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.