Second response to NISTIR 8351-DRAFT DNA Mixture Interpretation: A NIST Scientific Foundation Review

By the Institute of Environmental Science and Research Limited, New Zealand

8 November 2021

Summary: NIST Foundation Review - No problems found, no solutions offered.

It is worth clarifying the messaging of the draft NIST Foundation Review (NFR). NIST identify no error in any probabilistic genotyping software. They do not identify any unpublished limitation in any software, nor do they identify any deficiency in any validation.

They state that they cannot keep up with collation of the published literature and abandon this objective.

They table a suggestion to place partially processed data into the public domain to enable a desk audit against criteria that they do not specify.

They do not undertake to do the proposed audit and name no other body that has indicated a desire to do so.

In summary, NIST have identified no problems and offered no solutions.

Introduction

We have not found terms of reference for this review but NIST have stated that "In September 2016, both NCFS and PCAST requested that NIST examine the scientific literature and conduct technical merit evaluations and validation studies of forensic science methods and practices. The NCFS recommended that … "NIST's evaluation may include but is not limited to: a) research performed by other agencies and laboratories, b) its own intramural research program, or c) research studies documented in already published scientific literature."[1]

Submission: It is worth clarifying the messaging of the current draft NIST Foundation Review (NFR).

NIST do not identify any error in any software. No actual analysis has been undertaken by NIST that has uncovered any deficiency in any software.

NIST do not identify any published or unpublished limitation in any probabilistic genotyping (PG) software. Again, they have actually not undertaken any evaluation, hence they have not found anything either good or bad.

NIST do not identify any deficiency in any validation. As no evaluation is undertaken there is no finding.

NIST speculate on factors affecting reliability. Many of these seem reasonable but often, we believe, impact more on discrimination than reliability. We have no quantitative measure of

[1] https://nvlpubs.nist.gov/nistpubs/ir/2020/NIST.IR.8225.pdf accessed 2nd November 2021

reliability nor is one provided by NFR. This is pivotal. There is an insurmountable barrier to defining standards for validation until we know how to assess validity.

NIST suggest that validation should cover the range of samples likely to be encountered in casework but do not make any practical suggestion on density of coverage nor, short of redefining a fractional factorial design as bracketing, do they make any suggestion how the multidimensional volume is to be explored. The comments are self-contradictory in places, in some cases insisting on coverage and in others stating the obvious that dense coverage is impossible. There is an unevidenced but plausibly correct focus on number of contributors, template, mixture proportion, and allele sharing but no mention at all of triallelic patterns, non-resolution of peaks at capillary electrophoresis, and the shape of the tails of the distributions that determine the response to very bad PCR. The biggest single problem we encounter is input file errors and hence warning and safeguards here seem important.

NFR do not mention code quality, documentation of quality systems, nor audit and accreditation of programming activities. These are important aspects, we suggest, to reassure users.

They state that they cannot keep up with the collation of the published literature and abandon this objective. Again, this is pivotal. This is where the community have been disclosing material.

NFR table a suggestion to place partially processed data into the public domain to enable a desk audit against criteria that they do not specify (hereafter "The NFR hybrid"). They test the availability of data by, what has subsequently been shown to be, an ineffective internet search. They define the result as insufficient, but we would greatly value a statement of what would be sufficient. Only vague concepts are given of what to do with the output if sufficient data was available. They describe ROC plots but give no path from that, nor do we believe one exists, to any assessment of reliability. They very briefly mention calibration, whereas this does appear to have some hope of a path to assess reliability. We really need a much more practical and concrete path forward.

They do not undertake to do the proposed audit and name no other body that has indicated a desire to do so. Again, this is pivotal. The justice system will be left awaiting some analysis.

In this second submission we again concentrate on Key takeaway 4.3 which is the clause that raises the novel requirement. Key takeaway 4.4 also adds some detail to the NFR request for 'data' to be placed in the public domain, specifically adding that what data they have found they feel lacks detail.

KEY TAKEAWAY #4.3: "Currently, there is not enough publicly available data to enable an external and independent assessment of the degree of reliability of DNA mixture interpretation practices, including the use of probabilistic genotyping software (PGS) systems. To allow for external and independent assessments of reliability going forward, we encourage forensic laboratories to make their underlying PGS validation data publicly available and to regularly participate in interlaboratory studies."

The NIST Foundation Review concept encapsulated in Key Takeaway 4.3 is that developers or other groups should put large amounts of partially processed data into the public domain. Specifically, NIST ask for the data outlined in their box 4.1.

NFR declare that there are insufficient data, or insufficiently detailed data, available in the public domain to enable an external and independent review of PG systems. Additional data have been identified in the public domain (see Appendix 1). To our knowledge, neither the developers, the authors of any papers, nor any agencies were approached for access to their data. This was true when the draft was published and it is still true four months later.

STRmix™ is available to purchase by anyone who has undertaken training. This includes NIST who have had it since March 2014. This enables a much more complete and practical solution. Anyone wanting to test STRmix™ can simply perform any experiment they want and place the results in the public domain if they desire.

We have also offered to tailor experiments to NIST's desires. For example, in 2016 we asked John Butler and Eric Lander (PCAST) to specify what experiments they wanted but received no reply.

No "independent and external" organisation has asked for our data with the exceptions of Brooklyn Defender Services, New York and Forensic Aid, LLC. We have delivered the requested data to them but received no feedback nor have we seen any product of their investigations.

The NFR appeared June 2021. At the time of publication the suggestion to place large amounts of partially processed data in the public domain was additional, extending guidelines from SWGDAM [1], ISFG [2], IEEE [3], PCAST [4, 5], and the Forensic Science Regulator [6]. The NFR request for data is neither a Daubert not a Frye criterion.

Dr Butler is a signatory to the ISFG guidelines and is quoted as agreeing with PCAST. The NFR data sharing suggestion was not mentioned in either of these documents and we assume that Dr Butler has extended his thinking to include the postulated desk audit. However, this cannot be considered pivotal to an assessment of reliability. It is one possible suggestion amongst many that are possible and as yet, we have no agreed plan for how to turn these data into information about reliability.

In our own experience we can often identify when an answer is wrong. This is achieved in two ways:

1. Parallel calculation of an answer from the models, or;

2. Comparing the answer against subjective expectation. This is often started by looking at those data that are false exclusions or show high adventitious support. If the *LR* is much lower than expected from the ground truth status and template it is plausible that something is wrong. Examples of this appear in the paper by Cheng et al. [7]. We are not the only people who can do this. Most referrals from laboratories about anomalous results stem from them applying the same approach.

If we, and others, can define certain results as unsuitable given the inputs we feel it must be possible to define, in some sense, at least a range of answers that are not wrong. Some work

in this area could eventually lead to improvement in our concept of validity. Some very good progress was made working with Drs Peter Gill and Oyvind Bleka on a comparison of EuroForMix and STRmix™ [7]. This progress was made by detailed examination of the cause of unusual results.

The NFR request for data sharing represents an abrupt change of direction when compared with PCAST or ISFG. PCAST encouraged the community to publish more empirical work in the peer reviewed literature. NFR bypasses the peer reviewed publication step and asks for partially processed data. We are uncertain how partially processed data can be considered "external and independent" using the definition from NFR.

The NFR request appears intended to permit a summary audit from the desk of the auditor. It is potentially possible for us and the community to achieve NIST's expectations if we can focus them. We have already placed a large amount of data in the public domain (29 July 2021)[2]. Approaching us and others during the tenure of this project would have allowed us to provide the data to the public domain and NIST during their work period and this may have greatly increased the value of the NFR.

The data we have placed in the public domain exceeds 8,000 true donor tests and 128 million false donor tests. Additionally, over 60 laboratories have completed internal validation studies with the PG software STRmix™ from which data could have been requested to be considered within this foundational review. It may be that some of these could also have been placed in the public domain. Appendix 1 gives, what we think are, seven additional internal validation documents that were available in the public domain at the time of release of the draft NFR. To our examination these give extensive high-quality data.

We reiterate our willingness, previously expressed, to work constructively with organisations wishing to test STRmix™, including NIST. We will endeavour to increase the amount of data placed in the public domain for our research projects in the future. This placement of data has proven quite an unrewarding activity to date with resources applied and no usable feedback received.

We invite feedback from NIST on these data. The time since publication of our data is now about 3 months which, we feel, is plenty of time to have undertaken an analysis. May we appeal for some constructive interaction on this subject.

We have no indication that NIST intend to do anything with these or other data themselves. A letter asking NIST management to outline their intentions was peremptorily returned with the statement that we should submit it to the reopened comment period (see Appendix 2). This submission to the second comment period will not be timely for the multiple admissibility hearings that are quoting the NFR and are proceeding at this time and we appeal to NIST to be constructive in assisting courts with timely data.

We have suggested that NIST make mixtures and we would run them and hand the results back to NIST. This could have been completed by now.

[2] https://figshare.com/articles/dataset/ESR_response_to_NISTIR_8351_-_DRAFT_DNA_Mixture_Interpretation_A_NIST_Scientific_Foundation_Review/15062907

It is important that independence is not substituted for competence. We are concerned that NIST has a view that any conversation with us compromises their independence in some way. Any trained scientist, whether at NIST or in the various laboratories in the US or worldwide, is capable of assessing the value of information received. NIST have published three papers where they have used PG software [8-10]. We have investigated their work in detail. There are multiple technical concerns the largest of which was leaving in the input file artefacts that the version of EuroForMix used was not designed to handle. They have additionally used an unvalidated software, CleanIt, that appears to remove peaks that should be retained.

The primary method of analysis of empirical data given in the NFR are Receiver Operating Characteristic (ROC) curves. ROC curves quantify the discriminatory power of a continuous marker to predict a binary outcome. They are very ill suited to the task of assessing PG output. Consider the sets of *LR* curves in figures 1a-c.

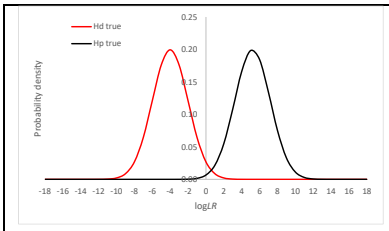

Figure 1a. A well calibrated *LR*

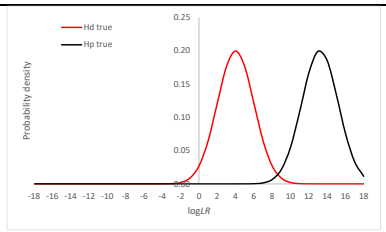

Figure 1b. A poorly calibrated *LR*

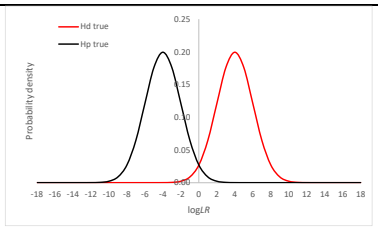

Figure 1c. A "reversed" set of *LR*s

The *LR*s shown in Figures 1a. through 1c. would have the same ROC plot (the reversed plot requires inversion of the classification parameter). ROC plots therefore do not inform on accuracy but do inform on discrimination. A referee did suggest we could adorn the ROC curves with multiple tags of *LR* values to recover the information lost in the process of making the ROC plot. Even with this, and a now overly cluttered figure, we still only comment on discrimination and not accuracy or reliability.

However, in an attempt to be constructive, we have developed some ROC curves from one of our biggest datasets. This appears here[3]. We have also attempted calibration here[4] and in this paper [11]. All of these were in the public domain during the tenure of the NFR. Feedback on these extensive efforts by us from NIST would be most welcome.

Key Takeaway 4.4 specifies the details of the data that NIST desire. In our first submission we mentioned some concerns about what was asked and what was omitted. We are not qualified to undertake a legal analysis of the disclosure of genetic data. However in the absence of any lead from NIST we note that:

3 https://research.esr.cri.nz/articles/report/The_discriminatory_power_of_STRmix_illustrated_by_ROC_curves/11833524

4 https://research.esr.cri.nz/articles/report/Calibration_of_STRmix_LRs_following_the_method_of_Hannig_et_al_/12324011

1. The United Nations Universal Declaration on the Human Genome and Human Rights[5] outlines a number of guidelines that appear to impact on the disclosure of genetic data both encouraging dissemination but suggesting strong safeguards such as informed consent.

2. The National Human Genome Research Institute webpage states[6]: "Federal laws like the Common Rule and the Health Insurance Portability and Accountability Act (HIPAA) aim to balance efforts to promote scientific progress and protect patient privacy. This is challenging for genomic data because, with the exception of identical twins, each person's DNA sequence is unique, which means a DNA sample can never be truly anonymized.

"… a study published in 2013 shows that research participants can be re-identified using genomic data from one such database paired with genealogical databases and public records."

It is not possible to treat the matter of disclosure of genotypes from a scientific desirability view in isolation of considering the wider ethical issues. Whilst at some future time we may be in a position to disclose some genetic data we are a long way away from having the ethical and legal framework in place at writing.

NFR state at line 532 "The findings described in this report are meant solely to inform future work in the field." However, it was inevitable that this report would be used in legal proceedings from the time a draft was first tabled. Some are proceeding at this time. The difficulty is compounded by the fact that NIST are unresponsive to direct questions (see Appendix 2). We therefore request NIST to take an open, constructive, and responsible approach. This involves:

1. Cognisance that vague, unevidenced or misevidenced concerns published by NIST may immediately be used in court, and
2. a timely response and feedback with respect to the data made available in response to requests. Feedback before the publication of the final report would allow us to respond to any amendments NIST desire, and
3. a more constructive approach to obtaining and sharing data going forward, and
4. practically implementable suggestions preferably tested in advance by NIST.

We are unable at this stage to discern what NFR would wish done beyond broad discussions of ROC plots, a very brief discussion of calibration, and contradictory comments regarding coverage. We appeal for a constructive conversation, preferably a detailed joint analysis of our data, designed to meet NIST's needs.

Have NIST met NCFS's and PCAST's requirements?

These were: "In September 2016, both NCFS and PCAST requested that NIST examine the scientific literature and conduct technical merit evaluations and validation studies of forensic science methods and practices. The NCFS recommended that … "NIST's evaluation may

[5] https://www.ohchr.org/en/professionalinterest/pages/humangenomeandhumanrights.aspx

[6] https://www.genome.gov/about-genomics/policy-issues/Privacy#research

include but is not limited to: a) research performed by other agencies and laboratories, b) its own intramural research program, or c) research studies documented in already published scientific literature."[7] We note that no technical merit review is reported and no validation studies were performed (except maybe Riman et al. [8], although that does appear to be a separate project). NIST have certainly not exhausted the data options listed by NCFS.

---

[7] https://nvlpubs.nist.gov/nistpubs/ir/2020/NIST.IR.8225.pdf accessed 2nd November 2021

Appendix 1
Internal validation data identified by internet search. If they are listed side by side they may be the same document.

NIST	Brooklyn defender's https://indefenseof.us/issues/kinship-problem
Erie County Central Police Services Forensic Laboratory (Buffalo, NY) STRmix v2.3 (PowerPlex Fusion, ABI 3500) https://johnbuckleton.files.wordpress.com/2016/09/strmix-implementationand-internal-validation-erie-fusion.pdf STRmix v2.3 (Identifiler Plus, ABI 3500) https://johnbuckleton.files.wordpress.com/2016/09/strmix-implementationand-internal-validation-erie-id-plus.pdf	
Michigan State Police (Lansing, MI) STRmix v2.3.07 (PowerPlex Fusion, ABI 3500/3500xl) https://johnbuckleton.files.wordpress.com/2016/09/strmix-summarymsp.pdf	
Office of Chief Medical Examiner Forensic Biology Laboratory (New York City, NY) STRmix v2.4 (PowerPlex Fusion, ABI 3130xl) https://www1.nyc.gov/site/ocme/services/validation-summary.page	
Palm Beach County Sheriff's Office (West Palm Beach, FL) STRmix v2.4.06 (PowerPlex Fusion, ABI 3500xl) http://www.pbso.org/qualtrax/QTDocuments/4228.PDF STRmix v2.6.2 (PowerPlex Fusion 6C, ABI 3500xl) https://www.pbso.org/qualtrax/QTDocuments/10787.PDF	Palm Beach County Sheriff's Office (PBSO) Laboratory - Internal Validation of STRmix v. 2.4 (FusionTM 5C)
San Diego Police Department Crime Laboratory (San Diego, CA) STRmix (GlobalFiler, ABI 3500), STRmix v2.3.07; STRmix v2.4.06 https://www.sandiego.gov/police/services/crime-laboratory-documents	
Virginia Department of Forensic	

Science (Richmond, VA)* TrueAllele Casework (PowerPlex 16, ABI 3130xl) https://epic.org/state-policy/foia/dna-software/EPIC-15-10-13-VA-FOIA-20151104-Production-Pt2.pdf	
Department of Forensic Sciences (Washington, DC) STRmix v2.3 parameters & validation report (Identifiler Plus, ABI 3500) https://dfs.dc.gov/page/fbu-validation-studiesperformance-checks STRmix v2.4 parameters & validation report (GlobalFiler, ABI 3500) https://dfs.dc.gov/page/fbu-validation-studiesperformance-checks	
	Los Angeles County Sheriff's Department, Scientific Services Bureau Biology Section - Validation of STRmixTM v. 2.5.11 using the Powerplex Fusion 6C Kit
	Jefferson County Regional Crime Laboratory - Internal Validation of STRmix™ v. 2.6 for the Analysis of GlobalFiler™ Profiles
	• Sacramento County District Attorney's Crime Laboratory - Internal Validation of STRmix™ v. 2.4
	Las Vegas Metropolitan Police Department - Internal Validation of STRmix™ v2.6
	Colorado Bureau of Investigation - Internal Validation of STRmix™ v. 2.5 for the CBI Forensic Laboratories
	••• Wisconsin State Crime Laboratory - Internal Validation Summary for STRmix™ Probabilistic Genotyping Software
	Oregon State Police, Forensic Services Division, Portland Metro Laboratory - Validation Study for STR Analysis Volume 67—2016

	Validation – STR Casework Analysis using GlobalFiler, the 3500xl, and STRmix

Appendix 2

19 October 2021

James K. Olthoff
Acting Director, National Institute of Standards and Technology
U.S. Department of Commerce
100 Bureau Drive
Gaithersburg, MD 20899
USA

By email: james.olthoff@nist.gov

Dear Dr. Olthoff,

Draft NIST Report- "DNA Mixture Interpretation: A NIST Scientific Foundation Review"

STRmix™ is a joint venture between the Governments of South Australia and New Zealand. It is in active use for the interpretation of DNA evidence for evidential purposes in about 52% of accredited US laboratories with a further 28% testing or implementing it.

We read the recent Draft NIST Report authored by Butler et al. with great concern. We found Chapter 4, Reliability of DNA Mixtures, Measurements and Interpretation particularly significant. Of serious concern was Key Takeaway 4.3 which states: "Currently, there is not enough publicly available data to enable an external and independent assessment of the degree of reliability of DNA mixture interpretation practices, including the use of probabilistic genotyping software (PGS) systems…" [Draft NIST Report, p. 75].

There are now a large number of admissibility hearings proceeding in the US that quote the Draft NIST Report as the reason that has been advanced for non-admission of PG evidence. This, obviously, has significant implications for the justice system in the US and represents a cost to us and many laboratories and District Attorneys.

In addition, a resolution was recently tabled at the New York State Forensic Science Commission calling for a moratorium on DNA testing pending the release of the finalised Report. This resolution was not seconded and, at writing, is not proceeding, but may do so after the final Report is published if the conclusions do not substantively change.

The key conclusion in the Draft NIST Report is that there is not enough publicly available data to enable an external and independent assessment of the degree of reliability of DNA mixture interpretation practices. As a response we have placed a large amount of data into the public

INSTITUTE OF ENVIRONMENTAL SCIENCE AND RESEARCH LIMITED
Mt Albert Science Centre: 120 Mt Albert Road, Sandringham, Auckland 1025 | Private Bag 92021, Auckland 1142, New Zealand
T: +64 9 815 3670 **F:** +64 9 849 6046
www.esr.cri.nz

domain[8]. They are not the type of data usually published and this is a new requirement added by the authors of the report. Further, the authors missed a lot of data in their internet search and realistically could have made much more effort to obtain data by contacting laboratories or us directly. We are aware, for example, that they could have inspected the FBI data at Quantico and this is the single most extensive dataset. More recently, NIST also has received these data at Gaithersburg and are repeating some of the interpretations.

We feel that NIST has made an observation with no clear way forward. Does NIST have a view which party is responsible to evaluate the data to verify the laboratory/ software claims beyond their validation obligations and hence reassure the justice system as to the reliability of PG software? Alternatively, is NIST planning to do this data analysis as part of its mission to advance measurement science, standards, and technology, to provide confidence to the community?

We urgently ask you to confirm whether or not NIST has any plans to do analysis on these data and if so whether there is a time frame for completion.

Further we urgently ask you to confirm whether or not NIST knows of any other organization that is planning to do NIST approved analysis, and if so what time frame is planned for that.

Kind regards

John Bone

General Manager STRmix Limited

---

[8] https://figshare.com/articles/dataset/ESR_response_to_NISTIR_8351_-_DRAFT_DNA_Mixture_Interpretation_A_NIST_Scientific_Foundation_Review/15062907

john.bone@esr.cri.nz

www.strmix.com

From: "Shyam-Sunder, Sivaraj (Fed)" <sivaraj.shyam-sunder@nist.gov>
Date: 23 October 2021 at 4:04:20 AM NZDT
To: John Bone <john.bone@esr.cri.nz>
Subject: FW: NIST DNA Mixture Report

Dear Mr. Bone,

Thank you for your letter to Dr. James Olthoff regarding the draft NIST DNA mixture report. NIST has re-opened the public comment period until November 19, 2021 to receive additional comments, new data, or information. You may submit your letter as well any other information for consideration by NIST in accordance with the process specified in the attached NIST announcement. Thank you.

Best regards,

Shyam

==========
Dr. S. Shyam Sunder
Director, Special Programs Office
and Chief Data Officer
National Institute of Standards and Technology
U.S. Department of Commerce
301-975-6713 (w) 301-943-4934 (m)

References

1. Scientific Working Group on DNA Analysis Methods (Swgdam). Guidelines for the Validation of Probabilistic Genotyping Systems. 2015 3 October 2016]; Available from: http://media.wix.com/ugd/4344b0_22776006b67c4a32a5ffc04fe3b56515.pdf.
2. 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, and 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, 2016, 25: 191-197.
3. Software and Systems Engineering Standards Committee of the Ieee Computer Society, IEEE Standard for System, Software, and Hardware Verification and Validation IEEE Std 1012™-2016. 2017: IEEE, 3 Park Avenue, New York, NY 10016-5997, USA.
4. President's Council of Advisors on Science and Technology. Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods. 2016 22 April 2017]; Available from: https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/PCAST/pcast_forensic_science_report_final.pdf.
5. President's Council of Advisors on Science and Technology. An addendum to the PCAST report on forensic science in criminal courts. 2016 20 July 2017]; Available from: https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/PCAST/pcast_forensics_addendum_finalv2.pdf.
6. Forensic Science Regulator. Software Validation for DNA Mixture Interpretation, FSR-G-223 Issue 1. 2018 11 September 2018]; Available from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/730994/G223_Mixture_software_validation_Issue1.pdf.
7. K. Cheng, O. Bleka, P. Gill , J. Curran, J.-A. Bright, D. Taylor, and J. Buckleton, A comparison of likelihood ratios obtained from EuroForMix and STRmix™ Journal of Forensic Sciences, , 2021, 66 (6): 2138-2155.
8. S. Riman, H. Iyer, and P.M. Vallone, Examining performance and likelihood ratios for two likelihood ratio systems using the PROVEDIt dataset. PLoS ONE, 2021, 16(9): e0256714.
9. S. Riman, H. Iyer, and P.M. Vallone, Exploring DNA interpretation software using the PROVEDIt dataset. Forensic Science International: Genetics Supplement Series, 2019, 7(1): 724-726.
10. J. Hannig, S. Riman, H. Iyer, and P.M. Vallone, Are reported likelihood ratios well calibrated? Forensic Science International: Genetics Supplement Series, 2019, 7(1): 572-574.
11. J.-A. Bright, M. Jones Dukes, S.N. Pugh, I.W. Evett, and J.S. Buckleton, Applying calibration to LRs produced by a DNA interpretation software. Australian Journal of Forensic Sciences, 2019: 1-7.