 53; 102518

104. Duncan Taylor, John Buckleton. Can a reference ‘match’ an evidence profile if these have no loci in common? Forensic Science International: Genetics 2021; 53. 102520

105. Maarten Kruijver, Duncan Taylor, Jo-Anne Bright. Evaluating DNA evidence possibly involving multiple (mixed) samples, common donors and related contributors. Forensic Science International: Genetics. 2021. 54. 102523.

106. Lucas Puliatti, Oliva Handt, Duncan Taylor. The level of DNA an individual transfers to untouched items in their immediate surroundings. Forensic Science International: Genetics 2021; 55. 102561

107. Kevin Cheng, Øyvind Bleka, Peter Gill, James Curran, Jo-Anne Bright, Duncan Taylor, John Buckleton. A comparison of likelihood ratios obtained from EuroForMix and STRmix™. Journal of Forensic Sciences. IN PRESS.

108. Peter Gill, Corrina Benschop, John Buckleton, Øyvind Bleka, Duncan Taylor. A Review of Probabilistic Genotyping Systems: EuroForMix, DNAStatistX and STRmix™. Genes 2021; 12 (10); 1559

109. Luke Volgin, Duncan Taylor, Jo-Anne Bright, Meng-Han Lin. Validation of a neural network approach for STR typing to replace human reading. Forensic Science International: Genetics, 2021; 102591

110. Tacha Hicks, Zane Kerr, Simone Pugh, Jo-Anne Bright, James Curran, Duncan Taylor, John Buckleton. Comparing multiple POI to DNA mixtures. Forensic Science International: Genetics, 2021; 52. 102481

111. Belinda Martin, Duncan Taylor, Adrian Linacre. Comparison of six commercially available STR kits for their application to touch DNA using direct PCR. Forensic Science International: Reports. 2021; 100243.

112. Duncan Taylor, Damien Abarno. Using big data from probabilistic genotyping to solve crime. Forensic Science International: Genetics, 2021; 102631

2022:

113. Duncan Taylor. Using a multi-head, convolutional neural network with data augmentation to improve electropherogram classification performance. Forensic Science International: Genetics, 2022; 102605

114. Claire Mercer, Julianne Henry, Duncan Taylor, Adrian Linacre. What's on the bag? The DNA composition of evidence bags pre-and post-exhibit examination. Forensic Science International: Genetics, 2022; 102652

115. Belinda Martin, Duncan Taylor, Adrian Linacre. Exploring tapelifts as a method for DUAL workflow STR amplification. Forensic Science International: Genetics, 2022; 102653

116. D Taylor, D Abarno. Using big data from probabilistic genotyping to solve crime. Forensic Science International: Genetics. 2022. 57, 102631

117. D Ward, J Henry, D Taylor. Analysis of mixed DNA profiles from the RapidHIT™ ID platform using probabilistic genotyping software STRmix™. Forensic Science International: Genetics, 2022. 102664

118. Sasha Carson, Luke Volgin, Damien Abarno, Duncan Taylor. The potential for investigator-mediated contamination to occur during routine search activities. Forensic Science Medicine and Pathology, 2022. IN PRESS

119. Rhianna Curtis, Denise Ward, Duncan Taylor and Julianne Henry. Investigation of X-STR haplotype diversity in the Australian Aboriginal population. Australian Journal of Forensic Science, 2022. IN PRESS

120. T Kalafut, JA Bright, D Taylor, J Buckleton. Investigation into the effect of mixtures comprising related people on non-donor likelihood ratios, and potential practises to mitigate providing misleading opinions. Forensic Science International: Genetics, 2022, 102691

121. John Buckleton, Jo-Anne Bright, Duncan Taylor, Richard Wivell, Øyvind Bleka, Peter Gill, Corina Benschop, Bruce Budowle, Mike Coble. Re: Riman et al. Examining performance and likelihood ratios for two likelihood ratio systems using the PROVEDIt dataset. Forensic Science International: Genetics, 2022. IN PRESS

122. Hannah Kelly, Jo-Anne Bright, Maarten Kruijver, Duncan Taylor, John Buckleton. The effect of a user selected number of contributors within the LR assignment. Australian Journal of Forensic Sciences; 2022, 54 (4), 450-463

123. John Buckleton, Jo-Anne Bright, Duncan Taylor, Richard Wivell, Øyvind Bleka, Peter Gill, Corina Benschop, Bruce Budowle, Michael Coble. Re: Riman et al. Examining performance and likelihood ratios for two likelihood ratio systems using the PROVEDIt dataset. Forensic Science International: Genetics; 2022. IN PRESS

Published books:

Forensic DNA Evidence Interpretation Second Edition. Editors John Buckleton, Jo-Anne Bright, Duncan Taylor. (2016) CRC Press. ISBN 9781482258899

Forensic Biology Evidence Evaluation: Utilizing Activity Level Propositions and Likelihood Ratios. Duncan Taylor and Bas Kokshoorn. (2022) CRC Press. ISBN 9781032225814

Contribution to books:

Parentage analysis and other applications of human identity testing (Chapter 82). Duncan Taylor, In I. Freckelton & H. Selby (Eds), *Expert evidence*. North Ryde, Australia; Thomson Lawbook Co.

'Complex Mixtures' (Chapter 19). Duncan Taylor, John Buckleton, Jo-Anne Bright, In Encyclopaedia of Forensic Science, Third Edition: Section 10023... Senior Editor Max Houck. Academic Press, Elsevier. ISBN 978-0-12-382165-2

Grants

2019 – 2023 – $27 000 – South Australian Police - Humphries M, Roughan M, Taylor D. "Recommender systems for forensic evidence triage"

2021 – $15 000 – Australian Academy of Forensic Science Research Fellowship Award. Duncan Taylor, Adrian Linacre, Russel Brinkworth. "Using machine learning to improve PCR"

Patents

63/037,475 - 10 June 2020 - provisional patent - Methods and systems for identifying nucleic acids

Contributions to Forensic Science

Technical developer of software STRmix™, used for DNA evidence interpretation in Australia, New Zealand and parts of USA. STRmix™ training courses provided in:

- Melbourne, Australia
- Auckland, New Zealand
- Manchester, England
- Washington, USA
- Las Vegas, USA
- Belfast, Northern Ireland
- Dublin, Ireland

2017 - Pioneered Activity level evaluation of DNA evidence in Australia. Member of the Australia New Zealand working group to produce the ANZPAA-NIFS "An introductory Guide to Evaluative Reporting".

2019 - Technical co-developer of common DNA donor analysis in software DBLR™.

2020 - Technical developer of Artificial Neural Network functionalist in FaSTR™, a DNA profile reading software.

2020 - Invited to be involved in the development of expert assessment and registration for activity level evaluation by Nederlands Register Grechtelijk Deskundigen (Netherlands Register for Court Experts).

2021 – Membership of Standards Australia Committees and Joint Standards Australia/Standards New Zealand Committees for work on ISO 21043

Work presented at conferences

Presented at the Australia and New Zealand Forensic Science Society symposium in Melbourne 2008:

- Y-chromosome short tandem repeat (Y-STR) diversity in South Australian Aboriginal and Caucasian populations – Duncan A. **Taylor**, Robert J. Mitchell, Roland van Oorschot, Nano Nagle, Julianne M. Henry.

Presented at the International Society of Forensic Genetics symposium in Buenos Aires 2009:

- Knowing your DNA database: issues with determining ancestral Y haplotypes in a Y-Filer database - Duncan A. **Taylor**, Robert J. Mitchell, Roland van Oorschot, Nano Nagle, Julianne M. Henry.

Presented at the Australia and New Zealand Forensic Science Society symposium in Sydney 2010:

- Modelling stochastic effects from empirical data to develop interpretational tools and guidelines – Duncan Taylor, Christopher Hefford – Won the award for best biology presentation 2010.
- Introducing rules to a staff DNA profile validation process to improve the detection rate of contamination events – Duncan Taylor, Christopher Hefford.

- Resolving the extent of admixture in an Australian Aboriginal Y-STR database - Duncan A. Taylor, Robert J. Mitchell, Roland van Oorschot, Nano Nagle, Julianne M. Henry.
- Novel and Rare Y-Chromosome Short Tandem Repeats At DYS456 And DYS635 In Australian Aborigines – Tegan E Collins, Michael Gardner, Julianne M Henry, Duncan A Taylor, Alison J Fitch, Amanda Goodman.
- The Derivation of a Paternity Index where the Mother and Father are Biological Brother and Sister - Damian Abarno, Duncan Taylor.

Presented at the International Society of Forensic Genetics symposium in Vienna 2011:

- Novel and rare Y-chromosome short tandem repeats in Australian Aborigines

Asian Forensic Sciences Network in 2011:

- Population frequency study for Y-STR loci for Brunei Darussalam Malay and Chinese

Presented at the Australia and New Zealand Forensic Science Society symposium in Hobart 2012:

- STRmix: sophisticated DNA profile analysis for forensic scientists (Keynote address)
- Evaluation and statistical analysis of data pertaining to the persistence of seminal components after sexual assault
- A brother comes to the rescue when a mother is not enough
- Quick and easy semi-automated DNA reporting using Microsoft Office
- DNA profiling of soils using next generation sequencing
- The impact of Aboriginal database admixture on weight of evidence calculations for uniparental and autosomal markers

Presented at International Society of Forensic Genetics symposium in Melbourne 2013:

- How certain are we about our statistics? - D. Taylor, J. Bright, J. Buckleton, J. Curran
- Going totally Bayesian: Lab experiences when moving to a continuous DNA interpretation model – D. Taylor

Presented at the Australia and New Zealand Forensic Science Society symposium in Adelaide 2014:

- Removing the need to specify a number of contributors for DNA interpretation - D. Taylor, J. Bright, J. Buckleton
- Using continuous DNA interpretation systems to revisit likelihood ratio behaviour - D. Taylor
- Contamination or coincidence: Determining the appropriate likelihood ratio threshold for contamination detection using STRmix™ - J. Henry, D. Abarno, D. Taylor
- The effectiveness of STRmix™ software to detect contamination of forensic samples by laboratory personnel - J. Henry, D. Abarno, D. Taylor
- Direct PCR improves the recovery of DNA from various substrates - Jennifer E.L Templeton, Renée Ottens, Oliva Handt, Duncan Taylor, Adrian Linacre

Presented at the 5th International Conference on Evidence Law and Forensic Science in Adelaide 2015:

- Using Bayesian Networks to put DNA findings in a greater case context

Presented at International Society of Forensic Genetics symposium in Krakow, Poland 2015:

- Using Hd true tests to inform on model performance and address adventitious matching – Duncan Taylor, John Buckleton, Ian Evett
- The interpretation of y chromosome mixtures - Moretti T. R., Myers S. P., Taylor D., Bright J. A., Buckleton J. S
- Interpreting mixed DNA profiles considering a range in the assigned number of contributors - Cooper S. J., McGovern C. E., Abarno D., Bright J. A., Taylor D., Buckleton J. S.
- Vectors of DNA transfer in a laboratory environment – Taylor D., Abarno D., Rowe E., Rask-Nielsen L.
- DNA profiles from fingermarks - Templeton J. E. L., Blackie R., Taylor D., Handt O., Linacre A.

Presented at the Australia and New Zealand Forensic Science Society symposium in Auckland 2016:

- Using sensitivity analyses on Bayesian networks to assess sampling uncertainty and direct further research – Duncan Taylor, Tacha Hicks, Christophe Champod (oral)
- Direct PCR: successes and limitations. Templeton J, Blackie R, Rowe E, Taylor D, Handt O, Linacre A (poster)
- Is standardisation of DNA profile interpretation achievable? Stuart Cooper. Laura Russell, Jo-Anne Bright, Catherine McGovern, Duncan Taylor, John Buckleton (oral)
- Direct PCR of Hair Samples – A success story? Oliva Handt, Mel Sifis, Duncan Taylor (oral)

Presented at the Australia and New Zealand Forensic Science Society symposium in Perth 2018:

- How much DNA accumulates on untouched items in the home? - Taylor D, Moroney M, Linacre A (poster)
- Evaluating mixed Y-STR profiles - Taylor D, Curran J, Buckleton J (oral)
- Introducing activity level reporting to casework - Taylor D (oral)
- Using artificial neural networks to read electropherograms - Taylor D, Harrison A, Kitselaar M, Powers D (oral)
- SNP panel DNA profiles from touched sample. Adrian Linacre, Duncan Taylor (oral presented by Linacre)
- Validation of the Qiagen Argus x-12 QS X-STR PCR kit for use in familial search candidate exclusionary work. Abarno DV, Pearce M, Rowe E, Scandrett L, Taylor DA, Linacre A (poster presented by Abarno)
- Musings on the first Australian conviction resulting from a familial search - Abarno DV, Summers C, Sobieraj TC, Taylor DA (oral presented by Abarno)
- DNA profiles from touched samples - Martin B, Blackie R, Kirkbride P, Taylor D, Linacre A (oral presented by Linacre)
- Assessment of changes to DNA database interrogation at forensic science SA - Collins T, Dubrich J, Stankovic D, Williams T, Windram R, Taylor DA, Abarno DV (oral presented by Collins)
- An introductory guide to evaluative reporting - Catoggio D, Bunford J, Taylor D, Wevers G, Ballantyne K, Morgan R (poster presented by Morgan)

Presented at International Society of Forensic Genetics symposium in Prague, Czezh Republic 2019:

- Applying autosomal STR probabilistic genotyping models to SNP data using hierarchical Bayesian modelling – Duncan Taylor, Julianne Henry, Catherine Hopkins, James Curran (poster)
- Modelling DNA transfers in complex scenarios - Duncan Taylor, Tacha Hicks-Champod, Christophe Champod (oral)
- From reference to mixture to mixture to mixture and beyond - Maarten Kruijver, Duncan Taylor (oral presented by Kruijver)
- Application of the GNano 31-plex ancestry prediction assay in an Australian context - Catherine Hopkins, Duncan Taylor, Kelly Hill and Julianne Henry (poster presented by Henry)
- Verification of the GNano 31-plex ancestry prediction assay for forensic casework - Julianne Henry, Catherine Hopkins, Kelly Hill and Duncan Taylor (poster presented by Henry)

Presented at the American Academy of Forensic Sciences in USA, 2021:

- A two-trace problem in probabilistic genotyping: should the evidence be combined or not - Maarten Kruijver, Duncan Taylor (oral presented by Kruijver)

Presented at the Australia and New Zealand Forensic Science Society symposium in Brisbane 2022:

- Do we need to read profiles anymore? Combining neural network profile processing and probabilistic genotyping - Duncan Taylor, John Buckleton (oral presented by Duncan Taylor)
- Examples of new investigative leads generated from large-scale, inter-case crime scene profile comparisons – Duncan Taylor, Damien Abarno (oral presented by Duncan Taylor)
- Covert sampling, familial searching and extradition. Solving the murder of Suzanne Poll – Anthony van der Stelt, Duncan Taylor (oral presented jointly by Anthony van der Stelt and Duncan Taylor)
- AI and machine learning for DNA evidence: addressing the practical legal issues – Kerry Ann Andresen, Caitlin Williams, Linzi Wilson-Wilde, Duncan Taylor (oral presented jointly by Kerry Ann Andresen and Caitlin Williams)
- Real data making big data: using Bayesian networks to build a digital twin for case exhibit submission – Louise Campbell, Melissa Humphries, Duncan Taylor (poster presented by Louise Campbell)

Level of incidental DNA transfer from bedding – Denise Ward, Oliva Handt, Duncan Taylor (poster presented by Denise Ward)

Presented at International Society of Forensic Genetics symposium in Washington DC, USA 2022:

- An experimental extension to the discrete Laplace method for Y-STR haplotype frequency estimation - Maarten Kruijver, Duncan Taylor, John Buckleton (oral presented by Maarten Kruijver)
- DNA transfer between exhibits, evidence bags & workspaces - Claire Mercer, Adrian Linacre, Duncan Taylor, Dr Julianne Henry (oral presented by Claire Mercer)

Lecturing Duties

Lectures given in:

2011 – present: Flinders University - BIOL3792 (Forensic Biology)

2017 – 2019: Murdoch University - Forensic Science Professional Practise

Student Supervision

Current:

– Claire Mercer – PhD – trace DNA transfer during DNA exhibit transport and analysis
– Louise Campbell – PhD – using recommender systems for forensic exhibit triage
– Isla Madden – Honours – Predicting probative levels of touch DNA on forensic DNA tapelifts using Diamond™ Nucleic Acid Dye
– Druvi Patel – Honours – Collection of DNA from Airspaces – a look at contactless DNA transfer
– Caitlyn McDonald – Honours – Applying machine learning to PCR conditions to improve DNA profiling

2022 – Honours – Ayesha Khalid Ahmed Khan - Improving PCR efficiency by using API

2022 – Kerry Andresen and Cailin Williams – Adelaide Uni Law School project - The application of artificial intelligence and machine learning to DNA profiling: admissibility under the rules of evidence issues paper
2022 – Bridget Alyward – 3rd year Adelaide Uni Law School project – Admissibility issues around the use of activity level evaluations in South Australian Courts

2021 – Lingchen Wang – Honour – adapting standard PCR thermocyclers to provide real-time feedback to a machine learning system
2018 – 2021 Belinda Matulick (nee Martin) – PhD – trace DNA analysis on improvised explosive devices

2020 – Lucas Puliatti – Honours – Investigating the level of DNA transfer from a brief visit
2020 – Sasha Carson – Honours – Investigating the potential for cross-contamination at a crime scene
2020 – Cara-Mae Shipley – Honours – Validation of the HIrisPlex SNP kit

2019 – 2020 – Partho Protim Gosh – Masters – Using ANN to determine number of contributors

2019 – Phola Ramos – Honours – DNA transfer to clothing during simulated sexual assaults

2017 – 2019 - Suni Edson – PhD student from Armed Forces DNA Identification Laboratory (AFDIL) dealing with the processing and profiling of bone samples.

2018 – Joshua Sweaney – honours – application of artificial neural networks to forensic biology DNA detection and analysis
2018 – Catherine Hopkins – honours – Creation of a GNano SNP database for Aboriginal Australians
2018 – Kirsten Heuer – honours – Development of an in-house Y-SNP ancestry assay for the enhancement of the FSSA Aboriginal Y-STR DNA database

2017 – Michael Kistelaar – Work placement student from Flinders University - Using deep learning neural networks for interpretation of complex electrophoretic data
2017 – Belinda Matulick – Honours – Developing a SNaPshot panel for the identification of ancestry Aboriginal Y-chromosomes

2017 – Monique Moroney – Honours – Analysis of trace DNA transfer and persistence to fill activity level reporting gaps
2017 – Jess Champion – Honours – DNA transfer and persistence

2016 – Ashleigh Harrison – Summer student from Flinders University - Using deep learning neural networks for interpretation of electrophoretic data
2016 – Melissa Drogemuller – Summer student from Flinders University - Using deep learning neural networks for forensic handwriting comparison
2016 – Renée Blackie (nee Ottens) – PhD student from Flinders University - Direct PCR as a means to generate DNA profiles from trace material such as hair and fibres
2013 - 2016 – Jennifer Templeton – PhD student from Flinders University – Studies on Low Template DNA for Forensic Human Identification

2011 – Renée Ottens - Honours Student from Flinders University - Novel Y-Chromosome Short Tandem Repeat Sequences.

2009 – Tegan Collins - Honours Student from Flinders University - Novel and Rare Y-Chromosome Short Tandem Repeats at DYS456 and DYS635 in Australian Aborigines.

2008 – Ankita Chitalia - Masters Student from Flinders University - Forensic DNA profiling technology: Driving trace DNA profiling to its technical limit; particularly with post-amplification procedures.

Invited speaking events & workshop presentations
2022 – presented series of six workshops on activity level evaluation as part of an ANZPAA-NIFS training workshop

2021 – 13th Asian Forensic Science Network Annual Meeting and Symposium – Discussion panel on “Reporting probabilistic genotyping in court; lessons from the stand”

2021 – 7th Annual STRmix workshop – ‘Y-STRs in STRmix (a.k.a. STRmixY)’

2021 – 3rd Annual Northeast Forensic Laboratory Probabilistic Genotyping Users Group Meeting - ‘Factoring uncertainty into evaluations—The HPD interval in STRmix’

2021 - 6th annual Questioning Forensics conference hosted by the DNA Unit of the Legal Aid Society in New York City DNA Unit speaking on Bayesian Networks and activity level reporting

2019 - 2020 – Lectures given in online course ‘DNA Interpretation given activity level propositions’ run by Tacha Hicks from Lausanne University

2019 – Ontario Centre of Forensic Sciences workshop on evaluative reporting – “Australian practitioner perspective on evaluative reporting”

2019 – Web series: Probabilistic Genotyping of Forensic Evidentiary Typing Results – “What can ‘big data’ tell us about performance? Multi-lab studies, PCAST, sensitivity/specificity and ROC plots”

2019 – Australian Defence Lawyers Alliance Conference – "What do the DNA results really mean?"

2018 – Australia and New Zealand Forensic Science Society symposium in Perth 2018 workshop – "Activity Level Inference in Forensic Genetics"

2018 – Adelaide Festival of Ideas "My teacher said I'd need maths one day: Mathematical techniques you never knew were being used by Forensic Science SA to solve crime"

2018 – Gordon Research Conference: Forensic Analysis of Human DNA. Maine USA – "Probabilistic genotyping software"

2017 – Open State. Future Forensics: Crime scene to courtroom discussion panel. The Dome. Victoria Square Adelaide.

2017 - Griffith University Innocence Project and the Griffith Law Criminal Justice Symposium: Lifting the Veil on DNA Evidence: What Do the Statistics Really Mean?

2016 – Document Examination Specialist Advisory group, Melbourne 2017 - Logical Reporting for Forensic Handwriting and Signature Examinations

2015– International Society of Forensic Genetics symposium in Krakow, Poland 2015 workshop – Interpretation of complex DNA profiles using a continuous model – an introduction to STRmix™

2014 – International Symposium on Advances in Legal Medicine – Fukuoka Japan – Invited to speak on the topic of Advances in DNA evidence interpretation

2013 – International Society of Forensic Genetics symposium in Melbourne 2013 – Lectured at the Basic and Advanced DNA interpretation workshops on population genetics, continuous DNA interpretation systems and implementation of continuous DNA interpretation systems

2013 – Australian Association of Crown Prosecutors in Adelaide – Invited to speak on Familial Searching and STRmix™

2013 – Magistrates Judicial Development – Invited to speak on STRmix™

2012 – Australian Association of Crown Prosecutors in Darwin – Invited to speak on Population genetics

Positions held

2021 – present: Member of the Australasian working group for activity level reporting

2021 – present: Expert and Assessor for the NRGD (Netherlands Register for Judicial Experts) in the field of DNA Activity Level evaluations

2020 – 2021: Associate Investigator member of the Australian Research Council Centre of Excellence for Mathematical and Statistical Frontiers (ACEMS)

2019 – present: Editorial board member of Forensic Science International Genetics journal

2016 – 2019: Member of the International Society of Forensic Genetics working group on evidence interpretation

2016: member of the ANZPAA-NIFS working group on evaluative reporting.

2015 – 2016: Member of the US Scientific Working Group on DNA Analysis Methods (SWGDAM) group formed to evaluate Y-STR evidence.

2014 – present: Member of the Australian and New Zealand Statistical Scientific Working Group

2013 – present: Associate Professor at Flinders University in Biological Sciences

2012 – present: Ad hoc university student thesis examiner and journal article reviewer

2012 – 2014: Chair of the Australian and New Zealand Statistical Scientific Working Group

2010 – 2012: Vice Chair of the Australian and New Zealand Statistical Scientific Working Group – An international group of statistical experts tasked with developing and reviewing statistical methodologies to be used by Forensic Laboratories throughout Australia and New Zealand.

Awards and recognitions

2021 – Awarded the Public Service Medal in the Australia Day Honours January 26th 2021

2020 & 2021 – Identified by in the compilation of the World's Top 2% Scientists by Stanford University (DOI: 10.13140/RG.2.2.18594.45767)

2020 – Identified as the field leader in forensic science in Australia in '*the Australian: RESEARCH*' 2020, edited by Tim Dodd.

2018 – New Zealand Prime Minister's Science Award – Awarded to the STRmix™ team

2018 – Flinders University Distinguished Alumni Award

2017 – SA Science Excellence Award winner in STEM Professional Category 2017

2015 – KiwiNet Research Commercialisation Award – finalist in PwC Commercial Deal category

Publication awards from the National Institute of Forensic Science (NIFS)

NIFS Best paper award 2021 – Best Paper – Capability Enhancement and Innovation – "Validation of a top-down DNA profile analysis for database searching using a fully continuous probabilistic genotyping model"

NIFS Best paper award 2021 – Highly commended in Best Technical Article or Note – "Probabilistic interpretation of the Amelogenin locus"

NIFS Best paper award 2020 – Best Paper – Capability Enhancement and Innovation – "Using Bayesian networks to track DNA movement through complex transfer scenarios"

NIFS Best paper award 2020 – Best Paper – Forensic Fundamentals – "Investigating the position and level of DNA transfer to undergarments during digital sexual assault"

NIFS Best paper award 2020 – Best New Publisher in a Refereed Journal – "Examining the additivity of peak heights in forensic DNA profiles"

NIFS Best paper award 2019 – Best Technical Article or Note – "Inter-sample contamination detection using mixture deconvolution comparison"

NIFS Best paper award 2019 – Highly commended in Best paper in a refereed journal – "Likelihood ratio development for mixed Y-STR profiles"

NIFS Best paper award 2019 – Highly commended in Best literature review – "Evaluation of forensic genetics findings given activity level propositions: a review"

NIFS Best paper award 2018 – Best paper in a refereed journal – "Internal validation of STRmix™ – A multi laboratory response to PCAST"

NIFS Best paper award 2018 – Highly commended in Best paper in a refereed journal – "A template for constructing Bayesian networks in forensic biology cases when considering activity level propositions."

NIFS Best paper award 2018 – Highly commended in Best case study – "Likelihood ratio formulae for disputed parentage when the product of conception is trisomic"

NIFS Best paper award 2017 – Best paper in a refereed journal – "Teaching artificial intelligence to read electropherograms"

NIFS Best paper award 2017 – Best Technical Article or Note – "Observations of DNA transfer within an operational Forensic Biology Laboratory"

NIFS Best paper award 2015 – Best paper in a refereed journal – "Testing likelihood ratios produced from complex DNA profiles"

NIFS Best paper award 2015 – Highly commended in Best paper in a refereed journal – "Toward male individualization with rapidly mutating Y-chromosomal short tandem repeats"

NIFS Best paper award 2014 – Best paper in a refereed journal – "The interpretation of single source and mixed DNA profiles"

NIFS Best paper award 2014 – Highly commended in Best Technical Article or Note – "Searching mixed DNA profiles directly against profile databases"

Employment

I have been employed at Forensic Science SA since 2005 and currently hold the position of Chief Scientist in Forensic statistics within the biology group. I have appeared in court to present expert evidence on approximately 100 occasions in the Magistrates, District and

Supreme Courts in states around Australia. These include appearance as a prosecution witness and a defence witness.

I work on various criminal matters including sexual assaults, homicides, cold cases and coronial investigation, involving both Autosomal and Y-Chromosome STR data, and activity level evaluations. I have conducted familial searches for several matters and provided informational sessions for stakeholders. I provide activity level reports for prosecution and defence council around Australia and have provided reports for international innocence project matters. I have also carried out calculations for complex kinship scenarios. I have carried out DNA database analysis for various organisations and have analysed population datasets of DNA allele frequencies, generated by the forensic laboratories across Australia.

Within Forensic Science SA I developed methodology, validated, wrote standard operating procedures and implemented the following:

- Probabilistic Genotyping using STRmix™
- Standardised and semi-automated reporting of DNA results
- Familial searching
- Searching of mixed DNA profiles against the searchable DNA database
- Activity level evaluation and reporting
- Complex kinship calculation
- Mixture to mixture analyses and reporting