Public comments received

on NISTIR 8351-DRAFT

DNA Mixture Interpretation: *A NIST Scientific Foundation Review*

Published December 3, 2021

NISTIR 8351-DRAFT: DNA Mixture Interpretation: *A NIST Scientific Foundation Review* was released for public comment on June 9, 2021. That draft document is available at: https://nvlpubs.nist.gov/nistpubs/ir/2021/NIST.IR.8351-draft.pdf.

Public comment periods were held from June 9, 2021 to August 23, 2021 and from October 22, 2021 to November 19, 2021. This document lists all 63 public comments (PC1 to PC63) in the chronological order in which they were received. When an attachment was provided with an email (e.g., PC2), then this material is labeled with an "a" (i.e., PC2a). A bookmark has been placed with each PC to help index the file. *Submitter email addresses and phone numbers have been redacted.* Note that there are several attachments available as part of PC12.

NIST hosted a webinar on July 21, 2021 to review the contents of our draft report and address questions (see https://www.nist.gov/news-events/events/2021/07/webinar-dna-mixtures-nist-scientific-foundation-review). The 83 questions/comments provided during the Q&A portion of this webinar are listed as W1 after PC6.

This PDF file was originally made available for download from https://www.nist.gov/dna-mixture-interpretation-nist-scientific-foundation-review.

National Institute of Standards and Technology
U.S. Department of Commerce

PC #	Commenter Name and Affiliation
PC1	**Daniel Myers** (New York State Police Forensic Investigation Center)
PC2	**Peter Gill** (University of Oslo, Norway)
PC3	**Norah Rudin** (forensic DNA consultant)
PC4	**Timothy Goble** (New York State Police Forensic Investigation Center)
PC5	**Mark Timken** (retired from California Department of Justice DNA Laboratory, Richmond)
PC6	**Meegan Fitzpatrick** (New York State Police Forensic Investigation Center)
W1	**WEBINAR Q&A** *(83 questions or comments received)*
PC7	**Lucy A. Davis** (LDH Consultants, LLC)
PC8	**Meegan Fitzpatrick** (New York State Police Forensic Investigation Center)
PC9	**Peter Gill** (University of Oslo, Norway)
PC10	**Bjorn Sutherland** (ESR/STRmix, New Zealand)
PC10′	**Bjorn Sutherland** (ESR/STRmix, New Zealand)
PC11	**Jeanette Kovari** (New York State Police Forensic Investigation Center)
PC12	**Jessica Charak Lehrner** (Las Vegas Metropolitan Police Department Biology/DNA Detail)
PC13	**Jennifer Thayer** (New Jersey State Police Office of Forensic Sciences)
PC14	**Amy McGuckian** (Palm Beach County Sheriff's Office)
PC15	**Greg Hampikian** (Boise State University)
PC16	**Dawn Moore Boswell** (University of North Texas Health Science Center, Center for Human Identification)
PC17	**Alex Biedermann** (University of Lausanne, Switzerland)
PC18	**Colleen Spurgeon & Jeanette Wallin** (California Department of Justice, Jan Bashinski Laboratory)
PC19	**Christina De La O** (no affiliation provided)
PC20	**Mark Perlin** (Cybergenetics/TrueAllele)
PC21	**Bruce Budowle, Michael Coble** (University of North Texas Health Science Center), **Fred R. Bieber** (Harvard Medical School)
PC22	**Nelson Bunn** (National District Attorneys Association)
PC23	**Erin P. Forry** (American Society of Crime Laboratory Directors, ASCLD Board of Directors)
PC24	**Jack Laird** (Centre of Forensic Sciences, Ontario, Canada)
PC25	**Kim E. Mooney** (Defense Forensic Science Center)
PC26	**Joel Sutton** (Defense Forensic Science Center; member of the Resource Group)
PC27	**Ray Wickenheiser** (New York State Police Forensic Investigation Center; member of the Resource Group)
PC28	**Timothy D. Kupferschmid**, **Craig O'Connor**, **Eugene Y. Lien** (New York City Office of Chief Medical Examiner; Eugene Lien served as a member of the Resource Group)
PC29	**Clark Jaw** (Washington DC Department of Forensic Sciences)
PC30	**Tiffany Roy** (written by Tacha Hicks on behalf of the Human Factors in Forensic DNA Analysis Testimony and Reporting Subgroup)

PC #	Commenter Name and Affiliation
PC31	**Roberto Puch-Solis** (Leverhulme Research Centre for Forensic Science, University of Dundee, UK)
PC32	**Lloyd Halsell III** (Houston Forensic Science Center)
PC33	**Gretchen DeGroot** & **Amber Rasmussen** (Wisconsin Division of Forensic Sciences)
PC34	**Beth Hewitt** (Jefferson County Regional Crime Laboratory)
PC35	**Jack Ballantyne** (National Center for Forensic Science, University of Central Florida; member of the Resource Group)
PC36	**Elizabeth Daniel Vasquez & Clinton Hughes** (Brooklyn Defender Services)
PC37	**Brad Jenkins** (Virginia Department of Forensic Sciences)
PC38	**Lauren Lu** (Michigan State Police, Forensic Science Division)
PC39	**Robert Griffith (**Miami-Dade Police Department, Forensic Services Bureau - Forensic Biology Section)
PC40	**Terri Rosenblatt** (The Legal Aid Society, DNA Unit)
PC41	**Mackenzie DeHaan** (North Carolina Department of Justice Crime Laboratory, Forensic Biology Section)
PC42	**Mary Lou Nicholson** (Royal Canadian Mounted Police)
PC43	**Cecilia von Beroldingen** (UC Davis Forensic Science Graduate Program)
PC44	**Ann Gross** (Minnesota Bureau of Criminal Apprehension Forensic Science Services Laboratory)
PC45	**Sarah Chu** (Innocence Project)
PC46	**Tiffany Roy** (forensic DNA consultant)
PC47	**Robyn Ragsdale** (Florida Department of Law Enforcement; on behalf of the OSAC Human Forensic Biology Subcommittee)
PC48	**Erin Murphy** (New York University School of Law; on behalf of a group of scientists and legal scholars who study the use and impact of forensic evidence)
PC49	**Tim Kalafut** (Sam Houston State University) and **Simone Gittelson** (Washington DC Department of Forensic Sciences)
PC50	**Antonio Onorato** (Scientific Working Group on DNA Analysis Methods, SWGDAM)
PC51	**Geoffrey Stewart Morrison** (forensic consultant)
PC52	**Ray Wickenheiser** (New York State Police Forensic Investigation Center; member of the Resource Group) **(2nd submission)**
PC53	**Bjorn Sutherland** (ESR/STRmix, New Zealand) **(2nd submission)**
PC54	**Raymond Valerio** (Queens County District Attorney's Office)
PC55	**Dawn Herkenham** (SWGDAM) **(2nd submission)**
PC56	**Jarrah Kennedy** (Kansas City Police Crime Laboratory)
PC57	**Thomas Grill** (DNA Technical Leader, Erie County Central Police Services Forensic Laboratory)
PC58	**Dennis McNevin** (University of Technology Sydney, Australia)

PC #	Commenter Name and Affiliation
PC59	**Jeanette Wallin** & **Colleen Spurgeon** (California Department of Justice, Jan Bashinski Laboratory) **(2nd submission)**
PC60	**Erica Wissolik** (Government Relations, Institute of Electrical and Electronics Engineers IEEE-USA)
PC61	**Elizabeth Daniel Vasquez & Clinton Hughes** (Brooklyn Defender Services) **(2nd submission)**
PC62	**Laura Sudkamp** (American Society of Crime Laboratory Directors) **(2nd submission)**
PC63	**Matthew Gamette** (Idaho State Police Forensic Services)

PC1

Comments on Chapter 2 of DNA Mixture Interpretation: A Scientific Foundation Review

MYERS, DANIEL (TROOPERS) [redacted]

Mon 6/28/2021 1:49 PM

To: ScientificFoundationReviews <ScientificFoundationReviews@nist.gov>

Chapter 2 reads very well, is understandable and the sections on likelihood ratio framework and transposed conditional are great. Analysts could always use examples and explanation of the transposed conditional, in not just likelihood ratios but match probabilities as well. I wouldn't argue if more were added.

Something that might be helpful, is further delineation between comparison and statistic at the beginning of section 2.2. The distinction is made clear on page 29 where the ISFG recommendations include comparison as an independent step before statistical analysis. A nuance that is often overlooked (especially by those newer to the community) is that statistics are necessarily dependent on the results of an analysis and comparison, rather than being or determining the results of an analysis and comparison.

Very good chapter.

Dan

PC2

mixtures review

Peter Gill [redacted]

Fri 7/9/2021 11:08 AM

To: ScientificFoundationReviews <ScientificFoundationReviews@nist.gov>

Please see my comments in relation to the foundation review on mixtures interpretation.

Best wishes

Peter

PC2a

Notes on NIST foundation review

Peter Gill

Congratulations to the authors on your comprehensive review. I can see that this has been a mammoth task for all those involved. Thankyou for the opportunity to input. I offer some thoughts below:

Key Takeaway #4.3: I agree that the lack of published data is a short-coming in this and most other areas of forensic genetics. This restricts the peer review process. The unavailability of data means that detailed checks are not possible. The way to deal with this is for journals such as FSIgen to be much more pro-active in insisting upon data publication as a precondition of acceptance (I completely agree with lines 1150-1154).

Line 1209: We adopt exploratory analysis where conditions such as the numbers of contributors are varied to see the effect upon the LR and the model parameters. In my experience, numbers of contributors is not a serious issue: for Hp=true samples, the LR tends to plateau as the number of contributors increases. A 'best-fit' model can be adopted (e.g. my book table 8.7) by choosing the model that minimises the log likelihood. Secondly, the model can be tested to ensure that it behaves in accordance with theoretical expectations (my book fig 8.21). Sometimes, the model may fail this test hence the analysis may be rejected.

Line 1635: Generally, decisions are not binary – a peak may consist of an allele plus background noise (that increases its size); a peak may be stutter and allele at the same time (PG can be used to evaluate the latter; setting a suitable AT will capture the former at the expense of discounting true allelic peaks).

Line 1877: Key takeway #2.5 "Continuous probabilistic genotyping software (PGS) methods utilize more information from a DNA profile than binary [and discrete] approaches (this would be consistent with principle 14).

Note that we should not be shy about advocating moving from discrete to continuous - the continuous methods have many advantages, hence we strongly encourage labs to move to the next stage, with suitable training of course.

Table 2.4: re untested brother – do you mean that the propositions are Hp: POI+unknown unrelated contributor vs Hd: an unknown sibling + unknown unrelated contributor? Clarify in the legend as it is not clear what you mean by "a possible untested brother….assumptions made" – which assumptions?

Line 2123: Many describe the LR as a "personal belief" that is conditioned upon many factors as described in the takeaway 26 box. Probabilities and LRs do not exist in nature, only states – i.e something is either true or false.

Line 2247: I am not sure what this means – are the two comparable? or why should they be compared? Peak heights are affected by stochastic effects. I think you mean "Peak positions can be used to accurately designate alleles whereas peak heights are subject to stochastic effects and are

variable" – and this leads to principle 10 – so I am not sure why you need to mention this in principle 9 as well.

Line 2350 in favor [of] s…

Line 2352: propositions are not exhaustive (i.e. we never consider the universe of possible numbers of contributors etc). Non exhaustive propositions are not misleading: https://doi.org/10.1016/j.fsigen.2020.102406. It may be that there are several sets of propositions to consider and each will give a different LR (my book section 6.2.5). Recently Hicks et al https://doi.org/10.1016/j.fsigen.2021.102481 published a method to combine different propositions so that they are effectively exhaustive – but this method is not widely used since it is only recently described.

Line 2402: Yes completely agree – such data can easily be made available electronically either on dedicated internet sites or as electronic supplements of publications – need to pressure journals to insist on this as a condition of publication. However, this is an issue not restricted to forensic genetics:

From: https://doi.org/10.1098/rsos.150547 (citing The Committee on Responsibilities of Authorship in the Biological Sciences, National Research Council) – also see discussion in this paper:

"[T]he act of publishing is a quid pro quo in which authors receive credit and acknowledgment in exchange for disclosure of their scientific findings. An author's obligation is not only to release data and materials to enable others to verify or replicate published findings (as journals already implicitly or explicitly require) but also to provide them in a form on which other scientists can build with further research. All members of the scientific community—whether working in academia, government, or a commercial enterprise—have equal responsibility for upholding community standards as participants in the publication system, and all should be equally able to derive benefits from it."

Line 2468: Consider the role of free pre-print servers like https://www.biorxiv.org/. This would be an effective way to distribute 'unpublishable material'. Widely used for this purpose in academia.

Line 3372: Allele sharing: False positive results may occur with related individuals, where the propositions are set under the assumption of unrelatedness.

Consequently, it is important that propositions are formulated to include assumptions of relatedness provided that the case circumstances dictate this.

Line 3425: Difficult for me to understand what is meant by takeaway 4.4: "Even when a comparable reliability can be assessed (results for a two-person mixed sample are generally expected to be more reliable than those for a four-person mixed sample, for example), there is no threshold or criteria established to determine what is an acceptable level of reliability."

This statement implies that there is some underlying 'true likelihood ratio' (defined earlier as "a result that is consistently accurate"), but at the same time we accept that the LR does not exist in nature (line 2777) and this generates a paradox in your thesis. Furthermore, there is no uncertainty associated with an LR assessment (line 2779) since it is regarded as "personal belief". To assist, The ENFSI guideline for evaluative reporting (ENFSI,

2015), page 16 recommends: “Such personal probability assignment is not arbitrary or speculative, but is based on a body of knowledge that should be available for auditing and disclosure.” In this context, we ideally wish to see if the personal beliefs of different scientists coincide when different methods are used. This can be ascertained by carrying out comparative studies, and this is certainly something to be encouraged. It is inevitable that different models will give different answers – the more different the models, the more extreme the differences. Take the example of discrete vs continuous model comparisons. The differences will be more divergent compared to comparisons of 2 continuous models (this is well published), but how can you define an "accepted level of reliability"? Each model will give a "correct" or the "right" answer within the context of assumptions and conditioning of the model. There is no underlying "true" LR, to compare the answer with, hence how can we ascribe one to be more reliable than another? Indeed, when models are based upon different assumptions, then there must be two different sets of criteria to test reliability (because there is no universal true LR to test against). Therefore, because of the paradox, the issue of reliability, seems difficult to resolve. Systems are often tested using ground truth Hp-true and Hd-true tests (lines 3576-82, ROC plots are also useful to assess the characteristics of the system), but this gives us information about the characteristics of systems in general and enables comparative studies. In case-work we know nothing about ground truth. To measure 'reliability' on a case by case basis, non contributor tests have been introduced, such as Turing tests (see section 12.17 of my book for discussion). See **https://doi.org/10.1002/wfs2.1321** for a discussion about developing scientific consensus about the use of different methods based on knowledge basis.

Line 3508: "There appears to be a general misconception that LR assessments made by different experts will be close enough to one another to not materially affect the outcome of a case. Although they may be close enough in many instances, this is not known for any particular case and it is not advisable to take this for granted."

This is an important point. The court process is usually dominated by the prosecution expert. Indeed the latter often formulates the defence propositions. However, it is essential at the court stage that the environment is conducive to allow exploration of the evidence (see my points made at the end of this submission – discussed in **https://doi.org/10.1002/wfs2.1321**). Defence experts need to be trained to the same level, and have access to the same or alternative software in order to carry out separate evaluations. In the UK jurisdiction, defence and prosecution experts are encouraged to collaborate to write joint statements that clearly describe where there is agreement, so that the court does not have to consider very complex issues of evidence (also see lines 3552-6). Therefore I argue that reliability is achieved when the court procedures are conducive to discovery and prosecution and defence experts collaborate towards a common purpose.

Line 4569: Definition of contamination is not quite right. You state: " consist of stray alleles arising from unknown sources or profiles or alleles from persons handling the items" this should state: "consists of DNA from investigators and scientists and other personnel at the crime scene and/or laboratory" i.e. unknown persons who handled the evidence (innocently) before the crime was committed do not contaminate the evidence, because this is part of the natural environment of the scene.

Also, it is necessary to distinguish between contamination and drop-in (also see line 5852). Drop-in is characterised from negative controls to measure the prevalence and can be accommodated probabilistically by PG models. Drop-in events are independent (alleles from different contributors). Contamination events are dependent, leading to multiple alleles from a single contributor – also dealt with by PG simply by conditioning an unknown contributor. See section 4.2.1 of my book

Line 4827: Source level is the body fluid (is it blood or saliva etc) – relatives would be assessed in sub-source propositions. Source is (mistakenly) substituted for sub-source elsewhere (line 4748, 4713).

Line 4861: I would add to takeaway 5.5:

"The fact that DNA transfers easily between objects does not negate the value of DNA evidence [at the sub-source level. However, this value cannot be carried over to other levels in in the hierarchy of propositions – i.e. source and activity levels, because different likelihood ratio calculations are needed for each. The value of DNA evidence at source and activity levels depends on the circumstances of the case.]"

Note that all statements need a caveat explaining the limitations of the LR assessment at sub-source level – a recommendation is needed to explain. From ISFG DNA Commission https://doi.org/10.1016/j.fsigen.2019.102186: "This report does not provide any information on the mechanisms or actions that led to the deposition of the biological material concerned. It only provides help regarding its origin. Should there be any issue regarding the activities that led to this material, we should be contacted to provide a new report."

General: Activity level assessments are difficult because of lack of consistency of experimental regimes. Insufficient work on reproducibility. Lack of suitable software to carry out calculations requiring Bayesian Networks. Much more effort, training of scientists, lawyers and judges in needed. For a salutary lesson, in a case where a consideration of activity level exonerated a convicted suspect see: Jones v R. [2020] EWCA Crim 1021. Retrieved from: https://www.bailii.org/ew/cases/EWCA/Crim/2020/1021.html, also described in supplement 6 of: https://www.sciencedirect.com/science/article/pii/S1872497321000478

Appendix A1.8 Thoughts on training and court environment:

It is not just the matter of educating the scientist who works for the prosecution. The likelihood ratio framework provides the means for constructive debate where the court is a forum for the exchange of views and the environment must be conducive for this to occur. This is where issues of reliability may be discussed by constructive dialogue and PG testing of samples using agreed sets of propositions. Different software may be used to evaluate evidence of a given case.

The court needs to foster an environment where complex issues may be discussed between experts. But the adversarial system does not encourage dialogue because it is combative in nature rather than collaborative (this is especially true in US courts). In addition, the reliance upon one or two experts in court does not mean that a scientific consensus is represented.

For example, the formulation of propositions may not be clear-cut in many cases. A dialogue between scientists that represent the defence and prosecution to agree those to test. The software is used to evaluate the evidence against competing sets of propositions that are agreed beforehand.

To facilitate, there is need for equivalent access for tools (databases, probabilistic models) for both defence and prosecution scientists. Access to software and open data are important in this respect because of the increased transparency and unlimited availability (note that it is no problem for experts to use different software solutions, this approach may be preferable since there is opportunity to check results using two different methods – which helps with reliability). More knowledge bases are needed.

It follows that there needs to be education at all levels of the criminal justice system, scientists acting for both prosecution and defence, along with lawyers and judges. This creates a level playing field. Scientists working for defence and prosecution need to collaborate with each other and write joint statements where there are disputes, e.g. see courts of England and Wales recommendations:

Forensic Science Regulator Codes of Practice and Conduct:
Development of Evaluative Opinions FSR-C-118 Issue 1, 2021,
https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/960051/FSR-C-118_Interpretation_Appendix_Issue_1__002_.pdf

With an example (mentioned earlier) at:
https://www.bailii.org/ew/cases/EWCA/Crim/2020/1021.html

You can find further thoughts on this at https://doi.org/10.1002/wfs2.1321

Gill, Peter. "Interpretation continues to be the main weakness in criminal justice systems: Developing roles of the expert witness and court." *Wiley Interdisciplinary Reviews: Forensic Science* 1.2 (2019): e1321.

PC3

NISTIR 8351-DRAFT comments

Norah Rudin [redacted]

Mon 7/19/2021 12:05 PM

To: ScientificFoundationReviews <ScientificFoundationReviews@nist.gov>

1. The review is comprehensive, well-considered and well-written. However, the disclaimer that "The findings described in this report are meant solely to inform future work in the field." stands to negate the entire body of work. First of all, this needs expansion and clarification. Does it mean that past work is to be given a free pass from criticism? Does it mean that no court case should be reviewed in light of this document? How far in the future? The moment the final version of this review is released? When persons or labs should reasonably be aware of its contents? Please consider whether such a disclaimer is necessary or even useful. None of the ideas and concepts discussed in this review are new or novel. While this is indeed a useful compilation, the forensic DNA community should already be aware of everything discussed. To suggest that the publication date of a single review be used as a demarcation for the requirement to adhere to best scientific practices and to apply critical thinking is to suggest that this is a paradigm shift rather than a summary of decades of evolution in scientific thinking. If you feel compelled to keep some version of a disclaimer, perhaps consider something more along the lines of the SWGDAM 2017 wording - "*… With the underlying assumption that work performed prior to the issuance of these revisions was appropriate and supported by validation, ….*" Better yet, please consider whether such a disclaimer is even necessary, and whether it actually contributes to the application of forensic DNA to the fair administration of justice.

2. The review perpetuates the unfortunate historical LR designations of Hp and Hd. First of all, the hypotheses may or may not be what prosecution or defense might propose as explanations for the evidence. Second, while laboratories may have some idea of the prosecution hypothesis, they generally will not have had a specific discussion with defense prior to generating LRs and writing a report. Thus this practice is more prejudicial to defense than prosecution and sets up yet another barrier to affording defendants fair access to the criminal justice system. Third, any number of pairs of hypotheses might be generated for any item of evidence. It starts bordering on the absurd to call the numerator and denominator for each Hp and Hd. This would be an excellent opportunity to suggest that the community switch to neutral numbered hypotheses: H1, H2, H3 etc. That would remove any hint of scientists guessing about legal explanations for evidence, and provides a much more ordered paradigm for multiple pairs of hypotheses.

3. The one thing this review fails to address is the practice of continuing to use historical approaches of interpreting and weighting evidence. A non-trivial number of laboratories have not switched to, and have no immediately plans to transition to, a PG approach. Thus, even a document that is forward-looking should address these issues. Because those publishing papers, giving presentations, and writing guidelines have, for some time, been concentrating on transitioning to PGS, a vacuum has been left with regard to guidelines commenting on best practices for binary approaches. Please consider adding a section that addresses this current lack of guidance.

Respectfully,

Norah Rudin

If you have received this e-mail in error, please delete it.

~~~~~~~~~~~~~~~~

Dr. Norah Rudin, Ph.D.
Forensic DNA Consultant
[redacted]
www.forensicdna.com
[redacted]

~~~~~~~~~~~~~~~~

Science over Fiction

~~~~~~~~~~~~~~~~

New Nature photos posted
~~~~~~~~~~~~~~~~

PC4

DNA Mixture Interpretation: A NIST Scientific Foundation Review - Comment

GOBLE, TIMOTHY (TROOPERS) [redacted]

Tue 7/20/2021 1:04 PM

To: ScientificFoundationReviews <ScientificFoundationReviews@nist.gov>
Cc: FITZPATRICK, MEEGAN (TROOPERS) [redacted]; MCGINNIS, CARRIE (TROOPERS) [redacted]; MYERS, DANIEL (TROOPERS) [redacted]

Good afternoon,

I reviewed Chapter 2: Principles and Practices and found it easy to follow and understand.

I did notice a typo in **Table 2.3** on page 35. I believe the top right column should say “Mathematical models”, instead of “Mathematically models”.

Thank you,

Timothy J. Goble
Supervisor of Forensic Services (DNA), Biological Science

New York State Police Forensic Investigation Center
1220 Washington Ave., Building 30 Albany, NY 12226-3000

[redacted]

For serology or DNA case inquiries please email: Biosci-CMGT@troopers.ny.gov

www.troopers.ny.gov

PC5

Gill reference in Forensic DNA Mixture Review

Mark Timken [redacted]

Wed 7/21/2021 1:12 AM

To: ScientificFoundationReviews <ScientificFoundationReviews@nist.gov>

On pp. 11-12 of the review, there is the following:

"Complications can also arise when reduced DNA template amounts are used in PCR, where random sampling, also known as stochastic effects, make it more difficult to confidently interpret the resulting DNA profile (e.g., Gill et al. 2000)."

An actual relevant reference to random allelic sampling and consequent stochastic effects is not to the Gill et al. 2000 paper but to the Gill et al. 2005 paper.

PC6

Comments on Appendix II

FITZPATRICK, MEEGAN (TROOPERS) [redacted]

Wed 7/21/2021 10:11 AM

To: **ScientificFoundationReviews <ScientificFoundationReviews@nist.gov>**
Cc: PIZZIKETTI, JULIE (TROOPERS) [redacted]; STREVELL, CHERYL (TROOPERS) [redacted]

Good morning,

I have reviewed Chapter 1 and Appendix II and have only 1 comment on Appendix II:

On p. 199, it mentions the FLSB emails, however, I am sure this document was written before Jeff accepted a position at the Global Forensic and Justice Center. This will need to be updated. I know I am looking forward to when this service is back on-line.

I do have a follow up question regarding the recommendation in Appendix II about having 5% of their paid time or two hours each week dedicated to continued learning and journal article review. Are you aware of labs who are currently doing this? I would be interested in discussing their programs and how they manage this. We are currently using our review period to build up our article library along with the SWGDAM recommended training journal articles I have been requesting.

Thank you!
Meegan

Public Comment Period for NIST Review on DNA Mixture Interpretation
EVENT DATE: WEDNESDAY, 21 JULY 2021 - 01:00 PM to 04:00 PM

Posted Questions* with Times Received

(Responses May Have Been Provided in a Different Order)

*Names listed were self-created by attendees; thus some names may be aliases

[01:25 PM]
Richard Wivell asked : Does the report reflect a consensus across all the scientists mentioned in Table 1.2?

[01:32 PM]
Tamyra Moretti asked : Dr. Butler mentioned, as expected given the use of PGS in the criminal justice system, that questions/concerns have been raised in a number of admissibility hearings. Please discuss the outcome of those hearings. What do Court decisions indicate overall as to the reliability of PGS for forensic use?

[01:37 PM]
Tamyra Moretti asked : Related to my question, what do you conclude given that individual laboratory validation studies have been reviewed as part of hearings in which the PG evidence was ruled admissible?

[01:56 PM]
Marla F Kaplan asked : In continuation of the below question, the recommendation implies that technical leaders have not done a quality job at designing these validation studies, interpreting the resulting data, and making the subsequent implementation decisions.

[02:02 PM]
Tamyra Moretti asked : So as not the misunderstand Dr. Butler's statement regarding lack of public access and review of data, all internal DNA validations by US government labs that participate in NDIS are externally peer-reviewed by qualified forensic DNA analysts, although they may not be in the public realm.

[02:04 PM]
Tamyra Moretti asked : Can you please describe an example of sufficiency in factor space? You say "bracket," which is the easy part - the extremes of the range. What does this space look like in terms of density?

[02:08 PM]
Laura asked : When validation studies (internal) are being done by an outsourced company how are the actual practitioners (and then anyone the data is affecting) benefiting from those validations?

[02:11 PM]
Tamyra Moretti asked : For transparency, please share with this audience what you noted as "challenges" for conducting this review: e.g., no defined threshold for establishing that "reliability" has been demonstrated, lack of data to conduct this foundation review, etc.

[02:21 PM]
Richard Wivell asked : Does this foundational review seek to determine if PGS is suitable for the interpretation of mixed profiles in casework- or- is it down to a lab by lab, case by case basis dependent on the internal validation studies done?

[02:27 PM]
Laura asked : With LR, if the analyst includes suspect in calculation then how does that not go against the presumption of innocence? As noted in State v Skipper 228 Comm 610 1994. Just positing two theories goes against the US criminal justice system as the defense does not have a burden. How do you get around?

[02:30 PM]
Marco Bosmans asked : Has the shift (that we see in the Netherlands), from using consensus profiles to using all detected alleles of multiple runs of DNA mixture profiles for LR calculations, ever been reviewed? Can you point me to relevant literature?

[02:33 PM]
Brad Maurer asked : In your opinion, can a court reasonably conclude that a PGS system has been reliably applied in a given complex mixture case without access to the relevant validation data?

[02:34 PM]
Tiffany Roy asked : A colleague of mine in Cognitive Science told me that the underlying data for all published studies is required for her publications. Why have forensic journals not required published underlying data?

[02:35 PM]
lara adams asked : Please expand on LRs are not measurements, but assigned by individuals. Comparisons and propositions are analyst dependent, but IMO the LR assigned by PGS (though variable through MCMC) is a mathematical measurement of the support for those propositions, from models algorithms and pop gen data sets.

[02:36 PM]
Ann Marie Gross asked : How many of the authors of the 60 articles on Prob Gen were contacted to see if they would release their data for the independent review for determining if there is sufficient scientific foundation for the use of prob gen

[02:37 PM]
Ann Marie Gross asked : How many years of hands on experience conducting mixture interpretation in an accredited forensic DNA laboratory do the authors of this report have?

[02:37 PM]
Brad Maurer asked : In your opinion, should a given DNA analyst be qualified as an expert witness in a case involving complex mixture(s) if that analyst's proficiency in interpreting complex mixtures has never been assessed?

[02:41 PM]
Bicka asked : How does the involvement of the developers of the PG systems directly in the internal validation of the programs impact the reliance that the scientific community puts on these studies since the developers have an interest in the PG systems being widely adopted?

[02:43 PM]
Marla F Kaplan asked : Dr. Butler, Section A.2.3.5 discusses the need for additional training for technical leaders in order to design validation experiments and statistical analysis. How many internal validation studies related to mixture interpretation did the panel review before making this recommendation?

[02:48 PM]
Sammy Bearkat asked : What value do ROC curves have in testing the reliability of PGS Software from Brand X if Brand X PGS software is used to generate the ROC curve? This seems like circular logic.

[02:50 PM]
Bess Stiffelman asked : Will you consider incorporating the internal validation studies on relatedness that were recently made public and can be found here: https://indefenseof.us/issues/kinship-problem

[02:50 PM]
Garon Foster asked : Please elaborate on Key Takeaway #4.7. Are the authors suggesting that validation data and/or validation summaries be included in a laboratory report (or just available for review to the user of the report in a public forum)?

[02:51 PM]
Tamyra Moretti asked : RE Key Takeaway 4.4 & sufficiency of data that you seek for your review. How would we meet your requirements to enable reliability based on your [undefined] criteria? I'm not sure what we are being asked to do. "More" is of course better but we must know your minimum criteria, if they exist?

[02:58 PM]
Mary Lou Nicholson asked : In key take away 4.3, it states that in order to determine the reliability of DNA mix interpretation practices, forensic labs should make their validation data public. What does mean specifically? What public forums are being proposed? How does this impact accreditation requirements?

[02:59 PM]
Mary Lou Nicholson asked : What data are you actually asking for to be made public?

[03:03 PM]
Jessica Charak Lehrner asked : Did the group consider how the results of internal validation studies were ultimately used to inform the adopted laboratory protocols, including the types of likelihood ratios that will be reported based on available case-specific info (e.g., HPD, unified), the use of report disclaimers, etc.?

[03:05 PM]
david lynch asked : Can the report please discuss the discrepancy between what PG gives the jurors v. what they need. Jurors are given "likelihood of seeing the evidence GIVEN defendant is guilty/innocent" but jurors need to know "likelihood def is guilty/innocent GIVEN the evidence".

[03:05 PM]
Dorothy Catella asked : If internal validation data is available upon request, what is the point in having it in a public forum.

[03:05 PM]
Brad Maurer asked : Is underlying validation data merely "helpful" to assessing the reliability of a given complex mixture interpretation in casework, or is it necessary?

[03:06 PM]
Susannah Kehl asked : Who is the intended end user of this report? What are you hoping to accomplish with it?

[03:07 PM]
Caitlin Rogers asked : Is the report recommending (in section 5.5) a shift to evaluative reporting and testimony?

[03:09 PM]
david lynch asked : Our lab is blocking us from seeing the underlying data for their validation because they used their own analysts' DNA and are claiming privacy concerns. How necessary is it for the underlying validation data to be released for review?

[03:09 PM]
Terri Rosenblatt asked : Could you please address whether NIST would include an assessment of the racial impact of PGS. People of color are predominately the subject of PGS testing in crime labs, largely as a result of policing decisions. Should any recommendations about PGS consider who the results are used against?

[03:11 PM]
Lynne K Burley asked : To your knowledge, has there been any dialogue or movement with PT providers (CTS, Forensic Assurance) regarding takeaway 4.5?

[03:12 PM]
Clinton Hughes asked : Would NIST be able to perform the algebraic reformatting mentioned in Box 4.1 to get past labs' privacy arguments about validation donor profiles, or would the individual labs have to do it? And how would that work with epgs?

[03:14 PM]
Mary Lou Nicholson asked : During your foundation review, did you look at contamination rates in forensic labs? How do labs approach reporting contamination rates?

[03:15 PM]
Brad Maurer asked : Can a lab's/analyst's conclusions about a given evidence sample be considered scientifically reliable if the factor space in which that sample lies falls outside the range of samples assessed in that lab's validation studies?

[03:19 PM]
Mary Lou Nicholson asked : Some forensic labs are using a Baysian approach to providing opinion evidence on DNA transfer. Yet there are no guidelines for this type of approach. As stated in this review paper, it states that more studies need to be done on this area. Based on your review, how many labs are doing this?

[03:21 PM]
Ray Wickenheiser asked : Is mixture interpretation reliable if there are good peak heights with no drop out, no degradation (i.e. a good quality profile)?

[03:25 PM]
Bess Stiffelman asked : When discussing relevance based on case contexts and sample quality, are you suggesting analysts should include in their report a conclusion about relevance? Most use a verbal scale for the LR, but that doesn't incorporate any other factors.

[03:25 PM]
Greg Hampikian, Boise State asked : Scientifically, is it valid to report results outside of your validated limits (amount of DNA, # of contributors, etc.)?

[03:25 PM]
Katrina Vetrano asked : A key takeaway is that complex mixtures are more difficult to interpret than simple mixtures. Does the report distinguish between the reliability of interpreting 3 person vs. 4 person mixtures?

[03:25 PM]
Tamyra Moretti asked : Is the issue "consisten[cy of] likelihood ratio?

[03:28 PM]
Joanne B Sgueglia asked : Given all labs need to create a set of validation samples for the mixture studies -- would it be possible to have a vendor or PT provider consider making a set of mixtures for labs to purchase. If NIST/group works to create what would be the best set--a standard set of validation samples could be ma

[03:28 PM]
Tiffany Roy asked : Will all the comments be made public? Will the people who made the comments be identified in published commentary?

[03:29 PM]
Tamyra Moretti asked : Can you please comment on the reliability and consistency of LRs close to 1, as distinguished from a low weight of evidence?

[03:30 PM]
Marko asked : Are the softwares currently used in USA suitable for LR calculation when the contributors are related and there are drop out events in mixture?

[03:30 PM]
Allison Lewis asked : The report refers to "complex mixtures" as more than 2p in some discussions but then also seems to address 2p or more in others. Can you clarify?

[03:31 PM]
Greg Hampikian, Boise State asked : I am not sure I understand your view on validating limits of analysis. Are you suggesting that labs should be allowed to extrapolate beyond their validated limits regarding the minimal amount of DNA they can accurately analyze?

[03:32 PM]
Bicka asked : Have you considered when a change in version of a PGS is enough to trigger new validation? How does the version and changes in the program figure into Factor Space?

[03:33 PM]
D asked : Is there a concern that this report can be used by the legal system to make assumptions about cases in which the authors were not intending--primarily from attorneys by misinterpreting the intent or true meaning of the key points? What considerations, if any, are being done to guard against this?

[03:35 PM]
Brad Maurer asked : What should labs look to in formulating policies regarding conditioning evidence samples on a given contributor?

[03:36 PM]
Tamyra Moretti asked : Refer to QAS regarding software upgrades and validation

[03:38 PM]
Lynne K Burley asked : In your opinion, what are the most imperative gaps that need to be filled further regarding takeaway 5.6?

[03:38 PM]
david lynch asked : Does/should NIST recommend a threshold LR for an "inconclusive" finding, i.e. the range 0.001-1,000 as recommended by Buckleton of STRmix? Or, a recommendation that labs not set this range narrower than the false positives they observed in their validation?

[03:40 PM]
ACHIN JANA asked : If LR varies according to various softwares then don't you think that prejudices justice delivery system?

[03:41 PM]
Mythri Jayaraman asked : This question has still not been addressed:
Where do the population allele/genetype frequencies come from? Are there recommendations relating to this? Ie, should the allele frequencies used be updated at certain intervals? How old are the frequencies we use now?

[03:41 PM]
Carlotta Lepingwell asked : When do you expect to publish the final report?

[03:42 PM]
Monica Sloan asked : Will the slides from this presentation be made publicly available? I'm going to want to re-read the report with this summary handy. Thanks!

[03:43 PM]
Robert Bever asked : Is this report only to be associated with STR data? can it be extended to MPS data... this should be stated in the report.

[03:43 PM]
Tamyra Moretti asked : FBI pops, 2016

[03:43 PM]
Mythri Jayaraman asked : would the LR be different, depending on which data set is used? Is there any recommendation in the report for labs to use a particular data set over others?

[03:46 PM]
peter.vallone asked : Thank you for your questions!

Not Posted (Declined) Questions

[01:11 PM]
ACHIN JANA asked : Would recording of the webinar be available afterwards?

[01:25 PM]
Ray Wickenheiser asked : Slides are not advancing.

[01:26 PM]
Casandra Setser asked : Are there technical difficulties on the presenter end right now? I'm trying to figure out if I should be seeing something on my screen yet.

[01:57 PM]
Asad Saeed asked : have the judiciary system in USA access on programming of PGS software(s) used for reporting?

[02:37 PM]
Marko asked : Do You consider interpreting mixtures with more than four contributors?

[02:43 PM]
Ariel Payan asked : Have you had a chance to review the LCO (Law Commission of Ontario) report? In that report they recommend that barring legislation amending the admission of AI generated evidence that the Canadian courts adopt a presumptive inadmissibility of the introduction of PGS systems evidence.

[02:44 PM]
ARASH asked : Is it mandatory to mention the PGS system used in our report? (in terms of our legal responsibility and transparency)

[02:45 PM]
Mythri Jayaraman asked : Where do the population allele/genetype frequencies come from? Are there recommendations relating to this? Ie, should the allele frequencies used be updated at certain intervals? How old are the frequencies we use now?

[02:47 PM]
Qazi Laeeque Ahmed asked : Please share url where recording of the webinar would be available.

[02:47 PM]
Qazi Laeeque Ahmed asked : Please share url for recording of this webinar.

[02:52 PM]
Bshar asked : Can we search for people entered into the database by entering the mixture so that the suspect can be identified?

[03:05 PM]
Hossam Tawaha asked : do you recommend to all forensic labs have PGS to analyze mixture samples?

[03:26 PM]
Mythri Jayaraman asked : Where do the population allele/genetype frequencies come from? Are there recommendations relating to this? Ie, should the allele frequencies used be updated at certain intervals? How old are the frequencies we use now?

[03:29 PM]
david lynch asked : Thank you and I like the text on your shirt ;)

[03:39 PM]
gem asked : Are there a "standard statement" that is being followed when doing the final reports in mixture interpretations? What should be included in every report done that makes your analyses accepted legally?

[03:41 PM]
Allison Lewis asked : thanks!! this was very helpful. (no need to post this but also great shirt, Dr Butler!)

[03:46 PM]
Chris Glaze asked : Who did it?

[03:47 PM]
Jayshree Patel asked : Thank you John and team.

[03:48 PM]
ACHIN JANA asked : How will we get certificates?

[03:48 PM]
Imma asked : Thanks for having us

PC7

COMMENTS ON NISTIF 8351 Draft DNA mixture Interpretation Scientific Foundation Review

Lucy A. Davis [redacted]

Wed 7/21/2021 4:37 PM

To: ScientificFoundationReviews <ScientificFoundationReviews@nist.gov>

Throughout the entire document when you reference and hyperlink to a published standard as ANSI/ASB *Year*. You routinely use the publication date (e.g., "ANSI/ASB 2020") that misrepresents the actual standard. The 2020 is the date the standard is published and ASB has published multiple standards in 2020. Recommend use the actual standard unique number e.g., "ANSI/ASB 018".

Definitions or around Lines 6245-6249 and 6358 – you use SDO and "standards" and the reader may benefit to understand what a "documentary standard" and what an SDO (Standard Developing Organization) does. If not included in the definition, somewhere within the document provide the information as to what a "documentary standard" is and the difference is between standard (e.g., FBI Quality Assurance Standards) and a published "consensus-based standard" or American National Standard (ANS).

Line 2977 – It might be useful to reference here ANSI/ASB Standard 018, Standard for Validation of Probabilistic Genotyping Systems. Standard 4.1.2 specifically discusses requirements a developmental validation shall include. "Developmental validation studies shall address the following: accuracy, sensitivity, specificity, and precision. These studies shall include case-type profiles of known composition that represent (in terms of number of contributors, mixture ratios, and total DNA template quantities) the range of scenarios that would likely be encountered in casework. Studies shall not be limited to pristine DNA samples but shall also include compromised DNA samples (e.g., low template, degraded, and inhibited samples)."

Line 6064 – use the proper standard number to identify all 3 ASB standards. "ANSI/ASB 2019" (actually ANSI/ASB 040) is now published.

Lines 6255 & 6662 – Identify "AAFS Standards Board" either has "Academy Standards Board" or "AAFS Standards Board (ASB)"

Line 6267 – "…being finalized through the AAFS Standards Board DNA Consensus Body with the SDO process". Although mentioning the Consensus Body name is nice, it is finalized by ASB and publication by ANSI as an American National Standard.

Line 6660 - The proper name to be referenced is: ANSI/ASB Standard 018, Standard for Validation of Probabilistic Genotyping Systems (2020). You use "ANSI/ASB 2020 PGS Validation Standard". ANSI/ASB 2020 does not accurately identify the standard.

Section A1.8.3 – It might be appropriate to add to training conducted by OSAC and ASB through Promega *ANSI/ASB Standard 018 - Standard for Validation of Probabilistic Genotyping Systems, First Edition 2020 (Part 4)* the describes the ANSI/ASB Standard 018 and provides information concerning implementation of the standard. It is available free of charge at https://event.on24.com/eventRegistration/EventLobbyServlet?target=reg30.jsp&partnerref=postwebpage&eventid=2880133&sessionid=1&key=8E38164BA0F3689D5C7F0355169CAFF7&regTag=&V2=false&sourcepage=register

Section A1.9 – Add either a new Key Takeaway or add to #A1.3 that published documentary standards are available related to Validation of PGS (ANSI/ASB 018), Validation of DNA mixtures (ANSI/ASB 020), Internal Validation of DNA methods (ANSI/ASB 038), and Interpretation and Comparison Protocols (ANSI/ASB 040. Also, additional standards being developed concerning training and Formulating Propositions for Likelihood Ratios in Forensic DNA Interpretations (ASB 041) and Standard for Training in the Use of Statistics in Interpretation of Forensic DNA Evidence (ASB 081).

Section A2.3 – Reference standards published related to training: ANSI/ASB Standard 040, ANSI/ASB Standard 022, Standard for Forensic DNA Analysis Training Programs, and being developed, Standard for Training in the Use of Statistics in Interpretation of Forensic DNA Evidence (ASB 081).

Lucy A. Davis
LDH Consultants, LLC
[redacted]

PC8

RE: Comments on Appendix II-possible typo

FITZPATRICK, MEEGAN (TROOPERS) [redacted]

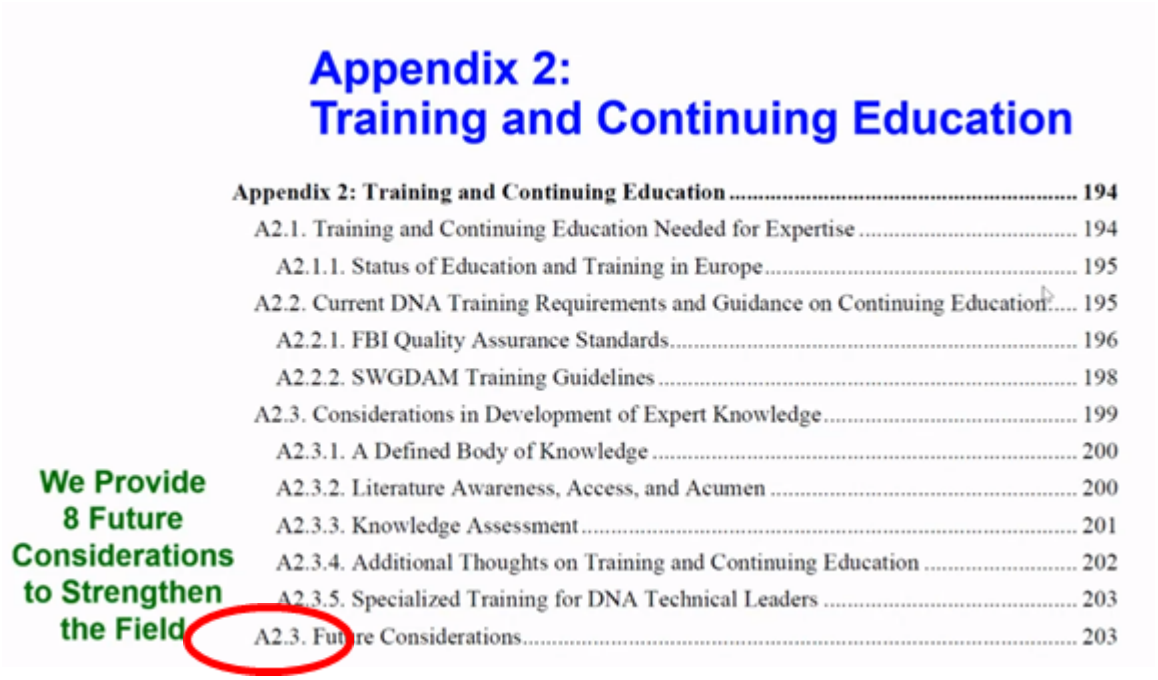
Thu 7/22/2021 7:50 AM

To: Butler, John M. (Fed) <john.butler@nist.gov>; **ScientificFoundationReviews <ScientificFoundationReviews@nist.gov>**
Cc: PIZZIKETTI, JULIE (TROOPERS) [redacted]; STREVELL, CHERYL (TROOPERS) [redacted]; KOVARI, JEANETTE (TROOPERS) [redacted]

Good morning Dr. Butler,

Thank you for doing the webinar yesterday. I thought it was a great overview and very helpful. And many of our staff loved your shirt decal.

One of my colleagues notice a potential typo in one of the presentation slides. Jeanette also noticed this was from the table of contents, so we wanted to reach out to you to mention it. We try to take screenshots of all of our continuing education webinars for our records and to have as references, and this is from our notes from your presentation yesterday .



Thank you!
Meegan

PC9

FW: mixtures review

Butler, John M. (Fed) <john.butler@nist.gov>

Thu 7/22/2021 10:10 AM

To: **ScientificFoundationReviews <ScientificFoundationReviews@nist.gov>**

Further information from Peter Gill

From: Peter Gill [redacted]
Sent: Thursday, July 22, 2021 9:35 AM
To: Butler, John M. (Fed) <john.butler@nist.gov>
Subject: Re: mixtures review

Hi John,

Congratulations on your marathon yesterday. Following the discussions, I have jotted down some further thoughts.

Best wishes

Peter

PC9a

1) There was some confusion about the meaning of "evaluative" reporting, where it was described as synonymous with "activity-level" reporting.

   A report can be either "investigative" or "evaluative". The former is applied when there is little information in a case and where a suspect is not immediately apparent. Typically an investigative search will be used to interrogate a national DNA database. Once candidate(s) are found, then the investigators will be provided with a list in order to discover the other non-DNA evidence in the case – eventually a candidate may become a suspect.

   Once a suspect is found, an evaluative report can be written using the agreed assumptions and propositions. This report is then forwarded to the court and forms the basis of the prosecution case.

   Investigative software (e.g. CaseSolver in EuroForMix) is used for complex cases where there are numerous crime stains (e.g. 100) and potential suspects (e.g. 30) to investigate. It is very time consuming to work through all possible permutations, hence software can be used to identify potential associations between suspects and crime stains using generalised propositions. Once associations are found, the analysis proceeds in the evaluative mode where it may be necessary to refine propositions and assumptions (e.g. frequency database; numbers of contributors).

   Searching cases for potential contamination from lab workers or police investigators is an extension of investigative software.

2) There needs to be clarification about the so called "relatedness problem". The "problem" is presented as follows: "I know that I will achieve a much reduced LR if a relative is conditioned. Therefore if the (unknown) truth state is that the donor of the crime stain is a sib of the defendant, and the defence proposition does not take this into account then the strength of the evidence is significantly overstated."

   The LR cannot tell us what the truth state is, it can only tell us which proposition is more likely. Indeed, both propositions may be false, yet a probative LR is still achieved. This is why it is important to ensure that the propositions are properly formed from a consideration of the case circumstances, taking account of views of both defence and prosecution. Propositions should as far as possible be formulated before the crime-stain DNA analysis is carried out in order to prevent any sub-conscious bias. At this stage, there will be a consideration of whether it is reasonable to consider sibs etc as an alternative suspect.

   Read the ISFG DNA commission 2018 paper, section 7.2 for an explanation of the impact of prior odds and details of the maths used in the following examples.

   To put into perspective, it is useful to consider the impact on the evidence in relation to prior odds. Suppose we have a crime stain where the propositions are

   Hp: The donor is Mr S

   Hd: The donor is an unknown individual.

And the LR=1bn in favour of Hp

Suppose there is no other information in the case, other than the perpetrator is a male from a population of the UK (approximately 30m individuals). My prior odds of picking the donor before the DNA evidence is 1:30m. The posterior odds are 1/30m x 1bn = 33.3:1

Now consider the alternative set of propositions where the defence propose a sib as the unknown perpetrator so that:

Hp: The donor is Mr S

Hd: The donor is a bother of Mr S

And the LR=1000 (considerably less than the previous unrelated example)

If an individual has two sibs, the population of suspects is considerably smaller. Hence the prior odds are now 1:2 (compared to 1/30m for the previous example) and the posterior odds are 1/2 x 500 = 250:1, which is greater than that shown in the previous example of unrelatedness.

Therefore, this shows that there must always be careful consideration to establish if close relatives should be included in calculations; it does not particularly help the defence to propose, without justification, that the alternative contributor is a sib of the defendant since the suspect population is narrowed to a handful of individuals and the priors for any given individual is high (indeed subsequent profiling may completely eliminate the other sibs from the suspect pool), whereas the suspect pool is always much larger for the unrelated example but the priors are much lower.

Of course we tend to avoid priors in casework, but it is pertinent to for the scientist to opine on the size of the population of suspects when there is very limited other information since this places the LR into perspective when relatedness issues are raised.

Recommendation 9 of ISFG DNA commission

It is crucial to outline that scientists do not give their opinion on who is the source of the DNA. There is a difference between the probability of the results given that the DNA is from an unknown person and the probability that the DNA is from an unknown person given the result. To equate one with the other is known as the transposed conditional, the prosecutor's fallacy, or the source probability error. It is thus important to explain what the likelihood ratio is and what it is not. This can be done by training or by providing a table with different odds, the LR and resulting posterior odds [75]. Because of the dangers of misrepresentation, it is essential to convey that scientists do not give opinions on the probability of propositions [25] and this is reinforced here.

3) On MPS please note following publications (we have SNP and STR MPS modules encoded into EuroForMix):

Bleka, Øyvind, et al. "Open source software EuroForMix can be used to analyse complex SNP mixtures." *Forensic Science International: Genetics* 31 (2017): 105-110.

Bleka, Øyvind, et al. "An examination of STR nomenclatures, filters and models for MPS mixture interpretation." *Forensic Science International: Genetics* 48 (2020): 102319.

Bleka, Øyvind, et al. "Automation of high volume MPS mixture interpretation using CaseSolver." *Forensic Science International: Genetics Supplement Series* 7.1 (2019): 14-15.

And chapter 13 of "Forensic Practitioner's guide to the interpretation of complex DNA profiles

PC10

Comment on NIST DNA Mixture Interpretation: A Scientific Foundation Review

Bjorn Sutherland [redacted]

Thu 7/29/2021 12:45 AM

To: ScientificFoundationReviews <ScientificFoundationReviews@nist.gov>

Dear Sir/Madam,

Please see attached a response to this review.
Kind regards
Björn

PC10'

RE: Comment on NIST DNA Mixture Interpretation: A Scientific Foundation Review

Bjorn Sutherland [redacted]

Thu 7/29/2021 4:59 AM

To: ScientificFoundationReviews <ScientificFoundationReviews@nist.gov>

Dear Sir/Madam,
Could you please disregard our previously sent response and replace it with the attached, which has a corrected website link.
Kind regards

Björn Sutherland MSc
STRmix Manager
Kenepuru Science Centre: 34 Kenepuru Drive, Kenepuru, Porirua 5022
PO Box 50348, Porirua 5240, New Zealand
[redacted]
www.strmix.com

STRmix.
EMPOWERING FORENSIC SCIENCE.

From: Bjorn Sutherland
Sent: Thursday, 29 July 2021 4:38 PM
To: 'scientificfoundationreviews@nist.gov' <scientificfoundationreviews@nist.gov>
Subject: Comment on NIST DNA Mixture Interpretation: A Scientific Foundation Review

Dear Sir/Madam,

Please see attached a response to this review.
Kind regards
Björn

PC10'a

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

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

29 July 2021

Executive summary

- NIST state that they cannot find enough data by an internet search to verify validity of any mixture analyses
- We have placed a large amount of data in the public domain here:
- Activity and sub-source level considerations should not be mixed

Background

In 2016 The President's Council of Advisors on Science and Technology (PCAST) [1, 2] published a report. We paraphrase PCAST's main findings on DNA mixtures here. PCAST accept that validity has been established up to three person mixtures in which the POI is at least 20% of the DNA. They call for more and broader testing and ask for full independence from the developers (pg 79) or inclusion of the developers with others (pg 81). They ask for the research to be in the peer reviewed literature. We note that PCAST ask (finding 3 pg 82) that DNA analysis of complex mixtures should move rapidly to more appropriate methods based on probabilistic genotyping and that "at present, published evidence supports the foundational validity of analysis, with some programs, of DNA mixtures of 3 individuals in which the [POI] constitutes at least 20 percent of the intact DNA in the mixture and in which the DNA amount exceeds the minimum required level for the method. The range in which foundational validity has been established is likely to grow as adequate evidence for more complex mixtures is obtained and published". PCAST are clear that their expectation is publication in scientific journals[1].

PCAST (@ pg 83) called on NIST to play a role in this process, by ensuring the creation and dissemination of materials and stimulating studies by independent groups through grants, contracts, and prizes; and by evaluating the results of these studies. This has not happened.

[1] PCAST pg 81 Because empirical evidence is essential for establishing the foundational validity of a method, PCAST urges forensic scientists to submit and leading scientific journals to publish high-quality validation studies that properly establish the range of reliability of methods for the analysis of complex DNA mixtures.

In October 2017 NIST announced the commencement of a study to "Assess the Reliability of Forensic Methods for Analyzing DNA Mixtures"[2] John Butler introduced this describing his conclusion prior to the study that "*Just in the past two years, there has been a huge rush to go into the probabilistic genotyping field, and people are jumping into this without really thinking about a lot of these issues: how sensitivity impacts what they're doing, how "transfer" and "persistence" of DNA can impact their results, and what they're doing in terms of the way that they set up their propositions that go into the likelihood ratios of their probabilistic genotyping programs*."[3]

Four years later and after summarising an extensive body of research Butler et al. report their current view (hereafter "The NIST foundational review" or "NFR"). We have divided our response into themes and address each below.

Lack of available supporting data (Key takeaway 4.3 line 741)

We focus initially on NFR clause #4.3 and Box 4.1. Clause 4.3 reads: "*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*."

If we read this correctly then the authors' position is that reliability has not been demonstrated by *an external and independent assessment* for any forensic DNA interpretation. Our assessment of the NFR is that in order to meet a new criterion of *external and independent assessment* some data requirements exist. We note that this criterion differs from that of PCAST which was publication in scientific journals studies performed by or *including* independent research groups[4].

PCAST noted that they consulted with John Butler who concurred with PCAST's finding. We make this note because the lead author of the NFR is Butler who now introduces new criteria differing markedly from those he had agreed with in 2016. The most obvious differences that we observe are a move from publication in the peer reviewed literature to the

---

[2] https://www.nist.gov/news-events/news/2017/10/nist-assess-reliability-forensic-methods-analyzing-dna-mixtures

[3] https://www.propublica.org/article/putting-crime-scene-dna-analysis-on-trial

[4] The exact text from PCAST is: "*Because empirical evidence is essential for establishing the foundational validity of a method, PCAST urges forensic scientists to submit and leading scientific journals to publish high-quality validation studies that properly establish the range of reliability of methods for the analysis of complex DNA mixtures.*
When further studies are published, it will likely be possible to extend the range in which scientific validity has been established to include more challenging samples. As noted above, such studies should be performed by or should include independent research groups not connected with the developers of the methods and with no stake in the outcome."

placement on the internet of partially processed data. The NFR at pg 48 suggests "*we believe for information to be considered foundational, it needs to be reasonably accessible to anyone who wishes to review it.*" The NFR practically interprets "*reasonably accessible*" as being findable on the internet.

This work was US government funded to find "*What established scientific laws and principles as well as empirical data exist to support the methods that forensic science practitioners use to analyze evidence*?" (line 127). This could have been greatly facilitated by requesting data from US government laboratories as part of this work. One of their key points KT#4.1 (line 732) is that empirical testing must be undertaken (and that the user must test the system in a manner they will apply the method in casework, see lines 2941-2943).

More than 60 laboratories are using STRmix™ in casework in the US. Each laboratory would have completed their own internal validation. NFR only reviewed data in the 'public domain' (8 laboratories) which represents less than 15% of the data they had a mandate to review. We note that the validation data from laboratories with individuals listed as Members of the DNA Mixture Resource Group in Table 1.2 (line 1193) has not been studied. There has been a very considerable effort by many people in the US to test Probabilistic Genotyping (PG) software thoroughly and it would have been valuable to recognise this.

At no time, during the tenure of this review or earlier, did any member of the review approach us for the information they desired. We, and many others, could have gone a long way to meet their needs had we been approached. We did write to John Butler and Eric Lander twice in 2016 asking them to specify an experimental design that they wanted to demonstrate validity of STRmix™ that we would do. We received no reply.

Finally, in this section we note that one of the original aims of NFR was to "*develop a comprehensive, curated bibliography on DNA mixtures*" (line 2456). This goal "*proved unfeasible as a result of the constantly growing literature*" implying that lack of peer reviewed data supporting the use of PG was not an issue.

Requirements for validity

We discuss here the practicality of implementing NFR's requirements given in clause 4.3. Clause 4.3 needs to be read in conjunction with Box 4.1 which we reprise here.

Box 4.1 **Desired Information for Reliability Assessments of LR Values in PGS Systems** in part reads:

1. Sample Number or Unique Identifier
2. Number of Contributors (NOC)
3. Target DNA Template Amounts
4. Degradation Status of DNA Template(s)
5. NOC used for Analysis (Apparent NOC)
6. H1 true? (Yes/No)
7. Person of Interest (POI) position in the mixture (if H1 is true)
8. Reported Log10(LR)
9. Mixture EPG results
10. POI profile
11. Known contributor A profile and any additional known contributors
12. Noncontributor profile (if H1 is not true): is this profile simulated or determined from

Box 4.1 does not include the multiplex, cycle number, or injection conditions. We hope that NFR have trialled this data format and that it achieves whatever it is they desire.

We discuss further points 8, 10 and 11 in Box 4.1 in more detail below.

Bullet point 8 asks for the "Reported log10(LR)." We think the best number to use for scientific purposes is the point assignment for the sub-source propositions assuming unrelated unknown donors. Even then there will be embedded variability in the choice of ethnic database and value for the coancestry coefficient, θ.

Bullet points 10 and 11 ask for the genotype of individuals. Broadly, we have available two sources of data. The public domain PROVEDIt dataset [3] and samples we have obtained largely from our own or other laboratories. There is neither problem nor need for us to disclose the PROVEDIt genotypes. The PROVEDIt data has limited coverage, however. For example, the target templates for GlobalFiler profiles were 0.5, 0.25, 0.125, 0.063, 0.031, 0.016, and 0.007 ng. The mixture ratios targeted were:

2 donor	3 donor	4 donor	5 donor
1:1	1:1:1	1:1:1:1	1:1:1:1:1
1:2	1:2:1	1:1:2:1	1:1:2:1:1
1:4	1:4:1	1:1:4:1	1:1:4:1:1
1:9	1:9:1	1:1:9:1	1:1:2:4:1
	1:2:2	1:2:2:1	1:1:2:9:1
	1:4:4	1:4:4:1	1:1:4:4:4
	1:9:9	1:9:9:1	1:1:9:9:9
		1:4:4:4	

We make no criticism of PROVEDIt and note that this is an extensive set, increased by considering other factors such as multiplex and degradation state. Any finite set must have limitations. Coverage of the samples space was always impossible, and this can be shown by considering, for example, the two-donor set. The smallest minor is 10% of the DNA template. Given the interest in the low tail of the distribution this will not be adequate.

There are other peculiarities within PROVEDIt. For example, Reference K41 and has a confirmed PBSM at locus D1S1656 (as reported by Alphonse et al. [3]). It is not unreasonable to have a PBSM in the set and the effect of this is diagnosable in the mixtures. However, it requires attention by the operator that is not always given by *external and independent assessment* that may be inexperienced with the software and the PROVEDIt data. Some of the mixtures show erratic amplification, such as the complete drop-out of the sister allele of a peak at 406 rfu (for the 15s injection). We do not know the reason for the prevalence of such events and to fully accommodate them would require bespoke modelling.

Our other source of samples is laboratory data. These samples have often been obtained with informed consent from the individuals concerned. We are working with the data from these laboratories in a trusted capacity and we honour that trust. Consent very rarely includes permission to share personal data publicly and we note this is perhaps why the laboratories with individuals listed as Members of the DNA Mixture Resource Group in Table 1.2 have also not released their own data.

NFR suggests that if the privacy of the profile genotypes is a concern then alleles could be coded in an alphabetic format (Box 4.1 and also line 5755). They reference Gill et al. [4]. Privacy protection was not the purpose of the use of these codes by Gill et al. [4] which was simply to label mixture types. For example, AA:BC was a homozygote not overlapping a heterozygote. This is clearly evidenced by Gill et al.'s table 1 where they include both the genotypes of the contributors and the code.

We tested one such substitution code (alleles to letters). All staff queried broke the code independently in under 30 minutes by calculating allele frequencies and referencing known population databases. We think it is simply too great a risk and an inappropriate suggestion. We could potentially reduce the risk by destroying the allele order and having a different code for each mixture, but this would not allow any consideration of stutter overlapping alleles. Any code substitution would not allow for the replication of likelihood ratios (LRs).

In any case, we simply will not be permitted to place the genotypes in the public domain with or without coding and we note the inappropriateness of the suggestion and associated pressure from NFR to ignore the ethics of genetic privacy.

We make constructive counterproposals:

1. We have supplied summary data for each profile for a number of different published papers online (https://research.esr.cri.nz/articles/dataset/ESR_response_to_NISTIR_8351_-_DRAFT_DNA_Mixture_Interpretation_A_NIST_Scientific_Foundation_Review/15062907), including a value for allelic overlap (see below), and
2. NIST have had STRmix™ since March 2014, they could, in that time, have made and processed as much data as they desired, or

3. NIST could make mixtures and send them to us for interpretation.

Allelic overlap

The NFR does not define a measure of allele sharing nor were we able to obtain one by writing to Butler or Iyer. However, we have attempted to assess what it is they may want. Our measures neglect dropout and stutter. We did explore other options that included these, but complexity rose markedly. Without further guidance from the review team, we have proceeded with the definitions below in our recently published data summaries.

For *true donors* we report the fraction of alleles shared between at least two donors. Examples are given below:

Locus	True donors			Count of shared alleles	Count of alleles
	1	2	3		
1	AB	CD	EF	0	6
2	AA	CD	EF	0	6
3	AB	BC	EF	2	6
4	BB	BC	EF	3	6
5	BB	BC	CC	6	6

For *false donors* we report the fraction of alleles shared between the false donor and any peak above the AT in the mixture.

External and independent data

We think that NFR's suggestion is that the developers and laboratories are to publish their validation data, specifically the raw electropherograms, references and LRs, in the public domain without formal peer review. We assume that NIST, or someone else, will then interpret these results and draw conclusions. We infer that they intend to create ROC curves although we would greatly prefer a calibration analysis [5-7]. NFR authors comment (line 3723) that "*tools for examining calibration accuracy of LR assignments are less widely known to forensic scientists*".

We have already made ROC curves and calibration analyses from the multi-laboratory response to STRmix [8] (hereafter "the 31 laboratories data"). A paper on the ROC curves was rejected, partly due to lack of novelty. We posted this work [9] on the online open access repository Figshare.

We also infer that they seek to explore coverage. It would be greatly helpful to have this confirmed. NFR state (line 2906): "*The level of "coverage" is also critical; a laboratory has to have tested more than one sample of a particular type*."

We draw the reader's attention to the broad nature of the "*more than one*" clause and the difficulty defining "*particular type.*"

What we do note here is that this will end up being a considerable investment of effort by us, and probably others, to get the data in a suitable format and in the public domain. The final result will be partially external and partially independent since we will still have produced the data. We take this moment to point out that neither ourselves nor those laboratories using

STRmix™ have a vested interest to exaggerate STRmix™'s capabilities. Given the extensive usage and testing such an exaggeration would be rapidly exposed and destroy STRmix™'s reputation or expose laboratories to significant scandal and sanction. We have both an interest and a policy to absolutely disclose openly any limitations.

The only full solution is for NIST to create and run the samples themselves however we note the lack of courtroom, casework, and PG experience in the NFR team.

Likelihood ratios

On line 1918, NFR state that "*In recent years, the LR framework (Jackson et al. 2006) has gained widespread acceptance in DNA mixture interpretation (e.g., NRC 1996, Gill et al. 2006b) as a way of reporting the strength of evidence (E) in support of one proposition (H_1 or H_p) over an alternative proposition (H_2 or H_d or H_a).*" We note that these papers are at least 15 years old. The use of the LR is well established. On line 775 KT#4.8 is a request for more funding to review a method that we feel is already well established globally and predates the use of PG software.

On line 677 and 2123 (KT#2.6) it is stated that "*Likelihood ratios are not measurements*." Whilst they are not *absolute* measurements, they do provide a logical means to assign the value of findings within a defined framework. The accepted information, framework and propositions are key here. The importance of "I" or information and the propositions themselves cannot be underestimated in the calculation of an LR. Using different propositions including conditioning or siblings as alternative source as opposed to unrelated individuals must give a different likelihood ratio from the same evidence evaluated using different propositions or evidence. That is the case even for two-person mixtures with a clear major and minor contributor that could be interpreted "by hand" outside of PG. Given knowledge of the population genetic model (all PG use NRC II recommendation 4.2), values for theta, allele frequencies, and propositions, a likelihood ratio can be replicated. These are simple checks that go some way towards assessing the validity of the PG software [10].

On lines 3545-3556, NFR describe variation in LR for the same evidence given subjective decisions by an analyst. Changing LRs due to differing propositions and assumptions demonstrates the power of likelihood ratio and how an LR approach can accommodate different considerations more eloquently.

At line 2350, Principle 16 overstates the requirement for 'exhaustive' propositions. When formulating propositions, it is helpful to have all the relevant information to assign alternatives however there is no requirement for exhaustive propositions. This is echoed by many different standards bodies:

- *The assignment of a likelihood ratio therefore requires a pair of mutually exclusive propositions that reflect two competing positions, for example: that of the prosecution and the defence. These do not need to be exhaustive, but should reflect the positions of both parties.* DNA Commission of the International Society for Forensic Genetics [11]
- *H1 and H2 are two mutually exclusive propositions, but not usually mutually exhaustive*. Draft ASB Standard 041, Assigning Propositions for Likelihood Ratios in Forensic DNA Interpretations [12]

- *… for forensic evaluation it is not necessary that they be exhaustive. That is, they do not need to cover all possibilities; it is sufficient that they represent the two competing positions of the prosecution and defence within an accepted framework of circumstances*. UK Forensic Science Regulator [13]
- *Though the considered propositions are those deemed most relevant, they do not need to be exhaustive, so both propositions could be false. The likelihood ratio says nothing about propositions other than the two that were considered.* European Network of Forensic Science Institutes [14].

Would the NIST approach validate a software?

We would be concerned that an emphasis on coverage and ROC curves, if indeed that is NIST's intention, would not achieve the necessary purpose. ROC curves provide an estimate of the rates of false inclusion and exclusion. This requires the choice of a threshold (or the investigation of many thresholds) for the inclusion and exclusion decisions, which is something no one intends to do. To even get these curves many data are needed and certainly way more than one per type of mixture. Even if these conditions are met the ROC curve by itself gives no indication of the accuracy of any particular LR.

Calibration can test the accuracy of LRs en masse. That is, it can determine if a group of LRs are accurate in general but each individual LR may be inaccurate.

On line 3566, NFR state that "*The accuracy of the LR assessment in any specific casework situation cannot be determined*." In actual fact some assessment can be undertaken using Hd true trials as previously published [15].

Our own view is that validation is based on:

1. Belief, founded on empirical evidence, that the models are adequate representations of casework reality,
2. Belief, based on repeat calculation and sound mathematical inference, that the LR is assigned properly from the data and the models,
3. Black box testing of very large-scale false donor tests, and
4. Comparison with other software coupled with investigation of the causes of any difference.

A valuable exercise would be to determine what needs to be done locally and what can be done globally.

Chapter 5 activity level propositions

We recognise the importance of the content of this chapter although there are many inaccurate statements in NFR. However, one cannot associate concern with a misunderstanding of the hierarchy of propositions and the incorrect presentation of DNA findings with a review of PG and mixture interpretation methods. PG and mixture interpretation as discussed in this review is very firmly aligned with sub-source level propositions only. It is confusing and wrong to conflate the two topics in this one document. The assessment of source and activity level propositions perhaps deserves its own review.

Chapter 6 Future technology

This chapter offers an interesting insight into the potential alternative and novel solutions to mixture interpretation but does not address the current perceived issues around the application of PG with Capillary Electrophoresis data.

References

[1] President's Council of Advisors on Science and Technology. Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods. 2016. https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/PCAST/pcast_forensic_science_report_final.pdf. Accessed: 22 April 2017.

[2] President's Council of Advisors on Science and Technology. An addendum to the PCAST report on forensic science in criminal courts. 2016. https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/PCAST/pcast_forensics_addendum_finalv2.pdf. Accessed: 20 July 2017.

[3] Alfonse LE, Garrett AD, Lun DS, Duffy KR, Grgicak CM. A large-scale dataset of single and mixed-source short tandem repeat profiles to inform human identification strategies: PROVEDIt. Forensic Science International: Genetics. 2018;32:62-70.

[4] Gill P, Sparkes RL, Pinchin R, Clayton T, Whitaker JP, Buckleton JS. Interpreting simple STR mixtures using allelic peak areas. Forensic Sci Int. 1998;91.

[5] Bright J-A, Jones Dukes M, Pugh SN, Evett IW, Buckleton JS. Applying calibration to LRs produced by a DNA interpretation software. Australian Journal of Forensic Sciences. 2019:1-7.

[6] Ramos D, Gonzalez-Rodriguez J. Reliable support: Measuring calibration of likelihood ratios. Forensic Science International. 2013;230:156-69.

[7] Ramos D. On the Calibration of Likelihood Ratios. 2011. http://arantxa.ii.uam.es/~dramos/files/2011_02_08_WIC_Ramos_calibrtionLRValues_v2.pdf . Accessed: 5 November 2018.

[8] Bright J-A, Richards R, Kruijver M, Kelly H, McGovern C, Magee A, et al. Internal validation of STRmix™ – A multi laboratory response to PCAST. Forensic Science International: Genetics. 2018;34:11-24.

[9] Bright J-A, Taylor D, Curran J, Buckleton JS. Testing methods for quantifying Monte Carlo variation for categorical variables in Probabilistic Genotyping. 2020. https://figshare.com/articles/report/Testing_methods_for_quantifying_Monte_Carlo_variation_for_categorical_variables_in_Probabilistic_Genotyping/13180610. Accessed: 15 March 2021.

[10] Bright J-A, Evett IW, Taylor D, Curran JM, Buckleton J. A series of recommended tests when validating probabilistic DNA profile interpretation software. Forensic Science International: Genetics. 2015;14:125-31.

[11] Gill P, Hicks T, Butler JM, Connolly E, Gusmão L, Kokshoorn B, et al. 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.

[12] AAFS Standards Board. ASB Standard 041 Assigning Propositions for Likelihood Ratios in Forensic DNA Interpretations. 2019. http://www.asbstandardsboard.org/wp-content/uploads/2019/03/041_Std_Ballot01.pdf. Accessed: 10 December 2020.

[13] Forensic Science Regulator. Software Validation for DNA Mixture Interpretation, FSR-G-223 Issue 1. 2018. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/730994/G223_Mixture_software_validation_Issue1.pdf. Accessed: 11 September 2018.

[14] European Network of Forensic Science Institutes. ENFSI Guideline for Evaluative Reporting in Forensic Science. 2015. http://enfsi.eu/news/enfsi-guideline-evaluative-reporting-forensic-science/. Accessed: 3 March 2017.

313 [15] Taylor D, Buckleton J, Evett I. Testing likelihood ratios produced from complex DNA
314 profiles. Forensic Science International: Genetics. 2015;16:165-71.

PC11

NIST Chapter 5_Comments

KOVARI, JEANETTE (TROOPERS) [redacted]

Thu 7/29/2021 2:09 PM

To: Butler, John M. (Fed) <john.butler@nist.gov>
Cc: FITZPATRICK, MEEGAN (TROOPERS) [redacted]; PIZZIKETTI, JULIE (TROOPERS) [redacted]; LEE, MELISSA (TROOPERS) [redacted]; WHITE, THERESA (TROOPERS) [redacted]; HURBANEK, NICHOLE (TROOPERS) [redacted]

Good afternoon, Dr. Butler,

Theresa Myers, Melissa Lee and I were assigned to read Chapter 5: Context and Relevance Related to DNA Mixture Interpretation. We wanted to provide you with some comments/feedback:

- **Law Enforcement Elimination Databases:**

  The idea of law enforcement elimination databases is great, and this is something that we should be moving toward in the United States. For years, the forensic community has recognized that there is potential for contamination at crime scenes, just as there is potential for contamination within the laboratory. Forensic laboratories provide a significantly more controlled environment in which to prevent contamination, yet contamination events still occur. In chapter 5, the authors identify the risks associated with not having law enforcement elimination databases, such as misleading the investigations and the waste of time and resources. We found it interesting that the authors did not seem to address the newer investigational tools that are available, such as Familial Searching and Investigative Genetic Genealogy, and we feel it would be interesting if the authors were to address how time-consuming and/or costly it would be to pursue these avenues, only to discover that the profile in question was a result of contamination.

- **Relevance:**

  One of the central themes in Chapter 5 is relevance, but the overall point seemed to have gotten lost due to the large amount of information that was being provided. Ultimately, the authors recognized that the mere presence of a DNA profile does not address the *when* or *how* it was deposited at the scene. The forensic community recognizes that there is risk in LEOs and DAs misinterpreting the weight-of-evidence statistics but that does not negate that the weight-of-evidence statistics are independent of the profile's "relevance" to the crime. The emphasis should be placed on the fact that the DNA results are one of many investigative tools and that further investigation is required upon receipt of the DNA results.

  Further, lines 4513 through 4515 state that, "While the traditional view is to focus on the major contributor to a mixture based on the assumption that the profile belongs to the last person to handle an item, some studies have shown this is not always the case (e.g., Cale et al. 2016, Buckingham et al. 2016, Goray et al. 2016)". However, this seems to conflict with the message being conveyed in Key Takeaway #5.2, which states that "Highly sensitive DNA methods increase the likelihood of detecting irrelevant DNA. When assessing evidence that involves very small quantities of DNA, it is especially important to consider relevance." The job of the forensic examiner must be to report the findings and provide training on the significance of those findings. However, the responsibility of determining relevance should lie with the investigating officer(s).

  Lastly, we were interested in the concept of the suggested Case Assessment and Interpretation (CAI) method and we wonder how that might contradict with what we've been taught about cognitive bias. Does the issue really lie with forensic examiners being "findings-led", especially given that there are several requirements to which we must adhere to document the CODIS eligibility (or ineligibility) of a DNA profile.

- **DNA Mixtures and Sequencing:**

  Given all of the information provided in Chapter 5 and the emphasis placed on relevance, we would be curious to know whether the authors had reviewed any publications about DNA sequencing and whether there is a benefit to deconvoluting 5 and 6 person mixtures. Since secondary and tertiary transfer has been documented, would the forensic community really benefit from attempting to continue to increase sensitivity and pull more information from a profile?

- **Public Knowledge:**

  We can appreciate the benefit to publishing internal and developmental validations that individual laboratories perform and that more research is needed, more specifically on DNA transfer and persistence as encountered with DNA mixture profiles. We also understand that it is important for individual laboratories to share the sensitivity of the methods utilized so that other laboratories can gather and gain knowledge and information from these publications. We feel that it is most important for the laboratories to communicate better and work together as a forensic community. This can help to improve upon the studies performed regarding DNA mixture interpretation and that there is possible risk and vulnerability working solely as individual laboratories, generating varying studies/results and potentially publishing data that could be used against the forensic community. It would be best to develop a united front with the mixture validations to be performed and the methods to be established to identify limitations and gain consistency amongst laboratories.

Thank you and your team for your hard work on this book.

Regards,
Theresa Myers, Melissa Lee and Jeanette Kovari

PC12

From: Jessica Charak Lehrner
To: ScientificFoundationReviews
Cc: Kim Murga; Cassandra Robertson; Kellie Gauthier
Subject: Comments and Additional Data for DNA Mixture Interpretation: Scientific Foundation Review
Date: Friday, July 30, 2021 8:11:20 PM
Attachments: LVMPD Validation Mixture Schemes_3500 INV24_STRmix v2.6.xlsm
LVMPD_NOC for Deconvolution_3500 INV24_STRmix v2.6.xlsx
200220_Internal Validation of STRmix v2.6 with 3500xl.pdf
200220_STRmix v2.6 Parameters for Investigator 24plex on 3500xl.pdf
200220_PC of Degradation Max Setting in STRmix.pdf

click on attachments to view files

Good Afternoon-

Attached you will find the genotypes, mixture schemes, and degrees of allele sharing utilized by the Las Vegas Metropolitan Police Department (LVMPD) Biology/DNA Detail during the internal validation of STRmix v2.3.06 (QIAGEN Investigator 24plex QS, 28 cycles, ABI 3500xl). These mixtures include two-, three-, four-, and five-person mixtures with varying degrees of allele share to include full sibling familial relationships. Both of the biological parents of the three siblings independently exhibit a high degree of homozygosity; therefore, the full sibling DNA profiles are all very similar to one another, despite there being a several year age difference between them. In addition, purposely differentially degraded mixtures were created in order to conduct a performance check of the impact of an increase to the degradation max setting in STRmix (see attached summary). All mixture schemes and ratios were amplified to target five different total template inputs representing the high (1.5 ng), ideal (1 ng), low (500 pg and 200 pg), and very low (50 pg) points of the LVMPD's internally validated and characterized dynamic range. Additional mixtures were amplified in duplicate or triplicate at select inputs and ratios. Refer to the different tabs in the attached "LVMPD Validation Mixture Schemes_3500 INV24_STRmix v2.6.xls" for all profile details.

Each mixture was analyzed in GeneMapper ID-X v1.6 and assessed for its "apparent" number of contributors (NOC) as would be routinely performed during casework. In the event the apparent NOC differed from the ground truth NOC, both deconvolutions were run in STRmix in order to characterize the impact on the Hp true and Hd true LR results. This resulted in 380 deconvolutions being performed during the course of the validation. Please refer to pages 3-6 of the attached internal validation document for a summary of the deconvolutions and samples run and to "LVMPD_NOC for Deconvolution_3500 INV24_STRmix v2.6.xlsx" for documentation of the NOCs assigned and interpreted for each mixture scheme/template.

All studies were purposefully designed to explore a large "factor space" of complexity in way of number of contributors, mixture ratio, total template, quality of template, and degree of allele sharing that may be encountered during casework. Most importantly, this data was utilized to directly inform the development of the LVMPD protocols and ensure that the limitations of interpretation of complex DNA mixtures are appropriately addressed. These adopted protocols are publicly available at the following address: http://www.lvmpd.com/en-us/Documents/ForensicLabManuals/Biology_DNA%20Procedures%20Manual%20%28published%2006.10.2021%29.pdf and http://www.lvmpd.com/en-us/Documents/ForensicLabManuals/Biology_DNA%20Quality%20Manual%20%28published%2006.10.2021%29.pdf.

Notably, the following procedures were adopted as a result of the validation factor space testing:

- The LVMPD has elected to continue to utilize an uninformative range as a direct result of the Hp true and Hd true internal validation testing. When a reference standard comparison results in an individual falling into this range, the numerical likelihood ratio value is provided in the report; however, it is followed with the statement:
  "The likelihood ratio value noted above for <name and Item #> is within the uninformative range. Therefore, this does not provide sufficient support for whether <name and Item #> is included or excluded as a contributor to this sample. Internal validations and published studies help inform the limits of STRmix to where a false positive or false negative result may possibly arise. Likelihood ratios with exponents between 10^{-3} and 10^{3} have the potential to support a false inclusion or exclusion based on LVMPD internal validation studies."
- The uninformative range was knowingly expanded by an order of magnitude from the value originally identified during validation when considering the only the Hp true and Hd true testing of *unrelated* contributors. This expansion was necessary due to the recognition that false support may be given to the inclusion or exclusion of biologically related individuals or due to the possibility of transfer or persistence of background DNA.
- Though the 99% 1-sided HPD is routinely reported, the unified and untested biological relationship LRs are also routinely calculated for every profile that is eligible to do so (i.e., when there is only a difference of one unknown between Hp and Hd). This information is technically reviewed and maintained in the case file for instances in which the possibility of an untested relative becomes a question.
- The LVMPD requires that the unified LR be 10^{4} or greater for an individual in order to support the re-deconvolution of an environmental sample using conditioning (e.g., steering wheel, bedding, items recovered from domestic locations, etc.). This extra layer of conservatism takes into consideration the question of false support with untested relatives based on a shared environment or the presence of background trace DNA.
- A report disclaimer is included in LVMPD reports which addresses biological relatedness and the possibility for false inclusions:
  "In DNA mixtures of closely related individuals (such as parents, offspring, and siblings), false inclusions of other closely-related family members can occur due to the elevated sharing of genetic information between relatives."
- A report disclaimer is also included in LVMPD reports regarding the questions DNA profiles can and cannot answer:
  "The reported DNA profile results can aid in answering questions regarding who may have deposited DNA on an item of evidence and where this DNA was deposited. However, the presence or absence of a DNA profile cannot answer questions with regards to the timeframe and/or circumstances in which the DNA was deposited on an item of evidence."
  - We agree with the recommendation to more prominently display this disclaimer to ensure it is not hidden in a notes section (page 130-131; lines 4536-4542) and will seek to implement this change.

Our hope is that the data provided will be considered and included in Table 4.5 (page 73; lines 3069-3074) when finalizing the NIST Scientific Foundation Review document. The attached documentation provides all elements requested for a complete population of the columns of the table, to include Laboratory PGS (version) STR Kit, ABI CE, NoC Range, # Samples, Total DNA Quantity

Range (pg), and Mixture Ratio Range. In addition, all elements referenced on lines 2638-2641 of page 55 of the report are included in the attached documentation and support the degree of "factor space" explored by the LVMPD during internal validation.

Additional comment:
Page 133, lines 4618-4620 state: *"The LR, as typically used when interpreting DNA mixtures, is based only upon the analytical properties of the DNA. It does not provide information about other important aspects of the evidence, such as the quantity of DNA or whether the cell type is known."* (emphasis added).

- A portion of this statement is misleading. Likelihood ratios calculated using continuous probabilistic genotyping do provide information regarding the quantity of DNA present in the mixtures via its biological modeling of the template per contributor. The results of this modeling directly impacts the final calculated LR. This is an important and relevant distinction that must be made for continuous vs. semi-continuous or binary likelihood ratios.

Respectfully-
Jessica Charak Lehrner

Jessica Charak Lehrner, MFS, ABC-MB
DNA Technical Leader
Las Vegas Metropolitan Police Department
5555 W. Badura Avenue, Suite 120
Las Vegas, NV 89118
[redacted]
[redacted]

Note: Correspondence referencing cases may be retained as part of the Forensic Laboratory's case record and are subject to Information Disclosure Requests.

PC13

Comments for NISTIR 8351-Draft

Jennifer Thayer [redacted]

Thu 8/5/2021 12:54 PM

To: ScientificFoundationReviews <ScientificFoundationReviews@nist.gov>
Cc: [redacted]

To all involved in the Draft document NISTIR 8351, congratulations on a very thorough foundation review on DNA mixture interpretation. I'm sure this was quite an involved and time consuming undertaking!

Our laboratory has just a few comments, of which I would greatly appreciate your consideration:

1. There are privacy and other considerations in making validation data itself publicly available, so we respectfully request you reconsider that recommendation in Key Takeaway #4.3 and supporting paragraphs.

2. Throughout the document, the lack of publicly available validation data is discussed, along with the benefits of having that data publicly available for peer review and potential reasons labs do not make their validation data public. It is not mentioned (at least that I recall), that all US accredited labs have validation data reviewed by external auditors/scientists in two subsequent external QAS audits once the validation has been completed and the instrument/technique/software implemented. We feel a mention of this as well, perhaps in section 4.4.4, would alert the readers of the document that there are at least those two external peer reviews, even if a laboratory's validation data is not made available to the general public. A similar mention is made regarding training programs and QAS auditor review in Appendix 2 (lines 6955-6956).

3. Throughout the document, we feel the words reliable and reliability are used too broadly and sometimes in an improper context. For example, we are concerned that Key Takeaway #4.4 implies, especially to a layperson, that likelihood ratios for higher-order mixtures are inherently unreliable due to the increased complexity of those mixture types. While there is no disagreement that an LR will decrease/become less informative as mixture complexity increases, this should not be considered a reflection of the reliability of the LR. For example, wouldn't it be more accurate to use *informative* or *discriminating* instead of *reliable* in the parenthetical "results for a two-person mixed sample are generally expected to be more reliable than those for a four-person mixed sample..."?

4. We believe the second sentence should be removed from Key Takeaway #4.7 ("To enable users of results to assess the degree of reliability in the case of interest, it would be helpful to include these validation performance results in the case file and report."), as we do not feel it is necessary to include validation performance results, which could be rather voluminous, in case files or reports. This is not something our lab has ever done, and we are unaware of this practice in other forensic laboratories; our validation data is available for review upon request of interested parties. The first sentence seems quite sufficient for this key takeaway.

We thank you for your consideration of our feedback.

- Sincerely,

Jennifer Thayer, FS3
DNA Laboratory Technical Leader

NJ State Police Office of Forensic Sciences
[redacted]

PC14

Public comment response to: DNA Mixture Interpretation: A NIST Scientific Foundation Review

McGuckian, Amy B. [redacted]

Thu 8/5/2021 3:53 PM

To: ScientificFoundationReviews <ScientificFoundationReviews@nist.gov>

Comments and feedback prepared in response to NIST's draft document," DNA Mixture Interpretation: A Scientific Foundation Review." These comments and feedback were prepared by experienced, practicing forensic DNA analysts and their leadership.

The authors state that the purpose of this Review is to report on what is known about the limits of DNA mixture interpretation methods, to include the relevant probabilistic genotyping software. While the authors make a grand attempt in their 250-page document to summarize these limitations, the publication falls short of accomplishing this goal. Instead, the authors present a document that is not only academically biased, but unusable by the relevant community in its current state. The authors regularly present paradoxes as limitations; presenting an issue that needs to be addressed and then later stating that there is no way to address it. As such, the document provides little to no actionable feedback. It is clear that NIST is generous in pointing out what they see as inadequacies in DNA interpretation, but fall short in providing any direction as to how to address their concerns. Additionally, chapter 5 (Context and Relevance Related to DNA Mixture Interpretation) is a complete deviation from the intended scope of this report. While an important topic, in and of itself, and worthy of discussion, it has no relation to the limits of DNA interpretation methods and should be removed from this publication. During the review of this document, analysts were left asking "how" and "what would you have us do," so much so that the overall opinion was that this document is uninformative to the forensic DNA community. Officially published in its current state, the document is certain to only serve as fodder to undermine the hard work and advancements made by the forensic DNA community on the topic of DNA mixture interpretation.

Please see specific comments below that may be worthy of review and attention:

Chapter 3: Data sources and Information

3.1.2 – Lines 2480-2483 – The paper brings up valid concerns on why detailed validation data were not included in most public available information ("privacy concerns around releasing genotype information from individuals") however *still* goes on to critique the lack of publically available data for review on validations (Key Takeaway 4.4).

Chapter 4: Reliability of DNA Mixture Measurements and Interpretation

4.1.2 lines 2774-2779. state that a likelihood ratio (LR) does not involve comparisons to any reference standard and that there is "no uncertainty associated with an LR assumption." This statement is counterintuitive as in conducting an LR, the genotypes of the 'reference,' be it a true contributor or non-contributor, are assessed and compared to the weights of the proposed genotypes during the deconvolution process.
There are uncertainties that are considered when generating propositions and calculating a LR. The calculations have layers of conservatism built in to account for these uncertainties (distributing allele frequencies, theta values, several LR calculations). Uncertainty in LR assumptions is taught during training on LRs (assigning propositions) – how factors such as conditioning individuals, multiple contributors in the numerator, or running individuals separate can affect the LR.

4.1.4 – Factor space and Factor space coverage.

Table 4.1 outlines many aspects that should be considered for a validation (Factor Space). Most of these influencing factors are considered in validations if the laboratory follows SWGDAM validation guidelines and QAS validation standards. Every item listed on this factor space is generally covered.

Lines 2893-2894 state that when laboratories do examine factor space during their validations "it is unlikely that laboratories have explored every possible region of this factor space…". What the authors appear to be arguing is every exhaustive scenario for each factor space must be thoroughly examined in order to make the validation

reliable. Later the authors state “it is practically impossible to demonstrate reliability across the full extent of any factor space (lines 3194-3195).

4.3.3- - lines 3074-3075 key take away #4.3 “…not enough publically available data to enable an internal independent assessment of the degree of reliability of DNA mixture interpretation practices…” NIST makes no mention of what criteria should be used to determine how to assess reliability or what the thresholds for determining reliability are. No recommendations to who is qualified or acceptable to conduct such and independent review are provided.

4.4 Discussion 3234-3240……. No guidance or recommendations are provided on how to establish reliability or what in the factor space is enough to demonstrate such. Earlier parts of the paper describe aspects to consider, which many are, but the authors give no further guidance on how to achieve the “reliability” they propose, as the current assessment is validation are not meeting “reliability.” The authors postulate that because they can’t see all the laboratories’ data it does not exist.

Lines 3217-3222 – describe trend analysis as “sanity checks” in PGS systems and these do not qualify as “specific reliability of an LR number.” The previous definition of reliability in this paper was “trustworthiness established through empirical assessment.” Due to the nature of the MCMC process utilized in some PGS programs, it is known that the exact LR number will not be reproduced if the same sample is run again. This is accepted as part of the PGS modeling. The current empirical assessment that the software is working as expected is to run a range of samples and extrapolate trends. These trends provide context when assessing an overall DNA profile or LR. No other alternative is offered for assessment of reliability in PGS data.

4.4.1 Lines- 3280- Few studies have explored 4 or 5 person mixtures…… The article references PCAST but does not reference Bright et. al Internal validation of STRmix; a multi laboratory response to PCAST. Forensic Science International: Genetics, 2018 vol 34, pg 11-24… This paper reports the internal validation data from 31 laboratories using or validating PGS (STRmix) where 2825 mixture samples of various template amount, contributor number (mixtures of 3,4,5 and 6 contributors were specifically targeted), and ratio were examined. In terms of factor space, forensic laboratories cannot actually test every possible circumstance or replicate the exact circumstance of any case. This paper provides no guidance on the factor space needed to assess and determine reliability.

4.4.2 Lines 3347 – 3350. Authors argue that developmental validations should assess lower level minor contributors since instrument sensitivity has improved over the years for laboratories. However, this is considered in Internal validations. The purpose of a developmental validation is to provide optimized conditions for generating DNA profiles.

4.4.3 Line 3366-3370- may not be N.E.S (not expressly stated) but the information regarding the number of samples tested can be gleaned from the information provided in the validation summaries. There is a difference between N.E.S and not there. Who is an appropriate individual or outside body to conduct the review? Data acquired and tabulated into tables, figures, and charts can generate thousands of pages’ worth of data. To what extend is publically available feasible and where will the data be housed?

4.4.4 – Table 4.9 PT recommendation to “require more challenging PT samples containing low level, degraded DNA and mixtures with more than two contributors” – This comment does not consider the testing lab specific sensitivity of instruments/kits. No guidance on how to implement is provided.

4.4.5 – lines 3468-3474 Bracketing approach – the idea asks an analyst “to use ground truth data from known samples similar to the casework sample of interest and study the result.” This is in essence what a validation does and provides the context for a casework analyst to consider where the casework sample fits in the range of samples tested to inform the analyst’s interpretation. This approach to interpretation is provided during DNA analyst training.

Key take away 4.8 (Line 3594) - how can a scientific foundation review on the topic of likelihood ratios in forensic science and how LRs are calculated etc. be conducted when there are no established or proposed measures to evaluate this against?

Chapter 5: Context and Relevance Related to DNA Mixture Interpretation

As previously stated, this entire chapter should be struck from the document as it is completely irrelevant to the authors' stated purpose of reporting on what is known about the limits of DNA mixture interpretation methods. Relevance of DNA test results in criminal investigations is a completely unrelated topic to the one at hand.

Key Takeaway 5.3 (Line 4578): This discusses highly sensitive method increasing the likelihood of detecting contaminating DNA that might affect the investigation, but how is this "contamination" differentiated from random low-level DNA that may have been deposited prior to the crime event? Both sides are critiqued with no real answer on how to deal with such.

5.4.2.2 (lines 4596 – 4597, 4601 – 4604, 4610 - 4614): Analysts know this to be true, but are often told to not be influenced by contextual bias and shouldn't know certain details about the case. This contamination (if ever detected) would come out in the courts. Is it the lab's responsibility or does this fall into the narrative of each side in courtroom arguments?

5.4.2.4 (lines 4651 – 4663): NIST would seek to add a framework for considering case context during interpretation known as the CAI (Case Assessment and Interpretation). Are the readers to understand that NIST is asking analysts to conduct LRs on activity propositions to produce a conservative assessment of the weight of evidence that is more useful to the court (also lines 4751 – 4754)? This would introduce subjectivity while the community is trying to come to a more standardized approach and no doubt result in more uncertainty and evidentiary hearings. (See lines 4777-4779)

Key Takeaway 5.5 (line 4861): Transfer of DNA does not negate the value of DNA evidence, however, the value of DNA Evidence depends on the circumstances of the case. Is it the lab's responsibility or does this fall into the narrative of each side in courtroom arguments? Yes, a good argument, but fails to address the Court's and court participants' role.

5.4.3 (lines 4932 – 4934): If it is accepted that there is variation in transfer studies because results are affected by the processes employed by each lab – their extractions, detection, amplification kits and parameters, and interpretation methods. How can proficiency tests be 'more like case samples' with more complex mixtures, low level, degradation, etc. and be graded as accurate between laboratories? Even section 2.3.3 (p.31) discusses that complexity increases with NoC and increased sensitivity. The idea has merit in theory, but the practical aspect of such an ask hasn't been considered.

Chapter 6: New Technologies: Potential and Limitations

The authors convey that the forensic DNA community relies heavily on commercial suppliers and ready-made solutions which can lead to limitations in product development and advancement. What the authors don't address is the "why" this is the case for most forensic laboratories; most notably a lack of resources (staff, money, etc.). While forensic laboratories *can* perform developmental validations for methods, most forensic laboratories (at least in the U.S) lack a true "R & D" department, instead having to focus solely on the analysis of casework.

Additionally, the authors do not make mention of the existing and in-development software solutions that can assist analysts with mixture interpretation. For example: PACE, FaSTR DNA, DBLR, NOCIt, etc.

PC15

DNA Mixture Interpretation: A Scientific Foundation Review

Greg Hampikian [redacted]

Thu 8/5/2021 7:25 PM

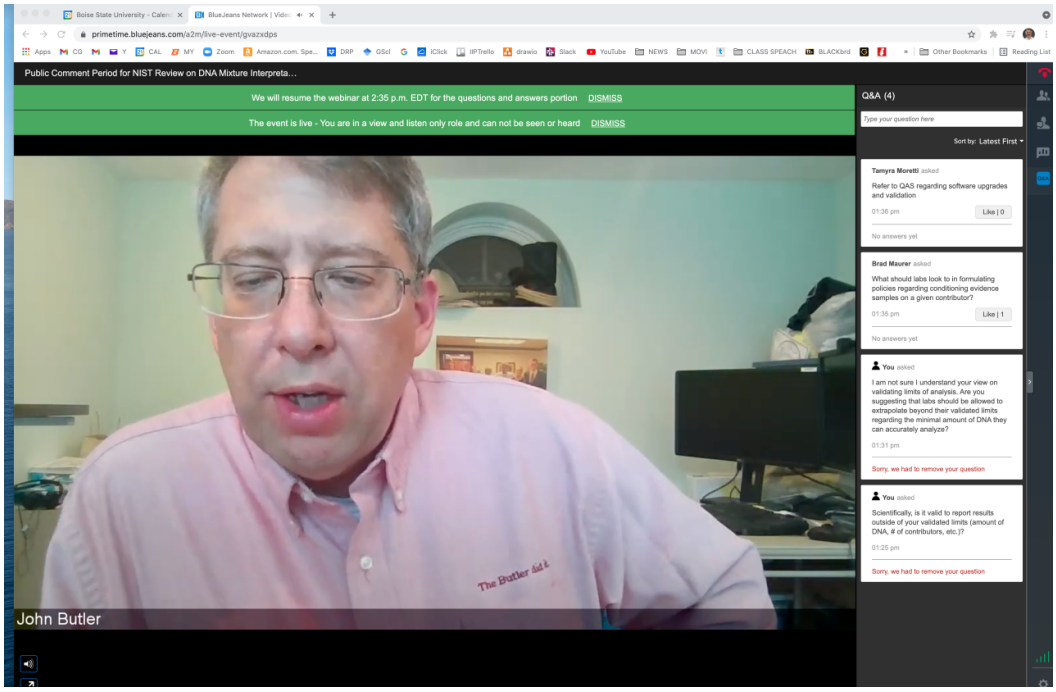
To: ScientificFoundationReviews <ScientificFoundationReviews@nist.gov>

1. The report and webinar do not support the foundational importance of validation. Labs must stick to their validated limits of interpretation, in terms of minimal amounts of DNA, number of contributors etc. Without proper validation and double-blind proficiency testing including samples that fall below validated levels, we are right where we started before this review: subjective, over extrapolative conclusions are still permissible. When I asked Dr. Butler about the necessity of staying within validated limits (screenshot below), he would only go so far as to say "It would be best," if labs stayed within their validated limits for mixture interpretation. I asked him twice, because I was perplexed. Is NIST abandoning the idea that validation limits are essential for forensic science? If labs can not determine their failure points, and establish appropriate limits in their Standard Operating Procedures, we will continue to have the terrible errors we saw in the Kerry Robinson wrongful conviction, and in the NIST Mix13 Problem 5 failures where more than 70% of accredited labs came to the wrong conclusion. These are both examples of the danger of not assessing and validating methods by identifying the failure points of those methods.

2. The high variance in statistics seen in interlaboratory studies such as Mix13 is not properly addressed in this review. Statistical values are used by jurors and judges to ascribe weight to evidence. They are usually given only the statistics calculated by a single lab. But we know from interlaboratory studies (like Mix 13) that accredited crime labs using the same data will calculate match statistics that vary by more than 12 orders of magnitude. The reliability and reproducibility of statistics in DNA mixture interpretation is foundational, since mixture interpretation is now entirely probabilistic. We need to include in all statistical reports information about the variance seen in interlaboratory studies.

3. The critical importance of probabilistic validation using genetic relatives must be stressed. In many forensic applications, likelihood ratios should include relatives of the accused in hypothesis testing. The validation of DNA mixtures from relatives needs to be thoroughly addressed, so that SOPs indicate under what conditions the methods become unreliable, or produce highly variable results.

4. Lack of Public Data, DNA Software Access It was encouraging to read and hear that NIST is clearly stating the need for public data from crime labs, without which defendants can't independently evaluate probabilistic genotyping findings. I would add that software used to evaluate data must be publicly available for lawyers and outside experts to explore and run alternative hypotheses. The gigantic expense of licenses for probabilistic genotyping programs, makes these "expert systems" impossible for most attorneys to effectively interrogate.



Greg Hampikian, Ph.D.
Professor of Biology, and Criminal Justice
Director of the Boise State Wastewater Virus Lab
Director of the Forensic Justice Project, Co-Director of the Idaho Innocence Project at BSU
[redacted] [redacted]
Links to papers and talks: http://biology.boisestate.edu/faculty-and-staff/faculty/greg-hampikian/
Office and Labs: Room SN-215, Science/Education, 2133 W Cesar Chavez Ln, Boise, ID
Mail: Biology Dept., 1910 University Drive, Boise, ID, 83725-1515

Idaho Innocence Project https://www.boisestate.edu/innocenceproject/
Follow us on Facebook https://www.facebook.com/IdahoIP/

PC16

Public Comment NISTIR 8351

Moore Boswell, Dawn [redacted]

Fri 8/6/2021 10:20 AM

To: ScientificFoundationReviews <ScientificFoundationReviews@nist.gov>

Observations on *NISTIR 8351-Draft, DNA Mixture Interpretation: A NIST Scientific Foundation Review*

First, thank you for endeavoring to undertake this crucial and daunting task. The reported observations were important and helpful. *NISTIR 8351 ["Draft 8351"]* identified many of the key questions surrounding areas of DNA analyses that may warrant further exploration. The compilation of information relating to some key DNA concepts in one accessible document could be of real value to both scientists and stakeholders.

While *Draft 8351* serves as a useful framework which appears to be consistent with NIST's published, articulated guidance for conducting these reviews, *NISTIR 8225: Scientific Foundation Reviews ["NISTIR 8225"]*, *Draft 8351* is not without its flaws. To begin with, the process employed to create this document has perhaps exposed a weakness in the adopted and/or as applied scientific foundation review process. Not coincidentally, some of the public comments offered in *NISTIR 8225* foreshadowed this current complication. *See generally*, *NISTIR 8225*, Public comments received on draft NISTIR 8225. Previously articulated public comments (summarized) suggested that:

- It could be difficult for NIST to make empirical evaluations without using unpublished/internal validation data, in-house data sets, and training materials (considered concurrently with comments that it would also be necessary to review the source and suitability of any used data);
- NIST should act as an online repository for reviewed data or provide a public online platform to ensure an open peer review process for foundation data; and,
- To be a complete and thorough foundation review, NIST would likely need to conduct its own intramural research, or use its already completed research, and/or acquire data it could access or request to supplement any gaps. Observations noted this as especially true considering the "casework-like" applications of many forensic disciplines.

The wisdom of these observations is now clear. As a result, NIST finds itself in the untenable position of attempting to answer whether some aspects of DNA are foundationally reliable without first having acquired or reviewed the information necessary to make such an evaluation. It is helpful that NIST specifically delineated the limitations of its own review -- clarifying that it relied on publications and information gleaned via "Google searches." *Draft 8351* at 50. Certainly, this is important contextual information for the forensic community and stakeholders to consider. The articulation of this limitation is also helpful in that it highlights some differences between NIST's review and the reliability assessments made by courts. In particular, stakeholders and judges typically access (or have the ability to access) laboratory specific validations and findings to aid in determinations of legal reliability. While the stated limitations of *Draft 8351's* are appreciated, it is still somewhat perplexing that *Draft 8351* would offer any assessments without first obtaining sufficient data.

Availability of Data for Conducting Review

Draft 8351 recites that a scientific foundation review seeks to "document and evaluate the foundations of a scientific discipline" and specifically answer the question of "what empirical data exist that speak to the reliability of the methods that forensic science practitioners use to analyze crime scene material?" *Draft 8351* at 14. *Draft 8351* conveys an intent to engage in this scientific foundation review by 1) identifying scientific principles; 2) reviewing the scientific literature; 3) gathering other empirical evidence from unpublished sources; and 4) collecting input from a group of leading forensic DNA practitioners and researchers. *Draft 8351* at 11. While *Draft 8351* identified scientific principles and reviewed selected scientific literature, it is unclear whether the gathering of "other empirical evidence from unpublished sources" component of its review was satisfied due to the focus on only data (published or unpublished) deemed "publicly available" or "reasonably accessible to anyone who wishes to review it." *Draft 8351* at 1-92, 185; 48.

Draft 8351 also emphasizes being constrained by the unavailability of data necessary to make complete assessments. In particular, it bemoans the lack of published data from laboratories on validation studies while simultaneously acknowledging that such information is not typically sought after or accepted publication material. Setting aside the fact that most laboratories have little opportunity to achieve the publication of such information, and may be restrained in some instances by data privacy concerns (as the report also notes), casework

laboratories have other operational demands that frequently position the publication of validation data as less pressing by comparison.

At a minimum, as a federal agency, NIST had the ability to request assistance from its federal laboratory colleagues. Additionally, NIST could have simply asked laboratories to voluntarily participate in providing data for review (along with offering any necessary data protections). There are other examples of laboratories "answering the call" to help clarify key issues. For example, Texas laboratories undertook the arduous task of reviewing DNA mixture interpretations and collaborating under the oversight of the Texas Forensic Science Commission. In one intra-laboratory STRmix™ study this type of laboratory cooperation likewise occurred. In fact, NIST itself has previously conducted studies with voluntary laboratory participation in order to gather needed information.

Simply put – it may have been more circumspect and supportive to the mission espoused by NIST had it just directly sought the needed information from laboratory sources. After all, having the actual *answers* to these important questions is what would best serve the needs and goals of both the forensic and legal communities – *whatever those answers may be*.

It is important to recognize (and give credit to) the articulated concerns and observations of *Draft 8351's* own consulted DNA Mixture Resource Group ("Resource Group"). *Draft 8351* at 17-18. The Resource Group's observations were insightful -- identifying the needs of the forensic community and salient questions demanding resolution. If NIST had addressed, or would be willing to address, those specific observations and offer a path forward in those articulated areas, it would result in immense progress. That Resource Group could act as an invaluable resource, offering concrete suggestions for NIST on how to gather the information needed to complete thorough scientific foundation reviews, as well as assisting in implementing concrete plans to conduct future assessments or create any required databases or sharing platforms.

Relevance

Draft 8351 found "relevance" to be one of the major challenges posed by DNA mixture interpretation. Observations regarding relevance seem to be an attempt to underscore the importance of understanding DNA limitations so that its use is appropriate. In particular, *Draft 8351* references "scientific" concerns regarding the potential for transfer and contamination. While not new considerations, these are arguably increased considerations due to more sensitive detection methods, more testing, and a wider variety of sample submissions. Unfortunately, in attempting to address these "relevance" assessments, *Draft 8351* appears to blur the lines between the realms of scientifically relevant information and legal relevance evaluations.

To begin with, it might be appropriate to consider whether it is even proper to address "relevance" in a scientific foundation review. Relevance of DNA evidence in the context of the legal setting necessarily relies on non-scientific facts and information often never known or provided to scientists. Addressing "scientifically" relevant factors without veering over into the "legal" relevance lane is inherently difficult. No matter how difficult, however, it is imperative to avoid conflating the respective roles. Accomplishing this task successfully takes very precise and detailed language, explanations, and cautions. While *Draft 8351* is replete with examples where those "lanes" are well defined, there are an equal number of instances throughout the document where the roles seem confused. *See and compare generally, Draft 8351 at* 38, 39, 44, 47, 63, 97, 132, 133. This confusion may have been exacerbated by the use of sometimes less precise language -- such as assigning dual meanings to the same terms/roles (user, provider) in some places while using different or more specific terms/roles (stakeholders) in other places. *Draft 8351* at 3, 5, 12, 13, 14, 18-19, 38, 39, 42, 44, 55, 59, 63, 64, 78, 82, 83-87, 89, 92, 94, 97, 135, 141. Even simply revising to use terms/roles uniformly and with clear intent may aid in alleviating some portions of this problem.

Perhaps, NIST might instead consider addressing this "relevance" component by emphasizing and impressing upon the scientific community the appropriate limitations of its role, the importance of communicating clearly with stakeholders, and the critical need for transparency in scientific and laboratory processes. Communicating with and educating stakeholders and triers of fact about potential "scientifically" relevant information directly impacts *the latter's* duty to assess *relevance*. Scientists must not be paternalistic or attempt to make these assessments independently. Science does not assess legal relevance – it provides information to those that do. This warrants additional clarification.

If NIST intends to include a component in this (or future) scientific foundation review(s) which involves legal considerations that may overlap with science considerations, it may wish to consider first consulting with legal participants familiar with the issues of forensic evidence in criminal cases on how to best communicate those interrelated conclusions in reports.

Call to Aid

Finally, for the legal community struggling with "critical decisions impacting life and liberty" that are "often based on the results of forensic analysis" it would be most meaningful if NIST would undertake to offer specific

assistance and direction that could actually serve to inform future work in the field. *Draft 8351* at 1. NIST could be enormously helpful to the community by developing and providing to laboratories a large pool of single source DNA standard reference materials that encompass relatedness, degradation and inhibition. Laboratories could then use the same standards to create mixtures to employ in various applicable scenarios. This could aid with consistency across laboratories; increase, streamline, and enhance robust validation work; and enable the public sharing of information while addressing any privacy, confidentiality, or legal implications associated with providing genetic information. Accepting responsibility for this task would be consistent with NIST's articulated purpose in conducting these reviews and well within NIST's area of expertise.

The real work of progress is difficult and complicated. Actually doing the real work, by providing tangible guidance and assistance on identified issues, is essential. NIST should continue to accept a national leadership role in this effort by first welcoming and truly considering sincere, constructive suggestions, followed by "rolling up its sleeves" to then help do that hard work.

Dawn Moore Boswell, J.D.
Director of Legal Forensics & Training
UNTHSC Center for Human Identification
3500 Camp Bowie, CBH 443
Fort Worth, TX 76107
[redacted]
Pronouns: she, her, hers

PC17

Comments on “NISTIR 8351-DRAFT DNA Mixture Interpretation: A NIST Scientific Foundation Review”

Alex Biedermann [redacted]

Sat 8/7/2021 10:43 AM

To: ScientificFoundationReviews <ScientificFoundationReviews@nist.gov>

Dear Writing Committee,

Please find, in the attached file, comments on the report "NISTIR 8351-DRAFT DNA Mixture Interpretation: A NIST Scientific Foundation Review", in reply to the call for public comments.

Best regards,

Alex Biedermann

_____

Alex Biedermann, Associate Professor
[redacted]
http://www.unil.ch/unisciences/alexbiedermann
http://www.unil.ch/forensicdecision

University of Lausanne
Faculty of Law, Criminal Justice and Public Administration
1015 Lausanne, Switzerland
www.unil.ch/esc

PC17a

Alex Biedermann, Ph.D., Associate Professor
University of Lausanne
Faculty of Law, Criminal Justice and Public Administration
School of Criminal Justice
1015 Lausanne (Switzerland)
[redacted]

By email to:
scientificfoundationreviews@nist.gov

Lausanne, August 7th 2021

Comments on "NISTIR 8351-DRAFT DNA Mixture Interpretation: A NIST Scientific Foundation Review" (National Institute of Standards and Technology, June 2021)

Dear Writing Committee,

Please find, hereafter, comments[1] on the above-mentioned draft report, in reply to the call for public comments.

Lines 1937–1951

Comments:
Figure 2.3 and the accompanying text in lines 1939–1946 can lead to misunderstandings. For example, representing a LR>1 in terms of a scale that tips towards the left could be misunderstood as meaning that the *case as a whole* is in favor of H_p rather than H_d. The problem here comes from the fact of "placing" probabilities in the scales. This is not how the use of the scale is conveyed in scientific literature, which is based on the logarithm. See e.g. Aitken et al. (2018): "When considering the scales of justice it is the logarithm of the probabilities of the evidence given each of the two competing propositions that should be put in the scales, not the probabilities."

For more details, see also:

evidence. The logarithm of the likelihood ratio has the pleasingly intuitive operation of additivity when converting the logarithm of the prior odds in favor of a proposition to the logarithm of the posterior odds in favor of the proposition.

$$\log\left\{\frac{\Pr(H_p \mid E)}{\Pr(H_d \mid E)}\right\} = \log\left\{\frac{\Pr(E \mid H_p)}{\Pr(E \mid H_d)}\right\} + \log\left\{\frac{\Pr(H_p)}{\Pr(H_d)}\right\}. \quad (1)$$

When considering the scales of justice it is the logarithm of the probabilities of the evidence given each of the two competing propositions that should be put in the scales, not the probabilities. Equation (1) can be rewritten as

$$\begin{aligned}&\log\{\Pr(H_p \mid E)\} - \log\{\Pr(H_d \mid E)\} = \\ &\quad \log\{\Pr(E \mid H_p)\} - \log\{\Pr(E \mid H_d)\} + \log\{\Pr(H_p)\} - \log\{\Pr(H_d)\} \\ &= [\log\{\Pr(E \mid H_p)\} + \log\{\Pr(H_p)\}] - [\log\{\Pr(E \mid H_d)\} + \log\{\Pr(H_d)\}]\end{aligned}$$

Expressions to the left of the negative sign in the last line are associated with one pan in the scales, expressions to the right with the other pan. Thus $\log(\Pr(E \mid H_p))$ is added to the prior log probability for H_p in one scale and $\log(\Pr(E \mid H_d))$ is added to the prior log probability for H_d in the other scale. The difference in the sums of the two pairs of log probabilities is a more intuitive characteristic of the evidence to which the term *weight* may be applied than the ratio of the probabilities of the evidence given the respective propositions.

C. Aitken et al., Commentary: Likelihood Ratio as Weight of Forensic Evidence: A Closer Look, Frontiers in Genetics, June 2018, Vol. 9, Article 224, p. 2, doi: 10.3389/fgene.2018.00224.

[1] The comments in this letter are of the author alone and do not represent the views of the University of Lausanne.

Lines 2776–2779:
"Assertions have been made that there is no true LR (e.g., Steele & Balding 2014, Gill et al. 2018). Some even hold the view that there is no uncertainty associated with an LR assessment (Berger & Slooten 2016; see also Biedermann et al. 2016a, Curran 2016, Morrison & Enzinger 2016, Taylor & Balding 2020)."

Comments:

- The choice of the terms "assertions" and "view" is unsuitable here because these terms suggest that the respective positions (i.e., no true LR and no uncertainty about the LR) are merely opinions. Instead, they are positions based on arguments drawn from foundational fields, such as statistics.
- The term "even" is unsuitable here because it is a subtle suggestion that the advocated position (i.e., that there is no uncertainty associated with a LR) is marginal and/or lacking credibility. Again, as mentioned above, the contrary is the case: the advocated position is based on arguments drawn from foundational fields, such as statistics.
- The presumably first paper in forensic literature providing a detailed rejection of the idea of "uncertainty" associated with an LR is: Taroni F. et al. (2016), Dismissal of the illusion of uncertainty in the assessment of a likelihood ratio, Law, Probability and Risk, 15, 1–16. For completeness, this paper should be added here as a reference.

Line 4469:

Addressing propositions	Biedermann et al. 2016, Hicks et al. 2015, Gittelson et al. 2016, Kokshoorn et al. 2017, Taylor et al. 2017d

Comment:
The references given here do not deal with "addressing propositions", because scientists do not address (i.e., opine on) propositions. Scientists address findings (results) *given* propositions, never the contrary. What the references deal with is the question of *how to define* propositions. Thus, "Defining propositions" or "Definition of propositions" should be written here instead of "Addressing propositions".

Lines 4770–4773:
"Some researchers have argued that, in that case, it would be appropriate to assign "subjective probabilities" (Biedermann et al. 2016a, ENFSI 2015), while others have argued that this would not be appropriate (Meakin & Jamieson 2013)."

Comment:
Opposing the references (Biedermann et al. 2016a, ENFSI 2015) and (Meakin & Jamieson 2013) is unsuitable here because the latter authors do not understand "(subjective) probability" in the same way as the former authors.
Meaking & Jamieson (2013) assert that the published data establish "the possibility, but not the probability [of transfer]" (p. 442), which refers to so-called probabilities of causes. These are, by definition, beyond the scientists' area of competence and not what Biedermann et al. 2016a

and ENFSI 2015 refer to. Scientists do/must not opine in terms of a probability that a given event transfer rather than another has occurred. See, for example, Recommendation 3 of the document "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" (Gill et al., 2020, Forensic Science International: Genetics, vol. 44, 102186):

> "Scientists must not give their opinion on what is the 'most likely way of transfer' (direct or indirect), as this would amount to giving an opinion on the activities and result in a prosecutor's fallacy (i.e. give the probability *that* X is true). The scientists' role is to assess the value of the *results* if each proposition is true in accordance with the likelihood ratio framework (the probability of the *results if* X is true and *if* Y is true)." [at p. 4]

What Biedermann et al. 2016a and ENFSI 2015 mean are probability of effects: i.e., the probability of observing DNA *given* varying alleged activities. These authors acknowledge that:

(i) probability assertions are based on more than experimental data alone,
(ii) that there is a hierarchy of data types, and
(iii) that scientists can use data to varying extents (i.e., information to be drawn from data is not a discrete "all or nothing" issue, but a function of the quality and quantity of data).

In particular, ENFSI (2015) emphasizes that probability assessments are transparent and based on arguments (i.e., justifications) that clarify the extent to which assigned probabilities are informed by data. Consider the following:

- "The **basis** for these assignments shall be **documented** on the case file. Relevant and appropriate published data will be used wherever possible. If appropriate published data are not available then data from unpublished sources may be used. **Regardless of the existence of sources** (published or not) of numerical data, personal data such as experience in similar cases and peer consultations may be used, provided that **the forensic practitioner can justify** the use of such data." (at p. 15; **emphasis added**)
- "Such personal probability assignment is not arbitrary or speculative, but is based on a body of knowledge that should be **available for auditing and disclosure**. The forensic practitioner should not mislead the recipient of expert information as to the basis of the personal assignment, and the extent to which the assignment is supported by scientific research." (at 16; **emphasis added**)

On the notion of "*justified* probability assessments", see also: Biedermann A., Bozza S., Taroni F., Aitken C. 2017, The meaning of justified subjectivism and its role in the reconciliation of recent disagreements over forensic probabilism, Science & Justice (Virtual Special Issue "Measuring and reporting the precision of forensic likelihood ratios"), 57, 477-483, doi.org/10.1016/j.scijus.2017.08.005.

In summary, thus, there is *no* suggestion in ENFSI 2015 and Biedermann et al. 2016a to assert probabilities *without* justificatory grounds.

Best regards,

Alex Biedermann, Ph.D.

PC18

Pre-publication comments for Review of DNA Mixture Interpretation Methods

Colleen Spurgeon [redacted]

Mon 8/9/2021 4:57 PM

To: ScientificFoundationReviews <ScientificFoundationReviews@nist.gov>
Cc: Jeanette Wallin [redacted]

After reviewing the draft document and seeing the presentation by John Butler on July 21, 2021, we know many in the forensic community would like to make comments and suggestions before the document is finalized. Given the length of this publication, its release in the summer months, and the short time interval between the presentation and the proposed end of public comments, we would suggest an extension to the review period, to allow more organizations and individuals to provide comments. (We see it has now been extended so thank you for that.)

Due to the short time frame, we will just focus on three items for our comments today. Please note that within this letter we have added underlining to some quoted paragraphs or sentences for emphasis.

Item 1. The authors need to be more clear and consistent about what this review is actually reviewing.

The authors start the document by stating in the preface (line 125) "Congress has appropriated funds for NIST to conduct scientific foundation reviews in forensic science. These reviews seek to answer the question "What established scientific laws and principles as well as empirical data exist to support the methods that forensic science practitioners use to analyze evidence?"

After listening to John Butler's presentation on July 21, 2021 and reviewing the draft document, it appears that this is not what they actually did. They instead looked at – What established scientific laws and principles as well as empirical data exist in the public forum to support the methods that forensic science practitioners use to analyze evidence?

The authors are clearer about what they actually reviewed in some places in Chapter 4. For example they state (line 2961), "As part of our assessment of the foundations of DNA mixture interpretation methods and practices, we examined factor space coverage in published articles describing STR kit developmental validation, PGS validation data, publicly available PGS internal validation summaries, ….", but unfortunately that language is not consistent, even within this chapter. Just one page before this it states (line 2912), "Thus, in this scientific foundation review we assess what information and data are available, what portion of the factor space this information and data cover, and what can be learned about reliability of DNA mixture interpretation from the available information and data". Clearly what they really mean is "what information and data are available in the public forum, and "what can be learned about reliability of DNA mixture interpretation from the publicly available information and data".

The authors noted that not all of the "factor space" information they wished to see was readily available in the 11 internal validations they found via "internet searches", but it appears they did not reach out to the labs involved to see if this data could be made available. They do not mention why they did not reach out to the labs (if they did reach out, that should be clarified in the review). This is particularly confusing because the information noted as missing in Table 4.5, or more specifically "not explicitly stated", such as "# of samples" or "total DNA quantity range" is information that labs would likely have gladly provided in this context.

Given NIST's resources and capabilities, it is unclear why no apparent effort was made to gather and evaluate the empirical data that does exist (granted not via internet searches). By publishing this review before making the effort to locate and truly evaluate the empirical data that does exist, NIST is doing unnecessary and potentially lasting damage to the forensic community.

One further point on this topic. When the authors note how many internal validations they found (line 3066) it reads the "eleven publically available internal validation summaries that could be found when these searches were performed". It would be helpful if they document when they conducted the search. They document the dates in Chapter 3 when they discuss this search, but it should also be documented with Table 4.5, where the information they collected is reported. In a rapidly changing area such as Probabilistic Genotyping, even a year can make a big difference in the number of articles or validations that are available, and the reader should be able to access how relevant the search that is referenced might be at the time of reading.

Item 2. Please provide specific details on what data you believe labs should make available and how they can safely do so. Why rush to publication without compiling and reviewing the data first?

If one of the goals is to access the degree of reliability of probabilistic genotyping, by the authors own assessments, the fact that they could not do this simply by reviewing published or publicly available data, should not have come as a surprise.

First, they apparently had not set any criteria by which to do so. (Line 3201) "Based on an examination of publicly available information reviewed during the time frame of this study, there is not enough information for the authors of this report to independently assess the degree of reliability of DNA mixture interpretation at any one point in the factor space. This is particularly true without an established and accepted criteria for reliability with complex mixtures involving contributors containing low quantities of DNA template or where there is a high degree of allele overlap among contributors"

Second, they either already knew or discovered early in review process that (line 3227) "However, to independently assess the degree of reliability of PGS models, metadata associated with specific sample results and the corresponding specific log(LR) value datapoints are needed. Data of this nature are not generally shared in publications or validation summaries."

At this point, instead of publishing a review that does not move forensic science forward, why not act as a resource for the community, and take some of the steps presented in the PCAST report and referenced in this review? They quote the PCAST report (line 3274), "The path forward is straightforward". As part of a **research facility**, NIST could follow the advice from the PCAST report stating, "The validity of specific [probabilistic genotyping] software should be validated by testing a diverse collection of samples within well-defined ranges", or even "extend the range in which scientific validity has been established to include more challenging samples" (line 3260). As the authors acknowledged, most forensic labs do not have the resources (nor typically the mandate) to do this kind of extensive research, so who better than NIST to do it?

Short of doing research, as the authors note in several locations throughout this review, laboratories have accumulated data while conducting internal validations of PG software. They note in KEY TAKEWAY #4-3 "….we encourage forensic laboratories to make the underlying PG validation data publicly available…" Why not request some of this data from labs, and then write this review based on that data?

If the authors believe that the data needs to be in the public forum, NIST could create a confidential repository and procedure for labs to share the specific data deemed necessary to assess the foundational validity while protecting the privacy rights of the donors (e.g., per GINA and HIPAA). As noted in Item 1, even reaching out to the 11 labs whose internal validation summaries they did have access to, would likely yield data that they could then evaluate.

Item 3. Please remove unnecessary, antagonistic language from this review

At least four areas within Chapter 4 come across as antagonistic to the forensic community.

For example (line 3436), "One explanation for this lack of public data is simply that there has been no expectation to provide it. Choosing not to make public the data underlying decisions that are made in laboratory protocols is generally without consequence, while giving public access carries a risk of increased scrutiny".

I do not see the value of adding the last sentence. Public labs and those who work in them are used to having their work scrutinized, as they undergo external audits at least every two years, and preparing discoveries of casework, protocols and procedures prior to court is a common occurrence. If the aim of this review is to add value, the first sentence should either stand-alone or could instead read "One explanation for this lack of public data is simply that a there has been no expectation to provide it and/or no forum to do so".

As stated earlier, this is where NIST could step in and provide a service. If the authors believe that more of the internal validation data needs to be available for review, please provide a forum for labs to do so, and clarify who would do this review. Without a specific list of what data is needed, and providing a way for labs to share that data, NIST is missing an opportunity to move forensic science forward.

Another example of language that is inappropriate is (line 3322) "Internal validation studies provide an opportunity for the user (e.g., DNA analyst) to understand performance of a method in their forensic laboratory environment rather than trusting the provider's (e.g., the software developer) claim that everything works fine." The last part of the sentence is unnecessary, and insulting to both the providers of the tools forensic scientists use and to the forensic labs that employ those tools. The authors could either end the sentence at "in their forensic laboratory environment." or add "in their forensic laboratory environment and have confidence that the method functions as intended/expected."

Developmental and internal validations are two sides of the same coin, and the developers and the forensic labs understand that both are necessary. All NDIS labs are audited to QAS standards, which include standards regarding both developmental and internal validations. Forensics labs use internal validations to understand the new technology and its limitations, to see how it works with their analysts and within their system, and to establish protocols and training programs, so to suggest that the labs would forgo this critical evaluation for any reason, is inappropriate.

The authors include yet another example of antagonistic language (line 3270) in stating "what is actually happening in casework settings". This implies some sort of hoodwink activity is occurring. Unless there is clear evidence of such activity, this statement requires rewording.

Additionally, footnote 23 implies lack of cooperation among the forensic community with the authors: "The willingness of journals to publish validation studies is a separate issue from the willingness of laboratories to make data available on their website for anyone to download or at least sharing full data sets with credible parties in a timely manner when requested." This footnote implies some level of stonewalling by the forensic community when this document does not make clear that data from the community was ever requested. Rather, it states that it **only considered data publically available**. This footnote should simply be removed as it provides no insight to the reader and nothing in the way of moving the field forward.

Colleen Spurgeon
Assistant Lab Director
CA DOJ, Jan Bashinski Laboratory

Jeanette Wallin
Assistant Lab Director
CA DOJ, Jan Bashinski Laboratory

PC19

Mixture Interpretation Review comments/questions

Christina De La O [redacted]

Mon 8/9/2021 8:24 PM

To: ScientificFoundationReviews <ScientificFoundationReviews@nist.gov>

Good evening,

Here are some of the questions I'd like to pose to the open forum for the Scientific Foundation Review for DNA Mixture Interpretation:

1) The majority of examples used in the reviews from 2009, 2015, and 2021 deal with stain samples. Is the mixture interpretation more complex when an alternate extraction method is used (like that from tissue, bone, or tooth)? Additionally, will the lab need validation studies indicating a stain technique used with on tissue allows for complete extraction of the DNA? Has the NIST explored the impact of labs utilizing incorrect extraction methods with various samples?

2) The NIST states the crime sample is to be extracted and amplified prior to testing the reference sample. Once markers are determined, and show a severe overlapping of alleles, should a kinship analysis be conducted to determine relatedness of the individuals?

3) It seems that validation studies are the gold standard to whether a lab can substantiate their DNA testing is following the scientific standards. What stronger language can be used to express the importance of a lab's cooperation and open communication with the scientific community to ensure that bad science does not continue to reach our courthouses.

Thank you for your time,
Christina De La O

PC20

NISTIR 8351-DRAFT comments

Dr. Mark W. Perlin [redacted]

Mon 8/9/2021 11:28 PM

To: ScientificFoundationReviews <ScientificFoundationReviews@nist.gov>

Dear NIST,

Attached are my comments on the NISTIR 8351-DRAFT report entitled “DNA Mixture Interpretation: A NIST Scientific Foundation Review”. Thank you.

Kind regards. - Mark

==========

Mark W. Perlin, PhD, MD, PhD
Chief Scientific and Executive Officer
[redacted]

Cybergenetics
Omega Building, Suite 210
160 North Craig Street
Pittsburgh, PA 15213
USA
412.683.3004
412.683.3005 FAX
www.cybgen.com

PC20a

Comments on NIST Report NISTIR 8351-DRAFT
"DNA Mixture Interpretation: *A NIST Scientific Foundation Review*"
https://doi.org/10.6028/NIST.IR.8351-draft

Mark W. Perlin, PhD, MD, PhD
Cybergenetics, Pittsburgh, PA
August 9, 2021

Introduction

I invented reliable genotyping of mixed or low-level deoxyribonucleic acid (DNA) evidence over twenty years ago [1]. My computer-based "probabilistic" genotyping (PG) approach has been widely adopted and tested by the forensic science community. Unlike ineffective human-based "deterministic" mixture interpretation methods, Cybergenetics' TrueAllele® PG system is accurate, objective, impartial, validated, reproducible, sensitive, specific, automated, and thorough. TrueAllele computing can use all short tandem repeat (STR) data, and consider all genotype combinations to explain the data, statistically accounting for polymerase chain reaction (PCR) amplification variation and artifacts.

A DNA forensic group at the National Institute of Standards and Technology (NIST) has issued a draft Internal Report (NISTIR 8351-DRAFT) on DNA mixture interpretation that claims to be a "scientific foundation review." It is not. The NISTIR is philosophically wedded to the old deterministic mixture interpretation paradigm based on human data review, and entirely misses the point of automated probabilistic genotyping. It is unfortunate that NIST has wasted public resources pursuing the wrong issues. By chasing phantoms, NIST undermines the scientists who understand this modern science, the practitioners who use it in their work, and the courts that need the DNA results to render fair criminal justice decisions.

This brief response first reviews PG and the likelihood ratio (LR), providing an accurate general framework missing from the NISTIR. I then discuss validation and error rates, crucial ideas for testing and reporting PG results that the NISTIR mangles or omits. Next, I review the scientific and legal standards for reliability, as they apply to PG and LRs.

The second half of this response begins by describing why ineffective "deterministic" misinterpretation of DNA mixtures cannot produce accurate LR results. Unfortunately, this antiquated world view lies at the heart of the NIST report's approach and definitions. I then comment on some of the NISTIR "key takeaways," highlighting where NIST goes wrong. Finally, I state conclusions, and offer some recommendations for NIST.

Probabilistic Genotyping

The PG method can help identify a suspect who left their DNA in a mixed or low-level DNA evidence sample. PG can also help determine if someone did not leave their DNA in the

evidence, which facilitates exonerations. This genotyping approach can be described in three main steps.

1. Produce DNA data that has PCR amplification variation artifacts

Suppose we have a set of given genotypes. These genotypes can be represented as a matrix **G**. Suppose the relative amounts of the genotypes are the vector weights **w**. Then combining **G** with **w** will produce a perfect pattern **G*w** that describes genotype allele amounts before PCR amplification.

PCR is an imperfect DNA amplifier that introduces distortions. The final post-PCR product allele amounts will differ from the pre-PCR pattern **G*w**. The variation of the PCR random process can be represented as a random error vector **e**.

The observed electropherogram (EPG) data vector **d** records relative allele sizes and amounts. The data **d** is the sum of the pre-PCR pattern **G*w** and the PCR variation **e**. That is [1], reading the equation from right to left, EPG data is formed from mixed genotypes as:

(1) $\mathbf{d} = \mathbf{G}*\mathbf{w} + \mathbf{e}$

2. Compute accurate genotype probability from the data using PCR variance

The genotyping task is to find the genotypes present in the DNA mixture, based on the data, ideally one for each contributor. That is, starting from EPG data **d** on the left of equation (1) above, proceed rightward to find the genotypes **G**. Given the PCR variation, more than one genotype value might be possible. Indeed, to solve the problem with Bayesian methods, every possible posterior (i.e., based on having examined the data) genotype value will be assigned some probability. The probabilistic solution will also determine the PCR variance in error term **e**, as well as the contributor weight **w** probability distribution.

What is key here is to recognize that PCR introduces a random element **e** into the laboratory experiment. The idealized genotype combination pattern **G*w** will not be the same as the observed data **d**. Rather, the starting point for our genotyping analysis is the actual data **d**, which has a random **e** deviation from the idealized pattern **G*w**. All the EPG data **d** must be used, without change. Since the genotype answer **G** is a probability distribution $q(\omega)$, all possible genotype values $\omega \in \Omega$ must be considered. (ω is one genotype value. Ω is the set of all genotype values.)

3. Identify a suspect from the DNA mixture using genotype probability

The support in the data for a suspect having left their DNA in the DNA mixture is quantified as a LR measurement. The greater the LR value is above 1, the more support there is for the hypothesis H that the suspect left their DNA. The smaller the LR value is below 1, the more support there is for the alternative hypothesis that it was someone else.

In most of science and mathematics, the LR of a hypothesis for data **d** is a ratio of two probabilities. For DNA mixture evidence **d**, at suspect genotype ω, the LR of the identification hypothesis is the *posterior probability* $q(\omega)$ after having seen the data, divided by the *prior probability* $p(\omega)$ before seeing the data. There are other LR formulations, all of which involve a ratio of probabilities, and produce the same numerical LR result [2].

The prior genotype probability $p(\omega)$ of coincidence is based on how often STR alleles occur in a human population. The posterior genotype probability $q(\omega)$ is found by PG software analysis of the evidence data. The PG software separates out each contributor genotype from the mixture, calculating $q(\omega)$ for every possible genotype ω. The LR is then calculated at the suspect's genotype ω as the numerical ratio $q(\omega)/p(\omega)$. To encompass the full range of possible LR values, whether tiny or huge, scientists often use a logarithmic scale, reporting the DNA match statistic in base ten as $\log_{10}[q(\omega)/p(\omega)]$.

Validation and Error Rates

Imagine one could compare an evidence contributor genotype q with every possible genotype reference ω, considering everyone who did not contribute their DNA to the evidence. We would obtain the logarithmic LR distribution for all possible noncontributors [3]. This noncontributor distribution is another representation of a probabilistic genotype q, showing the range of LR values the genotype can produce when compared with references.

The bell-shaped log(LR) distribution provides a frequency context for understanding a DNA match statistic. When the LR value lies in the middle of the bell curve, that match strength is not unexpected. But when it lies far to the right of the curve, then the LR is a rare event. The area under the curve to the right of the LR value supplies an error rate. This "probability of misleading evidence" (PME) is the chance that someone not in the DNA would match as strongly as the suspect. Jurors generally understand error rate probabilities better than LR measurements.

The specificity curves constructed in PG validation studies are simply an average of the noncontributor distributions for the component genotypes in the study [3]. That is, the LR distributions are an inherent property of the software-derived evidence genotypes, fully determined before any comparison is made to a suspect or another reference genotype. Since the point of specificity determination is to establish statistical error rates, no reference genotype is needed. In particular, "ground truth" studies (that require such reference genotypes) are entirely irrelevant to empirical validation and error rates.

Scientific and Legal Reliability

Scientific reliability is established by empirically testing software on input data. Validation studies test groups of genotypes, developed from either laboratory or casework samples. The LR distributions derived from those probabilistic genotypes are sufficient to establish the error

rates needed in validation studies. Hundreds of thousands of such sample DNA mixture genotypes have already been developed. Using probability and error rates, it is easy to determine the accuracy, sensitivity, specificity, reproducibility, and other properties of PG software systems through empirical genotype testing. No "known" answer is needed.

The predominant American legal reliability standard is *Daubert*, which has five prongs. First, PG systems are extensively *tested* by developers and users. The *error rate* of these systems is found using the LR distributions that are inherently part of the probabilistic genotypes produced. There has been extensive *peer review* publication of PG methods and their validation. *Standards* have been established by national groups for validating PG systems. The PG methods are *generally accepted* by the relevant scientific community of PG scientists and practitioners. TrueAllele has been found to be reliable in thirty admissibility challenges, primarily based on extensive validation studies and use of error rates.

Deterministic Misinterpretation

Early attempts to interpret DNA mixtures tried to extrapolate methods from simplistic EPG manipulations of single source sample data peaks. Such methods – peak height thresholds, stutter fraction cutoffs, heterozygote imbalance – involved throwing away crucial data needed for accurate genotyping. Human review methods tried to simplify the mixture problem by reshaping the data, making it conform to idealized models of what mixture data "should" look like. For most DNA mixtures, this simplification failed to give an answer [4]. The result was "inconclusive" DNA evidence lost to criminal justice.

This human review took a deterministic view. In the absence of PCR allele distortion, what would analysts think the allele peak heights should look like? How could they discard EPG data and genetic loci to make people feel more "comfortable"? This inflexibility ignored what the data were saying – that PCR variation held the key to accurate genotyping and match statistics. Policy groups like NIST got it wrong, ignoring the science along with the data. They threw away the PCR key by running away from variation, instead of embracing it.

NIST conducted an interlaboratory mixture study in 2005, but didn't publish a validation study. Instead, in 2010, NIST's Dr. Butler enforced stringent "stochastic thresholds" on the forensic DNA community. This second threshold discarded more STR data, losing even more LR information, without solving the real problem. Only PG – which measures the variation of the PCR amplification, and then uses that variance to calculate accurate genotype probabilities – could deliver accurate LR values for DNA mixtures.

NIST's understanding has not progressed beyond deterministic human review. They do not see PCR variation as a powerful tool. Uncomfortable with probability, NIST tries to cast data **d** into idealized **G*w** genotype patterns. When the error term **e** gets large, they advocate discarding the DNA evidence. But, in fact, all data are useful. Larger PCR distortion just means more data variation, leading to more genotype possibilities and lower LR match statistics. That's science.

NIST Report Definitions

Factor space

NIST proposes that DNA laboratories explore the entire "factor space" before using their PG software on similar cases. But science only requires sampling of representative test conditions. First, continuity of continuous parameters (e.g., mixture weight) and substitutability of discrete parameters (e.g., genotypes) ensures that far fewer tests are needed. Second, randomized sampling of genotypes and mixture weights covers the test conditions well [5, 6]. Third, casework studies [7, 8] and casework practice provide sufficient data for PG validation studies via LR distributions and error rates.

An analogy may help. Is a particular car make and model reliable? The manufacturer extensively tests the car before making it available for sale. Independent testing groups (e.g., Consumer Reports) release their findings. The consumer tries out the car by driving it. But nowhere does the government require the owner to test out every possible driving condition (road surface, visibility, speed, curves, rain or snow or sunshine, temperature, humidity, number of passengers, age of driver, weight of car, engine RPM, standard or automatic transmission, type of brakes, number of pistons, body composition, rear camera, city vs. country, highway vs. local road, etc., etc., etc.) before letting them take their car on a similar road. That would be ludicrous. As is NIST's resource-intensive "factor space" proposal.

Ground truth

The "correct answer" in forensic science is what can be inferred from the evidence. The "correct answer" resides in the DNA data, and cannot be "known by design." NIST's incorrect perspective has led them astray in their understanding of PG systems and how they work. The correct genotype answers are probability distributions and error rates, not some "known" answer.

Likelihood ratio

The LR is the ratio of two probability distributions, evaluated at a particular point. For example, in Bayesian PG, the LR is the ratio of posterior to prior genotype probability, evaluated the genotype of some person. Other formulations (e.g., the ratio of two likelihood functions) may aid computation, giving the same numerical LR value. The log(LR) is a standard measure, summarizing the information that evidence data contains in support of some hypothesis.

PCAST

NIST's citing the PCAST report is disingenuous. PCAST's PG recommendations drew their main support from NIST author Dr. John Butler. Moreover, PCAST advised authorizing fourteen million dollars in NIST funding for Dr. Butler's unnecessary "factor space" studies. NIST has a

serious conflict of interest with PCAST that their NISTIR report neglected to disclose. Any PCAST "recommendations" involving NIST or PG should be ignored as self-serving.

Key Takeaways

> *KEY TAKEAWAY #2.1: DNA mixtures, where the DNA of more than one individual is present in a sample, are inherently more difficult to interpret than single-source DNA samples.*

No, they are not. When using effective PG computing, all DNA samples entail the same level of effort – enter the data, run the software, output the result.

> *KEY TAKEAWAY #2.2: Generating a DNA profile involves measuring the inherent physical properties of the sample. Interpreting a DNA profile involves assigning values that are not inherent to the sample. To do this, the DNA analyst uses their judgment, training, tools (including computer software), and experience, and considers factors such as case context.*

The physical properties are indeed "inherent" to the PCR amplification and DNA fragment detection process that generates STR data. PG computing accounts for these properties by probability modeling. DNA analysts are not needed to do these things.

> *KEY TAKEAWAY #2.3: The process of generating a DNA profile can produce stochastic or random variation and artifacts that contribute to the challenge of DNA mixture interpretation.*

The random variation arises from the PCR amplification process. The amplification variation is part and parcel of STR data generation, and must be accounted for to separate genotypes and calculate match statistics. There is no "challenge" when computing probabilistically. But challenges do arise when deterministic methods cannot solve a probabilistic problem.

> *KEY TAKEAWAY #2.4: DNA mixtures vary in complexity, and the more complex the sample, the greater the uncertainty surrounding interpretation. Factors that contribute to complexity include the number of contributors, the quantity of DNA from each contributor, contributor mixture ratios, sample quality, and the degree of allele sharing.*

NIST's claimed "complexity" is an artificial concept only relevant to human deterministic interpretation. There is no "uncertainty" surrounding PG computer interpretation. The "factors" listed here can affect the concentration of a genotype probability distribution, or reduce the LR information measurement. There is a linear log-log (straight line) relationship between how much DNA a person contributes to a mixture, and the LR match information the mixture contains to that person [5, 9]. These "factors" tend to reduce to the measured information.

KEY TAKEAWAY #2.5: Continuous probabilistic genotyping software (PGS) methods utilize more information from a DNA profile than binary approaches.

More precisely, (a) probabilistic genotyping software (like TrueAllele) uses more of the STR data from a DNA sample, and (b) extracts more LR identification information, relative to deterministic binary approaches.

KEY TAKEAWAY #2.6: Likelihood ratios are not measurements. There is no single, correct likelihood ratio (LR). Different individuals and/or PGS systems often assign different LR values when presented with the same evidence because they base their judgment on different kits, protocols, models, assumptions, or computational algorithms. Empirical data for assessing the fitness for purpose of an analyst's LR are therefore warranted.

Every statement in this paragraph is incorrect. Of course likelihood ratios are measurements – they quantitatively measure the information contained in evidence data supporting a hypothesis. Automated PG systems of comparable modeling sophistication generally produce similar LR information output measurements when given the same STR evidence data input. The human "analyst" just operates the software, and doesn't play a major role in the process.

KEY TAKEAWAY #4.1: The degree of reliability of a component or a system can be assessed using empirical data (when available) obtained through validation studies, interlaboratory studies, and proficiency tests.

Assessing reliability using empirical data is pretty much the definition of science; no disagreement there. NIST's DNA mixture interlaboratory studies have been unhelpful, though, given their inability to publish results in a timely manner. While useful for other purposes, proficiency tests do not assess the reliability of a computer system.

KEY TAKEAWAY #4.2: To enable effective use of any information, responsibilities exist with both providers and users of that information. While a provider explains the relevance and significance of the information and data, only the user can assess the degree of reliability, validity, and whether that information is fit-for-purpose.

This statement is overly broad. For example, juries do not assess the "reliability" of a probabilistic genotyping computer system. A nice sound bite, but it lacks meaning.

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.

In addition to omitting my paper that unifies LR distributions, validation, and error rates [3], NIST didn't list the DNA mixture data set that I publicly posted along with that paper. The Rutgers resource NIST mentioned contains thousands of publicly available DNA mixture samples; more aren't needed for testing methods. NIST cannot make its own "external" assessments of PG reliability, since it doesn't understand the basic concepts of probabilistic genotyping, validation, or error rates. Moreover, NIST doesn't publish its interlaboratory studies in a timely manner, nor do they subject them to peer review, so their studies have little value to forensic science.

> *KEY TAKEAWAY #4.4: Additional PGS validation studies have been published since the 2016 PCAST Report. However, publicly available information continues to lack sufficient details needed to independently assess reliability of specific LR value produced in PGS systems for complex DNA mixture interpretation. Even when a comparable reliability can be assessed (results for a two-person mixed sample are generally expected to be more reliable than those for a four-person mixed sample, for example), there is no threshold or criteria established to determine what is an acceptable level of reliability.*

NIST's conflict of interest with PCAST was discussed above. It is actually easy "to independently assess reliability of specific LR value produced in PGS systems for complex DNA mixture interpretation" – just examine the genotype's LR distribution, and calculate the PME error rate [3]. NIST has an antiquated deterministic concept of "reliability", rather than a modern probabilistic understanding. They incorrectly apply that which is hard for people as an assessment criterion to that which is easy for PG computers.

> *KEY TAKEAWAY #4.5: Current proficiency tests are focused on single-source samples and simple two-person mixtures with large quantities of DNA. To appropriately assess the ability of analysts to interpret complex DNA mixtures, proficiency tests should evolve to address mixtures with low-template components or more than two contributors – samples of the type often seen in modern casework.*

Proficiency tests of how people perform may not be all that relevant to automated PG computer systems. However, it may be helpful for PG users to run every mixture through their software at least twice, since that duplication can help establish the reproducibility of the system's LR results.

> *KEY TAKEAWAY #4.6: Different analysts and different laboratories will have different approaches to interpreting the same DNA mixture. This introduces variability and uncertainty in DNA mixture interpretation. Improvements across the entire community are expected with an increased understanding of the causes of variability among laboratories and analysts.*

Groups running the same PG software using the same protocols should produce similar genotype and LR results. The main variability I have seen arises from artificial protocol cutoffs (e.g., on LR values) that have little to do with the software's genotyping capability. The

"uncertainty" in DNA mixture analysis arises from the STR laboratory experiment (e.g., PCR amplification variation), not from PG software that can exploit that variability.

> *KEY TAKEAWAY #4.7: The degree of reliability of a PGS system when interpreting a DNA mixture can be judged based on validation studies using known samples that are similar in complexity to the sample in the case. To enable users of results to assess the degree of reliability in the case of interest, it would be helpful to include these validation performance results in the case file and report.*

Based on TrueAllele peer-reviewed studies, this is untrue. First, known samples aren't needed when averaging the LR distributions of probabilistic genotypes – complete sensitivity and specificity LR frequency distributions, error rates, and reproducibility results can be found without any "known samples" [3]. Second, DNA information vs. amount analysis of covariance showed predictable log(LR) results having the same slopes, regardless of the "complexity of the sample" [5]. Third, the LR distribution of an evidence genotype, along with a PME error rate calculated for the suspect's LR value, provides the requisite information and frequency context. Finally, "the degree of reliability" is not meaningful in this context; NIST confuses "reliability" with less LR information residing in the evidence data.

> *KEY TAKEAWAY #4.8: We encourage a separate scientific foundation review on the topic of likelihood ratios in forensic science and how LRs are calculated, understood, and communicated.*

Likelihood ratios – and how they "are calculated, understood, and communicated" – are well understood by the information scientists who develop and compute them. This is particularly true in the standard Bayesian framework (e.g., TrueAllele) where a hypothesis and its alternative must be mutually exclusive and exhaustive. Since NIST has demonstrated little comprehension of these fundamentals, their conducting yet another uninformed "scientific foundation review" would be wasteful, and possibly harmful to science.

Conclusions

Based on the NISTIR report, particularly the "key takeaways", it is clear that NIST does not understand the fundamental concepts of probabilistic genotyping. Their report tries to apply the outmoded ideas of deterministic human data review (a failed paradigm lacking validation) to modern probabilistic computer genotyping (a successful paradigm backed by extensive empirical testing). It reads like an old-time blacksmith harping on why these new-fangled automobiles can't possibly work.

If NIST had anything scientific to say, they would have conducted an empirical PG study. Instead, they wrote a book-length review to undermine the real work of hundreds of forensic scientists. The NISTIR report advocates transparency. Perhaps NIST should release twenty years of internal memos and emails, so that we can all better understand their thinking.

References

[1] Perlin MW, Method and system for DNA mixture analysis, *US patent application* 09/776,096. Filed February 2, 2001; published October 17, 2002.

[2] Perlin MW. Explaining the likelihood ratio in DNA mixture interpretation. *Promega's Twenty First International Symposium on Human Identification*; San Antonio, TX, 2010.

[3] Perlin MW. Efficient construction of match strength distributions for uncertain multi-locus genotypes. *Heliyon*. 2018;4(10):e00824.

[4] Perlin MW. When DNA is not a gold standard: failing to interpret mixture evidence. *The Champion*. 2018 May:50-6.

[5] Perlin MW, Hornyak J, Sugimoto G, Miller K. TrueAllele® genotype identification on DNA mixtures containing up to five unknown contributors. *J Forensic Sci.* 2015;60(4):857-68.

[6] Bauer DW, Butt N, Hornyak JM, Perlin MW. Validating TrueAllele® interpretation of DNA mixtures containing up to ten unknown contributors. *J Forensic Sci.* 2020;25(2):380-98.

[7] Perlin MW, Belrose JL, Duceman BW. New York State TrueAllele® Casework validation study. *J Forensic Sci.* 2013;58(6):1458-66.

[8] Perlin MW, Dormer K, Hornyak J, Schiermeier-Wood L, Greenspoon S. TrueAllele® Casework on Virginia DNA mixture evidence: computer and manual interpretation in 72 reported criminal cases. *PLoS ONE.* 2014;9(3):e92837.

[9] Perlin MW, Sinelnikov A. An information gap in DNA evidence interpretation. *PLoS ONE.* 2009;4(12):e8327.

PC21

Commentary on NISTIR 8351-DRAFT

Budowle, Bruce [redacted]

Wed 8/11/2021 9:59 AM

To: ScientificFoundationReviews <ScientificFoundationReviews@nist.gov>; ScientificFoundationReviews <ScientificFoundationReviews@nist.gov>
Cc: Bieber, Frederick R.,Ph.D. [redacted] Coble, Michael [redacted]

To whom it may concern:
We thank NIST for the opportunity to comment on the captioned DRAFT document titled "DNA Mixture Interpretation:
A NIST Scientific Foundation Review." Attached is our response with commentary and suggestions that we hope will be useful. If you have any questions, please feel free to contact us.
Thank you,
Bruce Budowle
Michael Coble
Fred R. Bieber

PC21a

Commentary and Feedback on initial release of

National Institute of Standards
NISTIR 8351-DRAFT Document

DNA Mixture Interpretation:
A NIST Scientific Foundation Review

https://doi.org/10.6028/NIST.IR.8351-draft

Submitted by:

Bruce Budowle, The University of North Texas Health Science Center, Ft. Worth, TX

Michael Coble, The University of North Texas Health Science Center, Ft. Worth, TX

Frederick R. Bieber, Harvard Medical School, Boston, MA

*The authors of this Commentary have published and lectured widely on the topic of forensic DNA mixture interpretation and declare no financial conflicts of interest.

Scientific foundational reviews are often helpful in demonstrating that extant data and work support the validity and reliability of scientific methods and technologies. Such reviews can point to the capabilities and limitations of such methods and suggest new directions and approaches to improve systems and their applications. NIST should be commended for taking on an important issue such as the use of probabilistic genotyping (PG) to interpret DNA mixture profiles. This topic is of vital importance to the forensic, public safety, and legal communities as a substantial portion of biological crime scene evidence involves a mixture of two or more contributors. Several software tools are now in widespread use in the U.S. and abroad for analysis, interpretation and reporting of forensic DNA mixtures.

We are grateful for the opportunity to provide Commentary and respectfully submit the following comments and feedback on the initial release of the above named NIST DRAFT document on the topic of DNA mixture interpretation. We trust that NIST will respond to our comments and will circulate them to the wider community as part of this review process. As the NIST document is initially released as a DRAFT, we submit that extensive revisions are required before final

publication. We find that the initial DRAFT release has some serious flaws, fails to take proper notice of a substantial body of literature in support of PG, and therefore fails to meet the highest standards of a scientific foundational review. The first part of our comments provides an overview of our general concerns followed by enumeration of specific points to be addressed in a revision.

Overall, we submit that this initial release DRAFT by NIST:

- fails to discuss important work that describes the strengths of the current technology,
- departs beyond a scientific foundational review when delving into topics such as relevance,
- did not make use of all available data relating to PG validation,
- contains inconsistent/conflicting recommendations (e.g., bracketing – see #11 and #12 below),
- fails to address some noteworthy constructive criticisms provided by public comment on a previous document prepared by NIST on foundational reviews (NISTIR 8225),
- excessively cites scientific literature which are not primary sources,
- fails to give due credit for practices that are well established (see #53 below), and
- fails to address and perpetuates misunderstandings about forensic DNA mixtures (e.g., the 20% donor threshold quoted in the PCAST Report).

This DRAFT release by NIST does not make use of some of the wealth of information that lays a solid foundation for interpretation of forensic DNA mixtures using PG. The NIST DRAFT has opined on the perceived gaps of PG without presenting and discussing any of its own data analyses. The wealth of PG data, both published and internal validation studies, that NIST did not access and analyze, may support (at least some of) the concerns of the authors of their initial DRAFT release. NIST needs to obtain data from forensic laboratories that have performed PG validation studies. NIST opines that a reason that such data may not be accessible is related to privacy concerns (Lines 2480-2482 - "Some laboratories provide summary information from their validation studies, but detailed data are often unavailable, in part because of privacy concerns around releasing genotype information from individuals"). We note that Material Transfer Agreements (MTAs) or Data Use Agreements (DUAs) are available that may obviate such concerns.

In fact, the work of Bright et al. (Forens. Sci. Int. Genet. 34:11-24, 2018) would suggest that privacy is not an impediment to PG data sharing. In response to criticisms of DNA mixture analysis in the PCAST Report, Bright and Buckleton coordinated with 31 laboratories which willingly participated and shared their respective PG data. Failing to obtain extant data is reminiscent of similar failings of the PCAST Report (https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/PCAST/pcast_forensic_science_report_final.pdf) and notably was a significant criticism of the NRC 1 Report (1992). We would like to note that neither the PCAST nor NRC I obtained all available relevant and accessible data before issuance of these respective reports.

Alternatively, if privacy is a concern and thus “all” validation data generated by forensic laboratories may not be accessible to the public, NIST, in its role, might be able to gain limited

access to the data to perform foundational validity studies. These data may be protected in some fashion against dissemination; however, disclosure and discovery issues may still apply to these data. Regardless, the DRAFT may gain forensic laboratory community support for sharing the data with NIST if the DRAFT included a proposed validation plan and clear minimum criteria that should be met for such foundational validation studies. Such an endeavor would have the ancillary benefit of projecting the needed resources to perform an appropriate foundational validation study (assuming that such a study is needed after all accessible data from the peer-reviewed literature have been evaluated).

NIST, however, appears to have ample access to empirical data on PG, since it has been researching (and presenting at meetings) for almost a decade, and yet their initial DRAFT release cites just three publications of its own recent work (Hannig et al 2019, Riman et al 2019b and 2021) – two are short summaries in a supplemental series (which provide little data) and one is a manuscript (which to our knowledge is yet be accepted for publication subject to peer-review). These studies may have value as they compare results between two PG software tools using PROVEDIT mixtures. For example, in the manuscript cited by Riman et al (2019b) the authors conclude:

> "The publicly available PROVEDIt database is a useful resource to investigate probabilistic genotyping software. The effects of software (STRmix and EuroForMix), NOC, and propositions on Log10(LR) assessment were examined. As expected, both software showed high degree of discrimination between Hp True and Hd True cases across different contributor ratios and treatments for 2 and 3 contributor samples."

In this next section of our Commentary, we seek to identify the most significant issues with the NIST initial DRAFT release in our considered opinion the NIST DRAFT requires extensive revision prior to more open public comment or eventual final publication. Adequate data analyses including an introspective review/assessment are integral to the goal of improving the NIST DRAFT document to be most relevant and applicable to the forensic community and other stakeholders. The revision should include a more balanced discussion of the strengths and limitations of PG. Accordingly, we respectfully invite NIST to comprehensively address our suggestions/comments (shown in **BOLD** below).

1. **NIST should place its own data in an open access site for analyses by interested parties (see Lines 526-528 "Information contained in this report comes from the authors' technical and scientific perspectives and review of information available to us during the time of our study."). NIST has made some attempt with the three publications mentioned above – although two are summaries), but not with much of its other data. We recommend that NIST should consider addressing the sponsorship of an open access website for such data. MTAs or DUAs can be put in place for such data sharing purposes.**

2. **Limitations on methods should not equate to a wholesale rejection of a methodology or system. We note, for example, that the molecular biology of forensic DNA typing has**

limitations, which are known and relied upon, so scientists can make sound judgments and interpretations. Understanding limitations can enable scientists to employ a tool and to not exceed the bounds of that tool. Indeed, the one NIST unpublished manuscript by Riman et al (2021) describes limitations or reasons for their observed results. This study and similar ones that have been peer-reviewed guide users to better understand systems to make informed decisions. Note NIST's position on the purpose of a foundational review.[1]

3. **The laws, principles, and empirical data that support the use of PG, and the substantial literature that demonstrate improvement over current manual practices should be discussed and assessed. While the draft accepts single source and simple mixture interpretation is considered robust (seemingly deferring to the 2016 PCAST Report), it lacks a thorough review of the strengths (i.e., trusted and established knowledge that supports and underpins) of how and why these "simple" profiles are robust and more so what are the scientific underpinnings of PG. There is a substantial body of literature that demonstrates improvement over current manual practices; yet these key studies are not discussed and assessed for their strengths (and limitations).**

4. **Not all citations in the DRAFT release are from primary sources (many are, instead, review articles and textbooks written by the first author of the NIST DRAFT release). Such citations cannot be considered foundational and therefore may not be appropriate for a foundational review. For example, see lines 5850-5854 "Attempts to recover information from low amounts of DNA present in evidentiary samples using LCN methods inevitably led to increased imbalance in heterozygotes, higher levels of stutter products, allele drop-out, and allele drop-in (contamination). These phenomena are artifacts of stochastic, or random sampling,**

[1]Lines 122-125 "A scientific foundation review, also referred to as a technical merit evaluation, is a study that documents and assesses the foundations of a scientific discipline, that is, the **trusted and established knowledge** (emphasis added) that supports and underpins the discipline's methods." Additionally, on lines 126-128 see "These reviews seek to answer the question: "What **established scientific laws and principles as well as empirical data exist** (emphasis added) to support the methods that forensic science practitioners use to analyze evidence?" Also see lines 1105-1112 "1.5. Why Conduct This Scientific Foundation Review? As described in our earlier publication (NISTIR 8225), a scientific foundation review is "a study that seeks to document and evaluate the foundations of a scientific discipline, that is, the trusted and established knowledge that supports and underpins the discipline's methods. These reviews seek to answer the question: 'What **empirical data exist that speak to the reliability of the methods** (emphasis added) that forensic science practitioners use to analyze crime scene material?'" See also Lines 2375-2380 This scientific foundation review seeks to document and **independently assess the empirical evidence that supports the reliable use of DNA mixture interpretation methods** (emphasis added). The sources of data and information used in conducting this review are described in this chapter. These sources include (1) peer-reviewed articles appearing in scientific journals, (2) published interlaboratory studies, (3) laboratory internal validation studies that are accessible online, and (4) proficiency test data available on test provider websites.

effects that occur in the early cycles of PCR amplification when there are a limited number of target molecules to amplify (Butler & Hill, 2010)." We note that these findings were described by many other scientists at least a decade earlier (for example Gill et al. 2001; Moretti et al 2001). Thus, this initial NIST DRAFT release should be revised to cite the rich resources and primary research that support the many developments and practices of the field.

5. On lines 3255-3257 it is stated in the DRAFT "Specifically, these methods appear to be reliable for three-person mixtures in which the minor constitutes at least 20 percent of the intact DNA in the mixture and in which the DNA amount exceeds the minimum level required for the method."

   This statement is a direct quote from the PCAST Report (2016). But NIST does not note that published studies (including internal validation studies) routinely analyze mixtures in which component contributors are well less than 20%. In fact, a substantial portion of the 2825 mixtures reported by Bright et al. (Forens. Sci. Int. Genet. 34:11-24, 2018) compiled from 31 internal validation studies exhibit DNA contributions of one or more donors of less than 20%. The statement by PCAST was incorrect when initially published and that Report was not a typical peer-reviewed document. It is worthy of note that there have been substantial criticisms of the PCAST Report, and there have been studies that address issues that were raised in it.

6. On lines 643-644 the DRAFT states "There remains a need to assess the fitness for purpose of an analyst's LR using empirical methods."

   While there are issues of concern with the use of LRs which the community should be made aware of, **NIST does not discuss the studies that support them being fit for purpose. The laws, principles, and empirical data that support the use of LRs should be addressed.**

7. Lines 658-662 "KEY TAKEAWAY #2.2: Generating a DNA profile involves measuring the inherent physical properties of the sample. Interpreting a DNA profile involves assigning values that are not inherent to the sample. To do this, the DNA analyst uses their judgment, training, tools (including computer software), and experience, and considers factors such as case context." Also see Lines 1501-1511 "Measurements reflect the physical properties of the sample while interpretation depends on the DNA analyst assigning values that are not inherent to the sample."

   It is unclear here and in other places in the DRAFT what "inherent" means as this is confusing.

8. Lines 741-746 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.

Based on the initial NIST DRAFT release, it appears that the authors of the DRAFT did not gather and review all the available data (i.e., in particular the internal validation studies performed by many laboratories and underlying data from peer-reviewed publications). As suggested above NIST should prepare a validation plan and describe the criteria to be met to persuade forensic laboratories to share their data; with buy-in NIST now might be able obtain and analyze these data to be in a better position to point out any perceived (or better yet identified) gaps and limitations.

9. NIST suggests a list of desired information in Box 4.1 to provide an independent reviewer with data to assess the reliability of a PG system. One critical missing piece of information in this list is the input file from an analysis of the data. Simply providing raw data files to different reviewers can create different input files that can influence the resulting LR (e.g. failure to remove oversaturation, dye blobs, and other artifacts). PG software is a tool for the analyst and not a black box where data are simply uploaded without a critical interpretation of the results.

   NIST should evaluate the search criteria it used to find data and re-investigate the extant data. For example, it appears that NIST's "Google search" missed some publicly available data. Table 3.2 lists available laboratory internal validation studies as of March 23, 2020. The Brooklyn Defender Services (https://indefenseof.us/issues/kinship-problem) lists six laboratories that have made their summary data available that NIST did not identify in its search. Four of the six summaries provide dates of the work and all precede the date of the search made by NIST.

10. Lines 1114-1119 Such a review can help identify knowledge gaps and provide guidance for future research. In addition, documenting foundational studies and core principles in a written report can assist laboratories in identifying appropriate limits for interpretation and contribute to the training of forensic practitioners. This report can also help investigators, officers of the court, and other users of forensic science to consider DNA test results in context and with awareness of their limitations so they can make informed decisions.

    Revisions to this initial DRAFT release should include recommendations and guidance on how to address the perceived needs or gaps. Furthermore, NIST might consider providing metrics to gather to address any concerns. We note "documenting foundational studies and core principles in a written report can assist laboratories in identifying appropriate limits for interpretation and contribute to the training of forensic practitioners". While documenting in a report is not a sound recommendation (see below); it is not clear how documentation in a report contributes to training of practitioners. The important point to stress is that understanding appropriate limits can aid in performing valid and reliable interpretations. NIST should expand on this concept in a revised document.

11. Lines 769-773 KEY TAKEAWAY #4.7: The degree of reliability of a PGS system when interpreting a DNA mixture can be judged based on validation studies using known samples that are similar in complexity to the sample in the case. To enable users of results to assess the degree of reliability in the case of interest, it would be helpful to include these validation performance results in the case file and report.

    This statement points to incongruent recommendations by NIST. First, the initial NIST DRAFT release should carefully define the purpose of a report before recommending what should be in a it. Second, NIST should recognize that validation studies can be obtained through normal discovery and disclosure mechanisms; so, there is no need to incorporate these voluminous studies into a report, which would make a report unwieldy. Third and maybe, the only part of this "Key Takeaway" that could be attributed to a scientific foundation is including samples that are "similar in complexity." Yet in another part of this DRAFT release (see lines 400-402) the DRAFT offers that "bracketing" (actually known as fractional factorial design) is a way to address the difficulty of context capturing the myriad possible ways that mixtures may present, i.e., NIST notes that validation studies cannot cover each situation and bracketing can be "a pragmatic way of understanding case-specific reliability of an interpretation system" (also see lines 2877-2878 and lines 3209-3210 with caveat). If bracketing is a simple option, then it is unclear how laboratories would have validation data that are "similar in complexity" to what is observed in the instant case. This NIST DRAFT recommendation conflicts with the authors' suggestion of "bracketing" to address the fact that a similar complexity may not be met with validation studies (as recommended by NIST).

12. On lines 3209-3215 NIST Suggests "A bracketing approach (discussed in Section 4.4.5) may provide a pragmatic way to infer reliability for DNA mixtures in a region of the factor space, *but will still require an element of trust in the DNA interpretation system used* since the entire factor space may not be covered with previously collected validation data. Yet even with a bracketing approach where there is not validation data defining every portion of the factor space, a user must trust in the DNA interpretation system enough to extrapolate assessment of reliability across gaps in the factor space covered."

    The NIST DRAFT does not take notice that bracketing has been practiced from the beginning of the use of PG – for example, see manufacturer's recommendations for STRmix modeling. Indeed, this established statistical practice called a fractional factorial design (FFD), noted above, substantially predates PG. Given that some of the explanatory variables (e.g., template is a continuous variable, and thus DNA/mixture profiles will never be the same. Use of FFD is therefore a key component of mixture analysis.

13. Lines 835-840 This statistic does not provide any information about how much DNA was present, or how or when the DNA was deposited. For instance, a large blood stain might produce a very similar likelihood ratio to a swab from a light switch, yet the two types of evidence might vary greatly in terms of their evidential value. Therefore, the likelihood ratio should not be used in isolation. It is imperative that the likelihood ratio be considered in the context of other evidence in the case. (also see Lines 1100-1101 "Focusing only on a statistic without considering context can be misleading." And Lines 4623-4625 While an LR value is

an expression of the strength of evidence under a pair of propositions, the result should be considered in context (i.e., the result represents the evidence for what?).

This recommendation ventures into relevance, which is beyond the scope of a scientific foundation review aside for the need to understand the potential of DNA transfer, which should be distinguished as "activity" as opposed to "relevance". Probative value cannot necessarily be equated to the amount of DNA or the LR. Moreover, probative value is the purview of the fact finder and triers of fact, not the forensic laboratory. If the NIST DRAFT ventures into the issue of context, then it could address the problematic issues of use of the co-called "verbal scale" to explain LRs. Use of a verbal scale to explain results of mixture analysis is problematic as it is based simply on the value of the computed LR and is not context driven.

14. Lines 856-858 KEY TAKEAWAY #5.1: DNA can be transferred from one surface or person to another, and this can potentially happen multiple times. Therefore, the DNA present on an evidence item may be unrelated (irrelevant) to the crime being investigated.

    As has been recognized for years, while "transfer DNA" may occur; the NIST revision could address what happens to the amount of DNA deposited in successive transfers.

15. Lines 972-974 Highly sensitive methods began moving from research centers into crime laboratories more than ten years ago, but the application of such methods to detect minor contributors in DNA mixtures has increased rapidly in recent years.

    Note that highly sensitive methods have been in crime labs for more than two-three decades. A revision could either correct or clarify to what the "ten years" refers.

16. Lines 1147-1154 We find merit in the perspective that "Dissemination is a critical part of the scientific process because it exposes our work to peer review and allows scientists to build upon the contributions of others. A study isn't complete until it's been published" (Martire & Kemp 2018). In addition, many published developmental validation studies do not include enough data for an independent assessment of performance. We believe that greater transparency through forensic laboratories openly sharing their supporting validation data, along with an independent review, would help strengthen the field of forensic DNA analysis.

    We agree that dissemination of data is essential to be informed and continue to constructively critique work (i.e., a scientific culture). But, NIST has a dilemma here and needs to address this recommendation with the same advocacy for transparency. While NIST has performed and presented on its own PG studies for several years, results have not been published except for reference to unpublished studies by Riman et al (2021); NIST has not made all its own data publicly available. Again, we respectfully submit that NIST should now consider making all its data publicly available via open access.

17. Lines 1353-1354 The evaluative uses of DNA information are held to a higher standard than investigative ones.

A revision of the NIST DRAFT could expand on this comment. There are situations where this statement may be true and others where it is not correct.

18. Lines 1409-1410 The EPG is the raw data that must be interpreted to draw conclusions from the sample.

This statement is not correct. The EPG already is a presentation of processed data.

19. Lines 1421-1423 When degraded, DNA molecules break into smaller pieces, such that some or all of the tested loci are no longer detectable by PCR.

PCR amplifies (or fails to amplify if fragments are too degraded for the assay design). PCR in the context of STRs does not "detect" loci.

20. Lines 1424 -1428 Swabs from so-called “touch evidence” samples, which typically have a relatively small quantity of biological material deposited (with perhaps tens of cells), are more likely to exhibit allele drop-out compared to visible blood or semen stains, which contain hundreds to thousands of cells.

This statement - that visible stains contain hundreds of thousands of cells - is a generalization and may not apply to stains that have been exposed to environmental insults. A revision to the NIST DRAFT could clarify that fresh samples that have not been exposed to the conditions that forensic samples encounter may contain hundreds to thousands of cells.

21. Lines 1445-1448 Instead, DNA mixture interpretation is an effort to (1) infer possible genotypes as detectable sample contributors (a process sometimes referred to as *deconvolution* of the mixture components) and (2) provide the strength of evidence for a POI to be included in an evidentiary DNA profile.

Please comment on "exclusions".

22. Lines 1485-1487 Results are then interpreted, compared to reference sample profiles along with a statistical estimate of the strength of evidence, and reported in a written summary.

Please consider or qualify that comparisons that support an exclusion may not require a statistical estimate; also, when an individual is assumed to be a contributor there is no estimate.

23. Lines 1556-1558 Those are alleles that (1) are unmistakable, (2) may be masked by an artifact such as stutter, and (3) have dropped out completely and are therefore not detected (Gill et al. 2006b).

A revision should **mention "allele sharing" among/between contributors.**

24. Some examples of citations that are clearly not primary sources:

For example, see Lines 1717-1718 "Stochastic effects can also cause alleles that are not present in the sample to "drop in" to the profile (e.g., Moore et al. 2020)." – Gill et al., 2001 describe "drop-in" almost 20 years earlier.

Another example is on Lines 2001-2003 "A number of software programs have been developed in recent years to assist analysts in performing DNA mixture interpretation by computing LR results using discrete or continuous approaches (Coble & Bright, 2019; Butler & Willis, 2020)." - n.b., Butler and Willis are not a primary source.

25. Table 2.2
"For mixtures, an assumption that the major contributor can be distinguished from minor components so that specific genotypes in the major can be inferred."

Please comment on minor contributor(s).

"All alleles for all contributors are all present at the reported loci (i.e., cannot cope with allele drop-out that is expected with low quantities of DNA)
An assumption as to the number of contributors and a specific pair of propositions"

A revision should mention the need to determine NoC for CPI.

26. Lines 2018-2022 A PGS system computes LR values based on the information provided (Figure 2.4), including (1) *modeling choices* made by the system architect(s), (2) *data input choices* made by the analyst regarding an analytical threshold for calling peaks as alleles, selecting the number of contributors to the mixture for use in PGS calculations, and sometimes categorizing artifacts (e.g., pull-up peaks),

Please note that some PG systems do not advocate use of an AT.

27. Lines 2029-2037 An increasing number of forensic laboratories are beginning to use PGS for DNA mixture interpretation. The UK Forensic Science Regulator shared seven perceived benefits of PGS compared to manual calculations (UKFSR 2018b, p. 8): (1) increased consistency within and between organizations utilizing the same software, (2) information available in the profile is used more efficiently, (3) deconvolution of genotypes enabling

database searches that would not otherwise be feasible, (4) improved reliability due to increased automation in processing, (5) reduced variability between analysts in deciding whether peaks are true alleles or artifacts, (6) increased range of DNA profiles suitable for interpretation, and (7) publication of statistical models in peer-reviewed journals.

Number 7 is stated differently in the UK document (below) with emphasis on acceptance. It seems that the NIST DRAFT does not include this part of the UK Regulator's benefits.

In the UK Document it states
"The benefits of using DNA mixture interpretation software compared with manual calculations are as follows: a. Consistency: Reduced scope for operator-to-operator variation in data input and interpretational approach, thereby increasing consistency within and between organizations utilizing the same software. b. Information utilization: Software enables more sophisticated modelling that utilizes the available information in the profile more efficiently. In principle, this leads to higher LRs in cases where Hp is true and smaller LRs in cases where Hd is true. c. Deconvolution of genotypes: This is far more effective with software, enabling database searches that would not otherwise be feasible. d. Improved reliability: There is a methodical approach with defined standards built on principles that have been tested and validated. Increased automation of processing reduces the risk of human error in manual data manipulation. e. Reduced variability between analysts: Less analyst decision-making in terms of determining whether peaks are true alleles or artefacts, making peak assignment more automated and reducing variability between analysts. f. Cost-effectiveness/ utility: Increases the range of DNA profiles suitable for interpretation, including low template and complex DNA mixtures, for which manual calculation is unfeasible. **g. Demonstrable scientific acceptance: Publication in peer-reviewed journals of the validation of the statistical models and software programs demonstrates scientific acceptance, as may be required by the courts and for compliance to BS EN ISO/IEC 17025:2017."(emphasis added)**

28. Lines 2097-2099 Finally, the lowest LR result in Table 2.4 comes from considering a possible untested brother rather than an unrelated individual in the assumptions made and calculations performed. Even considering only two loci, LR assignments can differ by several orders of magnitude.

A revision to the NIST DRAFT should note that differences in computed LRs are indeed expected to occur when propositions change, and such difference in themselves do not necessarily connote discordance. Indeed, the LRs should be different (it would more surprising if the LRs were similar).

29. Lines 2209-2210 There are a limited number of alleles at each locus, and even individuals who are not closely related will share alleles and genotypes.

We recommend that revision to the NIST DRAFT release could expand on the topic of allele sharing with unrelated individuals. and under what circumstances such allele sharing may impact DNA mixture interpretation.

30. Lines 2247-2258 Principle 9 [Measurement]: Peak positions more accurately reflect allele calls than peak heights represent relative allele amounts.

"Use of an internal size standard with each tested sample along with calibration to an allelic ladder enables accurate STR allele designations with electrophoresis separation and detection systems (e.g., Gill et al. 1997, Lazaruk et al. 1998). Peak heights and relative peaks heights, which do not use internal size standards to normalize stochastic variation, are not as reproducible as peak positions but do show trends by locus (e.g., Leclair et al. 2004, Debernardi et al. 2011). *This principle is a reminder that while alleles may be either present or absent (impacted by their peak heights and instrument detection thresholds), detected alleles are reproducible in terms of their designation (i.e., replicate testing does not show alleles shifting to a different allele, e.g., a "12" cannot become a "14" because peak position/sizing is stable).*"

A revision should address or describe why peak heights vary; it could also address "noise".

31. Lines 2262-2264 Low peak heights are a function of starting amount and quality of the DNA template. When sufficient quality and quantity of DNA template exist, reliable and unambiguous DNA profiles can be generated from crime scene evidence.

A revision should also address "High peaks". Also, DNA template could be of high quality and quantity, but in a mixture the individual components may be low.

32. Lines 2270-2272 Replicate amplification from aliquots of the same DNA extract have been used to improve the degree of reliability (Taberlet et al. 1996, Gill et al. 2000, Benschop et al. 2011). *This principle relates particularly to minor contributors in DNA mixtures.*

A foundational scientific review should encompass extant data and provide a balanced view. There are alternate views and findings – see Bille al (2021) and Griesdale and van Daal (2012)

33. Lines 2391-2394 PT provider websites or publicly available internal validation data summaries from individual laboratories. PT data provide insights into how individual analysts performed on specific tests while internal validation studies offer insights into how laboratories performed when analyzing a range of DNA mixtures of varying complexity.

The revisions to the NIST DRAFT should address the findings and conclusions of these PT data? All that is provided is a list. No analysis of the results was performed. Are the errors all attributed to PG? Are any of the errors attributed to PG? Are some of the errors not related to STRs? If so, what are they? Are the errors transcriptional or are they interpretational or a laboratory issue? Was the frequency of error different between non-

PG and PG analyses? These are a few examples of questions that the NIST DRAFT release could have addressed (especially since it takes the position of making use of accessible data). If such analyses are not possible, revisions to the NIST DRAFT document could convey that the root cause of the error is unknown and may be attributable to a number of (varying) factors.

34. Lines 2539 3.1.3.2. Bode Technology

Please elaborate on what Bode offers (not as descriptive as CTS).

35. Lines 2647-2649 *We note that the degree of reliability of a DNA mixture interpretation system, such as a DNA analyst using a probabilistic genotyping software program, depends on sample complexity.*

This statement might suggest that the degree of reliability is related to complexity. This point deserves to be addressed in more depth in a revision given the purpose of the DRAFT release is purportedly a scientific foundational review. As an example, a complex sample may yield a LR that trends towards 1. The low LR value does not necessarily equate to unreliability. It may equate to the data be uninformative and yet be highly reliable. The complexity may not allow for deconvolution and a low LR may reliably reflect the interpretation.

36. Lines 2862-2863 Table 4.1. Factor space that influences DNA mixture measurements and interpretations with probabilistic genotyping software (PGS) systems.

The NIST DRAFT release provides a list of influencing factors on portions of factor space (and in Table 4.3 lists the publications on PG and the factor space covered). However, the NIST DRAFT does not describe what studies that satisfy the factor space parameters. An in-depth review of the literature and what foundations the studies support is requisite for a scientific foundational review.

37. Lines 2900-2903 If a casework scenario is encountered with an eight-person mixture involving only 10 pg. total template DNA, then DNA analysts might refrain from interpreting such a sample because it has not been covered in any of their validation experiments.

Such a sample would not be analyzed by most labs. Please use a more realistic example to make your point.

38. Lines 2932-2933 - Thus, a user of information assesses the degree of reliability (trustworthiness) and determines validity (e.g., whether a method is fit-for-purpose). Lines 2945-2946 - In this case, the judge, jury, and lawyers determine whether sufficient information has been provided.

The "user" is far more complex. Prosecutors and defense attorneys typically rely on the scientists or other experts to tell them what is reliable or not. We note that interaction with the judicial system is important but propose that the purpose of this NIST DRAFT should carefully focus on scientific foundations and underpinnings.

39. Lines 3157-3159 Curiously, the single false inclusion came from a reference Item 2 to a single contributor evidence profile (Item 3, which was not a provided reference profile and was incorrectly classified as a two-contributor mixture by the submitter).

This whole part does not address what the findings mean? Also, we are confused why this is "curious" - Please explain.

40. Lines 3201-3207 Based on an examination of publicly available information reviewed during the time frame of this study, there is not enough information for the authors of this report to independently assess the degree of reliability of DNA mixture interpretation at any one point in the factor space. This is particularly true without an established and accepted criteria for reliability with complex mixtures involving contributors containing low quantities of DNA template (e.g., Benschop et al. 2015a) or where there is a high degree of allele overlap among contributors (e.g., Bright et al. 2018, Lin et al. 2020).

It is apparent that NIST may not have such data (except for the unpublished study by Riman et al); NIST could attempt to obtain the data or utilize part of its funding to conduct a study to generate an adequate volume of mixtures to explore the "factor space." There are "ground truth samples" and reference samples/cell line DNAs, and many labs use the NIST standard to calibrate (note this latter is not mentioned). What about the mixture database – PROVEDIt? NIST mentions PROVEDIt in Table 3.3 (and associated text in section 3.1.5) but does not address the value or inadequacy of this resource supporting PG (i.e., ""a large dataset would play a critical role in demonstrating the 2626 foundational validity and robustness of new or existing DNA identity testing technology" (lines 2626-2627). Yet, PROVEDIt is a resource accessed by Riman et al and their unpublished findings (and those of other studies) could be discussed in the NIST foundational review.

41. Lines 3281-3283 Because *the nature of overlap among alleles is a key issue*, it is critical to examine mixtures from various different sets of people.

Given that each study uses its own accessible samples, collectively there are different sets of individuals. But the issue is not amount of people (although sampling/power is important) but the additive effects/variance. This part does not need a lot of people but good replicate studies.

42. Lines 3301-3303 Now, with the perspective of an additional five years of reflection, what publicly available data exist? Locating and understanding this information have been an important part of this DNA mixture interpretation foundation review.

"Understanding the information" has not been described in this document.

43. Lines 3318-3322 It is not helpful for the provider to describe a method as "validated" without providing context around the method's use and access to data to support claims of validity and reliability. Instead, it might be more appropriate to state "the following developmental validation studies have been conducted and here is the complete collection of results obtained, which can be examined by users to make reliability judgments."

The statement is potentially misleading as several developers of PG have been quite prolific and have provided far more than a mere statement that their system is validated. The developers tend to follow developmental validation requirements, and publications describe summaries of such studies. Our experience is that, for example, upon our requests the developers of one software tool (STRmix) have provided substantial data to support their findings. Also, unless such publications are generated, most U.S. labs will not proceed with undertaking internal validation studies and implementation. In addition, QAS Standard 8 is clear regarding developmental validation studies and that laboratories have those references as initial support for internal validation. Revisions to the NIST DRAFT release should cite the studies by PG developers and be more explicit of what data are captured as well as describe what exchange is occurring in the community.

44. Line 3454 Table 4.9. Issues with available information for the data sources examined in this study.

"most previous studies are not relevant to PGS methods in use today"

We respectfully disagree on a fundamental level. Previous work does lay a scientific foundation. The understandings of today are grounded on the work of the past, especially in science. For example, additive effects with allele sharing, stochastic effects, and sampling variance are all aspects of traditional STR typing that assist in forming the bases of PG.

45. Lines 3508-3512 There appears to be a general misconception that LR assessments made by different experts will be close enough to one another to not materially affect the outcome of a case. Although they may be close enough in many instances, this is not known for any particular case and it is not advisable to take this for granted.

The NIST DRAFT provides no metrics or guidance on what would be considered close, when results may not be close, and what laboratories should do or have done to assess closeness in this review. A blanket statement like this is not helpful. Also, should NIST be assessing materiality or should NIST be assessing measurement uncertainty? This NIST DRAFT

release risks blurring the clear lines between the distinct and separate roles of science and the judicial systems.

46. Lines 3514-3519 In addition, there are a number of different LR values that can be generated by a PGS system, such as a highest posterior density (HPD) LR to adjust for sampling uncertainty, a unified LR to account for both related and unrelated individuals under the defense proposition, a population stratified LR to incorporate relative proportions of different subpopulations, a variable number of contributors (varNOC) LR estimation, or various combinations of these LR adjustments (Kelly et al. 2020).

We suggest that NIST studiously avoid advocating the use of the terminology of "defense hypothesis". It is no longer considered an acceptable terminology, especially from a laboratory perspective. When issuing a report, the analyst may not know what defense hypotheses, if any, may be offered.

47. Lines 3558-3564 Likelihood ratios must satisfy an internal consistency requirement (called the property of being well-calibrated or "calibration accuracy," for short) which can be empirically tested (Ramos & Gonzalez-Rodriguez 2013, Meuwly et al. 2017, Hannig et al. 2019). The scientific validity of any particular PGS system used in casework can be assessed, at least partly, by investigating (1) repeatability, (2) reproducibility, (3) calibration accuracy, and (4) the efficiency or discriminating power. Such an exercise will help identify the better-performing PGS systems for consideration in casework applications.

The NIST DRAFT lays out four criteria – repeatability, reproducibility, calibration accuracy, and discrimination power. NIST does mention calibration accuracy (see line 3558). Yet the DRAFT does not describe the foundational information from the current literature. A scientific foundational review as proposed by NIST should explore the published studies that have performed these validation studies and describe how they support PG (or are insufficient and why). This section is potentially misleading as it intimates that studies have not been performed.

Revisions to the NIST DRAFT could consider accessing and evaluating Jo-Anne Bright, M. Jones Dukes, S. N. Pugh, I. W. Evett & J. S. Buckleton (2021) Applying calibration to *LR*s produced by a DNA interpretation software, Australian Journal of Forensic Sciences, 53:2, 147-153.

48. Lines 3746-3747 In the end, the reliability of LR values produced by a PGS system means little if relevance of the DNA evidence has not been established first (see Chapter 5 in this report).

Relevance is outside the science component. Perhaps some of the confusion here and elsewhere in the draft is that the NIST DRAFT seems to blur the concepts of relevance and activity level. These are different concepts.

49. Lines 3793-3795 For example, with a visible blood or semen stain, the cell type could be determined, and the activity that caused a sample to be deposited could often be inferred, even by nonexperts.

Typically, the forensic DNA scientist avoids opining how the sample came to be on the item, even when a lot of sample is collected. Consider that an activity that caused a sample to be deposited could be, for example, a consent/no consent argument – certainly outside of a scientific foundational review. See next comment as well.

50. Lines 3802-3803 Our summary of the above papers is that the relevance of a DNA sample to the crime is often difficult to discern.

Please be reminded that relevance is the purview of legal system.

51. Lines 4199-4201 In a number of studies, the major profile was not always associated with the last person to handle an item (Cale et al. 2016, Buckingham et al. 2016, Goray et al. 2016). This may result from background DNA or from the handler depositing non-self-DNA.

There are certain conditions in which it is more probable to observe a major profile not deposited by the one who handled the item. The revised NIST DRAFT should describe the conditions based on the findings in the literature cited in this DRAFT.

52. Lines 4512-4513 The above three points apply to any low-level profile and therefore also apply to profiles containing mixtures.

Actually, they also apply to high quantity DNA.

53. Lines 4521-4522 - The highly sensitive DNA methods that have become common in recent years increase the likelihood of detecting irrelevant DNA.

    Also, Lines 4536-4537 - The full implications of these observations have not yet infiltrated the routine practice of DNA testing in many criminal investigations.

Such discussions have been ongoing for at least the last 20 years. The NIST DRAFT revisions should list the issues that have been raised in this regard and what approaches or data have been made or proffered to address the concern. Leaving this sentence as is suggests that the forensic community is unaware and has not considered the concern, which does not convey the state-of-the-knowledge (either supporting or lacking). Moreover, the NIST DRAFT release provides no data or information on how it concluded that these observations and

implications have not infiltrated routine practice. Did NIST contact the laboratories or other stakeholders? Did NIST perform a survey?

54. Lines 4719-4721 - Suspects and attorneys may overestimate the value of the DNA findings and accept a plea possibly even when innocent.

Plea negotiations are outside the purview of this DRAFT document. What part of this statement refers to the scientific foundations of PG?

55. Lines 4741-4742 When secondary transfer is alleged, the DNA match probability has less impact and variables associated with the donor are important.

In an adversarial legal system one side may allege transfer and in some circumstances that proposition could be quite probable and in other scenarios it could be highly improbable. Alleging "transfer" by itself is not necessarily sufficient to diminish the value of the DNA findings.

56. Lines 4747-4754 The LRs produced from activity propositions are generally much lower in numerical value than those produced from source propositions. An early paper illustrated this observation, showing an activity level LR of the order of 1000, in contrast to what the authors describe as an infinite LR in favor of a sub-source level proposition (Evett et al. 2002). Some have argued that, given that activity propositions produce more conservative assessments of weight of evidence and are more relevant to the issues of the court, their use is more appropriate (Biedermann et al. 2016b, Kokshoorn et al. 2017, Taylor et al. 2018, Szkuta et al. 2018).

The NIST DRAFT release does not assess or address that given the difference in the propositions and that the LRs would be different. Is it proper to compare LRs that are derived from very different hypotheses? If so, when and why? If not, when and why?

57. Lines 4842-4844 The "number" (LR value) is like seeing the highlight of an advertisement without reading the small print and considering the propositions behind the number. Kwong recognized this for DNA in a *Harvard Law Review* article:

The NIST DRAFT does not address that the propositions are provided in reports and case files contain such information routinely.

58. Lines 4851-4854 - It may also be necessary to do additional sampling, seek information about other genotypes in the mixture, or conduct *ad hoc* transfer experiments that apply to the particulars of the case. This has been referred to as "sense making" by Paul Roberts (Roberts & Stockdale 2018).

It is not recommended to do *ad hoc* transfer experiments. We suggest that revisions to the NIST DRAFT consider the consequences of relying on *ad hoc* experiments as opposed to relying on well-developed validation studies. It seems, to us, rather odd to stress the need for validation studies and then support *ad hoc* experiments.

59. Lines 4907-4909 Triers of fact should be made aware that the LR value addressing a sub-source level question is not sufficient evidence that the POI transferred his or her DNA to the knife at the time of the stabbing.

Revisions to the NIST DRAFT should focus on the role of the forensic scientist to properly explain the LR and further note that this concept is often addressed when testifying.

60. Lines 5290-5291 While generally unsuitable for samples containing DNA mixtures, array-based SNP genotyping data can also be used for genetic genealogy searches (Greytak et al. 2019).

We recommend that NIST perform a careful thorough review of the literature before making overly broad statements about utility of specific laboratory methods to interrogate DNA mixtures. We wish to call attention to work that may not lead to the conclusion that arrays are generally unsuitable for such purposes (see Homer et al (PLOS Genetics 2008).

61. Line 5789 “AmpFlSTR Blue,”

NIST could mention CTT and CTTV (kits from Promega Corp.) that precede this kit.

Comments on Definitions:

Allele: one of two or more versions of a genetic sequence; humans typically inherit one allele from each parent; however, sometimes three alleles, called tri-allelic patterns, are seen in STR analysis of a single-source DNA sample; genetic sequence at a particular location (a locus) in the genome alleles targeted in STR analysis can vary by sequence in addition to length

This definition is unfocused and needs rewording.

Allele (or locus) drop-out: loss of allele (or both alleles) information from a DNA profile; failure of an otherwise amplifiable allele to produce a signal above the analytical threshold because the allele was not present, or was not present in sufficient quantity, in the aliquot that underwent polymerase chain reaction (PCR) amplification

If an "allele" is not present, then it does not constitute allele "drop-out"; this definition must properly describe sampling issues.

Amplification: an increase in the number of copies of a specific DNA fragment; in forensic DNA testing laboratories, this refers to the use of the PCR technique to produce many more copies of DNA alleles at specific genetic loci

"Many more copies" could be re-worded to more scientific verbiage.

Binary method: an interpretation scheme in which there are only two values (possible or not possible) for each decision (e.g., a peak is either “an allele” or “not an allele,” or a genotype is “included” or “not included”)

Please clarify; this concept is not well-defined here and is inaccurate.

Contamination: the transfer of irrelevant DNA during an investigation; inadvertent introduction of biological material including DNA alleles into a DNA sample at any stage from collection to testing; it is sometimes easy to identify but has the potential to mislead

"it is sometimes easy to identify but has the potential to mislead" - **this sentence is incomprehensible and does not add to the description of contamination. It seems out of place here.**

Ground truth: information provided by direct observation (i.e., empirical evidence) as opposed to information provided by inference; a situation where the correct answer is known by design

Direct observation does not necessarily equal "ground truth".

Known samples: DNA samples with known genotypes, used for validating methods and assessing proficiency

In forensic science there are other uses for known samples, such as reference samples that are used to generate DNA profiles to compare to profiles for inclusion/exclusion purposes.

Microhaplotypes: regions of DNA containing two or more closely linked single nucleotide polymorphisms (SNPs) associated with multiple allelic combinations (haplotypes); these markers have been explored for mixture deconvolution using massively parallel sequencing due to lack of stutter artifacts

Microhaplotypes have other important features than just lack of stutter.

Proficiency test: a quality assurance measure used to monitor performance of a scientist and identify areas in which improvement may be needed; can be internal (produced by the agency undergoing the test) or external (produced by an outside test provider); external proficiency tests can be either open (where the scientist is aware the samples being tested are a proficiency test) or blind (where the scientist is unaware the samples being tested are a proficiency test)

A proficiency test can also measure/monitor performance of the laboratory system.

SWGDAM: Scientific Working Group on DNA Analysis Methods; formerly known as TWGDAM, Technical Working Group on DNA Analysis Methods; an FBI-sponsored group that develops quality assurance standards and guidelines for forensic DNA and DNA databasing laboratories in the United States and Canada

Correction: SWGDAM does not issue "standards"

PC22

NDAA Comments on DNA report

Nelson Bunn [redacted]

Thu 8/12/2021 1:36 PM

To: **ScientificFoundationReviews <ScientificFoundationReviews@nist.gov>**

Good afternoon,

Please find attached a letter from the National District Attorneys Association (NDAA), a national, non-partisan, non-profit membership association, to provide input during the comment period for the recently released NIST report draft entitled, *DNA Mixture Interpretations: A NIST Scientific Foundation Review*. Thank you for consideration of our input. Please confirm receipt if possible.

Nelson

NELSON O. BUNN, JR.
Executive Director
1400 Crystal Drive, Suite 330
Arlington, VA 22202
[redacted]

Learn more about our CLE Courses: https://ndaa.org/training-courses/

PC22a

National District Attorneys Association
1400 Crystal Drive, Suite 330
Arlington, VA 22202
www.ndaa.org

August 12, 2021

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

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

Dear Dr. Olthoff,

As the oldest and largest association representing state and local prosecutors in the country, the National District Attorneys Association (NDAA) advocates for the use of reliable forensics to exonerate the innocent and inculpate the guilty. With more than 5,000 members nationwide, NDAA is recognized as the leading source of national expertise on the prosecution functions, including forensic science, and is a valuable resource for the media, academia, government, and community leaders.

The admissibility and the weight given to DNA testimony in courts throughout this country are governed by rules of evidence as well as guidelines established by state and federal appellate courts including the United States Supreme Court. We do not believe that NIST should attempt to interfere with these rules and precedent. NDAA also does not support a fundamental change in the respective role of the jury and of experts. Understanding that experts often assume an out-sized presence in a trial under any circumstance, inviting them to both offer opinions and to expound on the value of their own opinions would be a seismic shift usurping the function of the jury. While we certainly support meaningful review and analysis of the accuracy of DNA testing and believe that the draft version of the National Institute of Standards and Technology's (NIST) DNA Mixture Interpretation: *A NIST Scientific Foundation Review* [NIST-Report] illuminates several areas in which further research and review may be fruitful, NDAA has some specific concerns as detailed below.

We read the NIST-Report with great interest. We found Chapter 4, Reliability of DNA Mixtures, Measurements and Interpretation, and Chapter 5, Context and Relevance Related to DNA Mixture Interpretation particularly significant. As to Chapter 4, of particular concern was 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 . . .." [NIST-Report, p. 75.]

The NIST-Report states the following, among other findings: 1) the study is limited because most laboratories do not publish validation data;[1] 2) published internal validation summaries from eight U.S. forensic labs and peer-reviewed articles describing validation experiments lacked sufficient data to allow independent assessment of reliability; 3) while review was limited to publicly available data, additional internal validations exist within individual laboratories; and 4) laboratories face obstacles to making data public, including time required to compile data and privacy rights attached to genotypes used in validation testing. The report also includes tables from the NIST "factor space"[2] review of 60 published, peer reviewed PGS validations and publicly available internal validations from eight public labs, in which each piece of missing data was explicitly noted.

The NIST review team's expressed concern regarding limited publicly available information prompts us to ask the following questions: knowing that publicly available information was limited, and that each laboratory validating a probabilistic genotyping software would have a full set of validation data, what efforts did NIST make to gather that data to conduct a truly informed review? Did NIST canvass the forensic labs, or send letters soliciting the data listed as missing? Better yet, did the NIST review team reach out to those labs and propose an on-site visit?

The benefits of an on-site visit cannot be overestimated. The NIST review team, after signing a non-disclosure agreement to protect private genetic information, could have reviewed the complete validation data, and filled in the missing pieces of data. On-site inspections are common practices at forensic DNA labs, as the accrediting process relies on periodic on-site assessments; labs are familiar with, and would no doubt welcome, such a visit if the goal was to present a complete, accurate and useful foundational review. As it stands, the review is fraught with missed opportunities, because NIST failed to seek the data that would have made a comprehensive review possible.

In addition, NIST treads on dangerous ground when it appears to suggest that the forensic expert should not only report the *results* of their scientific testing, but also opine on the *relevance* of those results to the guilt of the accused. In the Chapter 5 Summary, the NIST-Report says, "the relevance of DNA to a crime cannot be taken for granted and needs to be assessed, because when DNA transferred and whether it transferred directly or indirectly affect its relevance to the crime." It furthers states, "…one must consider the LR within the larger context of the case and ensure that stakeholders do not use the sub-source 'number' alone as an indication of the contribution of DNA to the case." [NIST-Report, p. 141.] For example, finding a defendant's DNA in a rape kit would have different relevance if the defendant admitted sexual contact but claimed consent than if he contested his identity as the assailant. Any suggestion that a witness assess the relevance of their own testimony would invade the province of the judge as the gatekeeper of admissible evidence

[1] This is because scientific journals do not consider such validations novel and therefore do not publish them. At the meeting of the National Commission on Forensic Sciences on January 9-10, 2017, Michael Peat, editor of the Journal of Forensic Sciences, indicated that he will not publish internal validations. [NIST-Report, fn. 23.]

[2] Factor space (and factor space coverage) describes the totality of scenarios and associated variables (factors) that are considered likely to occur in actual casework. [NIST-Report, p. 60.]

and ignore the significant role that jurors and the adversarial process have played in evaluating evidence in criminal courts for hundreds of years.

As our members who have 'boots on the ground' in courtrooms throughout the country can attest, trial by jury and the adversarial process, along with the standard of proof beyond a reasonable doubt, serve as an effective crucible to boil away overbroad assertions and unsupported claims, scientific or otherwise. We urge NIST to place faith in these quintessential American institutions as well and not task the expert with deciding how jurors should think about the evidence they (the experts) provide in the larger context of the case.

It is not the job of the expert to know or assess this context in any event since the accused has a 5th Amendment right to remain silent, the defense need not reveal strategy, if at all, until after the prosecution rests, facts are often contested, and a key role of the jurors is to assess the credibility of the witnesses as well as the content of their testimony. In this task, we are well served by the members of our community who as jurors, take time from their busy lives to perform an essential civic duty. As our membership throughout the country often finds, jurors are intelligent, conscientious, and well able to evaluate the relevance of the evidence for themselves. Bearing witness, perhaps, to the old proverb that 'a handful of common sense is worth a bushel of learning'.

On behalf of our membership, we respectfully request that NIST consider further efforts to work collaboratively with forensic laboratories to gather more data in order to conduct a more comprehensive review of DNA mixture interpretation. We further suggest that NIST consider the role that trial by jury, the adversarial system, and the standard of proof beyond a reasonable doubt play in establishing the relevance of the results of forensic DNA testing in the broader context of the case. Finally, we express our strong support for the use of probabilistic genotyping software in DNA mixture interpretation and recognize the invaluable role such software plays in ensuring a fair and just result by inculpating the offender, exonerating the innocent, and increasing consistency in mixture interpretation among analysts and laboratories.

Sincerely,

Billy West
President

PC23

ASCLD Comments on Draft of DNA Mixture Interpretation: A Scientific Foundation Review

Erin Forry [redacted]

Mon 8/16/2021 2:15 PM

To: ScientificFoundationReviews <ScientificFoundationReviews@nist.gov>
Cc: John Byrd [redacted]; Sudkamp, Laura (KSP) [redacted]

Dear NIST Colleagues,

The ASCLD Board of Directors has reviewed the draft of DNA Mixture Interpretation: A Scientific Foundation Review and has offered comments, attached, for your review and consideration.

Very respectfully,

Erin P. Forry
President
American Society of Crime Laboratory Directors
65 Glen Road, Suite 123
Garner, NC 27529
[redacted]

PC23a

AMERICAN SOCIETY OF CRIME LABORATORY DIRECTORS®, INC.

65 Glen Road, Suite 123, Garner, NC 27529

August 14, 2021

ASCLD Comments to Draft of DNA Mixture Interpretation: A NIST Scientific Foundation Review (NISTIR 8351-DRAFT)

The currently proposed Draft of *DNA Mixture Interpretation: A NIST Scientific Foundation Review* (NISTIR 8351-DRAFT) includes a tremendous amount of information about forensic DNA analysis and specifically, interpretation of DNA mixtures. ASCLD thanks the authors and the DNA Mixture Resource Group for their work on this review.

The following are specific comments on the draft NIST report, our issues, and explanations along with recommended changes, where applicable, to help provide a consensus document:

1. Comment: "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."

Issue: The proposed NIST draft implies that because there is not enough publicly available data, the use of PGS is unreliable.

Suggested Change: ASCLD respectfully requests the authors note publicly available data is available in both the Bright article (Bright, 2018) and the FBI article (Moretti, 2017) for over 3000 samples for DNA Mixture interpretation. These 3000 plus samples were analyzed using the STRmix™ program to determine that the DNA mixture interpretation, as employed by forensic laboratories, is reliable. Prior to implementing any technology, including PGS, an accredited forensic laboratory performs validation studies that encompass the types of samples routinely tested in laboratories. These validation studies also determine the limitations of the technology.

It is not uncommon for laboratories to use DNA samples from casework and laboratory staff, friends, and family. ASCLD requests the authors acknowledge in this report that due to legal and privacy issues surrounding

the sources of these samples, laboratories may not be able to freely share the data observed. It is unclear to the reader what "reasonably accessible" means. While true that data generated by laboratories may not be found via an internet search, laboratories may be able to share such data with those interested in researching and reviewing the data such as academic institutions, assessors, customers, and organizations such as NIST.
ASCLD also requests the authors note that accredited and CODIS participating laboratories are rigorously assessed including an evaluation of validation studies and the underlying data. Policies and procedures developed and used by a laboratory are evaluated against results obtained during validation studies to ensure they are within the scope of the validation.

2. Comment: "KEY TAKEAWAY #4.4: Additional PGS validation studies have been published since the 2016 PCAST Report. However, publicly available information continues to lack sufficient details needed to independently assess reliability of specific LR values produced in PGS systems for complex DNA mixture interpretation. Even when a comparable reliability can be assessed (results for a two-person mixed sample are generally expected to be more reliable than those for a four-person mixed sample, for example), there is no threshold or criteria established to determine what is an acceptable level of reliability."

Issue: The proposed NIST draft implies that the forensic laboratories' studies need to be held to a higher standard than any other science by publishing the underlying data as well.

Suggested Change: ASCLD respectfully requests that Key Takeaway #4.4 be removed. Data is available for review at forensic laboratories and is reviewed by independent auditors through the accreditation process. Validation studies are conducted following FBI Quality Assurance Standards and typically the SWGDAM guidelines. Laboratories determine a threshold or criteria for acceptable reliability dependent upon the various factors unique to each laboratory from their validation studies.

3. Comment: "KEY TAKEAWAY #4.5: Current proficiency tests are focused on single-source samples and simple two-person mixtures with large quantities of DNA. To appropriately assess the ability of analysts to interpret complex DNA mixtures, proficiency tests should evolve to address mixtures with low-template components or more than two contributors – samples of the type often seen in modern casework."

Issue: The use of lower-level complex DNA mixture proficiency tests is not a practical nor a feasible recommendation. Currently, accreditation bodies require proficiency tests to be scored in a binary manner (i.e., pass or fail). Due to the inherent variability of stochastic effects of PCR products of low-level input DNA, it would be impossible to score proficiency tests as pass/fail because the test results variability may not correlate to the proficiency and competency of the test taker.

Suggested Change: ASCLD respectfully requests that Key Takeaway #4.5 be removed or modified to change the key takeaway from "proficiency test" to "challenge test," which are not graded in the same manner.

4. Comment: 4.4 Discussion, line 3201-3204: "Based on an examination of publicly available information reviewed during the time frame of this study, there is not enough information for the authors of this report to independently assess the degree of reliability of DNA mixture interpretation at any one point in the factor space."

Issue: The DNA mixture factor space, as defined by the NIST draft, contains 26 variables (Table 4.1, page 69-70) and as such is exceedingly large and complex. Utilizing the factor space and user defined acceptability, with a potential 10 increments for each variable to cover the factor space, with 26 variables, this would require 403 septillion samples. Accreditation standards dictate that the determination of a method to be "fit for purpose" to meet the needs of the customer is the responsibility of the accredited forensic laboratory.

Suggested change: ASCLD requests that lines 3201-3204 be removed. Validation and determination of a method to be "fit for purpose" is the responsibility of forensic laboratories.

References:

Bright JA, et al., Internal validation of STRmix™ - A multi laboratory response to PCAST, *Forensic Science International: Genetics*, 6-1-2018, DOI: https://doi.org/10.1016/j.fsigen.2018.01.003

Moretti TR, Just RS, Kehl SC, Willis LE, Buckleton JS, Bright JA, Taylor DA, Onorato AJ (2017) Internal validation of STRmix for the interpretation of single source and mixed DNA profiles. *Forensic Science International: Genetics* 29:126-144.

PC24

Comment on

Laird, Jack (SOLGEN) [redacted]

Tue 8/17/2021 8:44 AM

To: ScientificFoundationReviews <ScientificFoundationReviews@nist.gov>

Hello,

Please see attached comments in response to the draft DNA Mixture Interpretation review.

Thank you.

Jack Laird
Senior Manager - Biology
Centre of Forensic Sciences
Ministry of the Solicitor General
25 Morton Shulman Avenue
Toronto, ON M3M 0B1

[redacted]

Ontario

Ministry of the Solicitor General

Centre of Forensic Sciences

25 Morton Shulman Avenue
Toronto ON M3M 0B1
Telephone: 647 329-1320
Facsimile: 647 329-1361

Ministère du Solliciteur général

Centre des sciences judiciaires

25, avenue Morton Shulman
Toronto ON M3M 0B1
Téléphone : 647 329-1320
Télécopieur : 647 329-1361

Ontario

PC24a

August 17, 2021

Via email to scientificfoundationreviews@nist.gov

National Institute of Standards and Technology
Attn: Special Programs Office – Scientific Foundation Review
100 Bureau Drive Stop 4701
Gaithersburg, MD 20899-4701

Re: Comment – DNA Mixture Interpretation: A scientific Foundation Review

To Whom It May Concern,

Thank you, on behalf of the Centre of Forensic Sciences, for the opportunity to provide input with respect to this draft report. We also extend our thanks to the Resource Group, a member of which is from our organization, and we look forward to continuing to contribute wherever we can to the continuous improvement of our discipline.

To that end, we have the following comments for your consideration. Rather than a line by line dissection of the report, we focus here on broader, key elements and thus this feedback is not exhaustive.

The Purpose of the Report

As we understand it, the report set out to assess the foundational validity of DNA mixture interpretation and is one among a series of planned reviews across different forensic scientific disciplines.

If the report's conclusion (i.e. that there is insufficient published data to draw a conclusion – Key Takeaway #4.3) was to be dependent on a review of internal laboratory validation studies, it must have been reasonably anticipated in advance. It is very well known that such studies are rarely published by laboratories, nor accepted for publication by peer-reviewed journals and yet are clearly foundational in the way any method/technology is applied.

Having set that standard, it is also difficult to envision any different an outcome for any future review in any other forensic discipline.

In light of this, assuming the report's conclusion stands, the report itself might be better framed as a vehicle for continuous improvement in the field rather than as a foundational review as to validity which, as it happens, it could never be.

The Random Match Probability

Discussion with respect to the application of the Random Match Probability is, in our view, lacking. For straightforward mixtures of two persons, for instance, it is often readily possible to successfully deconvolute genotypes for both major and minor contributions and to calculate random match probabilities for each, regardless of whether multiple genotype options must be considered at one or more loci for each contributor and including instances where the possibility of allelic dropout cannot be ruled out.

Whether intended or not, the draft report suggests that use of the RMP is limited to a narrower set of circumstances (e.g. lines 1795 to 1798). Perhaps this is due to the fact that many labs in the US historically embraced use of the CPI over the RMP where deconvolution was nevertheless possible. The SWGDAM Interpretation Guidelines for Autosomal STR Typing by Forensic DNA Testing Laboratories (2010 and 2017) include a more accurate description of the RMP and its use.

We believe the report should strive to support the shift toward continuous probabilistic genotyping systems in casework, and away from the use of CPI/CPE and its inherent shortcomings.

Chapter 5: Context and Relevance Related to DNA Mixture Interpretation

This chapter is a welcome addition to the draft report. In our view, it addresses some important elements but should acknowledge that context is the key driver in the relevance of everything we do as forensic scientists. Context is not merely a consideration once test results are generated, it should inform what items are examined in the first place, what if any tests are performed on those items, and how those results are interpreted and reported. And, it is not merely a consideration in instances where one is dealing with low-level DNA (e.g. Key Takeaway 4.2). While the prevalence of low-level DNA results has increased due to the sensitivity of analytical tests, and while the significance of these results must be carefully considered in the context of their respective cases, the same can be said of any forensic test result.

The value of forensic DNA testing as a tool in answering the question as to who is the/a source must also not be conflated with its far more limited value in answering the questions as to what bodily substance the DNA comes from, and when and how it came to be deposited. Developing a more disciplined mindset in relation to the concept of relevance as the driver of our examination strategies and interpreting and reporting results in that framework will help to ensure that we are always providing information of the highest value to the criminal justice system.

Low LRs and Reliability

The draft report, in our view, wrongly associates (or at the very least fails to clearly distinguish) the discrimination power of DNA test results with the notion of reliability (e.g. Box 4.1 suggests that reliability assessments could include assessment of discrimination ability). Low LRs, or high RMPs for that matter, are not unreliable. They are a proper reflection of the strength of the evidence which, most importantly, can only be fully considered in the context of other information outside the forensic scientist's domain.

On the face of it, a LR of 100 or a RMP of 1 in 100 is certainly less discriminating than what is routinely observed with DNA testing of single source samples yielding full profiles, but it is not necessarily less meaningful in the context of any given case.

While we, as forensic scientists, are accustomed to stakeholders making the argument that any LR not in the trillions or higher is somehow less valid, the fallacy must not be repeated or endorsed in a report of this weight. Our laboratory first implemented STR technology in the mid 1990's using the four-locus Quadruplex system developed in the UK. It was no less reliable than the kits we employ today, despite its relatively limited discrimination power.

What is a Complex Mixture?

It appears that what constitutes a complex mixture varies throughout the draft report (e.g. see line 408 vs line 1468). The definition in the glossary is also a subjective one.

Measurement vs Interpretation

In the draft report, measurement is rightly distinguished from interpretation (lines 2702-2706), as they relate to DNA mixture analysis. In our view, it may be beneficial to further consider interpretation as two distinct steps: 1) deconvolution and 2) comparison.

Among certain stakeholders and critics, we are seeing promulgation of the mistaken belief that the Likelihood Ratio (LR) is a new concept, unique to probabilistic genotyping and by linking the LR intrinsically to the notion of 'interpretation', the report does not help to correct this misunderstanding.

While it may be considered semantics, we have always considered the interpretation of DNA profiles, including mixtures, to occur prior to a comparison being undertaken.

Furthermore, Key Takeaway 2.2 states 'Generating a DNA profile involves measuring inherent physical properties of the sample. Interpreting a DNA profile involves assigning values that are not inherent to the sample'. In our view this statement is potentially misleading, as the process of interpretation has two distinct steps. The first step is deconvolution, and necessarily relies on

measurable factors (e.g. stutter, peak heights, peak height ratios at loci, and between them). It is the second step of interpretation, the comparison and assignment of an LR value, to which the distinction rightly applies.

Scientific Foundation Review on the Topic of LRs in Forensic Science

Key Takeaway #4.8 encourages a further 'scientific foundation review' on the topic of likelihood ratios in forensic science. While further study on how LRs are calculated, understood and communicated may indeed be of value to the forensic community, we see no merit in framing it as a foundation review when the LR has been a cornerstone of forensic science (albeit more commonly outside the United States), including DNA testing, for over 20 years.

Sincerely,

Jack Laird

Digitally signed by Jack Laird
DN: cn=Jack Laird, o=Centre of Forensic Sciences, ou=Biology Section, email=jack.laird@ontario.ca, c=CA
Date: 2021.08.17 08:35:12 -04'00'

Jack Laird
Sr Manager, Biology Section

PC25

Defense Forensic Science Center public comments on NISTIR 8351-Draft (UNCLASSIFIED)

Mooney, Kim E CIV USARMY DFSC (USA) [redacted]

Tue 8/17/2021 11:51 AM

To: **ScientificFoundationReviews <ScientificFoundationReviews@nist.gov>**
Cc: Glidewell, Debra E CIV USARMY DFSC (USA) [redacted]; Coursey, Jennifer A CIV USARMY DFSC (USA) [redacted]; Sutton, Joel D CIV USARMY DFSC (USA) [redacted]; Ortiz-Meyer, Delilah L CIV USARMY DFSC (USA) [redacted]; Ogle, Michael A CIV USARMY DFSC (USA) [redacted]; Hill, Michael E (Deputy Director) CIV USARMY DFSC (USA) [redacted]

CLASSIFICATION: UNCLASSIFIED

Good morning,

In response to the request for public comments on the recently published “DNA Mixture Interpretation: A Scientific Foundation Review” document draft, please see the attached, which contains the Defense Forensic Science Center’s (DFSC) comments on the current version of the report.

Respectfully,
Kim

Kim E. Mooney, PhD
Director
Office of Quality, Initiatives, and Training
Defense Forensic Science Center
[redacted]

CLASSIFICATION: UNCLASSIFIED

PC25a

Defense Forensic Science Center (DFSC) Public Comments on NISTIR 8351-Draft

1) Key Takeaway 2.2: "Generating a DNA profile involves measuring the inherent physical properties of the sample. Interpreting a DNA profile involves assigning values that are not inherent to the sample. To do this, the DNA analyst uses their judgement, training, tools (including computer software), and experience, and considers factors such as context."

DFSC comment: Key Takeaway 2.2 omits the important components for interpreting a DNA profile, especially as it relates to probabilistic genotyping software (PGS) systems. Although mentioned in chapter 2, the three scientific principles are fundamental components for PGS methods of interpretation [1]:

- To evaluate the uncertainty of a proposition, it is necessary to consider at least one alternative proposition.
- Evidence interpretation is based on questions of the kind 'What is the probability of the evidence given the proposition'
- Evidence interpretation is conditioned not only on alternative propositions, but also on the framework of circumstances within which they are to be evaluated.

These three principles are based on well-established laws of probability incorporated into Bayes theorem that form the bases for the likelihood ratios generated from these PGS systems. Although the analyst (aka DNA expert) is involved in the interpretation, the method used applies scientific reasoning (Bayes theorem) which provides the necessary scientific support in informing the expert's interpretation.

2) Key Takeaway 2.4: "DNA mixtures vary in complexity, and the more complex the sample, the greater the uncertainty surrounding interpretation. Factors that contribute to complexity include the number of contributors, the quantity of DNA from each contributor, contributor mixture ratios, sample quality, and the degree of allele sharing."

DFSC comment: Additionally the best way to scientifically address the uncertainty described here is with probability. This is specifically mentioned later in Chapter 2 of this report as it "…enables weighting of specific genotype contributions through biological and statistical models…that incorporate mathematical modeling…to reflect uncertainty in the mixture interpretation." This is the underlying principle of PGS software which is why it is the preferred method of choice by forensic laboratories to utilize in these scenarios.

3)Key Takeaway 2.6: "Likelihood ratios are not measurements. There is no single, correct likelihood ratio (LR)."

> **DFSC comment:** These statements are confusing as written and could easily be taken out of context by a non-scientist who may simply regurgitate this statement without reading further. By definition a likelihood ratio (LR) calculated as part of a DNA interpretation is not a precise measurement. However, this does not mean it is not a "correct LR", which would imply that it is not reliable or valid. Even in two person mixtures where the contributors are unknown LR analysis can be used to compare the probabilities of competing hypotheses. As the report correctly points out in chapter 2, "There are no 'true' likelihood ratios, just like there are no true models." This statement better illustrates the application of the LR in a Bayesian construct which is how they are calculated in PGS systems. The LR is simply a ratio of probability assessments for the evidence given competing propositions offered by the court (decision maker).

4)Key Takeaway 4.1: "The degree of reliability of a component or a system can be assessed using empirical data (when available) obtained through validation studies, inter-laboratory studies, and proficiency tests."

> **DSFC comment:** Validation studies are required for ISO 17025 laboratories utilizing PGS systems. These studies are usually available as a part of the discovery process allowing for challenges if inconsistencies are identified. A standard acceptable "degree of reliability" is challenging, as there are inherent sources of variability within the DNA testing protocols between laboratories that take place prior to interpretation (i.e., mentioned in key takeaway 2.6 and further defined as system reliability). These variations are codified within the laboratory procedures and are often referenced in the reports. Also, there is known variability with the different PGS models currently in use, based on individual modeling which will also yield different likelihood ratios in the end. In order to assess component reliability with interpretation using PGS systems, the forensic science community must first define and set the criteria for an acceptable "degree of reliability" to measure given this known variability between laboratories and systems. This criteria is not based on the interpretation or the specific PGS system it actually uses.

5)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 publicly available and to regularly participate in inter-laboratory studies."

DFSC comment: This takeaway is confusing. Did the NIST review team assess the degree of reliability for PGS systems themselves or is it more an individual laboratory's application of the PGS system? These are mutually exclusive determinations. One of the more important questions which was mentioned at the beginning of this report was if established scientific laws and principles exist for DNA mixture interpretation? PGS systems using the likelihood ratio framework and its underlying scientific principles is an appropriate method for DNA mixture interpretation? The lack of published data, does not s in and of itself make PGS methods not reliable. This is especially true where several laboratories have implemented its use and developed validation studies to authenticate its application to DNA interpretation. That these internal studies are not available publically but accessible through the court process, does not negate their existence. Additionally, as mentioned in Chapters 2 and 4, there is an abundance of information in many peer-reviewed scientific publications on PGS modeling discussing in great detail the established principles and computations for generating the results. It would be helpful for the report to clarify more if the PGS method itself has foundational validity based on the research it conducted.

6) Key Takeaway 4.4: "Additional PGS validation studies have been published since the 2016 PCAST report. However, publicly available information continues to lack sufficient details needed to independently assess reliability of specific LR values produced in PGS systems for complex DNA mixture interpretation…there is no threshold or criteria established to determine what is an acceptable level of reliability."

DFSC comment: This takeaway is confusing. The report clearly acknowledges here that there is no defined or generally accepted standard for laboratories to use in order to meet or comply with this expectation for an "acceptable level of reliability" referenced in this report. It further states that "Demonstrating reliability requires that the provider provide empirical data that is accessible to users (i.e., decision makers) of the information for independent assessments of reliability. Agreed-upon criteria from the user are also needed to establish an acceptable degree of reliability." The term "user" should be more specifically defined in the report. This review used one approach (i.e., factor space) for evaluating reliability of existing laboratory data, but this assessment was most likely different than what those laboratories actually did as part of its reliability testing.

For example, accredited laboratories currently use different sets of criteria for evaluating reliability of their preferred PGS system. The current FBI Quality

Assurance Standards for Forensic DNA Testing Laboratories [3] define reliability testing as “the process of testing a software program beyond its functional aspects to ensure it works appropriately in the laboratory environment. This may include testing multi-user or multi-site scenarios, direct-access and network/server-access scenarios, and interaction with other software programs.” Therefore even if the NIST review team determines that the factor space approach is the better method for assessing reliability in these instances, this determination by no means is a consensus view shared by the rest of the scientific community. Even more so when you consider that 1) it is acceptable for laboratories to develop their procedures for PGS usage and interpretation cut offs and 2) the accreditation requirement for reliability is defined completely different.

7) Key Takeaway 4.7: “The degree of reliability of a PGS system when interpreting a DNA mixture can be judged based on validation studies using known samples that similar in complexity to the sample in the case. To enable users of results to assess the degree of reliability in the case of interest, it would be helpful to include these validation performance results in the case file and report.”

DFSC comment: While it is agreed that these studies should be available pursuant to a court determination of relevancy, per se inclusion may be unnecessarily burdensome to the lab for those cases that are adjudicated short of trial. The discovery process mitigates the need to include this voluminous information in every case. Additionally, the lack of publically available studies encourages private, independent studies to further scientific discussion and rigor in the area of interpretation.

References:

[1] Evett IW, Weir BS. 1998. Interpreting DNA Evidence www.boistat.washington.edu/~bsweir/InterpretingDNAEvidence

[2] Gittelson, et al. A response to ‘Likelihood ratio as a weight of evidence: A closer look by Lund and Iyer’ Forensic Science International 288 (2018) e15-e19.

[3] FBI Quality Assurance Standards for Forensic DNA Testing Laboratories, effective July 1, 2020. www.swgdam.org

PC26

Comments on NISTIR 8351

Joel Sutton [redacted]

Thu 8/19/2021 9:00 AM

To: ScientificFoundationReviews <ScientificFoundationReviews@nist.gov>

I am commenting on this report as a member of the Resource Group who was asked to review and provide input early on with this project. First, I will say that the NIST review team have thoroughly addressed the current challenges with DNA interpretation, and cited a wealth of scientific publications, background information, and insight on the subject matter for the reader to reference. Thoughtful recommendations, key takeaways, and discussion points were also provided to help the forensic DNA community better frame the issues moving forward so that improvements in forensic DNA interpretation can hopefully be made.

Having said this, I am disappointed with the fact that the report falls short with fully addressing the original question it was asked to investigate - "What established scientific laws and principles as well as empirical data exist to support the methods that forensic science practitioners use to analyze evidence?" The executive summary and key takeaways mostly hone in on the second part of this question – that there are challenges and concerns with not being able to independently review validation data from the individual laboratories who are interpreting DNA mixtures with probabilistic genotyping software (PGS) systems. But what does this have to do with PGS systems themselves related to the first part of the question? The more important questions in my view for this group to have explored with foundational validity would be "Did the NIST review team find that established scientific laws and principles exist for DNA mixture interpretation? And if so, do they consider PGS systems using the likelihood ratio framework as the appropriate method to apply these same principles for DNA mixture interpretation?"

The document implies in several places that scientific principles do exist (e.g., chapters 2 and 5), as well as mentions a suitable method (i.e., probabilistic genotyping in chapter 2) is in place to assist the provider with this interpretation. These principles with PGS methods are well characterized in the published, peer-reviewed scientific literature. I would argue that there is global consensus in the forensic science community that PGS methods using a likelihood ratio approach to account for the uncertainty with results provide the most useful model currently available for interpreting DNA data. Additionally, many peer-reviewed scientific publications containing specific laboratory validation studies using PG were cited (e.g., chapter 3) demonstrating that much progress has been in recent years to publish data.

Realizing that improvements are needed with individual laboratories publishing their validation data for sufficient independent review, this does not negate the fact that scientific methods do exist for DNA interpretation if properly applied. I would still argue that foundational validity for PGS methods does exist and wish this would have been emphasized more in this report for clarity. Instead, the reader could misinterpret these findings as written such that it would seem foundational validity does not exist for DNA mixture interpretation which is just not true.

Sincerely,

Joel Sutton

PC27

Draft Report on DNA Mixture Interpretation Methods

WICKENHEISER, RAY (TROOPERS) [redacted]

Thu 8/19/2021 11:43 AM

To: ScientificFoundationReviews <ScientificFoundationReviews@nist.gov>

Good Morning,

As you are aware, I was an invited guest of NIST to serve on the Resource Group to provide input for this DNA mixture scientific foundation review. The Resource Group provided the perspective of forensic crime laboratory experience to NIST, however were not included in authorship in the NIST report.

This draft report contains concepts and material that were not previously reviewed by the Resource Group, as no draft materials have been shared in the last approximately 1.5 years. I have a number of concerns with the draft NIST report on DNA mixtures. Three of the most significant are as follows:

1. The data sample utilized by NIST in generating this report is too restrictive and does not accurately reflect validation data used by forensic laboratories. NIST is only reviewing data that is publicly available. Most forensic laboratory validation data is not made public, as it contains staff, friends and family profiles, and individuals providing the samples who did not provide informed consent to permit their DNA profiles to be released into the public domain. Forensic laboratories operate in a secure environment where data must be safeguarded, which runs contrary to NIST's determination that only data published or posted publicly qualify for their foundation review.

   NIST did not make a request to public laboratories to review their data. Much validation data is currently available for defense witnesses, laboratory auditors and assessors review at forensic laboratory premises and has been independently reviewed by these entities. Requiring data to be publicly available as a prerequisite to determining it is valid is an unprecedented requirement by NIST, which is not in place for many other scientific endeavors. Therefore, I feel NIST's requirement that only data that is in the public domain will be used to determine the scientific foundation for DNA mixture interpretation is too restrictive.

   Recommendation: NIST visit forensic laboratories and forensic DNA mixture interpretation vendors and review validation data on site. As an alternative, they could make requests to review such data with appropriate confidentiality measures in place.

2. NIST incorrectly contends that forensic laboratory data has not been independently reviewed. There are 60 publications including DNA mixture studies noted in the NIST report, including one with 1315 samples run by 31 different forensic laboratories [1]. All forensic lab DNA validation studies are reviewed by independent external auditors within their 2-year external audit FBI Quality Assurance Standards requirements, and also by independent auditors from the national accrediting board 4-year audit cycle to meet ISO 17025:2017 standard requirements. Additionally, some states have statutorily created bodies responsible for oversight of forensic laboratory accreditation and approval of such laboratories use of new scientific methodologies and technologies. Many of these bodies have panels of forensic experts who have independently reviewed data and approved probabilistic genotyping of DNA mixtures as fit for purpose. Therefore, in my opinion DNA mixture data validation studies and data have been independently reviewed by objective external forensic experts and been found to be fit for use for individual forensic laboratories.

   NIST authors do not have the necessary practical forensic experience of working laboratories. The stated purpose of the Resource Group was to provide forensic experience that is not possessed by the authors of

the NIST report. Within the Resource Group as well as throughout the forensic science laboratory and DNA mixture interpretation vendor community exists a wealth of forensic experience with forensic laboratory validations, data, forensic casework and samples. I feel the importance of this DNA mixture scientific foundation report warrants inclusion of this experience in review of data, determination of what defines scientific foundation and in authorship of the report.

Recommendation: NIST include individuals with appropriate practical forensic experience to assist with independent review of validation studies and data and co-authorship of the report.

3. The draft report recommends an impracticable standard for validation studies to meet. NIST defines a novel concept of "factor space" including 26 factors impacting DNA mixtures, stating that the publicly available data did not cover this factor space. If every factor were comprehensively covered in a single mixture's "factor space," each of these 26 variables would need to be changed while holding the rest constant to determine the impact of a single variable on the mixture's behavior. Assuming 10 increments for each of the 26 variables, this would require 403 septillion factor comparisons (10 x 26 factorial). This huge number of samples is not practical nor feasible. The factor space model is therefore not appropriate for demonstrating that DNA mixture interpretation as practiced by forensic laboratories is fit for purpose.

   Recommendation: NIST abandon the concept of factor space and develop a more practical measure of what is required to demonstrate fit for purpose and apply that measure to the review of on-site data with additional experts with forensic experience. NIST should then revisit their preliminary report, make the recommended changes herein and include forensic expertise in authorship of the next corrected version.

1. Bright JA, et al. (2018) Internal Validation of STRmix™—a Multi-Laboratory Response to PCAST. *Forensic Science International: Genetics* 34: 11-24.

Please note these comments are my own, and not representative of the New York State Police, ASCLD, SWGDAM, OSAC or any other agency or organization with which I am affiliated.

Regards,

Ray

Dr. Ray Wickenheiser **DPS MBA FAAFS**
Director, NYSP Crime Laboratory System

New York State Police
Forensic Investigation Center
1220 Washington Avenue, Building# 30
Albany, New York 12226-3000
[redacted]

[redacted]

NEW YORK STATE | State Police

www.troopers.ny.gov

PC28

NYC OCME Response to NISTIR 8351

Kupferschmid, Timothy (OCME) [redacted]

Fri 8/20/2021 5:32 AM

To: ScientificFoundationReviews <ScientificFoundationReviews@nist.gov>

Please find our attach response to NISTIR 8351.

Timothy D. Kupferschmid, MBA, MFS
Chief of Laboratories
Office of Chief Medical Examiner
421 East 26th Street
New York, New York 10016
Office: 212.323.1300
[redacted]
Web: www.nyc.gov/ocme

PC28a

Timothy D. Kupferschmid, MBA, MFS
Chief of Laboratories

Craig O'Connor, Ph.D.
Assistant Director / Technical Leader, Nuclear DNA Operations

Eugene Y. Lien, MS
Assistant Director

Charles S. Hirsch Center for Forensic Sciences
421 East 26th Street
New York, New York 10016
Telephone: 212-323-1300 Fax: 212-323-1590
Email: tkupferschmid@ocme.nyc.gov
Official Website: www.nyc.gov/ocme

TO: James K. Olthoff, Acting Director, National Institute of Standards and Technology, U.S. Department of Commerce, c/o scientificfoundationreviews@nist.gov

FROM: Timothy D. Kupferschmid, MBA, MFS, Chief of Laboratories
Craig O'Connor, Ph.D., Assistant Director / Technical Leader, Nuclear DNA Operations
Eugene Y. Lien, MS, Assistant Director, Member of NIST DNA Mixture Resource Group

DATE: August 19, 2021

RE: NIST Draft Report NISTIR 8351, "*DNA Mixture Interpretation: A NIST Scientific Foundation Review*"

The stated goal of this draft foundational review was "What established scientific laws and principles as well as empirical data exist that speak to support the methods that forensic science practitioners use to analyze crime scene material?" (127-128) Despite the coverage of publications, standards, and some history of forensic DNA mixture interpretation, unfortunately, the review failed to meet that goal, and a few areas of concern need to be addressed.

The fact that the review examined only 'publicly available data' (1146-1147, 2402-2405, 2487-2488) does not – as the review suggests it does – inherently imply that the methods used should now be called into question (3201-3207, 3235-3241, 3425, 3484-3486). The review fails to mention the important fact that extensive empirical data exist. Based on the volume of papers that NIST reviewed (to the point where it was stated a comprehensive bibliography was unfeasible to compile due to the constantly growing literature (2456-2458, 3047)), as well as the number of laboratories that have performed their own internal validations of multiplex kits and probabilistic genotyping systems, there is clearly a huge mass of data that exist (2492, 3047, 3057). While it is not publicly available, or all published – which NIST accurately points out is due to such challenges as privacy laws governing genotypes and publishing non-novel internal

validation studies within journals – this hardly negates the existence of extensive foundational data.

There should be a greater emphasis within this review that laboratories have reviewed their own internal validation data and implemented protocols where the analysts use PGS in conjunction with their own interpretations and training (2707-2710, 3204-3207, 3237-3241, 3425). Shared resources are available and utilized by analysts and laboratories that use PGS. There are email listservs, user group meetings, and individuals participating in the development of standards, guidelines and training in mixture interpretation and the use and validation of PGS systems. These components speak to a community that is invested in learning about and sharing data and training with other laboratories to ensure that analysts across the country are using the techniques in a reliable manner. Analysts routinely have access to internal validation data through the course of training and as they work within their laboratory. Stakeholders also are able to receive access to a laboratory's internal validation data for review by their own experts ('users') through the course of discovery requests during court proceedings. Internal validation studies performed by laboratories allow the users of the case data (DNA analysts, stakeholders such as attorneys and hired DNA experts) to evaluate reliability to their samples in comparison to ground truth data. Additionally, oversight bodies, such as accrediting bodies review such studies. In the case of New York, the DNA Subcommittee of the New York Commission on Forensic Science is required to review the validation data from the first laboratory before a new technology within the state is implemented.

Nowhere does it mention within the review as to who would perform independent assessment of reliability of complex mixture interpretation and PGS if data was made publicly available (3201-3204, 3235-3236). There is not even a suggestion that funding could be made available for entities wanting to do this type of review. Based on the recommendations from the PCAST report[1], "NIST should perform such evaluations and should issue an annual public report evaluating the foundational validity of key forensic feature-comparison methods." In addition, concerning the suggestion of sharing validation data between laboratories to increase the factor space coverage, how does this work in relation to current standards within the field? Can one laboratory use the data of another to justify interpreting a case in their jurisdiction?

It is mentioned in the report that "there is a danger of inadvertently viewing results from narrowly-focused studies as applicable to system reliability" (3727-3732). Yet this seems to ignore that several system reliability publications exist (included in the body of work that NIST reviewed (3047)) as well as numerous internal validation studies which would also be a test of system, rather than component, reliability for the user of the method. To whom does this danger apply? Both system and component reliability must be tested for an overall method to function. Studies pertaining to both have been published within peer-reviewed journals for PGS systems (3047).
In lines 3740-3743, the report states "regardless of sources of uncertainty and complexity of the samples, reliability of a PGS system boils down to checking its calibration accuracy and discriminating power at every conceivable scenario described by the factor space." This is neither feasible nor realistic. It also goes against the method that the authors recommend within

[1] President's Council of Advisors on Science and Technology. "*Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods*." August 2016.

the report, a 'bracketing' approach (3463-3474) which would not test every conceivable scenario but bracket the factor space that the user intended for use of the PGS system.

Further, we include in the note below certain clarifications concerning references by NIST to older OCME data as well as corrections to specific issues.[2]

In closing, we respectfully submit these public comments with the hope that revisions will be forthcoming in the final version of the document that will clarify and address some of these issues.

[2] Please note the following specific clarifications and corrections:

- In Table 4.3 (3047), line 31, the Total DNA Quantity Range (pg) is listed as 25 to 500, but the published paper indicates ranges from 6.25 pg to 500 pg as the coverage for determine drop-out rates.

- Table 4.5 (3073) NYC OCME validation of STRmix (v2.4) study is covered. Only a couple of the studies (covering single source and two person mixtures) are noted here as explicitly stating the # samples, total DNA, or mixture ratios. However, there are additional experiments (namely 4, 6, 8, 9, and 10) within the validation summary that state the number of mixtures, mixture ratios, and total DNA amounts that are not covered within this table in the NIST report.

- For the header notes for both Tables 4.3 and 4.5 (3042-3 and 3072), it is stated that the "†inclusion of ranges is not meant to imply that all combinations of DNA quantities and mixture ratios were covered." This should state "…were covered by NIST." Otherwise, this could be misinterpreted to indicate that the quantities and ratios listed were not covered by the authors of each individual study, when the opposite is true - that the entire factor space covered by the individual publication (such as those by OCME) are not mentioned within the NIST report.

- In the header for Table A1.3 (6789) the NYC OCME is noted as an acronym relating to the PG webinar series. The NYC OCME was not a presenter at this series and should be removed from the list.

PC29

comments on Review of DNA Mixture Interpretation Methods

Jaw, Clark (DFS) [redacted]

Fri 8/20/2021 10:14 AM

To: ScientificFoundationReviews <ScientificFoundationReviews@nist.gov>

General Comment:

The report states that the publicly available validation studies do not contain sufficient information to decide whether probabilistic genotyping systems are reliable enough for complex DNA mixture interpretation. However, the way these statements are worded in the report implies that the validations themselves are insufficient, which is not the same. Also, validation summaries are just that – they are summaries. The underlying data for the validation would need to be requested by the reviewers if they wish to examine the validation in greater depth and detail beyond the validation summary. Overall, a more constructive approach could/should be taken to encourage more laboratories to make their validation summaries available to the public and to encourage the "information seekers" to request further details directly from the laboratories.

Specifics from the Report

Line 1389 – extra italic letters

Table from 1618 – seems a little out of date. 2c may also need to address saturation. 2f – baseline noise can also be established using samples; using only baseline noise from negative controls and extraction blank samples to determine analytical threshold is impractical.

Line 2230 – "However, when small amounts of DNA are amplified, the results may not exactly represent the original DNA sample…" – this statement is true but it could be clarified that the difference is typically loss of information and not gain. This statement as-is may be misleading to a lay person.

Line 2350 – It should be "Assessing the strength…in favor **of** a proposition…" it's missing the word "of"

Lines 2403 – 2405 – "However, we believe for information to be considered foundational, it needs to be reasonably accessible to anyone who wishes to review it." This is true, but what do you consider to be "reasonably accessible?" Forensic labs do not conduct internal validations for the purpose of publishing validations for peer reviews. Also, internal validations are conducted for the primary purpose of demonstrating that the method/procedures perform as expected in the laboratory. As such, if a forensic lab chooses to make its internal validations accessible to the public as a courtesy, the validation summaries are usually what is being made available and the lab does so with an implicit understanding that whoever wishes to review the validation in greater depth and detail beyond the validation summary would at least reach out to the lab to request additional information or relevant data. Is the validation data not considered to be "reasonably accessible" this way?

Line 2492 – Table 3.2 – What is listed in the table for the Department of Forensic Sciences (Washington, DC) is incorrect. It should be "STRmix v2.3 parameters & validation report (Identifiler Plus, ABI **3130**)

Line 3073 – Table 4.5 – The information listed for the Department of Forensic Sciences (Washington, DC) is incorrect. DC DFS STRmix 2.3 was Identifiler Plus and the ABI **3130**.

Lines 3235 and 3236 – "However, LR results cannot be externally and independently demonstrated to be reliable without access to underlying performance data." Why did the reviewers not request the performance data from the labs? Do the reviewers feel an external and independent review could not be performed if they have to request data from the labs? Similar to how reading an executive summary of a report is not the same as reading an actual report, reviewing a validation summary should not be equated with reviewing a validation as a whole.

Line 3345 – "They could **detection** ~50% minor…" Detection should be "detect"

Line 3446 – please consider adding other potential reasons such as: (4) size and quantity of information/underlying data, (5) due to the complexity, inability of another to follow the analysis without guidance from the laboratory, (6) availability of the software to the requestor to perform the data analysis

Line 3487 – Key Takeaway 4.7 – this takeaway does not appear to take into account the fact that the recipients and users of forensic laboratory reports are generally court and investigative personnel without science background. It would be more helpful and practical if the takeaway is that the users of results should request the information from the forensic laboratory if and when they choose to conduct an expert review of validation performance results.

PC30

Foundation Review Comment

tiffany roy [redacted]

Fri 8/20/2021 5:51 PM

To: **ScientificFoundationReviews <ScientificFoundationReviews@nist.gov>**
Cc: Nathalie Hicks Champod [redacted]; Jarrah Kennedy [redacted]; Busey, Thomas Andrew [redacted]; Adele Quigley-McBride, Ph.D. [redacted]

To Whom It May Concern,
I have attached the comment submitted by the Human Factors in Forensic DNA Analysis Testimony and Reporting Subgroup.

As a group working on a large consensus document, we recognize the time and effort that has gone into the foundation review and wish to congratulate the team on this tremendous work.

PC30a

NIST foundation review comments from the Human Factors in Forensic DNA Analysis testimony and reporting subgroup written by Tacha Hicks Champod and reviewed and adopted by the subgroup members Jarrah Kennedy, Tom Busey, Adele Quigley McBride, and Tiffany Roy

First congratulation to the authors for this extensive review and thank you for giving us the opportunity to comment this thorough work.

In general, as forensic scientists, for evaluation of our results, should we not first refer to the principles of interpretation :
1. Scientists assess their results in a framework of case circumstances (also to chose the relevant method of analysis)
2. Scientist assess their results given two propositions
3. They do not give an opinion on what has happened but on the value of their results given what (allegedly) has happened. This ensures we have an approach that is balanced and based on logic.

page 13:
Uncertainty: definition
There are also events we are uncertain that have no true value and cannot be measured: for example that Mary Queen of Scots knew of the plot to murder her husband. see, C.E.H. Berger et al. / Science and Justice 51 (2011) 43–49.

Summary:
line 553: *or whether there is a trace amount of suspect or victim DNA make DNA mixtures inherently*: how can we know this? Do you mean DNA compatible with a person of interest (suspect, victim, other).

line 559 : *When laboratories analyze high-quality, single-source samples:* A single source is not necessarily more relevant than a mixture, and indeed contamination are usually not complex mixtures... See the Jama or Adam Scott case.

Line 563: *a wide range of variation in how specific DNA mixtures are interpreted.*
Yes, but also true for paternity cases see ISFG challenges...And applies to analysis as well.

Line 566: *Chapter 1 introduces the topic of DNA mixtures (samples that contain..:* A sample is a part of something, we hope that we do not sample crime stains.

Line 609: *Therefore, this scientific foundation review does not concentrate on interpretation of single-source DNA samples and two-person mixtures involving*: yes, however single stain again can lead to miscarriages of justice...One of the issues being that laypeople think a case can rest on DNA only.

Line 616: *Correlated with overlapping alleles*: having DNA that could also come from a relative makes things more difficult.

Line 632: *that is an evaluative interpretation of the strength of this association*: why not say a comparison: it is not necessarily an association

Line 639: not really *a tested DNA sample*: more an analyed DNA trace. Also, one important step is to examine the EPG to decide if there is value, you determine the profile of the trace and then you usually compare it by eye to a K.

Line 643: case information is also key when assigning a LR.

Line 648: also if the amounts are 1:1:1 it is more complex than 6:3:1

Line 649: can the use of replicates be useful to manage stochastic effects?

LIne 654 KEY TAKEAWAY #2.1: *DNA mixtures, where the DNA of more than one individual is present in a sample, are inherently*: are generally ...

Line 680: *judgment on different kits* -> on different results rather than kits only.

Line 698: *A highly reliable method is one that consistently produces accurate results.* Accurate? But how come if there is no true value?

Line 754: *there is no threshold or criteria established to determine* ...: one can use Turing's rule. uring stated that the expected LR for a false proposition is one if the model is correct [I.J. Good, Probability and the Weighing of Evidence, Charles Griffin & Company Limited, London, 1950, p. 72.]. Turing's rule informs us that the fraction of non-donors producing an LR x is expected to be at most x 1 . Mathematically stated: Pr(LR x|Ha) x , x > 0

Line 764: *This introduces variability and uncertainty in DNA mixture interpretation*: yes, variability, but not really uncertainty.

Line 772: *degree of reliability in the case of interest, it would be helpful to include these validation performance results in the case file and report...:* in the case file yes, but it is difficult to see how it would be of use in a report (statement).

Line 776: LRs are not really calculated, they are assigned on what we know, we are told and what we assume.

Line 781: can you define *relevance*?

Line 849: *There is a need for more structured research for assigning the value of the results*. Also, a strong need for education.

Line 868: KEY TAKEAWAY #5.4: *DNA statistical results such as a sub-source likelihood ratio*: there is no such thing as a Sub-source LR, it is a LR given sub-source propositions. It is not the hierarchy of LRs, but the hierarchy of propositions as we are not supposed to give an opinion on the sub-source, the court is.

Line 950: Why invoke uniqueness here.. everything is unique but how does this help? In fact, it might do more harm than good.

Line 956: *or whether there is a trace amount of suspect or victim DNA*: we cannot determine that, this would be transposing the conditional.

Line 1022: Another factor is a similar contribution (1:1)...There are also factors that allow to decrease complexity: taking into account all data (peak heights) and all case information (DNA profiles of all POIs).

Line 1030: it is difficult to be so general, it will depend on the information that is available regarding the different contributors. A complex mixture 10:3:2:1:1 will provide different information depending on the comparison. This will be reflected in the LR.

Line 1039: *produces accurate results*: is accuracy the key ? It is the order of magnitude that matters, is it not?

Line 1061: You might also have a trace in large quantity that is irrelevant all depends on the circumstances of the case.

Line 1079: *should be used during all stages of an investigation, including at the crime scene*: yes, but also the victim and hospital as we have seen.

Line 1101: *This is especially so in cases involving very small quantities of DNA*: we can also mention when there is a legitimate alternate activity that could explain for the presence of the DNA of the POI.

Line 1207: *regarding concerns in DNA mixture analysis*: one important point is the use of meaningful propositions

Line 1216: one can use VARNOC, if they have a version with that feature

Line 1218: *Addressing report writing and content*: yes, especially avoiding transposed conditionals.

Line 1236: we should be careful here: *Improved understanding of secondary transfer possibilities*, this is not really what we need to do as this would be a transposed conditional. We need understanding of our results given activities that imply secondary transfer.

Line 1337: *include (1) ability to identify an individual or associate a perpetrator with a crime scene.* DNA does not allow to identify a person, it contributes to identity. How can we ever identify a perpetrator? This is a decision of the court.

Line 1360: in order to provide meaningful answers, we need case information to formulate meaningful propositions. This is key!

Line 1485: we cannot estimate the value of evidence: there is no true value. We assign the value of the results

Line 1519: *available relevant case context information* (e.g., location from which the sample originated, body fluid screening results, quantity of DNA extracted, and overall quality of the DNA profile) Some of these are not case information: body fluid screening results, quantity of DNA extracted, and overall quality of the DNA profile are results that needs to be assessed. Case information is assessed by the court.

Line 1526: Buckleton et al. 2005), there is a new edition.

Line 1529: it helps distinguishing information (case context) and factors that we use in our interpretation (transfer, prevalence, rarity of a DNA profile). Case context is not really a factor, is it?

Line 1576: *Stutter products are the most influential artifacts in an EPG*: since then we have learnt a lot on stutters and can model them more easily that pull-ups which cannot be yet modelled in PGS.

Line 1585: a solution to pull ups is to put less DNA amount...

Line 1628: one very goofy side effect of PGS is that a lot more work for implementation is done compared to before PGS. Now, laboratories are more aware of the expected variation.

Line 1650: isn't the first step to decide whether the profile is of value or not?

Line 1727: *Interpretation methods need to be able to account for this ambiguity*: what about the analysis? Should one also do replicates?

Line 1765: yes, one can use RMP for single trace (or pseudo-single) however, would it not be easier to have just one metric for all DNA analysis? That is Kinship, missing persons, crime stains single and mixtures ?

Line 1782: why not say that RMNE are unhelpful for court purposes? It is now time to shift to more sensitive and specific methods of comparison.

Line 1787: we do not randomly select, we chose in a known manner.

Line 1887: the term 'support' is better than is favour: because it is not because the LR is a billion that the hypothesis favoured will be the first proposition. All depends also on the other information in the case.

Line 1892: depends more importantly on case information on which the propositions will be based.

Line 1911: explain that subjective is meant as personal (it does not mean arbitrary).

Line 1923: remove '*among other things*,' and replace by for example (indeed propositions need to be exhaustive in the context of the case) and also mutually exclusive of course.

Line 1937-53: be careful this is not really an ideal analogy, normally we have the results that have a given weight that will shift the scales of justice. But, all depends on what the scales look like without the DNA findings. Consider first showing a scale where panes are even, and then your two examples. What would be best is to see the shift. Do not use the term favour as it is ambiguous.

Line 1949: it is not ideal to say *'DNA from someone else*: who is that person? A twin? A sib? An unrelated person? It is best to say for example an unknown unrelated person as it is less vague (thus easier to assign the probability of the results given this proposition).

Line 1973: would be "*DNA evidence found on the item is one...* Best to say: The DNA results are..than "DNA evidence found on the item".

Line 1998: it might be worth saying that in the same lab, the same verbal equivalents (verbal scale) ought to be used.

Line 2012: as there is no true LR there cannot be an estimate. estimate of the statistical strength change to ->a statistical value of the DNA comparison in the context of the case.

Line 2083: it is sufficient that there is a possibility that the victim's DNA is present. See Buckleton et al. When evaluating DNA evidence within a likelihood ratio framework, should the propositions be exhaustive? 2021. It all depends on the case, it could lead to a smaller LR. What it does, is that we have better sensitivity and specificity.

Line 2089: we could also use Fst maybe indicate this as well.

Line 2093: it is not really meaningful to have LRs as precise (except when we want to compare models). Maybe note that it is the order of magnitude that is important. So labs will only give one significative number.

Line 2108: DNA contributes to identify but one cannot identify a person only based on DNA

(Tacha) Line 2124: *'Empirical data for assessing the fitness for purpose of an analyst's LR are therefore warranted.*' You have shown the importance of propositions, should not this be the key point? That propositions should be justified based on case information and a caveat should emphasise that if the information changes so will the value of the results? I do not understand how empirical data would be helpful.

(Tiffany) I do believe there should be some study on whether the LR is well suited for forensic DNA. It does not address the question "who could be/is the source of the DNA." Many of Tacha's comments allude to references in this in this very document where there are references to 'identifying whose DNA might be on an item', which under the framework the analyst cannot do. And this is with the technical assistance of the experts at NIST.

Line 2201: it is important to distinguish frequency (which is counted) from probability. One does count alleles, or haplotypes but one does not count genotypes.
For Genotypes, we use a model, with assumptions. It is therefore helpful to distinguish both concepts (frequencies of alleles, probability of a DNA profile).
By' population frequency calculations made' should one understand probability or frequency? Frequency are not really calculated, are they?

Same for line 2211.

Line 2225: can help answer (not directly answers).

Line 2350 and 2352 : in favour can be misunderstood (supporting is better). One should say 'exhaustive in the context of the case' one cannot be exhaustive in general.

Line 2363: *'The framework of circumstances includes the hierarchy of propositions'* this is unclear. The hierarchy of propositions is a concept that is useful for our thinking as forensic scientists it is not part of the framework. This point would read best in point 1. '---one alternative proposition. Propositions can be classified according to a hierarchy: sub-source, source, activity and offence level propositions.

Line 2720: an important step is whether the DNA profile has sufficient information and is interpretable.

Line 2726: given propositions not for...

Line 2864: *Case specific information (propositions) and assumptions*: this is the most important point, it should come first not last. (why decision? It is a given, there should be no decision).

(Tacha) Line 3074: KEY TAKEAWAY #4.3. This is quite a statement, there has been a lot of empirical testing published on PGS, a lot more than on RMP or RMNE (or the use of LRs for kinship or missing persons). For RMP has there been a lot of empirical testing? Our LRs are based on models, so it really depends on how much we trust (believe in) the model.

(Tiffany) The document distinguishes between data and interpretation very well. There have been studies performed on PGS systems, and published articles containing interpretations of the data from those studies are many. But the underlying data for the publications is still not *publicly* available for independent inspection. PGS developers suggest the data could be made available on request, but it's uncertain if each request would be treated equally, whether the data would be organized and identified the way the FR describes or dumped without explanation, stripping it of all utility. KEY TAKEAWAY 4.3 speaks more to transparency of data, not number of research studies or peer reviewed publications.

Line 3426: One can use Turing's bound for Hd true experiments.

Line 3460: "*analysts and different laboratories will have different*..." this does not introduce uncertainty, but certainly variability (but laboratories will also use different methods, so that is also a source variability, or will even collect the trace differently). The variation among laboratories and interpretation is true for kinship cases also: a key point is using the same information and knowledge (through education). Also, one must remember that it is the order of magnitude that is important.

Line 3594: The use of LRs is well established and has been for many decades. How they are understood is another matter.

Line 3737: correctly modelled (indicate meaningfully modelled) as there is no correct model.

Chapter 5: if there is an AT given, the type of sequencer should be indicated as well. One important point is when a person will decide whether the DNA profile is of value. How this is decided is highly variable.

Line 4469: Formulating propositions (the court adresses them)

Line 4469: Graham Jackson et al were the ones to suggest that we had two roles: one in helping the investigation and one helping with the court. See Jackson, G., Jones, S., Booth, G., Champod, C., & Evett, I. W. (2006). The nature of forensic science opinion - a possible framework to guide thinking and practice in investigations and in court proceedings. Science & Justice, 46, 33–44.

Line 4518: indicate given sub-source propositions.

Line 4688: it is not well understood that in fact source propositions help address also the 'who' and not the the what (what is the nature). Propositions that are called source (for all forensic disciplines) are as follows: A is the source of the blood and unknown is the source of the blood. One assumes it is blood. This is because when the hierarchy was 'invented' it was only possible to analyse DNA from known body fluids. Source regards question: is the POI the source of the DNA (it is not about association)
Activity: question being asked. The term transfer should not appear: one can assess the absence of a DNA profile.
Offence level propositions can be useful to combine different activities. Scientist cannot assess motive, opportunity and the like. But, they can assess their findings (from different disciplines or from different traces) given offence propositions (which might imply several activities).

Line 4692: this is not an appropriate way to formulate a proposition. One should not have the term 'deposit'. One could say: Mr A had vaginal-penile intercourse with Ms B or they only had social activities e.g. (eating and watching a movie). Activity level propositions are about activities and deposit is not an activity.

Line 4727: to calculate an activity-level LR -> to assign the value (i.e., LR) of the results given activity level propositions

Line 4728: no, the term transfer should not be there. "..*be that the DNA was transferred to the handle of a knife*" replace be that Mr S stabbed the victim. and the defence alternative that Mr S has nothing to do with the stabbing. (contamination or transfer should not be in the proposition).

Line 4733: saying LR activity can lead to misunderstandings with layperson, we do not assess the activity but our results given the activities.

Line 4764: no, we do not assigning probabilities to propositions...this is a transposed conditional

The only way to assess multiple samples (for example in a sexual assault kit) is to consider the activities. Else, we cannot combine results.

In Europe one can mention the certifying education, see E-learning initiatives in forensic interpretation: report on experiences from current projects and outlook
Biedermann A., Hicks T., Voisard R., Taroni F., Champod C., Aitken C., Evett I., 2013/07. Forensic Science International, 230 (1-3) pp. 2-7.

PC31

Comments

Roberto Puch-Solis (Staff) [redacted]

Sun 8/22/2021 6:01 AM

To: ScientificFoundationReviews <ScientificFoundationReviews@nist.gov>

Please find attached my comments on the Scientific foundation review on DNA mixtures.

PC31a

Comments on “DNA Mixture Interpretation: A NIST Scientific Foundation Review”

R. Puch-Solis

Leverhulme Research Centre for Forensic Science,

University of Dundee, UK

The NIST report promotes a deeper understanding and application of forensic DNA interpretation. I have some comments below.

1. Wording of Principle 10

In line 2260 the document states:

> “Principle 10 [Measurement]: Relative fluorescence unit (RFU) variance (uncertainty) is inversely proportional to DNA profile peak height”.

The principle can be written as:

> “Principle 10 [Measurement]: The variability (uncertainty) of peak height ratios (and heterozygote imbalance) increases as peak height decreases”.

The reasons behind the suggestion of wording are given below. Firstly, it is not the variability of peak heights that increases as peak height decreases. Peak height variability may increase with peak heights. It is the coefficient of variation (standard deviation divided by the mean) that increases as peak height decreases. This is reflected in the variability of peak height ratios and heterozygote imbalance. Secondly, the word proportional is not adequate. It has a suitable meaning in lay terms: “corresponding in size or amount to something else“, (Oxford English Dictionary). However the description of Principle 10 is on mathematical terms and “proportional” has a specific mathematical definition: “(of a variable quantity) having a constant ratio to another quantity” (Oxford English Dictionary).

2. Application of likelihood ratios

In Line 839 the report states “It is imperative that the likelihood ratio be considered in the context of other evidence in the case”. However, this important statement is not included as one of the key takeaways.

The report defines (line 1931) the likelihood ratio in line 1931:

$$LR = \frac{\Pr(E|H_1, I)}{\Pr(E|H_2, I)}.$$

However, it is important to show that the LR is part of the odds form of Bayes' theorem:

$$\frac{\Pr(H_1|E, I)}{\Pr(H_2|E, I)} = \frac{\Pr(E|H_1, I)}{\Pr(E|H_2, I)} \times \frac{\Pr(H_1|I)}{\Pr(H_2|I)}.$$

The LR quantifies the support to either H_1 or H_2 given by the DNA evidence E. However, this is combined with the support provided to the propositions by information heard before E,

$$\frac{\Pr(H_1|I)}{\Pr(H_2|I)}$$

In practice this number is not quantified and the jury uses their common sense to combine it with the LR. A key takeaway is that the LR is not a standalone number and that it is combined with previous support to either H_1 or H_2. Without this consideration, the likelihood ratio may be mistaken for a classifier: if it is greater than one an inference of inclusion is indicated, while if it is smaller than one an inference of exclusion is supported.

3. Not all PGSs use MCMC

In line 2013 the report states "Weighted genotype possibilities can be estimated using Markov chain Monte Carlo (MCMC) simulations to assess possible combinations of parameters considered in deconvoluting potential contributor genotypes". MCMC is not used in all systems, e.g. likeLTD and EuroForMix do not use MCMC. The sentence can be re-written by simply removing the MCMC part: "Weighted genotype possibilities can be estimated to assess possible combinations of parameters considered in deconvoluting potential contributor genotypes". It is also worth noting that not all PGSs display weighted genotypes as part of an LR calculation.

4. Evidence E in DNA interpretation

In line 1931, the likelihood ratio is defined as

$$LR = \frac{\Pr(E|H_1, I)}{\Pr(E|H_2, I)}.$$

E is introduced as the evidence in line 1920. Terms for the questioned and reference samples, Q and K, have been already introduced in lines 629-630. It would be helpful for the reader to mention that the DNA evidence E is Q and K together.

5. Additional reference

In line 2009 a list of examples of continuous models are given. The model of Graversen & Lauritzen (2015) should be included, as it is one of the most sophisticated Gamma models and precursor to EuroForMix. The reference is:

Graversen, T. & Lauritzen, S. (2015). Computational aspects of DNA mixture analysis. Statistics & Computing, 25(3), 527-541.

6. Application of the bracketing approach

In line 3649 the bracketing approach is described and suggested as a sensible approach to assess the reliability of a system:

> "A bracketing approach which considers results from samples that are more complex and less complex than the casework sample of interest, is a sensible way of understanding case-specific reliability of the system."

The application of this approach assumes that all laboratories are able and can afford to produce samples of about the same complexity to questioned samples encountered in casework. It would be worth asking laboratories whether this is the case before the publication of the report.

PC32

Comments on NISTIR 8351-DRAFT

Lloyd Halsell III [redacted]

Mon 8/23/2021 11:13 AM

To: ScientificFoundationReviews <ScientificFoundationReviews@nist.gov>
Cc: Cheron Maxwell [redacted]

On behalf of the Houston Forensic Science Center, I submit the following comments and questions concerning the NISTIR 8351-Draft, DNA Mixture Interpretation: A NIST Scientific Foundation Review.

Upon reading Chapter 4 and arriving at Key Takeaway 4.3, we did not reach the conclusion that "there is not enough publicly available data to enable an external and independent assessment." In fact, after reaching Key Takeaway 4.4 which states "Additional PGS validation studies have been published since the 2016 PCAST Report. However, publicly available information continues to lack sufficient details needed to independently assess reliability", we were left with the understanding that the quantity of data available is sufficient, but the quality and formatting of the data need some improvement.

Furthermore, after finishing Chapter 4 this conclusion was reinforced. In particular with Table 4.9 that makes the recommendation to "adopt a community-wide uniform approach to" both publishing PGS validation and sharing internal validation information.

Key Takeaway 4.3 is very misleading in its wording and location within the draft document. Would it be better to limit Key Takeaway 4.3 to encouraging forensic laboratories to make their underlying data publicly available while adopting this community-wide uniform approach?

Key Takeaway 4.4 then can be better used to promote this community-wide uniform approach for both publications and sharing of internal validations.

Finally, would the authors be willing to add language, to both the report and the Executive Summary, that emphasizes the Key Takeaways must be evaluated within the context of the entire report. We understand the authors ultimately cannot claim responsibility for how others will use the report. However, they should wish to know that if the report is used to support a position, the entire context of the report was done to do so.

Lloyd Halsell III
Operations Coordinator FBIO
[redacted]
Houston Forensic Science Center
500 Jefferson St. 13th floor
Houston, Texas 77002

Houston Forensic Science Center

Help us improve! We would like to ask you a few questions about HFSC and our services. Please take this short survey.

www.houstonforensicscience.org

PC33

Comment on Mixture Review

DeGroot, Gretchen A. [redacted]

Mon 8/23/2021 12:02 PM

To: ScientificFoundationReviews <ScientificFoundationReviews@nist.gov>
Cc: Rasmussen, Amber L. [redacted]

Response of Wisconsin DFS DNA Technical Leaders Gretchen DeGroot and Amber Rasmussen

In the presentation on July 21, 2021, it was stated that this document was not intended to be used to determine if PGS should be admitted in court. In a discovery demand sent to our lab on June 15, 2021, this document was cited in order to question the use of PGS in the case and included the quote: "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". Despite our willingness to share our validation summary, the inclusion of this quote from the draft document implies to the court that the validation was insufficient even before it is seen by the defense expert.

[4.1.2, lines 2776-2778] "there is no true LR …[and]… there is no uncertainty associated with an LR assessment."
PGS gives the lab an opportunity to incorporate multiple levels and areas of uncertainty into the LR calculations. It would seem appropriate to acknowledge that in this section.

[4.1.4]:
In the past, it was suggested that labs were including too many samples in their validations [Butler, 2006]. Do the authors still believe that previous validation studies included excessive numbers of samples or is this a revision of that opinion?

The "factor space" information was never previously included in any document to inform labs of the information needed (per this NIST group) to assess the validations or studies. This type of information is examined as part of a quality validation, but it is next to impossible to test the full extent of what is seen in evidence. It is also impossible to know the ground truth of the evidence profiles so laboratories can only provide the best estimate of the "factor space" covered in forensic casework. The sheer number of samples required to investigate the full extent of the variables outlined in Table 4.1 is unrealistic to expect in forensic internal validation studies.

It is clear based on the information provided in this report, that the criteria to assess the reliability of current mixture interpretation methods examined by the NIST group evolved during the study. These criteria have never been documented in any other location of which we are aware. This fact makes it impossible for those doing studies (validations and research papers) in the past or present to meet these unknown assessment criteria.

[Key Takeaway #4.3]:
The WSCL internal validation summary is freely provided upon request, as part of discovery demands and in open record requests. The validation study has been shared publicly by the Brooklyn Defenders Services at https://indefenseof.us/issues/kinship-problem#:~:text=An%20upcoming%20report,the%20links%20below. This document was obtained through an open record request. This is the same method that others could obtain our validation summary. In addition, the WSCL DNA TLs would have willingly sent the validation summary to NIST if a request had been received.

Underlying PGS validation data typically includes laboratory staff and family DNA profiles. Public release of this information would be detrimental to the lab's ability to obtain a wide array of samples to use in order to sufficiently cover the "factor space". Current algebraic methods to mask contributor profiles (as suggested in Box 4.1) have been demonstrated to be easily reversed (See response submitted by ESR). What criteria does this group use to state that the underlying data must all be available in the public domain, as opposed to any other area of scientific study in which publication is generally the only public release of data? A thorough validation summary should be sufficient to detail the type of testing performed and samples included without the need to release data generated from donor profiles. Accreditations and assessments within the forensic community provide the opportunity for outside forensic professionals to assess the underlying data and its representation in a summary.

[**4.4 Discussion]**
In casework, we on occasion have incest cases. We are unable to replicate the "factor space" of allele sharing we see in these cases. Are we not to apply knowledge of inheritance patterns to appropriately interpret the evidence in these cases? What evidence is there that increased allele sharing interferes with the validity of mixture interpretation or the ability of ProbGen systems to appropriately deconvolute these types of profiles? We have done extensive family studies. We clearly state we will consider and run other hypotheses, and we request eliminations standards from others who are thought to be involved. If we do not get this information, how are labs to respond? Labs are not the only ones responsible in our legal environment.

The same goes with the lab's formulated propositions. The lab gives the LR for the proposition that applies to the questions asked of them and may perform investigative LRs, but it is the responsibility of the two sides to propose other options to test if the proposition reported do not address other questions that are asked in trial.

[4.4.1, lines 3267-3268]:
There is a reason that forensic DNA analysts are qualified as expert witnesses. It is unrealistic to expect members of the court (judge, jury, and attorneys) to understand complex scientific principles without having the needed education, training and experience. It is most important that the expert witness has an appropriate understanding of the studies performed in order to accurately guide the court as to the reliability of the method.

[4.4.1, lines 3288-3299]:
Papers like the response to PCAST report by Bright et al obviously address the testing of "mixtures in which a sample is present at an extremely low ratio." Each lab uses its own sets of samples for testing, so there were at least 31 different sets of individuals used for the mixtures in this study. Based on Buckleton's website, there are 61 labs just using STRmix. Each lab used its own sets of samples and created its own mixtures; therefore, that is an even greater variety of profiles. When a laboratory includes low quality data that results in LRs close to 1, or even defines an uninformative result range based on known contribution to validation samples, that is demonstrating the effectiveness of PGS to assign appropriate weight to uninformative or unreliable underlying data.

[4.4.3]:
When desired information was not provided in the available documents, was there an attempt made to request this information from the authors? Most would gladly provide the solicited information upon request. In addition, per [2006 Butler paper], the type and range of samples tested is more important to testing a method than a simple count of the number of samples run. Most laboratories would include this information in their validation summaries if the greater forensic community deemed it an important factor in reliability determination. Editors of forensic journals should request this information in published studies if it is an area deemed to be of value.

[Key Takeaway #4.4]:
What specific details are typically missing from these summaries that the authors feel are required to appropriately assess reliability? When PGS methods incorporate varying levels of profile reliability into their calculations and assign appropriate weight based on that reliability, what would the authors suggest is used as a threshold or criteria for reliability?

[Key Takeaway #4.5]:
Given all that is involved with Proficiency Tests (PTs) and that an analyst's work is dependent on a successful PT and getting the "correct" answer, it is unlikely that labs will willingly agree to more challenging samples for PT's.

With the variability known to occur in samples of low-template or more than two contributors, how do the authors suggest PT results for samples of this type be consistently evaluated to result in a simple pass or fail? The manufactured aspect of current PT samples often results in sample processing difficulties not commonly observed in forensic casework samples. Do the authors have any suggestions for how to overcome these obstacles? Without any achievable plan to consistently evaluate results, this suggestion is unsupported by the reality of forensic DNA testing.

[Key Takeaway #4.6]:
Variability in interpretation is often related to the lab's validation, the instruments used, the cycling conditions, the kits, the lab's TL's tolerance for uncertainty, and even fluctuations in the weather. These differences will continue to provide variability.

It would be helpful to outline the expected improvements, how the authors expect those to be implemented, and the progress made towards those improvements to date.

[Key Takeaway #4.7]:
Including validation performance results in the case file and report is unrealistic. We can provide this information as part of discovery demands but our "customers" have made it very clear to us they want the reports short and to the point. We try to do that while still including all the needed information by including a technical appendix.

[4.4.6, lines 3523-3524]: "The degree of reliability or trustworthiness of a given PGS method in a given case is dependent upon the number of instances where that method has been tested."
This contradicts the suggestions made in [2006 Butler paper] that suggest sample type is more important than quantity. It would be helpful if the authors defined an acceptable method to determine the number of samples required to appropriately assess reliability.
[4.4.6, lines 3554-3556]: "judgments of reliability by decision makers or triers of fact will be helped by comparing LR assessments from multiple systems and made by multiple experts"
This seems to suggest an apples and oranges comparison. Since data generated in different labs may utilize a variety of kits with varying loci, discrimination power and amplification efficiencies, results obtained for the same extract processed in different labs are not expected to generate the same LR values. Similarly, each validation uses lab specific data to establish PGS parameters appropriate for their own data, resulting in the inability to accurately interpret data from other laboratories. An LR produced from a Profiler/Cofiler sample run on a 310 would be expected to result in a much lower LR than data produced from Fusion 6C and the 3500xl. Increased information will result in greater discrimination potential and increased LRs for true contributors. Individual labs are not typically in a position to support multiple systems and validate each of them.

[4.5.1, lines 3613-3614]: "important information is not explicitly stated in the referenced publication."

Presumably, the forensic community, to include the editors and reviewers of these publications, did not feel that the lack of this explicitly stated information was integral to the understanding and scientific value of the work performed. The peer review process inherent in publication is generally accepted in the greater scientific community as being the time and place for colleagues in the field to raise concerns about the adequacy of the data provided.

[Chapter 5.]:
The topic of chapter 5, although of value, does not belong in this document. It clouds the focus on mixture interpretation. I know of no labs in the US who are trying to report results above the sub-source level without additional data to support a possible source conclusion (i.e. semen or blood identified in sample).

Butler J.M. Debunking Some Urban Legends Surrounding Validation Within the Forensic DNA Community, *Profiles in DNA*, 2006, pg 3-6.

Gretchen DeGroot and Amber Rasmussen

GRETCHEN A. DEGROOT M.S.| DNA Technical Leader
State of Wisconsin Department of Justice
Division of Forensic Sciences
Wisconsin State Crime Laboratories
[redacted]

PC34

Public comment for NISTIR8351 draft

Beth Hewitt [redacted]

Mon 8/23/2021 12:28 PM

To: ScientificFoundationReviews <ScientificFoundationReviews@nist.gov>

To whom it may concern –
I have attached our laboratory's brief response/comment for the NISTIR 8351 draft (DNA Mixture Interpretation: A NIST Scientific Foundation Review).

Thanks.

Beth Hewitt
DNA Technical Leader
Jefferson County Regional Crime Laboratory
[redacted]

PC34a

This laboratory has concerns about the overall tone of the NISTIR 8351 DNA Mixture Interpretation: A NIST Foundational Review (NISTIR 8351), but the biggest issues revolve around one of its primary conclusions – Key Takeaway #4.3 (Page 75, Line 3074), which states in part, "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 (PGS) systems."

Firstly, the authors focus on a lack of validation data available from individual forensic DNA laboratories. This perceived lack of data is then misconstrued as the lack of TESTING performed by laboratories. The authors clearly did not conduct a thorough pursuit for laboratory data. They did not reach out directly to forensic laboratories, which should have been a relatively easy task given the resources and contacts available to the authors. Instead, their method to obtain the data was an internet search. This is hardly a sufficient method of obtaining data for an organization such as NIST; it's more akin to a middle school student researching information for a school project. It is curious why the NIST authors did not reach out to forensic laboratories to inquire about more data. There are numerous online forums used by practicing DNA analysts to collaborate with one another that could have been utilized to solicit for the desired data. The NIST authors have previously conducted inter-laboratory studies (MIX 05, MIX 13, etc.) and found ways to contact numerous laboratories for those, but the authors clearly neglected to do so for this report. While many laboratories still may not have allowed the data to be published due to: 1) the data is not novel and 2) there are privacy concerns with the data, they may have been willing to provide the requested data for the authors' review. Perhaps an online share site could have been established for this purpose, and such a site could be used for future inter-laboratory collaborations.

Secondly, the authors make no mention that laboratories that participate in the CODIS system (and many non-CODIS labs) undergo routine audits that include a review of internal validation studies. These audits constitute a community-level peer review of reliability. NISTIR 8351 implies that this review is insufficient, that the community cannot evaluate itself adequately. That begs the question of who or what could perform that review to the authors' satisfaction. The NIST document proposes no solution. Regardless, it would be beneficial if the authors acknowledged that laboratories conduct their own validation studies, which are reviewed during an external audit, to provide a foundation on which to build the interpretation and statistical procedures. These exercises provide a measure of scientific reliability to mixture interpretation. This fact strongly counteracts the message in NISTIR 8351 that mixture interpretation is being conducted without a sufficiently reliable foundation.

Finally, despite a list of published references to the contrary, the tone of this document suggests that there is insufficient data available to assess the reliability of the interpretation of ALL types of mixture profiles, including high quality profiles of two contributors. Even the PCAST report didn't go that far. The published literature referenced in this document offers thorough and in-depth testing of a wide range of mixture types and propositions and provides a framework for the expected ambiguity/reliability across the mixture range. Based on this literature, DNA analysts are fully aware that more caution is needed when interpreting and reporting (including testimony) low level and complex mixtures than for robust and simpler mixtures. Therefore, it is disingenuous for the authors to lump the interpretation of all mixtures together as 'unreliable.' Unfortunately, this is how the legal community (a stakeholder in the outcome of DNA mixture interpretation) is using this report.

PC35

Comments and Feedback to NIST Draft Report on Mixture Interpretation (NISTIR 8351-DRAFT)

John Ballantyne [redacted]

Mon 8/23/2021 12:46 PM

To: ScientificFoundationReviews <ScientificFoundationReviews@nist.gov>

Comments on the Draft *DNA Mixture Interpretation: A NIST Scientific Foundation Review*

Jack Ballantyne, PhD
Professor, Department of Chemistry
Associate Director (Research)
National Center for Forensic Science
12354 Research Parkway, Suite 225
Orlando, FL 32826 USA

August 23, 2021

The draft Report ("Report") provides a comprehensive review of the history, current state and future developments of DNA mixture interpretation and the transfer and persistence of DNA and its relevance to providing context to a DNA profile. Together with the extensive compendium of relevant literature references, it should be a valuable resource for training and educating forensic scientists as well as others in the criminal justice system including, lawyers, judges and law enforcement officials. The comments below are intended as suggestions made to improve the content and tone of the Report.

Notwithstanding the above, the Report has a number of assertions and somewhat significant omissions that, as NIST appreciates, will inevitably cause considerable disruption and confusion within the criminal justice system. Victims and falsely accused suspects from current and already adjudicated cases will likely be the ones most impacted as a result. The biggest omission in the Report arises from the key Report takeaway that NIST is unable, due to a dearth of publicly available validation data, to assess the degree of reliability of DNA mixture interpretation practices. The Report cites scores of validation studies attesting to the reliability of PG-based DNA interpretation methods arising from test mixtures prepared under a variety of experimental conditions. It then, somewhat incongruously, concludes that none of the listed studies and data appeared to be sufficient to demonstrate an appropriate degree of scientific reliability. The Report subsequently calls for more reviewable study data not only for mixtures but also from further studies on the transfer and persistence of DNA. The said omission is that no concrete prescriptive solutions are proffered regarding how much more data is required, with what factor space conditions and how will that newfound data inform the current models and approaches to DNA mixture interpretation and context evaluation such that reliability could be established. Indeed, in the meantime many stakeholders will regard the content and tone of the Report as being a proxy call for a de facto moratorium on the use of complex DNA interpretation methods since, in the absence of demonstrated reliability, it is seemingly not yet shown to be fit for purpose.

More detailed comments are provided below.

Limitations of the NIST draft Report in exploring limitations of PG mixture interpretation systems

All bio-analytical and interpretation systems have limitations and that includes PG. There is a legitimate need for studies to determine what the limitations of complex mixture interpretation via PGS are. Thus, the expectation was that one of the main goals of NIST's scientific foundation review would be to explore such limitations and provide its input as to what the limitations are and what the

quantifiable risks are when working at these limitations. Limitations aren't a problem so long as they are recognized, and the interpretation system reflects the limitations. In PG analysis does the magnitude of generated LRs adequately reflect system limitations? If not always, then under what circumstances does this occur? This type of analysis and supporting data would represent *bona fide* limitations of the PG system that would need to be recognized, further studied, and used to inform analysts' interpretation conclusions. Some of this work has already been done and reported. However, disappointingly, the Report seemingly did not explicitly address these limitations, instead concentrating on asserting that the factor space is so great and that there was insufficient reviewable data to show reliability for any complex mixture interpretation scenario. In this regard a disservice has been done to the forensic DNA community who were hopeful of a comprehensive, thoughtful exploration of the limits of PG-based mixture interpretation. It is patently incorrect to imply that any potential limitations of the system are likely to extend to every complex mixture.

Chapter 4. Reliability of DNA Mixture Measurements and Interpretation

Community variability of mixture interpretation

(4.1.2., line 2768)

"In the context of DNA mixture interpretation using PGS ....., a DNA analyst assesses the probability of the findings if one proposition (H1) were true and also the probability of the findings if another proposition (H2) were true. *This assessment is typically accomplished with the help of specialized knowledge of the discipline, training and experience, and the assistance of statistical models and computer programs."*

This predicts that since LR assessment has a subjective component (knowledge and training of an individual) LRs obtained from the same sample are expected to vary, and this phenomenon is further recognized in the Report by the maxim that "there is no true LR". However, the report describes "KEY TAKEAWAY #4.6: Different analysts and different laboratories will have different approaches to interpreting the same DNA mixture. This introduces variability and uncertainty in DNA mixture interpretation. Improvements across the entire community are expected with an increased understanding of the causes of variability among laboratories and analysts." LR variability of course is entirely expected. The Report does not address the issue of <u>what level of this expected variation is acceptable</u> from the NIST authors' perspective such that the vaunted goal of 'reliability' ('providing consistently accurate results') of DNA interpretation via PG can be seen to be met by the laboratories.

Factor space coverage

(4.1.4., line 2848)

The Report states that the factor space (scenarios and variables) coverage of admixed DNA samples studied in validation studies can affect the reliability of DNA mixture measurement and interpretation. Also, the Report says that the factor space explored by a laboratory is only a small part of the entire factor space. Indeed, the factor space explored by a laboratory is <u>always</u> going to be a minuscule portion of the potential factor space even with the suggested bracketing, the latter being a standard approach already used by many laboratories during their validation studies. In effect the factor space is so vast that it cannot practically be thoroughly, empirically explored. Given the Report's conclusion that, given the data reviewed, <u>no point of the factor space</u> has been adequately validated to demonstrate reliability of DNA mixture interpretation, even from simple 2 person mixtures that have been reported in scores of publications, it would seem that the goal of appropriate factor space coverage as a condition of determining reliability for the wide range of samples encountered by forensic casework laboratories may be impossible to achieve by individual laboratories. It would be useful and more transparent for the Report to more explicitly state the true extent of the theoretical factor space, given the influencing factors listed in Table 4.1, even with bracketing (in terms of the numbers of factor space points..is it hundreds, thousands, millions etc?), such that individual lab users and validators of the PG systems realize the practical nigh impossibility of meeting NIST's expectations

for demonstrating reliability for the vast mixture casework variables likely to be experienced in a specific laboratory.

Not enough publicly available data to enable... assessment of the degree of reliability of DNA mixture interpretation
(4.3.3., line 3074 Publicly Available PGS Internal Validation Data)
"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".
This key takeaway is incredulous, and disconcerting to those in the community who have established the reliability of DNA mixture interpretation via PGS, for those Courts that have similarly ruled in its favor and for the victims and falsely accused suspects who have benefited from its application in real cases.
The reason is that it is incredulous is that there are scores of peer reviewed articles that attest to its reliability (i.e. mixture interpretation via PG) within the parameters and circumstances studied. Typically, bioanalytical scientists, including forensic scientists, test the diagnostic sensitivity and specificity of an analytical system and, therefore, its accuracy. Many of the peer reviewed papers listed do just that for PG based mixture interpretation. In addition, the PG-based analytical/interpretation system's repeatability and reproducibility is tested so that, in combination with the specificity/sensitivity analysis, establishes its degree of reliability. On the other hand, NIST hasn't cited any peer reviewed journal articles in the primary scientific literature that, via empirical testing, disproves or seriously undermines the scientific reliability of the PG-based methodology. It is incredulous that NIST's analysis couldn't describe the 'factor space', however limited, for which reliability seems to have been reasonably established. No details are provided regarding the results of the reliability analysis conducted. For example, such analysis should, inter alia, consider error rates including the variation of the true score versus the variation of the observed score, variables that are usually readily extracted from the peer reviewed literature. Thus, a suggestion is for the Report to be transparent about the results of its reliability analysis. If there are specific gaps revealed by this analysis (rather than just a broad call for more publicly available data, although more data is always good) then that would inform the community as to the direction future studies should take.
It should be noted that there is no legal or ethical requirement to make internal validation studies publicly available for independent review since that is not their purpose. Indeed, there are ethical and potential legal reasons why, in many instances, they cannot be made public.

"Not enough information.... to independently assess the degree of reliability of DNA mixture interpretation at any one point in the factor space"
(4.4. discussion line 3201)
At first blush, the above statement seems incredulous and patently wrong, since simple two person mixtures are deemed in the Report (line 605) to be foundationally valid, in agreement with the PCAST report. Perhaps NIST should consider changing the title of the Report to "Complex Mixture Determination: a Scientific Foundation Review". If NIST does find that some 2-person mixtures are foundationally valid, however, then the precise factor spaces for which they are deemed to be valid should be better clarified in the text. The reason for the latter is that it might be mis-conceived by readers (deliberately or otherwise) that, according to NIST, no mixtures have yet been shown to have foundational validity. The Report does state in passing that it only concentrates on complex mixtures (line 613), thus further highlighting the desirability of changing the title or, if not, to explicitly define the 2-person mixture factor space that is deemed by NIST to have demonstrated reliability.

Chapter 5: Context and Relevance Related to DNA Mixture Interpretation

The perceived special need for context and relevance for low level contributors to mixtures

The Report emphasizes the importance of determining the contextual circumstances surrounding a mixed DNA profile, and the degree to which a profile or components thereof are relevant to the case in question. The argument is made that, due to the high sensitivity of current STR analysis, and the newfound ability to routinely infer/determine the presence of low- level DNA contributors, then context and relevance becomes even more important (KEY TAKEAWAY #5.2). "Highly sensitive DNA methods increase the likelihood of detecting irrelevant DNA. When assessing evidence that involves very small quantities of DNA, it is especially important to consider relevance". However, these high sensitivity methods increase the likelihood of detecting relevant DNA evidence too. DNA evidence, independent of whether it involves small, medium, or high DNA quantities, always requires an assessment of relevance. Thus, whether one decides to assess relevance should not be dependent upon the amount of DNA detected but carried out as thoroughly as is practicable in every case.

Absence of an existing demonstrably reliable method to statistically combine the DNA genotype LR information with other contextual information

The Report states that although there is a growing body of knowledge about DNA transfer and persistence, significant knowledge gaps remain (KEY TAKEAWAY #5.6). Readers of the Report will rightly seek, since it's given such prominence, to see whether they could apply the results of transfer and persistence studies, and their contextual relevance, to actual cases and provide a quantitative or qualitative estimate of evidentiary relevance. "The likelihood ratio should not be used in isolation (line 839)" implies that this is a goal of the interpretation process. Although several bold efforts of incorporating relevance either qualitatively or quantitatively along with a sub-source DNA profile LR have been made is this area, there is no validated 'reliable' method to do so. Therefore, it would be useful for the Report to acknowledge and state that, as the literature (or lack thereof) indicates, and despite our increasing knowledge of the transfer and persistence characteristics of DNA at the current time and desire to use it to inform the crime scene analysis, there is no demonstrably reliable method (as defined in the Report) currently available to be able to statistically combine the sub-source DNA genotype LR information with DNA transfer, persistence and recovery) contextual information.

The importance of context and relevance to interpretation of DNA profiles

The Report clearly emphasizes the importance of context and relevance to DNA interpretation and should be embraced by the forensic DNA community. The importance of such knowledge for casework forensic scientists is paramount to so that they can best serve the interests of justice. The need to use context to inform the analysis/interpretation process provides a strong counterpoint to those who would argue that forensic scientists be provided with no such relevant knowledge/meta-data due to so-called 'contextual bias'. In science, contextual bias is ameliorated by an acculturation process which inculcates the scientific method into trainee scientists of all disciplines, including forensic science. Scientists are trained, indeed indoctrinated, into objectively examining and interpreting data and its relevance to a question at issue. This education and training in the scientific method, that de facto takes into account confounding effects such as contextual bias, is far more important and relevant than the use of superfluous ad hoc approaches such as 'sequential unmasking' (line 4588). The latter approach is really nothing more than simply applying the scientific method to the case, which all competent forensic scientists should be naturally doing anyway. Notwithstanding the above, forensic scientists, like all scientists, need to remain vigilant in their application of the scientific method to their work.

Appendix 1: History of DNA Mixture Interpretation

Criticism of labs' validation studies as being task-driven rather than performance-based

A1.7.3 line 6342

"An observation made in conducting this scientific foundation review is that, historically, FBI QAS validation requirements and SWGDAM validation guidelines have become *task-driven* rather than *performance-based*. In other words, the requirements and guidelines may be treated by some as a checklist of studies that need to be completed to satisfy requirements rather than a demonstrated performance of the accuracy or reliability of results obtained using the method".

This appears to be a gratuitous comment made in anecdotal mode without providing a literature (or any other) source. Given the forensic genetics community's obvious commitment to getting things right in this arena, Things have changed substantially over the years with regards to validation requirements and the recognized need to use internal validation results to inform the lab's interpretation guidelines. The Report should cite sources for the 'observation' statement or be removed. Otherwise, it will be regarded as a derogatory prejudicial anecdote which detracts from the purpose of an independent scientific foundational review document.

Regards,

Jack Ballantyne, PhD
Professor, Department of Chemistry
Associate Director (Research)
National Center for Forensic Science
12354 Research Parkway, Suite 225
Orlando, FL 32826
USA

[redacted]

PC36

Comment on NISTIR 8351-DRAFT

Elizabeth Vasquez [redacted]

Mon 8/23/2021 1:57 PM

To: **ScientificFoundationReviews <ScientificFoundationReviews@nist.gov>**
Cc: Clinton Hughes [redacted]

Good afternoon,
Please find attached our Comment on NISTIR 8351-DRAFT—*DNA Mixture Interpretation: A NIST Scientific Foundation Review*.

Thank you,
Elizabeth

Elizabeth Daniel Vasquez
Director, Science & Surveillance Project
Brooklyn Defender Services
177 Livingston Street, 5th Floor
Brooklyn, New York 11201
[redacted]

PC36a

August 23, 2021

Via Electronic Mail
National Institute of Standards and Technology
Attn: Special Programs Office – Scientific Foundation Review
U.S. Department of Commerce
100 Bureau Drive Stop 4701
Gaithersburg, Maryland 20899-4701
scientificfoundationreviews@nist.gov

Re: Request for Comment on NISTIR 8351-DRAFT—*DNA Mixture Interpretation: A NIST Scientific Foundation Review*.

Dear NIST Scientific Foundation Review Team:

Brooklyn Defender Services ("BDS") submits these comments in response to the *DNA Mixture Interpretation: A NIST Scientific Foundation Review* draft, NISTIR 8351-DRAFT (hereinafter, "Draft Report"), published on June 9, 2021.

BDS is a full-service public defender organization in Brooklyn, New York, that provides multi-disciplinary and client-centered criminal defense, family defense, immigration, and civil legal services, along with social work and advocacy support. BDS represents low-income people in nearly 30,000 criminal, family, civil, and immigration proceedings each year. Over the last decade, we have witnessed firsthand the dramatic expansion of forensic DNA analysis to more and more cases. NIST is correct that these methods are now used as a matter of routine in everyday casework. In response to this development, BDS established a dedicated Science & Surveillance Project and Forensic Science Practice. This team focuses on remaining abreast of and responding to developments and issues of data, science, and technology in the criminal legal system.

We applaud NIST's critical work "to answer the question: 'What empirical data exist that speak to the reliability of'" DNA mixture interpretation methods? We are particularly heartened by (1) the Draft Report's emphasis on the lack of available data surrounding these methods' reliability as summarized in Key Takeaway #4.3 and (2) the Draft Report's call for a separate scientific foundation review for likelihood ratios as summarized in Key Takeaway #4.8. Before this Draft Report is published, however, we would like to see NIST directly address the kinship problem, a key limitation of current DNA mixture interpretation methods that has gone unacknowledged in practice.

I. While a lack of publicly available data overall undermines review of DNA mixture interpretation methods' scientific foundation, the data that *is* available undermines those methods' reliability.

Despite some comment criticisms to the contrary, we commend NIST's articulation of a very real problem in assessing forensic DNA analysis *practices*: the lack of publicly available data.

Some commenters appear to have misunderstood NIST's critical point: validation is not addressed to a monolithic question of reliability and, thus, cannot be established by completing a checklist. Instead, validation testing must attempt to cover as much of the factor space as possible and, at the very least, must cover the factor space occupied by the real-world scenario to which the method is applied in the individual case. As it currently stands, it is impossible (externally and independently) to analyze how much of the factor space validation testing has covered. "Science and secrecy do not sit comfortably together."[1]

a. Key Takeaway #4.3 diagnoses a very real problem in forensic DNA analysis practices.

Most forensic DNA laboratories do not publicly share critical information about testing performance, safety, and accuracy. For example, of the approximately 69 forensic labs in the United States using STRMix, fewer than 10 have made their internal validation *summaries* public. In the face of this environment of secrecy, on July 12, 2021, BDS itself published 8 summaries that we had institutionally collected over the course of the last two years. These summaries were posted *after* the Draft Report's publication. None of the US labs using STRMix have publicly disclosed their full validation studies (even with redactions of sensitive material), absent an affirmative freedom of information request. This is critically detrimental to transparency.

Outside of internal validation study publication, less than a quarter of those laboratories publicly disclose their STRMix protocols. While the protocols "should be supported by validation studies conducted with samples of known origin similar to the types of samples routinely accepted and tested by the laboratory," ANSI/ASB Standard 020 at 3.2 "Internal Validation," *none* of these laboratories provide any public explanation of how their internal validation studies support their protocols.

Secrecy extends to analyst training as well. None of the probabilistic genotyping labs publish their full probabilistic genotyping staff training materials, including competency testing materials. The developers' training materials are also not publicly available. Similarly, proficiency testing is not made available. The public does not have access to the proficiency testing content or information about analyst performance on those tests.

[1] Sheila Jasanoff, *Transparency in Public Science: Purposes, Reasons, Limits*, 69 LAW AND CONTEMPORARY PROBLEMS 21, 21 (2006).

We agree with NIST that "transparency and openness are the hallmarks of good science."[2]

b. But NIST's Draft Report should also address the indications of method limitations that ARE *apparent from the currently available literature.*

While the Draft Report's diagnosis of the problem of secrecy highlights a fundamental problem at the foundations of forensic DNA mixture analysis, the Draft Report should also capitalize on the opportunity to comment on red flags in the data that *is* available. As we have highlighted on our micro-site The Kinship Problem | In Defense Of, probabilistic genotyping systems "can get a critical portion of cases . . . badly wrong." Specifically, each of the labs that has studied the question of high-allele sharing by testing mixtures made up of first-order relatives has found alarming false positive rates.

The impact of this problem is profound. Forensic DNA analysis is the second most requested type of forensic analysis nationwide.[3] As the Draft Report notes, while DNA testing is still conducted in the most serious of cases, it can be and is also deployed against the most minor. The area that has seen the biggest growth in testing in many major cities is possessory gun offenses. This leads to an increase, for example, in items collected from search warrants executed in family homes being submitted for testing.

There is a real threat of false inclusions where related individuals are involved. As discussed below, the small subset of labs that have appropriately tested probabilistic genotyping systems on this issue have discovered that false positive inclusions of known non-contributor relateds *will* happen. The typical diagnostics advertised by the software developers do not flag this system failure. As so-called "touch" DNA samples rise, the risk of encountering a related-individual mixture increases exponentially. And, with that elevating risk, the system's vulnerability to falsely inculpatory scientific evidence also rises.

Despite the 2009 National Academy of Science Report's recommendation that Congress establish an oversight agency for forensic science, no such body has emerged. While the FDA oversees medical applications of DNA testing and other life-critical systems like transportation or infrastructure fall under clear governmental oversight programs, there is no federal or state agency tasked with "oversight and enforcement of operating standards, certification, accreditation, and ethics" in the forensic science space.

This absence of regulation leaves the industry free to peddle techniques, kits, and products that have not been proven safe, reliable or effective to the criminal legal system market. And, as with the pre-FDA world of toxic and mislabeled drugs and foods, the only warning system presently in place to catch misbehaving forensic methods is one of trial-and-error on real human lives.

---

[2] NIST, *NIST Scientific Foundation Reviews*, NISTIR 8225 (Dec. 2020), at 1.2, https://nvlpubs.nist.gov/nistpubs/ir/2020/NIST.IR.8225.pdf.

[3] Government Accountability Office, Report to Congressional Requestors: DNA Evidence (March 2019), https://www.gao.gov/assets/gao-19-216.pdf.

The long-standing disparities endemic to the modern criminal legal system ensure that *any* impact of that system will be disproportionately felt by communities of color and the poor. Compounding this reality, statistically, multi-generational family living arrangements are more common amongst communities of color, as well as the poor. The specific threat of falsely inculpatory scientific evidence arising from the new probabilistic genotyping softwares cannot escape these obvious implications for both racial and social justice. The threat of false inclusions here *will* disproportionately impact people of color and the poor.

i. Defining the Kinship Problem with data.

In Table 3.2, the Draft Report identifies eight laboratories with publicly available internal validation summaries. Since the publication of the draft report, as discussed above, we have made seven additional laboratories' summaries publicly available here. Amongst those sixteen labs, ten included in their internal validation study design some evaluation of system performance on mixtures made up of first-order relatives.[4]

The limited studies by these American laboratories indicate that false positive LRs can go into the *trillions* and *quadrillions* for non-contributors compared against mixtures of multiple first-order relatives (*See, e.g.*, Los Angeles County Sheriff's Department's Scientific Services Bureau Biology Section, Validation of STRmixTM v. 2.5.11 using the Powerplex Fusion 6C Kit at 58; Las Vegas Metropolitan Police Department, Internal Validation of STRmix™ v2.6 [QIAGEN Investigator® 24plex QS with 3500xl] at 7).

There have also been non-contributor LRs in the millions, billions, trillions and quadrillions when non-contributor relatives are compared against a variety of familial mixtures. (*See, e.g.*, Jefferson County Regional Crime Laboratory, Internal Validation of STRmix™ v. 2.6 for the Analysis of GlobalFiler™ Profiles at 23-24; Sacramento County District Attorney's Crime Laboratory, Internal Validation of STRmix™ v. 2.4 at 9-10 and 29; Colorado Bureau of Investigation, Internal Validation of STRmix™ v. 2.5 for the CBI Forensic Laboratories (GlobalFiler™, 3500xL CE) at 70; Wisconsin State Crime Laboratory, Internal Validation Summary for STRmix™ Probabilistic Genotyping Software at 14, 15, and 31; 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 at 34).

The Sacramento County validation summary might be specially significant; it indicates that in low template three-person mixtures, the presence of a single sibling can

[4] While DNA Labs International's internal validation summaries or studies are not publicly available, it *is* clear from multiple public presentations given by their analysts that they also have studied the impact of this kind of high allele sharing. It is also clear that their studies revealed the same problems described here. *See, e.g.,* DNA Labs International – From Training to Trial: A Reflection on Nearly Three Years of Probabilistic Genotyping (Jan. 31, 2018), https://indefenseof.us/uploads/DLI_-from-training-to-trial-a-yr-reflection-of-probabilistic-genotyping.pdf; DNA Labs International – Probabilistic Genotyping in the Courtroom – Admissibility, Families, Secondary Transfer, and Competing Statistics (2020), https://indefenseof.us/uploads/DLI_Probabilistic-Genotyping-in-the-Courtroom-Admissibility-Families-Secondary-Transfer-and-Competing-Statistics.pdf.

lead to non-contributor LRs as high as 59 trillion (Sacramento County District Attorney's Crime Laboratory, Internal Validation of STRmix™ v. 2.4 at 29).

Up until now, for inexplicable reasons, this problem has not been publicly aired. None of the studies in Tables 4.4 or 4.5 explicitly predict the high false positive LRs that are reported in the validation summaries from the Kinship Problem website. The labs discussed here have not taken steps to warn the public or the community.

This gaping hole of information creates a sort of cognitive dissonance in the use of DNA in our courtrooms. American juries are expected to understand and properly apply a non-intuitive statistic like the likelihood ratio to complex DNA samples. But those same juries are not being exposed to data known to the labs and now publicly available indicating that PG software can falsely implicate a non-donor, with an LR higher in value than the national debt. The NIST Draft Report should take a step toward remedying the impact of this silence by including these validation summaries in its consideration.

ii. Addressing the Kinship Problem.

The Draft Report should address these issues in five key places:

(1) ***Table 3.2***. We strongly recommend that you update Table 3.2, "Publicly available internal validation data from forensic laboratories located in Google searches," to include the PG Validation Summaries linked from The Kinship Problem | In Defense Of.

Those validations include major labs that performed limited study of the achilles heel of probabilistic genotyping analysis: the extreme allele sharing of related individuals in complex mixtures.

(2) In ***Key Takeaway #4.3***, your draft report has already articulated the need for the availability and transparency of validation data in assessing reliability. The inclusion of these kinship problem summaries will support the urgent need to implement NIST's recommendations. The fact that these summaries were not made public by the labs themselves, and that the kinship problem has not been publicly aired and addressed by the developers, as well as practitioners, is alarming to say the least.

(3) Their inclusion will also support your discussion on ***4.4.3.1. Degree of Allele Sharing***, and the dangers of not including that element in validation studies. Additionally, these studies and the alarming conclusions to be drawn from them should be included in the Draft Report, as well.

(4) ***Principle 6***: We recommend that you correct the language here to state the obvious – e.g. strike "then performing calculations assuming individuals are related may be helpful to decisionmakers," and instead note the data-driven conclusion that: "this seems to be the most dangerous area of DNA analysis, where both human analysis and PG systems can get it wrong, and where non-contributors can be falsely implicated."

(5) ***Key Takeaway #2.6*:** There may be no single, correct LR. But when it comes to related individuals, there are going to be plenty of incorrect LRs that mislead users in the criminal legal system. Given that dilemma, it may be useful to ask, how often do labs receive samples from law enforcement searches of family homes with multiple-related individuals? How often do these samples come from family cars? How often do allegations include multiple family members? If the LR only scrutinizes the strength of a single hypothesis uncalibrated to all the others, what utility does it have to factfinders, particularly when that tunnel-visioned analysis can so badly mislead?

II. Conclusion

In addition to our requests in this letter, we join in the comments of our defender colleagues calling for the NIST Review to openly acknowledge the racial justice implications of their findings. We agree that NIST should do more. It is time for NIST to issue a call for a moratorium and to initiate a racial impact assessment for DNA mixture analysis. As illustrated by the kinship problem itself, the stakes are too high to do less.

Sincerely,

/s/ *Elizabeth Daniel Vasquez*
Elizabeth Daniel Vasquez
Director, Science & Surveillance Project

/s/ *Clinton Hughes*
Clinton Hughes
Forensic DNA Attorney, Criminal Practice
Brooklyn Defender Services
177 Livingston Street, 7th Floor
Brooklyn, New York 11201

PC37

Virginia DFS comments

Jenkins, Bradford [redacted]

Mon 8/23/2021 2:46 PM

To: ScientificFoundationReviews <ScientificFoundationReviews@nist.gov>

Thank you for the opportunity to comment.

Brad

Brad Jenkins
Program Manager
Forensic Biology Section
Virginia Department of Forensic Science
[redacted]

PC37a

COMMONWEALTH of VIRGINIA

DEPARTMENT OF FORENSIC SCIENCE

OFFICE OF THE DIRECTOR
A Nationally Accredited Laboratory

August 23, 2021

700 NORTH 5TH ST.
RICHMOND, VIRGINIA 23219
(804) 786-2281 FAX (804) 786-6857

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

Thank you for the opportunity to comment on the *DNA Mixture Interpretation: A NIST Scientific Foundation Review*. The Virginia Department of Forensic Science (VA DFS) acknowledges and appreciates all of the time and effort associated with such an endeavour, but respectfully disagrees with some key takeaways listed in the document. We feel that the publication, as a whole, does not fully recognize the full body of work performed by forensic laboratories across the country undertaken to improve DNA mixture analysis and interpretation processes. VA DFS suggests that, because of the limited scope of the review, some conclusions offered as key takeaways may be incomplete.

Additionally, the review publication does not adequately address the ability of probabilistic genotyping software (PGS) systems to eliminate non-contributors to complex DNA mixture profiles. There are numerous published and unpublished studies that demonstrate that, even with highly complex data covering the "factor space" defined in the document (e.g., greater than 3 and 4 contributors, high degree of allele sharing, multiple mixture ratios, etc.), these systems reliably and reproducibly eliminate non-contributors. This is one of the great features of the PGS systems and is most critical in protecting the wrongly accused.

KEY TAKEAWAY #2.6: *Likelihood ratios are not measurements. There is no single, correct likelihood ratio (LR). Different individuals and/or probabilistic genotyping software (PGS) systems often assign different LR values when presented with the same evidence because they base their judgment on different kits, protocols, models, assumptions, or computational algorithms. Empirical data for assessing the fitness for purpose of an analyst's LR are therefore warranted.*

The VA DFS experience with different PGS systems (i.e.,TrueAllele® Casework (TA) and STRmix™) is that:

- Regardless of the level of complexity of the test data sets, the two systems converge on equivalent results more frequently than they diverge. In particular, when using challenging data sets, including those covering the "factor space" defined in the document, the results in the form of LR values are shown to be quite similar.
- When substantial differences are observed between the LR outputs of the two PGS systems, they are usually expected, predictable, and can be easily explained by some of the basic modeling differences.

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.*

As cited in the document, VA DFS was represented on the DNA Resource Group and provided feedback regarding PGS system validation and use in casework in a state forensic laboratory. Requests to review our underlying PGS data were not made.

- VA DFS provides copies of validation summaries, allows for the inspection of underlying raw data by attorneys or experts when requested through proper legal channels, and by other individuals on a case-by-case basis with appropriate non-disclosure agreements. Other crime laboratory systems have similar practices or may offer review of data through the execution of non-disclosure agreements. Through this process, the authors may access additional underlying data for review.

VA DFS contends that the suggestion to make the underlying data widely and publicly available is both impractical and unreasonable.

- It is impractical due to the sheer volume of data generated by such studies.
- It is unreasonable due to the DNA privacy concerns of those whose DNA profiles are used in the studies.

VA DFS agrees that interlaboratory studies are an important tool for assessing laboratory procedures and testing. However, VA DFS contends that such studies should use samples or data that closely mimic actual casework profiles and scenarios, of varying levels of difficulty and complexity, to provide the most information and encourage participation. Samples that are engineered to provide unrealistic challenges may erode confidence in the results of such studies and discourage future participation in this type of assessment.

Chapter 5: *Context and Relevance Related to DNA Mixture Interpretation*

VA DFS agrees that the relevance of a DNA profile developed from evidence is most assuredly a matter of utmost importance in the context of a criminal proceeding. However, it is the opinion of VA DFS that this topic warrants its own discussion and a separate publication from one designed to be a scientific foundation review of mixture interpretation and, therefore, should be removed.

Remarks in conclusion

The 2010 Scientific Working Group on DNA analysis methods DNA mixture committee made many recommendations on DNA mixture analysis including the recommendation to incorporate probabilistic genotyping as a mixture analysis tool for casework. This represented a turning point in forensic DNA mixture analysis and interpretation in the United States and furthered the adoption and use of PGS

systems throughout the country. Eleven years has now passed and the repeated reliability of validated and widely employed PGS systems used by trained forensic casework examiners has resulted in tremendous value of these systems to the judicial system, not just the prosecution of cases but also in the exoneration of innocent individuals.

During the past decade, there were no legal, validation, accreditation or implementation requirements to make the underlying validation data publicly available. Retroactively requesting such data be made publicly available now is problematic as noted previously. Nonetheless avenues do exist for laboratories to provide such data for this review. As drafted, the review casts doubt on the reliability of PGS systems. VA DFS strongly encourages NIST to utilize appropriate mechanisms to gain access to additional validation data so that a comprehensive review can be performed prior to finalizing this publication.

For additional information or questions please contact Brad Jenkins, Virginia DFS Biology Program Manager at brad.jenkins@dfs.virginia.gov.

PC38

NISTIR-8351 DRAFT- DNA Mixture Interpretation: A Scientific Foundation Review

Lu, Lauren (MSP) [redacted]

Mon 8/23/2021 3:14 PM

To: ScientificFoundationReviews <ScientificFoundationReviews@nist.gov>
Cc: Bowen, John (MSP) [redacted]; Larrison, Ryan M. (MSP) [redacted]; Nye, Jeffrey V. (MSP) [redacted]

I am respectfully submitting the attached comments regarding the NISTIR-8351 DRAFT entitled *DNA Mixture Interpretation: A Scientific Foundation Review*.

Regards,

Lauren Lu

Lauren Lu
Biology Technical Leader
Forensic Science Division
Michigan State Police
Forensic Science Division
7320 N. Canal Rd.
Lansing, MI 48913
[redacted]

"A **PROUD** tradition of **SERVICE** through **EXCELLENCE**, **INTEGRITY**, and **COURTESY**"

Your opinion is very important to us. Please take a few moments to complete a brief survey at the following link:
MSP Forensic Science Division Customer Service Survey

PC38a

GRETCHEN WHITMER
GOVERNOR

STATE OF MICHIGAN
DEPARTMENT OF STATE POLICE
LANSING

COL. JOSEPH M. GASPER
DIRECTOR

August 23, 2021

National Institute of Standards and Technology
Attn: Special Programs Office – Scientific Foundation Review
100 Bureau Drive Stop 4701
Gaithersburg, MD 20899-4701

Dear Dr. John Butler and authors of the draft *DNA Mixture Interpretation: A Scientific Foundation Review*:

The Michigan State Police is respectfully offering the following comments and suggestions in response to the draft NISTIR-8351 entitled *DNA Mixture Interpretation: A NIST Scientific Foundation Review*.

Reference Table 4.1 (Line 2862) Factor space that influences DNA mixture measurements and interpretations with probabilistic genotyping software (PGS) systems.

- Michigan State Police internal laboratory validations have thoroughly covered all of the factors influencing "DNA mixture measurement and interpretation". The following table addresses which internal validations have studied these factors, as well as provides the relevant decisions which resulted from those validation studies.

Coverage of Factors Influencing DNA Mixture Measurement and Interpretation Across Laboratory Internal Validations

Portion of Factor Space	Influencing Factors	Internal Validation Study Covering the Factor/Laboratory Decisions
Measurement of STR Alleles and Genotypes	Peak position for short tandem repeat (STR) alleles	Validation of the PowerPlex® Fusion STR Chemistry Kit Validation of the Applied Biosystems™ 3500/3500xl Genetic Analyzers and GeneMapper™ Id-x v. 1.4
	Peak morphology or resolution for STR alleles	Validation of the PowerPlex® Fusion STR Chemistry Kit Validation of the Applied Biosystems™ 3500/3500xl Genetic Analyzers and GeneMapper™ Id-x v. 1.4
	Peak height for STR alleles	Validation of the PowerPlex® Fusion STR Chemistry Kit Validation of the Applied Biosystems™ 3500/3500xl Genetic Analyzers and GeneMapper™ Id-x v. 1.4

	Relative peak heights for STR alleles pairs	Validation of the PowerPlex® Fusion STR Chemistry Kit Validation of the Applied Biosystems™ 3500/3500xl Genetic Analyzers and GeneMapper™ Id-x v. 1.4
	Presence of stutter products and their relative heights compared to associated STR alleles	Validation of the PowerPlex® Fusion STR Chemistry Kit Validation of the Applied Biosystems™ 3500/3500xl Genetic Analyzers and GeneMapper™ Id-x v. 1.4
Sample Complexity	Number of contributors	Validation: STRmix™ and PowerPlex® Fusion Validation: Use of STRmix™ Software for Profile Interpretation of Challenging Samples
	Degree of allele sharing among contributors	Validation: Use of STRmix™ Software for Profile Interpretation of Challenging Samples
	Presence of stutter products	Validation: STRmix™ and PowerPlex® Fusion
	Total DNA template and contributor template amounts	Validation: STRmix™ and PowerPlex® Fusion Validation: Use of STRmix™ Software for Profile Interpretation of Challenging Samples
	Mixture ratio of DNA from contributors	Validation: STRmix™ and PowerPlex® Fusion Validation: Use of STRmix™ Software for Profile Interpretation of Challenging Samples
	Sample quality including degree of degradation	Validation: Use of STRmix™ Software for Profile Interpretation of Challenging Samples
	Presence of stutter products and potential minor contributors	Validation: Use of STRmix™ Software for Profile Interpretation of Challenging Samples
Laboratory Specific Decisions	STR typing kit(s) used	PowerPlex® Fusion 5
	Capillary electrophoresis (CE) instrument used	Applied Biosystems™ 3500 and 3500xl
	Sample processing methods	Extraction: Organic (Stain and Differential), QIAsymphony®/ DNA

		Investigator Kit
		Quantitation: Applied Biosystems 7500/Plexor HY
		Amp Target Range: 0.5 – 1.0ng
	Number of PCR cycles	30
	Replicate testing	Replicates not used in casework
	Analytical threshold	250 RFU
	Population allele frequencies	NIST 2017
	Co-ancestry coefficient	Θ = 0.01
	Analyst training and experience (with lab protocols)	Completion of ESR-provided workshop/Completion of extensive training program including literature, training profiles/exercises, a written exam, and a practical competency
PGS Model Decisions	PGS model used	Continuous
	Laboratory-specific parameters for use in the PGS model	Validation: STRmix™ and PowerPlex® Fusion (Model Maker)
	Non-contributor data construction and testing	Validation: STRmix™ and PowerPlex® Fusion Validation: Use of STRmix™ Software for Profile Interpretation of Challenging Samples
Software Implementing the PGS Model	Choice of numerical methods for computing likelihood ratios	MCMC, Metropolis-Hastings
	Choice of the number of iterations or numerical integration parameters	100,000 Burn-in 500,000 Post burn-in
	Choice of diagnostic checks on the results	Primary: Estimated mixture proportions, per-locus likelihood ratios (LRs), genotype weights Secondary: Gelman-Rubin convergence, Avg (log) likelihood, allele variance, stutter variance Other potential diagnostic indicators: Effective sample size, Highest Posterior Density (HPD), relatedness LRs
Case-Specific Decisions	Propositions and assumptions	Will vary from profile to profile, even within a case, but are subject to review

		by defense experts and cross examination

Reference Key Takeaway #4.2: To enable effective use of any information, responsibilities exist with both providers and users of that information. While a provider explains the relevance and significance of the information and data, only the user can assess the degree of reliability, validity, and whether that information is fit-for-purpose.

- In terms of PGS, the provider is the developer of the software and the forensic laboratory is the user. The developer of STRmix, ESR (the provider), is very active in the provision of training, technical support, workshops and lectures at conferences, and publishing studies. Through all of these mediums, the relevance and significance of their software and its functions have been explained. The forensic laboratory, as a user, internally validates the software. More than 60 laboratories in the United States have already completed internal validation of STRmix™, and found it to be reliable, valid, and fit-for-purpose. NIST is not the user of this software and thus, by its own definition, is incapable of assessing the reliability and validity of PGS and whether it is fit-for-purpose.
- The reliability of probabilistic genotyping has been examined in-depth in legal proceedings across the U.S. for the past five years. Federal courts of appeals have written opinions in favor of admitting probabilistic genotyping evidence in trials (US v Gissantaner, Western District-Michigan). During these proceedings, forensic DNA laboratories have provided countless documents supporting the validity, reliability, and evidence of fit-for-purpose. These documents include laboratory reports and case file documentation, validation studies, software source code, peer-reviewed literature, training programs, analyst's written exams, interpretation protocols, and defense expert reports and analysis. As a whole, the DNA community has demonstrated the reliability of PGS repeatedly in legal proceedings throughout the United States.

Reference Table 4.2 (Line 2997) Summary of factor space coverage and findings for measurement experiments and DNA mixture studies from three developmental validation studies of commonly used commercial STR typing kits.

- Regarding the developmental validation studies from STR kits, much of the factor space covering sensitivity, sizing precision, reproducibility, concordance, heterozygote balance, and stutter has been covered during forensic DNA laboratory internal validations. The beginning stages of validating probabilistic genotyping software involve in-depth evaluations of all or most of these factors in order to inform the system with parameters. These parameters drive the expectations of the modelling and are used to compare to the observed data in validation profiles of known composition, and ultimately evidentiary profiles.

Reference Table 4.5 (Line 3069) Factor space coverage of information in internal validation studies listed in Table 3.2.

- The Michigan State Police has effectively covered much of the "factor space" recommended by this report, but not all of that work was publicly available. The report should be corrected to reflect the actual factor space covered. The profiles outlined here are all lab-created samples-these charts do not contain the additional testing conducted on adjudicated samples.

Coverage of Factor Space from Validation: STRmix™ and PowerPlex® Fusion

C Range	# Samples	Total DNA Quantity Range (pg)	Mixture Ratio Range
1	6	500	N/A
		600	N/A
		150	N/A

		75	N/A
		50	N/A
		25	N/A
2	18	500:500	1:1
		909:91	10:1
		882:118	7.5:1
		833:167	5:1
		714:286	2.5:1
		500:500	1:1
		714:286	2.5:1
		2,143:857	2.5:1
		909:91	10:1
		882:118	7.5:1
		833:167	5:1
		714:286	2.5:1
		500:500	1:1
		909:91	10:1
		882:118	7.5:1
		833:167	5:1
		714:286	2.5:1
		500:500	1:1
3	22	625:312.5:62.5	10:5:1
		833:83:83	10:1:1
		769:154:77	10:2:1
		625:312.5:62.5	10:5:1
		476:476:48	10:10:1
		454.5:454.5:91	10:10:2 [5:5:1]
		400:400:200	10:10:5 [2:2:1]
		333:333:333	10:10:10 [1:1:1]
		500:334:167	3:2:1

		351:234:117	3:2:1
		234:156:78	3:2:1
		174:116:58	3:2:1
		78:52:26	3:2:1
		833:83:83	10:1:1
		740:185:74	10:2.5:1
		625:312.5:62.5	10:5:1
		540:405:54	10:7.5:1
		476:476:48	10:10:1
		444:444:111	10:10:2.5 [4:4:1]
		400:400:200	10:10:5 [2:2:1]
		364:364:272	10:10:7.5 [4:4:3]
		333:333:333	10:10:10 [1:1:1]
4	19	588:294:59:59	10:5:1:1
		769:77:77:77	10:1:1:1
		588:294:59:59	10:5:1:1
		385:192:38	10:10:5:1
		468:351:234:117	4:3:2:1
		312:234:156:78	4:3:2:1
		232:174:116:58	4:3:2:1
		104:78:52:26	4:3:2:1
		769:77:77:77	10:1:1:1
		769:77:77:77	10:1:1:1
		714:143:71:71	10:2:1:1
		588:294:59:59	10:5:1:1
		455:455:45:45	10:10:1:1
		435:435:87:43	10:10:2:1
		384:384:192:38	10:10:5:1
		323:323:323:32	10:10:10:1
		312.5:312.5:312.5:62.5	10:10:10:2 [5:5:5:1]

		286:286:286:143	10:10:10:5 [2:2:2:1]
		250:250:250:250	10:10:10:10 [1:1:1:1]

Coverage of Factor Space from Validation: Use of STRmix™ Software for Profile Interpretation of Challenging Samples

NoC Range	# Samples	Total DNA Quantity Range (pg)	Mixture Ratio Range
1	31 (Degraded)	500	N/A
2	4 (Biological Relatives)	250:250	1:1
	4 (Biological Relatives)	400:100	4:1
	4 (Biological Relatives)	455:45	10:1
3	4 (Degraded)	167:167:167	1:1:1
	6 (Degraded)	357:71:71	5:1:1
	2 (Low-level, Heterozygosity)	385:96:19	20:5:1
	2 (Low-level, Heterozygosity)	454:23:23	20:1:1
	3 (Profile Rarity)	385:96:19	20:5:1
	3 (Profile Rarity)	454:23:23	20:1:1
	4 (Biological Relatives)	167:167:167	1:1:1
	4 (Biological Relatives)	333:83:83	4:1:1

	4 (Biological Relatives)	313:156:31	10:5:1
4	2 (Low-level, Heterozygosity)	370:93:19:19	20:5:1:1
	2 (Low-level, Heterozygosity)	434:22:22:22	20:1:1:1
	3 (Profile Rarity)	370:93:19:19	20:5:1:1
	3 (Profile Rarity)	434:22:22:22	20:1:1:1

Reference Key Takeaway #4.3 (Line 3074) 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.

- A vast wealth of probabilistic genotyping internal validation data exists, even if not publicly available. Further, forensic laboratory validation data is not required by any standard or accrediting body to be made publicly available. This is a new construct- a new requirement over and above all of the mechanisms by which DNA validations are already evaluated (audits, court discovery, Freedom of Information requests) and should be clearly acknowledged as such.
- The data supporting these validations consist of DNA profiles, and are thus, confidential. Making the DNA profiles associated with these validation studies publicly available is not feasible from the standpoint of privacy.

Reference Key Takeaway #4.7 (Line 3487): The degree of reliability of a PGS system when interpreting a DNA mixture can be judged based on validation studies using known samples that are similar in complexity to the sample in the case. To enable users of results to assess the degree of reliability in the case of interest, it would be helpful to include these validation performance results in the case file and report.

- This Key Takeaway considers the forensic laboratory in the place of the "provider" and the reader of the report as the "user". While that is indeed the dynamic, the "user" in this instance cannot be expected to understand the concepts to the same level as a scientist with the education, training, and experience to support their work in a forensic laboratory. This Key Takeaway equates the reader of a report with the scientist herself and undermines the provision of expert testimony in our legal system. It should be the responsibility of the scientist to express these concepts in terms understandable to the recipient of the report, but the "user" will only rarely have an educational or experiential background to understand scientific information to this level of detail.
- An accredited forensic laboratory is expected to provide the results of testing "accurately, clearly, unambiguously and objectively" (ISO/IEC 17025: 2017, Standard 7.8.1.2). Forensic DNA testing reports must be accurate but must also attempt to communicate effectively with the recipient, so as to generate a working understanding of the content. Typically, the recipient of a forensic DNA testing report is an investigating agency, prosecuting attorney, defense attorney, or defense expert. Only the defense expert would find benefit in the inclusion of validation performance

results in each DNA report. Since many DNA reports are lengthy, and can by nature become complicated, for most readers the addition of validation performance results serves only to lengthen and further complicate reporting. In addition, validation performance results are readily available on discovery to all of these parties.

Reference Key Takeaway #4.8 (Line 3594): We encourage a separate scientific foundation review on the topic of likelihood ratios in forensic science and how LRs are calculated, understood, and communicated.

- This report does not address the ways in which forensic laboratories have designed their procedures based on the validation studies conducted. The validation itself is only half of the picture, with the DNA interpretation and probabilistic genotyping procedures completely unaddressed. For instance, the MSP tested up to four-contributor mixtures during validation, and thus subsequently only interprets up to four-contributor mixtures. Replicate amplifications were not tested during validation and are consequently not utilized in DNA casework. Further, likelihood ratios are not applied to the concepts of the quantity of DNA present, how the DNA was deposited on the item, or when the DNA was deposited. No activity-level propositions are reported or provided during testimony.

Reference Box 4.1 Desired Information for Reliability Assessments of LR Values in PGS Systems (Line 3659)

- Few mechanisms are proposed for forensic DNA laboratories to meet the new constructs proposed in this report. The publication of forensic DNA testing laboratory internal validations has only recently become an option through the advent of FSI: Reports. Previously, there was no mechanism for the publication of a study that was not characterized as novel.
- The suggestions regarding emulating the efforts of the digital PCR community for standardizing the information included in studies is well taken. The standardization of the information for reliability assessment can only serve to aid in future undertakings. That this information has not previously been standardized, however, does not serve as evidence against the reliability of probabilistic genotyping. Again, a vast wealth of data has been generated in the last six years in the United States, simply not in the exact format suggested in this report.
- Regarding the abundance of data that already exists in support of the reliability of probabilistic genotyping, no reviewing body is suggested, nor is a central repository for forensic DNA laboratory validation data currently in existence. The call for data is not supported by any existing structure through which to funnel it or evaluate its merits.

Thank you for the opportunity to provide comment on the draft *DNA Mixture: A Scientific Foundation Review* report and for your thoughtful consideration.

Sincerely,

Lauren Lu
DNA Technical Leader
Michigan State Police
Forensic Science Division

PC39

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

Griffith, Robert T. [redacted]

Mon 8/23/2021 3:19 PM

To: ScientificFoundationReviews <ScientificFoundationReviews@nist.gov>

Good afternoon,

Please see the MDPD FSB response to *NISTIR 8351-DRAFT DNA Mixture Interpretation: A NIST Scientific Foundation Review* attached to this email.

Thank you,

Robert Griffith
DNA Technical Leader
Miami-Dade Police Department
Forensic Services Bureau - Forensic Biology Section

[redacted]
www.mdpd.com
www.miamidade.gov
"Delivering Excellence Every Day"

PC39a

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

Miami-Dade Police Department, Forensic Services Bureau

August 23, 2021

DNA Mixtures have become increasingly more complex as both instrument and PCR amplification kit sensitivities have increased. There is a spectrum of complexity of DNA mixtures seen in casework samples, and the more complex mixtures cannot be interpreted in the same manner as the less complex mixtures. However, these differences do not equate to a spectrum of reliability of DNA mixture data. Some mixtures are not amenable to current mixture deconvolution methods; as such, more complex mixtures may be prone to misapplication of currently available protocols. In a DNA casework laboratory, quality assurance processes are in place to review all data interpretation. For example, a case file is technically reviewed prior to a report being released; further, an accredited DNA casework laboratory also undergoes external FBI QAS audits every other year. These audits include a review of the laboratory's validation data; in addition, the audit team reviews the application of the laboratory's methods in casework via review of actual DNA case files. Additionally, all DNA Technical Leaders are strongly encouraged to attend training at the annual CODIS meeting to ensure consistency across laboratories. In Key Takeaway #4.6 of the draft report (line 763), the authors state that "Different analysts and different laboratories will have different approaches to interpreting the same DNA mixture. This introduces variability and uncertainty in DNA mixture interpretation. Improvements across the entire community are expected with an increased understanding of the causes of variability among laboratories and analysts." The interpretation of mixture data observed in casework is conducted by DNA analysts; protocols are set to achieve as much consistency as possible and to mitigate any potential bias in the interpretation. Further, it is impossible to write mixture interpretation guidelines that address every casework scenario. As stated by Butler, et. al. in 2018, *"As we look to the future, the community may ask if there are obvious improvements necessary to achieve more reliable mixture interpretation. Is it possible to produce a "standard" mixture approach that all laboratories can implement to achieve consistency across the United States or around the world? Probably not. Protocols for interpretation are developed depending on different chemistries, different capillary electrophoresis platforms, different philosophies on interpreting mixtures, and the experience and training of analysts in the laboratory. However, we should nevertheless strive to achieve consistency within each laboratory to avoid the possibility of different conclusions."* [1] The laboratory's interpretation guidelines (based on its own validation studies), the technical review process, as well as the FBI QAS external audits guide that consistency.

Further, the FSB supports transparency in scientific data; all internal validation data are available in the laboratory and, as previously stated, are reviewed internally as well as externally during accreditation audits. The FSB also participates in inter-laboratory validation and/or research studies where data is publicly released (i.e., MIX05, MIX13, NDIS validations, etc.). As such, the information is already available for consumption. Therefore, as with all previous validation studies, there is no need to add the validation data to a DNA report. It is unclear as to why the authors claim that this information would be helpful to understanding the report.

---

[1] Butler JM, Kline M, Coble M (2018). NIST Interlaboratory Studies Involving DNA Mixtures (Mix 05 and Mix 13): Variations Observed and Lessons Learned, *For Sci Int: Genetics* 37: 81-84.

Lastly, the authors state that the relevance of DNA evidence should be assessed. The FSB strongly disagrees. Evidence submission policies are in place to screen evidence prior to DNA analysis. DNA analysts conduct their analysis on the evidence items that are submitted, issue a report, and testify to their results. The question of relevance is not a question for the scientist to answer. The question of the relevance of an item is asked of the submitting entity at the point of evidence submission. The authors' comments that DNA analysts should determine potential relevance introduces the same bias that the FSB and laboratories nationwide have been working to mitigate. Furthermore, the ISO 17025 international accreditation standards followed by accredited crime laboratories emphasize the importance of ensuring that the scientific analyses are objective and free from undue influence. Line 1976 of the NIST draft report states, "It is always the trier-of-fact's final decision whether the DNA originates from a specific person or not and the relevance of this information."

NIST should withdraw this draft report and make the appropriate changes as detailed above.

PC40

Public Comment on DNA Mixture Interpretation: A Scientific Foundation Review

Rosenblatt, Terri [redacted]

Mon 8/23/2021 4:29 PM

To: ScientificFoundationReviews <ScientificFoundationReviews@nist.gov>

Please find attached a public comment on NIST's "DNA Mixture Interpretation: A Scientific Foundation Review," submitted by public defense providers.

Thank you for your consideration.

Very truly yours,

Terri S. Rosenblatt
Supervising Attorney, DNA Unit
The Legal Aid Society
199 Water Street, 5th Floor
New York, New York 10038
[redacted]

THE LEGAL AID SOCIETY CRIMINAL DEFENSE

PC40a

New York County Criminal Defense Office
49 Thomas St.
New York, NY 10013
(212) 732-5000

https://www.legalaidnyc.org/

John K. Carroll
President

Janet E. Sabel
Attorney-in-Chief
Chief Executive Officer

Justine M. Luongo
Attorney-in-Charge
Criminal Practice

David Loftis
Attorney-in-Charge
of Post-Conviction and Forensic Litigation

August 23, 2021

Via E-mail to scientificfoundationreviews@nist.gov

Dr. John Butler, *et al*.
National Institute of Standards and Technology
Attn: Special Programs Office – Scientific Foundation Review
100 Bureau Drive Stop 4701
Gaithersburg, MD 20899-4701

Re: NISTIR 8351-DRAFT
DNA Mixture Interpretation: *A NIST Scientific Foundation Review*
Public Comment of 11 Public Defenders and Defense Organizations

Dear Dr. Butler:

As providers of defense services to indigent clients charged with crimes, we read with great interest the NIST publication, *DNA Mixture Interpretation: A Scientific Foundation Review*[1]. We are encouraged by NIST's critical analysis, and its conclusion that more information is needed to determine the reliability of "probabilistic genotyping software," the commercial computer programs used by laboratories to interpret complex DNA mixtures.[2]

However, we believe that NIST has a further responsibility as our nation's leader in scientific standards-setting to recommend that laboratories impose a moratorium on the use of probabilistic genotyping software until:

(1) laboratories and developers provide sufficient data for NIST to complete an independent assessment of its reliability[3];
(2) laboratories demonstrate that their analysts are proficient in complex mixtures, not just one- and two-person samples[4];
(3) laboratories update their reporting to indicate where the questioned sample falls within its validation[5];

[1] *Available at* https://www.nist.gov/dna-mixture-interpretation-nist-scientific-foundation-review ("NIST Review").
[2] The NIST Review defines "complex mixtures" as "a DNA profile resulting from comingled DNA of two or more contributors that is difficult to interpret due to uncertainty in the determination of contributor genotypes; factors complicating mixture interpretation include, but are not limited to, low quantity DNA, low quality (degraded) DNA, the number of contributors, and the amount of allele sharing." NIST Review, p. ix.
[3] *See* NIST Review, p, 75 (Key Takeaway #4.3).
[4] *See* NIST Review p. 6 (Key Takeaway #4.5).
[5] *See* NIST Review, p. 6 (Key Takeaway #4.7).

(4) scientists determine valid and reliable methods to close the "knowledge gap" surrounding DNA transfer and persistence[6]; and
(5) a group of qualified experts, including impacted people, perform a racial impact assessment to determine how the current use of this software impacts historically oppressed groups.

Without these changes, stakeholders are left with little guidance on how to proceed from here. Whether and how courts or juries will be influenced by NIST's questioning of these systems will be left to individual jurists or triers of fact, leading to inconsistent and confusing results. The fairness of the forensic science used against any individual defendant, then, will be based on the jurisdiction where they are prosecuted, rather than on any consistent guidance. The only thing that will be consistent is that the people impacted by unfair or untested DNA evidence will be in the same group who are already victimized by over-policing and prosecution—people from historically oppressed racial and ethnic groups.

Forensic science has been plagued with a history of prioritizing the development of results for the court above conducting rigorous validation and testing procedures observed in other scientific disciplines. For example, hair microscopy was once considered a promising science. However, due to its limited testing and validation, it is now recognized as the cause of many wrongful convictions.[7] The use of bite-mark matching similarly was considered, without empirical support, to be pioneering in solving crime, but subsequent research revealed it was nothing more than non-science.[8] These methods, like the ones evaluated by NIST in its report, were also described as simply needing more research[9], but nevertheless were used in court in the meantime.

To be clear, traditional forensic DNA analysis is different from methods solely developed by crime labs. DNA occupies a special place in forensics. As the NAS report recognized more than a decade ago[10], DNA analysis stands out as the singular discipline developed by scientific researchers for non-forensic purposes. When used correctly, STR-based DNA analysis can be reliable and largely

---

[6] *See* NIST Review, p. 140 (Key Takeaway #5.6).
[7] *See* Spencer S. Hsu, *FBI admits flaws in hair analysis over decades*, The Washington Post (Apr. 18, 2015); Innocence Project Staff, *How Santae Tribble's Wrongful Conviction Prompted Review of the FBI's Use of Hair Analysis and Inspired the Innocence Project's Research* (Jul. 5, 2020), *available at* https://innocenceproject.org/santae-tribble-inspired-hair-analysis-review-work/.
[8] Elizabeth Ann Brown, *Most agree bitemark matching is junk science. Why is it still in courts?* The Legal Examiner (Dec. 23, 2020).
[9] *See* Committee on Identifying the Needs of the Forensic Sciences Community, *Strengthening Forensic Science in the United States: A Path Forward*, United States Department of Justice (2009) ("NAS Report"), p. 119("the microscopic hair analysis process must be subjected to performance and validation studies in which appropriate error rates can be defined and estimated."), p. 42 (highlighting bite marks as having "never been exposed to stringent scientific scrutiny" but asserting there is "logic behind" the technique).
[10] NAS Report, p. 7.

race-neutral. The NIST Review makes clear, however, that there is a marked difference between single-source, simple mixtures, and complex mixture comparisons. When it comes to these complex mixtures, the software used bears more of a similarity to hair and bite marks analysis than it does to traditional gold standard analysis. These programs, unlike traditional DNA, were created largely by crime labs for the sole purpose of securing criminal convictions.[11] In this way, these programs more closely resemble hair and bite mark analysis. Those disciplines, too, found their initial basis in biological sciences, but were later misused by crime labs and prosecutors.

We don't yet know the extent of wrongful convictions caused by probabilistic genotyping. In cases where complex mixtures are analyzed, there is no "ground truth" to compare against. Traditional DNA provided an objective basis to test the reliability of hair or bite marks results. But in the complex DNA context, there is, by definition, no similar single-source benchmark. In reported cases that do exist, there is cause for concern. In *People v. Oral Nicholas Hillary*, STRmix produced inculpatory results in a case where the accused had an alibi and lacked motive.[12] Mr. Hillary was acquitted at trial, but only after facing years of living under the accusation of murdering a child. In another example, research by members of the defense community reveals that false positives occur when a DNA mixture is comprised of biological relatives who also are related to the person of interest.[13] Seemingly inculpatory probabilistic genotyping results also can pressure people to accept plea bargains that avoid harsh mandatory minimum sentences. Faced with such evidence, even innocent clients, or clients who suffered at the hands of police misconduct, are put in an impossible position of either hoping that a judge or jury will understand the limitations of probabilistic genotyping, or accepting a plea bargain for a reduced sentence.[14]

The impact of these wrongful convictions is not borne equally. Black people disproportionately are the targets of unfair or unreliable forensics. Indeed, data compiled by the National Registry of Exonerations found that, although Black people comprise 13% of the United States population, they account for almost 50% of the known wrongful convictions.[15] Black people are also *seven times* more

---

[11] *See* NIST Report, p. 42.
[12] Jesse McKinley, *Oral Nicholas Hillary Acquitted in Potsdam Boy's Killing*, The New York Times (Sept. 28, 2016), *available at* https://www.nytimes.com/2016/09/29/nyregion/oral-nicholas-hillary-potsdam-murder-trial-garrett-phillips.html; Hank Stuever, *Who Killed Garrett Phillips? Is a masterful study in the evils of assumptions*, The Washington Post (Jul. 22, 2019), *available at* https://www.washingtonpost.com/entertainment/tv/who-killed-garrett-phillips-is-a-masterful-study-in-the-evils-of-assumptions/2019/07/21/97d947a0-aa4d-11e9-a3a6-ab670962db05_story.html.
[13] Brooklyn Defender Service, *Upcoming Report Publication: The Kinship Problem*, *available at* https://indefenseof.us/issues/kinship-problem (last visited Aug. 17, 2021).
[14] *See, e.g.*, Judge Jed Rakoff, *Why Innocent People Plead Guilty*, The New York Review (Nov. 20, 2014), *available at h*ttps://www.nybooks.com/articles/2014/11/20/why-innocent-people-plead-guilty/.
[15] National Registry of Exonerations, *Race and Wrongful* Convictions, *available at* https://www.law.umich.edu/special/exoneration/Pages/Race-and-Wrongful-Convictions.aspx (last accessed Aug. 17, 2021).

likely than white people to be wrongfully convicted of murder—one of the charges that most often utilizes probabilistic genotyping.[16] Probabilistic genotyping is also frequently used in weapons possession cases in New York City. The arrests overwhelmingly target Black people who, despite comprising 18% of New York's population, represent 78% of this category of arrest.[17] Therefore, those people are disproportionately impacted by the use of probabilistic genotyping and necessarily more likely to suffer from any wrongful results generated.

NIST has a singular opportunity to help prevent further harm and mitigate past wrongs by making the concrete recommendations we request, including undertaking an impact analysis in its final review. The NIST Review concludes that probabilistic genotyping needs more research. But it does not recommend that laboratories stop using the programs before this research is completed. The impact of this is that the subjects of probabilistic genotyping in its experimental stage are predominantly Black people. This echoes a shameful history of using untested science on this community.[18] If NIST deems these issues outside of its jurisdiction, it is relinquishing its responsibility as a standard-setting institute to prevent the use of scientific testing on a population that has historically been victimized by untested or unproven science.

Scientists and regulators in a number of disciplines –even supposedly "race blind" ones—are answering the call to include an ethical and racial impact assessment of their work. Last year, in the wake of the murder of George Floyd, more than 1,400 scientists and mathematicians wrote an open letter calling for scientists to meaningfully engage with the racial impact of their work.[19] In this letter, they urged scientists to stop collaborating with police departments, given the structural racism endemic in American policing. At least, they asserted, government work with law enforcement should be closely interrogated, as "It is simply too easy to create a 'scientific' veneer for racism."[20] Also last year, the United States House of Representatives introduced a bill that would require NIST to do the very thing we are asking it to incorporate in this report: "an assessment for the potential for disparate

[16] Niraj Chokshi, *Black People More Likely to Be Wrongfully Convicted of Murder, Study Shows,* The New York Times (Mar. 7, 2017), https://www.nytimes.com/2017/03/07/us/wrongful-convictions-race-exoneration.html

[17] *See* Avainash Nitin Samarth, *Brief of Black Attorneys of Legal Aid, et al, as Amicus Curiae in Support of Petitioner in* New York State Rifle and Pistol Ass'n, Inc, et al. v. Corlett, et al., Case No. 20-843, *available at* https://www.supremecourt.gov/DocketPDF/20/20-843/184718/20210723101034102_20-843%20Amici%20Brief%20revised%20cover.pdf, p. 14.

[18] *See* Ada McVean, *40 Years of Human Experimentation in America: The Tuskegee Study*, McGill University (Jan. 25, 2019), *available at* https://www.mcgill.ca/oss/article/history/40-years-human-experimentation-america-tuskegee-study. *See generally Alondra Nelson, Body and Soul: The Black Panther Party and the Fight against Medical Discrimination,* Minneapolis, University of Minnesota Press (2011).

[19] *See* Davide Castelvecchi, *Mathematicians urge colleagues to boycott police work in wake of killings*, Nature (June 19, 2020), *available at* https://www.nature.com/articles/d41586-020-01874-9.

[20] *Id.*

impact, on the basis of race, ethnicity, socioeconomic status, gender, and other demographic features" in the development and use of the computational forensic software.[21]

In addition to our requests in this letter, we join in the comments of our colleagues calling for the NIST Review to go further in recommending that NIST further interrogate issues with false inclusions due to relatedness, laboratories operating outside of validation, and police officers who refuse to provide elimination samples. These issues, and the many identified in the report are critical to improving the reliability of PGS. But unless or until those improvements are made, there is only one just conclusion: NIST should recommend a mortarium and assess the racial impact of probabilistic genotyping.

Very truly yours,

The Legal Aid Society
DNA Unit
199 Water Street
New York, New York 10038

Terri S. Rosenblatt
Supervising Attorney, DNA Unit

Office of the Appellate Defender
11 Park Place, 16th Floor
New York, NY 10007

Caprice R. Jenerson
President & Attorney-in-Charge

Appellate Advocates
111 John Street, 9th Floor
New York, New York 10038

Patricia Pazner, Esq.
Acting Attorney in Charge

The Bronx Defenders
360 East 161st Street
Bronx, NY 10451

Emily J. Prokesch, Esq.
Forensic Practice Director
Paul Vernon, Esq
DNA Staff Attorney
Hannah Rosenthal, Esq.
DNA Staff Attorney

Office of Capital & Forensic Writs
Stephen F. Austin Building
1700 N. Congress Avenue, Suite 460
Austin, Texas 78701

[21] *See* H.R. 4368, "Justice in Forensic Algorithms Act of 2021," available at https://www.congress.gov/bill/117th-congress/house-bill/2438/text#:~:text=Introduced%20in%20House%20(04%2F08%2F2021)&text=To%20prohibit%20the%20use%20of,Program%2C%20and%20for%20other%20purposes.

Center for Appellate Litigation
120 Wall Street, 28th Floor
New York, NY 10005

Marika Meis, Esq.
Co-Director, Forensic Science Project

Office of the Public Defender
6 St. Paul Street, Suite 1508
Baltimore, Maryland 21202

Jeff Gilleran, Esq.
Chief Attorney, Forensic Division
Andrew Northrup, Esq.
Forensic Division

Committee for Public Counsel Services
75 Federal Street, 6th Floor
Boston, MA 02110

Ira Gant, Esq.
Forensic Services Director

The Minnesota Board of Public Defense
331 2nd Avenue S
Unit 900
Minneapolis, MN 55401

Ginny Barron, Esq.
JD Schmid, Esq.
Assistants Public Defender

The Monroe County Public Defender
10 North Fitzhugh Street
Rochester, New York 14614

Timothy Donaher, Esq.
Monroe County Public Defender

Public Defender Service for the District of Columbia
633 Indiana Avenue NW
Washington, DC 20004

Jessica Willis, Esq.
Special Counsel

/

PC41

NISTIR 8351-DRAFT

DeHaan, Mackenzie [redacted]

Mon 8/23/2021 4:41 PM

To: ScientificFoundationReviews <ScientificFoundationReviews@nist.gov>

The following are comments to the Draft of the DNA Mixture Interpretation: A NIST Scientific Foundation Review (NISTIR 8351) along with recommended changes, where applicable. This review contains a lot of information and the effort and time spent to complete this draft is appreciated.

- Lines 531 and 532 of the executive summary as well as line 1160 both note that the findings in this report are meant to inform future work. Adding what this means for readers would be helpful due to the diverse backgrounds of the readers. This is evidenced by the fact that this Draft report has already been cited during a Frye hearing request in the State of New York. Perhaps the addition of a scope and applicability section would better inform readers.
- Line 629 discusses how "successful analysis and interpretation of DNA results depends on…**and** the availability of a reference sample" Per ISO 17025 standards as well as QAS standards, analysis and interpretation is independent on the existence of a reference sample. Recommendation is to change the word "and" or separate this sentence to be more appropriate based on standards relied upon by Forensic Caseworking laboratories.
- Key Takeaway #2.2: This definition of interpretation is broad. It also infers that case context has to be used for any interpretation to be performed. Recommendation: more detailed definition of interpretation, in a manner to be consistent with existing audit and accreditation standards already in place would be helpful.
- Key Takeaway #4.3: Defining "publicly available" as those that can be found from an internet search is not fully representative of the data that is available for review. Independent assessments are routinely performed during the audit processes. Inferring that the use of PG systems is unreliable based on a lack of data available on the internet is lacking context.
- Key Takeaway #4.7: Including the validation performance results in the case files and report would make laboratory reports that are already criticized as lengthy even more involved. This information is available in discovery process and it is difficult to ascertain what wording could be added to the report that would make validation results understandable to the lay persons who are the users of these reports.
- Key Takeaway #4.8: A separate scientific review of LRs may be helpful, however much of the information has been included in this review. If the intent was do additional reviews, then the portions regarding which LR is being reported would be better suited in a future study.
- In line 1918 the study states "in recent years", but then cites articles that are 15 years old (or greater). If anything, having data ranging back decades in a field that is as rapidly changing as Forensic DNA analysis supports the foundational use of the theory and applications. Either reword this sentence or use more recent citations.
- Section 2.5.2 (beginning on line 1937) has the intent on discussing the testimony and implications on incorrect understanding of the LR results. This section seems out of place in a foundational review of DNA interpretation, recommend removing it to place in future review.
- Chapter 5 appears out of place in a review about DNA mixture interpretation. Topics including the hierarchy of propositions, context of results, case assessment and interpretation, and activity propositions are important but factor more into the relevance of the result and not the interpretation and reporting of the results that were obtained during DNA analysis.
- Chapter 6 offers possible future technology suggestions. This section is not addressing the focus of this review, which was previously stated to be a foundational review to explore the limits of DNA mixture interpretation methods, including probabilistic genotyping systems.
- Appendix 2, line 6991 discusses a standardized competency testing. This does not give any guidance to how this could be accomplished between the numerous laboratories that have different submission acceptance and different validation studies allowing for interpretation limits of different number of contributor mixtures. If it is to be inferred that one standardized competency should exist for the community, then to what level. There are existing standards to have written training programs that detail

the training and competency required to be a qualified analyst within the lab. Competency testing is defined in standards as being written, oral, and practical, not simply practical as written in line 7008.

If I can clarify any comments, please feel free to ask.

Respectfully,

Mackenzie DeHaan, MS, ABC-MB
Forensic Scientist Supervisor/DNA Technical Leader/Special Agent
Forensic Biology Section
[redacted]
121 Tryon Road, Raleigh, NC 27603
ncdoj.gov

PC42

Feedback to the DNA Mixture Interpretation NIST Scientific Foundation Review (NISTIR 8351-DRAFT)

Nicholson, Mary Lou [redacted]

Mon 8/23/2021 4:51 PM

To: ScientificFoundationReviews <ScientificFoundationReviews@nist.gov>
Cc: Crossman, Christine [redacted]; Litzenberger, Greg [redacted]; Dare, Tanya [redacted]; Beltran, Gerard [redacted]; Neudorf, Shaun [redacted]; Lett, Marc [redacted]; Trudel, Isabelle (P&PS) [redacted]

Overall, there is a lot of good information in this review paper and it would be a good resource for DNA Analysts.

General Comments for Consideration:

- NIST is not a regulatory body, so cannot say what is necessary for laboratories to do (should be stated in limitations of study)
- NIST does not comment on admissibility which would be an important aspect to cover
- NIST should clarify that laboratories' validation data is available for external review upon request by the courts during the disclosure request process and external accreditation audits. These are the USER group for forensic laboratories (Provider).
- Provide a disclosure that the primary authors of the review paper have no forensic DNA casework experience. The DNA Resource group were not provided actual input into the writing of the paper.
- NIST can't comment on how a laboratory conducts its internal validation but if NIST was to actually contact laboratories to review some internal validations, NIST could advise which ones were conducted according to the guidelines issued by a scientific body (e.g. SWGDAM, ENFSI, etc)
- NIST commentary on certain items carries a lot of weight. Need to ensure that the comments in the review paper are not reckless or misconstrued (i.e. see comments below).
- Not certain as to why there was a blog post made on July 28th, 2021 by Dr. Butler on the NIST website discussing this review paper when it was considered only a draft and was still out for public comment. It would have been more appropriate to wait until after public comments were received/considered and the paper was deemed final.

Specific Comments for Consideration:

Line 125 (Preface), it would be nice to indicate how much funds were allocated to this review and the total amount of time devoted to the review.

Line 1510-1515: Mixture interpretation not only based on training and experience of the Analyst. In our laboratory, interpretation guidelines are based on internal validation of samples that mimic casework samples. Missing importance of validation studies in defining mixture interpretation guidelines in the laboratory.

Line 1528: Key Take Away 2.2: missing reliance of the Analyst on the laboratory's internal validation which defines mixture interpretation guidelines.

Line 2123: Key Take Away 2.6: LR's are used in forensic casework for relatedness type of cases (i.e. paternity, parentage) in N. America. Extensively used in Europe for forensic DNA casework with no major issues. When used in DNA casework, LR's are measurements just like the random match estimate is a measurement. Uses empirical data for allele frequencies to calculate the value. LR = 1/RMP. So technically, it is a measurement. The same can be said for an RMP used in the binary approach, different values of an RMP can be assigned based on different kits, protocols, models, assumptions, computational models.

Line 2350: should read "in favor of a".

Line 2352: Does not make sense as to why the propositions need to be exhaustive. It seems this may be drawn out from the third principle on line 2361. This third principle is talking about incorporating LR's beyond the sub-source level in the hierarchy of propositions. PGS and binary interpretation methods both only assess at a sub-source or sub-sub source level, and LR's are not generally offered for anything above this in forensic labs. Below the activity level proposition, the LR would not necessarily need to be exhaustive to be relevant and informative.

Line 3074: Takeaway 4.3: It is problematic for a forensic laboratory to release validation raw DNA data to the public. First, accreditation requirements for forensic laboratories requirements are quite stringent on who has access to the lab and the data generated by the lab. There are numerous safeguards put into place to ensure integrity of the data and that it isn't being compromised. Second, donors who donate their biological samples for the validation studies do so based on consent for purposes solely of the validation study. Most donors would not consent to having their DNA profiles released to the public. Privacy is a big factor in protecting the data. How do you ensure the "public" is not using the data for nefarious purposes?

Line 3425, Key Take Away 4.4: Although not explicitly stated that PGS is considered unreliable (due to lack of sufficient detail in published articles/public domain), there is an implication that PGS is unreliable. This has already been stated as such in a recent court challenge in the Supreme Court of the State of New York (People vs Daval Wright) that quotes from this draft version of the review paper.

Line 3460: Key Take Away 4.6: PGS systems should actually assist in ensuring less variability and uncertainty in DNA mixture interpretation, especially within a laboratory.

Table 5.3, number 12: "The risk of contamination must…" The word "must" Is split on two lines.

Line 4646: Key Takeaway 5.4: Using the RMP in a binary approach doesn't take into account DNA transfer also (see comment for Line 4765). RMP reports at the sub-source level too – establishing an association to the DNA profile. Same as the LR used in PGS. As forensic DNA scientists, we cannot provide this information as it is unknown. There is no sound empirical data to support doing so. Based on Bayes theorem, the forensic DNA scientist can only report the LR or 1/RMP and not the posterior or prior odds. Suggesting that the LR is misleading without context removes the burden from the trier of fact and places it on the scientist who doesn't have all of the information and should be seen as unbiased.

Line 4765: The CAI-based reasoning is troubling in it's attempt to assign probabilities to activity level propositions. This line is key, as we do not, and will not ever have a complete enough understanding of transfer and persistence outside of a laboratory, much less outside a specific case to accurately assign an LR related to activity level propositions. If takeaway 5.4 states sub-source level LR's can be misleading without context, then I'm not sure how assigning LR's at the activity level will in any way rectify this – I think this has a greater risk to make things more misleading, not less.

Thank you for your consideration.

Regards
Mary Lou

Mary Lou Nicholson, Ph.D.
Program Technical Leader, Biology Services
RCMP Forensic Science and Identification Services
1200 Vanier Pkwy, P.O. Box 8885
Ottawa, ON K1G 3M8
[redacted]

PC43

Response to NIST Report

Cecilia Von Beroldingen [redacted]

Mon 8/23/2021 5:12 PM

To: ScientificFoundationReviews <ScientificFoundationReviews@nist.gov>

Please find attached my comments on the draft NIST report on DNA Mixture Interpretation.

Sincerely,
Cecilia H. von Beroldingen, Ph.D.

PC43a

Comments on NIST Draft Report: DNA Mixture Interpretation

This is a monumental piece of work. It is an impressive and masterful description of the history, background, and current methods used in DNA mixture interpretation, including the potential limitations and pitfalls and, as such, will be a valuable resource to the community. Nevertheless, I do have issues with some of the key takeaways.

With regard to Key Takeaway 4.3, which states that "Currently, there is not enough publicly available data to enable an external and independent assessment of the degree of reliability of DNA mixture interpretation…", I would say that gaps in available data are not necessarily the result of labs not performing the appropriate studies. More effort should be made to collect data from internal validation studies, not just relying on an internet search. In addition, statements such as the one below are a disservice and even an insult to the community.

"*One explanation for this lack of public data is simply that there has been no expectation to provide it. Choosing not to make public the data underlying decisions that are made in laboratory protocols is generally without consequence, while giving public access carries a risk of increased scrutiny.*"

As a former member of two crime laboratory systems and an auditor of several others, as well as a past member of TWGDAM and SWGDAM, I am troubled by the implication that laboratories choose not to make the results of their validations public because they want to avoid the risk of public scrutiny. With a few exceptions, internal validation studies are not considered to be sufficiently novel to be published in scientific journals. Many laboratories present the results of their internal validation studies at regional and national meetings, but most are not published as a journal article. This is not because of the desire to avoid peer review but because journals do not typically accept such manuscripts.

Public and private crime laboratories are making a genuine effort to comply with existing standards and guidelines for the internal validation of DNA mixture interpretation procedures. They rely on inspection by accrediting bodies to determine whether their internal validation studies are in compliance with the accepted and published standards. I acknowledge that crime laboratories may approach internal validation studies from a task-driven rather than a performance-based perspective, but I am heartened to hear that standards and guidelines are becoming both more detailed and performance-based.

Probablistic genotyping software has been an important advance in DNA mixture interpretation. That the report did not compile and consider all the available data, not just what was readily available via an internet search, should not be cited as a basis upon which to pronounce the method unreliable.

The statement that there is a lack of sufficient publicly available data reminds me of the controversy surrounding the NRC I report. Practitioners of that era feared there would be a moratorium on forensic DNA testing because the NRC I authors felt that there was not sufficient population data to support match probability calculations (although they did offer the hyperconservative "ceiling principle" method). Instead, the report galvanized the community, under the auspices of the FBI, to collect and analyze the available population data which ultimately resulted in a more reliable statement on the

weight of the DNA evidence. Nevertheless, forensic labs did not stop performing DNA testing. Instead, DNA analysts became more aware of the potential limitations and incorporated strategies to address these limitations in their interpretation guidelines.

With regard to proficiency testing (Key Takeaway 4.5), It is very difficult to design proficiency test samples that mimic all the possible casework scenarios. To design a proficiency test containing a complex DNA mixture with multiple donors representing low-level and degraded DNA contributors would be difficult. One potential problem I foresee is that it would be challenging to produce uniform test samples with these attributes. Even small differences in DNA content among samples could result in one lab attaining the designed "correct" results whereas another does not.

With regard to Key Takeaway 4.7: "… it would be helpful to include these validation performance results in the case file and report." This recommendation is unclear to me. My experience in performing validation studies is that the amount of data compiled in these studies comprises several notebooks, which can be made available upon request. I do not believe it would be practicable to provide a copy of this data as part of a case file or report.

In conclusion, PGS has been an important advance in DNA mixture interpretation. The perceived deficiencies in available validation data should not be cited as criteria to pronounce uncertainty on the reliability of the method. I have two recommendations:

First, the report indicates that there is not sufficient data coverage in publicly available validation studies to be able to assess whether DNA mixture interpretation procedures are reliable. I would contend that if the criteria by which the validation studies are judged are known, laboratories might be able to provide this information.

Second, it would be of benefit to both the forensic science community and the criminal justice system to have a government agency such as NIST that would serve as a repository to collect validation data from forensic laboratories to get a more complete picture of what data is available.

To make a statement to the effect that there is not sufficient data available to assess the reliability of DNA mixture interpretation procedures without making more of an effort to collect this data will have a negative impact on the community's progress towards a more objective approach to mixture interpretation.

Finally, I want to acknowledge and thank the authors for all the excellent information presented in the report. As I mentioned earlier, this report will be a valuable resource.

Cecilia H. von Beroldingen, Ph.D.

PC44

comments

Gross, Ann M. (DPS) [redacted]

Mon 8/23/2021 6:01 PM

To: ScientificFoundationReviews <ScientificFoundationReviews@nist.gov>

Please see the attached document for comments from the Minnesota BCA Forensic Science Services Laboratory regarding the NIST Scientific Foundation Review draft document.

Thank you for the opportunity to comment.

Ann Marie

ANN MARIE GROSS | TECHNICAL LEADER | BIOLOGY SECTION

BUREAU OF CRIMINAL APPREHENSION
FORENSIC SCIENCE SERVICES

1430 MARYLAND AVENUE EAST, ST. PAUL, MN 55106

[redacted]

PC44a

DNA Mixture Interpretation: A Scientific Foundation Review

https://www.nist.gov/dna-mixture-interpretation-nist-scientific-foundation-review

Comment Form

Submitted By (Initials)	Line #	Comment	Suggestions
MN BCA Forensic Science Service Laboratory	1164-1166 (and 1172-1173)	NONE of the members of the review team have ANY hands on DNA mixture interpretation experience in an accredited Forensic DNA laboratory	Assembling a review panel without any currently qualified DNA scientists (with experience using probabilistic genotyping) was a tremendous oversight and a disservice to the forensic community. It is clear that this perspective is missing with the following recommendations: Key takeaway #4.7
MN BCA Forensic Science Service Laboratory	Key Takeaway #4.7	"Including validation performance results in the case file and report" – to include this information in a report is simply an absurd suggestion.	Any forensic practitioner that has written a report knows that it is imperative to keep the report as simple and succinct as possible. Understanding that the "client" (or "user" as NIST refers to them) are generally police officers. In addition, this completely goes against ISO/IEC 17025:2017, section 7.8.1.2 – "The results shall be provided accurately, clearly, unambiguously, and objectively… and shall include all info agreed with the customer and necessary for the interpretation of the results…" If this information is needed by a defense expert, it can be obtained in discovery.
MN BCA Forensic Science Service Laboratory	2848-2850	"factor space" and "factor space coverage" – these terms not generally used by forensic practitioners in discussing validation studies, why are they being introduced in a review publication. These terms are not used in any other document which laboratories use for guidance in validating new technologies (e.g. QAS, SWGDAM guidelines, OSAC standards).	Why are these new terms being introduced in a review publication? This information is covered in the QAS for internal validation, specifically Standard 8.3.2.1 Mixture interpretation validation studies shall include samples with a range of the number of contributors, template amounts, and mixture ratios expected to be interpreted in casework.
MN BCA Forensic Science Service Laboratory	2486-2488	Authors are acknowledging that additional internal validation data likely exists, but they chose to conduct the scientific foundation review using only publicly available information.	Choosing to not even try to obtain the data to look at is an irresponsible approach to conducting a scientific foundational review and is doing a dis-service to forensic science. . (During the public webinar, it was asked how many of the authors of

			the 60 Probabilistic Genotyping publications were contacted to see if the data was available to review, and J. Butler replied none of them had been contacted)
MN BCA Forensic Science Service Laboratory	2968	Why did the reviewers limit their analysis to publicly available data? It would be very rare for laboratories to make their data publicly available -	Re-do chapter four after looking at data now available. How can it be considered a review if sufficient data was not looked at.
MN BCA Forensic Science Service Laboratory	Key Takeaway #4.5	The authors are suggesting that more complex and/or low-template components be used in creating proficiency tests. This may be a good idea in theory, but in reality to prepare samples for distribution (consistent for the 100's of tests needed) and the "scoring" of this type of proficiency test – clearly is not easily implemented.	A suggestion to create this type of proficiency is meaningless unless the authors also define how these tests will be made, distributed and graded.
MN BCA Forensic Science Service Laboratory	Chapters 5 and 6	These two chapters do not belong in a foundational review of mixture interpretation.	Remove chapters. Interesting information, but for a separate review document, not for inclusion in a foundational scientific review on mixture interpretation. These can be separate review documents.
MN BCA Forensic Science Service Laboratory	7256-7259; 7271-7273	Who is this "advisory group"	This review states that an "advisory group" is necessary without proposing who the members are. It is imperative that the proposed "advisory group" be comprised of at least 50% currently qualified DNA scientists who conduct mixture interpretation, not just individuals who have theories about mixture interpretation or conducted it during the early days of DNA analysis.
MN BCA Forensic Science Service Laboratory	7289-7291	"Some portion of DNA analysts' paid time should be devoted to examining relevant books and articles published in the scientific literature."	This is already done as per the QAS, specifically Standards 16.1.1 and 16.1.2
MN BCA Forensic Science Service Laboratory	7291-7292	What is this statement based on? How many internal validation studies were looked at to determine that TLs "don't have sufficient training/experience to design validation experiments."	Is this statement based on anecdotal evidence? Did the review panel review 100's of validation studies to be able to make this statement? It is flippant and cavalier to make such a statement and is a slap in the face to the many TLs that design and conduct validation studies.

PC45

Public Comment re: DNA Mixture Interpretation: A NIST Scientific Foundation Review (NISTIR 8351-DRAFT)

Sarah Chu [redacted]

Mon 8/23/2021 6:06 PM

To: ScientificFoundationReviews <ScientificFoundationReviews@nist.gov>

Dear NIST Scientific Foundation Review report authors,

Thank you for the opportunity to provide feedback on the NISTIR 8351-DRAFT report. The Innocence Project's public comments are attached here.

Respectfully submitted,
Sarah Chu

--

Sarah Chu, MS
Sr. Advisor on Forensic Science Policy
Innocence Project
Pronouns: She/Her/Hers
[redacted]

Follow us
facebook icon linkedin icon twitter icon instagram icon

INNOCENCE PROJECT

PC45a

Executive Director
Christina Swarns, Esq.

Co-Founders & Special Counsel
Barry C. Scheck, Esq.
Peter J. Neufeld, Esq.

INNOCENCE PROJECT PUBLIC COMMENT ON
NISTIR 8351-DRAFT
DNA MIXTURE INTERPRETATION: A NIST Scientific Foundation Review
August 23, 2021

The Innocence Project is pleased to respond to the National Institute of Standards and Technology (NIST) call for public comments regarding the NISTIR 8351-DRAFT report, *DNA Mixture Interpretation: A NIST Scientific Foundation Review*. For nearly 30 years, the Innocence Project has worked to exonerate the innocent and to prevent wrongful convictions through systemic reform. The vast majority of our exonerations were achieved by the power and strength of forensic DNA evidence. Based on our decades of experience and success, we respectfully submit that as DNA analysis and interpretation becomes more complex, it must be applied with transparency and proper safeguards in order to ensure that forensic DNA technology serves its full potential to exonerate.

This scientific foundation review ("the report") is the first of the series produced by NIST and it may prove to be among the landmark publications in forensic science scholarship. In carrying out the work of this review, we commend the authors for operating with transparency, actively disseminating information regarding its process at conferences across the country, and now holding a public comment period to receive feedback. The feedback we respectfully offer addresses: (1) parts of the report we believe are critical to scientific rigor in forensic science and should therefore be retained, (2) concepts in the report for which stronger language or clearer directives are needed, and (3) parts of the report that we believe may be misinterpreted, manipulated, or create problems for the justice process without more context. Our comments below are listed using a chart that emulates the public comment process for standards development work and is organized by these three categories.

Critical Report Components to Retain

The following comments reference language in the report that are important to retain to ensure that the practice of forensic DNA testing is based on policies and protocols that promote a sound, quality, and just enterprise. With respect to concepts and language repeated or used multiple times throughout the report, we may reference a selection of excerpts but intend for our comments to apply globally. When new language or edits are suggested to resolve comments regarding excerpts

of the report, ~~strikethroughs~~ are used to indicate text that should be deleted and **[bracketed and bold text]** indicate text that should be added.

Lines	NISTIR 8351-DRAFT Language	Comment	Resolution
512-514	All scientific methods have limits. One must understand those limits to use a method appropriately. This is especially important in forensic science as critical decisions impacting life and liberty are often based on the results of forensic analysis.	This critical opening line in the Executive Summary sets the tone for the rest of the report.	This language should be retained in the final report.
699 and 1039 704-709	Reliability is not a yes or no question, but a matter of degree. Reliability centers on trustworthiness established through empirical assessments of available data to evaluate the degree of reliability of a system or its components. We use the term "factor space" to describe the factors that influence complexity, measurement, and interpretation reliability – these factors include the number of contributors, the degree of allele sharing, the ratios of mixture components, and the amount and quality of the DNA tested.	These lines denote the complexity of evaluating reliability of forensic science methods and are deeply connected to the reasons why forensic laboratories must practice transparency in every aspect of their work.	This language should be retained in the final report.
714-717	In addition, reliability cannot be established without validation tests using known samples of similar complexity. The results of such tests provide data that are considered accurate and reliable; only with such valid results can comparisons be made as to the reliability of unknown casework samples.	Validation studies are not typically provided in the discovery process. These lines demonstrate that one cannot properly interpret the findings of a test without first establishing the range of the forensic laboratory's valid testing capacity. It is important for courts to recognize this concept as attorneys seek validation study data. If validation studies were published online it would conserve criminal process resources and ensure that attorneys are able to access the data.	This language should be retained in the final report and a recommendation should be made for forensic laboratories to publish their validation studies online.
677-682 and 2123-2124	KEY TAKEAWAY #2.6: Likelihood ratios are not measurements. There is no single, correct likelihood ratio (LR). Different individuals and/or PGS systems often assign different LR values when presented with the same	Likelihood ratios (LRs) are difficult to communicate and are not well understood by the public. Great care needs to be taken when presenting LRs, especially when LR values are low and when these values are uninformative for	This language should be retained in the final report and a recommendation should be made that forensic laboratories set a threshold below which a

Lines	NISTIR 8351-DRAFT Language	Comment	Resolution
868-871 and 4646-4647	evidence because they base their judgment on different kits, protocols, models, assumptions, or computational algorithms. Empirical data for assessing the fitness for purpose of an analyst's LR are therefore warranted. Takeaway 5.4: DNA statistical results such as a sub-source likelihood ratio does not provide information about how or when DNA was transferred, or whether it is relevant to a case. Therefore, using the likelihood ratio as a standalone number without context can be misleading.	decision making. Forensic laboratories should heed these recommendations when developing guidelines and standards for LR testimony.	result should be reported as uninformative.
1141-1144	First, forensic genetics is an evolving field, and this study can only provide a snapshot of the state of the science at a particular moment in time. Therefore, the literature and empirical evidence we discuss in this review will be incomplete as soon as it is published, as is the case with evidence reviews in other evolving fields such as medicine and public health.	The authors model the scientific principle that knowledge accumulates, and that science is ever evolving and changes over time. This concept contrasts with legal traditions and its inclusion here is an important assertion of values that are central to science.	This language should be retained in the final report.
1495-1501	This overall process can be divided into two parts (Figure 2.1): (1) measurement that involves a series of steps to generate a DNA profile and (2) interpretation of the DNA profile to help fact finders understand the value of the evidence. The measurement steps result in an electropherogram (EPG), which is a representation of the DNA profile observed from the test sample at specific DNA locations. Interpretation of the EPG concludes with a written report describing a strength-of-evidence statistic for Q-to-K comparison with the POI(s), and in some cases, court testimony.	The separation of the DNA testing process into measurement and interpretation phases is important framing for communicating the critical importance of validation studies.	This language should be retained in the final report.
1510-1511	Measurements reflect the physical properties of the sample while	These lines communicate that interpretation is subjective. The	This language should be retained in the final report.

Lines	NISTIR 8351-DRAFT Language	Comment	Resolution
658-662 and 1528-1529	interpretation depends on the DNA analyst assigning values that are not inherent to the sample. KEY TAKEAWAY #2.2: Generating a DNA profile involves measuring the inherent physical properties of the sample. Interpreting a DNA profile involves assigning values that are not inherent to the sample. To do this, the DNA analyst uses their judgment, training, tools (including computer software), and experience, and considers factors such as case context.	measurement must first be done correctly, but the interpretation is based on many more factors. Analysts should be encouraged to be honest about the subjective nature of interpretation. During testimony, analysts frequently present their interpretation of the data as an objective process to avoid the accusation that they misinterpreted the evidence.	
775-777 and 3594-3595	KEY TAKEAWAY #4.8: We encourage a separate scientific foundation review on the topic of likelihood ratios in forensic science and how LRs are calculated, understood, and communicated.	Given the complexities of explaining the likelihood ratio and recommendations by Lund and Iyer (2017) to limit its use to personal decision making, a deeper evaluation of the LR would be beneficial to the field of forensic science.	We encourage NIST to conduct a separate scientific foundation review on the topic of LRs and their impact on the presumption of innocence.
2943-2946	On the other hand, when serving as an expert witness in a court setting, a forensic scientist is the provider of information while a trier of fact (judge or jury) and lawyers asking questions in the admissibility hearing or trial are users of the provided testimony. In this case, the judge, jury, and lawyers determine whether sufficient information has been provided.	The discussion of providers and users provides an important foundation for the discussion of transparency and data sharing. Importantly, the report also includes criminal justice stakeholders among the users of the data. Stakeholders missing from this list include defendants and persons of interest. People accused of crime should have access to all the data relevant to any evaluation of the evidence against them. Although a forensic scientist is responsible for answering the questions that are asked of them, when they are asked a question that is misleading or will lead a factfinder to an incorrect conclusion about the facts in a case, a forensic scientist should respond beyond the direct question asked to ensure the results are communicated accurately and clearly.	Defendants and persons of interest should be included among the defined “users” in section 4.1.5.
2991-2993	When publishing developmental	These passages demonstrate that as	This language should be

Lines	NISTIR 8351-DRAFT Language	Comment	Resolution
	validation results with a new STR typing kit, 2992 the goal of mixture studies is typically *to demonstrate detection of minor alleles rather than accuracy with interpreting and/or deconvoluting mixture profiles*.	currently produced, mixture studies are overly simplified and insufficient for assessing the reliability of a testing system. More attention needs to be paid to validation studies for complex samples.	retained in the final report.
3009-3011	With these developmental validation studies, rarely is more than a single two-person mixture examined with the mixture ratio being the primary variable explored.		
3412-3414	Information on DNA quantities examined, mixture ratios studied, and degree of allele sharing in these five-person mixture samples was not explicitly stated in the referenced public sources.		
3343-50	After comparing results from 15 contributing laboratories, all laboratories could only identify every minor allele in the prepared mixtures between mixture ratios of 2:1 and 1:2. They could detect ~50% minor alleles at a 9:1 ratio and ~17% at a 19:1 ratio (Krenke et al. 2002). Instrument and assay sensitivity have improved in the past two decades so it is expected that lower-level minor contributors are detectable now across multiple laboratories. This aspect has not been specifically explored in published STR typing kit developmental validation studies or DNA mixture interpretation interlaboratory studies.		
3192-3195	The factor space for DNA mixture interpretation is vast and increases significantly with more contributors (Lynch & Cotton 2018). It is therefore practically impossible to demonstrate reliability across the full extent of any factor space.	These passages are akin to setting off alarm bells on complex DNA mixture interpretations. The LRs derived from complex mixtures have not been demonstrated to be reliable because the factor space is so vast that it cannot be captured. Without established criteria for assessing reliability and when bracketing approaches and sanity checks do not solve the problem, the	These passages and Key Takeaways #4.3 and #4.4 are important to retain in the final report.
3201-3207	Based on an examination of publicly available information reviewed during		

Lines	NISTIR 8351-DRAFT Language	Comment	Resolution
	the time frame of this study, there is not enough information for the authors of this report to independently assess the degree of reliability of DNA mixture interpretation at any one point in the factor space. This is particularly true without an established and accepted criteria for reliability with complex mixtures involving contributors containing low quantities of DNA template (e.g., Benschop et al. 2015a) or where there is a high degree of allele overlap among contributors (e.g., Bright et al. 2018, Lin et al. 2020).	user must rely on "an element of trust in the DNA interpretation system used." This is a considerable ask in an adversarial system in which forensic laboratories produce DNA testing for law enforcement customers, and in an opaque system that does not freely share foundational data.	
3209-3212	A bracketing approach (discussed in Section 4.4.5) may provide a pragmatic way to infer reliability for DNA mixtures in a region of the factor space, *but will still require an element of trust in the DNA interpretation system used* since the entire factor space may not be covered with previously collected validation data		
3221-3222	However, such "sanity checks" with observed trends in LR values do not demonstrate the reliability of a specific LR number.		
868-871 and 4646-4647	KEY TAKEAWAY #5.4: DNA statistical results such as a sub-source likelihood ratio do not provide information about how or when DNA was transferred, or whether it is relevant to a case. Therefore, using the likelihood ratio as a standalone number without context can be misleading.	The vulnerability and the potential for miscommunication when using the LR as a standalone number exemplifies why transparency in criminal cases is so important. The context of a case cannot be understood without all the available information. Analysts should also keep in mind that tunnel vision or other confirmatory biases can inadvertently impact reporting and testimony. Transparency is a form of insurance to guard fulsome testimony.	This language should be retained in the final report.

INNOCENCE PROJECT

Concepts that Deserve Stronger Language or Clearer Directives

The following comments reference concepts in the report that deserve to be elevated or stated more directly and forcefully to ensure the weight of these passages is fully communicated to criminal justice stakeholders. With respect to concepts and language repeated or used multiple times throughout the report, we may reference a selection of excerpts but intend for our comments to apply globally. When new language or edits are suggested to resolve comments regarding excerpts of the report, ~~strikethroughs~~ are used to indicate text that should be deleted and **[bracketed and bold text]** indicate text that should be added.

Lines	NISTIR 8351-DRAFT Language	Comment	Resolution
1075-1083 4578-4579	Forensic laboratories have been using procedures to avoid contamination since the advent of DNA methods. However, because the likelihood of detecting contaminating DNA has increased with highly sensitive DNA methods, contamination avoidance in forensic laboratories is more important than ever. Furthermore, contamination avoidance procedures should be used during all stages of an investigation, including at the crime scene. Elimination databases that include DNA profiles of laboratory staff and police who go to crime scenes can help identify contamination and should be maintained. Therefore, relevance should be carefully assessed and considered by both the DNA analyst and users of the DNA results, especially when an evidence item contains very small amounts of DNA. KEY TAKEAWAY #5.3: Highly sensitive methods increase the likelihood of detecting contaminating DNA that might affect an investigation. Contamination avoidance procedures should be robust both at the crime scene and in the laboratory.	This recommendation should be strengthened into a takeaway. As an ethical matter, if the government can compel DNA samples from defendants for the purposes of crime investigations, then its own agents should also provide their DNA profiles for elimination databases. Elimination profiles may simplify the interpretation of a complex mixture, are important when minor donor(s) are present in small amounts, and can prevent wrongful convictions in the event a DNA mixture is inadvertently misinterpreted.	Strengthen the importance of elimination databases by integrating this concept into Key Takeaway #5.3: KEY TAKEAWAY #5.3: Highly sensitive methods increase the likelihood of detecting contaminating DNA that might affect an investigation. Contamination avoidance procedures should be robust both at the crime scene and in the laboratory. **[These procedures include the maintenance of elimination databases that include both analysts and police and improved protocols for evidence collection.]**
902-905	An overall assessment of 1) how a new technology works, 2) what its limitations are, and 3) how it might	In order to fully evaluate whether the implementation of a new technology is worthwhile, the overall assessment	These lines should be revised to state the following:

Lines	NISTIR 8351-DRAFT Language	Comment	Resolution
	specifically address the problem to be solved (e.g., DNA mixture interpretation) is important and a key component of evaluating whether implementation will be worthwhile.	must also include a fourth item —-the social impact of the technology, especially with regard to justice and equity. We should be asking ourselves what it would mean to implement a technology in an American policing system beset by structural racial biases and absence of oversight and whether there are sufficient mitigations or interventions that would ensure just and equitable implementation. Even if NIST does not believe that this fourth item is within its purview, it should reference social impact as an essential component.	An overall assessment of 1) how a new technology works, 2) what its limitations are, ~~and~~ 3) how it might specifically address the problem to be solved (e.g., DNA mixture interpretation)**[, and 4) whether this new technology can be justly and equitably implemented in an American policing system beset by structural racial biases and absence of oversight]** is important and a key component of evaluating whether implementation will be worthwhile.
1731-1742 664-666 and 1743-1744	Section 2.3.3. Mixture Complexity Increases as Number of Contributors Increase KEY TAKEAWAY #2.4: DNA mixtures vary in complexity, and the more complex the sample, the greater the uncertainty surrounding interpretation. Factors that contribute to complexity include the number of contributors, the quantity of DNA from each contributor, contributor mixture ratios, sample quality, and the degree of allele sharing.	These sections of the report recognize that the complexity of a mixture is impacted by various factors, however, it does not indicate when a sample should be deemed uninterpretable. This report does not take a position on the number of contributors to a sample that would make it uninterpretable. Previously, PCAST set a limit of interpretability at a 3-person mixture where the minor contributor was 20% of the mixture. Probabilistic genotyping software programs make varied claims with regard to the number of contributors they can deconvolute, with one vendor claiming to be able to deconvolute up to seven contributors. It would be helpful for this report to provide users with guidance on how to evaluate the reliability of those claims.	Include information on the type of data that users need to evaluate the claims of probabilistic genotyping software programs, how to use that data to do so, and the current limits of deconvolution software.
2891-2895	With higher-order DNA mixtures, the potential factor space becomes vast (e.g., consider one aspect of the factor space with possible genotyping combinations as described in Lynch & Cotton 2018).	Without clear requirements for the scope and breadth of factor spaces that need to be tested, the reliability of a forensic laboratory's testing cannot be assessed and their capabilities cannot be compared with other forensic	Key Takeaway #4.4 should be edited to state the following: KEY TAKEAWAY #4.4: Additional PGS validation

Lines	NISTIR 8351-DRAFT Language	Comment	Resolution
2915-2917	Therefore, it is unlikely that laboratories have explored every possible region of this factor space and may not be comfortable commenting on the degree of reliability with especially complex samples. It is recognized that each laboratory has to demonstrate their own degree of reliability and that we must be careful not to pool data from different sources that may come with different assumptions and caveats.	laboratories. Key Takeaway #4.4 should be modified to call for the establishment of comparable factor space criteria that should be assessed across forensic laboratories.	studies have been published since the 2016 PCAST Report. However, publicly available information continues to lack sufficient details needed to independently assess reliability of specific LR values produced in PGS systems for complex DNA mixture interpretation. **[Validation standards that articulate specific factor space testing are needed to establish comparable assessments of forensic laboratories' reliability.]** Even when a comparable reliability can be assessed (results for a two-person mixed sample are generally expected to be more reliable than those for a four-person mixed sample, for example), there is no threshold or criteria established to determine what is an acceptable level of reliability. **[These thresholds need to be defined.]**
2460-2493 2495-2584 2800-2804	3.1.2. Available Internal Laboratory Data 3.1.3. Available Proficiency Test Data and its subsections For DNA mixture interpretation, this means that samples with known genotypes, known number of contributors, known mixture ratios, known degrees of degradation, etc., have been tested using the process of measurement and interpretation, and results from such tests are available to provide the basis for stakeholders to assess the degree of reliability of the process	An insufficient number of forensic laboratories publish internal validation studies and proficiency test data. The availability of internal validation summaries as provided by Table 3.2 are insufficient as the lack of meta data makes independent evaluation impossible. Proficiency tests are woefully inadequate for fully assessing the performance of forensic analysts because they are simple and do not reflect the complexities of casework. Unless proficiency tests and their data are shared, users would not be able to assess the performance capabilities of a forensic laboratory.	Include an additional "Key Takeaway" in Chapter 3 that states: **[Transparency of laboratory data is a best practice and a professional obligation to the criminal legal system. Forensic laboratories have a professional responsibility to publish or share empirical data from their validation studies and proficiency tests in an accessible format and provide them upon request.]**

Lines	NISTIR 8351-DRAFT Language	Comment	Resolution
2909-2912	To assess reliability of any system, the factors that impact that system's performance need to be studied and evaluated. In attempting to address the question of reliability, we need to first understand what portions of the factor space have been explored and what were the experimental outcomes.	Given the overwhelming volume of content provided in the report regarding the need for transparency and how essential it is for users to interpret the reliability of a forensic laboratory's results, this section should include a "Key Takeaway" that states that transparency is not just a best practice, but a professional obligation. If data cannot be interpreted reliably without validation information and if performance of analysts can be demonstrated by proficiency tests, then there is no credible scientific or justice reason for hiding this data.	
2929-2931	A provider of information delivers this information and accompanying data in an accessible format to be used for assessment by the user. The provider also explains the relevance and significance of the information and data.		
3080-3082	If proficiency tests are representative of commonly seen casework in a forensic laboratory, then these results can also help assess what PCAST termed "validity as applied" (PCAST 2016).		
3127-3129	These CTS DNA mixture PTs involve single-source or two-person mixtures created from large quantities of DNA (hundreds to thousands of cells). In other words, the mixtures in the Forensic Biology, DNA Semen, and DNA Mixture PT exams (Table 4.6) are not complex.		
3190-3191	Demonstrating reliability requires that the provider provide empirical data that is accessible to users of the information for independent assessments of reliability.		
3061-3063	During our discussions on the topic of available data to assess PGS systems for DNA mixture interpretation performance, the DNA Resource Group (see Table 1.2) underscored that additional PGS data exists in forensic laboratories as part of their internal validation studies.		

Lines	NISTIR 8351-DRAFT Language	Comment	Resolution
3227-3230	However, to independently assess the degree of reliability of PGS models, metadata associated with specific sample results and the corresponding specific log(LR) value datapoints are needed. Data of this nature are not generally shared in publications or validation summaries.		
3367-3370	For example, many internal validation studies described in Table 4.5 do not clearly state the number of samples tested, making it difficult to assess the extent of the studies. The lack of availability of underlying data prevents independent assessments of reliability.		
3419-3424	Although more validation studies (see Tables 4.3 and 4.5) have been performed since the 2016 PCAST Report was released almost five years ago, in their present form, publicly available internal validation summaries often do not provide sufficient information to assess factor space coverage. Further, these summaries typically do not provide data points (e.g., LR values) and associated information (see Box 4.1) necessary to assess the degree of reliability and performance under potential case scenarios.		
3446-3451	Potential reasons why forensic laboratories choose not to make their internal validation data publicly available include: (1) the information from a study itself may not be publishable[23] due to lack of novelty (e.g., Buckleton 2009), (2) genotype data may include information from donors who did not consent to public sharing of their DNA profiles (e.g., Manabe et al. 2017), and (3) sharing foundational data is not required by current accreditation or guidance documents.		

Lines	NISTIR 8351-DRAFT Language	Comment	Resolution
	fn23: The willingness of journals to publish validation studies is a separate issue from the willingness of laboratories to make data available on their website for anyone to download or at least sharing full data sets with credible parties in a timely manner when requested.		
3315-3325	As described earlier in Section 4.1**.6, a determination of whether the amount and type of data available is satisfactory or sufficient to the user of the information is something that must be decided by the user of the information (e.g., the DNA analyst), not the provider (e.g., the software developer).** It is not helpful for the provider to describe a method as "validated" without providing context around the method's use and access to data to support claims of validity and reliability. Instead, it might be more appropriate to state "the following developmental validation studies have been conducted and here is the complete collection of results obtained, which can be examined by users to make reliability judgments." Internal validation studies provide an opportunity for the user (e.g., DNA analyst) to understand performance of a method in their forensic laboratory environment rather than trusting the provider's (e.g., the software developer) claim that everything works fine.	This very strong statement in the report is not integrated into a "Takeaway" or key principle. The general policy of forensic laboratories is not to provide public access to this information and Table 3.2 provides evidence of this fact. Among the approximately 400 publicly funded forensic laboratories in the United States (DuRose, 2016), the NIST report was only able to identify eight forensic laboratories or laboratory systems that post internal validation study information online. To ensure fairness, transparency, and equity among jurisdictions, forensic laboratory policies should default to full disclosure. From the strong language in this report, we would assert that data sharing extends to proprietary software as well. Reference: DuRose, M. R., Burch, A. M., Walsh, K., & Tiry, E. (2016). Publicly Funded Forensic Crime Laboratories: Resources and Services, 2014 (p. 12). U.S. Department of Justice, Office of Justice Programs, Bureau of Justice Statistics. https://www.bjs.gov/content/pub/pdf/pffclrs14.pdf	Revise Key Takeaway #4.2 to integrate the concept that data availability is based on user demand rather than provider prerogative: KEY TAKEAWAY #4.2: To enable effective use of any information, responsibilities exist with both providers and users of that information. While a provider explains the relevance and significance of the information and data, only the user can assess the degree of reliability, validity, and whether that information is fit-for-purpose. **[A determination of whether the amount and type of data available is satisfactory or sufficient to the user of the information is something that must be decided by the user of the information, not the provider. This provision applies to validation data as well as proprietary software.]**
732-755	KEY TAKEAWAY #4.1: The degree of reliability of a component or a system can be assessed using empirical data (when available) obtained through	Chapter 4 and Takeaways #4.1-4.4 discuss the critical need for and absence of empirical data. Users (defense attorneys) cannot evaluate the	Integrate language in this report to acknowledge that transparency of data is a professional obligation of all

Lines	NISTIR 8351-DRAFT Language	Comment	Resolution
	validation studies, interlaboratory studies, and proficiency tests. KEY TAKEAWAY #4.2: To enable effective use of any information, responsibilities exist with both providers and users of that information. While a provider explains the relevance and significance of the information and data, only the user can assess the degree of reliability, validity, and whether that information is fit-for-purpose. 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. KEY TAKEAWAY #4.4: Additional PGS validation studies have been published since the 2016 PCAST Report. However, publicly available information continues to lack sufficient details needed to independently assess reliability of specific LR values produced in PGS systems for complex DNA mixture interpretation. Even when a comparable reliability can be assessed (results for a two-person mixed sample are generally expected to be more reliable than those for a four-person mixed sample, for example), there is no threshold or criteria established to determine what is an acceptable level of	reliability of a system without data. For this reason, validation studies must be made public. This report stops short of calling for forensic laboratories to publish this data. While NIST may avoid mandates or the use of "shall" or "must" language, this report takes great pains to describe the scientific cost of not making data public. NIST should establish transparency as a professional obligation to crystallize this responsibility for providers and users.	public forensic laboratory scientists and customers of public forensic laboratories. These stakeholders have an ethical obligation to eliminate barriers to data transparency.

Lines	NISTIR 8351-DRAFT Language	Comment	Resolution
	reliability.		
769-773 and 3487-34888	KEY TAKEAWAY #4.7: The degree of reliability of a PGS system when interpreting a DNA mixture can be judged based on validation studies using known samples that are similar in complexity to the sample in the case. To enable users of results to assess the degree of reliability in the case of interest, it would be helpful to include these validation performance results in the case file and report.	Takeaway 4.7 represents critical information needed to assess the conclusions reached by the forensic laboratory. Validation data is not typically provided in the case file or the report and should become a regular practice as a matter of scientific principles, transparency, and justice. In *United States v. Gissantaner*, the trial court correctly identified that the Michigan State Police was conducting analyses outside the limits of what its DNA laboratory had appropriately validated. In *People v. Collins-Peaks*, the court excluded high sensitivity DNA testing and the NYC Office of Chief Medical Examiner's (OCME) Forensic Statistical Tool after a Frye hearing where the validation data for each method was assessed. However, it was not until the hearing that validation data was shared by the OCME. This language needs to be strengthened in proportion to the importance of validation studies on the assessment of the data. Validation studies are not "helpful," for interpretation, they are "essential" for evaluating the reliability of the results.	This takeaway should be edited to the following text: KEY TAKEAWAY #4.7: The degree of reliability of a PGS system when interpreting a DNA mixture can be judged based on validation studies using known samples that are similar in complexity to the sample in the case. **[Validation studies are essential to]** ~~To enable~~ users of results to assess the degree of reliability in the case of interest**[.]** ~~, it would be helpful to~~ **[Forensic laboratories that seek to fully and scientifically communicate their results will]** include these validation performance results in the case file and report.
3047-3048	*Table 4.3* includes a list of published validation data from peer-reviewed literature	The 2016 President's Council of Advisors on Science and Technology's report on forensic science raised the concern that published validation studies for probabilistic genotyping software were primarily authored by software developers and expressed the need for groups independent of these developers to publish such studies. Most of the publications referenced in Table 4.3 are published by software developers. Does NIST share PCAST's concern? <u>Reference:</u> President's Council of Advisors on	Include a discussion about the disproportionate availability of validation data from software developers compared to independent researchers or laboratories and how this may impact the body of knowledge in DNA mixture scholarship.

Lines	NISTIR 8351-DRAFT Language	Comment	Resolution
		Science and Technology. (2016). Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods. President's Council of Advisors on Science and Technology. https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/PCAST/pcast_forensic_science_report_final.pdf	
913-915 and 5344-5345	KEY TAKEAWAY #6.2: Implementation requires a thorough understanding of the benefits and limitations of the new technology as well as the practical investment of time and effort put forth for its adoption by the laboratory.	Implementation of a forensic technology does not take place in siloes, but rather in a criminal legal system beset by adversarial tensions, structural racism, and deep disparities in how communities of color are policed and criminalized. The social implications of a technology and how it will be implemented within the criminal process needs to be understood as well. Simply ensuring validity, reliability, and application within limits does not mean that a technology should be unleashed freely.	Revised Key Takeaway #6.2 to reflect the need to understand the social impact of technologies: KEY TAKEAWAY #6.2: Implementation requires a thorough understanding of the benefits**[,]** ~~and~~ limitations**[, and social implications]** of the new technology as well as the practical investment of time and effort put forth for its adoption by the laboratory **[and the American policing system in which it will be implemented]**.

Report Components that may be Misapplied in the Justice Process

The following comments reference specific passages in the report that are vulnerable to misinterpretation and can potentially be manipulated to serve a purpose beyond what we presume to be the authors' intent. When new language or edits are suggested as resolutions to comments regarding excerpts of the report, ~~strikethroughs~~ will be used to indicate text that should be deleted and **[bracketed and bold text]** will indicate text that should be added.

Lines	NISTIR 8351-DRAFT Language	Comment	Resolution
531-532	The findings described in this report are meant solely to inform future work in the field.	This report identifies critical best practices that forensic science service providers may not be following. These lines undermine the recommendations of	This sentence should be removed and replaced with the following text:

Lines	NISTIR 8351-DRAFT Language	Comment	Resolution
		the report, the ethical and legal obligations of forensic science service providers, and excuses forensic laboratories from taking responsibility and being held accountable for their past work. Not only do these lines appear to permit unacceptable work product that may exist, they dismiss the obligation of forensic laboratories to take action to remedy the problems as mandated by quality management principles. This line also offers attorneys and courts an argument against correcting mistakes or errors that impact the criminal legal system. As referenced in lines (513-514), this language may influence "critical decisions impacting life and liberty."	**[The findings in this report offer forensic laboratories an opportunity to reflect upon their current policies and protocols. Forensic laboratories should be reminded of their ethical and professional obligations when their policies and protocols fall short of best practices. Please refer to item #15 and #16 in the National Commission on Forensic Science's National Code of Ethics and Professional Responsibility for the Forensic Sciences.]**
1911-1914 1953-1956	That is, different people may assign different values to the same evidence. Concerns have been raised that the LR framework applies only to personal decision making and cannot automatically be used for the transfer of information from one expert to a separate decision maker (Lund & Iyer 2017). A common problem known as "transposing the conditional" (Evett 1995) or committing the "prosecutor's fallacy" (Thompson & Schumann 1987) can lead to a misunderstanding of the meaning of an LR result. In these situations, a user confuses "the probability of the evidence given the propositions" with "the probability of the propositions given the evidence."	These important passages help us understand that Bayesian frameworks and the use of LRs can have varying degrees of subjectivity or provide incomplete analyses of a phenomena. The difficulties of communicating the Bayesian framework and transposing the conditional is also described using the example of a cow and its four legs (Lines 1961-1964). Notably, however, Bess Stiffelman has written about how the Bayesian framework can violate presumptions of innocence in the forensic setting. Not only are results only given in degrees of inclusion, thereby eliminating the ability to exclude a person of interest, but one must assume that the prior probability of a defendant contributing to a DNA sample is greater than zero. Therefore, any presentation of results assumes a level of guilt and this is especially true when only LRs are presented. It is not clear how one can mitigate this problem given the legal and	Include a discussion of how the Bayesian framework impacts the presumption of innocence that include the following resources: Stiffelman, B. (2109). No Longer the Gold Standard: Probabilistic Genotyping is Changing the Nature of DNA Evidence in Criminal Trials. https://doi.org/10.15779/Z384Q7QQ6X Presser, J. R., & Robertson, K. (2021). AI Case Study: Probabilistic Genotyping DNA Tools in Canadian Criminal Courts. Law Commission of Ontario.

Lines	NISTIR 8351-DRAFT Language	Comment	Resolution
		mathematical complications of presenting a prior probability, but this complex issue should be raised in the report.	
2087-2089 2206-2215	Another possible source of variation in LR results comprises the estimated degree of co-ancestry in observed alleles, which involves using a subpopulation correction factor typically symbolized by the Greek letter theta (Balding & Nichols 1994, NRC 1996). Principle 6 [Genetics]: DNA profiles from close relatives are more similar than DNA from unrelated people DNA profiles from close relatives are expected to be more similar than DNA profiles from unrelated individuals (Li et al. 1993). There are a limited number of alleles at each locus, and even individuals who are not closely related will share alleles and genotypes. The frequency of occurrence of specific alleles and genotypes varies. *This principle is a reminder that while statistical models typically assume individuals are unrelated, if case context suggests closely related individuals may have contributed to the sample in question, then performing calculations assuming individuals are related may be helpful to decision makers.*	The problem of relatedness and the potential impact of this problem deserves more discussion. Forensic laboratories should account for relatedness in their testing protocols. Principle 6 should also explicitly discuss the real risk of falsely including a non-contributor relative in a DNA mixture.	Expand Principle 6 by adding the following text to (Line 2215): **[Not accounting for relatedness can increase the risk of falsely including a non-contributor relative in the DNA mixture.]**
3264-3270	We, the authors of this NIST report, emphasize that publicly available data from validation studies, whether or not this information has been published in a peer-reviewed journal, enable a user (e.g., the DNA analyst when the provider is the PGS developer or the court when the analyst is providing their results) to scrutinize the underlying data and supporting details for what is currently possible in research	The report goes into great detail regarding the inadequate scope and breadth of DNA mixture validation studies. This passage may be misused by stakeholders to erase the concerns raised by Key Takeaways #4.3 and #4.4. Taken out of context, the language can appear to convey that the state of DNA mixture interpretation is sound. This statement should be revised to more carefully reflect the intent of the authors' position and reduce its potential for manipulation	Revise this passage to the following: We, the authors of this NIST report, emphasize that publicly available data from validation studies, whether or not this information has been published in a peer-reviewed journal, **[can be used by]** ~~enable~~ a user

Lines	NISTIR 8351-DRAFT Language	Comment	Resolution
	settings (what PCAST terms "scientific or foundational validity") and what is actually happening in casework settings (what PCAST calls "validity as applied").	when quoted out of context.	(e.g., the DNA analyst when the provider is the PGS developer or the court when the analyst is providing their results) to scrutinize the underlying data and supporting details for what is currently possible in research settings (what PCAST terms "scientific or foundational validity") and **[ascertain whether]** what is actually happening in casework settings (what PCAST calls "validity as applied") **[reflects the claims of the validation data].**

PC46

Comment on DNA Mixture Interpretation: A Scientific Foundation Review

tiffany roy [redacted]

Mon 8/23/2021 6:09 PM

To: ScientificFoundationReviews <ScientificFoundationReviews@nist.gov>

Dr. Butler,

I have reviewed the document *DNA Mixture Interpretation: A Scientific Foundation Review* and my comments are attached.

PC46a

Dr. Butler,
I have reviewed the document *DNA Mixture Interpretation: A Scientific Foundation Review* and I want to congratulate the team at NIST on their hard work. It is extremely useful to have the literature compiled and organized. I have colleagues that have already used the document a resource.

In my work, I have reviewed DNA cases from laboratories across the country. What I have found is that most forensic DNA analysts do not have a research background. They seem lost and confused when it comes to validation, which is why there is a need for extensive guidance from ASCLD LAB, SWGDAM, ANAB, OSAC and other guidance bodies on validation requirements and construction. Forensic scientists seem even more confused when they need to draw conclusions from the data once it is generated, though this is required to create sound protocols. This is particularly exacerbated by outsourcing validations and data analyses rather than ensuring lab analysts understand the data and identify trends. For me, the most important point in the paper--one which I feel the community is missing--is the difference between data and interpretation. Essentially, the Foundation Review calls for greater transparency and to make more *data* (as opposed to published opinions/interpretations of the data) open for public inspection. Not just inspection those who created the data deem qualified. Not just forensic scientist to forensic scientist inspection. But inspection from other fields of science, other interested parties, and criminal justice stakeholders. The suggestion we should make large databases containing pooled data, developmental validation data and internal validation data *publicly* and in a standardized format has been the most controversial aspect of this document. More published studies are not going to help address this. Throwing a bunch of unidentified data into the public sphere is not going to address this. That is performative. Making the data that supports your conclusion available to anyone and understandable to everyone is the only way to address this.

I have been surprised at some of the industry reaction to that recommendation. I assume it was similar to the NAS and NCFS recommendations that laboratory Standard Operating Procedures and Quality Documents publicly accessible for anyone who would wish to inspect them.

"Why is that necessary?"
"Qualified analysts inspect our data every year in audits."
"That data can be compelled by court process."

To that I say, why not? What is everyone so afraid of? What possible reason could a scientist provide for being less transparent about their work product? These lines of reasoning were rejected by the National Commission when they recommended quality documents and procedures be made public. They should be rejected now. I work a great deal in Florida, which allows the death penalty. It is used regularly. What possible privacy interest could forensic DNA analysts have in their work product that is more important than human life?

For me, foundational validity is about more than just whether DNA mixtures CAN be reliably interpreted, but whether they are being reliably interpreted at labs across the country and around the world in any consistent way. NIST was clear from the start that they would be looking at the underlying research and the methods *as applied*. As the FR suggests, proficiency tests and interlaboratory mixture studies will be required to do that. All the published research in the world will not tell us whether people are getting it right. And whether they are being appropriately described in court to the "users" of this information in the most serious context. To quote from the report,

"We would add that the overall reliability of a method or practice is influenced by many things. Samples examined in forensic science practice vary in quantity, quality, and complexity. Use of a "foundationally valid" method is insufficient to establish trustworthiness without knowing how that method is used in practice under a specific case situation."

The only way to assess that is to assess whether labs are drawing sound conclusions from their data. The only way to do that is to have the data scrutinized by others to see if they can recreate the conclusions drawn. It makes it very hard to look retrospectively at the existing body of research and feel confident in the published conclusions which have been the bedrock of practice. We've essentially been trusting the authors of forensic genetic research that their stated conclusions are supported by data, regardless of association and conflicts of interest. Taking their word for it. Are the opinions stated in the research papers listed this review a reflection of the actual data? Or are we only seeing the trends the authors saw or wanted us to see?

Watching the field respond to this review demonstrates some of the complex issues that surround any constructive criticism the field receives. No one wants to think they may be doing something improper.

"Whose fault is this?"
"Are you saying the technical leaders are not doing a good enough job designing these validations?"

We are all to blame. We, as a community, have not demanded the type of transparency we owe to the seriousness of this work. In our publication process. In our laboratory validation data. In the development of our quality documents and our procedures. If we fail to demand that transparency going forward, we are all complicit in the problems that undoubtedly exist in this field.

Commentary suggests that some already want to reject the authors of this review as academics; research scientists with no actual DNA testing experience or any such "other" labels as if they are relevant to the work done and message being delivered. To quote from the abstract,

"The National Institute of Standards and Technology (NIST) is a scientific research agency that works to advance measurement science, standards, and technology and that has been working to strengthen forensic science methods for almost a century. **In recent years, several scientific advisory bodies [1-3] have expressed the need for scientific foundation reviews of forensic disciplines and identified NIST as an appropriate agency for conducting them."**

The scientific advisory bodies referenced are the National Research Council, the National Commission on Forensic Science and the President's Council of Advisors on Science and Technology. Distinguished members of these bodies formally identify NIST as an appropriate agency to conduct these scientific foundation reviews. Countless DNA labs across this country list your book(s) as required reading for new DNA analysts. It will be interesting to see how many of these same laboratory analysts now argue you are somehow now unqualified to perform this foundation review because you have never hit the "go" button on an extraction robot in a crime lab. It will be interesting to see how many members of the National Commission now feel that team members at NIST are no longer qualified to perform a foundation review, even after their formal recommendation.

What the forensic DNA community observed in interlaboratory studies Mix 05 and Mix 13 was an indication that labs are not on the same page. How many? What is the true extend of the problem? This information is still unknown. And to be sure, a problem was identified but never investigated, described, and resolved. Forensic DNA analysts should be operating in such a way that it is easy for others to see what actions are performed. Science demands that we operate with openness, communication, and accountability. Those in this field who oppose efforts toward this end weaken it, undermine it, and delegitimize it.

PC47

OSAC Human Forensic Biology comments to NIST Scientific Foundation Review on DNA Mixture Interpretation

Ragsdale, Robyn [redacted]

Mon 8/23/2021 9:31 PM

To: ScientificFoundationReviews <ScientificFoundationReviews@nist.gov>
Cc: Ordeman, Beth [redacted]

Please find attached the comments of the Human Forensic Biology Subcommittee of the Biology Scientific Area Committee of the OSAC. We respect the immense amount of hard work and time tthat went into creating this document and appreciate the opportunity to comment.

Sincerely,

Robyn Ragsdale, Ph.D.
Chair of the Biology Scientific Area Committee

PC47a

Human Forensic Biology

OSAC has a mission to strengthen forensic science through the development and promotion of the use of high quality technically sound standards and best practices for forensic laboratories. With regard to this mission, there are validation, reporting, testimony, and research needs related issues in the NISTIR 8351 draft report on DNA Mixture Interpretation that we feel should be addressed as identified by the OSAC Human Forensic Biology Subcommittee. We have outlined the main topics of concern below.

Chapter 2

1. Exhaustive propositions

Principle 16 (lines 2350-2354) states "Assessing the strength of evidence in favor a [sic] proposition (hypothesis) H1 requires at least one other proposition (hypothesis) H2. These propositions H1 and H2 are required to be mutually exclusive and exhaustive. Strength of evidence assessments depend on the framework of circumstances within which they are evaluated." Lines 3572-3575, appear to cite Gittelson et al. 2018 out of context: "...as it has been noted: "The truth lies in the propositions: either the prosecution proposition is true or the [defense] proposition is true" (Gittelson et al. 2018). The implicit assumption in this statement is that the propositions are exhaustive."

The assertion that propositions need to be exhaustive is incorrect. These propositions do not need to be exhaustive. They are more useful when they represent each of the competing views of the parties, i.e. "exhaustive within the context of the case", but this is not the same as "being exhaustive". This is explained in the literature, as well as in numerous guidance documents.

Following currently available guidance documents and literature, the OSAC-developed standard pending at ASB (ASB Standard 041) Assigning Propositions for Likelihood Ratios and a best practice document currently in development in the OSAC Human Forensic Biology Subcommittee regarding evaluative Forensic DNA testimony explain that the propositions do not need to be exhaustive.

Chapter 4

1. Bracketing and factor space related to validation

The attempt in this report to propose validation studies through the concepts of "factor space (coverage)" and "bracketing" are confusing and potentially misleading.

Section 4.1.4. seems to suggest validation experiments that test the entire factor space, which is potentially misleading. This idea is revoked later in Chapter 4, e.g., in lines 3465-3466: "It is unrealistic to obtain and examine the volume of samples needed in order to provide complete coverage of the potential factor space with DNA mixture interpretation." Instead, the authors propose "bracketing",

which is described as "consider(ing) results from samples that are both more complex or less complex than the casework sample of interest as a pragmatic way of understanding case-specific reliability of an interpretation system." If the authors are suggesting "case-type profiles (…) that represent (in terms of number of contributors, mixture ratios, and total DNA template quantities) the range of scenarios that would likely be encountered in casework" that "include compromised DNA samples (e.g., low template, degraded, and inhibited samples)", this is already covered in SWGDAM documents as well as ANSI/ASB Standard 020, Standard for Validation Studies of DNA Mixtures, and Development and Verification of a Laboratory's Mixture Interpretation Protocol, First Edition, 2018 and in ANSI/ASB Standard 018, Standard for Validation of Probabilistic Genotyping Systems, First Edition, 2020 (both published and on the OSAC Registry). If this is not the intent, the actual intent and reasoning as well as specific gaps that are not covered by the aforementioned documents should be specifically addressed.

2. ROC plots (i.e. ROC curves)

ROC plots are mentioned as an acceptable way to assess the reliability of PGS. In this document, the authors write that ROC plots "have been used in evaluation of PGS systems previously (e.g., Bleka et al. 2016b, You & Balding 2019)" (lines 3722-3723). This statement is confusing, as both of the cited studies use ROC plots only for the comparison of models on a specific dataset, and not for the evaluation of a PGS system per se. We were unable to find any study where a PGS system was validated through the presentation of ROC plots. ROC plots can be useful (e.g., for the comparison of different models all run on the same dataset), but the literature is full of alternative methods, such as violin plots, LOWESS plots, Hd-True testing, and more. However, with such stress on using ROC plots to test the reliability of PGS, it would seem counterintuitive to rely on the output of the PGS system being tested to generate the ROC plots.

ROC plots present many limitations. Currently, there are no SWGDAM standards, no OSAC-developed standards, or as far as we are aware, any other documents that specifically call for the use of ROC plots in PGS validation studies. There are a variety of valid statistical approaches and tools available for laboratories to use in order to understand the limits of PGS systems. A foundational review should discuss other reliable and valuable options found in the literature.

3. Independent assessment and data availability

Key Takeaway #4.3 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." Upwards of 60 published papers and studies are cited in this report, encompassing 1000's of samples and many millions of probabilistic comparisons. It would be helpful to provide information as to how much data would be enough. Additionally, along with all the cited publications, SWGDAM guidelines for the Validation of Probabilistic Genotyping Systems as well as ANSI/ASB Standard 018, Standard for Validation of Probabilistic Genotyping Systems, First Edition, 2020 have been published. If followed, these promote reliable DNA mixture interpretation practices and would also be assessed during accreditation.

The second part of this Key Takeaway #4.3 states "….we encourage forensic laboratories to make their underlying PGS validation data publicly available….". Traditionally, journals have not been receptive to the publication of multiple internal validation studies. This NIST report suggests addressing this issue by stating laboratories need to be willing to let anyone access their data, for example, from a website. Web hosting and curation could represent other resource limitations (budgetary and personnel) that many,

particularly smaller laboratories, cannot accommodate. There are also ethical concerns related to the publication of DNA profiles. Most validation samples are collected from volunteers giving informed consent, possibly following Institutional Review Board review, approval and documentation. There is no expectation that such volunteers would agree en masse to have their DNA profiles publicly available on the internet. Government collected samples must remain private according to the Privacy Act of 1974 (5, U.S.C. 522a). Placing validation samples on the internet disregards ethical considerations of genetic privacy.

4. Validation performance results in the case file and report

Key Takeaway #4.7 states, "… To enable users of results to assess the degree of reliability in the case of interest, it would be helpful to include these validation performance results in the case file and report."

We acknowledge that no current standard, guideline or best practice document from any organization or accrediting body suggests this but it is unclear if this means to include a validation summary or all of the validation data. In practical terms, providing gigabytes of data along with the report, is not practical. The legal mechanism of discovery is put in place for exactly this and there are numerous ethical issues as stated above.

The authors should instead consider suggesting the development of a best practice document addressing how to provide a reasonable summary of validation studies and limitations that could be provided in the case record or published with the report (i.e. such as an appendix to a lab report). Not only is this more realistic, but it would also provide a framework that could be followed by all forensic DNA laboratories.

5. Reliability

Key Takeaway #4.4 states, "Additional PGS validation studies have been published since the 2016 PCAST Report. However, publicly available information continues to lack sufficient details needed to independently assess reliability of specific LR values produced in PGS systems for complex DNA mixture interpretation. Even when a comparable reliability can be assessed (results for a two-person mixed sample are generally expected to be more reliable than those for a four-person mixed sample, for example), there is no threshold or criteria established to determine what is an acceptable level of reliability."

Key Takeaway #4.4 focuses on the difference in reliability of simple mixtures and more complex mixtures. However, the report does not mention that the uncertainty of such samples is accounted for in the magnitude of the LR. For example, in simple two-person mixtures, LRs are commonly of magnitudes outside the scope of the understanding of a lay person. For difficult 4-person mixtures with degraded DNA and drop-out issues, the LR can be in the hundreds or tens. The difference is clear in the magnitude of the LR.

Reliability is a two-way street. DNA has exonerated hundreds of falsely incarcerated individuals and perhaps millions of falsely accused persons. It is unclear if this report means to infer that those exonerations should not have occurred if anyone cannot independently assess the reliability of mixtures both more and less complex for the factor space of difficult mixtures. Numerous organizations have put forth guidelines and standards dealing with validation and mixture interpretation for forensic DNA crime laboratories. These documents cover what is required and laboratories following these are assessed as

such and are required to meet an acceptable level of reliability for accreditation. A foundational review raising questions about the reliability of mixture interpretation validations and PGS should reference specific gaps and expected outcomes not covered by current guidelines and standards It is ineffective to state that there is insufficient data to support reliability without defining the criteria necessary to make that determination.

Chapter 5

The majority of Chapter 5 relates to elements beyond mixture interpretation (such as the transfer of DNA and the hierarchy of propositions) and it is suggested that these items be removed from this document and considered for an additional document. It may be better to create a separate foundational review for source and activity level propositions as well as the transfer of DNA.

Although the hierarchy of propositions, Case Assessment and Interpretation (CAI) approach and Bayesian Networks (BN) are mentioned in this report, portions of Chapter 5 are in direct conflict with a document that is currently being drafted by the OSAC Human Forensic Biology Subcommittee regarding Best Practice Recommendations for Evaluative Forensic DNA Testimony which is in line with current published literature on this topic.

1. Hierarchy of propositions

Chapter 5 should draw a boundary between an evaluation for activity level propositions and an evaluation for offense level propositions.

This report appears to blend relevance, transfer, persistence and background DNA all together, making it unclear to know which discussions refer to activity level evaluations and what discussions refer to offense level evaluations. A clear, unambiguous distinction is crucial, because forensic scientists do not evaluate DNA results with regard to offense level propositions as the evaluation at the offense level is for the trier of fact alone.

The statement made at line 810: "Relevance should be assessed. If not, the evidence can be misleading." is an example of contradicting the individual levels of the hierarchy of propositions. This is outside the scope of the forensic scientist/expert witness. The judge is the gatekeeper for the admission of evidence at court, and the jury is the only one that determines relevance. We suggest re-writing this chapter specifically to remove discussion of "relevance" and replace it with clear discussion focusing on the activity level of the hierarchy.

2. Investigative vs. Evaluative mode

We recommend adding discussion related to the differences between the investigative role of the forensic expert and the evaluative role. All examples and concerns seem to address the investigative mode but this is not clear in this document.

Page 7, lines 792, 803, 806 use the terms "readily", "might have been", and "might pick up". These are all phrases that may be somewhat useful when a forensic expert is engaging with a case investigator, but we feel are inappropriate for the evaluative role of the expert at court. There is no discussion of this in the report.

In Key Takeaway #5.5, the transfer rate is not addressed in the statement "transfer easily between objects". Without discussion of the evaluative role of the forensics expert, there is a very real danger that investigative terms such as "easy transfer" will be used in court. As transfer depends on numerous things (case-specific and otherwise), we feel this is an overstatement and is a research gap (i.e. under what specific case information and propositions can one base the assessment of "easy transfer"? What is the value of an appropriate alternate proposition?) rather than a given occurrence. Additionally, Key Takeaway #5.5, is an example of a one-sided DNA approach. Although this Chapter discusses the CAI Approach, (P 140, line 4911) there should be clear examples of a balanced approach where the evidence is considered under competing propositions. See further discussion below.

3. The danger of the transposed conditional in evaluative mode

There are sections of Chapter 5 that run the risk of being incorrectly incorporated into an evaluative role at court instead of investigative. An example is found in the Executive Summary on Page 7 of the full report, Lines 802-804: "First, that DNA might have been deposited before or after the crime was committed and therefore may not be relevant to the crime."

There is no mention of investigative vs. evaluative mode related to this statement. If such a statement is made at court, this is an example of the transposed conditional; it is an evaluation of the following propositions: P1 = "before the crime" and P2 "after the crime". There is no evaluation of the DNA evidence "if" or "given" those propositions. Evaluations of the propositions like this are for the trier of fact.

4. Formulation of activity level propositions

Activity level propositions stated throughout Chapter 5 appear to be in conflict with the ISFG Guidelines published in 2020 (Gill et al., (2020), '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 Sci Int Genet 44, 102186) and with current OSAC documents in production (which are based on guidelines from ISFG, ENFSI, and others). On page 6 of the ISFG 2020 Guidelines, Recommendation 6 states: "Results or factors that scientists take into account in their evaluation should not be interwoven into the propositions. The scientist should avoid the use of the term 'transfer' in propositions. Instead, there should be a focus on the alleged activities." The following statements in this report are examples that appear to violate this recommendation:

> "For example, in the case of a stabbing, the prosecution hypothesis might be that the DNA was transferred to the handle of a knife during the activity of stabbing, while the defense hypothesis might be that the DNA was deposited due to contamination or secondary transfer." (page 136, lines 4727-4730)
>
> "An activity proposition might be, for instance, that DNA collected during a sexual assault examination was deposited during sexual activity, or that DNA found on the handle of a knife was deposited during the act of stabbing a victim." (page 135, lines 4691-4693)

The correct formulation of activity level propositions for the former case would be "The POI stabbed the victim with this knife" for the prosecution's proposition, and "The POI did not touch this knife" for the defense proposition. The correct formulation of activity level propositions for the latter statement

would be “that sexual activity occurred between the POI and the complainant, or that the POI stabbed the victim with this knife”.

5. Significant knowledge gaps

Key Takeaway #5.6 states, “There is a growing body of knowledge about DNA transfer and persistence, but significant knowledge gaps remain.” Unfortunately, it fails to give any specific details about where the significant knowledge gaps lie. One of the tasks of OSAC consists of identifying specific research needs. Without specific details, it is left to the reader to guess exactly what they are.

Conclusion:

In general terms, this NIST report reads more like a Standard or Best Practice document rather than a Foundational Review. We feel that OSAC would be able to incorporate the gaps established from a foundational review into best practice documents, standards and research needs to better serve the Forensic DNA community if the report was more specific about the validation requirements, research needs and knowledge gaps that exist. Finally, we believe a separate foundational review should be done for source and activity level propositions as well as the transfer of DNA in lieu of inclusion in this document.

Comment

Murphy, Erin [redacted]
Mon 8/23/2021 10:05 PM

To: ScientificFoundationReviews <ScientificFoundationReviews@nist.gov>

Please accept the attached comment on NISTIR 8351-DRAFT—*DNA Mixture Interpretation: A NIST Scientific Foundation Review*.

Thank you.

Erin Murphy
Norman Dorsen Professor of Civil Liberties
NYU School of Law
40 Washington Square South
New York, NY 10012
[redacted]

PC48

August 23, 2021

Via Electronic Mail

National Institute of Standards and Technology
Attn: Special Programs Office – Scientific Foundation Review
U.S. Department of Commerce
100 Bureau Drive Stop 4701
Gaithersburg, Maryland 20899-4701
scientificfoundationreviews@nist.gov

Re: Request for Comment on NISTIR 8351-DRAFT—*DNA Mixture Interpretation: A NIST Scientific Foundation Review*.

Dear NIST Scientific Foundation Review Team:

As scientists and legal scholars who study the use and impact of forensic science evidence in criminal cases and litigators specializing in forensic evidence, we write in support of the *DNA Mixture Interpretation: A NIST Scientific Foundation Review* draft, NISTIR 8351-DRAFT (hereinafter, "Draft Report"), published on June 9, 2021.

The National Research Council first called for a project of scientific foundation reviews in 2009.[1] In the years that followed, numerous additional scientific bodies have repeated that call, including PCAST[2] and the NCFS.[3] Consistent with fundamental scientific principles, foundation reviews survey available data to evaluate empirical support for a method and its applications. In a mission critical system like forensics, scientific integrity is enhanced by such projects.[4] And, amongst the available options, NIST is an appropriate entity to conduct such reviews.[5]

Not only is scientific integrity enhanced by the scientific foundation review project in general, but NIST also appropriately selected DNA mixtures as an area of focus for review. Requests for forensic DNA analysis are second only to controlled substance testing requests, and number more than 250,000 analyses per year across the country.[6] As the Draft Report appropriately

---

[1] National Research Council, *Strengthening Forensic Science in the United States: A Path Forward* (2009), https://www.ojp.gov/pdffiles1/nij/grants/228091.pdf [hereinafter "NRC, *A Path Forward*"].
[2] President's Council of Advisors on Science & Technology, *Report to the President: Forensic Science in Criminal Cases: Ensuring Scientific Validity of Feature-Comparison Methods* (2016), https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/PCAST/pcast_forensic_science_report_final.pdf.
[3] National Commission on Forensic Science, *Recommendation to the Attorney General: Technical Merit Evaluation of Forensic Science Method and Practices* (2016), https://www.justice.gov/archives/ncfs/page/file/905541/download.
[4] *See, e.g.*, Memorandum of Understanding (MOU) between NIST and the Department of Justice (DOJ) (2015), https://www.justice.gov/archives/ncfs/file/761051/download ("Scientifically valid and accurate forensic science strengthens all aspects of our justice system.").
[5] NCFS, *Recommendations* at https://www.justice.gov/archives/ncfs/page/file/905541/download.
[6] United States Government Accountability Office, Report to Congressional Requestors: DNA Evidence (March 2019), https://www.gao.gov/assets/gao-19-216.pdf.

notes, the world of the NRC Report's gold standard single-source and simple-mixture DNA analysis has now been stretched far beyond its original bounds. Modern testing has ventured into territory involving drastically increased sensitivity. Today, "[t]he use of expert testimony in American trials is widespread and accelerating,"[7] and, when it comes to DNA, that testimony is routinely reporting complex DNA mixture analysis.

The *NIST Scientific Foundation Reviews*[8] laid out a clear plan for conducting scientific foundation reviews. The Draft Report follows that plan while incorporating discipline-specific nuance. The Draft Report addresses each of its planned reporting areas. *NIST Foundation Reviews* at 1.3. Further, we agree with the non-controversial premise of NIST's evaluation criteria: Retrievability, Reliability, and Respectability. Transparency and openness are, indeed, the hallmarks of good science.[9]

The Draft Report effectively addresses the question of DNA mixture analysis's system reliability. The Draft Report's focus on "system reliability" is a critical reframing of a nuanced point: when analyzing the overall reliability of mission-critical systems—like the DNA mixture analysis process—review of the reliability of its components is necessary, but not itself sufficient. Thus it is no answer to NIST's focus on system reliability to point to peer-reviewed publications addressing solely the software component of the process, as some would claim. And the PCAST Report's emphasis on peer review is not to the contrary. The two reports address entirely different questions.

The Draft Report addresses "factor space coverage" i.e. exploration of the complexities that affect DNA mixture interpretation—during validation testing. The report appropriately connects the breadth of factor space coverage during testing with reliability limitations.

The Draft Report also makes an important contribution by recognizing that issues of DNA transfer and mixture interpretation are inextricably intertwined, and as such "the possibility of transfer cannot be ignored when interpreting DNA evidence" at the risk of misleading fact finders (p. 129). It is therefore appropriate that the report devotes a section to the scientific knowledge base (and limitations of the same) related to issues of DNA transfer. The report could be further strengthened by connecting principles articulated in other sections -- e.g. the importance of robust validation, interlaboratory, and proficiency test data to assessing reliability of a methodology (p. 62) -- to activity-level analyses. Activity-level analyses – where the analyst goes beyond conceding the possibility of DNA transfer to offering an opinion comparing the likelihood of the DNA results under opposing scenarios of DNA deposition -- are a separate form of expertise from sub-source level analyses. As such, they require targeted training, as well as competency and ongoing (and sufficiently rigorous) proficiency testing. *See, e.g.,* van Oorschot et al., Need for

---

[7] Edward J. Imwinkelreid, *The Admissibility of Scientific Evidence: Exploring the Significance of the Distinction between Foundational Validity and Validity as Applied,* 70 Syracuse L. Rev. 817, 817 (2020).
[8] NIST, *NIST Scientific Foundation Reviews*, NISTIR 8225 (Dec. 2020), https://nvlpubs.nist.gov/nistpubs/ir/2020/NIST.IR.8225.pdf.
[9] *Id*. at 1.2.

dedicated training, competency assessment, authorizations and ongoing proficiency testing for those addressing DNA transfer issues (2017). *Cf.* NIST draft report, p. 144, Table 6.1.

The current system of forensic science in the United States simply does not have the structures in place to ensure that any given analyst is truly "expert" in this realm, and recognizing this gap is crucial to avoiding erroneous or overstated testimony. Further, activity-level analyses are unquestionably methods, which like any other scientific method, need meaningful standards against which to assess their quality; clear and detailed SOPs guiding their application; and robust validation testing to establish the reliability of activity level conclusions. Unfortunately, there is currently a dearth of all of these, and as a consequence, there are enormous as well as undefined levels of uncertainty associated with any activity-level conclusion. This uncertainty is exacerbated when analysts import subjectivity into their analyses -- whether in selection of transfer studies for consideration, or the use of "subjective probabilities" (NIST draft report p. 137). While there is extremely little in the way of black box validation data, what literature does exist suggests that "expert" opinions in this domain have very high rates of error. van Oorschot, supra. This report's coverage of activity level analyses would be more complete if it made these varied foundational gaps clear.

Another concern with the Draft Report is, as one commenter has already noted, the following sentence from the executive summary could be used to undermine the entire report: "The findings described in this report are meant solely to inform future work in the field." An example from litigation following the publication of NRC, *A Path Forward* is illustrative. In that instance litigants used similar language to argue that the substance of the NRC report should not inform decisions by courts or others about the admissibility of evidence. Suggesting that this body of work is only for use by forensic DNA practitioners, and, by implication, therefore should not be considered by those outside of the field who rely upon, make decisions based on, and must evaluate what weight to give forensic DNA results simply does not make any sense. As the Honorable Harry T. Edwards[10], stated in a speech following the review of pleadings suggesting he was of the view that the NRC "report is not intended to affect the admissibility of any forensic evidence,": "I most certainly never said, or even suggested, that judges should not take into account the new information provided by the Report in assessing the validity and reliability of forensic evidence while making admissibility determinations. Claims to the contrary are without basis in fact and utterly absurd."[11] The entire legal system should take heed of this report and any suggestion that it should not would be "absurd".

Finally, as every true scientist knows and Karl Popper warned, "The game of science is, in principle, without end. He who decides one day that scientific statements do not call for any further

---

[10] Senior Circuit Judge and Chief Judge Emeritus, United States Court of Appeals for the D.C. Circuit and Co-Chair, Committee on Identifying the Needs of the Forensic Science Community, The National Academy of Sciences

[11] H.T. Edwards, *The NAS Report on Forensic Science – What it Means for the Bench and Bar* p.4. https://www.cadc.uscourts.gov/internet/home.nsf/AttachmentsByTitle/NAS+Report+on+Forensic+Science/$FILE/Edwards,+The+NAS+Report+on+Forensic+Science.pdf

test, and that they can be regarded as finally verified, retires from the game."[12] True science has never been a game of "trust us," but instead one of "here is the data, here is the experimental design, now prove us wrong." The Draft Report appropriately emphasizes the need for more testing and data, but most critically for more transparency and independence. And we support NIST's approach to creating transparency that both effectively protects and addresses privacy concerns while providing the necessary level of transparency consistent with sound science.

In evaluating a scientific system that has the power to take away both life and liberty, emphasis on the basic scientific principle of putting the system to the test is both a scientifically sound approach and consistent with the fair administration of justice. We wholeheartedly recommend that the Draft Report be published in its entirety and encourage NIST to address the specific concerns we have raised in this letter in the final report.

Sincerely,

Jerome F. Buting
Attorney

Keith A. Findley
Professor of Law
University of Wisconsin-Madison

Jennifer Friedman
Past President
California Public Defenders Association

Jo Handelsman
Director
Wisconsin Institute for Discovery
University of Wisconsin-Madison

Julia Leighton
Retired
Public Defender Service for the District of Columbia

Erin Murphy
Norman Dorsen Professor of Civil Liberties
NYU School of Law

M. Katherine Philpott
Virginia Commonwealth University
Department of Forensic Science

---

[12] Karl Popper, *The Logic of Scientific Discovery*, at 32 (Routledge Classics 2005) http://strangebeautiful.com/other-texts/popper-logic-scientific-discovery.pdf.

PC49

response to mixture report

Kalafut, Tim [redacted]

Mon 8/23/2021 10:16 PM

To: **ScientificFoundationReviews <ScientificFoundationReviews@nist.gov>**
Cc: 'Simone' [redacted]

Please find attached comments on the NIST draft mixture report from Tim Kalafut, PhD and Simone Gittleson, PhD.

PC49a

We are commenting on the NISTIR 8351-DRAFT report DNA Mixture Interpretation: A NIST Scientific Foundation Review and offering both editorial and content suggestions. Line number is followed by comment

486 Perhaps the document could define the term "relevance" here. The way it is used throughout this report leaves the overall impression that they are considering the evaluation of the DNA results with regard to offense level propositions.

677 Key Takeaway #2.6 "Likelihood ratios are not measurements. (…)". Likelihood ratios measure the degree of support the data provides for one proposition with regard to an alternative proposition. In other words, it is a measurement of information. Its measurement unit is the hartley, ban (deciban), or dit.

725-728 This sentence directly states that only the 6 primary authors are qualified to evaluate publicly available information while dismissing the "claims" and "understanding" of actual practitioners and researchers who evaluate DNA mixtures and the use of PGS in their day-to-day work environment. Some readers could interpret this as arrogance, and as a statement of condescension and dismissal of all other DNA practitioners and points of view.

741 This report cites over 60 papers, 1000's of samples, and millions of LRs – yet that's not "enough" – and there is no indication of what is enough, other than "to allow for external and inde pendent assessments". In the accompanying webinar for this document on July 21, 2021 (Webinar on DNA Mixtures: A NIST Scientific Foundation Review | NIST), a statement was made to the effect of "any interested stakeholder" should have access to such publicly available data in order to assess reliability. We know of no such organization or court system with this end point. Presumably this would include investigators, jurors, and competing private companies, among others. This is an impossible end point to reach.

759-760 Agree – proficiency tests should be much more realistic. However, there is an inherent conflict of interest here between the proficiency test providers and the labs that buy them. If th e tests are too hard, and failure becomes a realistic possibility, the labs will move to a different provider.

792-811 The term "readily" is meaningless and dangerously misleading. This section uses unhelpful terms such as "might have been" and speculations which are unacceptable statements for an expert to make while operating in evaluative mode (e.g., 803-805). Perhaps this language is intended to discuss an investigative role of the expert? If so, the text must clearly delineate the difference between the investigative and evaluative roles and the specific language that is appropriate for each. This document mentions both investigative and evaluative roles, but when the opinions of the authors are expressed, such as in this section, it is not clear which role is being addressed.

789-811 There is no delimitation nor differentiation between an activity level evaluation and an offense level evaluation. To see this, note the following fundamental difference between the activity and offense levels: a) The evaluation for activity level propositions considers transfer, persistence and the presence of background DNA; b) The evaluation for offense level propositions considers relevance. This distinction is crucial, because forensic scientists don't evaluate DNA results with regard to offense level propositions. This distinction should appear very clearly in this text to avoid any ambiguity between an activity level evaluation and an offense level evaluation.

809-811 This includes the phrase "Relevance should be assessed." This is the offense level of the hierarchy of propositions, and it is clear throughout the scientific literature that this is not the purview of the expert. The courts are clear that only the jury makes this assessment, with the judiciary playing the role of the gatekeeper. A discussion of "relevance" and "misleading" evidence is not appropriate in a scientific review destined for DNA analysts in the evaluative role. I suspect the intention is to ensure the sub-source level LR and the activity level evaluation of DNA data are not conflated, and to encourage a full and proper evaluation of DNA given activity level propositions. If this is the case, much of Chapter 5 needs significant revision.

839-847 and Key Takeaway #5.4 Presumably this section and key takeaway are written for the jury, as this is who ultimately must understand the difference between the levels of the hierarchy of propositions. See also comment above for lines 809-811.

856 Key Takeaway #5.1 includes the phrase "may be unrelated (irrelevant) to the crime..." Again, this is not appropriate in a foundational document for experts operating in evaluative mode (see comment above for lines 809-811). It is not the role of the expert to make this decision. This would be a form of bias and even violate the ethics statements of various organizations. Is this key takeaway meant for the jury?

930-933 The fact is that the high-sensitivity of DNA methods has been discussed since 1997 – almost 25 years ago. This review is disingenuous by making a distinction between "early decades" and "today" as though increased sensitivity, and the need for a proper and distinctly different evaluation given activity level propositions, is a new problem.

967 This document cites papers from 25 years ago that show this is nothing new, yet constantly stresses the problems of "today".

1059 This is again dealing with "relevance" and is inappropriate for this foundational review as written, as this is for the jury and the investigator. The ideas presented here are only appropriate during the investigative phase, not the evaluative phase (see ISFG and other guidelines). This section does mention "users" of the DNA results. It is true that such users need to understand the context, but unless this document is presented to juries, this should be removed. The overall impression of this document is that somehow the DNA expert is to make this assessment. If this is not the intent, significant effort must be put into re-writing this document to make clear the difference in roles between the laboratory (experts in the evaluative role), the investigators, and the judges/juries.

1187 It seems like the main authors did not care for the opinions of the resource group: "The Resource Group reviewed an early draft of this report and provided valuable feedback, insights, and suggestions during its development. However, they were not asked to provide consensus advice or recommendations, sign off on our final report, or endorse its conclusions. The NIST team is grateful for their dedication and contributions to our efforts." Some readers might consider this an arrogant comment, and dismissive of the resource group. (See 725-728.) Perhaps that section and this section would benefit by being re-written.

1469 This is not a helpful analogy. Yes, calculus is hard – for lots of people. However, it is a valid branch of mathematics, and without it we would never have made it to the moon. I have yet to read anything that says because calculus is difficult, we should not use it until "anyone" (see Line 2404 and footnote on page 87) can understand it, yet that seems to be the message of this foundational review with respect to forensic DNA. This type of hyperbole is not useful, and will only contribute to further misunderstandings and the potential twisting of DNA to fit various agendas.

1656 Stutter masking is discussed, but this foundational review does not mention allele specific filters that assist and have been published in the literature. It would seem a foundational review would want to discuss tools that can improve the situation.

1705 No discussion that the advances in DNA science have allowed for the exoneration of falsely accused persons – even from complex mixtures.

1766 This would appear to suggest the "simple, two-person mixtures" are OK for use, yet in the webinar, the lead author of this review made it clear that perhaps even two-person mixtures are unreliable.

1912-1916 The report refers to a non-peer reviewed article to seemingly promote a narrative, as Iyer is an author of both this report and an author of the non-peer reviewed article cited here. This is an example of bias, as there is no summary of a different view – only a statement that "comments on these concerns" have been published. As written, this could mean that others have the same concerns, yet those papers are direct rebuttals of Lund and Iyer.

1986 Emphasis added with no explanation as to why there is a need to do so.

2030 The use of the word "perceived" here might be considered as a form of bias in this report. The actual wording in the UKFSR document is "The benefits of using DNA mixture interpretation software..." The addition of the word "perceived" here dismisses these benefits. Some readers might consider this as a comment where the authors of this document have special knowledge or qualifications that allow them to have final judgement on various issues.

2217 (Principal 7) The last sentence in italics is not true. DNA cannot answer "who" questions by itself, ever. If so, there would be no need to calculate any weight of evidence. And with proper evaluation, DNA can be used to give information related to activity. This report dedicates an entire chapter to this concept. As this is written, some readers might think this report does not respect the hierarchy of propositions based on the language used. As worded, this sentence may cause a lot of issues if mis-used at court. To be stereotypical, prosecutors could argue that DNA can by itself identify a person, and defense could argue that DNA can never be used to evaluate activity, as all things are possible. Both would be wrong.

2289 Allele specific stutter filters can be helpful here, yet this foundational review does not mention published literature that addresses this very concept.

2352 (Principle 16): The propositions H1 and H2 do not need to be "exhaustive". They should be "exhaustive given the case information and assumptions" in order to be useful to the court. The lead author of this report is the third author of the ISFG's recommendations Part I that directly states "[propositions] do not need to be exhaustive...". That propositions do not need to be exhaustive is also stated in the UK's Forensic Science Regulator's guidance, the ENFSI Guideline for Evaluative Reporting in

Forensic Science, and Practitioner Guide No. 4, Case Assessment and Interpretation of Expert Evidence, Guidance for Judges, Lawyers, Forensic Scientists and Expert Witnesses. See Buckleton et al. (2021) [Forensic Science International: Genetics 50: 102406] for the direct quotes and for an excellent explanation on why propositions do not need to be exhaustive.

2478 There is a claim that the eight validation summaries found are not sufficient for "independent assessment". There are potential problems with this statement. Some readers might interpret this to mean that these NIST report authors claim to be the final gatekeepers of what is sufficient. Some might point out that there is no guidance in this document as to what realistic requirements would be sufficient. Some readers might suggest there is no understanding of very real privacy issues, IRB restrictions, and even various laws banning the release of genotypes.

2846 If all factor space must be bracketed, a definitive list of factors should be provided. Otherwise, laboratories and researchers will never have an endpoint, as some external independent reviewer can claim some new factor (or divide factors into sub-factors) and declare the validation as incomplete and therefore reliability is not established. It would be helpful if both a definitive list of factors and appropriate bracketing were proposed. In other words, this document should provide a rubric. Otherwise, we are left to consider a quote from Lord Byron: "When you can measure what you are speaking about, and express it in numbers, you know something about it, when you cannot express it in numbers, your knowledge is of a meager and unsatisfactory kind; it may be the beginning of knowledge, but you have scarcely, in your thoughts advanced to the stage of science."

2900 This is a straw man argument – no one does 8-person mixtures from 10pg total DNA- that is a fraction of a cell per person. However, perhaps this is suggesting that validations should include such samples to allow the most complex mixtures to have now been bracketed. The LRs would be essentially 1 +/- for all persons, so this would be of limited value, but would seem to meet the proposed requirement for bracketing.

3012/3030 The text discusses the concept of allele sharing in the validation of various PCR kits. Some deficiencies in this area are noted. Some might claim there is a difference between validating the actual kits and the validation of interpretation strategies, including those that use PGS. The implication here is that the kit manufacturers must also be PG and interpretation experts. This might be pointed out as an example of conflated concepts in this report that should be dealt with separately and clearly. There is a point here – that the interpretation of related persons (high allele sharing) is a challenge. But some readers might interpret the general tone of this section that all prior work is negated. It might be more constructive to simply state that allele sharing is an area of internal validation studies that should be explored more, and then give useful information and criteria that labs can use. For example, a suggestion for a way to measure and quantify allele sharing might have been offered.

3074 Key Takeaway #4.3 "Currently not enough data to enable external and independent assessment" – Some readers might interpret this as a statement of fact. Giving the benefit of the doubt to the authors, perhaps this should be couched in terms of their opinion as there is no discussion of what would be enough data. The authors have chosen not to share their opinion about how much might be enough, nor do they define "external and independent assessment". They do, however, reference approximately 80 published/available studies encompassing thousands of samples and millions of comparisons.

With regards to making data publicly available, some might think the authors are unaware of, and give no consideration to privacy rights, IRB issues, various laws, or the technical challenge of hosting websites that allow for public access of data. Should NGS become mainstream, the amount of data will feasibly be measured in terabytes.

This report neglects to mention and acknowledge that in any given year there are numerous workshops and seminars put on by both the creators of various PGS's and by independent users. If the authors have information that these are restricted to only some persons, perhaps they could call for such workshops to be attended by anyone that is interested, and provide, or call for, funding to cover the cost of those currently unable to have access.

3077 Proficiency results

Some readers may consider this to be the most problematic section of this entire report. It is fine to call for more realistic and challenging proficiency tests. However, as written the bulk of this section should perhaps be removed entirely. As outlined below, some might accuse this section of being factually incorrect.

1. CTS has never been designed to be a proficiency test related to PGS. The only reason PGS is relevant is that participants realized some comment must be made if they choose to report "elevated stutter" or all stutter peaks that their PGS software might consider, as stutter is not part of the expected profiles that CTS "grades" on.

2. CTS only added a specific response about PGS at the end of 2018. However, there is no requirement for participants to give a response. Some laboratories that use PGS to determine weight do not give a positive response to the PGS query, as they feel that PGS has nothing to do with the detection of alleles or inclusions/exclusions (what CTS "grades" on) and there is no requirement - nor ability at CTS to evaluate – evidential weight.

3. The authors claim that no participants used PGS prior to 2016. This is not correct. My old laboratory was using PGS for all cases – including CTS proficiency tests – in late 2014. However, as noted, there was no place to indicate this at the time. Depending on how a laboratory has implemented PGS into their workflow, PGS is not relevant to the CTS test and they never have, nor currently, make comment that PGS was used.

4. The authors close this section with a count of false positives and false negatives, and a statement that "In the past five years, the number of participants using PGS has grown." This would seem to be a non-sequitur. Some readers might consider this highly misleading and not based on an informed analysis of CTS results.

A. The data required to link false positives and false negatives to PGS does not exist in the CTS results summaries, if for no other reason than the data is incomplete about the use of PGS.

B. Although it is inappropriate to do so, as the necessary data does not exist nor is the available data in CTS summaries accurate, a rudimentary trend line of false positive and false negative results related to the use of PGS can be made. When comparing results prior to the claimed use of PGS by the report authors to the reported results in the PGS

era, the trend shows that PGS has reduced the number of expected errors. However, as stated, the data does not exist to actually do this, although the authors seem to be linking errors to PGS with the sentence cited above.

C. A casual view of the data in this table draws attention to test 19-5705 and 13 false negative results. Spending 20 minutes with the CTS summary shows that at least the first 3 false negative results were from participants that only completed the mtDNA section of the test. To include these errors in a discussion of PGS and proficiency tests might be considered inexcusable.

3201-3204 The authors claim that it is not possible for them to evaluate the reliability of even "any one point in the factor space." Some readers might consider this a rather incredible statement. Even though 60-80 sources were examined, not one single factor can be considered "reliable" in mixture interpretation? Perhaps part of the problem is that there is no "established and accepted criteria for reliability". What is noteworthy here is that the authors, about whom some readers might think make claim to have some special insight into what would be acceptable, do not fully define what factors need to be evaluated nor what constitutes an end point. Perhaps the second draft of this report could make some suggestions towards "established and accepted criteria for reliability" that laboratories and researchers can use.

3224-3232 This again mentions the need to "independently assess the degree of reliability of PGS models." There is no comment on who, or what qualifications are needed, to serve as this assessor. Some readers might think that perhaps the authors intend that they themselves are the last word in reliability assessment. This section contains one of the few clues about the criteria these authors might find useful. It seems that plots do not meet this criterion. "Metadata" is needed, but this metadata is not described. A paper is mentioned (Rodriguez et al. 2019) that seems to be endorsed as a model for providing this additional criteria. It would be helpful if the authors would explain why they think these supplementary tables are useful, but they are silent. As a ten year user of PGS, I have examined both the paper and all supplemental material. Clearly Rodriguez et al. have crunched a lot of data. But I do not see why their particular presentation and sharing is so much better than dozens of other studies. Perhaps this foundational review has missed an opportunity to educate practitioners on expectations for future work.

3243 However, PCAST did state that two-person mixtures and mixtures of three persons with at least 20% were reliable. The PCAST report neglected to include some studies that were available at the time it was written. There were numerous publications that addressed mixtures of three persons with less than 20% contribution of the minor. In fact, the Rodriguez et al. study that is modeled as containing the type of data the authors of this report wish to see use two-person mixtures where the minor donor is so low as to exhibit dropout.

In the accompanying webinar (Webinar on DNA Mixtures: A NIST Scientific Foundation Review | NIST), the lead author of this NIST report questioned even the reliability and use of two person mixtures with no quality issues (i.e., no dropout, no degradation, high rfu counts). We wish to point out that there are numerous standards and best practices documents along with many dozens of published peer-reviewed articles that disagree with that position. Some readers may question why the lead author of this NIST report did not acknowledge that two person mixtures can be reliable during the webinar.

3428 Some readers might accuse this report of ignoring the issues related to making data publicly available.

3463 The bracketing approach. There is only a partial list of factors that must be sufficiently bracketed, and no guidance is offered as to what this bracketing must look like nor some definitive list of factors. In practical terms, some readers might claim this is a game that cannot be won. Without any guidance as to what constitutes acceptable bracketing, ever increasing levels of granularity could be demanded by external, independent reviewers. If the authors are suggesting running a high-level single source profile and a mixture of extreme complexity (e.g., an 8-person 20pg sample) in order to "bracket" all the samples the laboratory would see in casework, this approach is unhelpful for informing a method's performance (i.e., we already know that the first LR will be very large and the second LR very close to 1) and does not validate a method. No standard setting organizations (OSAC, SWGDAM, ISFG, ENFSI, ANZPAA NIFS) have ever called for bracketing as discussed by this NIST report. Perhaps actual criteria and guidance could be offered for this novel concept not previously described.

3547 Here the authors appear to call for accuracy and reliability of a specific LR assignment, yet elsewhere (see 3576-3578) they admit that there is no true LR. Perhaps some guidance could be offered as to how a validation study should address this conundrum. The authors cite numerous papers in this report that give LR distributions for thousands of Hp true LRs and millions of Hd true LRs yet have stated all of this work has not yet resulted in the ability to determine the reliability of a single point in the factor space. This would suggest that their guidance in the details of designing validation studies is needed in the community. Dr. Butler has offered such implementable guidance before. For example, in a Profiles in DNA article entitled "Debunking Some Urban Legends Surrounding Validation Within the Forensic DNA Community" Dr. Butler has given useful advice such as:

> "Once the developer of a particular measurement technique demonstrates that it is robust, reliable and reproducible, validating the technique for use in your lab just requires establishing that it is working properly."
>
> "Unfortunately, some forensic DNA labs, often because they are driven by fear of auditors, are taking far too long or running far too many samples as part of their "validation studies". This over-validation can contribute to backlogs in already overburdened DNA laboratories..."
>
> "... the same conclusion could have been reached with far fewer experiments."
>
> "When conducting an internal validation, the SWGDAM Revised Validation Guidelines recommend running a total of at least 50 samples — not 50 samples per experiment."
>
> "URBAN LEGEND #2: VALIDATION IS UNIFORMLY PERFORMED THROUGHOUT THE COMMUNITY […] Auditors need to realize that variability can exist among validation studies."

Some readers might benefit from a review of all Urban Legends. This can be found here: https://www.promega.com/-/media/files/resources/profiles-in-dna/902/debunking-some-urban-legends-surrounding-validation-within-the-forensic-dna-community.pdf?rev=ac1404beab4544378e5f5b60adf76dea&sc_lang=en

3573-3578 This is a misquote and misuse of what this paper says. I am an author of the paper that was cited. The sentence directly following the quoted sentence reads, "But the truth or otherwise of each of

these..." (Gittelson et al., 2018). These lines were emphasizing that there is no true LR value, as "[t]he truth lies in the propositions". This is very different from stating the paper implicitly claims propositions are exhaustive. Furthermore, there is no requirement that either proposition is true. Nor should the expert be concerned about this. That is not the role of the expert. Finally, the ISFG guidelines on proposition settings (Part I) that is cited in this report specifically states that "These do not need to be exhaustive, but should reflect the positions of both parties." Note that the first author of this report is third author of these ISFG guidelines.

3594 Key Takeaway #4.8 "We encourage a separate scientific foundation review on the topic of likelihood ratios in forensic science and how LRs are calculated, understood, and communicated." -- Such a separate scientific foundation review on the topic of likelihood ratios in forensic science was published by three distinguished, leading mathematicians in 1908: *Rapport de MM. les experts Darboux, Appell et Poincaré. Affaire Dreyfus. La Révision du Procès de Rennes, Enquête de la Chambre Criminelle de la Cour de Cassation (5 Mars 1904 – 19 Novembre 1904). Tome Troisième. Ligue française pour la défense des droits de l'homme et du citoyen. Paris, 1908: p. 499-600.*

3658 Box 4.1 The AB, BC code in that paper has absolutely nothing to do with being an algebraic code used to protect the privacy of genotypes. While computers can encrypt data, I would suggest that doing so would make the data unusable to an external, independent reviewer. Some might consider this a naïve view of the situation, and I am personally already aware of situations where this generic form of "two heterozygotes sharing one allele" is being misunderstood as an acceptable substitution code.

3702 The authors claim that bracketing is a pragmatic approach. However, the reader might claim, they have given no list of factors that would need to be bracketed, nor guidance as to how granular the brackets would have to be. Is this suggesting that doing a high-level single source profile and a 1:1:1:1:1:1:1 10pg mixture of 6 persons would be sufficient bracketing? It would have been more helpful if there were guidelines offered on how to design a valid study that would be able to successfully bracket the non-defined sample space (although numerous factors have been mentioned) in a manner that is not 1000 samples. (See the article mentioned above (3547) where so many samples are not necessary and that there is no one way to structure or document a validation study.)

3721 ROC curves are mentioned as a way to examine discrimination efficiency, yet no other methods are discussed. There is what some readers might consider an unsupported statement that other tools "are less widely known to forensic DNA analysts." I am an author of a paper that describes a tool that incorporates the concepts of Ramos, and the tool is both publicly available and has been shared with well over a dozen laboratories that are either evaluating it or putting it into use. As a foundational review, one would think that a tool like this might be of interest. Yet this paper is only cited in what some might consider a dismissive fashion related to allele sharing (3389-3390). In addition, the creators of STRmix™ have developed and published on DBLR™ that serves a similar but expanded purpose. It is interesting to note that one of the authors of this report (Iyer) has written a non-peer reviewed white paper on ROC curves, and that is the only thing discussed to any degree in this report. In fact, the authors of this report overtly dismiss "aggregate graphs" (3588-3592) which some presume is why they do not discuss other published approaches to the examination of discrimination efficiency. They do not mention the limitations of presenting data in ROC curves (e.g., lack of dimension for an explanatory variable, difficult readability of LR values), nor do they explain that the two studies cited in "have been used in evaluation of PGS systems previously (e.g., Bleka et al. 2016b, You & Balding 2019)" (lines 3722-

3723) only use ROC curves to compare the performance of different models on a specific dataset, and not to validate a PGS system.

Some readers might wonder why ROC curves where the data points are calculated by a software that is being tested for reliability result in the best measure of reliability for software that is considered unreliable prior to the generation of the ROC curves.

3750 As written, some might suggest that Chapter 5 should be removed from this report. The language used seems to constantly and consistently mix interpretation, LRs, evaluation, investigation, and the levels of the hierarchy throughout Chapters 4 and 5.

3791-3795 This is the start of many discussions on "relevance". The forensic expert, in their investigative role, can sometimes offer comments on relevance. However, in our evaluative role, that is most definitely not the expert's purview. In the hierarchy of propositions, relevance is considered at the offense level, and this is the role of the jury. While this report discusses the investigative and evaluative roles, there are numerous sections of writing where it is not clear which role is being discussed. This must be corrected. If the intent is to make comments about the investigative role, it must be clear that the investigative role is the focus. If this is the evaluative role, these comments are wrong. When things are unclear, there is great risk of the statements made throughout Chapter 5 being accidentally or intentionally misused.

3794 This sentence as part of this section might be considered problematic. It seems to be addressed to both experts and non-experts, and seems to be discussing the actual evaluation of the activity level proposition, not the evaluation of the evidence given a pair of activity level propositions. If this is the evaluative role of the expert, this is the transposed conditional. Presumably the nonexperts mentioned are the jury members, who can do this evaluation. That is their role and would be an appropriate evaluation of the posterior probabilities or odds of the activities. Yet, this section reads as though the expert needs to be aware of this, which would imply an investigative role. If this is written to the expert in their investigative role, it should be more clear. This is an example of writing that perhaps should be reconsidered in this chapter.

3802 Some readers might consider this statement as an example of arrogance. These readers might suggest that it is not up to this group to decide, based on three papers, that what is relevant in a DNA case is often difficult to discern. There is a body of literature that might be understood to claim that it is not up to the forensic expert to decide relevance. Federal jury instructions might suggest that this role belongs to the jury, as relevance is taken into account at the offense level of the hierarchy of propositions. Notice also the terms used in this paragraph: "can detect", "could be recovered", and "can persist". There is a large body of literature that says these terms are not relevant when evaluating DNA evidence given activity level propositions. They are empty words that merely state the obvious. There are also voluminous writings in peer reviewed journals that the expert can help when activity level propositions are of interest at court. None of these discuss 'relevance', as 'relevance' is for the jury alone. If this is meant to suggest that when the DNA trace in question is of low level, poor quality, partial, or some combination of those attributes, a formal evaluation of such evidence should be done given relevant activity level propositions, perhaps this could be written that way.

4169, 4176 There are papers cited where 1-3 alleles or "detectable" alleles are found. Does this then not imply that when a major single source foreign profile is found, it may have some meaning, and this

evidence may be more likely if A happened than if B happened? Some readers might interpret much of this entire chapter as focusing only on what "could" happen or what is "detectable." There are frameworks and tools, such as CAI and BNs, that both acknowledge that a few alleles "could" be "detected" and take this into account to evaluate the evidence appropriately given activity level propositions. Further, when the authors share their views like this, some readers might claim they do not demonstrate the principles of robust, logical, transparent, and above all, balance in their evaluation of data when the question before the court is dealing with activity level propositions.

In Chapter 4, the authors claim that without metadata, the validation studies have no meaning. Some reader might wonder why in Chapter 5, the authors discuss many papers/studies lacking this type of metadata: "non-self DNA was detected", "foreign alleles detected", once in a while it says, "1-3 alleles". Metadata equivalent to what the authors call for at the sub-source level is often missing for activity level experiments. Why do the authors not ask for, and model the same thing here? The effect of this writing without such metadata is to imply "all things are possible – and equally so".

4207(-4221) "DNA transfers readily" might be considered an unhelpful statement. Some might wonder what the transfer rate for "readily" is. This is in the middle of a large section that seems to point out that studies can detect DNA. A discussion whish demonstrates how this information might be used in evaluating DNA evidence given two mutually exclusive propositions would be a lot more useful. In addition, some might point out an inconsistency between chapter 4, where there is a call for propositions to be both mutually exclusive and exhaustive (the latter of which is not called for by numerous publications and organizations cited in this report), yet in Chapter 5 there is merely a discussion of one proposition at a time such as:

> 4077 "secondary and higher-order transfers of skin cells are facilitated more by non-porous substrates"
>
> 4081 "participants holding glass, cloth, and wood found the likelihood of obtaining a DNA profile was approximately 9%"
>
> 4138 "detected DNA out of doors that had been deposited up to two weeks before"
>
> 4175 "simple minor everyday interactions involving only a few items in some instances lead to detectable DNA"
>
> 4182 "non-self-DNA was detected on 79% of hands"
>
> 4214 "... washing studies... DNA from family members was detected on children's underwear even in instances where semen was not placed..."
>
> 4342 "Many of the studies on transfer and persistence in which ground truth is known note the presence of alleles not associated with subjects of the study"
>
> Some readers might point out that all of these statements are lacking metadata ("detected", "presence of alleles", "sometimes", [undefined quality] "profile"), and only discuss one-sided evaluations of such data. Some might point out that if a high-level, full and complete profile is obtained from an item that matches a person of interest, studies such as these suggest that is rather low probability if participants held cloth, or deposited DNA two weeks ago, or only had

everyday interactions, or was a result of washing in the same laundry machine. Some readers might suggest that a discussion like this would be of use in a scientific foundation review.

4479 Informed readers that have spent much time and effort dealing with activity level propositions might suggest that this "one must separate relevant DNA from irrelevant DNA" is for the jury. It is not clear who the "one" is in this statement. Perhaps the authors are trying to state an objection that any given DNA result is always more likely given the activity proposed by the prosecution. It might be a stronger statement to point out that sometimes the evidence is more likely if an activity proposed by the defense occurred than that given by the prosecution. Or that sometimes the evidence is equally likely under both propositions and is not informative for answering the question the court is interested in.

4518 Perhaps this could be rewritten to directly state that the LR given sub-source level propositions, no matter the magnitude, should never be used as the weight of the evidence with regard to activity level issues. It is also true that the LR given sub-source level propositions is not equal to the LR given offense level propositions when the relevance of the DNA is uncertain, which may be a more useful statement than only discussing relevance again.

4546-4547 Key Takeaway #5.2 "(…) When assessing evidence that involves very small quantities of DNA, it is especially important to consider relevance." See comments for lines 809-811, 3791-3795 and 4479.

4601-4625 Perhaps this might be a place to address not only the probability of the evidence given competing propositions (at either sub-source or activity levels), but also discuss Bayes' theorem and that the LR presented by the scientist is just one part of this framework, which also takes into account all of the other evidence in the case (e.g., the prior odds in Bayes' theorem). We have noticed that although the United States have adopted the LR for assigning the weight of evidence, the adoption of a full Bayesian framework has not been widespread. Some might suggest that the authors of this report have not embraced the Bayesian framework, as there is no discussion in this report about how the LR should update the belief of the jury LR times from what their belief(s) are without the DNA evidence. Perhaps, a full Bayesian framework where the DNA expert educates the jury about DNA being only one piece of evidence that must be taken into account along with all the remaining evidence might have prevented the miscarriage of justice in the case that is cited here. It would appear that even the judge in this case would benefit from such information.

4691-4693 "An activity proposition might be, for instance, that DNA collected during a sexual assault examination was deposited during sexual activity, or that DNA found on the handle of a knife was deposited during the act of stabbing a victim." – These are examples of incorrect activity level propositions. They violate Recommendation 6 of the ISFG Guidelines published in 2020 (Gill et al., 2020). Note that the lead author of this NIST foundational review is the third author on these ISFG guidelines. The correct formulation for these propositions is: "...that sexual activity occurred between the POI and the complainant, or that the POI stabbed the victim with this knife". Interweaving factors that scientists take into account in their evaluation into the propositions is a grave mistake because it affects the entire evaluation that follows, producing a very misleading LR value.

4727-4730 "For example, in the case of a stabbing, the prosecution hypothesis might be that the DNA was transferred to the handle of a knife during the activity of stabbing, while the defense hypothesis might be that the DNA was deposited due to contamination or secondary transfer." – This is another example of an incorrect pair of activity level propositions that violates Recommendation 6 of the ISFG Guidelines (Gill et al., 2020). Note that the lead author of this NIST foundational review is the third author on these ISFG guidelines. The correct formulation for this pair of propositions is "The POI stabbed the victim with this knife" for the prosecution's proposition, and "The POI did not touch this knife" for the defense proposition. Interweaving the term 'transfer' and the factors that scientists take into account in their evaluation into the propositions is a grave mistake because it affects the entire evaluation that follows, producing a very misleading LR value.

4762 This section appears to state what the main message of Chapter 5 is. It would be clearer for the reader if this section (actually, most of 5.4.2) would be presented much earlier in the document, and then more attention paid to the language used in the rest of Chapter 5 to try and fit with what is discussed here.

4777-4779 Some might suggest that much stronger language be used here. If the authors are comfortable with making demands in Chapter 4 about factor space, bracketing, and ROC curves, perhaps they could make fairly strongly worded comments here as well. Perhaps forensic DNA laboratories should not be high-throughput labs if they do not consider DNA evidence given activity level propositions. Perhaps this section could call for CAI to be implemented at all forensic DNA laboratories, and for the evaluation of the DNA results given appropriate activity level propositions to be included in written reports.

4799 Since the authors now discuss investigation and evaluation, perhaps they could clarify much of the previous writing to make it clearer regarding the distinction between these roles.

4861 Key Takeaway #5.5 is a very sage takeaway indeed. Yet, much of the document prior to this leaves the reader rather gobsmacked to find this statement. Some readers may not make it this far into the document to ever find this statement.

4864 This section, and most of 5.4.2, is quite useful.

Tim Kalafut, Ph.D.
Simone Gittelson, Ph.D.

PC50

Comments on Scientific Foundation Review: DNA Mixture Interpretation

Anthony Onorato [redacted]

Thu 8/26/2021 4:20 PM

To: **ScientificFoundationReviews <ScientificFoundationReviews@nist.gov>**
Cc: Butler, John M. (Fed) <john.butler@nist.gov>; RUSSELL VOSSBRINK [redacted]; Anthony Onorato [redacted]

Dear Sir/Madame.

Please see the attached.

Regards

Anthony J. Onorato
Chair
Scientific Working Group on DNA Analysis Methods (SWGDAM)
FBI Laboratory

PC50a

SCIENTIFIC WORKING GROUP — SWGDAM — DNA ANALYSIS METHODS

August 23, 2021

Dr. John M. Butler
NIST Fellow
Special Assistant to the Director for Forensic Science
Special Programs Office
National Institute of Standards and Technology
100 Bureau Drive, Mail Stop 4701
Gaithersburg, MD 20899-4701

Dear Dr. Butler:

The Scientific Working Group on DNA Analysis Methods (SWGDAM) respectfully submits the following substantive comments on NISTIR 8351-DRAFT entitled *DNA Mixture Interpretation: A NIST Scientific Foundation Review*.

Because the draft Report summarizes a number of SWGDAM Guidelines documents in its Appendices, we offer the following background for our comments. As you know, SWGDAM is a forensic DNA working group, established over 25 years ago, to "serve as a forum to discuss, share and evaluate forensic biology methods, protocols, training, and research to **enhance** forensic biology services…" Congressional recognition of SWGDAM's predecessor, TWGDAM, is contained in the Federal DNA Identification Act with the provision establishing the TWGDAM Guidelines as national standards for participation in this National Index until the new Federal DNA Advisory Board recommended national quality assurance standards to the FBI Director for adoption.[1] The Federal DNA Advisory Board was responsible for recommending quality assurance standards, and revisions as necessary, to the Director of the Federal Bureau of Investigation (FBI), and when their statutory time period concluded, the Board charged SWGDAM with this responsibility.

[1] Enacted in 1994, the Federal DNA Identification Act is the enabling legislation for the National DNA Index System (NDIS); *see* 34 U.S.C.§12591 *et seq*. The provision relating to the use of the TWGDAM standards in the interim before the Federal DNA Advisory Board recommends standards to the FBI Director can be found at 34 U.S.C. §12591(a)(4).

SWGDAM is currently comprised of dedicated forensic scientists, from international, federal, state and local forensic DNA laboratories as well as guests representing academia and other Federal agencies. These forensic scientists serve as the DNA technical leaders or Combined DNA Index System (CODIS) Administrators for their laboratories and are able to offer the perspectives of practitioners in the areas of STR, Y STR, mitochondrial DNA and next generation sequencing (NGS) technologies. We are fortunate to also have invited guests attend each meeting to provide their specific expertise in areas such as population genetics, Rapid DNA, probabilistic genotyping, statistics, etc.

Laboratories performing forensic DNA analysis and participating in the FBI's National DNA Index, unlike other forensic disciplines, are subject to Federal statutory requirements relating to quality assurance and privacy. As early as 2006, accreditation was required for forensic DNA laboratories contributing DNA records to the National Index. In addition to generating these DNA records in accord with specific minimum standards for a quality assurance program, these laboratories also abide by limited disclosure and release rules for their DNA records. Federal law also requires that these laboratories undergo an external audit every two years to document their compliance with the minimum standards. In fact, the Quality Assurance Standards audit documents contain a separate appendix – "Appendix E: Approved Validations" – to document the evaluation and approval of developmental and internal validations during the audit.[2] Moreover, the U.S. Department of Justice' Office of the Inspector General conducts audits of CODIS laboratories for compliance with the Federal DNA Act requirements as well as NDIS Operational Procedures.[3] In addition to this rigorous program of audits, the FBI's CODIS Unit conducts assessments of the NDIS participating laboratories as part of their administration of the National DNA Index System.

Several states also have an additional level of oversight provided by their State forensic oversight boards that review, evaluate and approve new technologies or methods prior to their use by forensic DNA laboratories, such as, the Connecticut DNA Data Bank Oversight Panel, the District of Columbia's Scientific Advisory Board, the

[2] *See The FBI Quality Assurance Standards Audit for DNA Databasing Laboratories*, and *The FBI Quality Assurance Standards Audit for Forensic DNA Testing Laboratories*; available at https://www.swgdam.org/publications.

[3] *See Combined DNA Index System Audits,* available at https://oig.justice.gov/reports/codis-ext.htm.

Massachusetts Forensic Science Oversight Board, the New Mexico's DNA Oversight Committee, the New York State Commission on Forensic Science and DNA Subcommittee, the Texas Forensic Science Commission, and the Virginia Forensic Science Board and Scientific Advisory Committee.[4]

Since the overwhelming majority of our membership and invited guests represent NDIS participating laboratories, we wish to highlight the Federal and state statutory requirements for the confidentiality and privacy of the DNA data, in addition to those privacy protections afforded by the Genetic Information Nondiscrimination Act of 2008, the Health Insurance Portability and Accountability Act of 1996, and applicable State genetic privacy laws.[5] The Federal DNA Identification Act provides for limited access to and disclosure of the DNA samples and resulting DNA analyses generally to criminal justice agencies for law enforcement identification purposes. The Federal DNA Act responsibilities are further explained in the NDIS Operational Procedures Manual which limits access for anonymized DNA data **to criminal justice agencies** for a population statistics database, forensic identification, forensic research, forensic protocol development or quality control purposes. The FBI *Quality Assurance Standards for DNA Databasing and Forensic DNA Testing Laboratories* also emphasize the confidentiality of the DNA data in Standard 11.

In addition to the limited access requirements of Federal law, states also prescribe access to the DNA data in their State databases. The overwhelming majority of State laws restrict access to the DNA data in these law enforcement databases **to criminal justice agencies** for law enforcement identification purposes.[6] These State

[4] *See* National Conference of State Legislatures, *Legislative Study and Oversight of Forensic Services*; available at https://www.ncsl.org/research/civil-and-criminal-justice/dna-database-search-by-policy.aspx.)

[5] *See generally*, Genetic Information Nondiscrimination Act of 2008, P.L. 110-233 (Section 206 of GINA provides for the confidentiality of an employee's genetic information and specific limitations on disclosure), available at https://www.eeoc.gov/statutes/genetic-information-nondiscrimination-act-2008; and the Health Insurance Portability and Accountability Act, *see* P.L. 104-191 and the Final Omnibus Rule at Fed. Register, Vol. 78, No. 17 (2013) (the HIPAA Privacy Rule includes safeguards to protect personal health information); Smith, S., Nielson, P. S., Kennedy, B. (2011) *Genetic Privacy Laws: 50 State Survey*, Journal of Health & Life Sciences Law, Vol. 5, No.1. *See also,* States enacting protections for genetic information in 2021, such as South Dakota S.B. 178, Utah S.B. 227 (Chapter 361), Florida H.B. 833 (Chapter No. 2021-216).

[6] *See, for example*, Ala. Code §36-18-27; Alaska Stat. §44.41.035(F); Ariz. Rev. Stat. Ann. §13-610. I; Ark. Code Ann. §§12-12-1112, 12-12-1114; Cal. Penal Code 299.5(A), (F); Colo. Rev. Stat. Ann. §§16-

laws also penalize the unauthorized disclosure of DNA data as misdemeanor or felony offenses.[7]

SWGDAM has several general observations about the NISTIR 8351:DRAFT: (1) the lack of forensic science practitioner(s) expertise among the authors may have hampered consideration of the existing regulatory and statutory framework within which forensic DNA laboratories operate; (2) the introduction of a new requirement that internal validation be 'publicly available' in order to be considered within the foundational review appears to be based solely on the authors' belief, with no further justification, especially when such a requirement is in conflict with Federal and State privacy laws for DNA data; (3) its use of 'factor space' to suggest that existing internal validations are inadequate and/or incomplete fails to acknowledge existing practices and guidelines to ensure the limitations of these systems are tested; and (4) its failure to

11-102.4(5), 16-23-104; Conn. Gen. Stat. Ann. §§54-102i(B), 54-102j; Fla. Stat. Ann §943.325(14); Haw. Rev. Stat. Ann. §§844D-81, 82, 91; Id. Code §19-5514; 730 Ill. Rev. Stat. Ch. 5, Para. 5-4-3(F); Kan. Stat. Ann. §21-2511(h)(2); La. Rev. Stat. Ann. §§15:612, 15-616; Me. Rev. Stat. Ann. tit. 25 §1577; Md. Code Ann., Pub. Safety §§2-502(C)(4), 2-508; Mass. Ann. Laws Ch. 22E §§9, 10; Mich. Comp. Laws Ann. §28.176(2); Minn. Stat. §299c.155(3); Mo. Rev. Stat. §650.055 (7); N.J. Rev. Stat. §§53:1-20.24, 53:1-20.27; N.M. Stat. Ann. §29-16-8(A), (B); N.Y. Executive Law §995c(6); N.C. Gen. Stat. §§15a-266.8, 15a-266.12; N. D. Cent. Code §§31-13-06; Ohio Rev. Code Ann. §109.573(b)(2), (E); Okla. Stat. Tit. 74, §150.27a (D); R.I. Gen. Laws §§12-1.5-11, 12-1.5-16; S. C. Code Ann. §23-3-650(A); S. D. Codified Laws §§23-5a-22, 23-5a-23, 23-5a-25; Tenn. Code Ann. §38-6-113(d); Tex. Gov't Code Ann. §§411.147 (C), 411.153(A); Utah Code Ann. §53-10-406(3)(A); Vt. Stat. Ann. Tit. 20, §§1937, 1938, 1941; Va. Code Ann. §§19.2-310.4, 19.2-310.5; Wis. Stat. §165.77. States that may have more expansive access provisions, as a condition for participation in NDIS, agree to comply with the limited access and disclosure provisions of the Federal DNA Identification Act. *See* NDIS Operational Procedures Manual, Section 3.2, Version 10 (2021); available at https://www.fbi.gov/file-repository/ndis-operational-procedures-manual.pdf/view.

[7] *See, for example*, Ala. Code §36-18-28; Alaska Stat. §11.56.762; Ark. Code Ann. §12-12-1115; Cal. Penal Code 299.5; Colo. Rev. Stat. Ann. §24-72-309; Conn. Gen. Stat. Ann. §54-102k; Del. Code Ann. Tit. 29 §4713(l); Ga. Code Ann.§35-3-164; Haw. Rev. Stat. Ann. §844D-113; Id. Code §19-5514; 730 Ill. Rev. Stat. Ch. 5, Para. 5-4-3(f-5); Ind. Code Ann. §10-13-6-22; Iowa Code §81.6; Kan. Stat. Ann. §21-2511(n, o); La. Rev. Stat. Ann. §§15:617-618; Me. Rev. Stat. Ann. tit. 25 §1578; Md. Code Ann., Pub. Safety §2-512; Mass. Ann. Laws Ch. 22E §§12, 13; Mo. Rev. Stat. §650.055(5); Neb. Rev. Stat. §29-4110; N.H. Rev. Stat. Ann. §651-C:4; N.J. Rev. Stat. §53:1-20.26; N.M. Stat. Ann. §29-16-12; N.Y. Executive Law §995-f; N.C. Gen. Stat. §15a-266.11; N. D. Cent. Code §31-13-09; Ohio Rev. Code Ann. §§109.99, 109.573(G); Okla. Stat. Tit. 74, §150.27a (D); 44 Pa. Cons. Stat. §2332; R. I. Gen. Laws §12-1.5-15; S. C. Code Ann. §23-3-650; S. D. Codified Laws §23-5a-26; Tex. Gov't Code Ann. §411.153; Vt. Stat. Ann. Tit. 20, §1941; Va. Code Ann. §19.2-310.6; W. Va. Code §15-2B-12; Wis. Stat. Ann. §165.77(5); and Wyo. Stat. §7-19-404.

pursue constructive alternatives to facilitate a comprehensive scientific foundation review or to now consider additional available data to provide meaningful thresholds and guidance to the forensic DNA community in establishing the reliability of probabilistic genotyping systems.

SWGDAM collected comments from its members and invited guests (*see* attached SWGDAM Comments) to forward to NIST for consideration. There are several points within the Table that we wish to highlight for your review relating to the benefits of including forensic science practitioners on the authorship panel; the arbitrary decision to only review publicly available data; the introduction of a new proposition that each factor/nuance must be evaluated by laboratories in their internal validation of probabilistic genotyping systems in order to adequately understand its limitations; and the impact of failing to establish clear criteria/thresholds for reliability and the foundational review.

Forensic science practitioner input/involvement throughout the process

It appears that the regulatory environment under which forensic DNA laboratories operate was not fully considered and may have been misconstrued as lacking in transparency and/or accountability, when, in fact, internal validations conducted by such laboratories are subject to review during audits and assessments as well as throughout the criminal justice process. This apparent oversight is just one justification for having experienced practitioners of the relevant forensic discipline on the authorship team.

We appreciate the need for objectivity but suggest that the failure to include representatives of the discipline being evaluated does a disservice to the credibility and validity of the report findings as the critical practitioner perspective is not present. While the draft report acknowledges a DNA Mixture Resource Group that included forensic science practitioners, it notes that their involvement was limited to the early stages of the report and none of those individuals are listed as report authors. Significantly, the expertise and training of the forensic science practitioners would have provided invaluable insight on the issues of making internal validation data public as well as the practicalities of assessing factor space and including validation performance results in case files and reports. As described in the attached comments, a major strength of the NIST sponsored Organization of Scientific Area Committees is that it

has embedded stakeholders who are active in the writing of standards and who have a vote on the outcome/final products; such was not the case with this draft report.

Requirement for validation data to be 'publicly available'

The draft NISTIR 8351 emphasizes what is referred to as a lack of demonstrated reliability and deemphasizes (a) the challenges and inability to publish internal validation data; (b) the existence of significant internal validation data performed by forensic DNA laboratories; and (c) the independent and external audit system required for all NDIS participating laboratories as well as other audits, such as ISO/ANAB. The authors made no meaningful attempt to request data directly from these laboratories, a point not addressed in the draft report, using only web searches to determine what may have been 'publicly available.' We suggest that a review of materials available via a basic internet search are neither indicative of the scope nor the reliability of the community's work, especially when there is no stated requirement to publish internal validations. In fact, publication has been discouraged by scientific journals because such material is generally not considered novel. Dr. Michael Peat (editor of the AAFS Journal of Forensic Science) confirmed this at Meeting #12 (January 9, 2010) of the National Commission on Forensic Science. A laboratory, even if granted permission to use time and personnel resources towards publication efforts instead of more casework/public safety-focused endeavors, would likely have to publish in a "pay to publish" journal (such as Forensic Science International: Reports). However, it is not clear whether the authors would consider this sufficient, nor does this seem like a reasonable resource expenditure for most public crime laboratories.

The authors may wish to consider an approach to establishing foundational validity that is more practical than the mass publication of internal validation studies in peer reviewed journals while concurrently making all the underlying data publicly available. The authors should also clearly acknowledge the privacy and legal issues that surround public access to a laboratory's validation data, to avoid the perception that forensic DNA laboratories are simply being obstructionist rather than complying with Federal and State statutory requirements and obligations to safeguard the confidentiality of this DNA data.

Limiting the foundational review to only 'publicly available' data is an excessively restrictive measure and there is no statutory or scientific requirement for such a limitation, other than the author's belief.[8] The authors should be open to including a review of available validation summaries as a means to establish validity. To require a review of the data itself rather than a review of summarized results that already exist in both peer-reviewed publications and laboratories' validation summaries is contrary to established practice for assessing validity in other branches of science. And, as previously explained, NDIS participating laboratories are required by the Federal DNA Act to maintain the privacy of forensic samples. Donors of study samples also have an expectation of genetic privacy[9], so the "requirement" by the authors that genetic data be made 'publicly available' in order for an assessment of reliability to be performed contravenes statutory and regulatory requirements under which NDIS participating laboratories operate.

In addition to the mechanisms previously described for the review of the internal validation data (such as audits and assessments), there are also well-established avenues for the disclosure of this information, such as discovery and pre-trial hearings in criminal cases; processes within the criminal justice system designed to protect defendants' rights and also safeguard the privacy of the DNA data.

'Factor space'

The draft report introduces the term "factor space" into the evaluation of foundational validity of DNA mixture measurements and interpretation. As applied in the draft, "factor space" is meant to "describe the factors that influence complexity, measurement, and interpretation reliability – these factors include the number of contributors, the degree of allele sharing, the ratios of mixture components, and the amount and quality of the DNA tested." Despite the new terminology, this is not a new

[8] Beginning at line 2402, the draft report states, "We recognize that there are information and data collected in forensic laboratories that may not yet be publicly available or published. However, we believe for information to be considered foundational, it needs to be reasonably accessible to anyone who wishes to review it."

[9] Federal and State laws also protect individual's privacy rights to their genetic information, *see for example*, the Genetic Information Nondiscrimination Act of 2008, P.L. 110-233, available at https://www.eeoc.gov/statutes/genetic-information-nondiscrimination-act-2008. *See also*, Congressional Research Services (2008) *Genetic Information: Legal Issues Relating to Discrimination and Privacy,* (State Statutes); available at https://www.everycrsreport.com/files/20080428_RL30006_d82c17a8245827846e913181fa76b504bd9ba61d.pdf.

concept for the forensic DNA community. *The FBI Quality Assurance Standards for Forensic DNA Testing Laboratories* (*QAS*) Standard 8.3.2.1 specifically requires laboratories to include samples with a range of the number of contributors, template amounts, and mixture ratios expected to be interpreted in casework. This benchmark has been included in the *QAS* dating back to at least 2011. In addition, SWGDAM and OSAC included this concept in their documents for validation of probabilistic genotyping systems[10] and mixture interpretation validation.[11]

With regard to factor space, you have acknowledged in your Profiles in DNA article from September 2006[12], a common misconception (urban legend #1) is that hundreds or thousands of samples are required to fully validate an instrument or method. Given the wide range of mixture ratios and template amounts evaluated in the assessments of published and internal validation data, there seems to be ample information from which the authors could have made a determination regarding foundational validity without seeming to require an exhaustive amount of data. For example, in the publication *Internal validation of STRmix™ – A multi laboratory response to PCAST*,[13] the data from thirty-one laboratories was compiled that contained over 1,500 three person mixtures that represented over 200 unique combinations and over 1,100 four person mixtures that represented over 100 unique combinations were evaluated. In each set of data, a portion of the mixture contained minor contributors of less than 1%.

Reliability

It is difficult to summarize the comments received on the draft report's treatment of the issue of reliability. There is considerable discussion of this issue but, in the end, our members were generally disappointed that there was no definitive guidance for the

[10] *SWGDAM Guidelines for the Validation of Probabilistic Genotyping Systems* available at https://www.swgdam.org/publications, and ANSI/ASB Standard 018, *Standard for Validation of Probabilistic Genotyping Systems* (2020); available at http://www.asbstandardsboard.org/wp-content/uploads/2020/07/018_Std_e1.pdf.

[11] ANSI/ASB Standard 020, *Standard for Validation Studies of DNA Mixtures, and Development and Verification of a Laboratory's Mixture Interpretation Protocol* (2018); available at https://asb.aafs.org/wp-content/uploads/2018/09/020_Std_e1.pdf.

[12] Butler, J. (2006) *Debunking Some Urban Legends Surrounding Validation Within the Forensic DNA Community*, Profiles in DNA; available at https://projects.nfstc.org/workshops/resources/literature/debunking%20validation%20butler.pdf.

[13] Bright, J. et al. (2018) *Internal validation of STRmix™ – A multi laboratory response to PCAST,* For. Sci. Int. Genet. 34:11-24.

forensic DNA community. Because this is a foundational review, there was an expectation that NIST would have suggested some minimum criteria or threshold for reliability or alternatively, the degree of variation that would be considered acceptable. The treatment of reliability appeared to some, rather circular, as reliability could not be established because of a lack of data which was a direct result of the decision by the authors to only review 'publicly available' data; a decision for which they attribute to their "belief" that such data should be 'publicly available' although internal validation data is not legally or otherwise required to be published or publicly accessible. Additionally, some members/invited guests were concerned about the inclusion of comments that appear to suggest that forensic DNA laboratories are not being transparent in their operations because of the failure to make their internal validation data 'publicly available'. Such an inference is both inappropriate and misleading and could have unintended and undesirable consequences for laboratories that are simply acting in accord with Federal and State legal requirements.

The draft report does not provide clarification on the issue of reliability for the DNA community and appears to have missed an opportunity to do so and provide guidance on gaps or other legitimate areas that would benefit from a review of all available internal validation data.

Conclusion

SWGDAM offers the following general suggestions for the draft report: (1) include the appropriate forensic science practitioners among the author team when performing a foundational study of forensic science disciplines; (2) define the threshold for establishing reliability prior to beginning such a review; (3) perform such a review with all available information and without arbitrarily limiting the information to that which is 'publicly available'; and (4) limit the report to only that information relevant to the foundational review. Specifically, Chapters 5 and 6 of the draft report contain interesting comments, observations and recommendations but may be more appropriate as a separate publication for forensic science practitioners as they are not directly relevant to the foundational review, nor mixture interpretation generally.

Thank you for the opportunity to comment on NISTIR 8351-DRAFT entitled *DNA Mixture Interpretation: A NIST Scientific Foundation Review*.

Sincerely,

Anthony J. Onorato
SWGDAM Chair

Attachment: SWGDAM Comments (August 2021)

DNA Mixture Interpretation: A Scientific Foundation Review

https://www.nist.gov/dna-mixture-interpretation-nist-scientific-foundation-review

SWGDAM Comments (August 2021)

Line #	Comment	Suggestions
General	NIST should have included practical experience in authorship: None of the authors have any experience working cases in an accredited crime laboratory. Authors with this critical experience should have been included. Having these individuals at your disposal for participation, however, not include them in your drafts for the last 2 years is not an appropriate use of stakeholder input. Example: OSAC on Forensic Science (NIST) has embedded stakeholders who are active in the writing of standards and have a vote on the outcome. Those with accredited forensic experience did not in the authorship of this document.	Include those with practical accredited forensic laboratory DNA mixture interpretation experience as authors of the report.
General	NIST should not restrict data to that publicly available: Determining that data had to be publicly available or it cannot be used is not an appropriate decision for establishing validity.	NIST should expand its definition of data that is considered to establish foundational validity. Included in that definition should be data held internally in forensic DNA crime laboratories that is available for on site review.
General	NIST should visit labs to review data on site: No data was reviewed on site.	NIST should personally visit accredited forensic DNA laboratories and consider firsthand observation of data as a valid means to assess foundational validity.
General	NIST should expand input to include auditors: No DNA lab auditors who review internal validation data were interviewed.	NIST should include opinions from DNA auditors who have reviewed validation data.
General	NIST should expand review of documentation: No lab audit documents were reviewed nor taken into consideration. Better disclaimer statements to prevent	NIST should include improved disclaimer statements to include concise language to the fact that NIST is not saying DNA mixture interpretation is not valid.

	misuse or misunderstanding (deliberate or otherwise) of the report: The lack of drawing a conclusion will lead some within an adversarial system to determine that complex DNA interpretation is not supported, or labs are not doing a thorough job in validations, or labs have something to hide. NIST's description regarding what they are not saying and how their report should not be used should be better spelled out.	
General	A review of published literature is not indicative of the reliability of work, when there is no stated requirement to publish internal validations, and in fact publication was discouraged. An example is the journal of the AAFS (editor Michael Peat) sent a letter telling scientists that internal validations would no longer be published (approximately 2005).	NIST should acknowledge that using only published data is an excessively restrictive measure to require for data and open its interpretation considerations to include forensic laboratories unpublished data.
Overall	There is minimal acknowledgement of QAS, SWGDAM, and OSAC validation requirements and guidelines. While these documents don't tell how to do an experiment, as the local scientists should design experiments to be relevant to their sample types, etc., they do provide common framework that is readily apparent to stakeholders and auditors.	These should be more readily acknowledged throughout (e.g., in Table 4.9).
Overall	This document seems to emphasize what's referred to as a lack of reliability and deemphasize 1. The challenges and inability to publish validation data, 2. The existence of reams of data within forensic laboratories, 3. The independent and external audit system to which all CODIS-associated crime labs are held, as well as other audits, such as ISO/ANAB.	The authors should include these overlooked and/or minimized but rather quite important and impacting points.

	Also noteworthy is the fact that the role of the criminalist is largely to perform casework and testify when needed. It is not often (with some exceptions) to vet, procure, validate, and publish on new technology, despite the efforts of many to nevertheless squeeze these additional tasks in.	
126-128	It would appear based on the summary from NIST 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, including its use of probabilistic genotyping software (PGS) systems." However, this statement in my view doesn't fully answer the original question – Did NIST find or not find that "established scientific laws and principles exist to support the methods" practitioners are currently using for DNA mixture interpretation? I understand there were gaps found from this study in terms of gathering sufficient empirical data from laboratories in applying these methods, but is NIST also asserting with this report that the scientific laws and principles also do not exist?	I would argue that there is indeed general consensus in the scientific forensic community that underlying scientific principles do exist and are reliable to use in DNA mixture interpretation, and that PGS systems using a likelihood ratio construct that apply these same principles also exist and are available for laboratories to utilize. In fact, these same principles and their application in PGS systems are well characterized in the peer-reviewed scientific literature and also described/referenced at length in this report (chapters 2 and 5). Although it is certainly important to note that improvements are needed particularly in the area of making laboratory internal validation of these methods more publicly available for independent review, this does not negate the fact that the methods do exist and contain the underlying scientific principles to interpret DNA data. I would also argue that empirical data also exists but is not necessarily publicly available. Just because NIST could not evaluate this data does not mean that this data does not exist or does not demonstrate "reliability" as defined in this report. NIST could not know that since they did not have the empirical data to evaluate in the first place. My suggestion here would be to address this part of the question in its summary to emphasize this distinction for full context so it is not misinterpreted by the rest of the forensic community.
609-611	This document doesn't clearly address this point. Rather it is vague and over-shadowed by Takeaway 4.4.	This needs to be flushed out.
610, 1028	Draft report indicates that the review does not concentrate on interpretation of single source or two person mixtures involving significant quantities of DNA. In order for the readers of the report to determine exactly what the scope of this document is, please define what are considered "significant quantities of DNA".	Define "significant quantities of DNA"
722-724	The authors state, "Note that our original goal in this review was	

	external and independent assessment of reliability based on publicly available data that met our selection criteria." Each forensic lab is required by the DNA ID Act to maintain the privacy of forensic samples, and donors of study samples also have an expectation of genetic privacy, so the "requirement" by the authors that genetic data be made publicly available for external review in order for an assessment of reliability to even be performed is not feasible. Moreover, requiring that a lab's PG data be made publicly available for external review ensures that the authors' definition of reliability will not be achieved. Publication in peer-reviewed journals has been the hallmark of an independent assessment of the validity and soundness of basic scientific research and method development. Moreover, the PCAST report called for peer-reviewed publication of the "foundational validity" of forensic methods and suggested "that NIST explore with one or more leading scientific journals the possibility of creating a process for rigorous review and online publication of important studies of foundational validity in forensic science." This document would benefit from expanding the "assessment of reliability" to include the publication of internal validations, developmental validations, and inter-laboratory studies.	
724; 3204; 3250; 605-611; 754-755; 3255	…that met our selection criteria What are these criteria? They are never stated. Instead, it is later said (3204) that there were no criteria. PCAST apparently had criteria but the authors are unclear	Clarify what the intention was. It is confusing (and circular) to say the criteria were changing and that there were no criteria for reliability. Were the authors attempting to establish national thresholds and criteria for universal application? Are the authors challenging the conclusions of PCAST? Confusing.

	as to what that was. Nevertheless, the authors claimed to be in agreement with PCAST (605-611). What about 3PMs, designated reliable by PCAST (3255)? Are these authors in agreement with PCAST on 3PMs?	
741-746	I agree with this recommendation, but the report falls short with how laboratories and where to publish its validation data and in what format? Keep in mind making this data public comes with certain privacy concerns, informed consent, or even IRB considerations with the actual genetic profiles that were collected/generated in-house.	I would suggest that a more practical approach would be for laboratories to participate in an inter-laboratory survey using a universal data set (perhaps generated at NIST) seems much more reasonable and these results could be published without the privacy concerns related to the data. It would also serve to evaluate a laboratory's application of its PGS method and associated interpretation guidelines that could then be compared against other laboratories or even other PGS systems.
741-746	How do we ensure the reviewers or "users" have the expertise for a rigorous scientific review of each laboratory's PGS validation studies in order to establish an acceptable level of reliability? What are the requirements to be considered an expert user to conduct these reviews?	I believe this knowledge and expertise is equally important for the "users" to possess as well as the "providers" who publish this data. This should be included in this takeaway.
748-755	The critique of the published studies so far in this report is that they "lack sufficient detail", but this report falls short in recommending what would be "sufficient". What specific criteria is NIST looking for? Until it knows what the requirements are, how can a laboratory or its stakeholders ever feel confident that it has met "an acceptable level of reliability"?	The relevant scientific forensic community needs to further research this topic and develop standard criteria based on peer reviewed consensus (e.g., OSAC) for laboratories to reference. This should be included in this key takeaway.
754-755	Statement that no threshold or criteria established to determine an acceptable level of reliability	Guidance documents on internal validations do provide information on how to assess for reliability and then each laboratory performs testing and develops an appropriate SOP based on the reliability shown within their lab. Remove no threshold and say established within the laboratory's internal validation studies.
771-773	Include validation performance results in case files and reports	This is unnecessary, and the reason validation is done and then even summarized so that it can be reviewed to include the data when an appropriate party is authorized to come onsite for viewing which is done when requested. The information is available to review and just putting summaries in case files would not truly answer the reliability question as it's being tied to the actual data and not just the results.

769-773	It is unclear in this key takeaway and throughout the report what role if any laboratory accrediting bodies have as it relates to reviewing validation studies as part of an external assessment? Was this intentional? Memorialization of laboratory internal validations through external assessments has always been the means for laboratories to demonstrate compliance using internationally recognized and accepted accreditation standards and requirements but also to receive the much needed independent review with data that is not published.	Is it possible for NIST to provide specialized training to external auditors or even give the accrediting bodies the "reliability" metrics needed so that they in turn could adequately provide the "independent review" of this data through the standardized process that is already in place with laboratory accreditation? The elements of this validation review process could be standardized for all labs using the reliability criteria and thresholds discussed in this document. Even though this review would not necessarily be made public initially, any gaps or deficiencies could be made public to the court as part of the routine discovery process where its reliability would matter most in a forensic context. Additionally, NIST could monitor any trends with these accreditation reviews to determine what gaps or improvements would be needed. This would serve as a more reasonable solution in my view in the short term. My suggestion here would be to at least mention the role of accreditation and external assessments which are well-positioned for this task so long as the auditors have the subject matter expertise to rigorously conduct these independent validation reviews.
849-852	Individual laboratories would need to know how the sensitivity of methods… This type of information is mainly disseminated through the literature but may also be presented at conferences and shared through communication exchanges between colleagues. Practitioners can use that information to draw comparisons and expectations.	It's difficult to determine if the authors feel this is unavailable or they are just reiterating the existing availability. Clarify.
870-871	Using likelihood ratio as a standalone number without context can be misleading	LR is a statistic used like random match and CPI so why would LR now be viewed so differently? Stats have been utilized in reports for over 25 years and explained in court and to our customers. Standard guidance language for LRs has been provided so I don't see the need for this takeaway.
1146-1147	Indicates that 'most laboratories do not publish data from their validation studies.' However, it does not mention that laboratories would find it difficult to publish their internal validation studies due to the fact that they are not novel work.	Acknowledge that laboratories may find it challenging to publish internal validation studies due to the fact that they are not novel work, and may not be accepted for publication into peer-reviewed journals as is covered later within Chapter 3.
1164-1166 and 1172-1173	NONE of the members of the review team have ANY hands on DNA mixture interpretation experience in an accredited Forensic DNA laboratory	The review team is missing a key perspective in which to review the issues and that is from a currently qualified DNA examiner with experience in mixture interpretation
1346-1347	"DNA information can assist both law enforcement (investigative)	Add defense/exculpatory items and mass disaster identifications to this statement about usefulness of DNA information.

	and prosecutorial (evaluative) aspects of the criminal justice system." Does not mention that DNA information can also assist defense or provide exculpatory information for an accused person, and be useful in the course of mass disaster identifications.	
1756	DNA analysis is not based on "belief" but on analysis and evaluation	When a DNA analyst evaluates a mixture and determines that a major component…
2123	Empirical data to assess fitness of purpose of analyst's LRs	This is foundationally set with validation and SOPs within a lab and then applied to information specific to each case. It is exists so warranted makes it sound as if it does not.
2123-2124 (KT 2.6)	The authors state, "Likelihood ratios are not measurements. There is no single, correct likelihood ratio (LR). Different individuals and/or PGS systems often assign different LR values when presented with the same evidence because they base their judgment on different kits, protocols, models, assumptions, or computational algorithms. Empirical data for assessing the fitness for purpose of an analyst's LR are therefore warranted."	The implication is that since different LRs may be obtained from the same evidence, that the LR is not reliable since there is not one "correct" value. Peak heights are measurements. If the same sample is injected multiple times within the same lab, or in different labs, the peak height will not be the same. This is because of a variety of factors that impact peak height, however it doesn't make the peak height unreliable. Similarly, LRs determined from the same evidence will also differ within the same lab, or in different labs, due to a variety of factors (assumptions, kits, models, etc…) and this also doesn't automatically make the LR value unreliable. While there is certainly a need for additional inter-laboratory studies beyond the "multi-lab response to PCAST", this document would benefit from the acknowledgement of peer-reviewed publications of validation data that demonstrate how the LR is dependent on a laboratory's procedures/assumptions and how these studies demonstrated the reliability of an LR value.
2352	"These propositions H1 and H2 are required to be mutually exclusive and exhaustive." I don't believe that there are any requirements for a single H1 and H2 to be completely exhaustive. There may be multiple H1 or H2 propositions that could explain the evidence, resulting in multiple LRs to be run, and none may be completely exhaustive of all potential propositions	Remove the phrasing 'and exhaustive.'
2402-2405	Similar to 1146-1147, there is no mention that laboratories may have difficulty publishing internal validation data at this point in the document. In addition, there is no mention that NIST did not attempt	Remove these lines as the sentiment is already mentioned in other areas of the document. Mention in the previous paragraph 2399-2400 that NIST made no effort to solicit labs for data.

	to request data from laboratories, and that they only used web searches to deem what may have been 'available'.	
2486-2488	Authors are acknowledging that additional internal validation data likely exists, but they chose to conduct the scientific foundation review using only publicly available information.	Choosing to not even try to obtain the data to look at (during the public webinar, it was asked how many of the authors of the 60 prob gen publications were contacted to see if the data was available to review, and Dr. Butler replied none of them had been contacted). This is an irresponsible approach to conducting a scientific foundation review and is doing a dis-service to forensic science.
2487-2488	Should mention here that NIST did not make an attempt to solicit or evaluate laboratory data.	This scientific foundation review is limited to publicly available information, and no effort was made by NIST to request the underlying data from any laboratories.'
2490; Table 3.2	DC Dept of Forensic Sciences STRmix v2.3 validation is incorrectly stated as being ABI 3500. Samples were processed on the 3130xl. See page 1 of STRmix v2.3 Parameters report.	Identifiler Plus, ABI 3130xl
2707-2710	"In recent years, DNA analysts have increasingly relied on one of several available PGS systems to assign a numerical value to their mixture result based on a pair of propositions selected by the analyst." Should mention here similar to 2770-2772 that the assessment using PGS is used in conjunction with the analyst's interpretation and training and experience.	Indicate that PGS is used in conjunction with the analyst's interpretation and training as part of the laboratory's protocols.
2848	"factor space" and "factor space coverage" – these terms have never been used in discussing validation studies, why are they being introduced in a review publication.	
2901-2903	…an eight-person mixture involving only 10 pg total template DNA, then DNA analysts might refrain from interpreting such a sample because it has not been covered in any of their validation experiments This is an extreme, far-fetched example well beyond where forensic laboratories would test reliability.	Change to something more realistic, like 5PM with 100pg.

2922	Degree of reliability is assessed through empirical data	The data is made available by laboratories as requested through appropriate channels. If the entire foundation study was going to summed up by saying we just don't know the reliability of mixture interpretation due to the lack of publicly available data, then when did this become the way to measure and assess? NIST provided trainings on validations for years and never did I hear the major takeaway be that the data must be made publicly available to truly assess reliability. This seems so disingenuous to state when going into the foundational review, the writers knew the current state of validations and how they were handled as they conducted many workshops on the very topic over the years. So, make it available publicly and then who is reviewing it? What knowledge base does that individual/group contain to comment in a useful manner on reliability? It's important for individuals to be educated on the entire process and not just mark points of a plot for each mixture to cover space.
2956	Proficiency tests really don't ask how reliable or trustworthy a method is, except perhaps on a very basic level. Reliability is addressed in internal validation, before proficiency tests on a given method are initiated. Proficiency tests are meant to address analyst competency. When an analyst has an issue with a proficiency test, that is addressed directly with the analyst. It's possible that there could be an improvement to a written procedure to address the root cause of the issue. However, it's generally unlikely that such a situation is due to an unreliable method. This source of data seems minimally helpful to examine foundational reliability of a technology.	The authors should instead seek data that challenges the systems, e.g., internal validation data from forensic laboratories.
2968	Why did the reviewers limit their analysis to publicly available data? It would be very rare for laboratories to make their data publicly available -	Re-do chapter four after looking at data now available. How can it be considered a review if sufficient data was not looked at.
2983-2986	This is indeed the case but seems to be underemphasized throughout the document and is at the heart of the issue here – the review was NOT conducted b/c the committee did not have internal validation data from forensic laboratories.	Work with forensic laboratories to collect data to perform the review.
3021-3022	"a great deal more information is now available" implies that this	Remove sentence

	information was not available at the time of PCAST and that the 8 articles cited in PCAST were the only ones; however, 18 of the articles in table 4.3 were published prior to PCAST	
3069; Table 4.5	DC Dept of Forensic Sciences STRmix v2.3 validation is incorrectly stated as being ABI 3500. Samples were processed on the 3130xl. See page 1 of STRmix v2.3 Parameters report. Table 4.5 incorrectly states that mixture ratios were not explicitly stated in the summary document for the STRmix v2.3 validation. See page 9 of the STRmix v2.3 Validation summary where it states that a summary of the profiles analyzed for the sensitivity and specificity plots are in Appendix 3. Appendix 3 starts on page 38 of same document.	Identifiler Plus, ABI 3130xl For sensitivity and specificity studies, i.e., Section D studies: 17 single source, various DNA quantities (quantities listed in Table 4.5 are additional single source samples from Section A and Section B studies) 2 person mixture ratios: correctly listed in Table 4.5 3 person mixture ratios: 5:2.5:1, 20:1:1, 3:1:1, 20:10:1, 3:1.5:1, 10:1:1, 10:5:1, 5:1:1 4 person mixture ratios: 5:5:1:1, 10:5:2:1, 2:2:2:1, 10:1:1:1, 5:5:5:1, 1:1:1:1, 10:5:5:2, 5:2:2:1, 10:10:1:1, 5:2:1:1, 2:1:1:1, 10:5:5:5, 3:1:1:1, 10:10:10:1, 2:2:1:1, 5:1:1:1, 3:2:1:1, 3:2:2:1,
3069; Table 4.5	Table incorrectly states that total DNA quantity, and mixture ratios were not explicitly stated in the summary document for the DC Dept of Forensic Sciences STRmix v2.4 validation. See page 7 of the STRmix v2.4 Validation summary which states that a summary of the profiles analyzed for the sensitivity and specificity plots are in Appendix 3. Appendix 3 starts on page 37 of same document.	For sensitivity and specificity studies, i.e., Section D studies: 2 person mixtures: DNA quantities 300 and 600pg 2 person mixture ratios: 1:1, 1:2, 1:3, 1:5, 1:7, 1:10, 1:15, 1:20, 1:25 3 person mixtures: DNA quantities 200, 500, 900 pg 3 person mixture ratios: 3:1:1, 20:10:1, 3:2:1, 10:5:1, 5:1:1, 10:2:1 4 person mixtures: DNA quantities 100, 200, 400, 600, 700, 800, 900, 1000pg 4 person mixture ratios: 2:2:2:1, 20:5:2:1, 5:1:1:1, 5:2:1:1, 5:5:5:1, 4:3:2:1, 3:3:2:1, 10:5:3:1, 2:2:1:1, 20:10:1:1, 3:1:1:1, 7:1:1:1 5 person mixtures: DNA quantities 300, 600, 1000pg 5 person mixture ratios: 10:5:2:1:1, 5:4:3:2:1, 10:10:10:10:1, 10:10:5:1:1, 5:5:5:2:2, 20:1:1:1:1, 2:2:2:1:1, 3:1:1:1:1, 5:1:1:1:1
3069	Reference Table 4.5 (Line 3069) Factor space coverage of information in internal validation studies listed in Table 3.2. • The Michigan State Police has effectively covered much of the "factor space" recommended by this report, but not all of that work was publicly available. The report should be corrected to reflect the actual	Coverage of Factor Space from Validation: STRmix™ and PowerPlex® Fusion

factor space covered. The profiles outlined here are all lab-created samples- these charts do not contain the additional testing conducted on adjudicated samples.

C Range	# Samples	Total DNA Quantity Range (pg)	Mixture Ratio Range
1	6	500	N/A
		600	N/A
		150	N/A
		75	N/A
		50	N/A
		25	N/A
2	18	500:500	1:1
		909:91	10:1
		882:118	7.5:1
		833:167	5:1
		714:286	2.5:1
		500:500	1:1
		714:286	2.5:1
		2,143:857	2.5:1
		909:91	10:1
		882:118	7.5:1
		833:167	5:1
		714:286	2.5:1
		500:500	1:1
		909:91	10:1
		882:118	7.5:1
		833:167	5:1
		714:286	2.5:1
		500:500	1:1
3	22	625:312.5:62.5	10:5:1
		833:83:83	10:1:1
		769:154:77	10:2:1
		625:312.5:62.5	10:5:1
		476:476:48	10:10:1
		454.5:454.5:91	10:10:2 [5:5:1]
		400:400:200	10:10:5 [2:2:1]
		333:333:333	10:10:10 [1:1:1]
		500:334:167	3:2:1
		351:234:117	3:2:1
		234:156:78	3:2:1
		174:116:58	3:2:1
		78:52:26	3:2:1
		833:83:83	10:1:1
		740:185:74	10:2.5:1
		625:312.5:62.5	10:5:1
		540:405:54	10:7.5:1
		476:476:48	10:10:1
		444:444:111	10:10:2.5 [4:4:1]
		400:400:200	10:10:5 [2:2:1]
		364:364:272	10:10:7.5 [4:4:3]
		333:333:333	10:10:10 [1:1:1]
4	19	588:294:59:59	10:5:1:1
		769:77:77:77	10:1:1:1
		588:294:59:59	10:5:1:1
		385:192:38	10:10:5:1
		468:351:234:117	4:3:2:1
		312:234:156:78	4:3:2:1
		232:174:116:58	4:3:2:1
		104:78:52:26	4:3:2:1
		769:77:77:77	10:1:1:1
		769:77:77:77	10:1:1:1
		714:143:71:71	10:2:1:1
		588:294:59:59	10:5:1:1
		455:455:45:45	10:10:1:1
		435:435:87:43	10:10:2:1
		384:384:192:38	10:10:5:1
		323:323:323:32	10:10:10:1
		312.5:312.5:312.5:62.5	10:10:10:2 [5:5:5:1]
		286:286:286:143	10:10:10:5 [2:2:2:1]
		250:250:250:250	10:10:10:10 [1:1:1:1]

Coverage of Factor Space from Validation: Use of STRmix™ Software for Profile Interpretation of Challenging Samples

<table>
<tr>
<td></td>
<td></td>
<td>
<table>
<tr><th>NoC Range</th><th># Samples</th><th>Total DNA Quantity Range (pg)</th><th>Mixture Ratio Range</th></tr>
<tr><td>1</td><td>31 (Degraded)</td><td>500</td><td>N/A</td></tr>
<tr><td rowspan="3">2</td><td>4 (Biological Relatives)</td><td>250:250</td><td>1:1</td></tr>
<tr><td>4 (Biological Relatives)</td><td>400:100</td><td>4:1</td></tr>
<tr><td>4 (Biological Relatives)</td><td>455:45</td><td>10:1</td></tr>
<tr><td rowspan="8">3</td><td>4 (Degraded)</td><td>167:167:167</td><td>1:1:1</td></tr>
<tr><td>6 (Degraded)</td><td>357:71:71</td><td>5:1:1</td></tr>
<tr><td>2 (Low-level, Heterozygosity)</td><td>385:96:19</td><td>20:5:1</td></tr>
<tr><td>2 (Low-level, Heterozygosity)</td><td>454:23:23</td><td>20:1:1</td></tr>
<tr><td>3 (Profile Rarity)</td><td>385:96:19</td><td>20:5:1</td></tr>
<tr><td>3 (Profile Rarity)</td><td>454:23:23</td><td>20:1:1</td></tr>
<tr><td>4 (Biological Relatives)</td><td>167:167:167</td><td>1:1:1</td></tr>
<tr><td>4 (Biological Relatives)</td><td>333:83:83</td><td>4:1:1</td></tr>
<tr><td>4 (Biological Relatives)</td><td>313:156:31</td><td>10:5:1</td></tr>
<tr><td rowspan="4">4</td><td>2 (Low-level, Heterozygosity)</td><td>370:93:19:19</td><td>20:5:1:1</td></tr>
<tr><td>2 (Low-level, Heterozygosity)</td><td>434:22:22:22</td><td>20:1:1:1</td></tr>
<tr><td>3 (Profile Rarity)</td><td>370:93:19:19</td><td>20:5:1:1</td></tr>
<tr><td>3 (Profile Rarity)</td><td>434:22:22:22</td><td>20:1:1:1</td></tr>
</table>
</td>
</tr>
<tr>
<td>3074</td>
<td>Not enough publicly available data…</td>
<td>Who determined that reliability is only able to be assessed with publicly available data? To embark on a foundational study in this arena knowing that it was not standard practice in labs to make all their validation data publicly available is ridiculous quite frankly. Why would discussions not have occurred for many years (all their trainings/workshops offered) on the need to put data in a public forum? The labs are being held to criteria that seems nearly just invented for the direct purpose to dismiss PGS.</td>
</tr>
<tr>
<td>3074 and 3425</td>
<td>- Consider that within the factor space, which is very large, there is a continuum of data from the more reliable to the less reliable.
- Despite this size and complexity, there are components of the factor space that are reliable, even within many very complex mixtures. Within a complex mixture, components of that mixture can be reliable, while other components are not.</td>
<td>Acknowledge that DNA mixture interpretation is reliable for those mixtures with a major contributor above a certain ratio and those mixtures where all of a potential contributor's alleles are present above a certain threshold. Some component of the factor space, where mixtures are above a certain threshold, support DNA mixture interpretation that is reproductible and hence valid.</td>
</tr>
</table>

	- That among the more reliable data (good peak height, quality, etc.) that the major contributors and those with the greater reliability can be determined, not undermined by the minor contributors with less reliable data. - There is a component of the factor space that is reliable. In my opinion that is: o Major contributor: Major contributors for complex mixtures where there is sufficient separation of peak height o All alleles present: Inclusion with a statistic for validated systems where all peaks above a predetermined threshold are present (no drop out, all peaks from a person's profile are represented in a mixture) o Less weight or perhaps not determined (inconclusive) for the far more difficult mixtures, where some of the individual's profile is not represented (drop out, degradation, below threshold, etc.) - By declaring at least some of the factor space reliable, this acknowledges the obvious, while pinpointing the areas where additional work needs to be done. At present, the conclusion of nothing has sufficient data available makes it appear a thorough job has not been done, as the "Major contributor" component with sufficient peak height, and the "all alleles present" are included examples above have not been acknowledged.	
3135	Using CTS PT results	It should be noted that students and possibly other untrained individuals may participate in the PT so it's not an appropriate measure to gather the results as a whole and make any assessment
3201-3204	Again, a mention of lack of publicly available data, but no mention that NIST did not reach	Be clear about whether NIST would be able to be determined as the 'user' who could establish a degree of reliability if more data was made public as suggested.

	out to laboratories to obtain data. Would NIST have been able to perform an assessment of reliability if the data had been made available? Would NIST be considered the 'user' in this sense to be able to assess the degree of reliability, validity, and whether that information is fit-for purpose, in line with key takeaway 4.2?	
3204-3207, and within 3425 (KT 4.4.)	"This is particularly true without an established and accepted criteria for reliability with complex mixtures involving contributors containing low quantities of DNA template or where there is a high degree of allele overlap among contributors" From the key takeaway 4.4 "...there is no threshold or criteria established to determine what is an acceptable level of reliability." There is no suggestion of who would/could create such criteria for reliability. Or that laboratories that have evaluated and empirically tested their data have determined reliability within their own factor space and are applying it accordingly. This is also somewhat in conflict with key takeaway 4.1 where it says that "The degree of reliability of a component or a system can be assessed using empirical data obtained through validation studies, interlaboratory studies, and proficiency tests." The phrasing also has the potential to have a detrimental effect as commentary in relation to admissibility hearings. There is a prong of evaluation of error rates and reliability that is part the Daubert standard in determining whether a particular scientific technique (like PGS, which is still being evaluated in many	Add recognition that laboratories have established reliability as a user of PGS within their laboratories by performing internal validation studies on empirical data. This would be in line with key takeaway 4.1 from this chapter.

	jurisdictions). This makes it sounds as if there is no standard or even suggestion of how a laboratory would determine reliability. Despite recommendations from ISFG and guidelines from SWGDAM and others that state that a laboratory's internal validation is recommended in order to evaluate a PGS software in order to determine if it is reliable for application to casework.	
3209	Bracketing approach and factor space	This is a foundational study so I would not expect terms such as bracketing and factor space to be introduced here but would be better suited in a document for future improvements/studies. Use the terms that are established in our community and widespread b/c it's a foundational look. Bracketing is done in our validations but maybe the term was not directly named as such.
3222; 3541-3543;3580-3582	Reliability of a specific LR number… A specific LR is specific to the population databases used, the propositions, theta, etc. Approaches involving sampling (e.g., HPD) will not give you the same exact number either. The emphasis, as has been the case, should be on trends and variance among data with repeated tests. Statistics are estimates. Just because we would get the same number with repeated RMP or CPI tests, holding the approach and databases constant, doesn't make those numbers the true and right. They are all estimates and conveyed as such to the legal community. Trends establish expectations. This includes understanding false neg and false pos behavior. This is something explored by forensic laboratories in validation and additional studies, but this report seems to purport it as overlooked.	Per webinar given by Dr. Butler on July 21, 2021, LRs will be the subject of a future review. Perhaps these sections should be removed and addressed later as this topic seems incompletely vetted and explained in this document.
3227-3229; 3479-3482	1. These data are available in individual forensic laboratory validations and are frequently explained/addressed in publications as well (7486, Bright	Update to convey correct information. Add: Bright, Stevenson, Curran, Buckleton FSIG 2015 14:187 Bright, Turkington, Buckleton FSIG 2010 4:111

	et al FSIG 2015, Bright et al FSIG 2010 – latter two should be added). Without having examined actual internal validations, the authors seem unaware. 2. Defense experts are provided all data for independent review.	
3235-3237	LR results can be externally and independently evaluated to be reliable… By other laboratories generating their own data to test. When numerous laboratories, independent of one another, generate their own data and find LRs to be a reliable statistic, that provides greater strength to speak to reliability than a lab reanalyzing another laboratory's work. This work is shared and discussed at conferences and various other meetings and when novel, published.	Update to convey correct information.
3235-3236	States that "LR results cannot be externally and independently demonstrated to be reliable without access to underlying performance data." The report makes no suggestion as to what bodies would be available to do this external or independent reliability assessment.	Provide a suggestion, would NIST be able to do these independent reviews? Is there a suggestion that funding be provided to institutions to be able to do this type of assessment if data was made publicly available?
3237-3241	Later in the paragraph it mentions that "To establish and support clear reliability boundaries, data need to be available to users of the information (e.g. DNA analyst or stakeholders using their results) and acceptable levels of reliability must be decided upon by the user." This statement seems to imply that these users mentioned (DNA analysts and stakeholders) do not have access to data that helps to determine reliability. However, DNA analysts and stakeholders routinely have access to internal validation data in order to evaluate the data in relation to their case-specific results. This should be pointed out.	It should be mentioned that DNA analysts routinely have access to internal validation data through the course of training and working in a laboratory and stakeholders routinely receive access to internal validation data for review by their own experts through the course of discovery requests during court proceedings.

3237-3241	To establish and support clear reliability boundaries… Forensic laboratories generate and assess data sets addressing such questions. These are obviously the "users" of the information as well as stakeholders involved in the court system. All have access. Levels of reliability are established prior to implementation (verbal scales (SWGDAM), interpretation procedures, policies, etc.).	Update to correctly convey such information.
3264-3270	Enable a user to scrutinize…what is actually happening in casework settings. 1. Stakeholders have access to all the validation data they request (e.g., defense community through discovery) 2. If NIST is a "user," tasked by Congress to assess reliability, then NIST should have asked forensic laboratories for data. 3. Who else is a "user"? The authors seem to be creating an issue that there is an issue of access to "anyone." Who is anyone and why does anyone need access? 4. The language "what is actually happening in casework settings" seems inflammatory and to insinuate there is a concealed problem. 5. Here, there seems to be an inherent contradiction as published data are stated in lines 3217-3232 to not meet the authors' criteria for testing.	Who are the users? State who and why they need access. The raw data contain personal genetic information not for public viewing; this is protected by federal and state laws such as GINA and HIPAA. Furthermore, the data require purchased specialized software and expertise to process and understand. To actually perform the foundational review, the authors should ask forensic laboratories for what they need. Point 4 – this unnecessarily negative language that should be removed. Point 5 – clarify.
3288-3294; 3369-3370	This data exists in numerous forensic laboratories across the US. The catch-22 here is that it is not publishable, and therefore not publicly available in its various forms.	It seems the authors could arrange to have access with agreements to treat individual profiles confidentially (to protect the privacy of the study donors) to complete the Congressional task of a foundation review.
3302	Locating… The data exists and is in forensic laboratories.	This report doesn't address that it has located the vast data across the country. Rather, it states that it wasn't publicly available. If the study is to be conducted, the authors should simply organize efforts to obtain it. Otherwise, the congressional task has not been performed.

3336-3350	This paragraph is an oversimplification. A publication addressing sensitivity is not addressing mixture deconvolution. Unique alleles are only one aspect of mixture interpretation and represent a simple way to convey sensitivity in a publication. This is not intended to be used as a measure of deconvolution. While it may be true that publications on comparing old versus new system sensitivities are lacking, this is not necessary nor novel/publishable information. Typically, forensic laboratories make direct comparisons when needed to their own old versus new approaches, especially as related to the FBI QAS standards.	Delete entire paragraph. There is not helpful and does not move the field forward.
3367-3369; 3421-3422	Summaries are simply that. Here summaries are criticized for representing exactly what they intend to – a summary. Furthermore, the purpose of the summaries is associated with the FBI QAS. This drives the purpose and focus. The stakeholders are the laboratory staff, auditors, and the courts, all of whom have access to the actual validation.	This seems nonsensical. A summary is a summary. If there is a vision for a different layout that would result in some meaningful benefit to a named stakeholder (more details on "factor space"), that has not been shared in this document. Rather, simply providing a criticism seems unproductive and doesn't move the field forward.
3369-3370	Lack of availability of data	I feel like a broken record, but it is continually mentioned in this document. The data is made available when requested within the proper channels. Exactly who is doing these independent assessments? Have these individuals performed the testing and do they understand then how the laboratory decided to implement the data appropriately within their SOPs so it's not just simply an external data review.
3374-3375	Given that the authors didn't actually obtain any validation data (only a handful of documents summarizing work), it seems presumptuous to say that allele sharing is missing from many validations. In fact, the authors have no data to indicate one way or the other.	This statement is fact less and should be removed.
3412-3413	Report incorrectly states that DNA quantities and mixture ratios	See information in box above, which is included in the DC DFS STRmix v2.4 validation summary

	were not stated in the public documents	
3425	NIST states "there is no threshold or criteria established to determine what is an acceptable level of reliability."	NIST should have determined what reliability was, rather than leaving it to someone else. If they are not in a position to say what reliability for DNA mixture interpretation is, who is?
3425	Takeaway 4.4 - There is no threshold or criteria … The authors really don't know this without having seen any validation data and corresponding interpretation criteria of various forensic laboratories. For example, this statement is factually incorrect considering my laboratory.	These conclusions are invalid. Collect the data and evaluate before drawing fact-based conclusions.
3425	Takeaway 4.4 - Publicly available information lack… to independently assess… Reliability of foundational concepts is not typically assessed in this manner in other scientific areas. While this is changing for big data (MPS sequencing in certain fields), it has not been the case to "reanalyze" someone else's data. Rather, especially with peer-reviewed published results, it is up to the "user" to generate their own data to assess reproducibility, etc. A publication contains enough information to enable this long established practice for independent verification. Rather, the authors here seem to be interested in seeing whether they agree with each individual lab's nuances to the basic scientific approach. This is impractical and unnecessary, as this is the role of the courts (via discovery, defense experts, etc.). Furthermore, while this committee may have been unable to "publicly" access data, it is regularly shared through discovery requests.	Authors should address standard practices in place and why they feel this must be handled differently.
3439	Call for collaborative approach	Laboratories routinely share validation studies with each other. The reference paper is going a step further and attempting to streamline each

		lab's validation work by being able to possibly adopt the validation done by a lab and then run more of a performance check (dependent on what is being adopted). Data is shared routinely with labs and it's written as if making them public is the only avenue to have this effectively happen.
3441	Critically assessed by other scientists	Audits, outside experts are ways in which this is done and has been done for many years
3447	Footnote 23 1. Many forensic laboratories don't have the infrastructure to have websites, let alone manage large data storage in that format. 2. More importantly, it is a genetic privacy (GINA/HIPPA/etc.) violation that is brushed off in this document. 3. "credible parties in a timely manner when requested" sounds like there is a back story not being told here.	This footnote does not provide any helpful information and does not contribute to moving the field forward. If it cannot be reworded to avoid any negative implications, it should be removed.
3454	Table 4.9 - On PGS validations, collection of common data to demonstrate performance and ultimately reliability can actually be accomplished as Bright et al have shown.	The authors should address this.
3454	Table 4.9 - On recommendation for internal validation data sharing. This is a reasonable suggestion if it can be done in a way that complies with privacy laws.	NIST should vet and present this idea further to provide a practical solution to "move the field forward." For example, a CODIS-controlled site.
3454	Table 4.9 - On recommendation for more challenging proficiency testing. This is a reasonable recommendation but may not be practical for proficiency testing companies.	NIST should vet and present this idea further to provide a practical solution to "move the field forward." For example, NIST partner with the private companies to provide a model that allows for reproducibility across all samples in a lot and fair and expedient scoring of expected results. Seems like something NIST could play a major role in and is currently a missed opportunity.
3455	2nd recommendation	Constant need to publish data and have an independent assessment of PGS performance. Who is doing this assessment?
3458-3459 (KT 4.5)	The authors are suggesting that more complex and/or low-template components be used in creating proficiency tests. This reminds me of the "blind proficiency tests" suggestion from years ago. In theory, it's a good idea, but in reality the preparation (for consistency to distribute to the many test takers) and the "scoring" of this type of	If a suggestion like this is to be included in the document, then the authors need to define how these tests will be made, distributed and graded.

	proficiency test make it very difficult to conduct.	
3461-3462 (KT 4.6)	Stating that "improvements across the entire community are expected with an increased understanding of the causes of variability among laboratories and analysts" without giving guidance on how this is to be accomplished	
3474	Bracketing may be needed… In fact, laboratories already do this but rather than call it "bracketing of factor space," we record it in internal validation and laboratory procedures as limitations, policies, etc. This is a standards requirement! The authors may be unaware having not examined any of this data.	Work with forensic laboratories to understand the data, procedure, and policies in place. Then form opinions based on the facts and update this document to reflect the reality of such work.
3484-3486	Stating here that the determination of whether PGS systems are reliable depends on the coverage of factor space for that particular case sample of interest and coverage of the ground truth for assessing reliability. No mention of the fact that laboratories are doing this as part of their own internal validation. And also demonstrated by numerous admissibility hearings or trials where users (stakeholders and their own DNA experts) have also had a chance to evaluate and argue about the degree of reliability in a particular case based on a laboratory's internal validation data.	Suggest to add to the end of this paragraph that "Internal validation studies performed by laboratories allow the users of the case data (DNA analysts, stakeholders such as attorneys and hired DNA experts) to evaluate reliability in relation to their case samples in comparison to ground truth data.
3487	Takeaway 4.7 - Helpful to include these validation performance results in the case file and report… 1. Courts already have access to such information through the discovery process, 2. Law enforcement is largely not interested in such information but rather the bottom line and just a summary rather than the full reports currently provided,	This idea seems poorly considered and impractical and unhelpful. Furthermore, it does not directly tie into the proceeding text. This should be reconsidered and the authors should interview those working in the field to vet the topic.

	3. Case files include case-pertinent information and all other information (validation, instrument and reagent quality control, proficiency testing, corrective action, etc.) is available separately for various reasons including efficiency, and 4. Not all validation data is relevant to a case. This is impractical, unnecessary, and would negatively impact crime lab efficiencies. Given that <10% of cases even go to court, this is particularly illogical.	
3487-3488 (KT 4.7)	"Including validation performance results in the case file and report" – it is obvious that the authors do not work in a forensic DNA laboratory and know who the "customer" is that receives these reports. To include this information in a DNA laboratory report is absurd.	If the "user" needs this information to understand the report, they can request it in discovery.
3487 (KT 4.7)	This mentions that "The degree of reliability of a PGS system when interpreting a DNA mixture can be judged based on validation studies using known samples that are similar in complexity to the sample in the case. To enable users of results to assess the degree of reliability in the case of interest, it would be helpful to include these validation performance results in the case file and report." There is no suggestion or example given as to what this would look like, and how easy this would even be for stakeholders to understand, by taking a complex topic and adding additional complex information to a DNA report. Part of the role of the expert witness is to be able to explain and put into context the results from the testing. Putting in more complex validation results into a report may make it even	An example should be provided as to what is meant by putting validation performance results within a report. Is a link to publicly available validation summaries sufficient? Discovery material available to the receiver of the report that consists of validation summaries or underlying data?

	more challenging for a reader of a DNA report to understand the results as they are. In addition, laboratories are already reviewing their internal validation data in order to ensure that their reported results accurately reflect the data that was tested within a particular case.	
Lines 3614-3616 and Box 4.1, page 94	Overall, the practical considerations for how a laboratory would publicly share validation information and implement Box 4.1 for independent review/assessments does not appear to have been discussed by the report authors. Where are labs posting the data? Who is reviewing the data for the independent assessment? How much time does a lab give to an independent reviewer? Is this review being performed before or after the new technology is implemented in casework? Etc.	Authors should provide a method or roadmap for this process to be implemented. Otherwise, Box 4.1 reads as a task-driven checklist of data to be provided by labs, with no follow-up steps.
3727-3732	This paragraph is confusing in relation to how published data and studies are helping to determine system reliability. Previous recommendations (like PCAST as referenced in this report) as well as various standards and guidelines nationally and internationally state that the underlying components should be published. But this statement seems to undermine their value. In addition, and as referenced in this report, there are many publications that cover actual system reliability and not just component-level reliability. Both are important, and both have been published, to an extent where the authors of this report had difficulty keeping up with the amount of publications. In the last statement that states 'there is a danger of inadvertently viewing results from narrowly-	There should be recognition here that there are a large number of system reliability published studies, including numerous publications and internal validation studies, as referenced in this report. In addition, if the last statement is kept in the report it should be clarified and supported as to who is at risk for the dangers mentioned in the last sentence of this paragraph.

	focused studies as applicable to system reliability' – is their evidence that this is how these publications are being interpreted by someone? It is important to study each component and for analysts to understand both PGS systems both at the component and system levels.	
3740-3743	"Regardless of sources of uncertainty and complexity of the samples, reliability of a PGS system boils down to checking its calibration accuracy and discriminating power at every conceivable scenario described by the factor space." – stating that the reliability needs to be evaluated at 'every conceivable scenario' seems to go against the concept of bracketing the desired factor space as recommended earlier within the report. This sets a much higher threshold for being able to determine reliability.	This should be clarified as to how this is in line with the bracketing approach mentioned earlier in the report.
3742	Every conceivable scenario described by the factor space	This seems to go against the bracketing approach. Every conceivable scenario is unrealistic and not necessary to cover proper usage for casework. Establish a proper foundation/framework with limitations to implement SOPs and then they are modified over time as needed.
3765-3772 (Chapter 5)	Points 2, 3, and 5 – While we have much in the way of published studies and internal validation studies to offer and address such questions, this document fails to point out the lack of control a criminalist actually has over this. While a criminalist will "do their best" to ensure the information is not misused, the court system doesn't support this level of involvement in many instances.	The consideration of the relevance of the court process throughout the document is underwhelming. This is a significant miss as the courts dominate the role and impact of the practitioner. So while the points made are those which one strives to maintain, the adversarial counterbalance pulling at and restricting the practitioner is unaddressed.
3785; Chapter 5	Inadequate to consider only a single trace in isolation	Who is making this determination? It's truly case dependent on what is tested and there are so many factors that may affect multiple traces being detected so then single is just inadequate?
3802; Chapter 5	Relevance of a DNA sample to the crime is often difficult to discern	Asking forensically relevant questions when assessing the evidence to exam is key and done by labs. DNA evidence is a piece of the puzzle and not the puzzle. I completely disagree with saying often difficult to discern. The question it may answer is known before choosing to test the item so it's really the weight of it that is part of our judicial system to determine.

Chapter 5	The main focus of this review is on the application of LRs to DNA mixtures using PG software.	Chapter 5 shifts the focus of this review to DNA transfer and persistence. The topics discussed in Chapter 5 are relevant to the interpretation of all DNA data (single source and mixtures) and are not specific to DNA mixture interpretation. The inclusion of DNA transfer/persistence in this review conflates the issues of reliability and relevance as they relate to DNA mixture interpretation. The concept of DNA transfer/persistence is important, and the community would benefit from a "stand-alone" review of this subject matter.
4546 (Chapter 5)	Takeaway 5.2 - It is especially important to consider relevance… This is the job of the court system (the lawyers). The criminalists inform the lawyers and do assist the court with whatever scientific knowledge is available to shed light. While practitioners attempt to control /stop the transition from who to how/when (4639), court testimony is a restrictive framework limiting the expert. Furthermore, this seems to be in direct conflict with recommendations from Dr. Itel Dror on Cognitive Bias (https://www.ucl.ac.uk/~ucjtidr/).	The consideration of the relevance of the court process throughout the document is underwhelming. This is a significant miss as the courts dominate the role and impact of the practitioner. So while the points made are those which one strives to maintain, the adversarial counterbalance pulling at and restricting the practitioner is unaddressed.
4646 (Chapter 5)	Takeaway 5.4 - Without context… Is vague. Do you mean without propositions? Agreed. Or something else like case context? Disagree. Not directly a criminalist's role. The crim will inform the courts on scientific questions but the attorneys provide case context.	Clarify.
4647; Chapter 5	LR as a standalone number without context can be misleading	Statistics have been reported and never implied info on how DNA was transferred or if it's relevant to the case so why is this an issue with LRs?
Chapter 5	Although there is a lot of interesting information in this chapter, I am not sure it should be in a foundational scientific review on mixture interpretation.	
4468 (KT 5.4)	This takeaway is not limited to LRs. RMP and CPI do not provide information about how or when DNA was transferred.	If this takeaway is included, it should be more transparent to clarify that it applies to all DNA statistical results.
4682	Example is sub-source level proposition	"An example of a sub-source level proposition might be that the DNA mixture contains DNA from the POI and the victim."
4777-4784, and	The challenges of efficiency and throughput for DNA laboratories	More emphasis or examples could be put into this report in relation to how a lab could address this within reports, particularly if time constrained by

referencing 34770-4773 and KT 5.4 (4646)	is mentioned here, as well as that they are not always aware of case context. Per the PCAST report and NAS report, independence of laboratories from law enforcement (which is why many don't know case context) was emphasized as something that would be beneficial for crime laboratories in general. Is this going against that recommendation? Then the following paragraph is discussing that if the labs don't take the time to put the information into reports or evaluate relevance, implying that they are now going to bias one side or the other by doing this. But bias is what labs are trying to avoid by not diving too deep into the arguments made by one side or another within court, and by being an independent laboratory. In the previous paragraph (4770-4773), it is mentioned that there are arguments for and against assigning probabilities to activity-level propositions, but it is not mentioned how lengthy and complex a process assigning these probabilities is, and there is no mention of how challenging those assignments (ex. Using Bayesian networks) could be to explain in court to a stakeholder.	caseloads, backlogs, and local laws governing turnaround times for cases. As well as discussing the value of laboratories remaining independent of law enforcement and prosecutorial or defense entities, and how that lack of influence can hope to unbias scientific reporting. Assigning probabilities to activity-level propositions is highly subjective, as mentioned here, but it is also very time-consuming and typically does not result in exceptionally strong evidence in either direction for the propositions. The value of that time could be discussed here, and whether more general statement examples of how activity-level propositions could be put into reports to separate them from the sub-source level statements that are being reported could be given.
4961; Chapter 6	This chapter attempts to provide new technologies to assist with DNA mixture interpretation; however, presentation of cell separation techniques (such as laser-capture microdissection) as a method appears out-of-touch with casework demands in crime labs. The described cell separation techniques are laborious and unrealistic for crime labs that process hundreds, if not thousands, of samples a year.	Remove Chapter 6.

	In addition, the application of NGS to mixtures needs further assessment. NGS lacks available tools such as prob gen for interpreting mixtures. Lines 5173-5174 and 5240 of the draft report mention the issue of distinguishing low-level contributors from noise sequences. Separately, there is no consistent guidance on applying statistics to additional markers like XSTRs and SNPs, that may be part of a NGS panel. These need to be addressed before this technology can be adopted as an improvement to DNA mixtures or a solution to existing problems in the forensic community.	
4961	General comment on Chapter 6 - aligns nicely with the DEI document yet to be published, esp Takeaway 6.2.	
5213	Do you mean developmental, instead of internal?	"performed in a developmental validation"
5343	Regarding 3rd point, solving a problem is not always why changes are made. Often it's simply worthwhile improvements (e.g., for efficiency).	Incorporate.
6698	Retain results for exam by third parties	This is from 2020 and nowhere does it state provide your data in a public forum or it's not able to be assessed for reliability. Data is now available to third parties such as court requests to have outside experts come onsite and assess it.
7090-7091; 7207-7208	Appendix 2 - Virtual courses could be offered by the NIJ Forensic Technology Center of Excellence… Virtual courses are offered and should be cited here. For example, the 8 part lecture series on probabilistic genotyping. This comes up later (7207-7208) but should be mentioned here.	Appropriately update.
7146-7148; 7137-7138	Appendix 2 - Seems worth mentioning here that CA, which contains numerous crime labs and ~10% of the US population has a librarian on staff with the CA	Consider adding something that other states could model or leverage such services.

	Criminalistics Institute (https://oag.ca.gov/cci). CCI subscribes to numerous forensic journals and has an electronic library catalogue system. CA criminalists not within CA DOJ simply send their requests directly to the librarian who responds with literature pdfs. Furthermore, requests can be made of any journal, not just those to which CCI subscribes.	
7211-7212	A certificate of attendance by itself is not sufficient for demonstrating that training or continuing education materials have been understood. This point is understood though this is common practice in numerous other fields. Perhaps a missed opportunity here is to consider the offerings of the ABC.	Incorporate the offerings of the ABC and make viable suggestions, such as a recommendation to states to offer incentives or even mandates (Texas as an example) on passing their certification test, and building out ABC to include CE exams.
7256-7259	"forensic community" and "advisory group"	Who is this "advisory group" – the make up of this proposed group needs to be defined if it is going to be a consideration. And it is imperative that an advisory group have individuals who are practitioners, not just theorists
7270-7283	What is this actually saying?	
7271-7273	"Consensus decisions from an advisory group"	Who is this "advisory group" – this needs to be defined if it is going to be a consideration. (see above)
7285-7287	This is already covered by the QAS in that the TLs must approve the analysts training	
7291-7292	What is this statement based on? How many internal validation studies were looked at to determine that TLs don't have sufficient training/experience to design validation experiments.	

PC51

comments on NISTIR 8351-DRAFT 68 (June 2021)

Geoffrey Stewart Morrison <[redacted]>

Sun 10/24/2021 4:34 AM

To: ScientificFoundationReviews <ScientificFoundationReviews@nist.gov>

Please find attached my comments on NISTIR 8351-DRAFT 68 (June 2021)

PC51a

Comments on:

Butler J.M., Iyer H., Press R., Taylor M., Vallone P.M., Willis S. (2021). DNA Mixture Interpretation: A NIST Scientific Foundation Review. National Institute of Standards and Technology Internal Report (NISTIR) 8351-DRAFT 68 (June 2021). https://doi.org/10.6028/NIST.IR.8351-draft

by:

Geoffrey Stewart Morrison

http://geoff-morrison.net/

Disclaimer:

All opinions expressed in the present submission are those of the author, and, unless explicitly stated otherwise, should not be construed as representing the policies or positions of any organizations with which the author is associated.

Comments:

I would like to draw the report authors' attention to Morrison et al. (2021) *Consensus on validation of forensic voice comparison*, and to the body of literature cited therein. The *Consensus* was developed on a foundation of two decades of work in a branch of forensic science in which the data are continuously valued and have substantial within source variability. This is a context in which the magnitudes of log-likelihood-ratio values are much closer to zero than in single-source DNA interpretation, which is also the case for DNA-mixture interpretation. Although the scope of the *Consensus* is forensic voice comparison, with minor wording changes it would be applicable to many

other branches of forensic science, including DNA-mixture interpretation.

There are multiple recommendations in the *Consensus*, some of which parallel recommendations in the draft report (although sometimes with different emphasis), and some which could potentially become additions to the report.

As potential additions, I would like to highlight the metric, log-likelihood-ratio cost (C_{llr}), and the graphic, Tippett plot, used for representing the results of a validation of the performance of a system that outputs likelihood ratios. I would suggest that these are an appropriate metric and an appropriate graphic for this purpose, whereas the graphic mentioned in the draft report (in the paragraphs beginning lines 1051, 3599, and 3720) and metrics mentioned/described in the draft report (in the paragraphs beginning lines 2816 and 3571) are not.

The *Consensus* places a much stronger emphasis on the calibration of systems that output likelihood ratios than does the draft report. The draft report includes scattered mentions of such calibration (in the paragraphs beginning lines 1051, 3558, 3599, 3655, 3720, and 3734), but no comprehensive treatment. Some of these mentions (e.g., in the paragraph beginning line 3720) are specifically about measuring degree of calibration rather than the process of calibration. The *Consensus* places an emphasis on the need for the system to have been calibrated. Elsewhere (Morrison, 2021), I have argued that, once a system has been calibrated, there is no longer any meaningful metric of degree of calibration for that system.

KEY TAKEAWAY #2.6 of the report is correct, but, overall, the report places what I consider to be an unwise emphasis on the subjectivity of the assignment of likelihood-ratio values (e.g., in the paragraphs beginning lines 1899, 1910, and 2781). DNA-mixture interpretation using probabilistic-genotyping systems and human-supervised-automatic approaches to forensic voice comparison share the properties of being systems that apply statistical models to quantitative measurements. For such systems (as emphasized in §2.6 of the *Consensus*), subjective judgements have to be made about

whether the data used for training the models (including the calibration model) are appropriate and about whether the data used for validating the system are appropriate, i.e., whether they are sufficiently representative of the relevant population and sufficiently reflective of the conditions of the case. These subjective judgements are made before the forensic practitioner has analyzed the properties of the questioned-source and known-source items, thus the practitioner does not know what effect their judgements will have on the likelihood-ratio value actually calculated for the comparison of the questioned-source and known-source items. The process of making these subjective judgements is therefore resistant to cognitive bias. Once these judgements have been made, the remainder of the system is automated and therefore is not susceptible to cognitive bias (note that the remainder of the system calculates a likelihood ratio or a Bayes' factor). Contrast this with likelihood ratios, Bayes' factors, or other strength-of-evidence conclusions that are directly the output of a subjective-judgement process. The appropriateness of the practitioner's judgement as to whether the training and validation data are appropriate is a pre-empirical question. It is therefore a legitimate topic of debate before the decision maker. If the decision maker is satisfied that the data used by the practitioner are appropriate, how good the judgements were about what measurements to use, what models to use, and (for a Bayesian approach) what prior parameter distributions to use, can then be assessed via empirical testing. If the latter judgements were good the performance of the system will be relatively good, if the latter judgements were poor the performance of the system will be relatively poor. Yes, the likelihood ratio or the Bayes' factors output by such systems are subjective, but the subjectivity can be relegated to judgements about data, which should be disclosed and can be debated, and judgements whose goodness can be assessed via empirical validation. This message is what should be emphasized. This is a message that can potentially instill trust in probabilistic-genotyping systems by laypeople such as lawyers. The generic message that likelihood ratios are subjective will instead instill skepticism about the trustworthiness of probabilistic-genotyping systems.

There are two topics addressed in the draft report that are each related to the previous topic:

One topic relates to the precision of likelihood ratios, referred to in the paragraph beginning at line 2774. This paragraph is biased toward a subjectivist Bayesian perspective that emphasizes the wrong message (see the discussion of the previous topic). The paragraph cites a number of papers, including several papers from a journal special issue of which I was guest editor, but the majority of the papers cited come from only one side of the debate. Other papers that could be cited to correct the imbalance would include Sjerps et al. (2016) and Morrison (2017a). The latter paper points to a practical solution that is further expanded upon in Morrison & Poh (2018): Rather than emphasize an esoteric philosophical disagreement, use procedures that both sides of the debate agree would have the desired effect that as the amount of data available decreases the value of the likelihood ratio or Bayes' factor shrinks toward the neutral value of one. Indeed, the desirability of this solution is already mentioned in the report in the paragraph beginning at line 3532.

The other topic relates to arguments made in Lund & Iyer (2017), which are referenced in the paragraph beginning at line 1910. These arguments are strawman arguments. The strawman arguments may have been based on misunderstandings of the literature (which may not always have been clear), but were also biased toward theoretical subjectivist Bayesian perspectives, and ignored the more pragmatist perspectives discussed above. My own response (Morrison, 2017b) was primarily in response to the NIST press release which sensationalized the strawman arguments made in Lund & Iyer (2017). I would suggest that NIST should not continue to endorse these strawman arguments. As I stated in my response (which, with ellipsis, was favourably quoted by Gittelson et al., 2018): "Transparent implementation of the likelihood ratio framework using relevant data and statistical models with empirical testing of performance under casework conditions is actually the solution to the problem." In general, the problem of whether courts should trust the output of forensic-evaluation systems. My statement

seems to be in line with the overall message of the draft report, especially KEY TAKEAWAY #2.6, whereas Lund & Iyer (2017) appears to attempt to undermine that message.

Finally, I make a comment regarding so-called verbal scales (discussed in paragraphs beginning lines 1980 and 1991). Some of my thoughts on this matter are stated in Morrison & Enzinger (2016), which is already cited elsewhere in the draft report, and Marquis et al. (2016), also cited, provide an excellent critical discussion. If one has made use of a probabilistic-genotyping system trained (including calibrated) using relevant data (and one has adopted procedures that shrink the likelihood-ratio values toward a neutral value of one), and one has validated that system under conditions reflecting those of the case under investigation, then it makes no sense to substitute or supplement the transparently-calculated likelihood-ratio output of that system, a number that has an explicit meaning with respect to the competing hypotheses, with an arbitrary verbal expression whose meaning is at best vague. I recommend that this not be encouraged.

References:

Gittelson S., Berger C.E.H., Jackson G., Evett I.W., Champod C., Robertson B., Curran J.M., Taylor D., Weir B.S., Coble M.D., Buckleton J.S. (2018). A response to “Likelihood ratio as weight of evidence: A closer look” by Lund and Iyer. *Forensic Science International*, 228, e15–e19. https://doi.org/10.1016/j.forsciint.2018.05.025

Lund S.P., Iyer H. (2017). Likelihood ratio as weight of forensic evidence: A closer look. *Journal of Research of National Institute of Standards and Technology*, 122, article 27. https://doi.org/10.6028/jres.122.027

Marquis R., Biedermann A., Cadola L., Champod C., Gueissaz L., Massonnet G., Mazzella W.D., Taroni F., Hicks T. (2016). Discussion on how to implement a verbal scale in a forensic laboratory: Benefits, pitfalls and suggestions to avoid misunderstandings. *Science & Justice*, 56, 364–370. http://dx.doi.org/10.1016/j.scijus.2016.05.009

Morrison G.S. (2017a). What should a forensic practitioner's likelihood ratio be? II. *Science & Justice*, 57, 472–476. http://dx.doi.org/10.1016/j.scijus.2017.08.004

Morrison G.S. (2017b). A response to: "NIST experts urge caution in use of courtroom evidence presentation method". Available from: https://arxiv.org/abs/1710.05878

Morrison G.S. (2021). In the context of forensic casework, are there meaningful metrics of the degree of calibration? *Forensic Science International: Synergy*, 3, article 100157. https://doi.org/10.1016/j.fsisyn.2021.100157

Morrison G.S., Enzinger E. (2016). What should a forensic practitioner's likelihood ratio be? *Science & Justice*, 56, 374–379. http://dx.doi.org/10.1016/j.scijus.2016.05.007

Morrison G.S., Enzinger E., Hughes V., Jessen M., Meuwly D., Neumann C., Planting S., Thompson W.C., van der Vloed D., Ypma R.J.F., Zhang C., Anonymous A., Anonymous B. (2021) Consensus on validation of forensic voice comparison. *Science & Justice*, 61, 229–309. https://doi.org/10.1016/j.scijus.2021.02.002

Morrison G.S., Poh N. (2018). Avoiding overstating the strength of forensic evidence: Shrunk likelihood ratios / Bayes factors. *Science & Justice*, 58, 200–218. http://dx.doi.org/10.1016/j.scijus.2017.12.005

Sjerps M.J., Alberink I., Bolck A., Stoel R.D., Vergeer P., van Zanten J.H. (2016). Uncertainty and LR: To integrate or not to integrate, that's the question. *Law Probability and Risk*, 15, 23–29. http://dx.doi.org/10.1093/lpr/mgv005

PC52

Second Comment on the for NISTIR-8351-DRAFT Report

WICKENHEISER, RAY (TROOPERS) [redacted]

Fri 11/5/2021 11:42 AM

To: ScientificFoundationReviews <ScientificFoundationReviews@nist.gov>

Good Morning,

Further to your solicitation "to receive additional comments, new data, or information" found at https://content.govdelivery.com/accounts/USNIST/bulletins/2f8b05e, Second Public Comment Period for NISTIR-8351-DRAFT Report: Oct. 22 to Nov. 19, 2021, I'm providing the following information to supplement my original comments (see also original emailed comments below).

The New York State Police Forensic Investigation Center (FIC) is currently validating the use of the Probabilistic Genotyping STRMix and expects to complete the validation study in the spring of 2022. The FIC will make its validation study data available to your team for on-site review at the FIC in Albany, NY, upon execution of a memorandum of understanding of confidentiality similar to that provided by external auditors who routinely perform audits at our laboratory for accreditation purposes. This stipulation is necessary to ensure protection of the DNA data used for the validation study, which includes DNA profile data provided by FIC staff for purposes of internal validation studies.

We believe our completed validation study will demonstrate that DNA mixture interpretation performed by the FIC using probabilistic genotyping will be fit for purpose to reliably interpret DNA mixtures samples within the scope of our validation. Conversely, if your team feels there are additional measures necessary to demonstrate fit for purpose, we welcome your feedback and appreciate any specifics you can provide on any additional studies evaluating such measures.

In advance of your on-site visit, we request you provide the criteria upon which you will evaluate DNA mixture validation studies. Also, please include the names of individuals on your team with forensic DNA casework and auditing experience, which is routinely provided for external audits nationwide.

We further request that you reach out to authors of the many DNA mixture publications you list as having reviewed to request and review the data with the same defined evaluation criterion.

We look forward to your on-site visit.

Regards,

Ray

Dr. Ray Wickenheiser DPS MBA FAAFS
Director, NYSP Crime Laboratory System

New York State Police
Forensic Investigation Center
1220 Washington Avenue, Building# 30
Albany, New York 12226-3000

[redacted]

[redacted]

NEW YORK STATE | State Police

www.troopers.ny.gov

From: WICKENHEISER, RAY (TROOPERS)
Sent: Thursday, August 19, 2021 11:44 AM
To: scientificfoundationreviews@nist.gov
Subject: Draft Report on DNA Mixture Interpretation Methods

Good Morning,

As you are aware, I was an invited guest of NIST to serve on the Resource Group to provide input for this DNA mixture scientific foundation review. The Resource Group provided the perspective of forensic crime laboratory experience to NIST, however were not included in authorship in the NIST report.

This draft report contains concepts and material that were not previously reviewed by the Resource Group, as no draft materials have been shared in the last approximately 1.5 years. I have a number of concerns with the draft NIST report on DNA mixtures. Three of the most significant are as follows:

1. The data sample utilized by NIST in generating this report is too restrictive and does not accurately reflect validation data used by forensic laboratories. NIST is only reviewing data that is publicly available. Most forensic laboratory validation data is not made public, as it contains staff, friends and family profiles, and individuals providing the samples who did not provide informed consent to permit their DNA profiles to be released into the public domain. Forensic laboratories operate in a secure environment where data must be safeguarded, which runs contrary to NIST's determination that only data published or posted publicly qualify for their foundation review.

   NIST did not make a request to public laboratories to review their data. Much validation data is currently available for defense witnesses, laboratory auditors and assessors review at forensic laboratory premises and has been independently reviewed by these entities. Requiring data to be publicly available as a prerequisite to determining it is valid is an unprecedented requirement by NIST, which is not in place for many other scientific endeavors. Therefore, I feel NIST's requirement that only data that is in the public domain will be used to determine the scientific foundation for DNA mixture interpretation is too restrictive.

   Recommendation: NIST visit forensic laboratories and forensic DNA mixture interpretation vendors and review validation data on site. As an alternative, they could make requests to review such data with appropriate confidentiality measures in place.

2. NIST incorrectly contends that forensic laboratory data has not been independently reviewed. There are 60 publications including DNA mixture studies noted in the NIST report, including one with 1315 samples run by 31 different forensic laboratories [1]. All forensic lab DNA validation studies are reviewed by independent external auditors within their 2-year external audit FBI Quality Assurance Standards requirements, and also by independent auditors from the national accrediting board 4-year audit cycle to

meet ISO 17025:2017 standard requirements. Additionally, some states have statutorily created bodies responsible for oversight of forensic laboratory accreditation and approval of such laboratories use of new scientific methodologies and technologies. Many of these bodies have panels of forensic experts who have independently reviewed data and approved probabilistic genotyping of DNA mixtures as fit for purpose. Therefore, in my opinion DNA mixture data validation studies and data have been independently reviewed by objective external forensic experts and been found to be fit for use for individual forensic laboratories.

NIST authors do not have the necessary practical forensic experience of working laboratories. The stated purpose of the Resource Group was to provide forensic experience that is not possessed by the authors of the NIST report. Within the Resource Group as well as throughout the forensic science laboratory and DNA mixture interpretation vendor community exists a wealth of forensic experience with forensic laboratory validations, data, forensic casework and samples. I feel the importance of this DNA mixture scientific foundation report warrants inclusion of this experience in review of data, determination of what defines scientific foundation and in authorship of the report.

Recommendation: NIST include individuals with appropriate practical forensic experience to assist with independent review of validation studies and data and co-authorship of the report.

3. The draft report recommends an impracticable standard for validation studies to meet.
NIST defines a novel concept of "factor space" including 26 factors impacting DNA mixtures, stating that the publicly available data did not cover this factor space. If every factor were comprehensively covered in a single mixture's "factor space," each of these 26 variables would need to be changed while holding the rest constant to determine the impact of a single variable on the mixture's behavior. Assuming 10 increments for each of the 26 variables, this would require 403 septillion factor comparisons (10 x 26 factorial). This huge number of samples is not practical nor feasible. The factor space model is therefore not appropriate for demonstrating that DNA mixture interpretation as practiced by forensic laboratories is fit for purpose.

Recommendation: NIST abandon the concept of factor space and develop a more practical measure of what is required to demonstrate fit for purpose and apply that measure to the review of on-site data with additional experts with forensic experience. NIST should then revisit their preliminary report, make the recommended changes herein and include forensic expertise in authorship of the next corrected version.

1. Bright JA, et al. (2018) Internal Validation of STRmix™—a Multi-Laboratory Response to PCAST. *Forensic Science International: Genetics* 34: 11-24.

Please note these comments are my own, and not representative of the New York State Police, ASCLD, SWGDAM, OSAC or any other agency or organization with which I am affiliated.

Regards,

Ray

Dr. Ray Wickenheiser DPS MBA FAAFS
Director, NYSP Crime Laboratory System

New York State Police
Forensic Investigation Center
1220 Washington Avenue, Building# 30
Albany, New York 12226-3000
[redacted]

[redacted]

NEW YORK STATE | State Police

www.troopers.ny.gov

PC53

Comment on NIST DNA Mixture Interpretation: A Scientific Foundation Review

Bjorn Sutherland [redacted]

Sun 11/7/2021 6:16 PM

To: ScientificFoundationReviews <ScientificFoundationReviews@nist.gov>
Cc: John Bone [redacted]; John Buckleton [redacted]; Jo Bright [redacted]; Richard Wivell [redacted]

Dear Sir/Madam,

Please see attached a second response to this review.
Kind regards
Björn

Björn Sutherland MSc
STRmix Manager
Kenepuru Science Centre: 34 Kenepuru Drive, Kenepuru, Porirua 5022
PO Box 50348, Porirua 5240, New Zealand

[redacted]
www.strmix.com

PC53a

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.

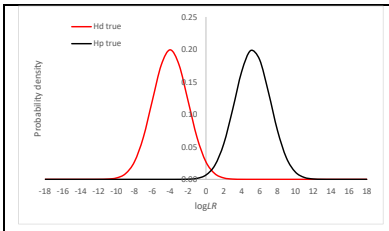


Figure 1a. A well calibrated *LR*

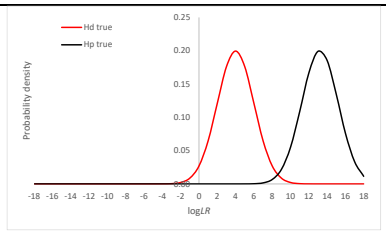


Figure 1b. A poorly calibrated *LR*

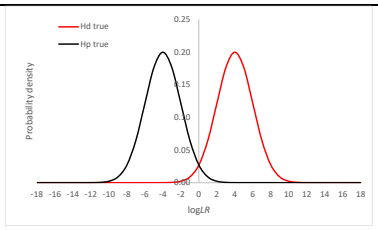


Figure 1c. A "reversed" set of *LR*s

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

[redacted]

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 [redacted]
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.

PC54

Second Public Comment Period for NISTIR-8351-DRAFT Report: Oct. 22 to Nov. 19, 2021

Raymond Valerio [redacted]

Mon 11/15/2021 12:11 PM

To: ScientificFoundationReviews <ScientificFoundationReviews@nist.gov>

To Whom it May Concern,

Attached to this email is public comment on the Draft NISTIR 8315 Report. Please acknowledge receipt.

Thank you for your consideration,
Raymond Valerio

****

A.D.A. Raymond Valerio
Director, Forensic Sciences

Office of the Queens County District Attorney
125-01 Queens Boulevard, 2nd Floor
Kew Gardens, NY 11415
[redacted]

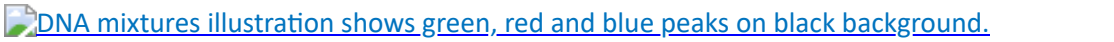
From: National Institute of Standards and Technology (NIST) <subscriptions@service.govdelivery.com>
Sent: Friday, October 22, 2021 9:55 AM
To: Raymond Valerio [redacted]
Subject: [EXTERNAL] Second Public Comment Period for NISTIR-8351-DRAFT Report: Oct. 22 to Nov. 19, 2021

nist

View as a Web Page

Green rectangle says: Updates

Second Public Comment Period for NISTIR-8351-DRAFT Report: Oct. 22 to Nov. 19, 2021


DNA mixtures illustration shows green, red and blue peaks on black background.

The National Institute of Standards and Technology (NIST) is reopening the comment period on the NIST Internal Report 8351-DRAFT, *DNA Mixture Interpretation: A NIST Scientific Foundation Review,* to receive additional comments, new data, or information.

Written comments and related material must be submitted to scientificfoundationreviews@nist.gov by 11:59 p.m. Eastern Standard Time on Nov. 19, 2021. All comments received to date will be considered and need not be resubmitted. If any commenters who have already submitted comments wish to provide supplemental or updated comments, we encourage them to do so.

Read the Draft Report

Connect with us

facebook twitter youtube linkedin instagram

Received this email from a friend? Subscribe here.

Privacy Policy

NIST

PC54a

QUEENS COUNTY DISTRICT ATTORNEY
125-01 QUEENS BOULEVARD
KEW GARDENS, NEW YORK 11415-1568

MELINDA KATZ
DISTRICT ATTORNEY

718.286.6000
WWW.QUEENSDA.ORG

November 15, 2021

Dear Dr. Butler & the NISTIR 8351 Review Team,

Thank you for the opportunity to submit further public comment. The following comment calls for external peer review of NISTIR 8351 and the use of non-public data sources for NISTIR 8351.

External Peer Review:

NISTIR 8225 outlines the reasons and goals for NIST's scientific reviews. In lines 331-332 it says, "In fiscal year 2018, Congress appropriated funding for NIST to conduct 'technical merit evaluations.' NIST scientific foundation reviews are intended to fulfill this mandate." Therefore, it follows that one of the NIST Scientific Foundation Review goals on lines 710-712 is, "(6) sharing findings with the scientific and criminal justice communities to **convey the capabilities and limitations of studied forensic disciplines** to practitioners, judges, lawyers, jurors, and other stakeholders." (emphasis added). At bottom, the essence of a technical merit evaluation is to determine the validity of the forensic discipline. See OSAC Technical Merit Worksheet. There is nothing more influential than this.

NIST defines "influential" and "influential scientific information" at https://www.nist.gov/nist-information-quality-standards.

- "Influential scientific information means scientific information the agency reasonably can determine will have or does have a clear and substantial impact on important public policies or private sector decisions."
- "Influential, when used in the phrase "influential scientific, financial, or statistical information," means that the agency can reasonably determine that dissemination of the information will have or does have a clear and substantial impact on important public policy and private sector decisions." See https://www.nist.gov/nist-information-quality-standards, last visited 11/15/21.

In practice, there is a growing body of evidence that the NISTIR 8351 is "influential scientific information."

- At the New York State Commission on Forensic Science Meeting on 9/17/21, NIST's Review was discussed (1:16:23-1:57:50). As a direct result of NIST's Review, one defense attorney commissioner motioned to suspend the use of probabilistic genotyping software ("PGS") in New York State (https://www.youtube.com/watch?v=uNUczp9LFgs) ("…the only appropriate action is to place a moratorium on the use of PG in New York State for complex DNA mixtures…"). She said NIST's review of PGS was "troubling" and "such a huge red flag". Such a moratorium would cause nearly all courtroom use of DNA to cease.

- In California, defense attorneys have already cited to the draft NISTIR Review to challenge the reliability of PGS. In *People v. Alvin Davis*, Case No. C089567, the defense attorney asked the Third District Court of Appeal to take judicial notice of NIST's Review.
- In New York City, NIST's review is likewise being cited by defense attorneys. For example, in a homicide case, a well-known defense DNA expert wrote a letter to the Court to pause the trial until NIST finalizes this Review. In another NYC case, *People v. Daval Wright*, Indictment 3167/2019, the defense attorney filed a *Frye* motion challenging the reliability of STRmix™ based partly on the NIST Review.

In the NIST Information Quality Standards, there is guidance regarding peer review of "influential scientific information." Based on the conclusions and opinions of the NISTIR 8351 Foundation Review, NIST's definition of "influential" is satisfied. Therefore, according to the White House OMB Memorandum "m05-03", it appears that peer review (external, independent) is necessary for the NISTIR 8351 Foundation Review.

The OMB Memorandum specifically states that "Peer review should not be confused with public comment and other stakeholder processes" and "The mere existence of a public comment process (e.g., notice-and-comment procedures under the Administrative Procedure Act) does not constitute adequate peer review or an "alternative process," because it does not assure that qualified, impartial specialists in relevant fields have performed a critical evaluation of the agency's draft product" (See OMB Memo pp. 4, 28 of 45).

- What peer review will be conducted for the NISTIR 8351 Foundation Review?
- Will the peer review be external and independent?
- If it is your position that peer review is not required, can you please explain why not?

Use of Non-Public Data Sources in NISTIR 8351:

NISTIR 8225 describes NIST's published approach to conducting scientific foundation reviews, including data sources used, evaluation criteria, and expected outputs. As such, NISTIR 8225 seems to negate one of the central takeaways from NISTIR 8351—that there isn't enough publicly available information to evaluate the reliability of PGS (Key Takeaway 4.3). NISTIR 8225 states:

"1.1. What Data Sources Will We Use?
176
177 Because peer-reviewed publications are essential building blocks of a respected edifice of
178 scientific knowledge, studies that address the reliability of forensic methods would ideally be
179 present in a discipline's published, peer-reviewed, and well-cited scientific literature.
180 However, a focus on peer-reviewed literature alone may not provide a complete picture of a
181 discipline's available body of knowledge. **For instance, data from laboratory validation**
182 studies may not be publicly available or published. Therefore, NIST scientific foundation
183 reviews are designed to seek input by:
184
185 • collecting and evaluating the peer-reviewed literature
186 • assessing available data from interlaboratory studies, proficiency tests, and laboratory
187 validation studies
188 • exploring other available information including **position statements and non-peer**
189 reviewed literature
190 **• obtaining input from members of the relevant community through interviews,**
191 workshops, working groups and other formats for the open exchange of ideas and

192 information.
193
194 Obtaining input from experts outside of NIST is an integral component of a NIST scientific
195 foundation review. This will help ensure that these reviews capture the full breadth of
196 knowledge that forensic practitioners and researchers consider foundational to their
197 discipline."
(emphasis added)

Since NIST is using NISTIR 8225 to guide NISTIR 8351, non-public data sources such as interviews with laboratories, experts, and working groups should be conducted before NIST publishes the final version of 8351. And, therefore, Key Takeaway 4.3 should be modified.

Thank you for taking the time to consider these points on external peer review and non-public data sources.

Sincerely,

Raymond Valerio
Assistant District Attorney
Director, Forensic Sciences
Queens County District Attorney's Office
125-01 Queens Boulevard
Kew Gardens, NY 11415

[redacted]

PC55

SWGDAM Comments on NISTIR 8351 - DNA Mixture Interpretation: A Scientific Foundation Review

Dawn Herkenham [redacted]

Wed 11/17/2021 8:14 AM

To: ScientificFoundationReviews <ScientificFoundationReviews@nist.gov>
Cc: Anthony Onorato [redacted]; Dawn Herkenham [redacted]

Hi,

Please accept the attached updated comments from the Scientific Working Group on DNA Analysis Methods (SWGDAM) on the draft report entitled *DNA Mixture Interpretation: A Scientific Foundation Review*.

If you have any questions, please contact the SWGDAM Chair, Anthony Onorato, at ajonorato@fbi.gov.

Thank you,
Dawn Herkenham
Executive Secretary, SWGDAM

ECS

Dawn Herkenham
CODIS Operations and Support Services

- - - - - - - - - -

[redacted]

PC55a

SCIENTIFIC WORKING GROUP — SWGDAM — DNA ANALYSIS METHODS

August 23, 2021
Revised and Resubmitted November 15, 2021

C/O scientificfoundationreviews@nist.gov
Dr. John M. Butler
NIST Fellow
Special Assistant to the Director for Forensic Science
Special Programs Office
National Institute of Standards and Technology
100 Bureau Drive, Mail Stop 4701
Gaithersburg, MD 20899-4701

Dear Dr. Butler:

The Scientific Working Group on DNA Analysis Methods (SWGDAM) respectfully submits the following substantive comments on NISTIR 8351-DRAFT entitled *DNA Mixture Interpretation: A NIST Scientific Foundation Review*.

Because the draft report summarizes a number of SWGDAM Guidelines documents in its Appendices, we offer the following background for our comments. As you know, SWGDAM is a forensic DNA working group, established over 25 years ago, to "serve as a forum to discuss, share and evaluate forensic biology methods, protocols, training, and research to **enhance** forensic biology services…" Congressional recognition of SWGDAM's predecessor, TWGDAM, is contained in the Federal DNA Identification Act with the provision establishing the TWGDAM Guidelines as national standards for participation in this National Index until the new Federal DNA Advisory Board recommended national quality assurance standards to the FBI Director for adoption.[1] The Federal DNA Advisory Board was responsible for recommending quality assurance standards, and revisions as necessary, to the Director of the Federal Bureau of Investigation (FBI), and when their statutory time period concluded, the Board charged SWGDAM with this responsibility.

[1] Enacted in 1994, the Federal DNA Identification Act is the enabling legislation for the National DNA Index System (NDIS); *see* 34 U.S.C.§12591 *et seq.* The provision relating to the use of the TWGDAM standards in the interim before the Federal DNA Advisory Board recommends standards to the FBI Director can be found at 34 U.S.C. §12591(a)(4).

SWGDAM is currently comprised of dedicated forensic scientists, from international, federal, state and local forensic DNA laboratories as well as guests representing academia and other Federal agencies. These forensic scientists serve as the DNA technical leaders or Combined DNA Index System (CODIS) Administrators for their laboratories and are able to offer the perspectives of practitioners in the areas of STR, Y STR, mitochondrial DNA and next generation sequencing (NGS) technologies. We are fortunate to also have invited guests attend each meeting to provide their specific expertise in areas such as population genetics, Rapid DNA, probabilistic genotyping, statistics, etc.

Laboratories performing forensic DNA analysis and participating in the FBI's National DNA Index, unlike other forensic disciplines, are subject to Federal statutory requirements relating to quality assurance and privacy. As early as 2006, accreditation was required for forensic DNA laboratories contributing DNA records to the National Index. In addition to generating these DNA records in accord with specific minimum standards for a quality assurance program, these laboratories also abide by limited disclosure and release rules for their DNA records. Federal law also requires that these laboratories undergo an external audit every two years to document their compliance with the minimum standards. In fact, the Quality Assurance Standards audit documents contain a separate appendix – "Appendix E: Approved Validations" – to document the evaluation and approval of developmental and internal validations during the audit.[2] Moreover, the U.S. Department of Justice' Office of the Inspector General conducts audits of CODIS laboratories for compliance with the Federal DNA Act requirements as well as NDIS Operational Procedures.[3] In addition to this rigorous program of audits, the FBI's CODIS Unit conducts assessments of the NDIS participating laboratories as part of their administration of the National DNA Index System.

Several states also have an additional level of oversight provided by their State forensic oversight boards that review, evaluate and approve new technologies or methods prior to their use by forensic DNA laboratories, such as, the Connecticut DNA Data Bank Oversight Panel, the District of Columbia's Scientific Advisory Board, the

[2] *See The FBI Quality Assurance Standards Audit for DNA Databasing Laboratories*, and *The FBI Quality Assurance Standards Audit for Forensic DNA Testing Laboratories*; available at https://www.swgdam.org/publications.

[3] *See Combined DNA Index System Audits,* available at https://oig.justice.gov/reports/codis-ext.htm.

Massachusetts Forensic Science Oversight Board, the New Mexico's DNA Oversight Committee, the New York State Commission on Forensic Science and DNA Subcommittee, the Texas Forensic Science Commission, and the Virginia Forensic Science Board and Scientific Advisory Committee.[4]

Since the overwhelming majority of our membership and invited guests represent NDIS participating laboratories, we wish to highlight the Federal and state statutory requirements for the confidentiality and privacy of the DNA data, in addition to those privacy protections afforded by the Genetic Information Nondiscrimination Act of 2008, the Health Insurance Portability and Accountability Act of 1996, and applicable State genetic privacy laws.[5] The Federal DNA Identification Act provides for limited access to and disclosure of the DNA samples and resulting DNA analyses generally to criminal justice agencies for law enforcement identification purposes. The Federal DNA Act responsibilities are further explained in the NDIS Operational Procedures Manual which limits access for anonymized DNA data **to criminal justice agencies** for a population statistics database, forensic identification, forensic research, forensic protocol development or quality control purposes. The FBI *Quality Assurance Standards for DNA Databasing and Forensic DNA Testing Laboratories* also emphasize the confidentiality of the DNA data in Standard 11.

In addition to the limited access requirements of Federal law, states also prescribe access to the DNA data in their State databases. The overwhelming majority

[4] *See* National Conference of State Legislatures, *Legislative Study and Oversight of Forensic Services*; available at https://www.ncsl.org/research/civil-and-criminal-justice/dna-database-search-by-policy.aspx.)

[5] *See generally*, Genetic Information Nondiscrimination Act of 2008, P.L. 110-233 (Section 206 of GINA provides for the confidentiality of an employee's genetic information and specific limitations on disclosure), available at https://www.eeoc.gov/statutes/genetic-information-nondiscrimination-act-2008; and the Health Insurance Portability and Accountability Act, *see* P.L. 104-191 and the Final Omnibus Rule at Fed. Register, Vol. 78, No. 17 (2013) (the HIPAA Privacy Rule includes safeguards to protect personal health information); Smith, S., Nielson, P. S., Kennedy, B. (2011) *Genetic Privacy Laws: 50 State Survey*, Journal of Health & Life Sciences Law, Vol. 5, No.1. *See also,* States enacting protections for genetic information in 2021, such as South Dakota S.B. 178, Utah S.B. 227 (Chapter 361), Florida H.B. 833 (Chapter No. 2021-216).

of State laws restrict access to the DNA data in these law enforcement databases **to criminal justice agencies** for law enforcement identification purposes.[6] These State laws also penalize the unauthorized disclosure of DNA data as misdemeanor or felony offenses.[7]

SWGDAM has several general observations about the NISTIR 8351-DRAFT: (1) the lack of forensic science practitioner(s) expertise among the authors may have hampered consideration of the existing regulatory and statutory framework within which forensic DNA laboratories operate; (2) the introduction of a new requirement that internal validation be 'publicly available' in order to be considered within the foundational review appears to be based solely on the authors' belief, with no further

[6] *See, for example*, Ala. Code §36-18-27; Alaska Stat. §44.41.035(F); Ariz. Rev. Stat. Ann. §13-610. I; Ark. Code Ann. §§12-12-1112, 12-12-1114; Cal. Penal Code 299.5(A), (F); Colo. Rev. Stat. Ann. §§16-11-102.4(5), 16-23-104; Conn. Gen. Stat. Ann. §§54-102i(B), 54-102j; Fla. Stat. Ann §943.325(14); Haw. Rev. Stat. Ann. §§844D-81, 82, 91; Id. Code §19-5514; 730 Ill. Rev. Stat. Ch. 5, Para. 5-4-3(F); Kan. Stat. Ann. §21-2511(h)(2); La. Rev. Stat. Ann. §§15:612, 15-616; Me. Rev. Stat. Ann. tit. 25 §1577; Md. Code Ann., Pub. Safety §§2-502(C)(4), 2-508; Mass. Ann. Laws Ch. 22E §§9, 10; Mich. Comp. Laws Ann. §28.176(2); Minn. Stat. §299c.155(3); Mo. Rev. Stat. §650.055 (7); N.J. Rev. Stat. §§53:1-20.24, 53:1-20.27; N.M. Stat. Ann. §29-16-8(A), (B); N.Y. Executive Law §995c(6); N.C. Gen. Stat. §§15a-266.8, 15a-266.12; N. D. Cent. Code §§31-13-06; Ohio Rev. Code Ann. §109.573(b)(2), (E); Okla. Stat. Tit. 74, §150.27a (D); R.I. Gen. Laws §§12-1.5-11, 12-1.5-16; S. C. Code Ann. §23-3-650(A); S. D. Codified Laws §§23-5a-22, 23-5a-23, 23-5a-25; Tenn. Code Ann. §38-6-113(d); Tex. Gov't Code Ann. §§411.147 (C), 411.153(A); Utah Code Ann. §53-10-406(3)(A); Vt. Stat. Ann. Tit. 20, §§1937, 1938, 1941; Va. Code Ann. §§19.2-310.4, 19.2-310.5; Wis. Stat. §165.77. States that may have more expansive access provisions, as a condition for participation in NDIS, agree to comply with the limited access and disclosure provisions of the Federal DNA Identification Act. *See* NDIS Operational Procedures Manual, Section 3.2, Version 10 (2021); available at https://www.fbi.gov/file-repository/ndis-operational-procedures-manual.pdf/view.

[7] *See, for example*, Ala. Code §36-18-28; Alaska Stat. §11.56.762; Ark. Code Ann. §12-12-1115; Cal. Penal Code 299.5; Colo. Rev. Stat. Ann. §24-72-309; Conn. Gen. Stat. Ann. §54-102k; Del. Code Ann. Tit. 29 §4713(l); Ga. Code Ann.§35-3-164; Haw. Rev. Stat. Ann. §844D-113; Id. Code §19-5514; 730 Ill. Rev. Stat. Ch. 5, Para. 5-4-3(f-5); Ind. Code Ann. §10-13-6-22; Iowa Code §81.6; Kan. Stat. Ann. §21-2511(n, o); La. Rev. Stat. Ann. §§15:617-618; Me. Rev. Stat. Ann. tit. 25 §1578; Md. Code Ann., Pub. Safety §2-512; Mass. Ann. Laws Ch. 22E §§12, 13; Mo. Rev. Stat. §650.055(5); Neb. Rev. Stat. §29-4110; N.H. Rev. Stat. Ann. §651-C:4; N.J. Rev. Stat. §53:1-20.26; N.M. Stat. Ann. §29-16-12; N.Y. Executive Law §995-f; N.C. Gen. Stat. §15a-266.11; N. D. Cent. Code §31-13-09; Ohio Rev. Code Ann. §§109.99, 109.573(G); Okla. Stat. Tit. 74, §150.27a (D); 44 Pa. Cons. Stat. §2332; R. I. Gen. Laws §12-1.5-15; S. C. Code Ann. §23-3-650; S. D. Codified Laws §23-5a-26; Tex. Gov't Code Ann. §411.153; Vt. Stat. Ann. Tit. 20, §1941; Va. Code Ann. §19.2-310.6; W. Va. Code §15-2B-12; Wis. Stat. Ann. §165.77(5); and Wyo. Stat. §7-19-404.

justification, especially when such a requirement is in conflict with Federal and State privacy laws for DNA data; (3) its use of 'factor space' to suggest that existing internal validations are inadequate and/or incomplete fails to acknowledge existing practices and guidelines to ensure the limitations of these systems are tested; and (4) its failure to pursue constructive alternatives to facilitate a comprehensive scientific foundation review or to now consider additional available data to provide meaningful thresholds and guidance to the forensic DNA community in establishing the reliability of probabilistic genotyping systems.

SWGDAM collected comments from its members and invited guests (*see* attached SWGDAM Comments) to forward to NIST for consideration. There are several points within the Table that we wish to highlight for your review relating to the benefits of including forensic science practitioners on the authorship panel; the arbitrary decision to only review publicly available data; the introduction of a new proposition that each factor/nuance must be evaluated by laboratories in their internal validation of probabilistic genotyping systems in order to adequately understand its limitations; and the impact of failing to establish clear criteria/thresholds for reliability and the foundational review.

Forensic science practitioner input/involvement throughout the process

It appears that the regulatory environment under which forensic DNA laboratories operate was not fully considered and may have been misconstrued as lacking in transparency and/or accountability, when, in fact, internal validations conducted by such laboratories are subject to review during audits and assessments as well as throughout the criminal justice process. This apparent oversight is just one justification for having experienced practitioners of the relevant forensic discipline on the authorship team.

We appreciate the need for objectivity but suggest that the failure to include representatives of the discipline being evaluated does a disservice to the credibility and validity of the report findings as the critical practitioner perspective is not present. While the draft report acknowledges a DNA Mixture Resource Group that included forensic science practitioners, it notes that their involvement was limited to the early stages of the report and none of those individuals are listed as report authors. Significantly, the expertise and training of the forensic science practitioners would have

provided invaluable insight on the issues of making internal validation data public as well as the practicalities of assessing factor space and including validation performance results in case files and reports. As described in the attached comments, a major strength of the NIST sponsored Organization of Scientific Area Committees is that it has embedded stakeholders who are active in the writing of standards and who have a vote on the outcome/final products; such was not the case with this draft report.

Requirement for validation data to be 'publicly available'

The draft NISTIR 8351 emphasizes what is referred to as a lack of demonstrated reliability and deemphasizes (a) the challenges and inability to publish internal validation data; (b) the existence of significant internal validation data performed by forensic DNA laboratories; and (c) the independent and external audit system required for all NDIS participating laboratories as well as other audits, such as ISO/ANAB. The authors made no meaningful attempt to request data directly from these laboratories, a point not addressed in the draft report, using only web searches to determine what may have been 'publicly available.' We suggest that a review of materials available via a basic internet search are neither indicative of the scope nor the reliability of the community's work, especially when there is no stated requirement to publish internal validations. In fact, publication has been discouraged by scientific journals because such material is generally not considered novel. Dr. Michael Peat (editor of the AAFS Journal of Forensic Science) confirmed this at Meeting #12 (January 9, 2010) of the National Commission on Forensic Science. A laboratory, even if granted permission to use time and personnel resources towards publication efforts instead of more casework/public safety-focused endeavors, would likely have to publish in a "pay to publish" journal (such as Forensic Science International: Reports). However, it is not clear whether the authors would consider this sufficient, nor does this seem like a reasonable resource expenditure for most public crime laboratories.

The authors may wish to consider an approach to establishing foundational validity that is more practical than the mass publication of internal validation studies in peer reviewed journals while concurrently making all the underlying data publicly available. The authors should also clearly acknowledge the privacy and legal issues that surround public access to a laboratory's validation data, to avoid the perception that forensic DNA laboratories are simply being obstructionist rather than complying with

Federal and State statutory requirements and obligations to safeguard the confidentiality of this DNA data.

Limiting the foundational review to only 'publicly available' data is an excessively restrictive measure and there is no statutory or scientific requirement for such a limitation, other than the author's belief.[8] The authors should be open to including a review of available validation summaries as a means to establish validity. To require a review of the data itself rather than a review of summarized results that already exist in both peer-reviewed publications and laboratories' validation summaries is contrary to established practice for assessing validity in other branches of science. And, as previously explained, NDIS participating laboratories are required by the Federal DNA Act to maintain the privacy of forensic samples. Donors of study samples also have an expectation of genetic privacy[9], so the "requirement" by the authors that genetic data be made 'publicly available' in order for an assessment of reliability to be performed contravenes statutory and regulatory requirements under which NDIS participating laboratories operate.

In addition to the mechanisms previously described for the review of the internal validation data (such as audits and assessments), there are also well-established avenues for the disclosure of this information, such as discovery and pre-trial hearings in criminal cases; processes within the criminal justice system designed to protect defendants' rights and also safeguard the privacy of the DNA data.

'Factor space'

The draft report introduces the term "factor space" into the evaluation of foundational validity of DNA mixture measurements and interpretation. As applied in

[8] Beginning at line 2402, the draft report states, "We recognize that there are information and data collected in forensic laboratories that may not yet be publicly available or published. However, we believe for information to be considered foundational, it needs to be reasonably accessible to anyone who wishes to review it."

[9] Federal and State laws also protect individual's privacy rights to their genetic information, *see for example*, the Genetic Information Nondiscrimination Act of 2008, P.L. 110-233, available at https://www.eeoc.gov/statutes/genetic-information-nondiscrimination-act-2008. *See also*, Congressional Research Services (2008) *Genetic Information: Legal Issues Relating to Discrimination and Privacy,* (State Statutes); available at https://www.everycrsreport.com/files/20080428_RL30006_d82c17a8245827846e913181fa76b504bd9ba61d.pdf.

the draft, "factor space" is meant to "describe the factors that influence complexity, measurement, and interpretation reliability – these factors include the number of contributors, the degree of allele sharing, the ratios of mixture components, and the amount and quality of the DNA tested." Despite the new terminology, this is not a new concept for the forensic DNA community. *The FBI Quality Assurance Standards for Forensic DNA Testing Laboratories* (*QAS*) Standard 8.3.2.1 specifically requires laboratories to include samples with a range of the number of contributors, template amounts, and mixture ratios expected to be interpreted in casework. This benchmark has been included in the *QAS* dating back to at least 2011. In addition, SWGDAM and OSAC included this concept in their documents for validation of probabilistic genotyping systems[10] and mixture interpretation validation.[11]

With regard to factor space, you have acknowledged in your Profiles in DNA article from September 2006[12], a common misconception (urban legend #1) is that hundreds or thousands of samples are required to fully validate an instrument or method. Given the wide range of mixture ratios and template amounts evaluated in the assessments of published and internal validation data, there seems to be ample information from which the authors could have made a determination regarding foundational validity without seeming to require an exhaustive amount of data. For example, in the publication *Internal validation of STRmix™ – A multi laboratory response to PCAST*,[13] the data from thirty-one laboratories was compiled that contained over 1,500 three person mixtures that represented over 200 unique combinations and over 1,100 four person mixtures that represented over 100 unique combinations were

[10] *SWGDAM Guidelines for the Validation of Probabilistic Genotyping Systems* available at https://www.swgdam.org/publications, and ANSI/ASB Standard 018, *Standard for Validation of Probabilistic Genotyping Systems* (2020); available at http://www.asbstandardsboard.org/wp-content/uploads/2020/07/018_Std_e1.pdf.

[11] ANSI/ASB Standard 020, *Standard for Validation Studies of DNA Mixtures, and Development and Verification of a Laboratory's Mixture Interpretation Protocol* (2018); available at https://asb.aafs.org/wp-content/uploads/2018/09/020_Std_e1.pdf.

[12] Butler, J. (2006) *Debunking Some Urban Legends Surrounding Validation Within the Forensic DNA Community*, Profiles in DNA; available at https://projects.nfstc.org/workshops/resources/literature/debunking%20validation%20butler.pdf.

[13] Bright, J. et al. (2018) *Internal validation of STRmix™ – A multi laboratory response to PCAST,* For. Sci. Int. Genet. 34:11-24.

evaluated. In each set of data, a portion of the mixture contained minor contributors of less than 1%.

Reliability

It is difficult to summarize the comments received on the draft report's treatment of the issue of reliability. There is considerable discussion of this issue but, in the end, our members were generally disappointed that there was no definitive guidance for the forensic DNA community. Because this is a foundational review, there was an expectation that NIST would have suggested some minimum criteria or threshold for reliability or alternatively, the degree of variation that would be considered acceptable. The treatment of reliability appeared to some, rather circular, as reliability could not be established because of a lack of data which was a direct result of the decision by the authors to only review 'publicly available' data; a decision for which they attribute to their "belief" that such data should be 'publicly available' although internal validation data is not legally or otherwise required to be published or publicly accessible. Additionally, some members/invited guests were concerned about the inclusion of comments that appear to suggest that forensic DNA laboratories are not being transparent in their operations because of the failure to make their internal validation data 'publicly available'. Such an inference is both inappropriate and misleading and could have unintended and undesirable consequences for laboratories that are simply acting in accord with Federal and State legal requirements.

The draft report does not provide clarification on the issue of reliability for the DNA community and appears to have missed an opportunity to do so and provide guidance on gaps or other legitimate areas that would benefit from a review of all available internal validation data.

Postscript on the use of the NISTIR 8351-Draft Report

SWGDAM members are grateful for and benefitted from your presentation on the draft report during the July 2021 SWGDAM Virtual Meeting Round Table. It was at this presentation that you highlighted the inclusion of the following language on page 1 to respond to questions about the use of the draft report:

> "Where our findings identify opportunities for additional research and improvements to practices, we encourage researchers and practitioners to take action toward strengthening methods used to move the field forward. The findings described in this report are meant solely to inform future work in the field."

Attendees at the meeting were understandably concerned about the use of the draft report for purposes of challenging scientific evidence in the criminal justice system.[14] And, as it turns out, justifiably so. SWGDAM is aware that the draft report is being cited as the basis for motions to challenge the admissibility of scientific evidence generated using probabilistic genotyping systems.[15] In fact, there has been a reported decision in a Delaware case, *State v. Hudson*, in which the defendant copied the "KEY TAKEAWAYS" in the beginning of the draft report for the proposition that the expert testimony is based on "unfounded and unsupported pseudoscience."[16] In this case, the Court found the STRmix results satisfied the factors and that "the method utilized by STRmix is reliable as there are internal and external validations checks to insure reliability."[17]

SWGDAM takes no position on the use of the draft report for these purposes. However, in light of the fact that the draft report is being cited in support of challenges to expert testimony in the criminal justice system, SWGDAM requests that NIST publish the public comments now in the interests of fairness and transparency. SWGDAM understood that it was originally NIST's intent to publish the public comments at the conclusion of the comment period but that was postponed to coincide with the issuance of the final report. The public comments could provide valuable information/research for addressing the use of the draft report as justification for claims to exclude expert scientific testimony relating to the probabilistic genotyping

[14] *See* SWGDAM July 2021 Semi Annual Report available at www.swgdam.org.
[15] *See, for example, U.S. v. Quantavious Arnold*, No.: 1-20-CR-244-LMM-CMS (U.S. Dist. Ct., N.D., GA); *People of New York v. Daval Wright*, Indict No. 3167-2019 (Sup. Ct., N.Y. Co.); *People v. Alvin Davis*, Case No. C089567 (Ct. of Appeal, 3rd App. Dist.); *State v. Karim*, Case No. 1816-CR00139-01 (Circuit Ct., Jackson Co. MO). *See also State v. Resiles*, Case No. 14-12657CF10A (Circ. Ct. 17th Jud. Dist. Broward Co.) November 2021
[16] *State v. Hudson*, 2021 Del. Super. LEXIS 636, 640 (New Castle Co. Superior Ct.) October 2021
[17] *State v. Hudson*, 2021 Del. Super. LEXIS 636, *28 (New Castle Co. Superior Ct.) October 2021

systems. Accordingly, SWGDAM heartily requests that the public comments be published as soon as possible, considering that the draft report is being cited as the basis for motions to exclude expert scientific testimony contrary to its intended purpose. SWGDAM also recommends the NIST authors provide additional direction and clarification that the draft report does not suggest that the interpretation of complex mixtures lacks scientific foundation, to stem the tide of court cases seeking to repurpose the NIST authors' intent to prevent introduction of evidence in court.

Conclusion

SWGDAM offers the following general suggestions for the draft report: (1) include the appropriate forensic science practitioners among the author team when performing a foundational study of forensic science disciplines; (2) define the threshold for establishing reliability prior to beginning such a review; (3) perform such a review with all available information and without arbitrarily limiting the information to that which is 'publicly available'; and (4) limit the report to only that information relevant to the foundational review. Specifically, Chapters 5 and 6 of the draft report contain interesting comments, observations and recommendations but may be more appropriate as a separate publication for forensic science practitioners as they are not directly relevant to the foundational review, nor mixture interpretation generally. Additionally, in accordance with the postscript comments, SWGDAM respectfully requests that NIST publish the public comments received to date so that all relevant information may be considered in connection with motions citing the draft report in support of excluding scientific evidence.

Thank you for the opportunity to comment on NISTIR 8351-DRAFT entitled *DNA Mixture Interpretation: A NIST Scientific Foundation Review* and also provide the postscript to our original comments.

Sincerely,

Anthony J. Onorato
SWGDAM Chair

Attachment: SWGDAM Comments (August 2021)

DNA Mixture Interpretation: A Scientific Foundation Review

https://www.nist.gov/dna-mixture-interpretation-nist-scientific-foundation-review

SWGDAM Comments (August 2021)

Line #	Comment	Suggestions
General	NIST should have included practical experience in authorship: None of the authors have any experience working cases in an accredited crime laboratory. Authors with this critical experience should have been included. Having these individuals at your disposal for participation, however, not include them in your drafts for the last 2 years is not an appropriate use of stakeholder input. Example: OSAC on Forensic Science (NIST) has embedded stakeholders who are active in the writing of standards and have a vote on the outcome. Those with accredited forensic experience did not in the authorship of this document.	Include those with practical accredited forensic laboratory DNA mixture interpretation experience as authors of the report.
General	NIST should not restrict data to that publicly available: Determining that data had to be publicly available or it cannot be used is not an appropriate decision for establishing validity.	NIST should expand its definition of data that is considered to establish foundational validity. Included in that definition should be data held internally in forensic DNA crime laboratories that is available for on site review.
General	NIST should visit labs to review data on site: No data was reviewed on site.	NIST should personally visit accredited forensic DNA laboratories and consider firsthand observation of data as a valid means to assess foundational validity.
General	NIST should expand input to include auditors: No DNA lab auditors who review internal validation data were interviewed.	NIST should include opinions from DNA auditors who have reviewed validation data.
General	NIST should expand review of documentation: No lab audit documents were reviewed nor taken into consideration. Better disclaimer statements to prevent	NIST should include improved disclaimer statements to include concise language to the fact that NIST is not saying DNA mixture interpretation is not valid.

	misuse or misunderstanding (deliberate or otherwise) of the report: The lack of drawing a conclusion will lead some within an adversarial system to determine that complex DNA interpretation is not supported, or labs are not doing a thorough job in validations, or labs have something to hide. NIST's description regarding what they are not saying and how their report should not be used should be better spelled out.	
General	A review of published literature is not indicative of the reliability of work, when there is no stated requirement to publish internal validations, and in fact publication was discouraged. An example is the journal of the AAFS (editor Michael Peat) sent a letter telling scientists that internal validations would no longer be published (approximately 2005).	NIST should acknowledge that using only published data is an excessively restrictive measure to require for data and open its interpretation considerations to include forensic laboratories unpublished data.
Overall	There is minimal acknowledgement of QAS, SWGDAM, and OSAC validation requirements and guidelines. While these documents don't tell how to do an experiment, as the local scientists should design experiments to be relevant to their sample types, etc., they do provide common framework that is readily apparent to stakeholders and auditors.	These should be more readily acknowledged throughout (e.g., in Table 4.9).
Overall	This document seems to emphasize what's referred to as a lack of reliability and deemphasize 1. The challenges and inability to publish validation data, 2. The existence of reams of data within forensic laboratories, 3. The independent and external audit system to which all CODIS-associated crime labs are held, as well as other audits, such as ISO/ANAB.	The authors should include these overlooked and/or minimized but rather quite important and impacting points.

	Also noteworthy is the fact that the role of the criminalist is largely to perform casework and testify when needed. It is not often (with some exceptions) to vet, procure, validate, and publish on new technology, despite the efforts of many to nevertheless squeeze these additional tasks in.	
126-128	It would appear based on the summary from NIST 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, including its use of probabilistic genotyping software (PGS) systems." However, this statement in my view doesn't fully answer the original question – Did NIST find or not find that "established scientific laws and principles exist to support the methods" practitioners are currently using for DNA mixture interpretation? I understand there were gaps found from this study in terms of gathering sufficient empirical data from laboratories in applying these methods, but is NIST also asserting with this report that the scientific laws and principles also do not exist?	I would argue that there is indeed general consensus in the scientific forensic community that underlying scientific principles do exist and are reliable to use in DNA mixture interpretation, and that PGS systems using a likelihood ratio construct that apply these same principles also exist and are available for laboratories to utilize. In fact, these same principles and their application in PGS systems are well characterized in the peer-reviewed scientific literature and also described/referenced at length in this report (chapters 2 and 5). Although it is certainly important to note that improvements are needed particularly in the area of making laboratory internal validation of these methods more publicly available for independent review, this does not negate the fact that the methods do exist and contain the underlying scientific principles to interpret DNA data. I would also argue that empirical data also exists but is not necessarily publicly available. Just because NIST could not evaluate this data does not mean that this data does not exist or does not demonstrate "reliability" as defined in this report. NIST could not know that since they did not have the empirical data to evaluate in the first place. My suggestion here would be to address this part of the question in its summary to emphasize this distinction for full context so it is not misinterpreted by the rest of the forensic community.
609-611	This document doesn't clearly address this point. Rather it is vague and over-shadowed by Takeaway 4.4.	This needs to be flushed out.
610, 1028	Draft report indicates that the review does not concentrate on interpretation of single source or two person mixtures involving significant quantities of DNA. In order for the readers of the report to determine exactly what the scope of this document is, please define what are considered "significant quantities of DNA".	Define "significant quantities of DNA"
722-724	The authors state, "Note that our original goal in this review was	

	external and independent assessment of reliability based on publicly available data that met our selection criteria." Each forensic lab is required by the DNA ID Act to maintain the privacy of forensic samples, and donors of study samples also have an expectation of genetic privacy, so the "requirement" by the authors that genetic data be made publicly available for external review in order for an assessment of reliability to even be performed is not feasible. Moreover, requiring that a lab's PG data be made publicly available for external review ensures that the authors' definition of reliability will not be achieved. Publication in peer-reviewed journals has been the hallmark of an independent assessment of the validity and soundness of basic scientific research and method development. Moreover, the PCAST report called for peer-reviewed publication of the "foundational validity" of forensic methods and suggested "that NIST explore with one or more leading scientific journals the possibility of creating a process for rigorous review and online publication of important studies of foundational validity in forensic science." This document would benefit from expanding the "assessment of reliability" to include the publication of internal validations, developmental validations, and inter-laboratory studies.	
724; 3204; 3250; 605-611; 754-755; 3255	…that met our selection criteria What are these criteria? They are never stated. Instead, it is later said (3204) that there were no criteria. PCAST apparently had criteria but the authors are unclear	Clarify what the intention was. It is confusing (and circular) to say the criteria were changing and that there were no criteria for reliability. Were the authors attempting to establish national thresholds and criteria for universal application? Are the authors challenging the conclusions of PCAST? Confusing.

	as to what that was. Nevertheless, the authors claimed to be in agreement with PCAST (605-611). What about 3PMs, designated reliable by PCAST (3255)? Are these authors in agreement with PCAST on 3PMs?	
741-746	I agree with this recommendation, but the report falls short with how laboratories and where to publish its validation data and in what format? Keep in mind making this data public comes with certain privacy concerns, informed consent, or even IRB considerations with the actual genetic profiles that were collected/generated in-house.	I would suggest that a more practical approach would be for laboratories to participate in an inter-laboratory survey using a universal data set (perhaps generated at NIST) seems much more reasonable and these results could be published without the privacy concerns related to the data. It would also serve to evaluate a laboratory's application of its PGS method and associated interpretation guidelines that could then be compared against other laboratories or even other PGS systems.
741-746	How do we ensure the reviewers or "users" have the expertise for a rigorous scientific review of each laboratory's PGS validation studies in order to establish an acceptable level of reliability? What are the requirements to be considered an expert user to conduct these reviews?	I believe this knowledge and expertise is equally important for the "users" to possess as well as the "providers" who publish this data. This should be included in this takeaway.
748-755	The critique of the published studies so far in this report is that they "lack sufficient detail", but this report falls short in recommending what would be "sufficient". What specific criteria is NIST looking for? Until it knows what the requirements are, how can a laboratory or its stakeholders ever feel confident that it has met "an acceptable level of reliability"?	The relevant scientific forensic community needs to further research this topic and develop standard criteria based on peer reviewed consensus (e.g., OSAC) for laboratories to reference. This should be included in this key takeaway.
754-755	Statement that no threshold or criteria established to determine an acceptable level of reliability	Guidance documents on internal validations do provide information on how to assess for reliability and then each laboratory performs testing and develops an appropriate SOP based on the reliability shown within their lab. Remove no threshold and say established within the laboratory's internal validation studies.
771-773	Include validation performance results in case files and reports	This is unnecessary, and the reason validation is done and then even summarized so that it can be reviewed to include the data when an appropriate party is authorized to come onsite for viewing which is done when requested. The information is available to review and just putting summaries in case files would not truly answer the reliability question as it's being tied to the actual data and not just the results.

769-773	It is unclear in this key takeaway and throughout the report what role if any laboratory accrediting bodies have as it relates to reviewing validation studies as part of an external assessment? Was this intentional? Memorialization of laboratory internal validations through external assessments has always been the means for laboratories to demonstrate compliance using internationally recognized and accepted accreditation standards and requirements but also to receive the much needed independent review with data that is not published.	Is it possible for NIST to provide specialized training to external auditors or even give the accrediting bodies the "reliability" metrics needed so that they in turn could adequately provide the "independent review" of this data through the standardized process that is already in place with laboratory accreditation? The elements of this validation review process could be standardized for all labs using the reliability criteria and thresholds discussed in this document. Even though this review would not necessarily be made public initially, any gaps or deficiencies could be made public to the court as part of the routine discovery process where its reliability would matter most in a forensic context. Additionally, NIST could monitor any trends with these accreditation reviews to determine what gaps or improvements would be needed. This would serve as a more reasonable solution in my view in the short term. My suggestion here would be to at least mention the role of accreditation and external assessments which are well-positioned for this task so long as the auditors have the subject matter expertise to rigorously conduct these independent validation reviews.
849-852	Individual laboratories would need to know how the sensitivity of methods… This type of information is mainly disseminated through the literature but may also be presented at conferences and shared through communication exchanges between colleagues. Practitioners can use that information to draw comparisons and expectations.	It's difficult to determine if the authors feel this is unavailable or they are just reiterating the existing availability. Clarify.
870-871	Using likelihood ratio as a standalone number without context can be misleading	LR is a statistic used like random match and CPI so why would LR now be viewed so differently? Stats have been utilized in reports for over 25 years and explained in court and to our customers. Standard guidance language for LRs has been provided so I don't see the need for this takeaway.
1146-1147	Indicates that 'most laboratories do not publish data from their validation studies.' However, it does not mention that laboratories would find it difficult to publish their internal validation studies due to the fact that they are not novel work.	Acknowledge that laboratories may find it challenging to publish internal validation studies due to the fact that they are not novel work, and may not be accepted for publication into peer-reviewed journals as is covered later within Chapter 3.
1164-1166 and 1172-1173	NONE of the members of the review team have ANY hands on DNA mixture interpretation experience in an accredited Forensic DNA laboratory	The review team is missing a key perspective in which to review the issues and that is from a currently qualified DNA examiner with experience in mixture interpretation
1346-1347	"DNA information can assist both law enforcement (investigative)	Add defense/exculpatory items and mass disaster identifications to this statement about usefulness of DNA information.

	and prosecutorial (evaluative) aspects of the criminal justice system." Does not mention that DNA information can also assist defense or provide exculpatory information for an accused person, and be useful in the course of mass disaster identifications.	
1756	DNA analysis is not based on "belief" but on analysis and evaluation	When a DNA analyst evaluates a mixture and determines that a major component…
2123	Empirical data to assess fitness of purpose of analyst's LRs	This is foundationally set with validation and SOPs within a lab and then applied to information specific to each case. It is exists so warranted makes it sound as if it does not.
2123-2124 (KT 2.6)	The authors state, "Likelihood ratios are not measurements. There is no single, correct likelihood ratio (LR). Different individuals and/or PGS systems often assign different LR values when presented with the same evidence because they base their judgment on different kits, protocols, models, assumptions, or computational algorithms. Empirical data for assessing the fitness for purpose of an analyst's LR are therefore warranted."	The implication is that since different LRs may be obtained from the same evidence, that the LR is not reliable since there is not one "correct" value. Peak heights are measurements. If the same sample is injected multiple times within the same lab, or in different labs, the peak height will not be the same. This is because of a variety of factors that impact peak height, however it doesn't make the peak height unreliable. Similarly, LRs determined from the same evidence will also differ within the same lab, or in different labs, due to a variety of factors (assumptions, kits, models, etc…) and this also doesn't automatically make the LR value unreliable. While there is certainly a need for additional inter-laboratory studies beyond the "multi-lab response to PCAST", this document would benefit from the acknowledgement of peer-reviewed publications of validation data that demonstrate how the LR is dependent on a laboratory's procedures/assumptions and how these studies demonstrated the reliability of an LR value.
2352	"These propositions H1 and H2 are required to be mutually exclusive and exhaustive." I don't believe that there are any requirements for a single H1 and H2 to be completely exhaustive. There may be multiple H1 or H2 propositions that could explain the evidence, resulting in multiple LRs to be run, and none may be completely exhaustive of all potential propositions	Remove the phrasing 'and exhaustive.'
2402-2405	Similar to 1146-1147, there is no mention that laboratories may have difficulty publishing internal validation data at this point in the document. In addition, there is no mention that NIST did not attempt	Remove these lines as the sentiment is already mentioned in other areas of the document. Mention in the previous paragraph 2399-2400 that NIST made no effort to solicit labs for data.

	to request data from laboratories, and that they only used web searches to deem what may have been 'available'.	
2486-2488	Authors are acknowledging that additional internal validation data likely exists, but they chose to conduct the scientific foundation review using only publicly available information.	Choosing to not even try to obtain the data to look at (during the public webinar, it was asked how many of the authors of the 60 prob gen publications were contacted to see if the data was available to review, and Dr. Butler replied none of them had been contacted). This is an irresponsible approach to conducting a scientific foundation review and is doing a dis-service to forensic science.
2487-2488	Should mention here that NIST did not make an attempt to solicit or evaluate laboratory data.	This scientific foundation review is limited to publicly available information, and no effort was made by NIST to request the underlying data from any laboratories.'
2490; Table 3.2	DC Dept of Forensic Sciences STRmix v2.3 validation is incorrectly stated as being ABI 3500. Samples were processed on the 3130xl. See page 1 of STRmix v2.3 Parameters report.	Identifiler Plus, ABI 3130xl
2707-2710	"In recent years, DNA analysts have increasingly relied on one of several available PGS systems to assign a numerical value to their mixture result based on a pair of propositions selected by the analyst." Should mention here similar to 2770-2772 that the assessment using PGS is used in conjunction with the analyst's interpretation and training and experience.	Indicate that PGS is used in conjunction with the analyst's interpretation and training as part of the laboratory's protocols.
2848	"factor space" and "factor space coverage" – these terms have never been used in discussing validation studies, why are they being introduced in a review publication.	
2901-2903	…an eight-person mixture involving only 10 pg total template DNA, then DNA analysts might refrain from interpreting such a sample because it has not been covered in any of their validation experiments This is an extreme, far-fetched example well beyond where forensic laboratories would test reliability.	Change to something more realistic, like 5PM with 100pg.

2922	Degree of reliability is assessed through empirical data	The data is made available by laboratories as requested through appropriate channels. If the entire foundation study was going to summed up by saying we just don't know the reliability of mixture interpretation due to the lack of publicly available data, then when did this become the way to measure and assess? NIST provided trainings on validations for years and never did I hear the major takeaway be that the data must be made publicly available to truly assess reliability. This seems so disingenuous to state when going into the foundational review, the writers knew the current state of validations and how they were handled as they conducted many workshops on the very topic over the years. So, make it available publicly and then who is reviewing it? What knowledge base does that individual/group contain to comment in a useful manner on reliability? It's important for individuals to be educated on the entire process and not just mark points of a plot for each mixture to cover space.
2956	Proficiency tests really don't ask how reliable or trustworthy a method is, except perhaps on a very basic level. Reliability is addressed in internal validation, before proficiency tests on a given method are initiated. Proficiency tests are meant to address analyst competency. When an analyst has an issue with a proficiency test, that is addressed directly with the analyst. It's possible that there could be an improvement to a written procedure to address the root cause of the issue. However, it's generally unlikely that such a situation is due to an unreliable method. This source of data seems minimally helpful to examine foundational reliability of a technology.	The authors should instead seek data that challenges the systems, e.g., internal validation data from forensic laboratories.
2968	Why did the reviewers limit their analysis to publicly available data? It would be very rare for laboratories to make their data publicly available -	Re-do chapter four after looking at data now available. How can it be considered a review if sufficient data was not looked at.
2983-2986	This is indeed the case but seems to be underemphasized throughout the document and is at the heart of the issue here – the review was NOT conducted b/c the committee did not have internal validation data from forensic laboratories.	Work with forensic laboratories to collect data to perform the review.
3021-3022	"a great deal more information is now available" implies that this	Remove sentence

	information was not available at the time of PCAST and that the 8 articles cited in PCAST were the only ones; however, 18 of the articles in table 4.3 were published prior to PCAST	
3069; Table 4.5	DC Dept of Forensic Sciences STRmix v2.3 validation is incorrectly stated as being ABI 3500. Samples were processed on the 3130xl. See page 1 of STRmix v2.3 Parameters report. Table 4.5 incorrectly states that mixture ratios were not explicitly stated in the summary document for the STRmix v2.3 validation. See page 9 of the STRmix v2.3 Validation summary where it states that a summary of the profiles analyzed for the sensitivity and specificity plots are in Appendix 3. Appendix 3 starts on page 38 of same document.	Identifiler Plus, ABI 3130xl For sensitivity and specificity studies, i.e., Section D studies: 17 single source, various DNA quantities (quantities listed in Table 4.5 are additional single source samples from Section A and Section B studies) 2 person mixture ratios: correctly listed in Table 4.5 3 person mixture ratios: 5:2.5:1, 20:1:1, 3:1:1, 20:10:1, 3:1.5:1, 10:1:1, 10:5:1, 5:1:1 4 person mixture ratios: 5:5:1:1, 10:5:2:1, 2:2:2:1, 10:1:1:1, 5:5:5:1, 1:1:1:1, 10:5:5:2, 5:2:2:1, 10:10:1:1, 5:2:1:1, 2:1:1:1, 10:5:5:5, 3:1:1:1, 10:10:10:1, 2:2:1:1, 5:1:1:1, 3:2:1:1, 3:2:2:1,
3069; Table 4.5	Table incorrectly states that total DNA quantity, and mixture ratios were not explicitly stated in the summary document for the DC Dept of Forensic Sciences STRmix v2.4 validation. See page 7 of the STRmix v2.4 Validation summary which states that a summary of the profiles analyzed for the sensitivity and specificity plots are in Appendix 3. Appendix 3 starts on page 37 of same document.	For sensitivity and specificity studies, i.e., Section D studies: 2 person mixtures: DNA quantities 300 and 600pg 2 person mixture ratios: 1:1, 1:2, 1:3, 1:5, 1:7, 1:10, 1:15, 1:20, 1:25 3 person mixtures: DNA quantities 200, 500, 900 pg 3 person mixture ratios: 3:1:1, 20:10:1, 3:2:1, 10:5:1, 5:1:1, 10:2:1 4 person mixtures: DNA quantities 100, 200, 400, 600, 700, 800, 900, 1000pg 4 person mixture ratios: 2:2:2:1, 20:5:2:1, 5:1:1:1, 5:2:1:1, 5:5:5:1, 4:3:2:1, 3:3:2:1, 10:5:3:1, 2:2:1:1, 20:10:1:1, 3:1:1:1, 7:1:1:1 5 person mixtures: DNA quantities 300, 600, 1000pg 5 person mixture ratios: 10:5:2:1:1, 5:4:3:2:1, 10:10:10:10:1, 10:10:5:1:1, 5:5:5:2:2, 20:1:1:1:1, 2:2:2:1:1, 3:1:1:1:1, 5:1:1:1:1
3069	Reference Table 4.5 (Line 3069) Factor space coverage of information in internal validation studies listed in Table 3.2. • The Michigan State Police has effectively covered much of the "factor space" recommended by this report, but not all of that work was publicly available. The report should be corrected to reflect the actual	Coverage of Factor Space from Validation: STRmix™ and PowerPlex® Fusion

	factor space covered. The profiles outlined here are all lab-created samples- these charts do not contain the additional testing conducted on adjudicated samples.	

C Range	# Samples	Total DNA Quantity Range (pg)	Mixture Ratio Range
1	6	500	N/A
		600	N/A
		150	N/A
		75	N/A
		50	N/A
		25	N/A
2	18	500:500	1:1
		909:91	10:1
		882:118	7.5:1
		833:167	5:1
		714:286	2.5:1
		500:500	1:1
		714:286	2.5:1
		2,143:857	2.5:1
		909:91	10:1
		882:118	7.5:1
		833:167	5:1
		714:286	2.5:1
		500:500	1:1
		909:91	10:1
		882:118	7.5:1
		833:167	5:1
		714:286	2.5:1
		500:500	1:1
3	22	625:312.5:62.5	10:5:1
		833:83:83	10:1:1
		769:154:77	10:2:1
		625:312.5:62.5	10:5:1
		476:476:48	10:10:1
		454.5:454.5:91	10:10:2 [5:5:1]
		400:400:200	10:10:5 [2:2:1]
		333:333:333	10:10:10 [1:1:1]
		500:334:167	3:2:1
		351:234:117	3:2:1
		234:156:78	3:2:1
		174:116:58	3:2:1
		78:52:26	3:2:1
		833:83:83	10:1:1
		740:185:74	10:2.5:1
		625:312.5:62.5	10:5:1
		540:405:54	10:7.5:1
		476:476:48	10:10:1
		444:444:111	10:10:2.5 [4:4:1]
		400:400:200	10:10:5 [2:2:1]
		364:364:272	10:10:7.5 [4:4:3]
		333:333:333	10:10:10 [1:1:1]
4	19	588:294:59:59	10:5:1:1
		769:77:77:77	10:1:1:1
		588:294:59:59	10:5:1:1
		385:192:38	10:10:5:1
		468:351:234:117	4:3:2:1
		312:234:156:78	4:3:2:1
		232:174:116:58	4:3:2:1
		104:78:52:26	4:3:2:1
		769:77:77:77	10:1:1:1
		769:77:77:77	10:1:1:1
		714:143:71:71	10:2:1:1
		588:294:59:59	10:5:1:1
		455:455:45:45	10:10:1:1
		435:435:87:43	10:10:2:1
		384:384:192:38	10:10:5:1
		323:323:323:32	10:10:10:1
		312.5:312.5:312.5:62.5	10:10:10:2 [5:5:5:1]
		286:286:286:143	10:10:10:5 [2:2:2:1]
		250:250:250:250	10:10:10:10 [1:1:1:1]

Coverage of Factor Space from Validation: Use of STRmix™ Software for Profile Interpretation of Challenging Samples

<table>
<tr>
<td></td>
<td></td>
<td>
<table>
<tr><th>NoC Range</th><th># Samples</th><th>Total DNA Quantity Range (pg)</th><th>Mixture Ratio Range</th></tr>
<tr><td>1</td><td>31 (Degraded)</td><td>500</td><td>N/A</td></tr>
<tr><td rowspan="3">2</td><td>4 (Biological Relatives)</td><td>250:250</td><td>1:1</td></tr>
<tr><td>4 (Biological Relatives)</td><td>400:100</td><td>4:1</td></tr>
<tr><td>4 (Biological Relatives)</td><td>455:45</td><td>10:1</td></tr>
<tr><td rowspan="8">3</td><td>4 (Degraded)</td><td>167:167:167</td><td>1:1:1</td></tr>
<tr><td>6 (Degraded)</td><td>357:71:71</td><td>5:1:1</td></tr>
<tr><td>2 (Low-level, Heterozygosity)</td><td>385:96:19</td><td>20:5:1</td></tr>
<tr><td>2 (Low-level, Heterozygosity)</td><td>454:23:23</td><td>20:1:1</td></tr>
<tr><td>3 (Profile Rarity)</td><td>385:96:19</td><td>20:5:1</td></tr>
<tr><td>3 (Profile Rarity)</td><td>454:23:23</td><td>20:1:1</td></tr>
<tr><td>4 (Biological Relatives)</td><td>167:167:167</td><td>1:1:1</td></tr>
<tr><td>4 (Biological Relatives)</td><td>333:83:83</td><td>4:1:1</td></tr>
<tr><td rowspan="4">4</td><td>4 (Biological Relatives)</td><td>313:156:31</td><td>10:5:1</td></tr>
<tr><td>2 (Low-level, Heterozygosity)</td><td>370:93:19:19</td><td>20:5:1:1</td></tr>
<tr><td>2 (Low-level, Heterozygosity)</td><td>434:22:22:22</td><td>20:1:1:1</td></tr>
<tr><td>3 (Profile Rarity)</td><td>370:93:19:19</td><td>20:5:1:1</td></tr>
<tr><td>3 (Profile Rarity)</td><td>434:22:22:22</td><td>20:1:1:1</td></tr>
</table>
</td>
</tr>
<tr>
<td>3074</td>
<td>Not enough publicly available data…</td>
<td>Who determined that reliability is only able to be assessed with publicly available data? To embark on a foundational study in this arena knowing that it was not standard practice in labs to make all their validation data publicly available is ridiculous quite frankly. Why would discussions not have occurred for many years (all their trainings/workshops offered) on the need to put data in a public forum? The labs are being held to criteria that seems nearly just invented for the direct purpose to dismiss PGS.</td>
</tr>
<tr>
<td>3074 and 3425</td>
<td>- Consider that within the factor space, which is very large, there is a continuum of data from the more reliable to the less reliable.
- Despite this size and complexity, there are components of the factor space that are reliable, even within many very complex mixtures. Within a complex mixture, components of that mixture can be reliable, while other components are not.</td>
<td>Acknowledge that DNA mixture interpretation is reliable for those mixtures with a major contributor above a certain ratio and those mixtures where all of a potential contributor’s alleles are present above a certain threshold. Some component of the factor space, where mixtures are above a certain threshold, support DNA mixture interpretation that is reproductible and hence valid.</td>
</tr>
</table>

	- That among the more reliable data (good peak height, quality, etc.) that the major contributors and those with the greater reliability can be determined, not undermined by the minor contributors with less reliable data. - There is a component of the factor space that is reliable. In my opinion that is: o Major contributor: Major contributors for complex mixtures where there is sufficient separation of peak height o All alleles present: Inclusion with a statistic for validated systems where all peaks above a predetermined threshold are present (no drop out, all peaks from a person's profile are represented in a mixture) o Less weight or perhaps not determined (inconclusive) for the far more difficult mixtures, where some of the individual's profile is not represented (drop out, degradation, below threshold, etc.) - By declaring at least some of the factor space reliable, this acknowledges the obvious, while pinpointing the areas where additional work needs to be done. At present, the conclusion of nothing has sufficient data available makes it appear a thorough job has not been done, as the "Major contributor" component with sufficient peak height, and the "all alleles present" are included examples above have not been acknowledged.	
3135	Using CTS PT results	It should be noted that students and possibly other untrained individuals may participate in the PT so it's not an appropriate measure to gather the results as a whole and make any assessment
3201-3204	Again, a mention of lack of publicly available data, but no mention that NIST did not reach	Be clear about whether NIST would be able to be determined as the 'user' who could establish a degree of reliability if more data was made public as suggested.

	out to laboratories to obtain data. Would NIST have been able to perform an assessment of reliability if the data had been made available? Would NIST be considered the 'user' in this sense to be able to assess the degree of reliability, validity, and whether that information is fit-for purpose, in line with key takeaway 4.2?	
3204-3207, and within 3425 (KT 4.4.)	"This is particularly true without an established and accepted criteria for reliability with complex mixtures involving contributors containing low quantities of DNA template or where there is a high degree of allele overlap among contributors" From the key takeaway 4.4 "…there is no threshold or criteria established to determine what is an acceptable level of reliability." There is no suggestion of who would/could create such criteria for reliability. Or that laboratories that have evaluated and empirically tested their data have determined reliability within their own factor space and are applying it accordingly. This is also somewhat in conflict with key takeaway 4.1 where it says that "The degree of reliability of a component or a system can be assessed using empirical data obtained through validation studies, interlaboratory studies, and proficiency tests." The phrasing also has the potential to have a detrimental effect as commentary in relation to admissibility hearings. There is a prong of evaluation of error rates and reliability that is part the Daubert standard in determining whether a particular scientific technique (like PGS, which is still being evaluated in many	Add recognition that laboratories have established reliability as a user of PGS within their laboratories by performing internal validation studies on empirical data. This would be in line with key takeaway 4.1 from this chapter.

	jurisdictions). This makes it sounds as if there is no standard or even suggestion of how a laboratory would determine reliability. Despite recommendations from ISFG and guidelines from SWGDAM and others that state that a laboratory's internal validation is recommended in order to evaluate a PGS software in order to determine if it is reliable for application to casework.	
3209	Bracketing approach and factor space	This is a foundational study so I would not expect terms such as bracketing and factor space to be introduced here but would be better suited in a document for future improvements/studies. Use the terms that are established in our community and widespread b/c it's a foundational look. Bracketing is done in our validations but maybe the term was not directly named as such.
3222; 3541-3543;3580-3582	Reliability of a specific LR number… A specific LR is specific to the population databases used, the propositions, theta, etc. Approaches involving sampling (e.g., HPD) will not give you the same exact number either. The emphasis, as has been the case, should be on trends and variance among data with repeated tests. Statistics are estimates. Just because we would get the same number with repeated RMP or CPI tests, holding the approach and databases constant, doesn't make those numbers the true and right. They are all estimates and conveyed as such to the legal community. Trends establish expectations. This includes understanding false neg and false pos behavior. This is something explored by forensic laboratories in validation and additional studies, but this report seems to purport it as overlooked.	Per webinar given by Dr. Butler on July 21, 2021, LRs will be the subject of a future review. Perhaps these sections should be removed and addressed later as this topic seems incompletely vetted and explained in this document.
3227-3229; 3479-3482	1. These data are available in individual forensic laboratory validations and are frequently explained/addressed in publications as well (7486, Bright	Update to convey correct information. Add: Bright, Stevenson, Curran, Buckleton FSIG 2015 14:187 Bright, Turkington, Buckleton FSIG 2010 4:111

	et al FSIG 2015, Bright et al FSIG 2010 – latter two should be added). Without having examined actual internal validations, the authors seem unaware. 2. Defense experts are provided all data for independent review.	
3235-3237	LR results can be externally and independently evaluated to be reliable… By other laboratories generating their own data to test. When numerous laboratories, independent of one another, generate their own data and find LRs to be a reliable statistic, that provides greater strength to speak to reliability than a lab reanalyzing another laboratory's work. This work is shared and discussed at conferences and various other meetings and when novel, published.	Update to convey correct information.
3235-3236	States that "LR results cannot be externally and independently demonstrated to be reliable without access to underlying performance data." The report makes no suggestion as to what bodies would be available to do this external or independent reliability assessment.	Provide a suggestion, would NIST be able to do these independent reviews? Is there a suggestion that funding be provided to institutions to be able to do this type of assessment if data was made publicly available?
3237-3241	Later in the paragraph it mentions that "To establish and support clear reliability boundaries, data need to be available to users of the information (e.g. DNA analyst or stakeholders using their results) and acceptable levels of reliability must be decided upon by the user." This statement seems to imply that these users mentioned (DNA analysts and stakeholders) do not have access to data that helps to determine reliability. However, DNA analysts and stakeholders routinely have access to internal validation data in order to evaluate the data in relation to their case-specific results. This should be pointed out.	It should be mentioned that DNA analysts routinely have access to internal validation data through the course of training and working in a laboratory and stakeholders routinely receive access to internal validation data for review by their own experts through the course of discovery requests during court proceedings.

3237-3241	To establish and support clear reliability boundaries… Forensic laboratories generate and assess data sets addressing such questions. These are obviously the "users" of the information as well as stakeholders involved in the court system. All have access. Levels of reliability are established prior to implementation (verbal scales (SWGDAM), interpretation procedures, policies, etc.).	Update to correctly convey such information.
3264-3270	Enable a user to scrutinize…what is actually happening in casework settings. 1. Stakeholders have access to all the validation data they request (e.g., defense community through discovery) 2. If NIST is a "user," tasked by Congress to assess reliability, then NIST should have asked forensic laboratories for data. 3. Who else is a "user"? The authors seem to be creating an issue that there is an issue of access to "anyone." Who is anyone and why does anyone need access? 4. The language "what is actually happening in casework settings" seems inflammatory and to insinuate there is a concealed problem. 5. Here, there seems to be an inherent contradiction as published data are stated in lines 3217-3232 to not meet the authors' criteria for testing.	Who are the users? State who and why they need access. The raw data contain personal genetic information not for public viewing; this is protected by federal and state laws such as GINA and HIPAA. Furthermore, the data require purchased specialized software and expertise to process and understand. To actually perform the foundational review, the authors should ask forensic laboratories for what they need. Point 4 – this unnecessarily negative language that should be removed. Point 5 – clarify.
3288-3294; 3369-3370	This data exists in numerous forensic laboratories across the US. The catch-22 here is that it is not publishable, and therefore not publicly available in its various forms.	It seems the authors could arrange to have access with agreements to treat individual profiles confidentially (to protect the privacy of the study donors) to complete the Congressional task of a foundation review.
3302	Locating… The data exists and is in forensic laboratories.	This report doesn't address that it has located the vast data across the country. Rather, it states that it wasn't publicly available. If the study is to be conducted, the authors should simply organize efforts to obtain it. Otherwise, the congressional task has not been performed.

3336-3350	This paragraph is an oversimplification. A publication addressing sensitivity is not addressing mixture deconvolution. Unique alleles are only one aspect of mixture interpretation and represent a simple way to convey sensitivity in a publication. This is not intended to be used as a measure of deconvolution. While it may be true that publications on comparing old versus new system sensitivities are lacking, this is not necessary nor novel/publishable information. Typically, forensic laboratories make direct comparisons when needed to their own old versus new approaches, especially as related to the FBI QAS standards.	Delete entire paragraph. There is not helpful and does not move the field forward.
3367-3369; 3421-3422	Summaries are simply that. Here summaries are criticized for representing exactly what they intend to – a summary. Furthermore, the purpose of the summaries is associated with the FBI QAS. This drives the purpose and focus. The stakeholders are the laboratory staff, auditors, and the courts, all of whom have access to the actual validation.	This seems nonsensical. A summary is a summary. If there is a vision for a different layout that would result in some meaningful benefit to a named stakeholder (more details on "factor space"), that has not been shared in this document. Rather, simply providing a criticism seems unproductive and doesn't move the field forward.
3369-3370	Lack of availability of data	I feel like a broken record, but it is continually mentioned in this document. The data is made available when requested within the proper channels. Exactly who is doing these independent assessments? Have these individuals performed the testing and do they understand then how the laboratory decided to implement the data appropriately within their SOPs so it's not just simply an external data review.
3374-3375	Given that the authors didn't actually obtain any validation data (only a handful of documents summarizing work), it seems presumptuous to say that allele sharing is missing from many validations. In fact, the authors have no data to indicate one way or the other.	This statement is fact less and should be removed.
3412-3413	Report incorrectly states that DNA quantities and mixture ratios	See information in box above, which is included in the DC DFS STRmix v2.4 validation summary

	were not stated in the public documents	
3425	NIST states "there is no threshold or criteria established to determine what is an acceptable level of reliability."	NIST should have determined what reliability was, rather than leaving it to someone else. If they are not in a position to say what reliability for DNA mixture interpretation is, who is?
3425	Takeaway 4.4 - There is no threshold or criteria … The authors really don't know this without having seen any validation data and corresponding interpretation criteria of various forensic laboratories. For example, this statement is factually incorrect considering my laboratory.	These conclusions are invalid. Collect the data and evaluate before drawing fact-based conclusions.
3425	Takeaway 4.4 - Publicly available information lack… to independently assess… Reliability of foundational concepts is not typically assessed in this manner in other scientific areas. While this is changing for big data (MPS sequencing in certain fields), it has not been the case to "reanalyze" someone else's data. Rather, especially with peer-reviewed published results, it is up to the "user" to generate their own data to assess reproducibility, etc. A publication contains enough information to enable this long established practice for independent verification. Rather, the authors here seem to be interested in seeing whether they agree with each individual lab's nuances to the basic scientific approach. This is impractical and unnecessary, as this is the role of the courts (via discovery, defense experts, etc.). Furthermore, while this committee may have been unable to "publicly" access data, it is regularly shared through discovery requests.	Authors should address standard practices in place and why they feel this must be handled differently.
3439	Call for collaborative approach	Laboratories routinely share validation studies with each other. The reference paper is going a step further and attempting to streamline each

		lab's validation work by being able to possibly adopt the validation done by a lab and then run more of a performance check (dependent on what is being adopted). Data is shared routinely with labs and it's written as if making them public is the only avenue to have this effectively happen.
3441	Critically assessed by other scientists	Audits, outside experts are ways in which this is done and has been done for many years
3447	Footnote 23 1. Many forensic laboratories don't have the infrastructure to have websites, let alone manage large data storage in that format. 2. More importantly, it is a genetic privacy (GINA/HIPPA/etc.) violation that is brushed off in this document. 3. "credible parties in a timely manner when requested" sounds like there is a back story not being told here.	This footnote does not provide any helpful information and does not contribute to moving the field forward. If it cannot be reworded to avoid any negative implications, it should be removed.
3454	Table 4.9 - On PGS validations, collection of common data to demonstrate performance and ultimately reliability can actually be accomplished as Bright et al have shown.	The authors should address this.
3454	Table 4.9 - On recommendation for internal validation data sharing. This is a reasonable suggestion if it can be done in a way that complies with privacy laws.	NIST should vet and present this idea further to provide a practical solution to "move the field forward." For example, a CODIS-controlled site.
3454	Table 4.9 - On recommendation for more challenging proficiency testing. This is a reasonable recommendation but may not be practical for proficiency testing companies.	NIST should vet and present this idea further to provide a practical solution to "move the field forward." For example, NIST partner with the private companies to provide a model that allows for reproducibility across all samples in a lot and fair and expedient scoring of expected results. Seems like something NIST could play a major role in and is currently a missed opportunity.
3455	2nd recommendation	Constant need to publish data and have an independent assessment of PGS performance. Who is doing this assessment?
3458-3459 (KT 4.5)	The authors are suggesting that more complex and/or low-template components be used in creating proficiency tests. This reminds me of the "blind proficiency tests" suggestion from years ago. In theory, it's a good idea, but in reality the preparation (for consistency to distribute to the many test takers) and the "scoring" of this type of	If a suggestion like this is to be included in the document, then the authors need to define how these tests will be made, distributed and graded.

	proficiency test make it very difficult to conduct.	
3461-3462 (KT 4.6)	Stating that "improvements across the entire community are expected with an increased understanding of the causes of variability among laboratories and analysts" without giving guidance on how this is to be accomplished	
3474	Bracketing may be needed… In fact, laboratories already do this but rather than call it "bracketing of factor space," we record it in internal validation and laboratory procedures as limitations, policies, etc. This is a standards requirement! The authors may be unaware having not examined any of this data.	Work with forensic laboratories to understand the data, procedure, and policies in place. Then form opinions based on the facts and update this document to reflect the reality of such work.
3484-3486	Stating here that the determination of whether PGS systems are reliable depends on the coverage of factor space for that particular case sample of interest and coverage of the ground truth for assessing reliability. No mention of the fact that laboratories are doing this as part of their own internal validation. And also demonstrated by numerous admissibility hearings or trials where users (stakeholders and their own DNA experts) have also had a chance to evaluate and argue about the degree of reliability in a particular case based on a laboratory's internal validation data.	Suggest to add to the end of this paragraph that "Internal validation studies performed by laboratories allow the users of the case data (DNA analysts, stakeholders such as attorneys and hired DNA experts) to evaluate reliability in relation to their case samples in comparison to ground truth data.
3487	Takeaway 4.7 - Helpful to include these validation performance results in the case file and report… 1. Courts already have access to such information through the discovery process, 2. Law enforcement is largely not interested in such information but rather the bottom line and just a summary rather than the full reports currently provided,	This idea seems poorly considered and impractical and unhelpful. Furthermore, it does not directly tie into the proceeding text. This should be reconsidered and the authors should interview those working in the field to vet the topic.

	3. Case files include case-pertinent information and all other information (validation, instrument and reagent quality control, proficiency testing, corrective action, etc.) is available separately for various reasons including efficiency, and 4. Not all validation data is relevant to a case. This is impractical, unnecessary, and would negatively impact crime lab efficiencies. Given that <10% of cases even go to court, this is particularly illogical.	
3487-3488 (KT 4.7)	"Including validation performance results in the case file and report" – it is obvious that the authors do not work in a forensic DNA laboratory and know who the "customer" is that receives these reports. To include this information in a DNA laboratory report is absurd.	If the "user" needs this information to understand the report, they can request it in discovery.
3487 (KT 4.7)	This mentions that "The degree of reliability of a PGS system when interpreting a DNA mixture can be judged based on validation studies using known samples that are similar in complexity to the sample in the case. To enable users of results to assess the degree of reliability in the case of interest, it would be helpful to include these validation performance results in the case file and report." There is no suggestion or example given as to what this would look like, and how easy this would even be for stakeholders to understand, by taking a complex topic and adding additional complex information to a DNA report. Part of the role of the expert witness is to be able to explain and put into context the results from the testing. Putting in more complex validation results into a report may make it even	An example should be provided as to what is meant by putting validation performance results within a report. Is a link to publicly available validation summaries sufficient? Discovery material available to the receiver of the report that consists of validation summaries or underlying data?

	more challenging for a reader of a DNA report to understand the results as they are. In addition, laboratories are already reviewing their internal validation data in order to ensure that their reported results accurately reflect the data that was tested within a particular case.	
Lines 3614-3616 and Box 4.1, page 94	Overall, the practical considerations for how a laboratory would publicly share validation information and implement Box 4.1 for independent review/assessments does not appear to have been discussed by the report authors. Where are labs posting the data? Who is reviewing the data for the independent assessment? How much time does a lab give to an independent reviewer? Is this review being performed before or after the new technology is implemented in casework? Etc.	Authors should provide a method or roadmap for this process to be implemented. Otherwise, Box 4.1 reads as a task-driven checklist of data to be provided by labs, with no follow-up steps.
3727-3732	This paragraph is confusing in relation to how published data and studies are helping to determine system reliability. Previous recommendations (like PCAST as referenced in this report) as well as various standards and guidelines nationally and internationally state that the underlying components should be published. But this statement seems to undermine their value. In addition, and as referenced in this report, there are many publications that cover actual system reliability and not just component-level reliability. Both are important, and both have been published, to an extent where the authors of this report had difficulty keeping up with the amount of publications. In the last statement that states 'there is a danger of inadvertently viewing results from narrowly-	There should be recognition here that there are a large number of system reliability published studies, including numerous publications and internal validation studies, as referenced in this report. In addition, if the last statement is kept in the report it should be clarified and supported as to who is at risk for the dangers mentioned in the last sentence of this paragraph.

	focused studies as applicable to system reliability' – is their evidence that this is how these publications are being interpreted by someone? It is important to study each component and for analysts to understand both PGS systems both at the component and system levels.	
3740-3743	"Regardless of sources of uncertainty and complexity of the samples, reliability of a PGS system boils down to checking its calibration accuracy and discriminating power at every conceivable scenario described by the factor space." – stating that the reliability needs to be evaluated at 'every conceivable scenario' seems to go against the concept of bracketing the desired factor space as recommended earlier within the report. This sets a much higher threshold for being able to determine reliability.	This should be clarified as to how this is in line with the bracketing approach mentioned earlier in the report.
3742	Every conceivable scenario described by the factor space	This seems to go against the bracketing approach. Every conceivable scenario is unrealistic and not necessary to cover proper usage for casework. Establish a proper foundation/framework with limitations to implement SOPs and then they are modified over time as needed.
3765-3772 (Chapter 5)	Points 2, 3, and 5 – While we have much in the way of published studies and internal validation studies to offer and address such questions, this document fails to point out the lack of control a criminalist actually has over this. While a criminalist will "do their best" to ensure the information is not misused, the court system doesn't support this level of involvement in many instances.	The consideration of the relevance of the court process throughout the document is underwhelming. This is a significant miss as the courts dominate the role and impact of the practitioner. So while the points made are those which one strives to maintain, the adversarial counterbalance pulling at and restricting the practitioner is unaddressed.
3785; Chapter 5	Inadequate to consider only a single trace in isolation	Who is making this determination? It's truly case dependent on what is tested and there are so many factors that may affect multiple traces being detected so then single is just inadequate?
3802; Chapter 5	Relevance of a DNA sample to the crime is often difficult to discern	Asking forensically relevant questions when assessing the evidence to exam is key and done by labs. DNA evidence is a piece of the puzzle and not the puzzle. I completely disagree with saying often difficult to discern. The question it may answer is known before choosing to test the item so it's really the weight of it that is part of our judicial system to determine.

Chapter 5	The main focus of this review is on the application of LRs to DNA mixtures using PG software.	Chapter 5 shifts the focus of this review to DNA transfer and persistence. The topics discussed in Chapter 5 are relevant to the interpretation of all DNA data (single source and mixtures) and are not specific to DNA mixture interpretation. The inclusion of DNA transfer/persistence in this review conflates the issues of reliability and relevance as they relate to DNA mixture interpretation. The concept of DNA transfer/persistence is important, and the community would benefit from a "stand-alone" review of this subject matter.
4546 (Chapter 5)	Takeaway 5.2 - It is especially important to consider relevance… This is the job of the court system (the lawyers). The criminalists inform the lawyers and do assist the court with whatever scientific knowledge is available to shed light. While practitioners attempt to control /stop the transition from who to how/when (4639), court testimony is a restrictive framework limiting the expert. Furthermore, this seems to be in direct conflict with recommendations from Dr. Itel Dror on Cognitive Bias (https://www.ucl.ac.uk/~ucjtidr/).	The consideration of the relevance of the court process throughout the document is underwhelming. This is a significant miss as the courts dominate the role and impact of the practitioner. So while the points made are those which one strives to maintain, the adversarial counterbalance pulling at and restricting the practitioner is unaddressed.
4646 (Chapter 5)	Takeaway 5.4 - Without context… Is vague. Do you mean without propositions? Agreed. Or something else like case context? Disagree. Not directly a criminalist's role. The crim will inform the courts on scientific questions but the attorneys provide case context.	Clarify.
4647; Chapter 5	LR as a standalone number without context can be misleading	Statistics have been reported and never implied info on how DNA was transferred or if it's relevant to the case so why is this an issue with LRs?
Chapter 5	Although there is a lot of interesting information in this chapter, I am not sure it should be in a foundational scientific review on mixture interpretation.	
4468 (KT 5.4)	This takeaway is not limited to LRs. RMP and CPI do not provide information about how or when DNA was transferred.	If this takeaway is included, it should be more transparent to clarify that it applies to all DNA statistical results.
4682	Example is sub-source level proposition	"An example of a sub-source level proposition might be that the DNA mixture contains DNA from the POI and the victim."
4777-4784, and	The challenges of efficiency and throughput for DNA laboratories	More emphasis or examples could be put into this report in relation to how a lab could address this within reports, particularly if time constrained by

referencing 34770-4773 and KT 5.4 (4646)	is mentioned here, as well as that they are not always aware of case context. Per the PCAST report and NAS report, independence of laboratories from law enforcement (which is why many don't know case context) was emphasized as something that would be beneficial for crime laboratories in general. Is this going against that recommendation? Then the following paragraph is discussing that if the labs don't take the time to put the information into reports or evaluate relevance, implying that they are now going to bias one side or the other by doing this. But bias is what labs are trying to avoid by not diving too deep into the arguments made by one side or another within court, and by being an independent laboratory. In the previous paragraph (4770-4773), it is mentioned that there are arguments for and against assigning probabilities to activity-level propositions, but it is not mentioned how lengthy and complex a process assigning these probabilities is, and there is no mention of how challenging those assignments (ex. Using Bayesian networks) could be to explain in court to a stakeholder.	caseloads, backlogs, and local laws governing turnaround times for cases. As well as discussing the value of laboratories remaining independent of law enforcement and prosecutorial or defense entities, and how that lack of influence can hope to unbias scientific reporting. Assigning probabilities to activity-level propositions is highly subjective, as mentioned here, but it is also very time-consuming and typically does not result in exceptionally strong evidence in either direction for the propositions. The value of that time could be discussed here, and whether more general statement examples of how activity-level propositions could be put into reports to separate them from the sub-source level statements that are being reported could be given.
4961; Chapter 6	This chapter attempts to provide new technologies to assist with DNA mixture interpretation; however, presentation of cell separation techniques (such as laser-capture microdissection) as a method appears out-of-touch with casework demands in crime labs. The described cell separation techniques are laborious and unrealistic for crime labs that process hundreds, if not thousands, of samples a year.	Remove Chapter 6.

	In addition, the application of NGS to mixtures needs further assessment. NGS lacks available tools such as prob gen for interpreting mixtures. Lines 5173-5174 and 5240 of the draft report mention the issue of distinguishing low-level contributors from noise sequences. Separately, there is no consistent guidance on applying statistics to additional markers like XSTRs and SNPs, that may be part of a NGS panel. These need to be addressed before this technology can be adopted as an improvement to DNA mixtures or a solution to existing problems in the forensic community.	
4961	General comment on Chapter 6 - aligns nicely with the DEI document yet to be published, esp Takeaway 6.2.	
5213	Do you mean developmental, instead of internal?	"performed in a developmental validation"
5343	Regarding 3rd point, solving a problem is not always why changes are made. Often it's simply worthwhile improvements (e.g., for efficiency).	Incorporate.
6698	Retain results for exam by third parties	This is from 2020 and nowhere does it state provide your data in a public forum or it's not able to be assessed for reliability. Data is now available to third parties such as court requests to have outside experts come onsite and assess it.
7090-7091; 7207-7208	Appendix 2 - Virtual courses could be offered by the NIJ Forensic Technology Center of Excellence… Virtual courses are offered and should be cited here. For example, the 8 part lecture series on probabilistic genotyping. This comes up later (7207-7208) but should be mentioned here.	Appropriately update.
7146-7148; 7137-7138	Appendix 2 - Seems worth mentioning here that CA, which contains numerous crime labs and ~10% of the US population has a librarian on staff with the CA	Consider adding something that other states could model or leverage such services.

	Criminalistics Institute (https://oag.ca.gov/cci). CCI subscribes to numerous forensic journals and has an electronic library catalogue system. CA criminalists not within CA DOJ simply send their requests directly to the librarian who responds with literature pdfs. Furthermore, requests can be made of any journal, not just those to which CCI subscribes.	
7211-7212	A certificate of attendance by itself is not sufficient for demonstrating that training or continuing education materials have been understood. This point is understood though this is common practice in numerous other fields. Perhaps a missed opportunity here is to consider the offerings of the ABC.	Incorporate the offerings of the ABC and make viable suggestions, such as a recommendation to states to offer incentives or even mandates (Texas as an example) on passing their certification test, and building out ABC to include CE exams.
7256-7259	"forensic community" and "advisory group"	Who is this "advisory group" – the make up of this proposed group needs to be defined if it is going to be a consideration. And it is imperative that an advisory group have individuals who are practitioners, not just theorists
7270-7283	What is this actually saying?	
7271-7273	"Consensus decisions from an advisory group"	Who is this "advisory group" – this needs to be defined if it is going to be a consideration. (see above)
7285-7287	This is already covered by the QAS in that the TLs must approve the analysts training	
7291-7292	What is this statement based on? How many internal validation studies were looked at to determine that TLs don't have sufficient training/experience to design validation experiments.	

PC56

NIST Internal Report 8351-DRAFT Comments

Kennedy, Jarrah R [redacted]

Thu 11/18/2021 10:08 AM

To: ScientificFoundationReviews <ScientificFoundationReviews@nist.gov>

The only recommendations I found in the document were in table 4.9. that stemmed from this review. In the appendices there are references to previous recommendations on Mixture Interpretation which are very clear in support of the LR (2006) and then transitioning to PG and models that incorporate drop-out/in, peak heights, and stutter and this is summarized in **Key Takeaways #A1.1 and #A1.2.**

Is there a reason why there is not a clear recommendation to the community to transition away from binary and towards PG?

On that topic ….

Key Takeaway #2.5
I am wanting for another portion to this takeaway: and laboratories should transition to these mixture interpretation approaches ?

The review is pretty clear in section 2.4.3 that contains this key takeaway that PG has advantages over binary. I am not sure how to highlight this more as it did seem to get lost in the (appropriate) critique regarding more transparency for the disclosure of validation data and other published data. I think that we can, in the same document, really highlight that PG is a much better approach than binary and that it should be implemented as soon as practical while also calling for continual improvement regarding the sharing of data for a widespread (independent) evaluation of the reliability.

In closing, I would say that the publicly available reliability data for previous binary approaches pales in comparison to what we have for PG.

Key Takeaway #4.8 – Yes- absolutely agree.

Chapter 5: Context and Relevance Related to DNA Mixture Interpretation
First, I want to acknowledge the importance of writing about this topic in this foundational review. As the US transitions to the use of PG and LRs – we must not be satisfied with only an improved LR framework for assigning sub-source LRs. With continued education and training – we can elevate the evaluations of the findings to higher levels of the hierarchy that are more aligned with the questions the court has.

My brain is stuck on the word "relevance." I'd love to see this chapter called what it is:
The Evaluation of Findings Given Activities: To Consider Transfer, Persistence, Prevalence, and Background

This is an advanced method of DNA interpretation that requires a higher level of expertise, training, and competency and the much more information than is typically needed at the sub-source level. These advanced methods evaluate the findings, whether biological or DNA – or lack thereof – given propositions which consider *specific* activities. This chapter is really all about the components necessary to evaluate the findings given activity level propositions.

Key Takeaway #5.4: The idea of relevance is a good one- but this word reads like a decision. It is not our job to say whether an item of evidence is relevant, but rather to provide an evaluation of the results (a strength of the evidence) given questions of activities which would help the factfinder decide if the DNA results are relevant!
Last sentence- considering inserting "sub-source" again.

5.4.2.6: Properly formulated propositions that address activities will be required to be much more specific than just “direct” vs. “indirect” transfer. Further – the word “transfer” should not appear in the proposition as this is a component to be evaluated (or have an assigned probability) given the more specific questions regarding how the DNA got on the item.

5.4.2.7 – Lines 4781-84: I would argue also that this is also not transparent (these opinions are not delineated in a report with supportable data). They are also dangerous regarding opining on the propositions themselves (transposing the conditional) and explaining results (ISFG recommendations) should be avoided as these are imbalanced and findings-led statements.

5.5. Summary:
I would like to reiterate that the DNA analyst cannot determine relevance – but what we can do is assess the results given the disputed facts (activities) – which can help the jury/factfinder make a decision about the “relevance” of the DNA.

I would also argue that the sub-source LR considers more than the rarity of the profiles though from a statistical standpoint I understand what is being stated. I’d like to see a re-emphasis on the importance of case information and proposition setting which leads to the assigned LR (it is conditional!). It is correct that a sub-source LR only addresses the origins of the DNA and not the activities associated with its deposition – but the continued use of relevance here may not be the best fit.

I appreciate the last sentence – DNA findings are but one piece of the puzzle and can only contribute to the decisions that need to be made by those who have all of the information.

Jarrah Kennedy
Senior Criminalist – Biology/DNA Section
Kansas City Police Crime Laboratory
2645 Brooklyn Ave.
Kansas City, MO 64127
[redacted]

PC57

Erie County CPS Forensic Lab Comments

Grill, Thomas [redacted]

Thu 11/18/2021 1:27 PM

To: ScientificFoundationReviews <ScientificFoundationReviews@nist.gov>
Cc: Schmitz, Michelli [redacted]

Please see the attached letter for comments from the Erie County CPS Forensic Lab.

Thank You,

--

Thomas Grill | Forensic Biologist IV
Erie County | Central Police Services
45 Elm St., | Buffalo, NY 14203
[redacted]

PC57a

County of Erie

MARK POLONCARZ
COUNTY EXECUTIVE

DEPARTMENT OF CENTRAL POLICE SERVICES

JAMES JANCEWICZ
COMMISSIONER

November 18, 2021

Dr. John M. Butler
National Institute of Standards and Technology
100 Bureau Drive, Mail Stop 4701
Gaithersburg, MD 20899-4701

Dear Dr. Butler,

This letter submitted by the Erie County Central Police Services (CPS) Forensic Laboratory is in response to the NISTIR 8351-Draft entitled, "DNA Mixture Interpretation: A NIST Scientific Foundation Review." We are aware that the scope of the NIST document is much more encompassing than a review of one specific probabilistic genotyping system; however, the Erie County CPS Forensic Laboratory's comments below are specific to the processes and software employed by our laboratory.

As you are aware, before a new method or technology can be implemented by an accredited laboratory, a thorough internal validation must be completed which includes the examination of a variety of sample types. There are published recommendations, guidelines, and standards on how this should be accomplished. New York State is unique in that it is one of the few states that has an additional level of oversight where forensic laboratories are also monitored by the New York State Commission on Forensic Science and the DNA Subcommittee. These two entities are responsible for evaluating and approving new methods and technologies before they are used on casework samples within a laboratory in New York State.

The Erie County CPS Laboratory has been using the probabilistic genotyping software STRmix™ since July of 2015. Our original internal validation was performed using the Identifiler™ Plus DNA typing kit. As noted above, before the laboratory could begin using the software, our internal validation study was required to go before the DNA Subcommittee for review. The DNA Subcommittee found our internal validation to be satisfactory and therefore approved us for the use of STRmix™ in casework. They in turn provided their recommendation to the New York State Commission on Forensic Science who subsequently accepted the use of the new technology. Erie County's Central Police Services Identifiler™ Plus internal validation study has also been reviewed during FBI QAS and ASCLAD LAB audits and has been approved by the same. In addition, in March of 2016, a Frye Hearing was held regarding the use of STRmix™ in our laboratory. The hearing took place in Niagara County Court in the case of The

People of the State of New York vs Vincent Bullard-Daniel. The court ruled in favor of the admissibility of STRmix™.

Since 2015, the Erie County CPS Forensic Laboratory has performed additional internal validations on the STRmix™ software using the Minifiler™ and PowerPlex® Fusion 5C typing kits. Finally, our laboratory has implemented numerous versions of the STRmix software as they have become available. Internal performance checks were completed with each upgrade to a newer version. All additional validations and performance checks have undergone reviews during FBI QAS, ASCLAB LAB, and ANAB audits as per accreditation requirements.

The Erie County CPS Forensic Laboratory alone has analyzed an abundance of samples during our three validations and numerous performance checks of STRmix™. The resulting data has been reviewed by professionals in the field of forensic DNA analysis, as well as the Niagara County Courts, all of which approved the use of the software in our laboratory.

Sincerely,

Thomas Grill
DNA Technical Leader
Erie County CPS Forensic Laboratory

PC58

NIST Internal Report 8351-"DNA Mixture Interpretation: A NIST Scientific Foundation Review"

Dennis McNevin [redacted]

Thu 11/18/2021 7:18 PM

To: ScientificFoundationReviews <ScientificFoundationReviews@nist.gov>

To whom it may concern

I would like to submit a response to NIST Internal Report 8351 - "DNA Mixture Interpretation: A NIST Scientific Foundation Review".

Key takeaway #4.5 calls for proficiency tests appropriate for complex DNA mixtures in order to assess the reliability of probabilistic genotyping (PG). Key takeaway #4.6 outlines the causes of variability and uncertainty in DNA mixture interpretation.

I would like to alert you to a recently published paper that address these key takeaways:
McNevin et al (2021) Proposed framework for comparison of continuous probabilistic genotyping systems amongst different laboratories, *Forensic Sciences*, 1(1):33-45. https://doi.org/10.3390/forensicsci1010006

This paper proposes a framework for an independent interlaboratory comparison of continuous PG systems applied to the interpretation of a specific class of DNA mixtures supplied to each laboratory, regardless of which STR multiplex assay is employed and regardless of which instruments are used to generate electropherograms. To use the language of the NIST review, we have restricted some factor spaces (specified mixture proportion, specified propositions, specified loci, specified population allele frequencies, specified co-ancestry coefficient) in order to make valid a comparison independent of other factor spaces (STR typing kit, PCR cycles, analytical threshold, PGS model decisions, PGS software, etc).

I hope it may contribute to the aims of the NIST review.

Sincerely

Professor Dennis McNevin
Forensic Genetics

Centre for Forensic Science
School of Mathematical & Physical Sciences
Faculty of Science
University of Technology Sydney
[redacted]
PO Box 123 Broadway NSW 2007 Australia
uts.edu.au

I acknowledge the Gadigal People of the Eora Nation and the Boorooberongal People of the Darug Nation upon whose ancestral lands the campuses of UTS now stand

forensic sciences

MDPI

Communication

Proposed Framework for Comparison of Continuous Probabilistic Genotyping Systems amongst Different Laboratories

Dennis McNevin [1,*], Kirsty Wright [2], Mark Barash [1,3], Sara Gomes [4], Allan Jamieson [4] and Janet Chaseling [5]

1 Centre for Forensic Science, School of Mathematical & Physical Sciences, Faculty of Science, University of Technology Sydney, Ultimo, NSW 2007, Australia; mark.barash@sjsu.edu
2 Centre for Genomics and Personalised Health, Genomics Research Centre, School of Biomedical Sciences, Institute of Health and Biomedical Innovation, Queensland University of Technology, Brisbane City, QLD 4000, Australia; k28.wright@qut.edu.au
3 Department of Justice Studies, San Jose State University, San Jose, CA 95192, USA
4 The Forensic Institute, Glasgow G1 2LW, UK; sarag@theforensicinstitute.com (S.G.); allanj@theforensicinstitute.com (A.J.)
5 School of Environment and Science, Griffith University, Nathan, QLD 4111, Australia; j.chaseling@griffith.edu.au
* Correspondence: dennis.mcnevin@uts.edu.au; Tel.: +61-2-9514-3902

Academic Editor: Manfred Kayser
Received: 31 March 2021; Accepted: 3 June 2021; Published: 10 June 2021

Abstract: Continuous probabilistic genotyping (PG) systems are becoming the default method for calculating likelihood ratios (LRs) for competing propositions about DNA mixtures. Calculation of the LR relies on numerical methods and simultaneous probabilistic simulations of multiple variables rather than on analytical solutions alone. Some also require modelling of individual laboratory processes that give rise to electropherogram artefacts and peak height variance. For these reasons, it has been argued that any LR produced by continuous PG is unique and cannot be compared with another. We challenge this assumption and demonstrate that there are a set of conditions defining specific DNA mixtures which can produce an aspirational LR and thereby provide a measure of reproducibility for DNA profiling systems incorporating PG. Such DNA mixtures could serve as the basis for inter-laboratory comparisons, even when different STR amplification kits are employed. We propose a procedure for an inter-laboratory comparison consistent with these conditions.

Keywords: forensic DNA analysis; probabilistic genotyping; likelihood ratio; DNA mixture; inter-laboratory comparison; reproducibility

1. Introduction

As forensic short tandem repeat (STR) genotyping assays have become more sensitive, DNA samples that may once have been classified as single source (assessed as being derived from a single DNA donor) may instead be classified as having multiple contributors as low-level alleles are now detected. The presence of multiple contributors has significant implications for propositions involving DNA transfer, persistence, prevalence and recovery (TPPR) [1]. Estimating the weight of evidence of these mixtures with the combined probability of inclusion/exclusion (CPI/E) has proved limiting, mostly because of problems with the treatment of allele drop in and drop out [2–4]. As a result, in many jurisdictions, probabilistic genotyping (PG) has become the default process for generating likelihood ratios (LRs) for forensic analysis of DNA mixtures [5]. Continuous PG algorithms model the probability distributions of observed peak heights in STR electropherograms (epgs) under

different scenarios. These can then be used to generate likelihoods for propositions which can in turn be combined into LRs. There are a number of continuous PG algorithms available including DNA·VIEW® [6], TrueAllele® Casework [7,8] (Cybergenetics, Pittsburgh, PA, USA), STRmix™ [9] (Institute of Environmental Science and Research, Forensic Science South Australia, Adelaide, SA, Australia), EuroForMix [10] and DNAxs [11]. The latter two PG systems are an extended version of the model proposed by Cowell et al. [12] which is open source while the other three require commercial licences [13–15]. Until relatively recently [16], there has been little evidence that continuous PG is reproducible amongst different laboratories, and little attempt has been made to define credible intervals for the LRs produced.

Swaminathan et al. [17] collated the LRs for 30 × one-person samples, 82 × two-person mixtures and 90 × three-person mixtures generated by four variations of their CEESIt continuous PG algorithm [18]. The four variations included different permutations of models for "mixture ratio" (also known as the "mixture proportion" [19] and as a "mass parameter" [9,20]), peak height distribution and forward stutter designed to mimic the diversity of available continuous PG algorithms. LRs were calculated five times for each mixture to assess intra-model variance resulting from the Markov chain Monte Carlo (MCMC) simulation procedure. In all four models, intra-model variability increased with an increase in the number of contributors and with a decrease in the contributors' template mass. The LRs were binned into ranges corresponding with verbal expressions for the weight of evidence according to the Association of Forensic Science Providers [21] ranging from "weak" for an LR between 1 and 10 to "extremely strong" for an LR > 10^6. For 9% of intra-model comparisons, LRs did not fall in the same bin for the same mixture, and for 1.5%, LRs were more than one bin apart. For 16% of inter-model comparisons (where two or more of the four models yielded LRs in the same bin for all five runs), LRs from one model fell in a different bin from one or more other models, and 11% were more than one bin apart.

Bright et al. [22] originally proposed and demonstrated a series of tests for validating PG systems using single source, simulated major/minor (3:1) mixtures and simulated balanced (1:1) mixtures. The LRs generated by PG were compared with those expected under theoretical modelling in Excel. Input electropherograms had peak heights adjusted so that there was:

- No possibility of drop in and drop out;
- No possibility of drop in but some possibility of drop out;
- No possibility of drop out but with artificial alleles added to mimic the possibility of drop in.

Replicate analyses were employed to test for reproducibility. The results of their tests showed good agreement between expected results, continuous PG and semicontinuous PG for single source and balanced profiles, although for the latter, continuous PG yielded higher LRs than semicontinuous PG, as expected. This is because of the extra peak height information considered by continuous PG. For major/minor profiles, agreement between continuous and semicontinuous PG only occurred when the major contributor was manually extracted. This is because continuous PG is able to take advantage of the peak height information in an unbalanced mixture while semicontinuous PG does not (putting aside manual interpretation by an analyst of stutter peaks, for example). All electropherograms were simulated from single source profiles derived from the same capillary electrophoresis instrument, and only one continuous PG algorithm (STRmix) was employed.

There have been other attempts to compare the reproducibility of outputs amongst different PG systems, but most of these (e.g., [23–27]) have involved submitting the same epgs from the same STR amplification kits to different PG algorithms. Benschop et al. [28] describe a validation of one PG system (DNAxs) in five laboratories using STR genotype data (alleles and peak heights) generated within each laboratory from different STR assays. Each laboratory shared its genotyping results with the others, and LRs were mostly within an order of magnitude for the same genotype data. However, the same DNA samples were not processed in each individual laboratory so that the LRs were all generated from the same epgs. Alladio et al. [29] showed that it was possible to compare the reproducibility of

LRs from different PG systems and different STR assays. The LRs generated from DNA·VIEW, STRmix and EuroForMix were reproducible for high DNA template amounts over a wide range of mixtures with different numbers of contributors and mixture ratios. Once again, the LRs were all generated from the same epgs. Different STR assays produced LRs that differed by many orders of magnitude, as expected. This is because different STR assays employ different STR loci and different numbers of loci. While this might seem like an impediment to inter-laboratory comparisons, we demonstrate that it can be overcome.

Inter-laboratory comparisons are a standard feature of forensic DNA analysis methods [16,26,30–32]. They indicate the reproducibility of a particular method amongst different laboratories and the variance of quantitative results. It is a reasonable expectation that they be undertaken. They serve to calibrate amongst laboratories, which helps to ensure equality of justice outcomes amongst jurisdictions. The US President's Council of Advisors on Science and Technology (PCAST) "believes that test-blind proficiency testing of forensic examiners should be vigorously pursued, with the expectation that it should be in wide use, at least in large laboratories" [33]. The US National Institute of Standards and Technology (NIST) states: "Inter-laboratory tests are the means by which multiple laboratories compare results and demonstrate that the methods used in one's own laboratory are reproducible in another laboratory. These tests are essential to demonstrate consistency in results from multiple laboratories" (quoted from [34]).

McNevin et al. [35] have previously suggested a method for assessing reproducibility and defining credible intervals for LRs derived from the same DNA extracts (not electropherograms) and calculated by STRmix in particular and continuous PG in general. This was met with some scepticism by Buckleton et al. [36] who contend, firstly, that there are "multiple reasonable answers in the case of evidence from one extract" [36,37] and, secondly, that it is sufficient to calibrate the LRs generated by PG from multiple laboratories using the method of Ramos and Gonzalez-Rodriguez [38]. In summary, this last method uses the LRs and a prior odds ratio from known numbers of contributors and non-contributors submitted by multiple laboratories to calculate a posterior odds ratio. The posterior ratio is compared with the relative frequencies of contributors. The number of non-contributors with LR above a certain threshold should reflect the number expected given the numbers of contributors and non-contributors [39]. This is a reasonable test of the bulk or macro properties of the LR from multiple laboratories; however, it does not provide any indication of the variance in LRs amongst laboratories for the same sample or whether an individual laboratory is producing reasonable LRs. For example, in a multi-laboratory comparison, a laboratory that consistently produces large LRs might be balanced by a laboratory that consistently produces small LRs without perturbing the bulk or macro properties of all the LRs produced. It also requires a large number of contributors and non-contributors for many mixtures.

We argue that there is a true test of each laboratory's ability to produce reasonable LRs, consistent with McNevin et al. [35] and regardless of the instrumentation and STR assays used to produce epgs. Here we provide a formal proof that such a test exists, and we define the conditions under which such a test could be performed.

2. The Likelihood Ratio Produced by Probabilistic Genotyping

We start with the general formulation of the LR for a DNA mixture as a ratio of two conditional probabilities:

$$\mathrm{LR} = \frac{\mathrm{P}(E|H_1)}{\mathrm{P}(E|H_2)} \tag{1}$$

We will loosely follow the notation of Taylor, Bright and Buckleton [9,20] in their descriptions of PG systems while acknowledging that other notations exist (e.g., [12,19]). The evidence, E, is an electropherogram (epg) from a crime trace (G_C) exhibiting a mixture of known reference profiles (G_R) and unknown profiles (G_U). There is also a person or persons of interest (POI or POIs). In general, one proposition, H_1, is that a particular reference genotype (or genotypes) from a POI or POIs (G_P)

is a contributor to the DNA mixture, while the alternate proposition, H_2, is that the contributors are two or more known (G_R) or unknown (G_U) genotypes not including the POI(s). The propositions can take various forms, but H_2 will always differ from H_1 in that the genotype of at least one POI (G_P) is replaced with an unknown genotype (G_U), for example:

$$H_1 = \{G_P, G_{R1}, G_{R2}, G_{R3}, \ldots, G_{U1}, G_{U2}, G_{U3}, \ldots\}$$

$$H_2 = \{G_{R1}, G_{R2}, G_{R3}, \ldots, G_{U0}, G_{U1}, G_{U2}, G_{U3}, \ldots\}$$

Cowell et al. [12] show that, under the assumption of Hardy–Weinberg equilibrium (HWE), the LR for a mixture for which G_P in H_1 is replaced by an unknown profile (G_{U0}) in H_2 can never be greater than the LR for a single source profile for the POI responsible for G_P. This places an upper limit on the LR under these circumstances.

The epg reveals M genotype sets S of possible explanatory genotype combinations from N contributors that could give rise to the DNA mixture at any locus. The likelihood ratio becomes:

$$\mathrm{LR} = \frac{\sum_{m=1}^{M} p(G_C|S_m)\mathrm{P}(S_m|H_1)}{\sum_{m=1}^{M} p(G_C|S_m)\mathrm{P}(S_m|H_2)} \quad (2)$$

where S_m is the mth possible explanatory genotype combination for N contributors and $p(G_C|S_m)$ is a conditional probability density (distinguishing it from a point probability, P). As an example, consider an epg at a locus where there are four alleles (A, B, C, D) detected above an analytical threshold. Possible genotype sets for two presumed contributors include {AB, CD}, {AC, BD}, {AD, BC}. There may also be genotype sets that do not include all detected alleles, for example, {BC, BD}, {BC, CD}, {BD, CD}, {BB, CD}, {BD, CC}, {BC, DD}, with A as an artefact (e.g., drop in or stutter). For three presumed contributors, possible genotype sets include {AA, BB, CD}, {AA, BC, DD}, {AB, CC, DD}, etc. There may also be genotype sets that include undetected alleles, for example {AB, CD, AE}, {AB, CD, BE}, {AB, CD, CE}, {AB, CD, DE}, etc, with E as a drop out. A "weight", w_m, can be used to describe the conditional probability density for observing the mixture profile given S_m:

$$w_m = p(G_C|S_m) \quad (3)$$

where:

$$\sum_{m=1}^{M} w_m = 1 \quad (4)$$

The normalised weights vary from 0 to 1 and account for the possibilities of allele drop in and allele drop out. For continuous PG, they also account for the possibilities of stutter, peak height stochasticity, peak height degradation and peak height variations as a result of allele overlap (shared alleles). Semicontinuous PG does not consider peak height information, although stutter must be differentiated from true alleles by the analyst. The weights for continuous PG are modelled using what Taylor et al. [9,20] refer to as "mass parameters" including a template DNA amount for each contributor, a degradation level for each contributor, an assay-specific locus amplification efficiency for each locus and a replicate amplification efficiency for each replicate. The last two parameters account for inter-locus and inter-replicate variabilities, respectively. The likelihood ratio then becomes:

$$\mathrm{LR} = \frac{\sum_{m=1}^{M} w_m \mathrm{P}(S_m|H_1)}{\sum_{m=1}^{M} w_m \mathrm{P}(S_m|H_2)} \quad (5)$$

3. A Reproducible Subset of Likelihood Ratios from Probabilistic Genotyping

The values for w_m will vary from laboratory to laboratory. This is because each laboratory must model epg artefacts and peak height variance for the particular conditions in their laboratory, and these

models inform the various w. At first glance, and this is certainly the view of Buckleton et al. [36], this suggests that LRs reported by different laboratories cannot be compared. While it is true that not all LRs can be compared, we can define specific conditions for which a subset of LRs can be compared. These conditions exist when the values of w_m are the same for different laboratories.

The weight or likelihood, w_m, for any genotype set, S_m, will vary from almost impossibility ($w_m \to 0$) to almost certainty ($w_m \to 1$). We distinguish between genotype sets with at least one allele not belonging to any of the contributors or without all contributor alleles present (S_i) and those with all alleles belonging to contributors and no others (S_j):

$$\mathrm{LR} = \frac{\sum_i w_i \mathrm{P}(S_i|H_1) + \sum_j w_j \mathrm{P}\left(S_j\middle|H_1\right)}{\sum_i w_i \mathrm{P}(S_i|H_2) + \sum_j w_j \mathrm{P}\left(S_j\middle|H_2\right)} \tag{6}$$

For our four-allele example, let us assume that the contributors have genotypes BC and CD (A is an artefact). Genotype sets S_i include any genotypes with allele A (AA, AB, AC, AD) or without at least one of B, C and D, while genotype sets S_j include all of B, C and D but not A (or any undetected alleles). We wish to restrict w_i so that each laboratory finds $w_i \to 0$. Under these conditions, for any PG system:

$$\lim_{w_i \to 0} \mathrm{LR} = \frac{\sum_j w_j \mathrm{P}\left(S_j\middle|H_1\right)}{\sum_j w_j \mathrm{P}\left(S_j\middle|H_2\right)} \tag{7}$$

We then extract the unique genotype set S^* that corresponds with the contributors to the mixture:

$$\lim_{w_i \to 0} \mathrm{LR} = \frac{\sum_k w_k \mathrm{P}(S_k|H_1) + w^* \mathrm{P}(S^*|H_1)}{\sum_k w_k \mathrm{P}(S_k|H_2) + w^* \mathrm{P}(S^*|H_2)} \tag{8}$$

where w^* is the weight assigned to S^* and the new subscript k is used because S^* has been separated from the other S_j. For our four-allele example, S^* is {BC, CD} which is now distinguished from {BC, BD}, {BD, CD}, {BB, CD}, {BD, CC}, {BC, DD}, etc. When H_1 corresponds with the contributors only (H_1 true) then $\mathrm{P}(S_k|H_1) = 0$, $\mathrm{P}(S^*|H_1) = 1$ and:

$$\lim_{w_i \to 0} \mathrm{LR} = \frac{w^*}{\sum_k w_k \mathrm{P}(S_k|H_2) + w^* \mathrm{P}(S^*|H_2)} \leq \frac{1}{\mathrm{P}(S^*|H_2)} \tag{9}$$

Note that there is an upper limit for LR which occurs if all $w_k \to 0$. This is essentially the same result obtained by Cowell et al. [12] for continuous PG but generalised for multiple contributors. When H_1 corresponds with non-contributors (H_1 false) where at least one allele of a non-contributor is not shared with a true contributor then $\mathrm{P}(S_k|H_1) \to 0$, $\mathrm{P}(S^*|H_1) \to 0$ and $\mathrm{LR} \to 0$.

For semicontinuous PG, we have no way to reduce uncertainty amongst S_k and S^* (because peak height information is not considered). All remaining genotype sets are equally likely. Hence, $w^* = w_k = w$. In this case, Equation (8) becomes:

$$\lim_{w_i \to 0} \mathrm{LR} = \frac{w \sum_k \mathrm{P}(S_k|H_1) + w\mathrm{P}(S^*|H_1)}{w \sum_k \mathrm{P}(S_k|H_2) + w\mathrm{P}(S^*|H_2)} = \frac{\sum_k \mathrm{P}(S_k|H_1) + \mathrm{P}(S^*|H_1)}{\sum_k \mathrm{P}(S_k|H_2) + \mathrm{P}(S^*|H_2)} \tag{10}$$

When H_1 corresponds with the contributors only (H_1 true):

$$\lim_{w_i \to 0} \mathrm{LR} = \frac{1}{\sum_k \mathrm{P}(S_k|H_2) + \mathrm{P}(S^*|H_2)} \tag{11}$$

This is the minimum performance expected of continuous PG. We therefore have an upper and lower bound for the LR from continuous PG if:

a. $w_i \rightarrow 0$ (i.e., uncertainty is minimised between genotype sets with all alleles belonging to contributors and no others and those with at least one allele not belonging to contributors or without all contributor alleles present) and;
b. H_1 is true (i.e., H_1 corresponds with the contributors only).

The range of expected values is given by:

$$\frac{1}{\sum_k \mathrm{P}(S_k|H_2) + \mathrm{P}(S^*|H_2)} < \lim_{w_i \rightarrow 0} \mathrm{LR} < \frac{1}{\mathrm{P}(S^*|H_2)} \quad (12)$$

The lower bound is the LR for the same mixture derived from semicontinuous PG, and the upper, aspirational bound is the LR that would be possible if uncertainty could be eliminated amongst the true contributor genotype set, S^*, and all others. To move from the lower bound to the upper bound requires increasing w^* beyond the average weight used for semicontinuous PG. Indeed, this is the goal of continuous PG, and the relative ability to increase w^* over all other weights is a performance measure for continuous PG systems.

4. Conditions for Achieving Reproducible LRs from Probabilistic Genotyping

The conditional probabilities, $\mathrm{P}(S^*|H_1)$ and $\mathrm{P}(S^*|H_2)$ are match probabilities defined by true contributor reference profiles, population genetic models, population allele frequencies and two alternative propositions. As long as any two laboratories have reference profiles for the same contributors, consider the same propositions and use the same models (e.g., Hardy–Weinberg proportions, NRC II recommendation 4.1, NRC II recommendation 4.2), the same population allele frequencies and the same θ for the same loci, they should obtain the same values for $\mathrm{P}(S^*|H_1)$ and $\mathrm{P}(S^*|H_2)$. This defines our first conditions for an inter-laboratory comparison of LRs:

1. The same standard mixtures should be examined.
2. The same propositions should be considered.
3. The same loci should be employed.
4. The same population allele frequencies should be employed.
5. The same population genetic model and sub-structure correction, θ, should be employed (e.g., $\theta = 0$).

Satisfying Equation (7) requires reducing the probabilities of genotype sets with at least one allele not belonging to any of the contributors or those without all contributor alleles present such that $w_i \rightarrow 0$. This will occur when there is little uncertainty between:

- No allele and allele drop out;
- A (low peak height) contributor allele and allele drop in;
- A (low peak height) contributor allele and a stutter peak;
- A single allele and shared ("stacked") alleles, either of which may or may not include allele drop in and stutter peaks.

We consider each of these in turn.

The greater the amount of contributor DNA in the mixture, the higher contributor allele peaks are likely to be. The higher the allele peaks, the less likely that drop out will occur. Similarly, high allele peaks are unlikely to be confused with (low peak height) allele drop in. Stochastic variation in peak heights will also be minimised with increasing peak height. Heterozygote peak height imbalance has been shown to decrease as average peak height (APH) increases [40,41]. Continuous PG algorithms model allele peak height and stutter peak height to reflect observations that variance decreases with peak

height. EuroForMix and TrueAllele use gamma [10] and normal distributions [7], respectively. STRmix models allele peak and stutter peak height variation according to a log normal distribution [9,20,42]:

$$\log_{10}\left(\frac{O_a}{E_a}\right) \sim \mathrm{N}\left(0, \frac{c^2}{E_a}\right) \quad (13)$$

$$\log_{10}\left(\frac{O_{a-1}}{E_{a-1}}\right) \sim \mathrm{N}\left(0, \frac{k^2}{O_a}\right) \quad (14)$$

O and E refer to observed and expected peak heights for alleles (a) and stutter ($a-1$), and c^2 and k^2 are locus-specific random variables which are in turn modelled by gamma distributions. For both allele and stutter peaks, the variance is inversely related to peak height (E_a and O_a, respectively) such that stochastic variation will be reduced with increasing peak height.

Too much DNA, however, will result in overloading of the epg with split peaks, pull ups and other artefacts, after which true allele peaks can be confused with these artefacts. This provides our next condition for an inter-laboratory comparison of LRs:

6. The DNA template from true donors should be maximised to a point within the linear range and below saturation of the epg.

The optimal amount of DNA defined by condition 6 may be difficult to assess. One way to achieve it is to amplify a dilution series of DNA such that there is a range of DNA template amounts previously tuned to the optimum amongst them [29]. The LR will approach a maximum for H_1 true as DNA template amount increases and as $\omega_i \to 0$. This is demonstrated by Bauer et al. [44] in their Figure 1 (originally in [8]) and defines our next condition:

7. Each laboratory is presented with aliquots of the same dilution series of DNA solutions which then undergo analyses to produce epgs for each solution according to each laboratory's standard practice (according to which the PG system was validated in that laboratory).

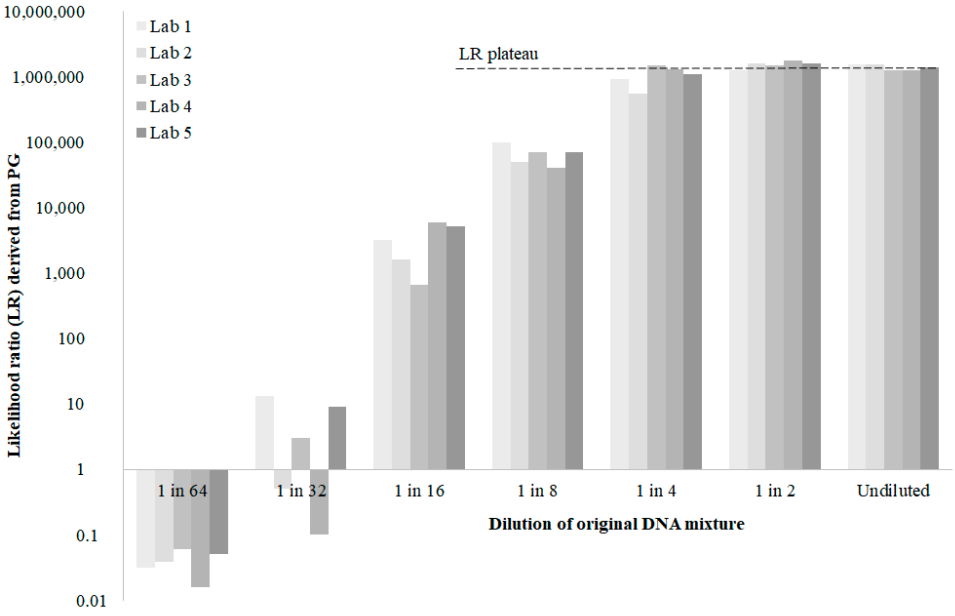


Figure 1. Idealised results of a hypothetical inter-laboratory trial demonstrating inter-laboratory reproducibility of probabilistic genotyping where each laboratory is provided with extracted DNA from a dilution series of a DNA mixture defined by conditions 1 to 8. Higher concentrations of DNA (**right**) will reduce ambiguity in epgs with less peak height stochastic variation, less drop out and less drop in.

Stutter artefact peak heights will scale with true allele peak heights approximately according to a stutter ratio. Hence, conditions 6 and 7 are insufficient on their own to reduce uncertainty between stutter peaks and smaller true allele peaks. Similarly, they will not reduce uncertainty between single alleles and stacked alleles. If the contributors to a DNA mixture are present in equal proportion, however, this uncertainty is minimised, and different labs and different PG systems will tend to find the same w_j. Cheng et al. [45] have recently demonstrated that peak heights are additive and proportional to the donor contributions in a DNA mixture epg. This means that if an allele is shared by two donors, then it should have double the height expected from an allele belonging to a single donor if both donors' DNA templates are not degraded and are present in equal proportion. If it is shared by three donors, it should have triple the height expected from an allele belonging to a single donor if all three donors' DNA is present in equal proportion, and so on.

Stutter peak heights are typically 15% or less of the parent allele peak height, depending on the length of the longest uninterrupted repeat chain [46]. Uncertainty between a stutter peak and a true allele will occur if one contributor is present in the mixture with this order of magnitude relative to another donor (15% or less). When all contributors to a mixture are present in equal proportion, then the size of each donor's allele peak should be approximately 100% relative to all other donors' peaks (albeit with stochastic variance and taking account of degradation) and thus less likely to be confused with a stutter peak. Our next condition for an inter-laboratory comparison of LRs is:

8. All known donors are present in equal proportion by DNA template amount.

We now have the eight conditions for an inter-laboratory comparison originally suggested by McNevin et al. [35]. Such a comparison should produce the results described by them in their Figure 1 and by Bauer et al. [44] in their Figure 1 where the maximised value of the LR corresponding with the plateau in both cases is given by Equation (7) for all propositions and Equation (9) for H_1 true (Figure 1). We would go so far as to say that Equation (12) defines the "expected" LR range under our eight conditions, in the same way that the reciprocal of the random match probability is the expected LR for a high quality single source profile. We acknowledge that there is debate here, including a special issue in *Science & Justice* devoted entirely to measuring (or not) the reproducibility of LRs [47], but the upper bound for the LR defined in Equation (9) is certainly aspirational.

We add two final conditions that should be employed for any inter-laboratory comparison consistent with best scientific practice. These are:

9. The trial should be blinded. Laboratories presented with a dilution series of DNA solutions to be analysed should not know which is which.
10. The trial should be facilitated by an entity not associated with the PG systems under comparison.

LR > 1 from semicontinuous PG will nearly always be less than the LR from continuous PG for the same mixture for H_1 true, except at low DNA template amounts when stochastic effects dominate. This is because more information (peak heights) is being used by continuous PG resulting in LRs further from 1. Exceptions may occur when a sample has an unlikely peak height that greatly deviates from the expected height, possibly due to extreme stochastic variation or a primer sequence polymorphism (null allele). This can lead to very low weights for S^* and thus a lower LR than for semicontinuous PG. Such exceptions notwithstanding, Equation (12) defines a theoretical range for the LR from continuous PG where the lower bound is the LR from semicontinuous PG and the upper bound represents no uncertainty amongst the true contributor genotype set, S^*, and all others. The greater the number of equal-proportion contributors, the lower the LR and the lower the theoretical range defined by Equation (12) will be. This is because there are greater numbers of allele permutations that could explain contributor genotype sets, S_j, and hence the weight, w_j, assigned to each one is lower.

We now address the questions of peak height imbalance and degradation (the typical "ski slope" of DNA profiles). STRmix (and, indirectly, other continuous PG systems) model these phenomena using the so-called mass parameters and then assign w_m according to how far the observed peaks

deviate from the modelled peaks. Allele decay is modelled as a function of molecular weight where longer alleles will have lower peak heights than shorter alleles. Different manufacturers of forensic STR assays will have different amplicon sizes for each of the loci and so the relative decay amongst loci will vary. At any particular locus, there will also be allele-specific variation leading to heterozygote imbalance, for example. For STRmix, this is modelled by Equations (13) and (14). If we consider two non-shared alleles in a genotype set, S_j, the further they are from equality (balanced), the lower the weight assigned to a heterozygous genotype in S_j, all other weights being equal.

Peak height variance and degradation have been posited by Buckleton et al. [36] as another reason LRs cannot be compared amongst laboratories. However, at any particular locus for any particular kit, we are restricting w_m such that each laboratory finds $w_i \to 0$ and w_j are the same for all laboratories. A true heterozygous genotype may have two unbalanced and unshared alleles, but the heterozygous genotype will still have a much higher probability, w_j, than other possible genotypes under our eight conditions, all other weights being equal.

5. An Inter-Laboratory Comparison

Our proposed conditions and trial will not provide a comparison point for every possible LR generated by continuous PG. This is because LRs produced by continuous PG are subject to variance. However, we have specified conditions that minimise this variance. Even less variance is possible if we specify conditions that minimise uncertainty between one POI and all other contributors (i.e., $w_i \to 0$, $w_k \to 0$, $w^* \to 1$), but this is the trivial case where one contributor is present at much higher proportion than all others, approaching the case of a single source profile.

It may be argued that our set of conditions 1 to 8 is restrictive and does not test the reproducibility of PG systems when w_i is not close to 0. However, by including a dilution series, we can see how the variance in LR increases from its minimum (at high average peak height, APH) as APH decreases. Swaminathan et al. [17] found that this variance increased with a decrease in the contributors' template mass for all four of their representative continuous PG model variations. Conditions 1 to 8 therefore provide for a minimum performance measure. Our condition 8 is a strenuous test because, as Bucketon et al. [43] point out: "testing two low-level contributors with similar APHs (a 1:1 mixture) presents more of a challenge to the software than does a 1:20 mixture, as the genotype of the higher contributor has less uncertainty and helps to inform the genotype of the lower contributor". This would equally be the case at high APHs.

An inter-laboratory comparison employing our conditions will provide the following information:

- The position of the plateaued, maximum LR from any laboratory within the theoretical range defined by Equation (12). This is a measure of performance, if not accuracy.
- The range of plateaued, maximum LRs reported by laboratories. This is an indication of the credible interval for LRs reported under the best possible conditions designed to minimise variance in LRs. This credible interval would suggest a minimum as we would expect the variance amongst laboratories to increase the further they are from conditions 1 to 8.
- Outlier laboratories. This would provide guidance on which laboratories (if any) might need to re-validate their PG system.
- Outlier PG systems. This would provide guidance on which PG systems (if any) do not model allele peak height variance adequately according to the procedures in a particular laboratory.
- The minimum template amounts at which fortuitous LRs are encountered for any laboratory (LR > 1 for a non-contributor, LR < 1 for a contributor). As DNA template amounts decrease in the dilution series, LRs for contributors and non-contributors will approach 1 but may actually overshoot.

We now define the procedure for an inter-laboratory comparison consistent with our conditions 1 to 10:

1. Identify participating laboratories. They are required not to communicate with each other concerning the trial.

2. Identify reported loci in common amongst participating laboratories. Longer loci, where Equation (7) might not be expected to hold, could also be excluded (with agreement). These excluded loci should not be used either to estimate parameters such as mixture proportions or to calculate LRs. In practice, any laboratory could nominate a locus to be excluded. A comparison between PG systems could, theoretically, be made with as little as one locus but, of course, more loci will increase the stringency of any trial.
3. Identify a trial facilitator not associated with any of the PG systems to be used. This could be a university, a centre of excellence or a national forensic regulator, for example.
4. The trial facilitator collects samples from reference cell lines or consenting volunteers and performs DNA extraction and quantitation for each sample.
5. The DNA concentration for each sample is normalised according to the quantitation results and assessed as being of a suitable (high) quantity and quality.
6. A single source STR profile for each donor is generated according to best practice. These are the contributor reference profiles. Non-contributor reference profiles can also be generated.
7. Equal volume and equal concentration aliquots of high abundance DNA are combined from various donors to create mixtures of 2, 3, 4, ... and N contributors in equal proportion by DNA amount.
8. For each mixture, a dilution series is created (e.g., undiluted, 1 in 2, 1 in 4, 1 in 8, etc.).
9. Aliquots of the various dilution series (one dilution series per mixture) are distributed to the participating laboratories, labelled randomly such that the laboratory does not know the concentration of DNA in any sample. For one, two, three, four and five contributors each at seven different dilutions, for example, a total of 35 samples would be supplied.
10. Each participating laboratory produces an STR epg for each aliquot according to the standard procedures for that laboratory.
11. The participating laboratories are also supplied with the following:

    - Reference profiles.
    - Allele frequencies from a defined population.

12. The following propositions are also provided to each of the participating laboratories:

    - H_1: The donor of reference profile X is a contributor to the mixture which also consists of N other known but unrelated contributors (where all N+1 reference profiles are supplied);
    - H_2: The donor of reference profile X is not a contributor to the mixture which consists of an unrelated, random member of the (defined) population and N other known but unrelated contributors.

    These can be applied to both contributor and non-contributor reference profiles.
13. Each laboratory is asked to provide a LR according to Equation (1). The laboratories are instructed to use the allele frequencies provided from the defined population without any population substructure corrections and using a consistent population genetic model (e.g., Hardy–Weinberg proportions).
14. The LRs are collated and compared by the trial facilitator.

6. Conclusions

We propose a procedure to allow comparison amongst PG systems and laboratories. The LR defined by Equation (7) and the LR range defined by Equation (12) and enabled by our conditions 1 to 8 will not depend on either the PG system or the laboratory if each PG system calculates LR according to Equation (5) and calculates w_m according to maximum likelihood and if each laboratory has calibrated their PG system appropriately. Kelly et al. [40] state that LR "variance is more profile specific than laboratory specific" if c^2 and k^2 in Equations (13) and (14), respectively, are adequately modelled.

Proposing that a PG system can be calibrated according to the procedures and instruments of a particular laboratory raises the question: Can that calibration be tested? We believe it can and that there is no reason to avoid inter-laboratory comparison of PG systems, even when different STR amplification kits are employed. Differences in STR assay, PCR thermal cycling, capillary electrophoresis or profile analysis settings will all be manifested in peak height variances which are modelled by PG systems. How well it is modelled will be determined by where any generated LR sits in the range defined by Equation (12) and, indeed, whether it falls in this range at all. We hope that our proposed study builds upon existing validation of continuous PG and provides another step towards establishing a standardised, best practice approach for DNA mixture analysis.

Author Contributions: Conceptualization, D.M., K.W. and J.C.; methodology, D.M., K.W., A.J. and J.C.; writing—original draft preparation, D.M.; writing—review and editing, K.W., M.B., S.G., A.J., J.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The authors are grateful for feedback on various drafts from a number of individuals.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. van Oorschot, R.A.H.; Szkuta, B.; Meakin, G.E.; Kokshoorn, B.; Goray, M. DNA transfer in forensic science: A review. *Forensic Sci. Int. Genet.* **2019**, *38*, 140–166. [CrossRef]
2. Perlin, M.W. Inclusion probability for DNA mixtures is a subjective one-sided match statistic unrelated to identification information. *J. Pathol. Inf.* **2015**, *6*, 59. [CrossRef]
3. Bieber, F.R.; Buckleton, J.S.; Budowle, B.; Butler, J.M.; Coble, M.D. Evaluation of forensic DNA mixture evidence: Protocol for evaluation, interpretation, and statistical calculations using the combined probability of inclusion. *BMC Genet.* **2016**, *17*, 125. [CrossRef] [PubMed]
4. Curran, J.M.; Buckleton, J. Inclusion probabilities and dropout. *J. Forensic Sci.* **2010**, *55*, 1171–1173. [CrossRef]
5. Coble, M.D.; Bright, J.-A. Probabilistic genotyping software: An overview. *Forensic Sci. Int. Genet.* **2019**, *38*, 219–224. [CrossRef]
6. Brenner, C.H. DNA·VIEW User's Manual. Charles Brenner, UC Berkeley, 6801 Thornhill Drive Oakland, California, USA. 2019. Available online: http://dna-view.com/downloads/documents/manuals/DNAVIEW%202019%20US.pdf (accessed on 8 June 2021).
7. Perlin, M.W.; Legler, M.M.; Spencer, C.E.; Smith, J.L.; Allan, W.P.; Belrose, J.L.; Duceman, B.W. Validating TrueAllele® DNA mixture interpretation. *J. Forensic Sci.* **2011**, *56*, 1430–1447. [CrossRef]
8. Perlin, M.W.; Sinelnikov, A. An information gap in DNA evidence interpretation. *PLoS ONE* **2009**, *4*, e8327. [CrossRef] [PubMed]
9. Taylor, D.; Bright, J.-A.; Buckleton, J. The interpretation of single source and mixed DNA profiles. *Forensic Sci. Int. Genet.* **2013**, *7*, 516–528. [CrossRef] [PubMed]
10. Bleka, Ø.; Storvik, G.; Gill, P. EuroForMix: An open source software based on a continuous model to evaluate STR DNA profiles from a mixture of contributors with artefacts. *Forensic Sci. Int. Genet.* **2016**, *21*, 35–44. [CrossRef] [PubMed]
11. Benschop, C.C.G.; Hoogenboom, J.; Hovers, P.; Slagter, M.; Kruise, D.; Parag, R.; Steensma, K.; Slooten, K.; Nagel, J.H.A.; Dieltjes, P.; et al. DNAxs/DNAStatistX: Development and validation of a software suite for the data management and probabilistic interpretation of DNA profiles. *Forensic Sci. Int. Genet.* **2019**, *42*, 81–89. [CrossRef]
12. Cowell, R.G.; Graversen, T.; Lauritzen, S.L.; Mortera, J. Analysis of forensic DNA mixtures with artefacts. *J. R. Stat. Soc. Ser. C* **2015**, *64*, 1–48. [CrossRef]

13. Get More Information from DNA Mixtures with TrueAllele®Casework. Available online: https://www.cybgen.com/products/casework.shtml (accessed on 16 May 2021).
14. STRmix™. Empowering Forensic Science. Available online: https://www.strmix.com/ (accessed on 16 May 2021).
15. Brenner, C. What is DNA•VIEW®? An Integrated Software Package for DNA Identification. Available online: http://dna-view.com/dnaview.htm (accessed on 16 May 2021).
16. Butler, J.M.; Kline, M.C.; Coble, M.D. NIST interlaboratory studies involving DNA mixtures (MIX05 and MIX13): Variation observed and lessons learned. *Forensic Sci. Int. Genet.* **2018**, *37*, 81–94. [CrossRef] [PubMed]
17. Swaminathan, H.; Qureshi, M.O.; Grgicak, C.M.; Duffy, K.; Lun, D.S. Four model variants within a continuous forensic DNA mixture interpretation framework: Effects on evidential inference and reporting. *PLoS ONE* **2018**, *13*, e0207599. [CrossRef] [PubMed]
18. Swaminathan, H.; Garg, A.; Grgicak, C.M.; Medard, M.; Lun, D.S. CEESIt: A computational tool for the interpretation of STR mixtures. *Forensic Sci. Int. Genet.* **2016**, *22*, 149–160. [CrossRef] [PubMed]
19. Gill, P.; Bleka, Ø.; Hansson, O.; Benschop, C.; Haned, H. *Forensic Practitioner's Guide to the Interpretation of Complex DNA Profiles*; Academic Press: Cambridge, MA, USA, 2020.
20. Taylor, D.; Bright, J.-A.; Buckleton, J.S. The continuous model. In *Forensic DNA Evidence Interpretation*, 2nd ed.; Buckleton, J.S., Bright, J.-A., Taylor, D., Eds.; CRC Press: Boca Raton, FL, USA, 2016.
21. Association of Forensic Science Providers. Standards for the formulation of evaluative forensic science expert opinion. *Sci. Justice* **2009**, *49*, 161–164. [CrossRef] [PubMed]
22. Bright, J.-A.; Evett, I.W.; Taylor, D.; Curran, J.M.; Buckleton, J. A series of recommended tests when validating probabilistic DNA profile interpretation software. *Forensic Sci. Int. Genet.* **2015**, *14*, 125–131. [CrossRef]
23. You, Y.; Balding, D. A comparison of software for the evaluation of complex DNA profiles. *Forensic Sci. Int. Genet.* **2019**, *40*, 114–119. [CrossRef] [PubMed]
24. Manabe, S.; Morimoto, C.; Hamano, Y.; Fujimoto, S.; Tamaki, K. Development and validation of open-source software for DNA mixture interpretation based on a quantitative continuous model. *PLoS ONE* **2017**, *12*, e0188183. [CrossRef]
25. Riman, S.; Iyer, H.; Vallone, P.M. Exploring DNA interpretation software using the PROVEDIt dataset. *Forensic Sci. Int. Genet. Suppl. Ser.* **2019**, *7*, 724–726. [CrossRef]
26. Buckleton, J.S.; Bright, J.-A.; Cheng, K.; Budowle, B.; Coble, M.D. NIST interlaboratory studies involving DNA mixtures (MIX13): A modern analysis. *Forensic Sci. Int. Genet.* **2018**, *37*, 172–179. [CrossRef]
27. Bright, J.-A.; Cheng, K.; Kerr, Z.; McGovern, C.; Kelly, H.; Moretti, T.R.; Smith, M.A.; Bieber, F.R.; Budowle, B.; Coble, M.D.; et al. STRmix™ collaborative exercise on DNA mixture interpretation. *Forensic Sci. Int. Genet.* **2019**, *40*, 1–8. [CrossRef]
28. Benschop, C.C.G.; Hoogenboom, J.; Bargeman, F.; Hovers, P.; Slagter, M.; van der Linden, J.; Parag, R.; Kruise, D.; Drobnic, K.; Klucevsek, G.; et al. Multi-laboratory validation of DNAxs including the statistical library DNAStatistX. *Forensic Sci. Int. Genet.* **2020**, *49*, 102390. [CrossRef]
29. Alladio, E.; Omedei, M.; Cisana, S.; D'Amico, G.; Caneparo, D.; Vincenti, M.; Garofano, P. DNA mixtures interpretation—A proof-of-concept multi-software comparison highlighting different probabilistic methods' performances on challenging samples. *Forensic Sci. Int. Genet.* **2018**, *37*, 143–150. [CrossRef]
30. Eduardoff, M.; Santos, C.; de la Puente, M.; Gross, T.E.; Fondevila, M.; Strobl, C.; Sobrino, B.; Ballard, D.; Schneider, P.M.; Carracedo, Á.; et al. Inter-laboratory evaluation of SNP-based forensic identification by massively parallel sequencing using the Ion PGM™. *Forensic Sci. Int. Genet.* **2015**, *17*, 110–121. [CrossRef] [PubMed]
31. Steensma, K.; Ansell, R.; Clarisse, L.; Connolly, E.; Kloosterman, A.D.; McKenna, L.G.; van Oorschot, R.A.H.; Szkuta, B.; Kokshoorn, B. An inter-laboratory comparison study on transfer, persistence and recovery of DNA from cable ties. *Forensic Sci. Int. Genet.* **2017**, *31*, 95–104. [CrossRef] [PubMed]
32. Köcher, S.; Müller, P.; Berger, B.; Bodner, M.; Parson, W.; Roewer, L.; Willuweit, S. Inter-laboratory validation study of the ForenSeq™ DNA Signature Prep Kit. *Forensic Sci. Int. Genet.* **2018**, *36*, 77–85. [CrossRef] [PubMed]
33. President's Council of Advisors on Science and Technology. *Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods*; Executive Office of the President of the United States: Washington, DC, USA, 2016.

34. Butler, J.M. *Forenisc DNA Typing*, 2nd ed.; Academic Press: Cambridge, MA, USA, 2005.
35. McNevin, D.; Wright, K.; Chaseling, J.; Barash, M. Commentary on: Bright et al. (2018) Internal validation of STRmix™—A multi laboratory response to PCAST, Forensic Science International: Genetics, 34: 11–24. *Forensic Sci. Int. Genet.* **2019**, *41*, e14–e17. [CrossRef]
36. Buckleton, J.S.; Bright, J.-A.; Ciecko, A.; Kruijver, M.; Mallinder, B.; Magee, A.; Malsom, S.; Moretti, T.; Weitz, S.; Bille, T.; et al. 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 Sci. Int. Genet.* **2020**, *44*. [CrossRef]
37. Bright, J.-A.; Stevenson, K.E.; Curran, J.M.; Buckleton, J.S. The variability in likelihood ratios due to different mechanisms. *Forensic Sci. Int. Genet.* **2015**, *14*, 187–190. [CrossRef]
38. Ramos, D.; Gonzalez-Rodriguez, J. Reliable support: Measuring calibration of likelihood ratios. *Forensic Sci. Int.* **2013**, *230*, 156–169. [CrossRef]
39. Bright, J.-A.; Jones Dukes, M.; Pugh, S.N.; Evett, I.W.; Buckleton, J.S. Applying calibration to LRs produced by a DNA interpretation software. *Aust. J. Forensic Sci.* **2021**, *53*, 147–153. [CrossRef]
40. Kelly, H.; Bright, J.-A.; Kruijver, M.; Cooper, S.; Taylor, D.; Duke, K.; Strong, M.; Beamer, V.; Buettner, C.; Buckleton, J. A sensitivity analysis to determine the robustness of STRmix™ with respect to laboratory calibration. *Forensic Sci. Int. Genet.* **2018**, *35*, 113–122. [CrossRef] [PubMed]
41. Moretti, T.R.; Just, R.S.; Kehl, S.C.; Willis, L.E.; Buckleton, J.S.; Bright, J.-A.; Taylor, D.A.; Onorato, A.J. Internal validation of STRmix™ for the interpretation of single source and mixed DNA profiles. *Forensic Sci. Int. Genet.* **2017**, *29*, 126–144. [CrossRef] [PubMed]
42. Taylor, D.; Buckleton, J.; Bright, J.-A. Factors affecting peak height variability for short tandem repeat data. *Forensic Sci. Int. Genet.* **2016**, *21*, 126–133. [CrossRef] [PubMed]
43. Buckleton, J.S.; Bright, J.-A.; Gittelson, S.; Moretti, T.R.; Onorato, A.J.; Bieber, F.R.; Budowle, B.; Taylor, D.A. The probabilistic genotyping software STRmix: Utility and evidence for its validity. *J. Forensic Sci.* **2019**, *64*, 393–405. [CrossRef]
44. Bauer, D.W.; Butt, N.; Hornyak, J.M.; Perlin, M.W. Validating TrueAllele® interpretation of DNA mixtures containing up to ten unknown contributors. *J. Forensic Sci.* **2020**, *65*, 380–398. [CrossRef] [PubMed]
45. Cheng, K.; Bright, J.-A.; Kerr, Z.; Taylor, D.; Ciecko, A.; Curran, J.; Buckleton, J. Examining the additivity of peak heights in forensic DNA profiles. *Aust. J. Forensic Sci.* **2020**, 1–15. [CrossRef]
46. Brookes, C.; Bright, J.-A.; Harbison, S.; Buckleton, J. Characterising stutter in forensic STR multiplexes. *Forensic Sci. Int. Genet.* **2012**, *6*, 58–63. [CrossRef]
47. Morrison, G.S. Special Issue on Measuring and Reporting the Precision of Forensic Likelihood Ratios. *Sci. Justice* **2016**, 56. [CrossRef]

PC59

Pre-publication comments for NISTIR 8351-DRAFT

Jeanette Wallin [redacted]

Fri 11/19/2021 10:55 AM

To: ScientificFoundationReviews <ScientificFoundationReviews@nist.gov>
Cc: Colleen Spurgeon [redacted]

Dear NISTIR 8351-DRAFT authors,

It seems NISTIR 8225, per lines 128-130 of NISTIR 8351-DRAFT, was the guiding framework within which the DNA mixture interpretation foundational review was to be conducted. Given this, the following points are in question.

1. Per NISTIR 8225, the fundamental question to be answered is: *"**What empirical data exist to support the methods that forensic science practitioners use to evaluate evidence?**"* This is reiterated in NISTIR 8351-DRAFT: "*These reviews seek to answer the question: **'What established scientific laws and principles as well as empirical data exist to support the methods that forensic science practitioners use to analyze evidence?***'"

   Moreover, per NISTIR 8225, "peer-reviewed publications are essential building blocks of a respected edifice of scientific knowledge..." It further states, "…studies that address reliability of forensic methods would ideally be present in a discipline's published, peer-reviewed, and well-cited scientific literature."

   Such studies are well represented in the publicly available, peer-reviewed body of scientific literature - from DNA extraction to sampling and PCR variation, to stutter and other artifact behavior characterization, to electrophoretic measurements and kinetic injection behavior, through mixture proportions and intra-locus and inter-locus peak height variation, to where we are now with modeling all this biology using probabilistic genotyping. In fact, NISTIR 8351-DRAFT states, "*Thousands of articles pertaining to forensic DNA methods have been published in dozens of peer-reviewed scientific journals in the past 35 years.*" One may find this vast body of literature overwhelming for such a review but these studies and conclusions represent most of the empirical data providing the foundations of today's practices. This necessarily includes literature beyond the keyword searches in PubMed (pubmed.ncbi.nlm.nih.gov) of "DNA" and "mixture" (see Section 3.1.1), and certainly outside the constraints of 2009-2018. A simple search of "mixture" and "forensic" and "DNA" in PubMed restricted to 2019-current, where this draft report leaves off, turns out another 240 publications. A search of "stochastic" and "forensic" and "DNA" in PubMed results in 113 publications. Scanning through this list to remove irrelevant papers to autosomal STR testing (*e.g.,* SNPs, mtDNA, methylation), leaves approximately 59 relevant publications that address DNA stochastic effects – one of many integral topics when considering forensic DNA mixture interpretation. In fact, NISTIR 8351-DRAFT points out the significance of stochastic affects at least twice: 1.) per lines 2888-2889, "*Stochastic variation when testing small amounts of DNA also impacts sample complexity*", and 2.) "*…this review focuses on methods for interpreting data from complex DNA mixtures, which we define as samples that contain comingled DNA from two or more contributors in which stochastic effects or allele sharing cause uncertainty in determining contributor genotypes*." **Why was this vast published body of literature so narrowly restricted for a *foundational* review?**

2. NISTIR 8351-DRAFT states "*PCAST used the phrase 'foundational validity' to reflect whether something was based on reliable principles and methods and 'validity as applied' to reflect whether the individual performing the work was applying these principles and methods reliably…*

In this chapter, we explore the basis for reliability in DNA mixture measurements and interpretation with a focus on what PCAST termed foundational validity." The chapter referred to is Chapter 4. Considering *foundational* reliability of forensic DNA mixture interpretation – *the established scientific laws and principles* - the *"trusted and established knowledge that supports and underpins the discipline's methods"* is predominantly found in the thousands of publicly available, peer-reviewed published journal articles. It seems the focus was more on the lack of publicly available internal validations. However, as pointed out in NISTIR 8351-DRAFT, internal validations are reflective of "*validity as applied*," not foundational validity. Furthermore, per NISTIR 8351-DRAFT, "*PT data provide insights into how individual analysts performed on specific tests while internal validation studies offer insights into how laboratories performed when analyzing a range of DNA mixtures of varying complexity.*" **Why was the focus of this *foundational* review instead on internal validations and proficiency test data, which reflect "validity as applied"?**

3. Per NISTIR 8225: *"For each area studied, the NIST proposal involved (1) assembling a NIST review team with a range of expertise in order to view issues from multiple perspectives, (2) seeking input on issues to consider from a variety of outside experts, (3) examining the scientific literature to evaluate available support for claims made, (4) conducting interlaboratory studies where appropriate and possible, (5) publishing a written report of findings and recommendations, and (6) sharing findings with the scientific and criminal justice communities to convey the capabilities and limitations of studied forensic disciplines to practitioners, judges, lawyers, jurors, and other stakeholders.*"

   Key Takeaway 4.3 of NISTIR 8351-DRAFT regarding publicly available internal validation data (*"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."*) does not reflect the stated framework of NISTIR 8225, which doesn't seem to require internal validation data for a foundational review. It rather calls for *examining the scientific literature* and *conducting interlaboratory studies*. **Again, why were internal validations the focus?**

4. Per NISTIR 8351-DRAFT, probabilistic genotyping, which seems to be the main focus of Chapter 4 given its newfound popularity, is defined by this vast body of literature coming together in a computerized format: *"PGS: probabilistic genotyping software; a computer program that utilizes statistical genetics, biological models, computer algorithms, and probability distributions to infer genotypes and assign likelihood ratios using either discrete or continuous approaches."* The scientific laws and principles of the biology are the same, whether applied manually, in PGS, or other computer-assisted approaches. The statistical genetics and likelihood ratios are also well established, both within the forensic field and other scientific disciplines; notably, PGS has made likelihood ratio calculations more readily available. The computer algorithms like MCMC may be new to forensics but they are not new to science; a quick "MCMC" search in PubMed returns 2,294 citations. Probability distributions of genotypes are a quantitative measure that objectively quantifies what we as forensic scientists have been doing simplistically with a calculator and qualitatively (based on "*analyst training and experience"*) for decades. **In considering *foundational* validity of DNA mixture interpretation, should it not cover the scientific laws and principles applied to the biology and the statistical genetics applied to the various DNA mixture interpretation approaches?** Instead, the review drilled down to narrowly focus on "*60 published articles on PGS and associated validation studies*" focusing on probabilistic genotyping programs, published during or earlier than 2018, and then declared a lack of available information in Key Takeaway 4.3.

5. Per NISTIR 8225, "*we will evaluate whether the selected features are characterized and measurable; to what extent the discriminating power of those features is known;…*" What features were chosen? Characterized? Measured? Was the extent of the discriminating power tested? From what is presented in NISTIR 8351-DRAFT, the foundational measurable features are not clearly apparent. Sixteen "principles" are declared in Chapter 2, but untested/not evaluated in this review despite having measurable and foundational underpinnings to forensic DNA mixture interpretation. For example, "Principle 8 [Measurement]: PCR amplification is… This principle is a reminder that STR results are a copy of the recovered DNA in a tested sample and depend on the accuracy and efficiency of the copying process. PCR artifacts increase uncertainty for the genotype possibilities of contributors to complex DNA mixtures." Sampling and the PCR process represent a treasure trove of measurement evaluations pertaining directly to the reliability of DNA mixture interpretation. **Why were such foundational measurements not characterized per the published literature?**

6. Per NISTIR 8225, Section 1.2 is "*How will we evaluate the data?*" It goes on to say the following three criteria will be evaluated: 1. Retrievable, 2. Reliable, and 3. Respected. The questions associated with each criterion don't seem to be answered, despite the vast body of publicly available, peer-reviewed published journal articles. **Why is that?** NISTIR 8225 states, "*We believe that for something to be considered foundational, it must be reasonably accessible to anyone who wishes to review it.*" This is a reasonable ask of foundational building blocks and is the case per the *"thousands of articles pertaining to forensic DNA methods"* in the publicly available, peer-reviewed literature.

7. Per NISTIR 8225, a foundational scientific review should include "*obtaining input from… working groups…for the open exchange of ideas and information.*" **To what extent were the following U.S. working group bodies included: SWGDAM, OSAC, and ASCLD?** It is not apparent that these groups of practitioner stakeholders participated *in an open exchange of ideas and information*.

    a. Of significance, an ASCLD/LAB board clarification was published in the July 31, 2015 newsletter for ASCLD accredited laboratories, in part stating: "DNA mixture interpretation procedures must be tested on mixture profiles from known contributors representing the range of mixture types (*e.g.,* different numbers of contributors, mixture proportions, and template quantities) to which the procedure will be applied in casework. The results of this validation must be used to define the capabilities and limitations of the procedure and to verify that it produces the expected results (*e.g.,* inclusions and exclusions)." This does not seem to be acknowledged anywhere in the historical archival appendices nor anywhere else in NISTIR 8351-DRAFT. This seems a miss considering the significance of this reliability requirement in the ASCLD/LAB accreditation standards, and the large percentage of government laboratories in the U.S. with ASCLD accreditation at the time.

8. Paraphrased from NISTIR 8225, NIST will communicate the following in a foundational review: identify methods built on a solid scientific foundation, identify parts that would benefit from strengthening, promote a shared understanding of critical concepts, and identify the discipline's foundational literature to develop a shared understanding of core principles. Such a review would be very valuable, but this does not comport with NISTIR 8351-DRAFT.

Under a NIST and U.S. DOJ MOU, according to NISTIR 8225, "*NIST's scientific foundation reviews fulfill the responsibilities outlined in the fourth element of that MOU,*" which states "*test and validate select forensic science practices and standards as appropriate.*" NISTIR 8225 goes on to say, "*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." "NCFS felt that their request fell within NIST's agreed upon responsibilities under the MOU."* NISTIR 8225 and NISTIR 8225-DRAFT share concerns NIST had related to this task. Perhaps if a more pragmatic and collaborative approach had been developed, the execution of the task would have aligned better with the intended benefits.

Regards,

Jeanette Wallin, M.P.H.
Assistant Lab Director
CA DOJ, Jan Bashinski DNA Laboratory

Colleen Spurgeon, M.S.
Assistant Lab Director
CA DOJ, Jan Bashinski DNA Laboratory

PC60

IEEE-USA's comments on NIST Internal Report 8351-DRAFT “DNA Mixture Interpretation: ..."

Erica Wissolik [redacted]

Fri 11/19/2021 3:07 PM

To: ScientificFoundationReviews <ScientificFoundationReviews@nist.gov>
Cc: Marc Canellas [redacted]; William L Robinson [redacted]

Please find attached, IEEE-USA's comments in response to NIST's requests for reader comments and feedback on the agency's Internal Report 8351-DRAFT “DNA Mixture Interpretation: A NIST Scientific Foundation Review.”

Thank you and please don't hesitate to contact me if you have questions.

Sincerely,
Erica Wissolik
Sr. Program Manager, Government Relations
IEEE-USA
2001 L St, NW, Suite 700
Washington, DC 20036
[redacted]

PC60a

18 November 2021

National Institute of Standards and Technology (NIST)
U.S. Department of Commerce
100 Bureau Drive
Gaithersburg, MD 20899
Via Email: scientificfoundationreviews@nist.gov

Re: RFC Response: NIST Internal Report 8351-DRAFT *DNA Mixture Interpretation: A NIST Scientific Foundation Review*

IEEE-USA is pleased to submit these comments on the above-captioned, Request for Comment on NIST's DNA Mixture Interpretation: A NIST Scientific Foundation Review (8351-DRAFT, "the Review").

IEEE-USA represents approximately 150,000 engineers, scientists, and allied professionals in United States, many of whom are actively conducting research and development into artificial intelligence, software engineering, cybersecurity, and advanced computing, as well as other foundational and emerging technologies. We are the American component of the IEEE – the largest organization of technology professionals in the world, representing more than 400,000 engineers, scientists, and allied professionals worldwide.

The IEEE Standards Association (IEEE-SA), the leading developer of global technical standards used in power and energy, telecommunications, biomedical and healthcare, information technology, transportation, and information assurance products and services, is developing technical standards and frameworks that show how professionals can and should prioritize ethical considerations in the design, development, and deployment of artificially intelligent and autonomous systems (hereinafter referred to collectively as AI systems).[1] Of note, IEEE is developing IEEE P3119 Standard for the Procurement of Artificial Intelligence and Automated Decision Systems which aims to address the needs of government workers, policymakers, and technologists to make meaningful, accountable, and transparent choices regarding the socio-technical considerations and impact of AI products, services, and/or systems encountered by the public.[2]

[1] See, e.g., The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, First Edition. IEEE, 2019. https://standards.ieee.org/content/ieee-standards/en/industry-connections/ec/Autonomous-systems.html (IEEE Ethically Aligned Design); IEEE P7000 Series Standards and Projects addressing topics including transparency, data privacy, and algorithmic bias https://ethicsinaction.ieee.org/p7000/; IEEE Model Process for Addressing Ethical Concerns During System Design, IEEE Standard IEEE 7000-2021; IEEE Recommended Practice for Assessing the Impact of Autonomous and Intelligent Systems on Human Well-Being, IEEE Standard 7010-2020.
[2] IEEE SA Working Group, "Process Model and Requirements Aimed at AI Procurement in a New IEEE Standard," (20 October 2021) https://beyondstandards.ieee.org/process-model-and-requirements-aimed-at-ai-procurement-in-a-new-ieee-standard/.

IEEE-USA | 2001 L Street, N.W., Suite 700, Washington, D.C. 20036-4928 USA
Office: +1 202 785 0017 | Fax: +1 202 785 0835 | E-mail: ieeeusa@ieee.org | Web: http://www.ieeeusa.org

IEEE-USA believes that we stand at an important juncture that pertains less to what new levels of efficiency AI systems can enable, and more to whether these technologies can become a force for good in ways that go beyond efficiency. We have a critical opportunity to use AI systems to help make society more equitable, inclusive, and just; make government operations more transparent and accountable; and encourage public participation and increase the public's trust in government. When used according to these objectives, AI systems can help reaffirm and protect our democratic values.

If, instead, we miss the opportunity to use these technologies to ensure protection of human values and trustworthiness, we risk reinforcing disparities in access to goods and services, discouraging public participation in civic life, and eroding the public's trust in government. Put another way: responsible development and use of AI systems to further safeguard human values and ensure trustworthiness is an approach that leads to a sustainable ecosystem of innovation. It is this type of approach that our society will trust and accept.

IEEE-USA believes that the software and hardware used to perform DNA mixture interpretation, including probabilistic genotyping systems (PGS) (hereinafter referred to collectively as DNA software), are automated decision-making systems that impact the life and liberty of individuals, and should be governed by the same rigorous standards and requirements as other automated decision-making systems such as AI systems.[3]

While there is no single test for determining whether a software system is an 'artificially intelligent' system, modern DNA software, including PGS, is inarguably complex scientific software that leverages the power of computing to automate portions of forensic DNA analytical and decision-making processes. The correct development, verification, validation, and use of DNA software requires specialized technical understanding of complex mathematical, statistical, and computing methods. Under the federally codified definition of AI, DNA software undoubtedly meets the definition of AI.[4] Additionally, DNA software, like many forensic technologies, is an engineered product incorporating scientific models and numeric methods. Too often, DNA software is narrowly seen only as a forensic technology governed by forensic analysts, when it is also software and hardware to be governed by software engineers and other related professionals. Many of those professionals are represented by IEEE-USA. In sum, IEEE-USA believes that general concerns, requirements, standards, and policies regarding AI systems should and do apply to modern DNA software.[5]

---

[3] M. Canellas, "Defending IEEE Software Standards in Federal Criminal Court." Computer, vol. 63, no. 6, pp. 14-23, 2021. doi: 10.1109/MC.2020.3038630.

[4] Section 238(g) of the John S. McCain National Defense Authorization Act for Fiscal Year 2019, Pub. L. No. 115-232, 132 Stat. 1636, 1695 (Aug. 13, 2018) (codified at 10 U.S.C. § 2358, note), defined AI to include the following:

1) Any artificial system that performs tasks under varying and unpredictable circumstances without significant human oversight, or that can learn from experience and improve performance when exposed to data sets.
2) An artificial system developed in computer software, physical hardware, or another context that solves tasks requiring human-like perception, cognition, planning, learning, communication, or physical action.
3) An artificial system designed to think or act like a human, including cognitive architectures and neural networks.
4) A set of techniques, including machine learning, that is designed to approximate a cognitive task.
5) An artificial system designed to act rationally, including an intelligent software agent or embodied robot that achieves goals using perception, planning, reasoning, learning, communicating, decision-making, and acting.

[5] While this Comment focuses particularly on DNA software, the same general concerns, requirements, standards, and policies apply to other forensic technologies that serve as automated decision systems such as face recognition, image recognition, fingerprint identification, and predictive policing.

The most important line in NIST's current draft is KEY TAKEAWAY #4.3 which concludes that "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." Because this is the reality for a process and for software used in decisions of whether to deprive people of their rights and liberties, is an indictment of the lack of trustworthiness for these systems and software. With this lack of ability to determine the reliability of DNA mixture interpretation practices including PGS, neither these systems, nor their results can be considered reliable or trustworthy. To remedy this issue, we draw upon our collective expertise in line with the goals of the RFC to provide the following recommendations:

1. **DNA mixture interpretation using DNA software should only be deemed reliable based on objective information gathered through independent verification and validation as determined by IEEE Standard 1012.**

The use of DNA software in criminal court can result in catastrophic failures through false imprisonment and the deprivation of people's rights. Scientists and engineers have long demanded that safety-critical software and hardware be the right systems built the right way. Therefore, DNA software should be independently verified and validated (IV&V) prior to deployment, or prior to informing decisions in the legal system, law enforcement, governance, and related compliance. Specifically, DNA software ought to be independently verified and validated in accordance with IEEE Standard 1012, IEEE Standard for System, Software, and Hardware Verification and Validation,[6] and be subject to recurring post-deployment audit, including with respect to their operators. IEEE-USA encourages NIST uphold these same requirements.

Sponsored by the IEEE Computer Society, IEEE Standard 1012 is a universally applicable and broadly accepted process for ensuring that the right product is correctly built for its intended use. It is used to verify and validate Department of Defense nuclear weapons systems and NASA manned space systems and critical space exploration probes, among many others.

IV&V are interrelated and complementary processes that build quality into any system. Verification is focused on a product, providing objective evidence for whether the product conforms to requirements, standards, and practices. Validation is focused on customers and stakeholders, providing evidence for whether a product is accurate and effective, solves the right problem, and satisfies the intended use and user needs in the operational environment. In short, verification ensures that a product is correctly built, while validation ensures that the right product is built.

In the context of DNA software, IV&V answers the following types of questions:[7] Is the model of DNA analysis used by the software the best available, coded as designed, and appropriate for the problem? Does DNA software systematically favor including defendants? How likely are false negatives and false positives? Would outside experts agree with the software's results at each stage of analysis?

[6] IEEE Standard for System, Software, and Hardware Verification and Validation, IEEE Standard 1012-2016, Sept. 2017 (hereinafter referred to as IEEE Standard 1012).

[7] N. Adams, R. Koppl, D. Krane, W. Thompson, and S. Zabell, "Letter to the editor — appropriate standards for verification and validation of probabilistic genotyping systems," J. Forensic Sci., vol. 63, no. 1, pp. 339–340, 2018. doi: 10.1111/1556-4029.13687.

To appropriately perform IV&V, IEEE Standard 1012 requires that each software and hardware component be assigned an integrity level that increases depending on the likelihood and consequences of a failure: negligible, marginal, critical (causing "major and permanent injury, partial loss of mission, major system damage, or major financial or social loss"), and catastrophic (causing "loss of human life, complete mission failure, loss of system security and safety, or extensive financial or social loss").[8] As the integrity level increases, so too does the intensity and rigor of the required IV&V tasks.

DNA software analysis tools, like all software, should undergo IV&V according to its integrity level as defined by IEEE Standard 1012. Because a thorough and public conversation is yet to take place, there is presently no consensus on such an integrity level. However, the likelihood of DNA software to cause wrongful convictions in the criminal justice system clearly constitutes catastrophic failure, and therefore should be held to the highest integrity level, the level where IV&V should be performed independently.

The IV&V process must be independent to avoid conflicts of interest that could lead to catastrophic failure. To this end, IEEE Standard 1012 requires technical, managerial, and financial IV&V when testing software and hardware where catastrophic consequences could occasionally occur and where critical consequences will probably occur. Moreover, letting developers certify their own software is a clear conflict of interest, and the IEEE/Association for Computing Machinery Code of Ethics for Software Engineers is clear about the obligation of developers to manage competing aims.[9] Full definitions of technical, managerial, and financial independence from IEEE Standard 1012 are below, but, in brief, the following must all be independent from the group that oversees the design and building of software: personnel, problem formulation, test and analysis tools for IV&V (technical), responsibility for IV&V (managerial), and control of the budget for IV&V (financial).[10]

Specifically, technical independence "[r]equires the IV&V effort to use personnel who are not involved in the development of the system or its elements. The IV&V effort should formulate its own understanding of the problem and how the proposed system is solving the problem."[11] "Technical independence means that the IV&V effort uses or develops its own set of test and analysis tools separate from the developer's tools."[12] And if sharing tools is necessary, "IV&V conducts qualification tests on tools to assure that the common tools do not contain errors that may mask errors in the system being analyzed and tested."[13] This independence requires the exclusion of parties with a stake in the outcome, which for forensic technologies includes forensic labs that, while not financially dependent on developers, have a shared interest in software's acceptance.

Managerial independence "[r]equires that the responsibility for the IV&V effort be vested in an organization separate from the development and program management organizations. Managerial independence also means that the IV&V effort independently selects the segments of the software, hardware, and system to analyze and test, chooses the IV&V techniques, defines the schedule of IV&V activities, and selects the specific technical issues and problems to act on."[14] The IV&V effort must be

[8] IEEE Standard 1012, p. 196.
[9] D. Gotterbarn, K. Miller, and S. Rogerson, "Computer society and ACM approve software engineering code of ethics," Computer, vol. 32, no. 10, pp. 84–88, 1999. doi: 10.1109/MC.1999.796142.
[10] IEEE Standard 1012, p. 198.
[11] IEEE Standard 1012, p. 198.
[12] IEEE Standard 1012, p. 198.
[13] IEEE Standard 1012, p. 198.
[14] IEEE Standard 1012, p. 198.

"allowed to submit to program management the IV&V results, anomalies, and findings without any restrictions (e.g., without requiring prior approval from the development group) or adverse pressures, direct or indirect, from the development group."[15]

Financial independence "[r]equires that control of the IV&V budget be vested in an organization independent of the development organization. This independence prevents situations where the IV&V effort cannot complete its analysis or test or deliver timely results because funds have been diverted or adverse financial pressures or influences have been exerted."[16]

It is clear from these definitions that peer-reviewed publications, while a priceless tool for scientific inquiry, are not a substitute, nor a valid approximation of IV&V when determining reliability or trustworthiness of a deployed system. Peer reviewed publications form the foundation of scientific advancement, but peer reviewers of scientific publications are not tasked with answering questions like "Should the DNA software be admissible in court? Is the DNA software fit for the evidence in this legal case?" Peer reviewers do not have access to the system itself and are not tasked with assessing its reliability. Peer reviewers are assessing whether a publication deserves the attention of the scientific community, whether the results described deserve the attention of other scientists. With respect to specific legal cases, any individual case could go well beyond the bounds of the published studies. For example, a case could involve more contributors, a smaller evidence sample, or a different version of the software than was examined in the peer-reviewed studies.

Moreover, as exemplified by the Review's Table 4.3, most peer-reviewed studies of probabilistic genotyping software are not independent, violating a fundamental tenet of IV&V, and making them insufficient to determine reliability. The peer-reviewed studies of TrueAllele or STRmix were almost exclusively authored by employees of the companies who developed the software, or laboratories who have at least an implied conflict of interest since the software they use needs to be viewed as reliable in order for it to be admissible under law. The lack of independent review raises serious concerns about the reliability of the studies themselves and was the chief criticism of DNA software in a report by the President's Council of Advisors on Science and Technology (hereinafter the "PCAST report").[17] The PCAST Report called for more testing that "should be performed by or should include independent research groups not connected with the developers of the methods and with no stake in the outcome."

To address the fact that peer-reviewed studies are no substitute for IV&V and the lack of independence in those peer-reviewed studies, IEEE-USA specifically recommends that:

- TAKEAWAY #4.1 be rewritten to state that the "The degree of reliability of a component or a system can be assessed using empirical data obtained through *technically, managerially, and financially independent verification and* validation studies, *and post-deployment audits*" (changes emphasized).
- The analysis of validation experiments in Chapter 4 includes managerial, technical, and financial independence as "influencing factors" in their analysis.

[15] IEEE Standard 1012, p. 198.
[16] IEEE Standard 1012, p. 198.
[17] President's Council of Advisors on Science and Technology, Report to the president Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods, Washington DC, 2016. https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/PCAST/pcast_forensic_science_report_final.pdf.

- NIST state that IEEE Standard 1012 is applicable to DNA mixture interpretation, especially when DNA software or hardware are used.
- NIST state that DNA software should undergo IV&V according to its integrity level as defined by IEEE Standard 1012.

2. **DNA software should be tested against and governed by standards adhering to principles of due process, openness, consensus, balance, and right of appeal.**

By definition, standards are "published documents that establish specifications and procedures designed to maximize the reliability of the materials, products, methods, and/or services people use every day."[18] They are the basis on which the safety and credibility of new products and markets are verified, making them fundamental to the modern economy.[19] Because standards have such a profound effect, standards-setting organizations (SSOs), like IEEE SA, have significant legal obligations regarding the standards they develop and the processes by which they craft those guidelines, including contract, intellectual property, and antitrust law.[20] Among the many U.S. Supreme Court opinions dealing with SSOs, there are two particularly relevant rules the organizations must abide by to avoid liability: fair processes and independence.[21]

As a result, the IEEE SA standards-development process follows a well-defined and documented path, from concept to completion, guided by a set of five basic principles and imperatives that ensure fairness and good practices during the development cycle.[22]

- Due process: having highly visible procedures for standards creation and following them.
- Openness: ensuring that all interested parties can participate and are not restricted to a particular type or category.
- Consensus: requiring a supermajority of a group to approve a draft standard (75% of the ballots must be returned, with 75% of them voting yes).
- Balance: ensuring that voting groups include all interested participants and avoid an overwhelming influence by any one party.
- Right of appeal: allowing anyone to appeal a standards development decision at any point, before or after approval.

IEEE-USA recommends that NIST distinguish between guidelines and standards, and evaluate whether the guidelines or standards from the Scientific Working Group on DNA Analysis Methods (SWGDAM), the International Society for Forensic Genetics (ISFG), the European Network of Forensic Science Institutes (ENFSI), the UK Forensic Science Regulator, the American National Standards Institute and AAFS Standards Board (ANSI/ASB), and the IEEE Standards Association adhere to the five principles of standard-development process and address the potential for conflicts of interest.

---

[18] "Developing standards." IEEE Standards Association. https://standards.ieee.org/develop/index.html (accessed Nov. 14, 2021).
[19] "Developing standards." IEEE Standards Association. https://standards.ieee.org/develop/index.html (accessed Nov. 14, 2021).
[20] A. Updegrove. "Laws, cases and regulations in the essential guide to standards." ConsortiumInfo, 2013. https://www.consortiuminfo.org/essentialguide/laws.php (accessed Nov. 14, 2021).
[21] A. Updegrove. "Laws, cases and regulations in the essential guide to standards." ConsortiumInfo, 2013. https://www.consortiuminfo.org/essentialguide/laws.php (accessed Nov. 14, 2021).
[22] "Developing standards." IEEE Standards Association. https://standards.ieee.org/develop/index.html (accessed Nov. 14, 2021).

3. **There must be standards and certifications for DNA software and their operators, and recurring benchmarking exercises and independent studies to ensure DNA software's effectiveness, competence, inclusiveness, accountability, and transparency in operation.**

The Review highlights a significant lack of guidance for testing and evaluating DNA software and its protocols[23] and the lack of publicly available data.[24] To address these deficiencies, IEEE-USA believes, and NIST should recommend, that governments should make the reports documenting the required IV&V and audits of their DNA software public. Furthermore, we believe that governments, including NIST, should encourage, develop, and update standards and certifications for DNA software and their operators, and fund recurring benchmarking exercises and independent studies to ensure their effectiveness, competence, inclusiveness, accountability, and transparency in operation. Specifically, we believe these standards, certifications, exercises, and studies should address:

- The requirements for informed trust by the general public in DNA software (see Recommendation #6 below) and the development of metrics that are immediately and easily accessible by experts and non-experts alike;
- The existence or absence of reliable and unbiased underlying scientific principles and methods in DNA software;
- The requirements for recurring testing and auditing of the operation of DNA software, including the operators, field conditions, testing data, environments, methodologies, and performance metrics;
- The requirements for publicly available documentation by developers and testers of DNA software, and of the use of DNA software in individual and aggregate cases and decisions;
- The requirements for certification or loss of certification of operators and DNA software, and for their validation for DNA software already in use;
- The requirements for individuals to be able to access, review, contest, and correct the data about them, to review and contest the decisions that affect them, and to request human review of such data and decisions;
- The requirements for operation in an ethical manner; and,
- The requirements for identifying and addressing vulnerabilities and threats to security, safety, and privacy such as spoofing, evasion attacks, transfer learning attacks, and data poisoning.

4. **Determining the reliability and trustworthiness of forensic technologies like DNA software requires evaluating them in their operational environments, their use in legal proceedings and how fit the technology was for those uses.**

IV&V is predicated on the value of testing technology in operational environments. No software or hardware is "generally" reliable -- any technology is only fit for certain purposes. Even technologies that are widely considered to be reliable have known failure modes. For example, cellular telephones are widely considered to be reliable but are not classified as "generally" reliable because they do not work effectively in tunnels or underground. Further, the desire for a technology to be classified as "generally" reliable rather than to consider its reliability in a particular case is misguided. A core premise of labeling

[23] KEY TAKEAWAY #A1.3 "Limited information has been provided in guidance documents, such as the FBI Quality Assurance Standards or the SWGDAM guidelines, regarding suggested or required studies to inform mixture interpretation protocols."
[24] 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."

a product or process as "well-engineered" is that these operating conditions are specifically defined, tested against pre-defined standards, and accompanied with estimated rates of failure. Systems like DNA software are engineered products incorporating scientific models and therefore require not only the perspective of researchers who have published proofs-of-concept but also engineers who have used product trials and operational testing and evaluation to demonstrate system performance in operating conditions, against predefined standards, and estimated rates of failure.

Therefore, a scientific foundation review of the reliability and trustworthiness of forensic technologies cannot be effective if detached from an analysis of how the technology is used in legal proceedings, in the forensic technology's operational environment -- yet that is exactly what this Review is purporting to do. The Review examines the peer-reviewed and forensic laboratory studies but does not compare that to any of the thousands of criminal cases where DNA software has been used. Notwithstanding the concerns over peer-reviewed studies discussed above, if the DNA profiles and contribution proportions analyzed in legal proceedings are not similar to the samples used in the peer-reviewed studies, the studies have little value. This is particularly true for many types of DNA software, especially PGS, because of its non-continuous nature, meaning that a small set of inputs cannot be reliably interpolated into cases involving different sets of inputs.

To determine the reliability of DNA software systems, IEEE-USA recommends NIST catalog and evaluate how DNA software is being used in legal proceedings and how fit the technology is for those uses.

5. **Users of DNA mixture interpretation include far more than forensic scientists, attorneys, judges, and juries. They include the public upon whom these systems are used, litigants, academics, journalists, and other researchers. For those users to assess the degree of reliability, validity, and whether that information is fit-for-purpose, they need appropriate access to the software.**

IEEE-USA believes that users are too often inappropriately denied access or forced to overcome improper and unnecessary barriers to access DNA software in order to determine the degree of reliability, validity, and whether that information is fit-for purpose. It is true that providers and users have responsibilities as described in KEY TAKEAWAY #4.2 and Section 4.1.5 but there are many more users of DNA mixture interpretation than merely forensic scientists, judges, or juries. Independent testing of proprietary or government DNA software by litigants, academics, journalists, and other researchers is needed to ensure that DNA software are properly vetted and held accountable. NIST should recommend governments clarify whether and how proprietary DNA software may be reverse engineered, modified, and evaluated under laws such as the Computer Fraud and Abuse Act and the anti-circumvention provision of the Digital Millennium Copyright Act, and rules of procedure and evidence. More broadly, NIST should recommend governments take steps to affirmatively promote awareness, access, research, and testing including:

- Ensuring accountability and transparency in government procurement and contracting for DNA software;
- Identifying and disclosing the DNA software used by the government;
- Adopting clear procedures relating to collection, usage, storage and sharing of personal information used by DNA software;

- Providing constituents notice about DNA software decisions, explanations for those decisions, and processes for challenging decisions or data; and,
- Specifically, in legal disputes, tribunals should permit disclosure under appropriate protective orders of intellectual property related to DNA software when necessary to obtain evidence in compliance with other judicial requirements, including constitutional requirements, discovery laws, or subpoenas.

6. Trustworthiness is determined by more than reliability, and therefore, to determine trustworthiness, one must assess the processes and procedures where these systems are deployed.

Technical assessments of reliability as surveyed in the Review are not the sole determination of trustworthiness. There are eight principles for creating and operating AI systems that further human values and ensure trustworthiness:[25] (i) human rights: AI systems shall be created and operated to respect, promote, and protect internationally recognized human rights; (ii) well-being: AI system creators shall adopt increased human well-being as a primary success criterion for development; (iii) data agency: AI system creators shall empower individuals with the ability to access and securely share their data, to maintain people's capacity to have control over their identity; (iv) effectiveness: AI system creators and operators shall provide evidence of the effectiveness and fitness for the purpose of AI systems; (v) transparency: the basis of a particular AI system decision should always be discoverable; (vi) accountability: AI systems shall be created and operated to provide an unambiguous rationale for all decisions made; (vii) awareness of misuse: AI system creators shall guard against all potential misuses and risks of AI systems in operation; and, (viii) competence: AI system creators shall specify and operators shall adhere to the knowledge and skill required for safe and effective operation.

Therefore, IEEE-USA recommends that NIST evaluate more than the technical assessments of reliability to determine trustworthiness. Below we list additional requirements for ensuring the trustworthiness of AI systems in general which includes the automated decision systems such as DNA software and many of the forensic technologies used today. These requirements should be used as factors of analysis in Chapter 4 when evaluating the various forensic laboratories and when developing an overall assessment for forensic technologies as described in Chapter 6. While Chapter 6 recommends assessing (i) how a new technology works, (ii) what its limitations are, and (iii) how it might specifically address the problem to be solved, these three factors are insufficient to ensure trustworthy use of AI systems, DNA software or any forensic technology. If those providing or using DNA software or any other forensic technologies do not adhere to these requirements, then they should not be deemed trustworthy or fit for their use in determining or affecting people's rights and liberties.

To ensure trustworthiness of AI systems, DNA software, and other forensic technologies, IEEE-USA believes that governments and forensic laboratories should be required to:

a. Ensure awareness, access, and research on the existence, fairness, safety, security, privacy, and ethical and societal impacts of DNA software.

[25] The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, First Edition. IEEE, 2019. https://standards.ieee.org/content/ieee-standards/en/industry-connections/ec/Autonomous-systems.html

Governments should: (i) publicly identify and disclose the DNA software used by the government; (ii) conduct and publicly disclose a methodological validation study that establishes the value of using new DNA software in place of existing practices prior to deploying DNA software; (iii) adopt clear procedures relating to the collection, usage, storage, and sharing of personal information in the context of developing, using, and validating a given DNA software in a privacy-preserving manner; and (iv) prevent intellectual property, confidentiality claims, lack of funding, or lack of an designated independent body within government to monitor compliance from impeding duly limited independent validation and verification and publicly disclosed review of the fairness, safety, security, privacy, and ethical and societal impacts of DNA software. DNA software ought to be submitted voluntarily, to the agency performing validation and verification thereof, and the agency using related private intellectual property or proprietary data in its evaluation must adopt rules to protect such private rights from misappropriation.

Users and the public should be allowed to (i) request and receive an explanation of how a government determination using DNA software was reached; (ii) determine whether the DNA software used in government decision-making disproportionately impacts a protected class; and (iii) rectify, challenge, or complete inaccurate or incomplete personal data that is part of the DNA software system or decision.

b. Commit to removing barriers to parties' access to information needed to ascertain relevant evidence about and from DNA software in legal disputes.

Specifically, in legal disputes where judges, juries, and lawyers are the users of DNA software results, barriers to parties' access to information needed to ascertain relevant evidence about and from DNA software should be eliminated.[26] Intellectual property protections should not be used as a shield to prevent duly limited disclosure of information needed to ascertain whether DNA software meets acceptable standards of effectiveness, fairness, and safety. Specifically, in legal disputes, tribunals should permit disclosure under appropriate protective orders of intellectual property related to DNA software necessary to obtain evidence in compliance with other judicial requirements, including constitutional requirements, discovery laws, or subpoenas. Furthermore, laws, procedures, and public funding should not make it more difficult for non-government parties in legal disputes to develop, obtain expertise regarding, or gain access to evidence from DNA software than for government parties to do so.

c. Ensure accountability and transparency in procurement and contracting for DNA software.

To support awareness, access, and research on the existence, fairness, safety, security, privacy, and ethical and societal impacts of DNA software, there must be accountability and transparency in

[26] For example, when source code is ordered to be provided, "information needed" requires providing sufficient information for the recipient to build, run, and test the software themselves including, at minimum:

- All software dependencies including third-party code libraries, toolboxes, plugins, frameworks, and databases;
- Software engineering and development materials describing the development, deployment, and maintenance of the version(s) of the software system used in the instant case, including software engineering documents and build instructions;
- All records of software glitches, crashes, bugs, or errors encountered during the developmental validation study;
- Software version numbers of the components of the system used for the developmental validation study; and,
- All records of unexpected results, including false inclusions, false exclusions and the conditions under which the unexpected results were achieved.

When source code is ordered to be provided, "access" requires, at minimum, that the source code be made available for inspection, in a format allowing it to be reasonably reviewed, searched, and tested, during normal business hours or at other mutually agreeable and reasonable times, and at mutually agreeable and reasonable locations.

government procurement and contracting for DNA software. The government should not procure DNA software that (i) require the governmental entity to indemnify vendors for any and all negative outcomes; (ii) do not adhere to the eight principles in IEEE's Ethically Aligned Design for creating and operating DNA software that further human values and ensure trustworthiness (as may be reflected in articulated guidelines, standards, certifications, audits, and other sound documentation);[27] (iii) do not comply with federal, state, and local anti-discrimination laws; or, (iv) are shielded from independent validation and verification, and public review.

7. **To ensure DNA software is reliable and trustworthy, governments should provide sufficient funding for testing, evaluation, certification, and investigation of DNA software.**

Throughout the document, the Review highlights the value of government funding in the development of research on DNA software (e.g., Sections 3.1.5, A2.3, and A2.3.2). IEEE-USA believes NIST should go further by including a KEY TAKEAWAY recommending that governments provide sufficient funding for testing, evaluation, certification, and investigation of DNA software. The adoption and acceptance of DNA software requires developing and sustaining public confidence in their quality, reliability, and compliance with regulations and social norms. Increased government funding for government and independent third-party evaluation and certification of DNA software is essential to ensure efficacy, transparency, traceability, accountability, and competency. Development of design requirements, methods, metrics, and environments so that DNA software can be tested and evaluated for interactions with different autonomous agents, including humans, and adversarial exploitation is critical in the adoption and acceptance of DNA software. To this end, mechanisms must be developed for identifying and accounting for the features of DNA software that could cause current testing, evaluation, certification, and investigation methods to misinform decision makers or the public about the risk of system deployment or the causes of system malfunction.

IEEE-USA thanks NIST for considering these comments in the agency's revisions to the Request for Comment on NIST's DNA Mixture Interpretation: A NIST Scientific Foundation Review. We would welcome any further discussions with the agency on these matters. If you have questions, please do not hesitate to contact Erica Wissolik at (202) 530-8347 or e.wissolik@ieee.org.

Sincerely,

William Robinson
IEEE-USA Vice President, Government Relations

[27] The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, First Edition. IEEE, 2019. https://standards.ieee.org/content/ieee-standards/en/industry-connections/ec/Autonomous-systems.html

PC61

RE: Comment on NISTIR 8351-DRAFT

Elizabeth Vasquez [redacted]

Fri 11/19/2021 5:44 PM

To: ScientificFoundationReviews <ScientificFoundationReviews@nist.gov>
Cc: Clinton Hughes [redacted]

Good evening,
Please find attached our brief follow-up comment on NISTIR 8351-DRAFT.

Thank you,
Elizabeth

Elizabeth Daniel Vasquez
Director, Science & Surveillance Project
Brooklyn Defender Services
177 Livingston Street, 5th Floor
Brooklyn, New York 11201
[redacted]

From: Elizabeth Vasquez
Sent: Monday, August 23, 2021 1:56 PM
To: scientificfoundationreviews@nist.gov
Cc: Clinton Hughes [redacted]
Subject: Comment on NISTIR 8351-DRAFT

Good afternoon,
Please find attached our Comment on NISTIR 8351-DRAFT—*DNA Mixture Interpretation: A NIST Scientific Foundation Review*.

Thank you,
Elizabeth

Elizabeth Daniel Vasquez
Director, Science & Surveillance Project
Brooklyn Defender Services
177 Livingston Street, 5th Floor
Brooklyn, New York 11201
[redacted]

PC61a

November 19, 2021

Via Electronic Mail
National Institute of Standards and Technology
Attn: Special Programs Office – Scientific Foundation Review
U.S. Department of Commerce
100 Bureau Drive Stop 4701
Gaithersburg, Maryland 20899-4701
scientificfoundationreviews@nist.gov

Re: Follow-up to Request for Comment on NISTIR 8351-DRAFT—*DNA Mixture Interpretation: A NIST Scientific Foundation Review*.

Dear NIST Scientific Foundation Review Team:

Brooklyn Defender Services ("BDS") submits these comments in response to the re-opening of the comment period for *DNA Mixture Interpretation: A NIST Scientific Foundation Review* draft, NISTIR 8351-DRAFT (hereinafter, "Draft Report"), published on June 9, 2021.

As we previously noted, BDS is a full-service public defender organization in Brooklyn, New York, that provides multi-disciplinary and client-centered criminal defense, family defense, immigration, and civil legal services, along with social work and advocacy support. BDS represents low-income people in nearly 30,000 criminal, family, civil, and immigration proceedings each year. Over the last decade, we have witnessed firsthand the dramatic expansion of forensic DNA analysis to more and more cases. NIST is correct that these methods are now used as a matter of routine in everyday casework. In response to this development, BDS established a dedicated Science & Surveillance Project and Forensic Science Practice. This team focuses on remaining abreast of and responding to developments and issues of data, science, and technology in the criminal legal system.

We write to briefly clarify that ESR's recent comment regarding our organization is factually incorrect. In their supplemental comment, ESR wrote:

> 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.

This sentence is misleading; in truth, it appears to refer not to the work of Brooklyn Defender Services,[1] but instead to the work of NIST's own Expert Working Group on Human Factors in Forensic DNA Interpretation and a presentation describing troubling research conducted by ESR staff that was given to that body by ESR's Dr. John Buckleton.

Specifically, one of our staff attorneys is a member of the Expert Working Group on Human Factors in Forensic DNA Interpretation. Dr. John Buckleton – whom BDS has hired in the past as a defense expert – gave a presentation on research ESR staff have conducted that bears directly on the kinship problem at a meeting of that group.

Following his presentation to the Human Factors, members of that Human Factors group reached out to Dr. Buckleton for the underlying data in an as-yet unpublished study by Dr. Buckleton and others involving mixtures of simulated siblings. The results, as presented by Dr. Buckleton, have been discussed extensively within the Human Factors group. We at BDS are anxiously awaiting the publication of this study by ESR, as we expect the results to show the terrifying danger of not accounting for relatedness when analyzing complex DNA mixtures.

Despite this interest in ESR's research and publication of this critical data, BDS did not receive data from ESR. If we had, we would have shared that data publicly because of its importance for our clients and its impact on avoiding miscarriages of justice.

Sincerely,

/s/ Elizabeth Daniel Vasquez
Elizabeth Daniel Vasquez
Director, Science & Surveillance Project

/s/ Clinton Hughes
Clinton Hughes
Forensic DNA Attorney, Criminal Practice
Brooklyn Defender Services
177 Livingston Street, 7th Floor
Brooklyn, New York 11201

[1] As previously noted in our August comment, the data obtained by Brooklyn Defender Services is publicly-available here.

PC62

ASCLD Comments - 11-19-21

Sudkamp, Laura B (KSP) [redacted]

Fri 11/19/2021 7:32 PM

To: ScientificFoundationReviews <ScientificFoundationReviews@nist.gov>

Attached to this email are comments and an offer from the American Society or Crime Laboratory Directors to the draft report: DNA Mixture Interpretation: A Scientific Foundation Review. We truly appreciate the opportunity.

Laura

Laura B. Sudkamp
Division Director
Kentucky State Police
Forensic Services
[redacted]

PC62a

AMERICAN SOCIETY OF
CRIME LABORATORY DIRECTORS, INC.®

65 Glen Road, Suite 123, Garner, NC 27529

November 19, 2021

RE: ASCLD Supplemental Comments to Draft of DNA Mixture Interpretation: A NIST Scientific Foundation Review (NISTIR 8351-DRAFT)

The American Society of Crime Laboratory Directors (ASCLD) has over 700 members comprised of crime laboratory directors, managers, and supervisors from the United States and around the world. The membership consists of scientists and law enforcement officers whose major function is the management of a forensic science laboratory, as well as educators and instructors in forensic science. As such, we are well-versed in the need and requirement to validate new technologies before incorporating them into standard laboratory protocol and practice. Validation of DNA mixture interpretation, including software, is no exception. ASCLD provided public comments on the first NIST public solicitation regarding this report. NIST then posted a second solicitation "to receive additional comments, new data, or information" found at: https://content.govdelivery.com/accounts/USNIST/bulletins/2f8b05e, "Second Public Comment Period for NISTIR-8351-DRAFT Report: Oct. 22 to Nov. 19, 2021." The NIST draft report attempts to conduct a foundational review of the methods used to interpret DNA mixtures. ASCLD made initial comments in a previous document dated August 14, 2021 and is providing additional public comments in this document.

The authors of the draft report conducted a "google search" and reviewed only data which was publicly available. This is not an effective method of acquisition as it relates to data from forensic science service providers. When forensic laboratories conduct validation or performance verification studies, they typically create ground truth datasets from items they have purchased or can directly verify the source. This includes the various bodily fluids used in the validation and development of the DNA mixture interpretation methods. Typically, laboratory staff and organizational employees and their families are the source of these fluids for validation purposes. The public release of the genetic information of our

laboratory staff and other sources is legally prohibited per the federal Genetic Information Nondiscrimination Act of 2008 (GINA). We do not object to making the data available for review by qualified experts with legitimate interest; however, we advocate doing this in an informed and responsible manner that is based upon compliance with federal law, state and local law and policy, case law, and state records request statutes. Laboratories have published executive summaries and journal articles referencing the source data, but are subject to federal, state, and local law regarding public release of the data itself.

A very brief, informal, preliminary poll of ASCLD member laboratories resulted in twelve (12) state, six (6) county, and five (5) city laboratories willing to make the data available for review as long as the underlying validation data is protected from public disclosure per GINA regulations. These volunteer laboratories extend from the East Coast to the West, and include several of the largest states by population. A more formal poll may increase that number significantly.

ASCLD would like to invite the NIST foundational study authors, along with team members or contracted staff who have forensic DNA mixture casework and auditing experience, to visit ASCLD member laboratories on-site to view their validation data. This exercise would surely assuage the authors' concerns about a lack of data in this area. Because of the ongoing need to protect the genetic information of staff members and other participants who provided samples for validation purposes without disclosure agreements, we would ask that prior to the onsite visit, individuals reviewing the data sign an MOU. This practice has been effectively used by laboratories during external audits to protect the security and confidentiality of casework data that is reviewed by external auditors. In addition to a signed MOU, we also request that if the authors intend to review this data with evaluation criteria that differs from that already used and discussed in the draft DNA mixture publication, that NIST provide specific criterion upon which DNA mixture validation studies will be evaluated. That allows the laboratory to efficiently provide the NIST researchers with the exact data they are seeking in a usable format. Finally, ASCLD requests that the NIST or NIST contracted reviewers and auditors provide the results of the individual laboratory evaluations to the individual forensic laboratory outside of the public forum to facilitate constructive dialog between the ASCLD member laboratory and the NIST representative regarding what is considered acceptable. ASCLD recognizes that there is no single authority on how to conduct validation work, and there can be differing, but equally valid, opinions among experts. Because of this, ASCLD member laboratories welcome feedback from NIST; however, it is not productive to have these types of discussions in a public setting.

In the future, NIST should consider creating an anonymized, national clearinghouse of validation data. NIST might also provide a validation outline and sample requirements, or provide the physical samples themselves, so all parties evaluate the methods and software in the same way. Common source materials would eliminate GINA issues, allow identification of individual participant issues, and provide more result consensus. NIST is highly skilled at providing quality reference materials for numerous technologies. As it is very difficult for

forensic science services providers to obtain biological sample sets, a NIST mixture interpretation sample set would be welcomed by forensic science service providers.

Validation of mixture interpretation methods for use in forensic laboratories has been extensive and appropriate. Data is available, just not in a public forum for legal reasons. We encourage the authors of the draft report to contact ASCLD for coordination with volunteer laboratories to provide access to the data within the provisions of federal law and in a manner that protects the confidential, personal, genetic information.

Thank you for the opportunity to comment on NISTIR 8351-DRAFT entitled *DNA Mixture Interpretation: A NIST Scientific Foundation Review* and also provide supplemental comments to our original submission.

Sincerely,

Laura Sudkamp

Laura B. Sudkamp
President
ASCLD

PC63

Idaho State Police Forensic Services Comments on NISTIR-8351 Draft Report

Gamette, Matthew [redacted]

Fri 11/19/2021 11:48 PM

To: ScientificFoundationReviews <ScientificFoundationReviews@nist.gov>

1 attachments (2 MB)

ISP Lab STRmix V2.8 Validation Final.pdf;

Idaho State Police Forensic Services (ISPFS) is an accredited DNA laboratory system with a forensic biology/DNA section that is an NDIS participating laboratory. ISPFS relies on robust and critically reviewed validation studies to support our mixture interpretation protocols and procedures. ISPFS was one of the first laboratory systems in the country to implement probabilistic genotyping software into laboratory methods. As a part of that validation process, ISPFS was a participant in a peer reviewed multi-laboratory publication in the journal "Forensic Science International Genetics." This article highlights the analysis of 2825 mixtures from 31 laboratories.

Your second solicitation, "to receive additional comments, new data, or information" found at https://content.govdelivery.com/accounts/USNIST/bulletins/2f8b05e, Second Public Comment Period for NISTIR-8351-DRAFT Report: Oct. 22 to Nov. 19, 2021, specifically requests new data. ISPFS has performed several validation/performance verifications of our DNA methods, including the use of the probabilistic genotyping software STRmix™. We have initial validation data when we put STR testing online, validation data when we upgraded to STRmix™, and data from last year when we moved to Applied Biosystems® 3500 instrumentation and redid our mixture interpretation data as part of the validation. While federal and state Genetic Information Nondiscrimination Act (GINA) laws prohibit us from openly sharing this data on our website, we have attached the executive summary of our latest validation to show the extent of work and the quality of the reports we generate from validation work. ISPFS is committed to putting all validation studies on our website, and has started that process with new validations. While we cannot share the raw genetic information and data on our website, we would be happy to share that data with NIST under a cooperative agreement where the provisions of GINA are addressed.

Our validations have been reviewed by many experts, both inside and outside the forensic science community. We would welcome the opportunity to have NIST review our data and provide us with constructive feedback. We have full confidence that our data will show the robust nature of our validation studies and the great effort we take to ensure that the protocols and software are scientifically robust and reliable. We post all of our scientific methods on our public website and welcome NIST or any other entity to review our standard operating procedures at any time. Attached to this email we have included a recent example of a recent validation summary of the STRmix™ software to demonstrate the robust nature of the validations performed by our laboratory.

We join with Dr. Ray Wickenheiser in the following concerns shared with NIST:

1. The data sample utilized by NIST in generating this report is too restrictive and does not accurately reflect validation data used by forensic laboratories. NIST is only reviewing data that is publicly available. Most forensic laboratory validation data is not made public, as it contains staff, friends and family profiles, and individuals providing the samples who did not provide informed consent to permit their DNA profiles to be released into the public domain. Forensic laboratories operate in a secure environment where data must be safeguarded, which runs contrary to NIST's determination that only data published or posted publicly qualify for their foundation review.

NIST did not make a request to public laboratories to review their data. Much validation data is currently available for defense witnesses, laboratory auditors and assessors review at forensic laboratory premises and has been independently reviewed by these entities. Requiring data to be publicly available as a prerequisite to determining it is valid is an unprecedented requirement by NIST, which is not in place for many other scientific endeavors. Therefore, we feel NIST's requirement that only data that is in the public domain will be used to determine the scientific foundation for DNA mixture interpretation is too restrictive.

Recommendation: NIST visit forensic laboratories and forensic DNA mixture interpretation vendors and review validation data on site. As an alternative, they could make requests to review such data with appropriate confidentiality measures in place. Idaho State Police would welcome discussions with NIST about reviewing our validation studies and data with the appropriate provisions to comply with federal and state law.

2. NIST incorrectly contends that forensic laboratory data has not been independently reviewed. There are 60 publications including DNA mixture studies noted in the NIST report, including one with 1315 samples run by 31 different forensic laboratories. All forensic lab DNA validation studies are reviewed by independent external auditors within their 2-year external audit FBI Quality Assurance Standards requirements, and also by independent auditors from the national accrediting board 4-year audit cycle to meet ISO 17025:2017 standard requirements. Additionally, some states have statutorily created bodies responsible for oversight of forensic laboratory accreditation and approval of such laboratories use of new scientific methodologies and technologies. Many of these bodies have panels of forensic experts who have independently reviewed data and approved probabilistic genotyping of DNA mixtures as fit for purpose.

Recommendation: NIST include individuals with appropriate practical forensic experience to assist with independent review of validation studies and data and co-authorship of the report. Idaho State Police participated in the publication referenced above and would welcome discussions with NIST about reviewing our validation studies and data with the appropriate provisions to comply with federal and state law.

3. The draft report recommends an impracticable standard for validation studies to meet. NIST defines a novel concept of "factor space" including 26 factors impacting DNA mixtures, stating that the publicly available data did not cover this factor space. If every factor were comprehensively covered in a single mixture's "factor space," each of these 26 variables would need to be changed while holding the rest constant to determine the impact of a single variable on the mixture's behavior. Assuming 10 increments for each of the 26 variables, this would require 403 septillion factor comparisons (10 x 26 factorial). This huge number of samples is not practical nor feasible. The factor space model is therefore not appropriate for demonstrating that DNA mixture interpretation as practiced by forensic laboratories is fit for purpose.

Recommendation: NIST abandon the concept of factor space and develop a more practical measure of what is required to demonstrate fit for purpose and apply that measure to the review of on-site data with additional experts with forensic experience. NIST should then revisit their preliminary report, make the recommended changes herein and include forensic expertise in authorship of the next corrected version.

Finally, ISPFS would like to express concern in regard to key takeaway #4.7 in the draft report. Lines 769 through 773 suggest that applicable validation performance results would be helpful to include in the case file and report. As previously stated, including aspects of validation results in individual case files would be a violation of federal and state privacy laws. That aside, validations are already available to the appropriate legal entities and case agents through the discovery process. Additionally, ISPFS is making great effort to make as much validation information available on our website as possible and allowable per federal and state law. The addition of this information in the case file would simply add length and confusion for the average customer.

We appreciate the opportunity to provide comment on this report.

Matthew Gamette, M.S., C.P.M.
Laboratory System Director
Idaho State Police Forensic Services
[redacted]
matthew.gamette@isp.idaho.gov

PC63a

Internal Validation of STRmix™ V2.8 for the ISP Laboratory

(Fusion 5C™ 3500)

STRmix™ internal validation

This document describes the internal validation of STRmix™ V2.8 within the ISP Laboratory. STRmix™ has undergone extensive developmental validation. This involved, in part, the complete 'by hand' confirmation of the calculations used within the software. The results of developmental validation are detailed within the STRmix™ User's Manual. In addition, a summary of the developmental validation activities undertaken has been published [1].

Internal validation describes activities the ISP Laboratory has undertaken in-house before the implementation of STRmix™ v.2.8.0 into routine casework. This document follows the internal validation section of the SWGDAM Guidelines for the Validation of Probabilistic Genotyping Systems [2]. This included the use of known and non-probative evidence samples to investigate precision, sensitivity, and specificity. In addition, this document also describes the experiments undertaken to internally validate the use of PCR replicates (including multi-kit replicates), the variable number of contributors (varNOC), the user-informed mixture proportion priors, and Mix to Mix matching features of STRmix™ within the ISP Laboratory. Appendix 1 cross-references the sections of this report that discuss specific SWGDAM guidelines.

The DNA profiling data described in this report was generated at the ISP Laboratory following the casework protocols for Fusion 5C™ amplification and separation using 3500 CE instrumentation followed by analysis of the raw data performed with GeneMapper® *ID-X* V1.6. The results of all experiments related to the internal validation of STRmix™ within the ISP Laboratory are retained within the laboratory's quality system. This validation was undertaken with the assistance of the STRmix™ Scientific Support team of ESR in New Zealand.

STRmix™ parameters

Unless otherwise stated, the parameters described in the document 'Estimation of STRmix™ V2.8 parameters for the ISP Laboratory (Fusion 5C™ 3500)' were used for all internal validation checks presented in this report. The STRmix™ developers have optimised all other run parameters.

Section A: Single-source profiles

<u>Inspection of weights:</u>

This section covers the following recommendations:

> 4.1.5. Single-source specimens
>
> 4.2.1.2. For single-source specimens with high quality results, genotypes derived from non-probabilistic analyses of profiles above the stochastic threshold should be in complete concordance with the results of probabilistic methods

It is demonstrated within this section how the weights assigned by STRmix™ to different genotype combinations are appropriate. The weights are one of the primary outputs of the deconvolution process and should be intuitively correct, with the most supported genotypes being assigned the

highest weights. In contrast, genotypes that provide a poor explanation of the profile should be assigned relatively low weight (or no weight at all).

The addition of information to an analysis can aid in the ability to deconvolute a DNA profile. For example, use of PCR replicates and/or conditioning profiles may reduce ambiguity and increase weightings for individual genotype sets. It has previously been demonstrated that the use of more information improves the ability of STRmix™ to discriminate between true donors and non-contributors [3].

A series of single-source profiles of varying quality was prepared. PCR DNA template amount ranged from 1 ng to 7.8 pg. Allelic drop out was observed in profiles 62.5 pg and below.

The profiles were interpreted in STRmix™ and a likelihood ratio (*LR*) assigned to the known donor. The propositions considered were:

H_p: The DNA originates from the person of interest

H_d: The DNA originates from an unknown, unrelated individual

Sub-source *LR*s were assigned using the FBI Caucasian allele frequencies [4]. A correction for population sub-structure was made using a theta (θ) value of 0.01. A plot of log(*LR*) versus PCR DNA template amount is provided in Figure 1. The dashed line represents the log(*LR*) expected for a full, unambiguous single-source profile where the weight is 1 (i.e. 100%) for a single genotype at each locus and theta = 0. The *LR*s assigned by STRmix™ should never exceed this value.

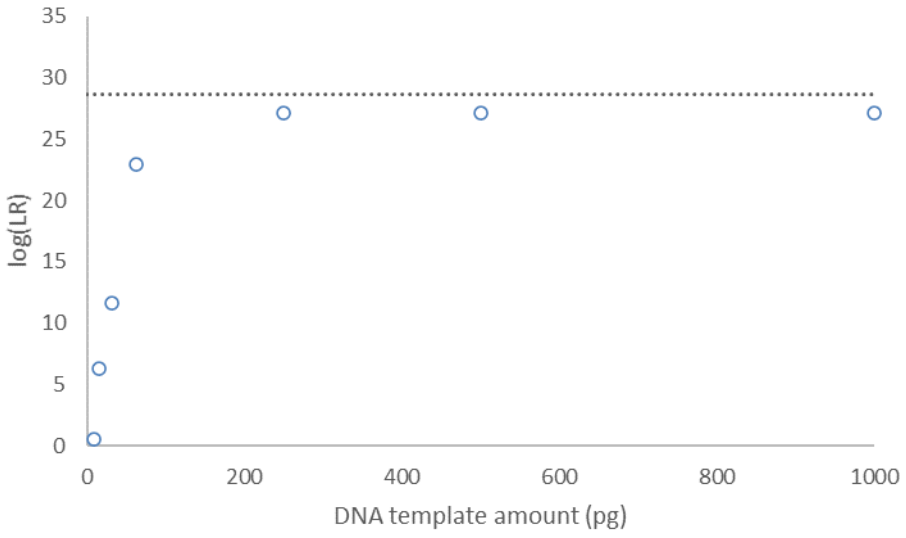


Figure 1: Plot of log(*LR*) versus PCR DNA template amount (pg). The dashed line represents the log(*LR*) expected for a full, unambiguous single-source profile from the known donor when theta = 0.

Inspection of Figure 1 shows that the log(*LR*) increased as DNA template amount increased. Template amounts of 250 pg and above resulted in unambiguous genotype weightings. This led to a log(*LR*) slightly under that expected for a full, unambiguous single-source profile from the known donor when theta =0, as expected. Weights for genotypes considering dropout (indicated as a Q

allele within the STRmix™ report) increased as template amount decreased. As peak height decreased, STRmix™ also accepted genotypes that include drop-in, albeit with quite low weight. Finally, the average posterior template amount reported within the STRmix™ report (measured in relative fluorescent units, rfu) was observed to decrease with decreasing PCR template amount. These are the expected results.

Reproduction of single-source *LR*:

There is a small sub-set of profiles where the 'answer' is known or can be estimated relatively easily given the models employed within STRmix™ [5]. These include unambiguous single-source profiles. The *LR* was calculated 'by hand' at each locus for such a single-source profile and the locus *LR*s compared with the corresponding STRmix™ results. The FBI Caucasian allele frequencies were used within these calculations. This was undertaken twice: once using an F_{ST} (or θ) value of 0 and once with F_{ST} = 0.01. Setting θ to zero returns the product rule where:

$2p_ip_j$ for heterozygous loci Equation (1)

p_i^2 for homozygous loci Equation (2)

When θ > 0, the Balding and Nichols formulae are applied [6]. These appear as equations 4.10 within the NRCII report [7]. For single-source profiles:

$$\frac{2\left[\theta+(1-\theta)p_i\right]\left[\theta+(1-\theta)p_j\right]}{(1+\theta)(1+2\theta)}$$ for heterozygous loci Equation (3)

$$\frac{\left[3\theta+(1-\theta)p_i\right]\left[2\theta+(1-\theta)p_i\right]}{(1+\theta)(1+2\theta)}$$ for homozygous loci Equation (4)

In the above formulae, p_i is the allele frequency for allele i, p_j is the allele frequency for allele j, and θ is the F_{ST} value.

The allele frequencies used within equations 1 through 4 are posterior mean frequencies. These are calculated using the following equation:

$$\frac{x_i+\frac{1}{k+1}}{N_a+1}$$ Equation (5)

Within Equation 5, x_i is the number of observations of allele i within the allele frequency database, k is the number of allele classes with non-zero observations at the locus in question, and N_a is the total number of alleles typed at that locus.

The *LR*s calculated 'by hand' and those calculated using STRmix™ were identical and are summarised in Table 1. As expected, the *LR*s decrease when a theta value of 0.01 is applied.

Table 1: Comparison of 'by hand' *LR*s (calculated using Microsoft Excel) with those calculated using STRmix™. *LR*s were calculated following interpretation of a full, unambiguous single-source profile. The FBI Caucasian allele frequencies were used with θ = 0 and θ = 0.01.

Locus	MS Excel θ=0	STRmix™ θ=0	MS Excel θ=0.01	STRmix™ θ=0.01
D3S1358	12.387	12.387	11.780	11.780
vWA	9.1275	9.1275	8.8106	8.8106
D16S539	8.9821	8.9821	8.6353	8.6353
CSF1PO	2.5056×10^{03}	2.5056×10^{03}	5.2029×10^{02}	5.2029×10^{02}
TPOX	3.5944	3.5944	3.5680	3.5680
D8S1179	7.3063	7.3063	7.1043	7.1043
D21S11	54.224	54.224	435.41	435.41
D18S51	27.485	27.485	24.978	24.978
D2S441	6.1798	6.1798	6.0565	6.0565
D19S433	5.1415	5.1415	5.0686	5.0686
TH01	7.3096	7.3096	7.1167	7.1167
FGA	63.964	63.964	52.990	52.990
D22S1045	4.3516	4.3516	4.3126	4.3126
D5S818	3.4946	3.4946	3.4841	3.4841
D13S317	32.310	32.310	27.338	27.338
D7S820	15.112	15.112	14.249	14.249
SE33	70.926	70.926	57.129	57.129
D10S1248	20.467	20.467	18.886	18.886
D1S1656	10.530	10.530	10.087	10.087
D12S391	47.850	47.850	41.142	41.142
D2S1338	36.076	36.076	32.127	32.127
Total	**4.0089×10^{28}**	**4.0089×10^{28}**	**1.2560×10^{27}**	**1.2560×10^{27}**

The results in Table 1 demonstrate that STRmix™ is giving the expected answer based on the population genetic model used.

Section B: Use of peak heights

This section covers the following recommendation:

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

STRmix™ uses a fully continuous model that considers peak height to inform the genotype combinations of contributors to DNA profiles. As template amount decreases, genotypes that include allelic dropout begin to be accepted by STRmix™. This leads to increased genotype uncertainty with a corresponding decrease in the *LR* for known donors. This relationship can be observed in Figure 1 and is the expected result. The effect of peak height on sensitivity and specificity is further examined in Section D below.

In this section the ability of STRmix™ to interpret profiles that include off-scale peaks is explored. It is not recommended that grossly saturated profiles are interpreted within STRmix™. Peak heights within such profiles are unlikely to have been accurately recorded by the CE camera, causing sub-optimal performance of the models within STRmix™. This typically results in the observation of

larger than expected stutter peak heights leading to elevated stutter peak height variance parameters. In extreme cases, the genotypes accepted by STRmix™ may not be intuitive.

Five over-amplified single-source profiles were prepared by the ISP Laboratory. The target PCR template amount was 4.0 ng. Peaks exceeding the ISP Laboratory's saturation threshold of 30,000 rfu were observed. Following analysis, each profile was interpreted within STRmix™ under the assumption of a single contributor. STRmix™ assigned a weight of 1.0 (100%) to the genotype of the known donor at each locus in four of the profiles. The remaining profile had a GR = NaN and unintuitive genotypes at several loci. The *LR* to the known contributor was lower than the *LR* obtained for a non-saturated profile. The posterior mean back stutter variance was also extremely elevated (data point at ~150 in Figure 2).

The posterior mean back stutter variance parameters for all saturated samples are plotted in Figure 2, overlaid on the back stutter prior distribution. These results indicate that STRmix™ can interpret saturated samples and diagnostics will indicate when the interpretation has not progressed as expected, however the routine interpretation of saturated samples using STRmix™ is not recommended.

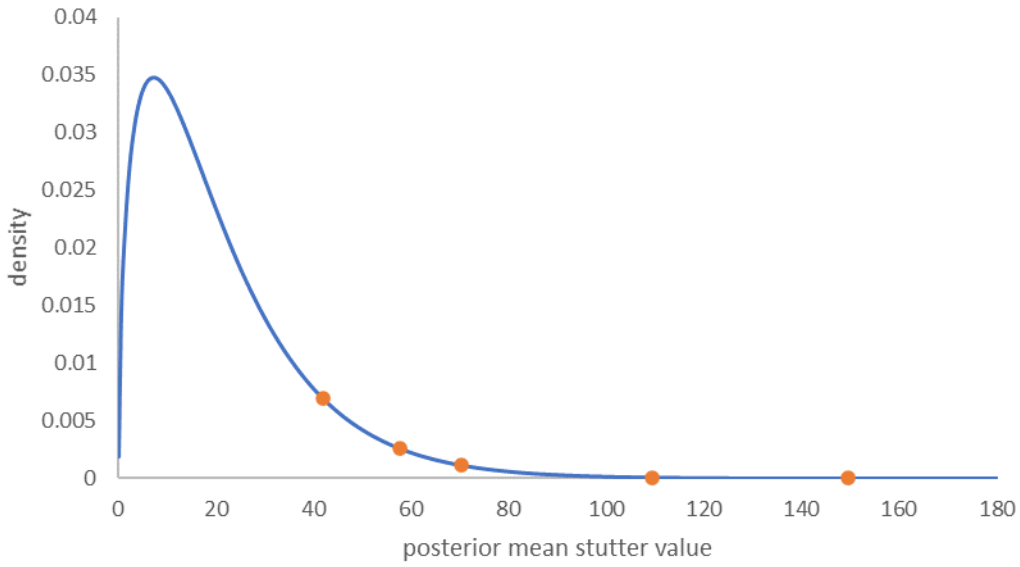


Figure 2: Posterior mean stutter values plotted on the back stutter prior distribution

Section C: Weights

This section covers the following recommendation:

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

The genotype weights are one of the primary outputs of STRmix™. They are determined using Markov chain Monte Carlo (MCMC) and feed into any subsequent *LR* calculations. The weights may be used as a diagnostic of the deconvolution process and should be intuitively correct. Genotypes

that explain the profile well should be assigned relatively high weight. In contrast, genotypes that offer a poor explanation should be assigned little or no weight.

A two-person mixture series was constructed by the ISP Laboratory in the following ratios: 25:1, 15:1, 9:1, 5:1, 3:1, 2:1, and 1:1. The total amount of DNA added to the PCR for each mixture sample was approximately 1 ng. The profiles were interpreted in STRmix™ assuming two contributors. *LR*s were assigned for each of the known donors under the following propositions:

> H_p: The DNA originates from the person of interest (known major or minor donor) and an unknown, unrelated individual
>
> H_d: The DNA originates from two unknown, unrelated individuals

The FBI Caucasian allele frequencies were used with θ = 0.01. The sub-source *LR*s assigned were recorded and are plotted in Figure 3 and Figure 4 below (in $\log_{10}$ format). Within these plots, the log(*LR*) expected for a full, unambiguous single-source profile where theta = 0 from each donor is plotted as a dashed line. The *LR*s assigned by STRmix™ should never exceed this value.

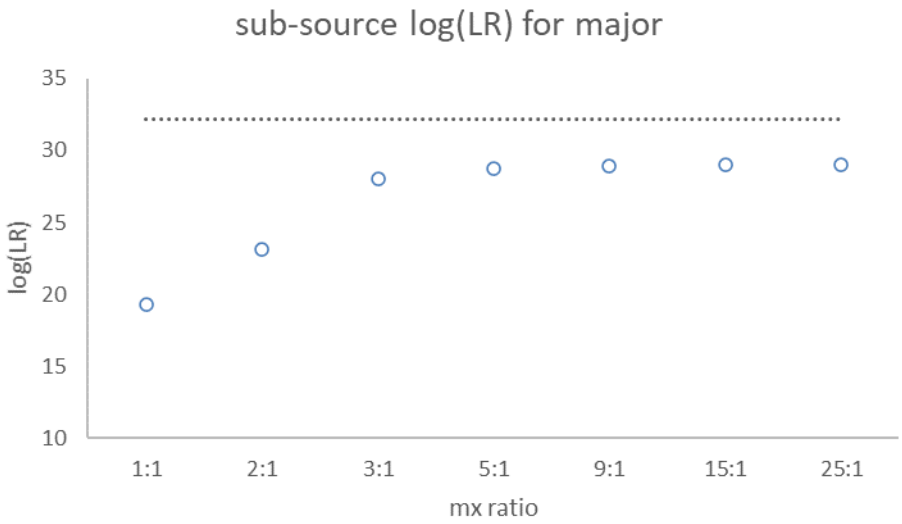


Figure 3: Log(*LR*)s assigned for the major contributor to a series of mixed DNA profiles originating from two contributors. The mixture ratios ranged from 25:1 to 1:1. The dashed line indicates the log(*LR*) assigned for a full, unambiguous single-source profile originating the major contributor when theta = 0.

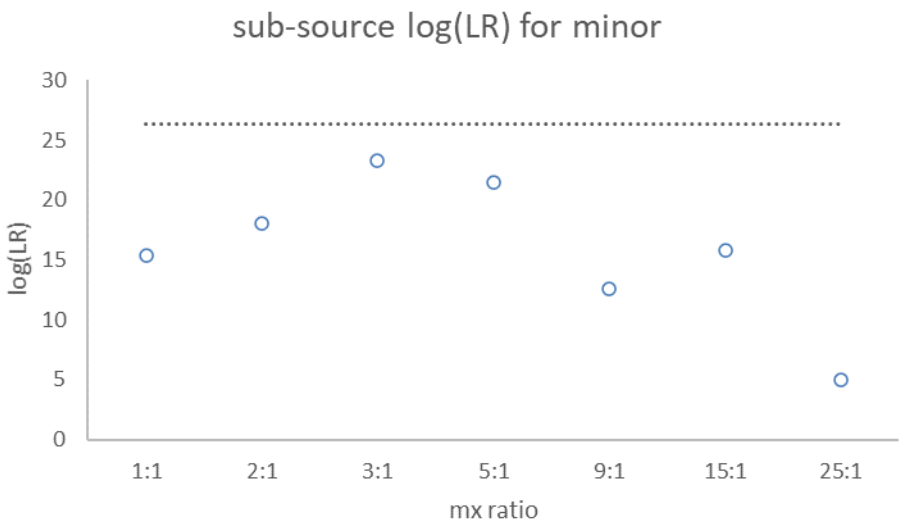


Figure 4: Log(*LR*)s assigned for the minor contributor to a series of mixed DNA profiles originating from two contributors. The mixture ratios ranged from 25:1 to 1:1. The dashed line indicates the log(*LR*) assigned for a full, unambiguous single-source profile originating from the minor contributor when theta = 0.

Inspection of Figure 3 and Figure 4 demonstrates that the *LR*s behaved as expected. Where the major contributor was well resolved, the *LR* assigned was almost equivalent to the single-source *LR* with no sub-population correction. Note that due to the use of a sub-source and theta correction, the *LR* assigned will never equal that of the single-source value, even where the major contributor is fully resolved. Where the donors contributed similar amounts of DNA (e.g. a 1:1 mixture), STRmix™ accepted multiple genotype combinations at all loci as expected. This reduction in genotype certainty resulted in a decrease to the *LR* for the major donor by several orders of magnitude. Regarding the minor contributor, the largest *LR* was assigned following deconvolution of the 3:1 mixture. As the amount of DNA from the minor contributor decreased, so too did the *LR* assigned, except for the 15:1 mixture vs the 9:1. In the 15:1 mixture there was more information available for the minor at some loci, (Penta E and vWA). Therefore, a higher LR relative to the 9:1 mixture is expected. Like the major donor, the *LR* for the minor contributor was also observed to decrease where both donors contributed similar amounts of DNA. In all cases, the *LR* assigned for the minor contributor was below the value assigned for a full, unambiguous single-source profile from the relevant donor.

Section D: Sensitivity and specificity of mixed DNA profiles

This section covers the following recommendations:

4.1.2. Hypothesis testing with contributors and non-contributors

4.1.6. Mixed specimens

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

4.1.6.2. Various total DNA template quantities

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

4.1.6.5. Sharing of alleles among contributors

4.1.7. Partial profiles, to include the following:

4.1.7.1. Allele and locus drop-out

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

A demonstration of sensitivity and specificity for a range of mixtures prepared by the ISP Laboratory was undertaken as per [3], using average peak height (APH) instead of input template.

With respect to interpretation methods, sensitivity is defined as the ability of the software to reliably resolve the DNA profile of the known contributor(s) to a DNA profile for a range of starting DNA templates. The log(*LR*) for known contributors (i.e. H_p true) should be high and should trend to 0 as less information is present within the profile. In this context, 'information' includes the amount of DNA from the contributor of interest, the use of conditioning profiles during interpretation (for example, the complainant's DNA on intimate samples), the use of PCR replicates, and decreased profile complexity. Specificity is defined as the ability of the software to reliably exclude non-contributors (i.e. H_d true) within a DNA profile for a range of starting DNA templates. The log(*LR*) should trend upwards to 0 as less information is present within the profile.

A series of mixed DNA profiles ranging from two to four contributors was prepared by the ISP Laboratory. These mixtures cover a broad range of template amounts and mixture proportions and are likely to be representative of DNA profiles recovered during casework analysis. The contributors include homozygous and heterozygous alleles and there is varying amounts of allele sharing across the different loci (recommendation 4.1.6.5). Given the template amounts, allele and/or locus dropout was expected to occur within the profiles containing lower DNA amounts (recommendation 4.1.7.1). The mixtures prepared are summarised in Table 2 below. Replicates were prepared for all samples. In total, 143 mixtures were prepared. Following amplification and CE, the profiles were analysed within GeneMapper® *ID-X* V1.6 using the ISP Laboratory's Fusion 5C 3500 casework analysis method.

Table 2: Summary of mixtures prepared by the ISP Laboratory to examine STRmix™ sensitivity and specificity.

Mixture type	Mixture ratio	DNA amount of lowest contributor (pg)
2-person	19:1, 10:1, 5:1, 3:1	100, 50, 25, 12.5, 6.25
	1:1	500, 300, 100, 50, 25
2-person degraded	25:1, 15:1, 9:1, 5:1, 3:1, 2:1, 1:1	1000, 500
3-person	10:5:1, 3:2:1	100, 50, 25, 12.5, 6.25
	1:1:1	500, 300, 100, 50, 25
4-person	4:3:2:1, 10:5:2:1, 1:1:1:1	500, 300, 100, 50, 25

Following analysis, each mixture was interpreted within STRmix™. The experimentally designed number of contributors (NOC) was assumed when setting up the interpretations. Likelihood ratios were assigned to true and non-contributors by searching each deconvolution against a database that contained the DNA profiles of the known donors (100 profiles) as well as 200 non-contributor profiles. The non-contributor profiles were simulated from the FBI Caucasian allele frequencies. An *LR* was assigned for each database individual considering the following propositions:

H_p: The DNA originates from the database individual and N-1 unknown, unrelated individuals

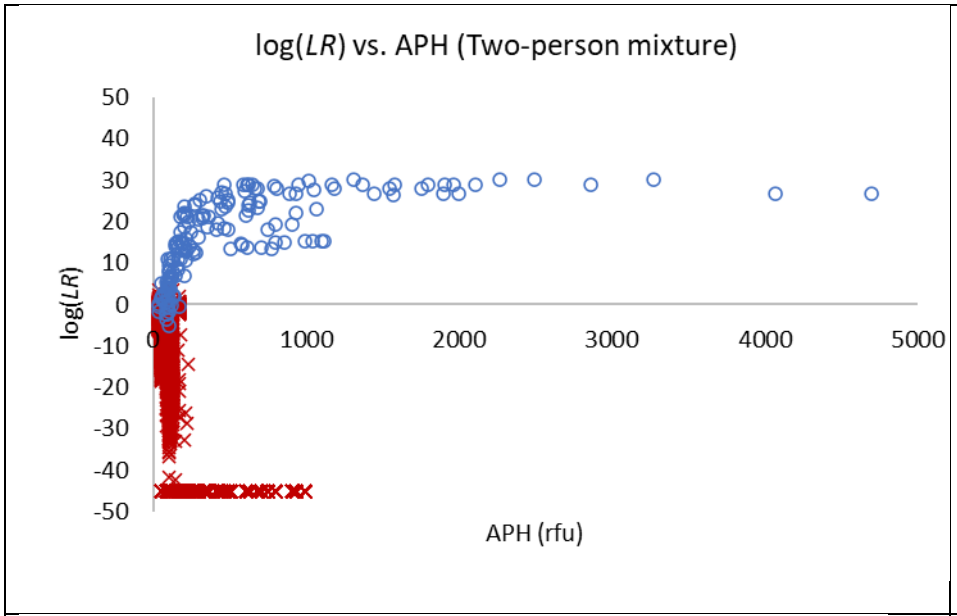
H_d: The DNA originates from N unknown, unrelated individuals

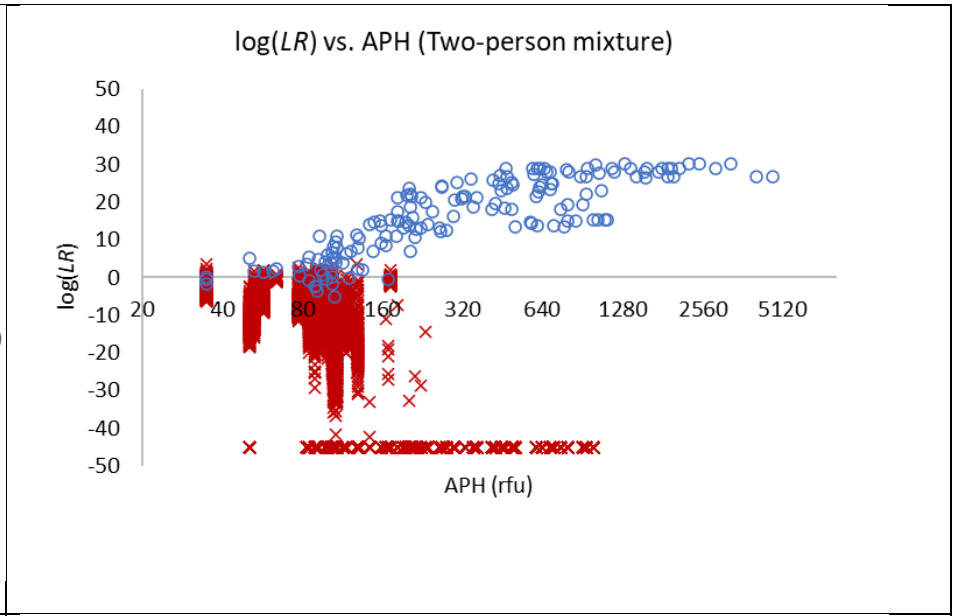
Where N is the experimentally designed NOC. *LR*s were assigned using the FBI Caucasian allele frequencies with $\theta = 0.01$ and the sub-source *LR* used as the point of comparison.

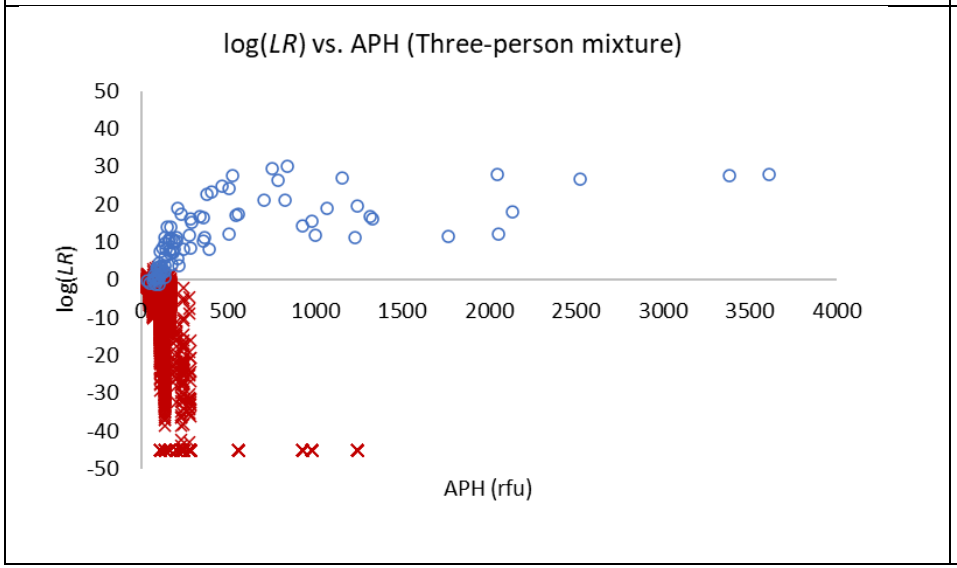
Plots of log(*LR*) versus per contributor average peak height (APH) for the two-, three-, and four-person mixtures are given in Figure 5. Each plot has been reproduced with the scale of the x-axis adjusted to better display data points for low APH. These plots follow the approach used in [3] with the exception that the log(*LR*)s assigned have been plotted against per contributor APH rather than per contributor template amount. This was done as APH is more readily estimated from forensic casework DNA profiles. APH was calculated using unmasked, unshared, and non-stutter-affected alleles for each contributor in the mixture. Where no such peaks were present, a value of half the AT was instead used. For non-contributors, the log(*LR*) was plotted against the lowest APH across all known donors to a mixture. Exclusions (*LR* = 0) are plotted as log(*LR*) = -45.

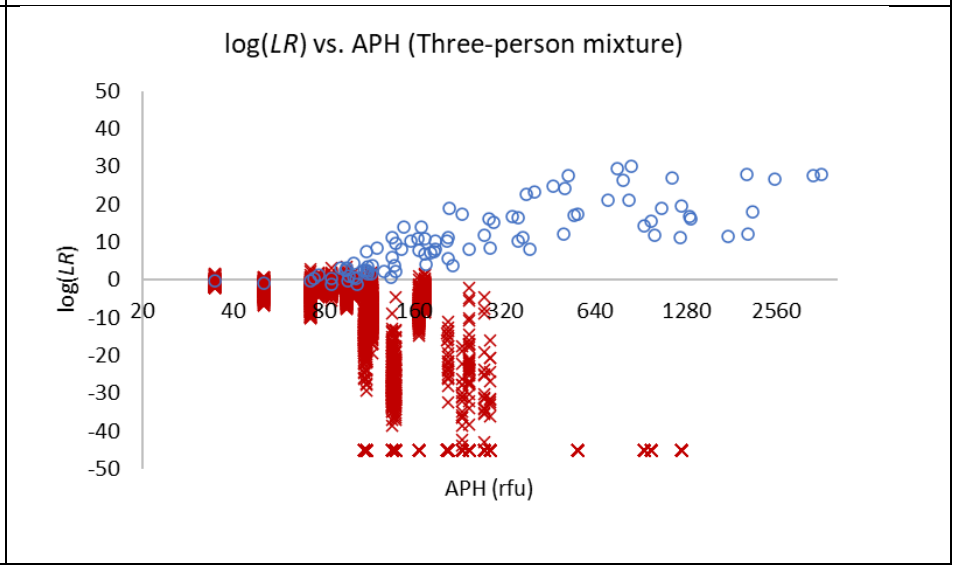
Inspection of the plots in Figure 5 shows that as APH increases, the *LR*s assigned for known donors and non-contributors diverge. As template decreases, *LR*s for known donors and non-contributors trend to log(*LR*) = 0. A log(*LR*) of zero may be considered to be 'uninformative', or 'neutral'. The plots in Figure 5 demonstrate that STRmix™ was able to reliably distinguish between true donors and non-contributors, even where per contributor APH was relatively low.

The plots in Figure 5 can help inform the limits of STRmix™, particularly the lower limit of DNA where an H_p true hypothesis still results in a log(*LR*) greater than 0 and the limit where false positives may arise (a log(*LR*) greater than 0 where H_d is true).


log(LR) vs. APH (Two-person mixture)
log(LR)
APH (rfu)
50
40
30
20
10
0
-10
-20
-30
-40
-50
0
1000
2000
3000
4000
5000


log(LR) vs. APH (Two-person mixture)
log(LR)
APH (rfu)
50
40
30
20
10
0
-10
-20
-30
-40
-50
20
40
80
160
320
640
1280
2560
5120


log(LR) vs. APH (Three-person mixture)
log(LR)
APH (rfu)
50
40
30
20
10
0
-10
-20
-30
-40
-50
0
500
1000
1500
2000
2500
3000
3500
4000


log(LR) vs. APH (Three-person mixture)
log(LR)
APH (rfu)
50
40
30
20
10
0
-10
-20
-30
-40
-50
20
40
80
160
320
640
1280
2560

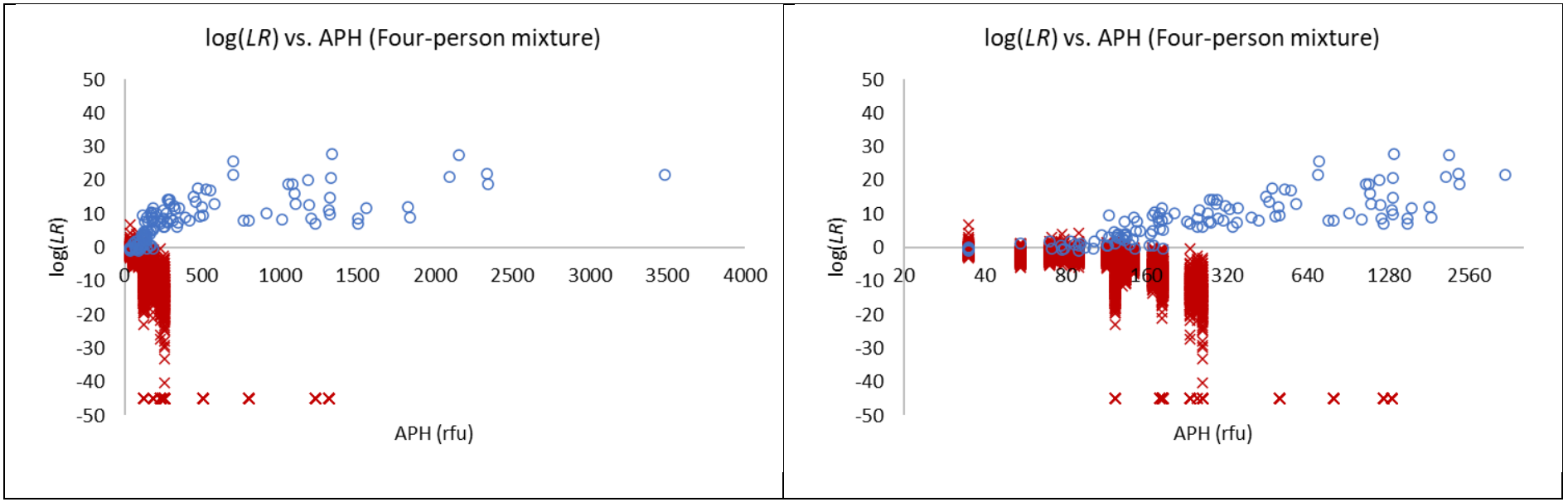


Figure 5: Log(*LR*) versus average peak height (APH) for known donors (plotted using circles) and non-contributors (plotted using crosses). Separate plots are provided for the two-, three-, and four-person mixtures examined. Each plot has been reproduced with the scale of the x-axis adjusted to better display data points for low-template contributors.

Review of specific results

The largest false inclusion of a non-contributor was approximately log(*LR*) 18.25 (*LR* ≈ 1.76 x 10^{18}) at APH 171 rfu for sample 3.1 C3 (3). The contributor resulting in this *LR* was K74. A review of this profile (labelled as a two-person mixture with a mixture ratio of 3:1 with 25 pg of DNA for the minor contributor) showed some oddities. The electropherogram (epg) showed that the experimentally designed known contributors (K32 and K90) were likely not present (also indicated by the STRmix™ log(*LR*)s of -1.09 and -2.31). A review of the electropherogram of the replicate PCR for this sample showed that the replicates were significantly different, as shown in Figure 6. Further investigation of the sample revealed that the sample had been renamed and the sample that was injected was sample 5:1 C4 (Figure 7), that originates from contributors K74 and K98. Comparison to known reference profile K74 supports this. Hence on this occasion, the large *LR* to K74 is expected. The *LR*s to K74, K32, K90 and K98 for this sample were subsequently removed from the sensitivity and specificity plots shown in Figure 5. The original results can be discovered in the supplementary materials if required.

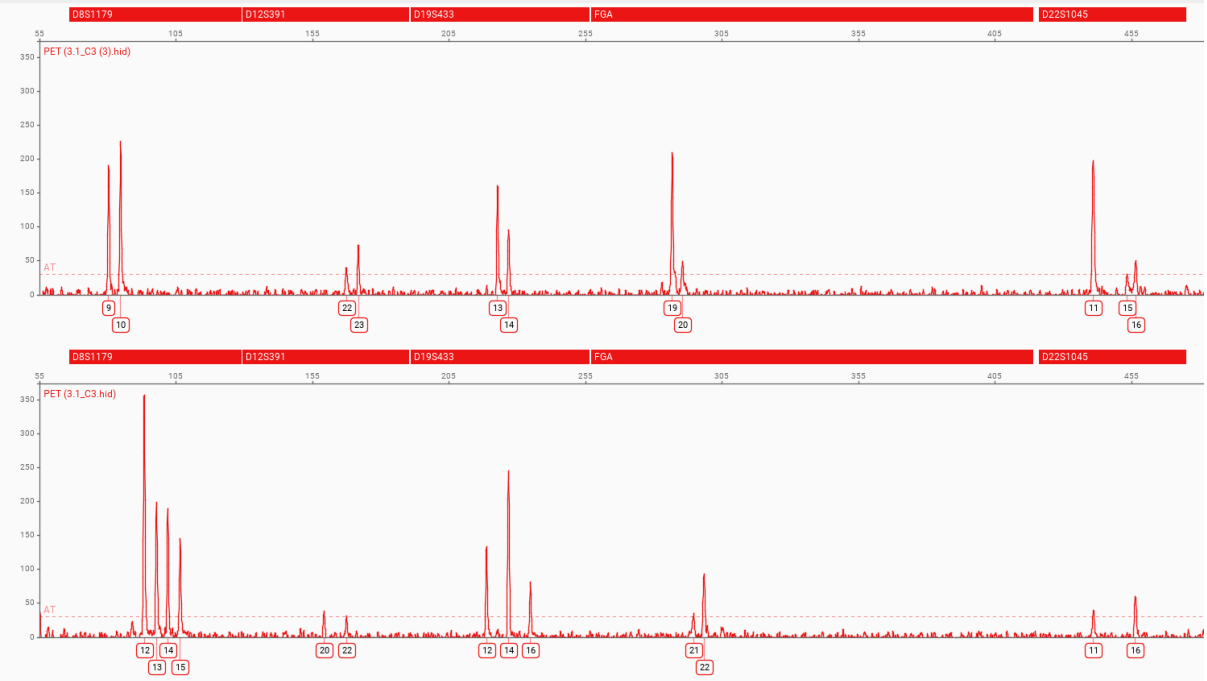


Figure 6: The red channel of two replicate profiles labelled'3:1 C3'. The problematic epg is displayed in the top pane.

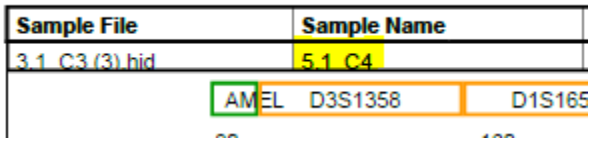


Figure 7: Screen grab from GeneMapper showing the sample 5.1 C4 was renamed to 3.1 C3 (3)

The second largest false inclusion of a non-contributor was approximately log(*LR*) 6.64 (*LR* ≈ 4 million). This result was observed following interpretation of a low template four-person mixture (4:3:2:1 C5 (2)). A review of the STRmix™ interpretation report showed the contributor position giving the highest *LR* considering K50 (the non-donor) was contributor position one, that had been assigned roughly 101 rfu template in the STRmix™ deconvolution. The experimentally designed

major contributor to this mixture was K55 and the *LR* when considering them as a contributor was log(LR) 9.44.

Through a comparison of K50 and K55 it was noted that there was about 73% similarity between the profiles with at least one allele being in common at each locus. This, combined with an allowance for dropout due to the low level of the profile, resulted in an inclusionary *LR* for the non-contributor. This is not a limitation of the software: STRmix™ is performing as expected. Rather, this result is better classified as an adventitious match arising from the fact that the non-contributor has many alleles in common with the known donors used to construct the mixture.

No false exclusions of true contributors (*LR* = 0) were observed in the ISP data. When false exclusions do occur, previous studies [8] have identified common causes as under-assignment of the number of contributors to the mixture and missing peak height data due to poor 1 bp CE resolution.

Review of Run Diagnostics

STRmix™ includes a number of diagnostics within its reports. These have been deliberately included to assist the user when evaluating the reliability of an interpretation. These may be conveniently categorised into 'primary' and 'secondary' diagnostics. Primary diagnostics include the mixture proportions, genotype weights, and locus *LR*s. Secondary diagnostics include the average log(likelihood), the Gelman-Rubin convergence diagnostic, and the posterior mean variance parameters. In instances where non-intuitive primary diagnostics are observed, the STRmix™ results should be closely scrutinised however elevated secondary diagnostics do not necessarily invalidate an interpretation. Provided that the primary diagnostics are intuitive, the results are likely still reliable. The secondary diagnostics reported by STRmix™ following interpretation of the mixtures described in Table 2 were examined and are discussed further in Appendix 2.

Section E: Alternative propositions

This section covers the following recommendation:

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

A selection of the profiles examined in Section D, representing a range of template, mixture proportion and complexity, were re-interpreted in STRmix™ with alternative propositions. Two series of propositions were tested. The first investigated the effect on the *LR* of conditioning on a known contributor and the second looked at the effect of including more than one POI under the prosecution proposition while retaining all unknown donors under the defence proposition.

Test 1:

Seven interpretations were repeated conditioning on a major known donor then *LR*s assigned for the remaining known donor(s) as well as 200+ non-contributors. The propositions considered were:

H_p: The DNA originates from the known (conditioned) individual, the database individual, and *N*-2 unknown individuals

H_d: The DNA originates from the known (conditioned) individual and *N*-1 unknown individuals

Where *N* is the experimentally designed number of contributors. *LR*s were assigned in the same manner as previously described then compared against the corresponding values from the unconditioned interpretations performed in Section D above. A plot of the log(*LR*) assigned for conditioned versus unconditioned interpretations is provided in Figure 8.

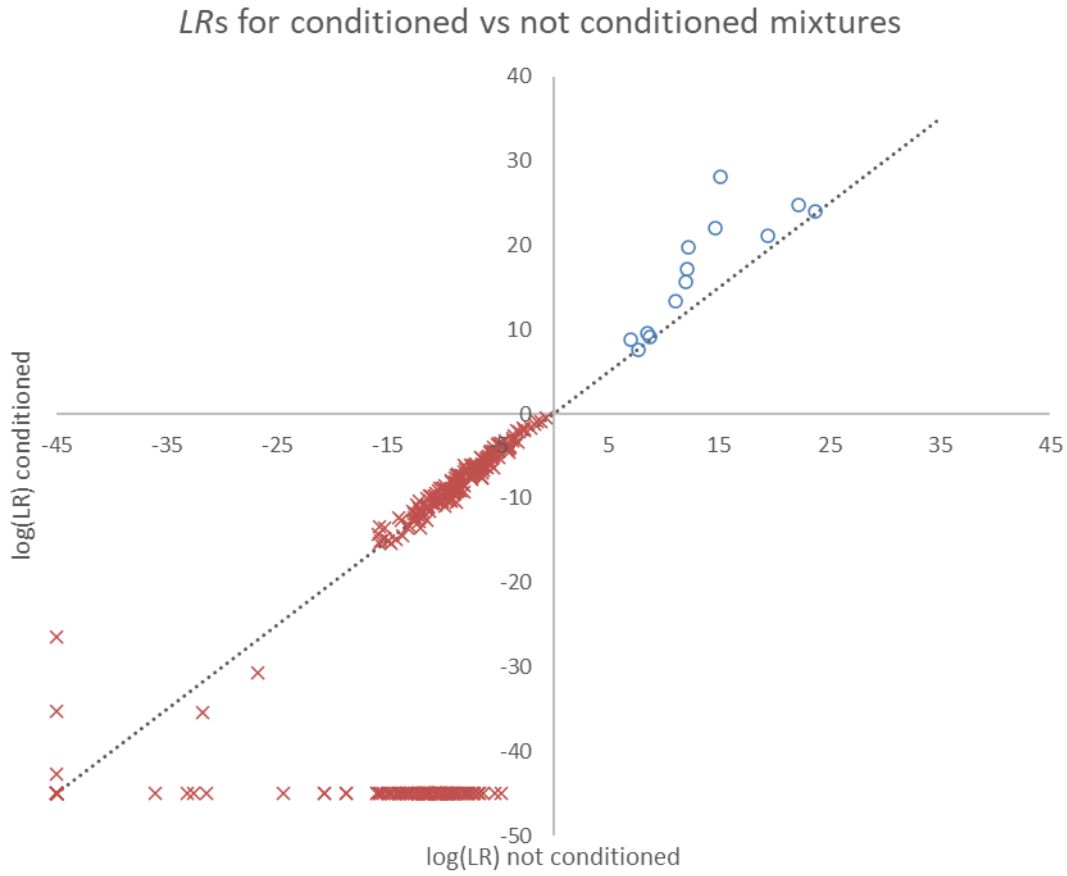


Figure 8: Comparison of log(*LR*) assigned following interpretation with and without use of a conditioning profile. Log(*LR*)s assigned for known donors are plotted using circles whilst those assigned for non-contributors are plotted using crosses. Exclusions (*LR* = 0) have been plotted as log(*LR*) = -45. A dashed line at $x = y$ has been added to assist with interpretation.

Values above the line at $x = y$ indicate the *LR* increased when a conditioning profile was used during the interpretation. Values below this line indicate the *LR* decreased. In general, when a conditioning profile was used true donors gave larger *LR*s supporting inclusion whilst non-contributors gave *LR*s with increasing support for exclusion or moved to outright exclusion. This demonstrates that the addition of more relevant information (such as the use of known donor

profiles) improves the ability of STRmix™ to discriminate between true and false donors. This is the expected result and is in line with [3].

Test 2:

The selection of samples identified above were investigated using the *LR* from previous function in STRmix™ and the interpretations from Section D. Proposition sets were built by sequentially adding one more POI to the prosecution proposition whilst retaining the number of unknowns under H_d. For example, the propositions for a three-person mixture are demonstrated below:

Set 1: H_p: The DNA originates from POI1, and two unknown individuals

H_d: The DNA originates from three unknown individuals

Set 2: H_p: The DNA originates from POI1 and POI2 and one unknown individual

H_d: The DNA originates from three unknown individuals

Set 3: H_p: The DNA originates from POI1, POI2 and POI3

H_d: The DNA originates from three unknown individuals

Propositions that include more than one person of interest under H_p and the maximum number of unknowns under H_d may be referred to as compound proposition sets.

As in Section D a theta value of 0.01 was applied and the FBI extended Caucasian allele frequencies were used. The resulting *LR*s are shown in Figure 9. These plots are designed to demonstrate the proportion that each contributor adds to the *LR*. For example, for mixture 4.3.2.1 the log(*LR*) considering POI1 and POI2 as co-contributors is ~40. For the same mixture, the log(*LR*) considering only POI1 is ~19.

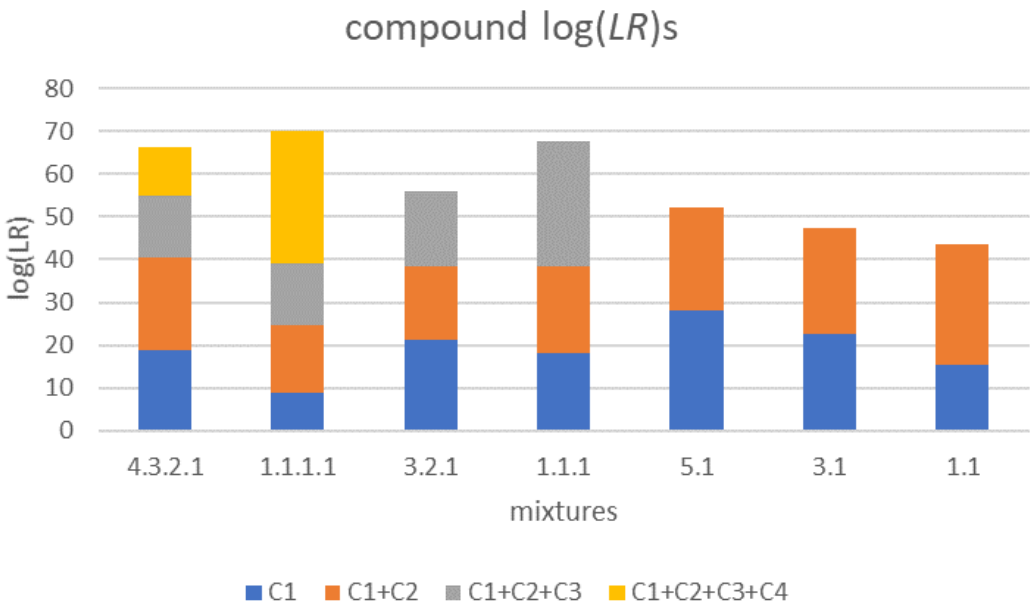


Figure 9: Compound log(*LR*)s showing the increase in LR as each co-contributor is added to H_p

Section F: Assigning number of contributors

This section covers the following recommendation:

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

The true number of contributors to a casework profile is always unknown. Analysts may add one or more additional contributors in the presence of an artefact peak, inflated stutter, or peak height imbalance due to stochastic effects. The assumption of one fewer contributor than actually present may be made when contributors are at very low levels and dropping out or when interpreting mixtures of individuals with similar DNA profiles (e.g. mixtures of DNA from closely related individuals). The risk of under-assigning the number of contributors to a mixture also increases as profile complexity increases [9, 10].

The effect of under- and over-assigning the number of contributors within STRmix™ has previously been investigated [11, 12]. The inclusion of an additional contributor beyond that present in the profile had the effect of lowering the *LR* for known low level trace contributors within the profile. STRmix™ was observed to add the additional (unseen) profile at trace levels which interacted with the known trace contribution, diffusing the genotype weights and lowering the *LR*. There was no significant effect on the *LR* of the major or minor contributor within the profiles. Under-assigning the number of contributors was occasionally observed to result in the false exclusion of known donors to a profile. Generally, the weakest/smallest contributor was excluded, with minimal effect observed on the *LR*s of the stronger contributors.

The effect of variation to the number of contributors assumed during a STRmix™ interpretation was examined within the present study. This was undertaken using profiles from Section D where the

apparent NOC differed from the experimentally-designed value. The results of the testing performed are described below.

Subtraction of one contributor

Eighty profiles from Section D were able to run in STRmix™ as *N*-1 contributors, where *N* is the experimentally-designed NOC. The remaining profiles had too many alleles at any one locus for the interpretation to progress as *N*-1 contributors. *LR*s were assigned for the known donors and non-contributors as in Section D. The resulting *LRs* (*N*-1) were compared with the equivalent comparisons performed within Section D (*N*). The *LR*s assigned are plotted in Figure 10 (in $\log_{10}$ format). Data points above the $x = y$ line represent an increase in log(*LR*) when assuming *N*-1 contributors. Data points below this line represent a decrease in log(*LR*) when assuming *N*-1 contributors.

From Figure 10 it can be seen that the log(*LR*)s assigned for many known donors remained relatively unchanged (note that some variation about the $x = y$ line is expected due to the MCMC process). There were however some known donors who either gave exclusionary *LR*s or were excluded under the assumption of *N*-1 contributors. This is the expected result given the interpretation has been run under the assumption of *N*-1 contributors. There were some large *LR*s that decreased significantly under the assumption of N-1 contributors. The largest differences come from 10:5:1 C3 mixtures where there are low peak heights across the profile (i.e. for all contributors). Under the assumption of N contributors, multiple combinations considering drop-out are accepted (which reflects what is seen in the electropherogram – that peaks from all contributors have dropped out given we know the ground truth samples). Under the assumption of N-1 contributors, in order for the profile to be explained, there is limited potential for any alleles to have dropped out. The result is that the deconvolution assuming N-1 contributors does not reflect the true contributor's profiles as well as the deconvolution assuming N contributors (i.e the deconvolution assuming N-1 contributors has lower weights for the genotypes of the true contributors than the deconvolution considering N contributors). This is also an expected result.

The true contributors can still be included, but with lower weights than considering the evidence under N contributors.

Log(*LR*)s assigned for non-contributors typically gave increased support for exclusion under the assumption of *N*-1 contributors. These results are in line with previous studies [12]. A review of the run diagnostics showed that some samples gave very low log(likelihoods) (the lowest value observed was -17) coupled with high posterior mean values for allele variance. These two diagnostics together can indicate that the assumed *N* is under assigned. As discussed with Section D data, secondary diagnostics should be evaluated relative to the profile being considered and it is recommended where there is uncertainty in the number of contributors that biological approaches to address this uncertainty such as reamplification is undertaken.

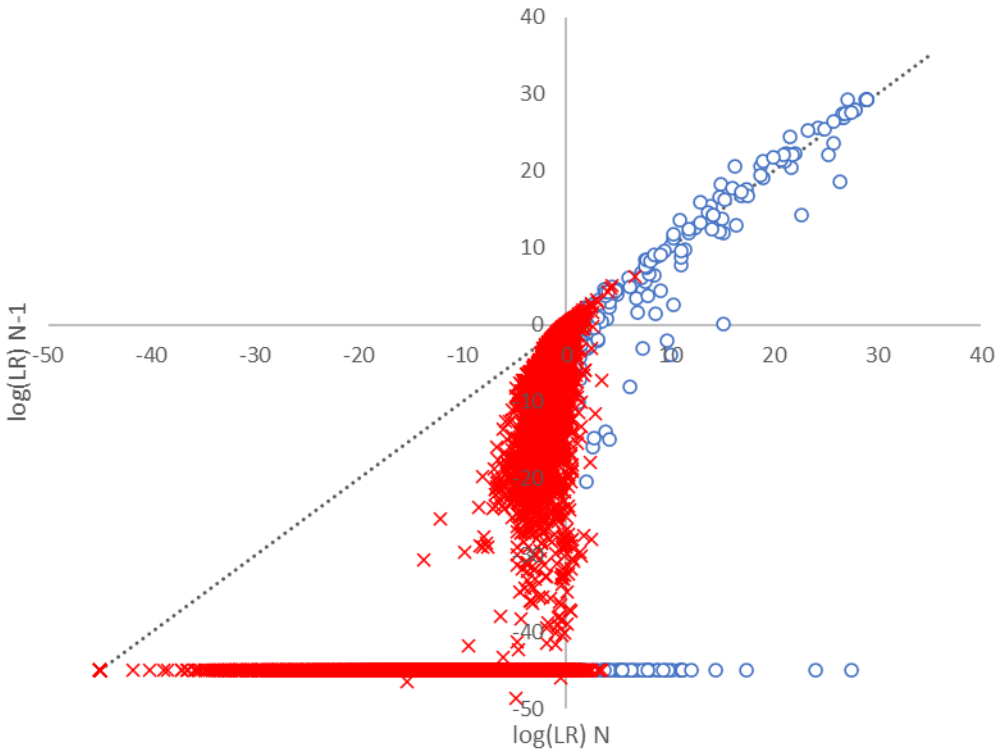


Figure 10: Comparison of log(*LR*)s assigned when assuming *N*-1 contributors versus *N* contributors, where *N* is the experimentally-designed number of contributors. Log(*LR*)s assigned for known donors have been plotted using circles whilst those assigned for non-contributors have been plotted using crosses. Exclusions (*LR* = 0) have been plotted as log(*LR*) = -45. A dashed line at $x = y$ has been added to assist with interpretation.

Addition of one contributor

A selection of profiles from Section D were re-assigned as originating from *N*+1 contributors, where *N* is the experimentally-designed NOC. These profiles were re-interpreted within STRmix™. *LR*s were assigned for the known donors and non-contributors as before; these were then compared with the equivalent comparisons performed within Section D. The *LR*s assigned are plotted in Figure 11 (in $\log_{10}$ format).

From Figure 11 it can be seen that the log(*LR*)s assigned for most known donors remained relatively unchanged. A few H_p true *LR*s were observed to decrease by several orders of magnitude. The data point at log(LR) ~24 (*N*) and log(LR) ~18 (*N*+1) is considering the known minor contributor (K55) to mixture 10.1 C2. The data point at log(LR) ~14 (*N*) and log(LR) ~9 (*N*+1) is considering the known minor contributor (K55) to mixture 10.1 C3. This is in line with other studies [11, 12] which demonstrated that the assumption of *N*+1 contributors tends to have the biggest impact on the *LR*s of the weakest/smallest contributor(s) to a mixture.

There were also a few H_p true *LR*s that increased by several orders of magnitude. The datapoint at log(*LR*) ~21 (*N*) and log(*LR*) ~26 (*N*+1) was considering the minor component in mixture 19.1 C1. An investigation into why the *LR* increased when considering *N*+1 showed that the driver was the locus D8S1179. In the evidence there is an 11 peak which cannot be attributed to either of the donors and is likely a double back stutter from the major contributor (Figure 12). Double back stutter is not

modelled in this kit and therefore for the true donor's genotype to be accepted by STRmix™ the peak must be modelled as drop-in (when considering the evidence originates from two contributors). This genotype is accepted with a low weight (Figure 13). When the evidence is interpreted considering it originated from three contributors, this peak can be attributed to the third contributor and therefore, the genotype of the true contributor is accepted with a greater weight in this deconvolution.

The data point at log(*LR*) ~13 (*N*) and log(*LR*) ~17 (*N*+1) is for mixture 19.1 C1(2), the replicate sample investigated above. The same artefact is present at D8S1179 and the same effect on the *LR* is observed.

The log(*LR*)s assigned for non-contributors generally increased under the assumption of *N*+1 contributors (observed in Figure 11 as data points above the line at $x = y$). Aside from the highest LR discussed in Section D, the largest log(*LR*) for a non-contributor was 4.745 assigned considering contributor K64 and sample 15.1 deg 1ng. The log(*LR*)s of the known donors to this sample were 28.814 (D1Fdeg) and 14.665 (K2). On investigation it was found that the highest *LR* considering K64 was obtained when they were compared with contributor position two – the position that K2 aligns with. A comparison of reference profiles of K2 and K64 showed that they share 70% of allelic calls. This combined with the low level nature of the second contributor position (the acceptance of QQ and Q genotypes/alleles) allowed for the adventitious match. Of note is that the DNA evidence provides slight support for the proposition that the DNA mixture originates from a sibling of the POI (vs the POI) (Figure 14).

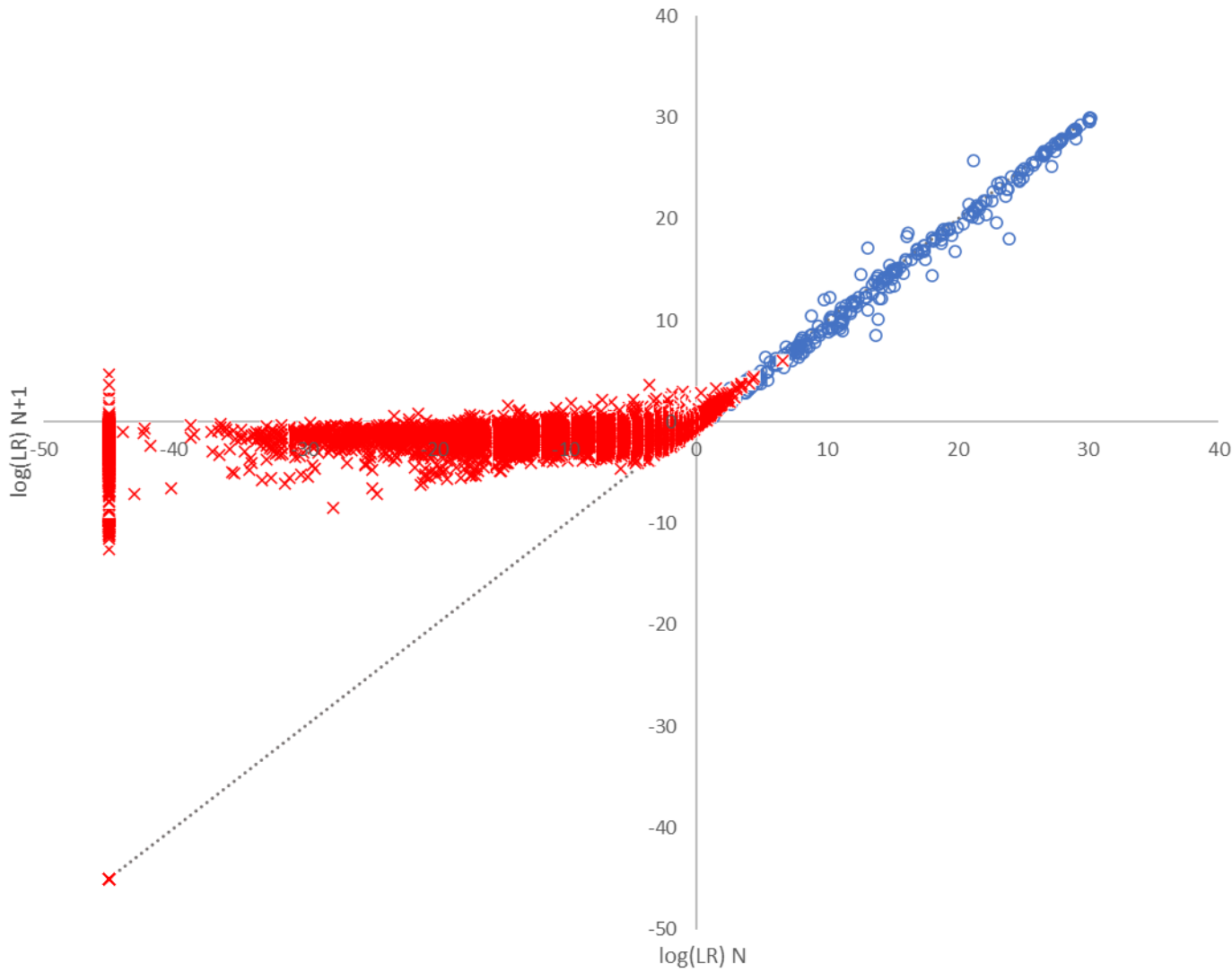


Figure 11: Comparison of log(*LR*)s assigned when assuming *N*+1 contributors versus *N* contributors, where *N* is the experimentally-designed number of contributors. Log(*LR*)s assigned for known donors have been plotted using circles whilst those assigned for non-contributors have been plotted using crosses. Exclusions (*LR* = 0) have been plotted as log(*LR*) = -45.

D8S1179	11	71	89
	12	1177	93
	13	15264	98
	14	654	102

Figure 12: Evidence input for sample 19.1 C1

D8S1179	13, 13	12, 14		9.6726E-1
	13, 13	12, 13		3.2462E-2
	13, 13	12, 12		1.8609E-4
	13, 13	11, 14		6.8990E-5
	13, 13	13, 14	11	1.4855E-5
	13, 13	14, 14	11	8.7015E-6

Figure 13: The deconvolution of 19.1 C1 at D8S1179 considering the evidence originates from two contributors. The genotype set of the known contributors is indicated by the red box.

SUMMARY OF LR

Sub-source LR. No HPD calculated.

LR	FBI_EXTENDED_CAUC
PROPORTION	1
Children per family	0
Population size	0
Relation of unknown in Hd to POI	
Unrelated	5.5594E4
Sibling	1.6741E-1
Parent/Child	1.9691E0
Half Sibling	8.0324E1
Grandparent/Grandchild	8.0324E1
Uncle or Aunt/Niece or Nephew	8.0324E1
Cousin	1.4903E3

Figure 14: Summary of LR table from the STRmix PDF report for 15.1 deg 1ng considering non-contributor K64

Section G: Drop-in

This section covers the following recommendation:

4.1.8. Allele drop-in

Proxy drop-in parameters have been assigned for the ISP Laboratory, refer to Table 3 (below) reproduced from the document 'Estimation of STRmix™ V2.8 parameters for the ISP Laboratory (Fusion 5C 3500)'.

Table 3: Drop-in parameters for the ISP laboratory, Fusion 5C 3500 DNA profiling kit.

Drop-in cap	250 rfu
Drop-in rate parameter	0.0007
Drop-in parameters α, β	0,0 (Uniform)

To examine the performance of these settings four experiments were undertaken. Specifically, modifications were made to the input file of a single-source profile that had previously been interpreted in Section A using STRmix™(DNA1_500pg.hid). The updated input files were used to reinterpret the profile in STRmix™ assuming the profile originated from a single source. For reference, the original input file is provided in Table 4 below (D3S1358 and CSF1PO only).

Table 4: STRmix™ input file for a single-source profile prior to modification (D3S1358 and CSF1PO only).

Locus	Allele	Height	Size
D3S1358	14	152	116
D3S1358	15	2037	120.33
D3S1358	17	160	129.01
D3S1358	18	1676	133.17
CSF1PO	8	85	326.98
CSF1PO	9	2326	331.04

First, a realistically sized drop-in peak was added at D3S1358. Specifically, a 12 peak with height of 75 rfu was added to the input file. The known donor is heterozygous at this locus with a genotype of 15,18. As expected, after re-interpretation STRmix™ modelled the 12 peak as drop-in given that two allelic peaks with relatively large peak height were detected at the locus (Figure 15). In this instance, STRmix™ was able to treat the 12 peak as drop-in given that its height is below the ISP Laboratory drop-in cap of 250 rfu. The 12 peak was listed as drop-in within the weights section of the STRmix™ output (Figure 15). The resulting *LR* was identical to that produced for the unedited profile (Table 5).

WEIGHTS

LOCUS	CONTRIBUTORS	DROP-IN ALLELES	WEIGHT
	1 (100%)		(HIGHLIGHT ≥ 0.99)
D3S1358	15, 18	12	1

Figure 15: Image of the STRmix PDF report showing that the 12 peak was accepted as drop-in

Table 5: *LR*s considering evidence with drop-in below the drop-in cap and without drop-in

Sub-source *LR* without drop-in peak	Sub-source *LR* with drop-in peak (75 rfu) at D3S1358	Sub-source *LR* with drop-in peak (75 rfu) at CSF1PO
1.2560×10^{27}	1.2560×10^{27}	1.2560×10^{27}

Next, the height of the 12 peak was increased to 251 rfu. STRmix™ was unable to progress the re-interpretation under the assumption of one contributor given that there were now three peaks at D3S1358 that could not be explained as either stutter or drop-in and therefore must have some allelic contribution. Peaks detected at a peak height equal to or above the drop-in cap will not be proposed as a drop-in event within STRmix™. The error message displayed by STRmix™ in this instance is provided in Figure 16 below.

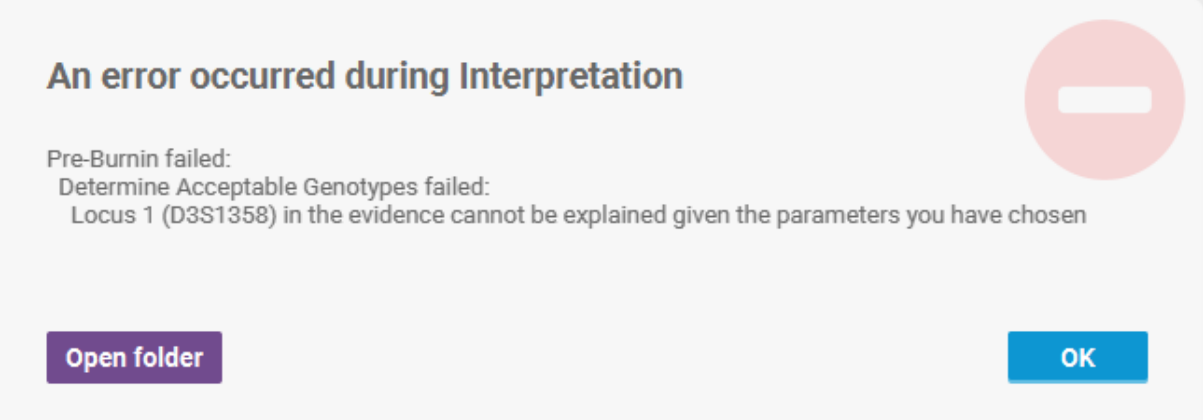


Figure 16: Error message displayed by STRmix™ when attempting to interpret a single-source profile following addition of a 12 peak at D3S1358 with peak height of 251 rfu. The profile could no longer be explained as originating from a single contributor.

The original input file was then edited at CSF1PO. The known donor is homozygous at this locus with a genotype of 9,9. Specifically, a 12 peak with height of 75 rfu was added. After re-interpretation, STRmix™ modelled the 12 peak as drop-in, assigning all weight to a genotype of 9,9 (Figure 17). The resulting *LR* was identical to that produced for the unedited profile (Table 5).

WEIGHTS

LOCUS	CONTRIBUTORS	DROP-IN ALLELES	WEIGHT
	1 (100%)		(HIGHLIGHT ≥ 0.99)
D3S1358	15, 18		1
D1S1656	11, 17.3		1
D2S441	11, 14		1
D10S1248	14, 15		1
D13S317	10, 11		1
Penta E	12, 16		1
D16S539	12, 13		1
D18S51	13, 17		1
D2S1338	16, 23		1
CSF1PO	9, 9	12	1

Figure 17: Image of the STRmix PDF report showing that the 12 peak was accepted as drop-in

The height of the 12 peak was then increased to 251 rfu and the profile re-interpreted. Under the assumption that the profile originates from a single donor STRmix™ was forced to treat the 12 peak as allelic and pair it with the 9 allele, leading to the exclusion of the known donor (Figure 18). Given the drop-in parameters used, this is the expected result. Review of the diagnostics included within the STRmix™ report should indicate to the user that the results require further scrutiny. In this instance, an *LR* of zero was assigned at a single locus. Furthermore, the posterior mean allele variance parameter was elevated due to the peak height imbalance introduced at CSF1PO (Figure 19: Posterior mean value for allele variance when a peak is included above drop-in cap at CSF1PO)

The results observed indicate that drop-in modelling is performing as expected within STRmix™.

PER LOCUS LIKELIHOOD RATIOS

	FBI_EXTENDED_CAUC 0.01b(1.0, 1.0)		
LOCUS	Pr(E|Hp)	Pr(E|Hd)	LR
D3S1358	1	8.4886E-2	1.1780E1
D1S1656	1	2.4306E-2	4.1142E1
D2S441	1	1.6511E-1	6.0565E0
D10S1248	1	9.9137E-2	1.0087E1
D13S317	1	3.6579E-2	2.7338E1
Penta E	1	1.7504E-2	5.7129E1
D16S539	1	1.1580E-1	8.6353E0
D18S51	1	4.0035E-2	2.4978E1
D2S1338	1	9.9769E-3	1.0023E2
CSF1PO	0	2.4935E-2	0
Penta D	1	5.2949E-2	1.8886E1
TH01	1	1.4051E-1	7.1167E0
vWA	1	1.1350E-1	8.8106E0
D21S11	1	2.2967E-2	4.3541E1
D7S820	1	7.0180E-2	1.4249E1
D5S818	1	2.8702E-1	3.4841E0
TPOX	1	2.8027E-1	3.5680E0
DYS391			
D8S1179	1	1.4076E-1	7.1043E0
D12S391	1	3.1127E-2	3.2127E1
D19S433	1	1.9729E-1	5.0686E0
FGA	1	1.8871E-2	5.2990E1
D22S1045	1	2.3188E-1	4.3126E0
SUB-SUB-SOURCE LR			0
SUB-SOURCE LR			0

Figure 18: Per locus LRs when a peak above drop-in is included in the evidence at CSF1PO

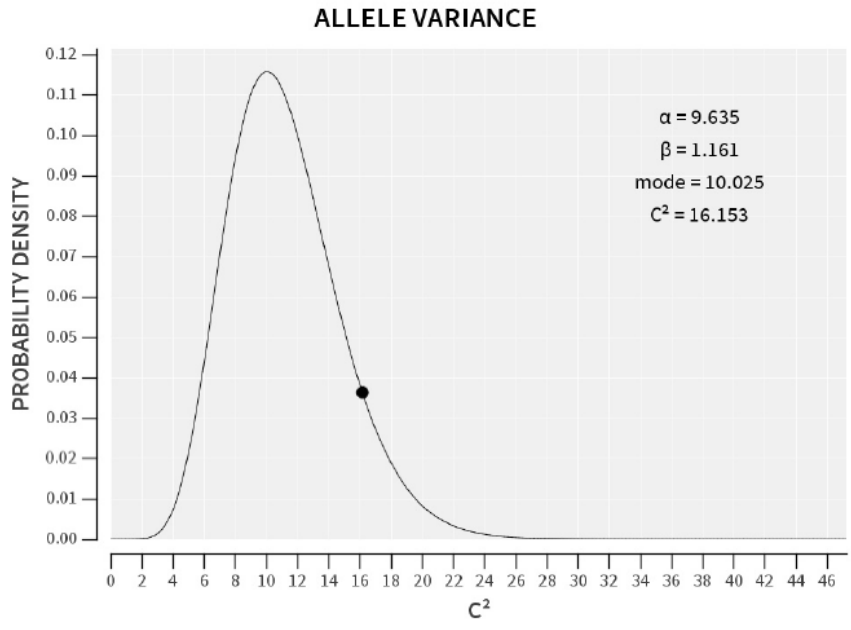


Figure 19: Posterior mean value for allele variance when a peak is included above drop-in cap at CSF1PO

Section H: Forward and reverse stutter

This section covers the following recommendation:

4.1.9. Forward and reverse stutter

Within STRmix™ V2.6 and later, any type of stutter observed by a laboratory may be modelled. This is referred to as *generalised stutter modelling*. The stutter models used within STRmix™ are both locus specific and allele specific and may be based on:

- Allelic designation,
- Longest uninterrupted stretch of repeats (LUS),
- Per allele average observed stutter ratio, or
- Per locus average observed stutter ratio.

The ISP Laboratory have elected to model back stutter and forward stutter at all autosomal loci. Refer to the document 'Estimation of STRmix™ V2.8 parameters for the ISP Laboratory (Fusion 5C 3500)' for further information regarding the stutter models developed.

Stutter peak labels must be retained during profile analysis and included within the STRmix™ input file. Stutter modelling was reviewed throughout the present validation study and found to be intuitive and in line with expectation. This can be seen in the interpretation of single source profiles (see Section A) where stutter peaks are retained at interpretation. As part of the MCMC process they may be proposed as allelic but those genotype combinations are not accepted and therefore receive no weight. In mixed DNA profiles, where the minor contributor is of a similar height as the stutter peaks they start to be considered as possible minor alleles. This behaviour can be seen within the mixture interpretations undertaken as part of Section D.

Section I: Intra locus peak height

This section covers the following standard:

4.1.10. Intra-locus peak height variance

STRmix™ models the variability of single peaks. The variance of this model is determined by directly modelling laboratory data. This is undertaken within STRmix™ using the Model Maker function.

Traditionally heterozygote balance (*Hb*) is investigated, which can be thought of as the variability of two alleles at a heterozygous locus. A plot of log(*Hb*) versus average peak height (APH) of a locus demonstrates that the variability in *Hb* decreases as APH increases.

The performance of Model Maker was checked by plotting the bounds informed by the ISP laboratory's Model Maker results. A plot of log(*Hb*) versus APH for the ISP Model Maker dataset is provided in Figure 20 below. The expected 95% bounds are indicated within the plot using dashed lines. The bounds were calculated as: $\pm\sqrt{2}\times 1.96\times\sqrt{\frac{c^2}{APH}}$

Within the above equation, c^2 = 10.802 (the 50th percentile from the allelic peak height variance prior distribution). Under the assumption of a normal distribution ~95% of data points are expected to fall within +/- 2 standard deviations (95% bounds) of the mean. The 95% bounds encapsulate sufficient data (coverage = 95.0%) indicating that the ISP laboratory's variance parameters are sufficiently optimised.

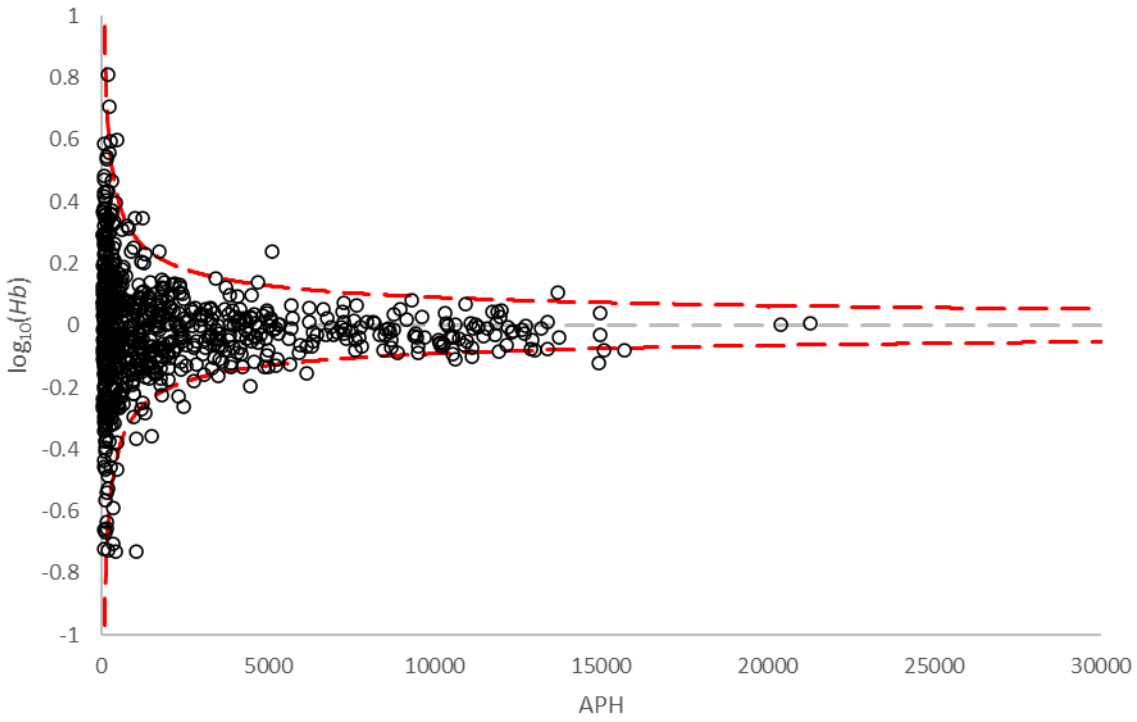


Figure 20: Plot of log(*Hb*) versus average peak height (APH) for Fusion 5C™ 3500 data within the ISP Laboratory.

Section J: Inter-Locus peak heights

This section covers the following standard:

4.1.11. Inter-locus peak height variance

Inter-locus peak variance is modelled in STRmix™ using locus specific amplification efficiencies (LSAE). The LSAE model reflects the observation that even after DNA template amount, degradation, and variation in peak height within loci are modelled, peak heights *between* loci are still more variable than predicted. The variance of this model is determined by directly modelling laboratory data, again using Model Maker. LSAE values for each STRmix™ interpretation appear within the results. The posterior mean LSAE variance parameter is also provided along with a plot of the LSAE variance prior distribution. These may be useful as diagnostics when evaluating a STRmix™ interpretation. A series of tests were performed to examine inter-locus peak height variance within STRmix™ as described below.

Single-source profile (unedited)

A single-source Fusion 5C™ profile (DNA1_500pg.hid) was interpreted within STRmix™. In Figure 21 the LSAE and APH for each locus are plotted, arranged in order of increasing molecular weight. For high quality profiles such as this, the LSAE and APH values are expected to have similar trends and the LSAE variance value to sit within the body of its prior distribution. The degradation curve and LSAE variance plots from the STRmix™ output are also reproduced in Figure 22 to provide a baseline for comparison when evaluating the subsequent tests described below.

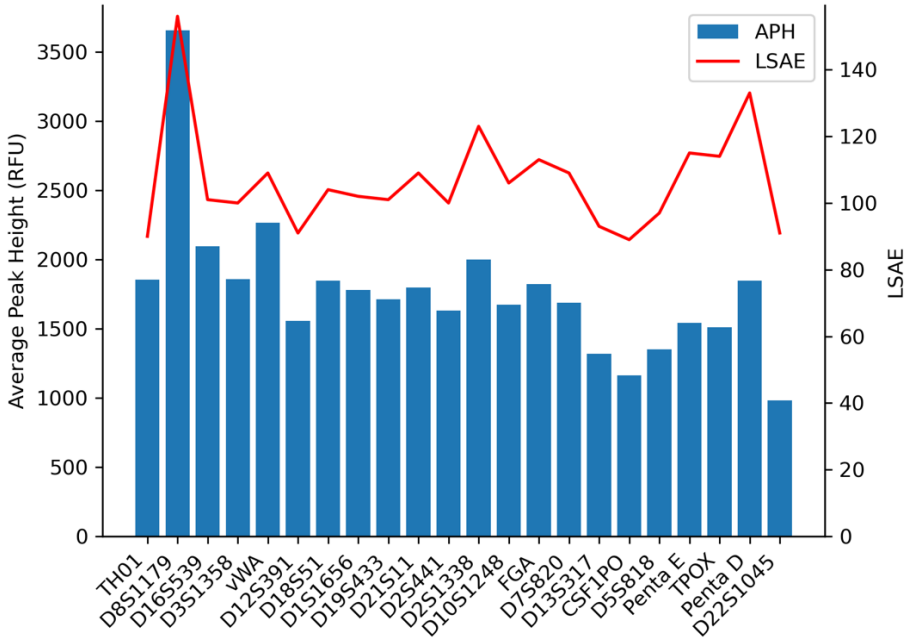


Figure 21: Plot of LSAE and APH values for each locus following interpretation of a single-source Fusion 5C™ profile. Loci are arranged in order of increasing molecular weight.

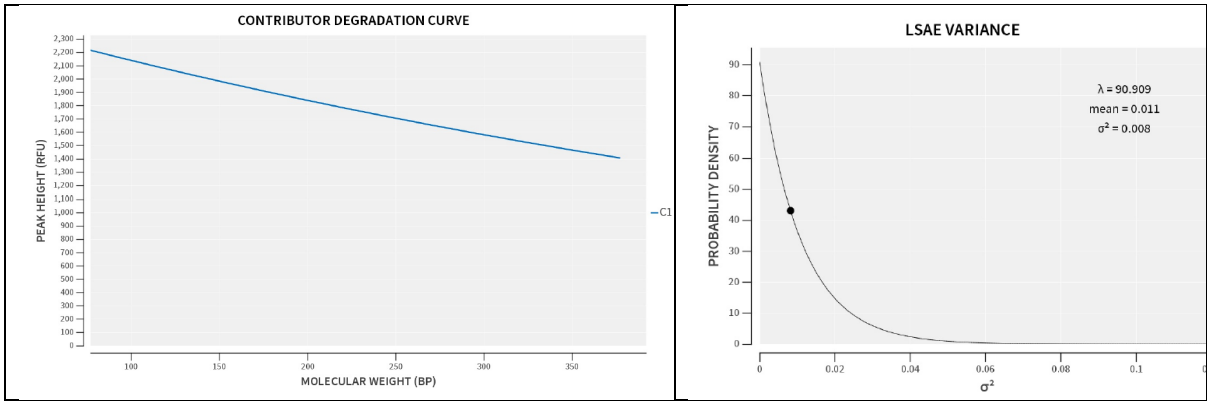


Figure 22: Contributor degradation curve and LSAE variance plot from STRmix™ report following interpretation of a single-source Fusion 5C™ profile. The posterior mean LSAE variance parameter following interpretation is displayed on the prior distribution as a black circle.

Single-source profile (degraded)

The STRmix™ input file of the single-source profile used above was edited *in silico* to simulate the effects of DNA degradation. Peak heights were artificially decreased such that the effect became more pronounced with increasing molecular weight. As LSAE is modelled independently of

degradation, similar LSAE values to the unedited profile are expected to be observed. In contrast, APH is expected to decrease with increasing molecular weight. As before, we plot LSAE and APH for each locus (Figure 23). Inspection of Figure 23 reveals that the expected trends were observed. The degradation curve and LSAE variance plots from the STRmix™ output are also provided in Figure 24. The posterior mean degradation parameter reported by STRmix™ was elevated relative to the unedited profile (1.5131×10^{-3} (unedited) versus 6.3129×10^{-3}(degraded)). The posterior mean LSAE variance value reported by STRmix™ was similar to the unedited profile (0.008 (unedited) versus 0.022 (degraded). These are the expected results.

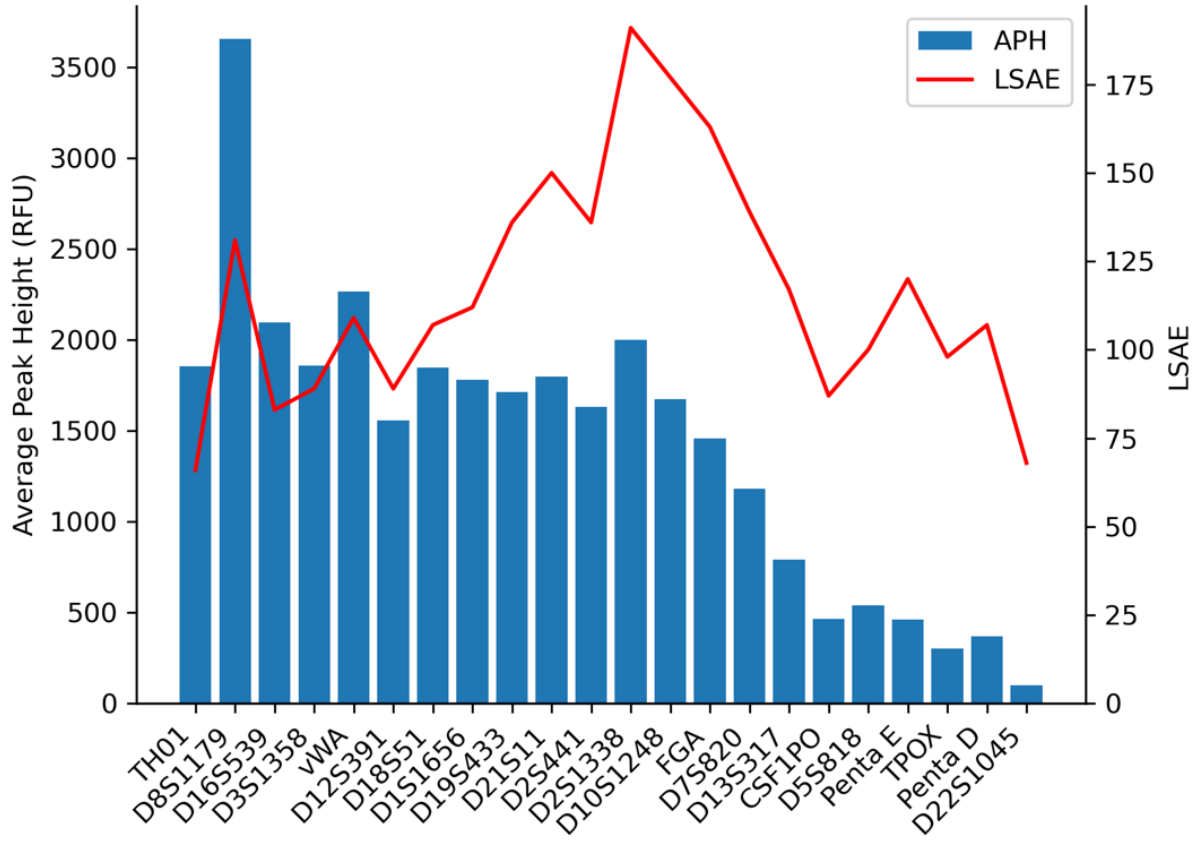


Figure 23: Plot of LSAE and APH values for each locus following interpretation of a single-source Fusion 5C™ profile. The STRmix™ input file was edited to simulate the effects of degradation, with peak heights reduced at the high molecular weight loci. Loci are arranged in order of increasing molecular weight.

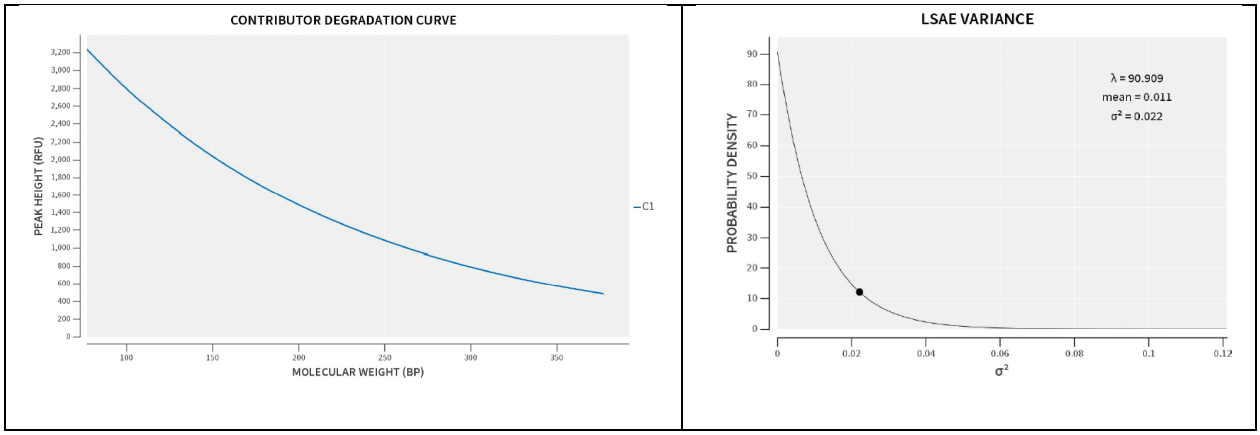


Figure 24: Contributor degradation curve and LSAE variance plot from STRmix™ report following interpretation of a single-source Fusion 5C™ profile. The STRmix™ input file was edited to simulate the effects of degradation. The posterior mean LSAE variance parameter following interpretation is displayed on the prior distribution as a black circle.

Single-source profile (inhibited)

A final test was performed to investigate the effect of inhibition on locus amplification efficiency. The input file of the single-source profile used above was edited *in silico* to simulate the effects of inhibition. In particular, peak heights were substantially reduced at the D3S1358, D16S539, and TH01 loci. A plot of LSAE and APH is provided in Figure 25 below. With the exception of the edited loci, similar results to the unedited profile were observed. Reduced LSAE and APH values were observed at the edited loci, as expected. The posterior mean LSAE variance parameter was also elevated relative to the unedited profile (0.008 (unedited) versus 0.072 (inhibited)). The posterior mean degradation parameter (Figure 26) reported by STRmix™ was similar to that of the unedited profile (1.5131×10^{-3} (unedited) versus 5.3234×10^{-4} (inhibited)). These are the expected results.

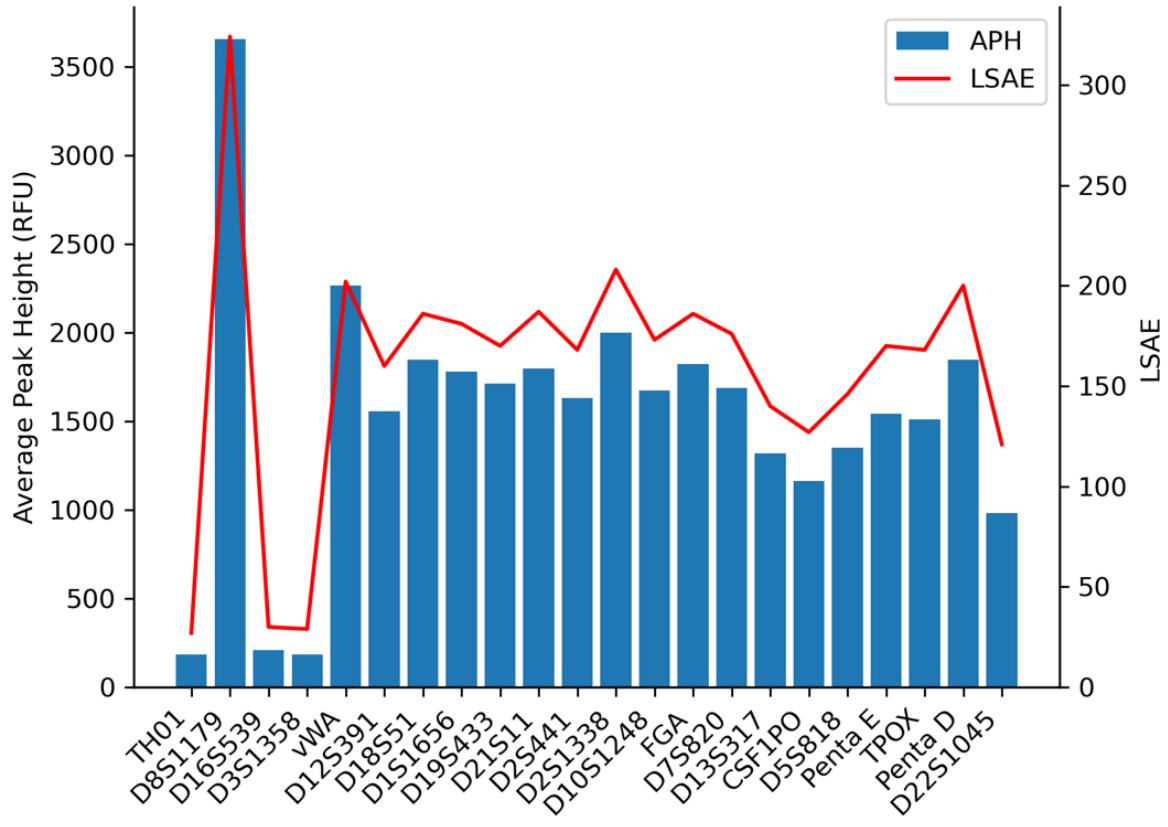


Figure 25: Plot of APH and LSAE values for each locus following interpretation of a single-source Fusion 5C™ profile. The STRmix™ input file was edited to simulate the effects of inhibition, with peak heights substantially reduced at several loci (D3S1358, D16S539, and TH01). Loci are arranged in order of increasing molecular weight.

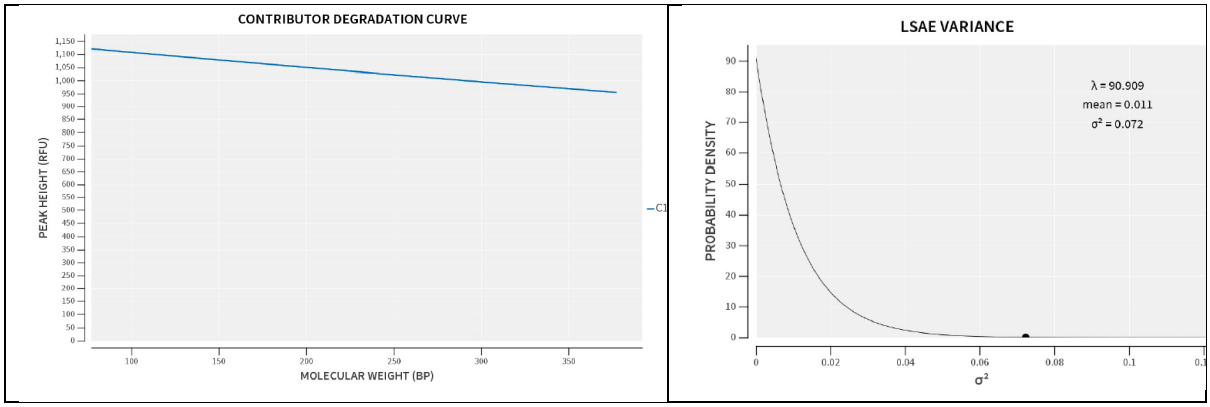


Figure 26: Contributor degradation curve and LSAE variance plot from STRmix™ report following interpretation of a single-source Fusion 5C™ profile. The STRmix™ input file was edited to simulate the effects of inhibition. The posterior mean LSAE variance parameter following interpretation is displayed on the prior distribution as a black circle.

Section K: Additional challenge testing

This section covers the following recommendation:

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

STRmix™ requires that only numeric values are retained within the input files. Any values that are not numeric (such as OL labels not removed during profile analysis, or custom peak labels inclusive of a symbol e.g. '>15') will cause STRmix™ to halt the interpretation. The presence of a non-allelic peak that has sized within an allelic bin position and had its numeric label retained within the input file can lead to a number of outcomes. These include:

- An exclusionary *LR*. A false exclusion may be observed if the artefact peak is modelled as an allelic peak having originated from a contributor of interest.
- No effect. If drop-in modelling is enabled within STRmix™, the artefact may be modelled as a drop-in peak if it is less than the drop-in cap.
- Failure to interpret. The additional peak may artificially increase the minimum number of contributors required to explain the profile. For example, an artefact at a heterozygous locus in a single-source profile will increase the minimum number of contributors by one if the peak cannot be modelled as either stutter or drop-in. In this example, STRmix™ will not progress an interpretation that assumes only one contributor.

Each of these expected outcomes was demonstrated by editing a single-source input file and assigning an *LR* to the known donor within STRmix™. The tests performed are described in further detail below.

Inclusion of OL peak:

An off ladder peak with a peak label of 'OL' was added to the STRmix™ input file at D3S1358 (DNA1_500pg.hid). The peak was given a height of 100 rfu and size of 135 base pairs. The edited input file is provided in Table 6 (D3S1358 only). An error message was produced by STRmix™ after importing the edited input file (Figure 27). An interpretation could not be progressed unless D3S1356 was ignored.

Table 6: Excerpt from STRmix™ input file for a single-source profile (D3S1358 only). The input file was manually edited with an OL peak being added at D3S1358 with height of 100 rfu (highlighted).

Locus	Allele	Height	Size
D3S1358	14	152	116
D3S1358	15	2037	120.33
D3S1358	17	160	129.01
D3S1358	18	1676	133.17
D3S1358	OL	100	135

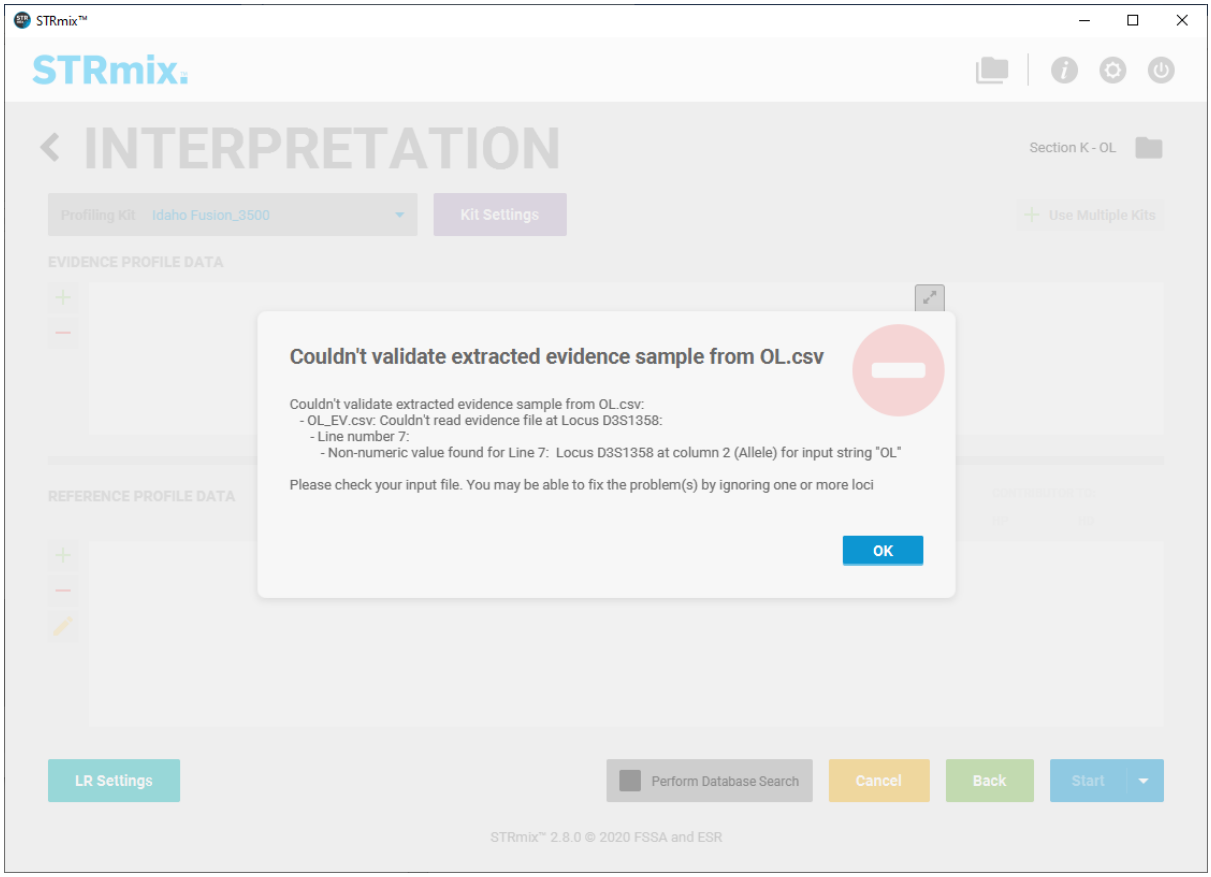


Figure 27: Error message displayed by STRmix™ after attempting to add an input file that contained an off ladder (OL) peak.

Refer to Section G for the demonstration of a false exclusion due to the inclusion of an artefact peaks (drop-in peak above drop-in cap at a homozygous locus), the inclusion of an artefact peak

resulting in no effect and the inclusion of an artefact resulting in a "failure to interpret" message in STRmix™.

While the effect of artefact peaks in these single-source tests was readily identifiable, the situation encountered when interpreting mixed DNA profiles may be more subtle. Every effort should be made to remove artefact peaks during profile analysis. As demonstrated above, the retention of artefact peaks may have no effect on the interpretation or may lead to the false exclusion of a true donor. In some cases, an interpretation will not progress at all. In other cases, it may result in elevated secondary run diagnostics. Close examination of the primary and secondary diagnostics following interpretation may help to identify instances where artefact peaks have erroneously been retained (for example, an *LR* of zero at a single locus or acceptance of non-intuitive genotypes).

Section L: Comparison to previous casework profile interpretation methods

This section covers the following recommendations:

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

> 4.2.1. The laboratory should determine whether the results produced by the probabilistic genotyping software are intuitive and consistent with expectations based on non-probabilistic mixture analysis methods
>
> > 4.2.1.1. Generally, known specimens that are included based on non-probabilistic analyses would be expected to also be included based on probabilistic genotyping
>
> 4.1.7. Partial profiles, to include the following:
>
> > 4.1.7.2. DNA degradation
> >
> > 4.1.7.3. Inhibition

STRmix™ probabilistic genotyping software was compared to ISP's manual methods during its original validation in 2015. During the original validation, proficiency test and training samples covering a range of scenarios, conclusions, and statistics calculations were reinterpreted with STRmix™ v2.3. The STRmix™ interpretations were shown to be generally consistent with those using the existing manual interpretation procedures at the time. Refer to the "Casework Profiles" section of the STRmix™ v2.3 validation completed in November of 2015 for more information.

Section M: Precision

This section covers the following recommendation:

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

Refer to section D above for details of sensitivity and specificity tests.

STRmix™ utilises Markov chain Monte Carlo (MCMC) to generate genotype weights during profile interpretation. MCMC is a standard mathematical process that relies on random sampling. Therefore, the genotype weights produced by STRmix™ will vary if the interpretation is repeated. As these weights feed into any subsequent *LR* calculation, variability in the weights will result in variability in the *LR*. The variability in *LR*s between replicate interpretations has previously been explored [14]. Typically, the level of variation observed is less than an order of magnitude and is unlikely to be large enough to change the general conclusions drawn from the *LR*. The MCMC process was shown to be a small source of variability compared with other laboratory variables including the PCR and CE processes. *LR* variability due to the size of the allele frequency database and the MCMC process is taken into account within STRmix™ using the highest posterior density (HPD) method [15-17]. This method calculates a probability interval, which may be thought to be similar to a confidence interval.

Parameters within STRmix™ that affect run to run variability include the number of MCMC accepts, the number of chains, and the random walk standard deviation (RWSD). Within STRmix™ V2.8 the default number of accepts per chain is set to 10,000 burn-in accepts and 50,000 post-burn-in accepts. By default, STRmix™ utilises eight independent chains. The settings controlling the number of Markov chains and RWSD were optimised during developmental validation of STRmix™ and it is recommended that these are not modified [1]. These settings should be suitable for the majority of casework profiles encountered by a laboratory. Decreasing the number of accepts will improve run time but may mean that STRmix™ does not reach convergence during burn-in, leading to increased variability. Increasing the number of accepts may mean convergence is achieved (if it has not already) but will almost certainly increase run time. The number of accepts selected by the user is therefore a trade-off between precision and run time.

The extent of STRmix™ run variability was investigated using a four-person 1:1:1:1 mixture where the previous Section D interpretation demonstrated ambiguity in the genotype weightings. The profile was interpreted in STRmix™ five times under the following conditions:

- 1,000 burn-in accepts, 5,000 post-burn-in accepts per chain (0.10x default values)
- 10,000 burn-in accepts, 50,000 post-burn-in accepts per chain (default values)
- 100,000 burn-in accepts, 500,000 post-burn-in accepts per chain (10x default values)

Following interpretation, an *LR* was assigned for one of the known donors (K54) using a theta value of 0.01 and the FBI extended Caucasian allele frequencies. The sub-source and HPD log(*LR*)s were recorded and are plotted below Figure 29).

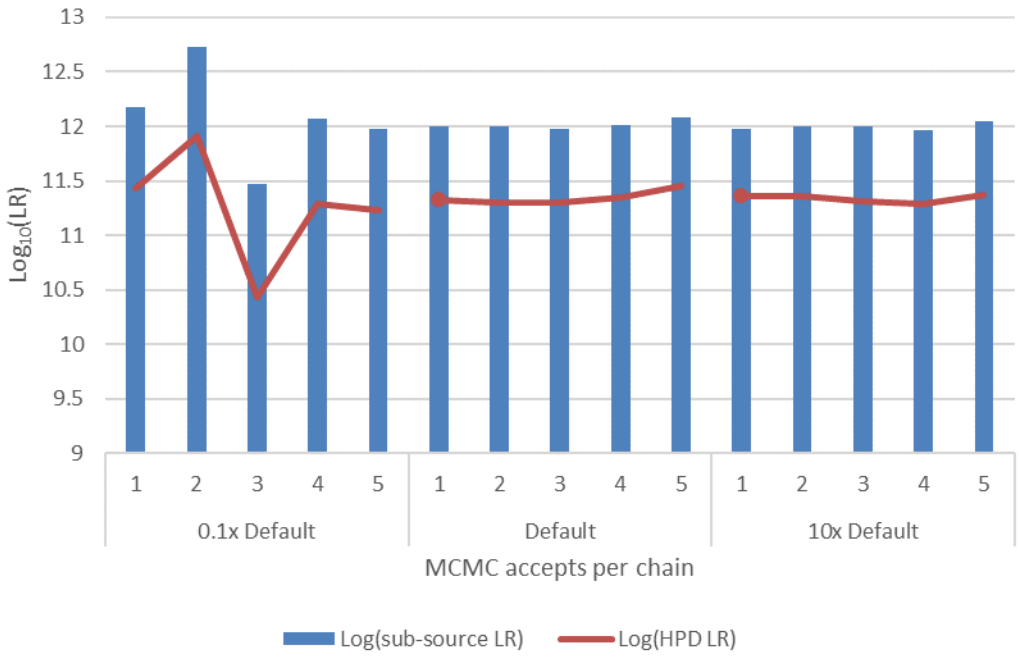


Figure 28: Log(*LR*)s assigned following repeat interpretation of 1:1:1:1_C1 with LRs to known donor K54

The difference in sub-source log(*LR*)s between the maximum and minimum *LRs* for 0.1 times the default MCMC accepts per chain was 1.26 (i.e. greater than an order of magnitude difference in *LR*s), the difference using the default settings was 0.11 (i.e. less than an order of magnitude difference in *LR*s) and the difference using 10 times the default settings was 0.08 (also less than an order of magnitude of difference). The average run-time for the 0.1 times the default settings was ~44 seconds, the average run-time using the default MCMC accepts was ~4 minutes and the average run-time using 10 times the default settings was ~36 minutes.

In general, increased MCMC accepts led to reduced variability in the *LR* with the trade off of increased run-time. From the plots above, it can be seen that the HPD *LR* for any given interpretation was always less than the corresponding sub-source *LR*. This is the expected result given that the HPD *LR* represents a lower bound whilst the sub-source *LR* may be considered to be a point estimate. These results demonstrate that the default MCMC accepts are working as expected using ISP's STRmix™ kit and data.

Conclusion

This document describes the internal validation of STRmix™ V2.8 within the ISP Laboratory. It has been shown that it is suited for its intended use for the interpretation of profiles generated from crime scene samples.

Signatures

Laboratory STRmix™ implementation manager
(ISP Laboratory Quality Manager)

ISP Laboratory Technical Leader

This work has been reviewed and it has been determined that STRmix™ V2.8 is suitable for its intended use for interpretation of crime profiles within the ISP Laboratory. The project work has met the validation requirements as required by A2LA.

References

1. Bright, J.-A., et al., *Developmental validation of STRmix™, expert software for the interpretation of forensic DNA profiles.* Forensic Science International: Genetics, 2016. **23**: p. 226-239.
2. Scientific Working Group on DNA Analysis Methods, *Guidelines for the Validation of Probabilistic Genotyping Systems*. 2015.
3. Taylor, D., *Using continuous DNA interpretation methods to revisit likelihood ratio behaviour.* Forensic Science International: Genetics, 2014. **11**: p. 144-153.
4. Moretti, T.R., et al., *Population data on the expanded CODIS core STR loci for eleven populations of significance for forensic DNA analyses in the United States.* Forensic Science International: Genetics, 2016. **25**: p. 175-181.
5. Bright, J.-A., et al., *A series of recommended tests when validating probabilistic DNA profile interpretation software.* Forensic Science International: Genetics, 2015. **14**: p. 125-131.
6. Balding, D.J. and R.A. Nichols, *DNA profile match probability calculation: how to allow for population stratification, relatedness, database selection and single bands.* Forensic Science International, 1994. **64**: p. 125-140.
7. National Research Council II, *National Research Council Committee on DNA Forensic Science, The Evaluation of Forensic DNA Evidence*. 1996, Washington, D.C.: National Academy Press.
8. Bright, J.-A., et al., *Internal validation of STRmix™ – A multi laboratory response to PCAST.* Forensic Science International: Genetics, 2018. **34**: p. 11-24.
9. Coble, M.D., et al., *Uncertainty in the number of contributors in the proposed new CODIS set.* Forensic Science International: Genetics, 2015. **19**: p. 207-211.
10. Curran, J.M. and J. Buckleton, *Uncertainty in the number of contributors for the European Standard Set of loci.* Forensic Science International: Genetics, 2014. **11**(1): p. 205-206.
11. Bright, J.-A., J.M. Curran, and J.S. Buckleton, *The effect of the uncertainty in the number of contributors to mixed DNA profiles on profile interpretation.* Forensic Science International: Genetics, 2014. **12**: p. 208-214.
12. Bright, J.-A., et al., *Searching mixed DNA profiles directly against profile databases.* Forensic Science International: Genetics, 2014. **9**: p. 102-110.
13. Moretti, T.R., et al., *Internal validation of STRmix for the interpretation of single source and mixed DNA profiles.* Forensic Science International: Genetics, 2017. **29**: p. 126-144.
14. Bright, J.-A., et al., *The variability in likelihood ratios due to different mechanisms.* Forensic Science International: Genetics, 2015. **14**: p. 187-190.

15. Taylor, D., et al., *An illustration of the effect of various sources of uncertainty on DNA likelihood ratio calculations.* Forensic Science International: Genetics, 2014. **11**: p. 56-63.

16. Triggs, C.M. and J.M. Curran, *The sensitivity of the Bayesian HPD method to the choice of prior.* Science & Justice, 2006. **46**(3): p. 169-178.

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

Appendix 1: Cross-reference for document sections and SWGDAM recommendations

Recommendation	Text	Refer to section
4.1	Test the system using representative data	Preamble
4.1.1	Specimens with known contributors	Preamble
4.1.2	Hypothesis testing with contributors and non-contributors	D
4.1.2.1	More than one set of hypotheses	E
4.1.3	Variable DNA typing conditions	Preamble
4.1.4	Allelic peak height, to include off-scale peaks	B
4.1.5	Single-source specimens	A
4.1.6	Mixed specimens	D
4.1.6.1	Various contributor ratios	D
4.1.6.2	Various total DNA template quantities	D
4.1.6.3	Various numbers of contributors	D
4.1.6.4	Both correct and incorrect number of contributors (i.e. over- and under-estimating)	F
4.1.6.5	Sharing of alleles among contributors	D
4.1.7	Partial profiles	D
4.1.7.1	Allele and locus dropout	D
4.1.7.2	DNA degradation	L
4.1.7.3	Inhibition	L
4.1.8	Allele drop-in	G
4.1.9	Forward and reverse stutter	H
4.1.10	Intra-locus peak height variance	I
4.1.11	Inter-locus peak height variance	J
4.1.12	In-house parameters	Preamble
4.1.13	Sensitivity, specificity, and precision	D and M
4.1.14	Additional challenge testing	K
4.2	Compare the results of probabilistic genotyping and of manual interpretation	L
4.2.1	Intuitive and consistent with expectations	L
4.2.1.1	Known specimens that are included based on non-probabilistic analyses would be expected to also be included based on probabilistic genotyping	L
4.2.1.2	Concordance of single-source specimens with high quality results	A
4.2.1.3	Generally, as the analyst's ability to deconvolute a complex mixture decreases, so does the weighting of a genotype set determined by the software	C

Appendix 2: Review of Secondary Run Diagnostics

Secondary diagnostics are a useful guide to provide confidence the STRmix™ interpretation has progressed as expected. Individual secondary diagnostics may indicate whether a more comprehensive review is warranted, however analysts should not rely on these diagnostics alone. Elevated values for one of these diagnostics may not necessarily mean the results are unfit for purpose. To put in context the range of diagnostic values that can be expected from ISP data, a discussion of the secondary run diagnostics obtained from the Section D interpretations is provided below.

Average log(likelihood):

STRmix™ uses a biological model to generate an expected DNA profile which is then compared with the observed profile. When assessing the fit of the expected profile with the observed, STRmix™ calculates a 'grade', referred to as a log(likelihood). The average log(likelihood) diagnostic reported in the STRmix™ output is the average of the log(likelihood) values across all post-burn-in iterations. The larger this value is, the better STRmix™ has been able to describe the observed data. A low or negative value suggests that STRmix™ has not been able to describe the data very well given the information it has been provided with. Reasons why this value may be low or negative include:

1. The profile is simply low level and there is very little data making up the likelihood,
2. There are large stochastic events in the STRmix™ run (e.g. large heterozygote peak imbalances or variation in mixture proportions across the profile). These may be forced by mis-assignment of the number of contributors,
3. Data has been removed that was real, in particular stutter peaks, and must now be described within STRmix™ by dropout, and
4. Artefactual peaks have been left labelled and must now be accounted for within STRmix™ by e.g. drop-in.

As per point 1 above, it is important to note that low or negative average log(likelihood) values may legitimately be produced when interpreting low level DNA profiles. As such, low or negative average log(likelihood) values do not necessarily indicate that the STRmix™ results are unreliable.

The average log(likelihood) diagnostic for each Section A and D interpretation is plotted against NOC in Figure 30 below.

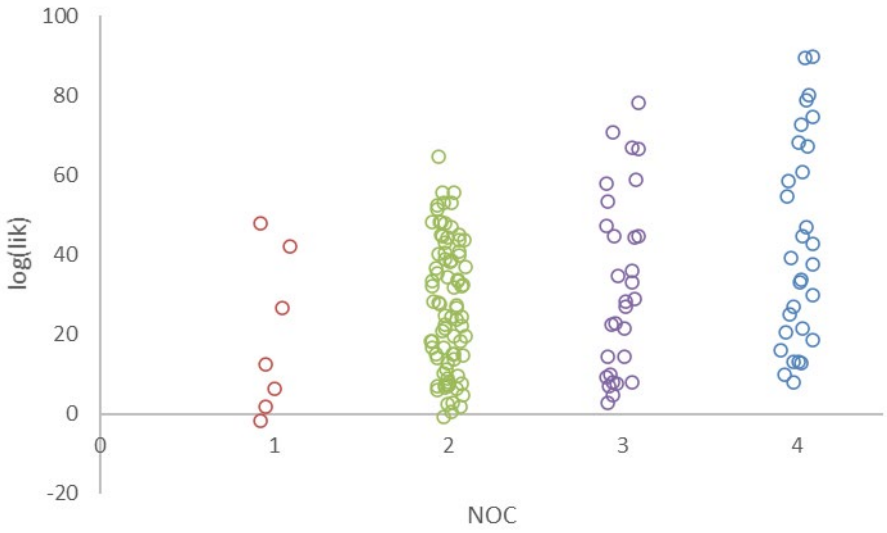


Figure 29: Plot of average log(likelihood) diagnostic versus number of contributors.

Gelman-Rubin convergence diagnostic:

Ideally, each MCMC chain will reach the area of high probability within the sample space during burn-in and will continue to sample from this space during the post-burn-in MCMC. This is referred to as 'convergence'. If the chains spend their time in different spaces during the post-burn-in MCMC then it is likely that the analysis has not been run for long enough. The Gelman-Rubin (GR) convergence diagnostic included in the STRmix™ report can indicate to the user if the Markov chains have not sufficiently converged. If the chains have been sampling from the same space, then the GR diagnostic should be close to 1.0. Notionally, values above 1.2 indicate that the chains may not be nearing convergence. It is important to note that the GR diagnostic output by STRmix™ is a summary statistic: values less than 1.2 do not guarantee that all parameters have converged whilst values greater than 1.2 do not necessarily indicate that the results are unreliable.

In rare instances, one (or more) chain(s) may fail to find the area of high probability space altogether. This is referred to as a wandering chain and typically leads to substantially elevated GR diagnostics. Often, the genotypes accepted at one or more loci will not be intuitive in instances where there has been a wandering chain. Simply re-running the interpretation will typically recover the GR and produce sensible results. However, not all causes of an elevated GR can be addressed in this way, therefore as with all run diagnostics it is recommended that both the input and primary and secondary outputs of runs with excessive values are closely scrutinized.

The GR convergence diagnostic for each Section A and D interpretation is plotted against NOC in Figure 31 below. The largest GR observed was approximately 1.37 and was produced following interpretation of a complex four-person (1:1:1:1) mixture. The genotype weights reported by STRmix™ were reviewed and found to be intuitive. Inclusionary *LR*s were produced for all known donors to this mixture. All of the non-donors compared with this profile gave an *LR* of 0. In this instance, the elevated GR is likely due to the complexity of the mixture. These findings demonstrate that reliable results may still be produced in circumstances where an elevated GR diagnostic has been produced.

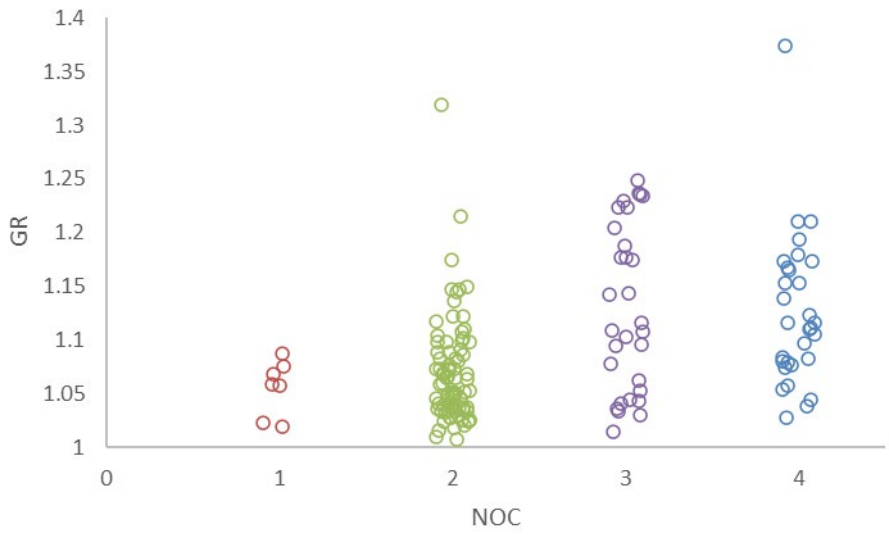


Figure 30: Plot of Gelman-Rubin (GR) convergence diagnostic versus number of contributors.

Posterior variance parameters:

Within the STRmix™ report, the posterior mean variance parameters are overlaid on the relevant prior distributions. Ideally, each of the posterior variance parameters should sit within the body of the relevant prior distribution. Values that fall in the right hand tail of the prior distribution may warrant further investigation. A large allele variance parameter in conjunction with a low or negative average log(likelihood) diagnostic may indicate that the number of contributors to the profile has been mis-assigned. Excessive stutter variance parameters may be due to the inadvertent application of a stutter filter during CE profile analysis. As with the other secondary diagnostics described above, elevated variance parameters do not necessarily invalidate the results. Provided that the primary diagnostics are intuitive, the STRmix™ results are likely reliable.

The posterior variance parameters for each Second D interpretation along with their prior distributions are provided in Figure 32 below. One back stutter posterior mean variance value (66.08) sat out in the right-hand side of the prior distribution and the sample was investigated. It was found that there was a missing back stutter peak at D22S1045 which was picked up during the input file error checks STRmix undertakes and prints to the PDF report (Figure 33). All of the other values reported by STRmix™ were acceptable and did not warrant further investigation.

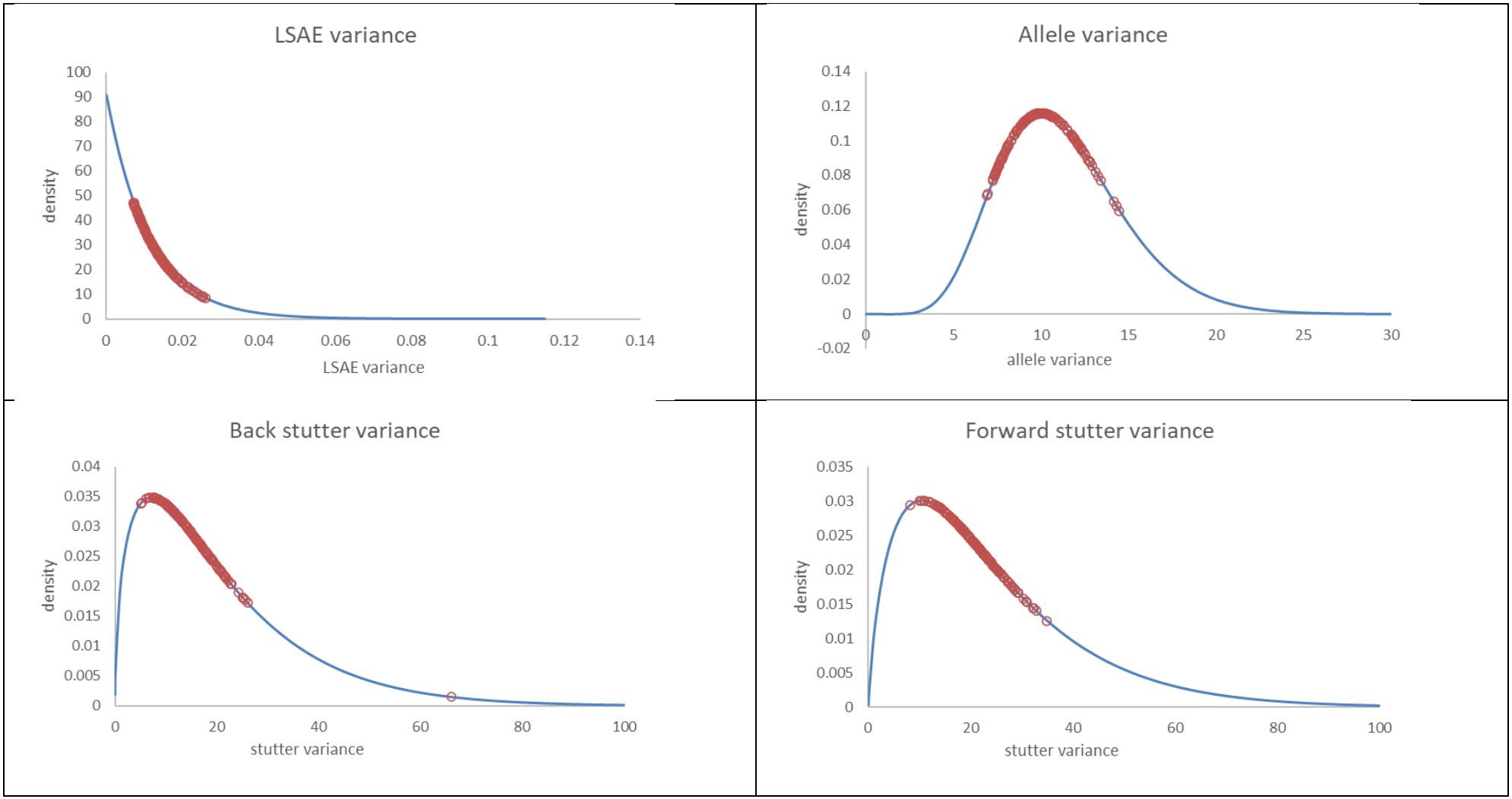


Figure 31: Plots of peak height and LSAE variance. The prior distributions for each parameter are plotted in blue. The posterior mean variance parameters are plotted in red. These are the optimised values reported by STRmix™ following profile interpretation.

EVIDENCE PEAK ISSUES

10.5.1_C1.HID_EV.CSV

LOCUS	PEAK	ISSUE	DECISION
Missing Stutter Peaks			
D22S1045	15	Allele 16 is missing Back Stutter at position 15 (expected height of 342 RFU)	-

Figure 32: Error check undertaken in STRmix showing the missing back stutter peak

Mixture proportions:

Within the STRmix™ interpretation report, the template values per contributor (rfu) and mixture proportion are given. The template values approximate the allelic peak heights from each contributor at the left-hand side of the electropherogram (before degradation can be seen in the profile). The mixture proportions are calculated using the template values. These values can be used to check that the interpretation has progressed as expected. The template and mixture proportion values for Section D were reviewed and the mixture proportions for one sample were thought to be significantly different than expected. Sample 19.1 C4 was experimentally designed with mx ratios of 19:1 (with 12.5 pg of DNA for the minor contributor) but the resulting STRmix™ ratio was 3:1.Whilst there is limited information present within the input file to assist with the approximation of mixture ratio, the STRmix™ posterior proportions do appear more intuitive given the information apparent at some loci. This sample was re-interpreted in STRmix™ using the M_x priors function. The M_x priors function informs STRmix™ of the template/mx proportions of each contributor based on analyst expectation. STRmix™ may explore other values but will be penalised the further it moves away from the prior expectation. The settings used for each contributor are in Figure 34. The resulting mix ratio when using the mx priors function was 14:1 (the template values were 496 rfu and 36 rfu). The log(*LR*) for each of the known contributors changed from; 25.73 to 25.33 and 3.94 to 4.85 indicating that the mx priors function may have helped STRmix to "see" the trace contributor somewhat, however there was no large shift in *LR* for either contributor.

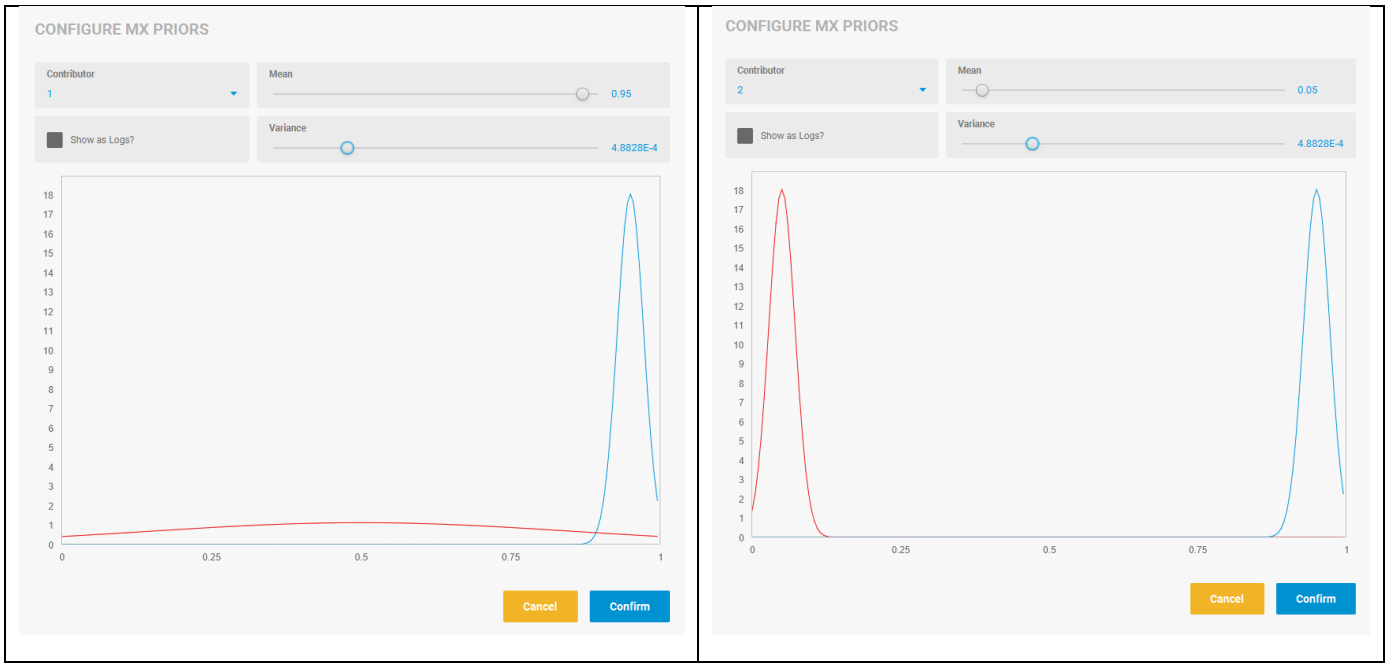


Figure 33: Mx priors settings