Case 7 Excerpt of Post-Validation Performance Testing of Previously Untested High Allele Sharing between Non-contributors and True contributors using the Forensic Statistical Tool (FST)

From a software engineering perspective, the validation of any probabilistic genotyping should include the highest levels of safety measures to avoid miscarriages of justice. Dr. Jeanna Matthews is a Professor of Computer Science at Clarkson University and member of IEEE[1]-USA, an organization representing roughly 150,000 engineers, scientists, and allied professionals in the United States, including those conducting research in emerging technologies. Dr. Matthews is also co-chair of IEEE-USA's AI Policy Committee. In a recent statement on DNA mixture interpretation, the organization articulated its position: "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."[2]

The IEEE-USA has also emphasized that "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."[3] This independent verification and validation should conform to the same high standard – commonly known as IEEE 1012 – as is "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."[4]

Independent verification and validation (IV&V), in the context of DNA software, should be able to address the following: "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?"[5]

Unsurprisingly, it is the position of the IEEE-USA that "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."[6]

As noted in her previous research, Dr. Matthews and her team of students developed a batch-processing testing harness for FST in order to test the effect of a hidden function that dropped a locus LR if allele frequencies exceeded 97 percent.[7] While of an extremely limited nature, her work with the hidden FST function was at least a beginning to apply IV&V to FST 2.5.

[1] The Institute of Electrical and Electronics Engineers is a not-for-profit organization. IEEE is "the world's largest technical professional organization dedicated to advancing technology for the benefit of humanity." https://www.ieee.org/.

[2] 11/18/2021 IEEE-USA letter, RFC Response: NIST Internal Report 8351-DRAFT DNA Mixture Interpretation: A NIST Scientific Foundation Review at 2.

[3] *Id.* at 3.

[4] *Id.*

[5] *Id.*

[6] *Id.* at 4.

[7] Matthews et al., *The Right To Confront Your Accusers: Opening the Black Box of Forensic DNA Software,* AIES'19, January 27–28, 2019, Honolulu, HI, USA, available at https://www.aies-conference.com/2019/wp-content/papers/main/AIES-19_paper_72.pdf.The Forensic Statistical Tool or FST was developed in-house at the New York City Office of Chief Medical Examiner. Spurred by litigation in federal court, in 2016 the lab withdrew a demand for a protective order and the source code and supporting documentation for FST v. 2.5 was made publicly available on GitHub by the news organization

Another problematic area that calls for IV&V is the FST's inability to take into account the phenomenon of high allele sharing. Allele sharing occurs when multiple contributors like relatives share the same variations of genes at any particular location. When validating probabilistic genotyping software, it is critical to evaluate the effect of allele sharing. "From a performance-driven perspective, it is quite important to include an evaluation of varying degrees of allelic overlap, both because mixtures with high levels of allele sharing are known to be exceptionally challenging and prone to erroneous interpretation under a traditional interpretation framework (Butler et al. 2018), and because situations where high allelic overlap could occur—such as the possibility of multiple family members contributing to a DNA mixture—are commonly encountered in forensic casework."[8] Section 4.1.6 of the validation guidelines of the Scientific Working Group for DNA Analysis Methods (SWGDAM) lists certain criteria that should be addressed with respect to mixed samples, including 4.1.6.5, "sharing of alleles among contributors."[9]

More recently, scientists at the National Institute of Standards and Technology have noted, "[a]n important missing element from many validation studies is the degree of allele sharing that has been tested."[10] They went on to note that "[i]f validation studies are conducted using mixtures that do not explore the complexity induced by allele sharing, the user may inadvertently extrapolate validation results and apply methods beyond the limits of the validation studies conducted."[11] NIST scientists advise that "[p]articular attention should be paid to validation data for DNA mixture interpretations that are expected to have a high degree of uncertainty, for example, when a contributor of interest has contributed very low DNA template quantities, or there are large amounts of allele sharing, or many contributors in the sample."[12]

In its published paper, the OCME scientists who developed FST admitted that "correlation among genotypes of contributors to mixtures is not considered, which means the calculation is based on unrelated individuals. Each unknown person's genotype is treated as independent from the genotypes of all others in the model. Hypotheses that involve unknown individuals who are related cannot be explicitly modeled and for example, FST cannot accommodate a prosecution hypothesis that includes the victim and the suspect along with a defense hypothesis that includes the victim and the suspect's brother."[13] However, the FST developers never incorporated intentional allele sharing into any of the FST validation. Until the source code was made public, there was no validation and verification whatsoever – internal, independent or otherwise –to intentionally test the effect of high allele sharing on FST results. The effect was to increase the danger, as NIST scientists have

ProPublica. *See also When Trusted Black Boxes Don't Agree: Incentivizing Iterative Improvement and Accountability in Critical Software Systems,* Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, February 2020, Pages 102–10, available at https://dl.acm.org/doi/10.1145/3375627.3375807.

[8] Krane D.E., Philpott M.K., (2022) Using Laboratory Validation to Identify and Establish Limits to the Reliability of Probabilistic Genotyping Systems. In Dash H.R., Shrivastava P., Lorente J.A. (eds) Handbook of DNA Profiling. Springer, Singapore. https://link.springer.com/referencework/10.1007/978-981-15-9364-2.

[9] Scientific Working Group on DNA Analysis Methods (SWGDAM) (2015) Guidelines for the Validation of Probabilistic Genotyping Systems. Available at https://1ecb9588-ea6f-4feb-971a-73265dbf079c.filesusr.com/ugd/4344b0_22776006b67c4a32a5ffc04fe3b56515.pdf.

[10] NISTIR 8351-DRAFT, DNA Mixture Interpretation: A NIST Scientific Foundation Review, at 86, available at https://nvlpubs.nist.gov/nistpubs/ir/2021/NIST.IR.8351-draft.pdf.

[11] *Id.*

[12] *Id.* at 89.

[13] Mitchell et al., *Validation of a DNA mixture statistics tool incorporating allelic drop-out and drop-in,* Forensic Science International: Genetics 6 (2012) 749–761, at 760.

recently warned, that "the user may inadvertently extrapolate validation results and apply methods beyond the limits of the validation studies conducted."[14]

To test the efficacy of FST applied to mixtures with high allele sharing, Dr. Matthews and her team employed their batch-processing testing harness to run FST comparisons with 10 mixtures from the original FST false positive study conducted by the OCME in 2010. Dr. Matthews only had access to version 2.5 of FST, which had been made publicly available after litigation in 2016.[15]

Ten mixtures were chosen because of the estimated computing time to run 1000 simulated siblings of true contributors to those mixtures. All ten mixtures were low copy number, with less than 100 picograms amplified by the lab. Each of the mixtures also showed concordance between the true contributors' LRs produced back in 2010 by the OCME and FST v. 2.5 as employed by Dr. Matthews' batch-processing testing harness. In other words, these ten mixtures gave the same LRs for FST 1.0 and FST 2.5. Therefore, these mixtures were not subject to any alterations made to FST by the OCME after its 2010 validation.

The ten mixtures were also chosen where there were at least two contributors with inclusionary likelihood ratios. The true contributor with the highest LR was chosen as the basis for generating the simulated profiles, as well as the true contributor with the lowest inclusionary LR. In that way, the simulated profile sets related to a "major" contributor and to a "minor" contributor could be explored.

To create the sets of 1000 high allele-sharing simulated profiles, the open source program Sibulator was used.[16] The use of simulating profiles is a commonplace in the study of probabilistic genotyping software. In fact the use of simulated relatives was used by the California Department of Justice in its validation of the commercial software STRmix.[17]

Plots were created of the ten mixtures showing the distribution of the 1000 simulated profiles. Plots were also created to compare LRs from non-contributor profiles used by the OCME in August of 2010, with LRs in this study from simulated profiles with high allele sharing. The OCME had developed a "bulk calculator" for FST, but that mode apparently had issues and was never used in casework. The lab used the bulk calculator in 2010 to create LRs from the comparison of 700 NIST profiles, 546 morgue body profiles, and an unknown number of lab personnel profiles. However, in its validation binders the OCME only preserved the first two pages of its bulk calculator runs for its final phase of testing. These first two pages listed the top 86-89 non-contributor LRs produced in the 2010 OCME testing. Since there were only 86-89 of the highest 2010 OCME profiles, they were compared to the same top number of the highest simulated profile LRs.

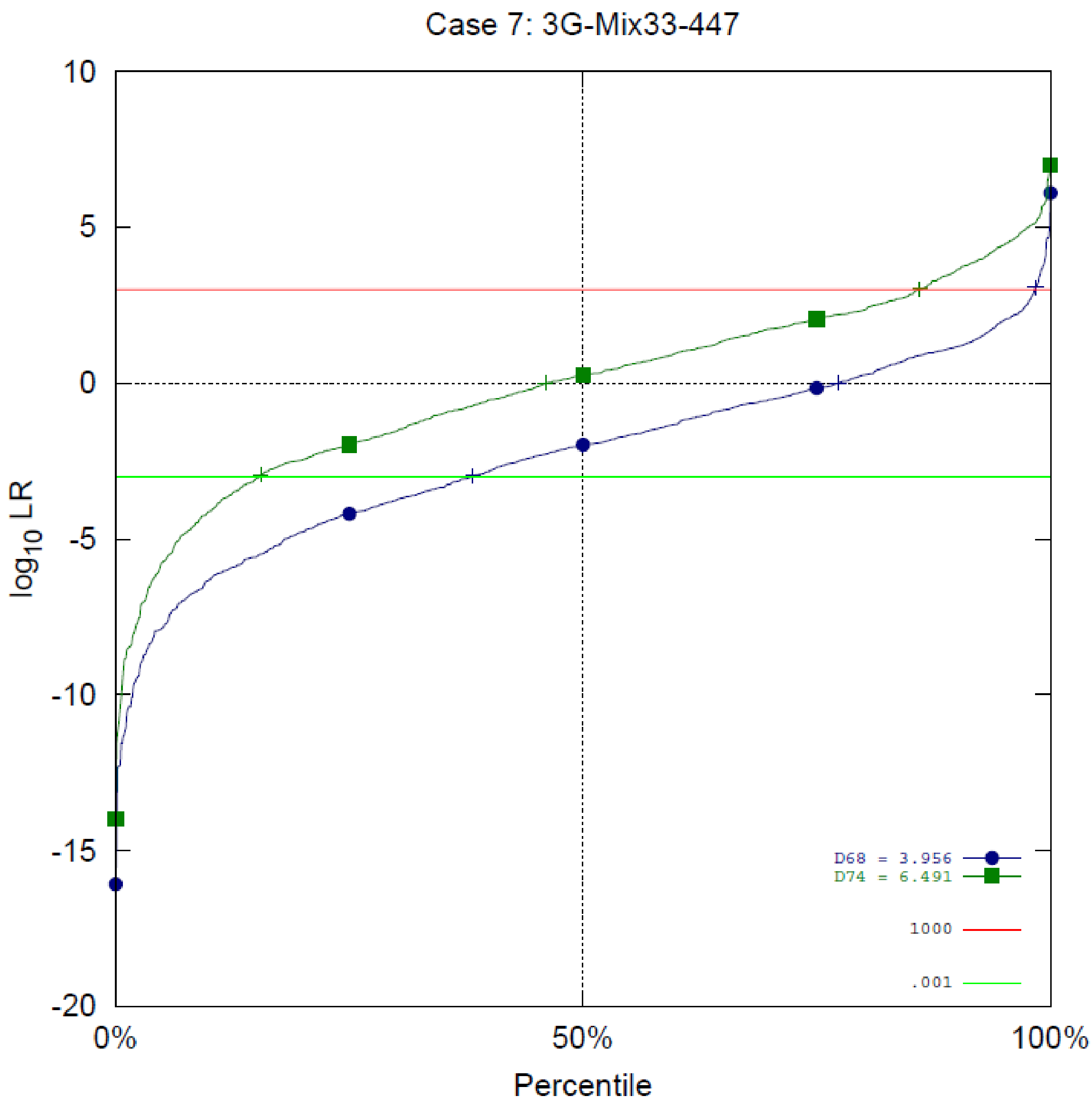
The following are plots and discussion of Case 7, arguably the most extreme of the ten selected samples.

[14] *See* Footnotes 11-14, *supra*.

[15] Software available at https://github.com/propublica/nyc-dna-software.

[16] The program, its methodology, user license and instructions are available at https://github.com/wesleyyuan17/LAS-Sibulator. The profiles in this study were generated using the NIST Caucasian allele frequency option.

[17] STRmix V2.0.6 BFS Casework Internal Validation Summaries, EPIC-16-02-02-CalDOJ-FOIA-20160219-STRmix-V2.0.6-Validation-Summaries, at 6-7 ("For each relationship category, 10,000 profiles were modeled in relation to the tested contributor's reference profile. The following relationships were simulated: parent/child, full sibling, half-sibling (which has the same degree of autosomal genetic relatedness as uncle/aunt, niece/nephew, grandparent, and grandchild), first cousin, second cousin, and unrelated"). Available at https://archive.epic.org/state-policy/foia/dna-software/EPIC-16-02-02-CalDOJ-FOIA-20160219-STRmix-V2.0.6-Validation-Summaries.pdf.



Case 7 was **Mix 33 from Study 3G,** a 45 picogram deducible three-person mixture. The more “major” true contributor **D74** had an LR of 3.1 million ($\log_{10}$LR = 6.491), while the more “minor” **D68** had an LR of 9,045 ($\log_{10}$LR = 3.956). Over 53 percent of LRs for siblings for “major” D74 were above 1, while “minor” D68 had 22.7 percent of LRs for siblings above 1.

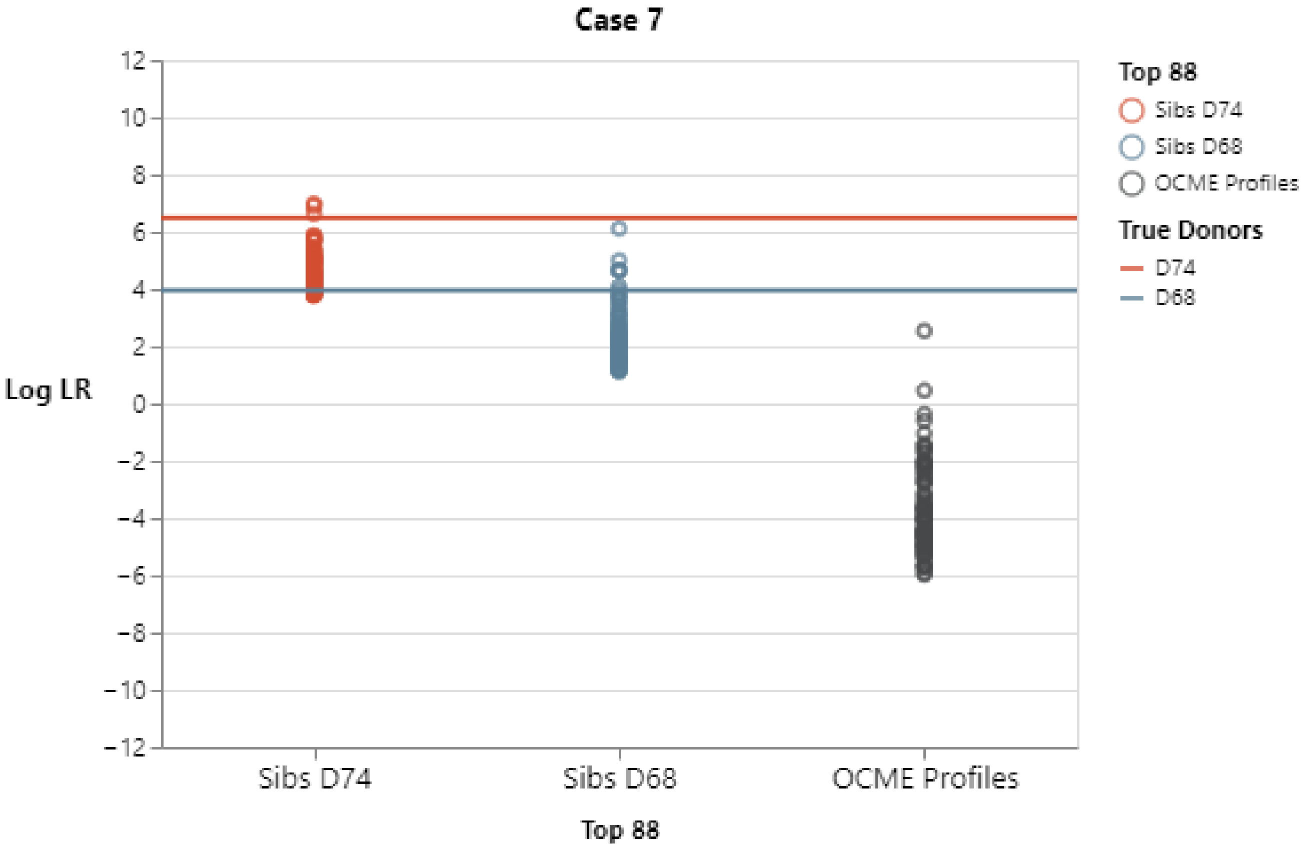
Case 7: 3G-Mix33		
LR Values	**"Major" Profiles**	**"Minor" Profiles**
< .001	15.5%	38.2%
< 1	46%	77.3%
< 1000	86%	98.4%
< True Donors	99.7%	99.4%

The scatter plot below shows the top 88 simulated profiles for Case 7, compared to the 88 OCME profile LRs printed in the FST validation binders.

The "major" D74 simulated set included three profile LRs above that of D74; the highest non-contributor LR was 9.78 million ($\log_{10}LR = 6.99$). All top 88 LRs from the D74-simulated set were above 1000, and most (72 out of 88) were above 10,000.

All of the simulated profiles derived from the "minor" D68 produced LRs above 1, with six non-contributor LRs above that of D68, the highest being 1.29 million ($\log_{10}LR = 6.11$).

Only two of the OCME's top 88 profiles – again, used by the lab to help establish its false positive rate in 2010 – generated LRs above 1 (348 and 2.47, respectively). In other words, for this particular mixture, the profiles used by the OCME in its 2010 false positive study were deceptive when considering a scenario involving high allele sharing.



Discussion

First, it should be noted how deceptive the LRs of unrelated non-contributors are when applied to a case involving high allele sharing, like those involving potential genetic relatives. In its validation studies, the OCME specifically reported as to the FST's ability to achieve separation between true contributors and true non-contributors. However, the OCME's testing in this area failed to explore the space of high allele sharing. Instead, the OCME used a limited number of actual profiles in false positive testing. Our independent exploration of the impact of high allele sharing has revealed that the results of OCME's original limited testing were, at best, misleading.

Second, robust validation and verification requires entities like the OCME to maintain complete records of their software development process, including the results of any internal validation and verification testing. Here, the OCME failed to conform to this generally accepted principle of scientific testing and software development. For example, our independent exploration of the impact of high allele sharing would have benefitted from access to the entire distribution of LRs for each sample used in the false positive study. But the OCME chose only to print and preserve the first two pages of their bulk run sheets, or between 86 and 89 profiles. That kind of arbitrary validation recording can inhibit independent or adversarial testing. Despite these limitations, however, the stark difference of the top 86-89 LRs in all ten of the samples here demonstrates that profiles reflecting high degrees of allele sharing will consistently cause FST to generate inclusionary LRs for non-contributing profiles.

In general, true contributors had higher LRs than that of their simulated siblings, but there were exceptions. One disturbing example is Case 7, a deducible mixture – meaning that an OCME analyst during the validation reported that they could separate out and identify the major contributor's profile. Since true contributor D74 had the highest LR of the three contributors, it stands to reason that D74's profile could be deduced out. However, FST reported three simulated siblings with higher LRs than D74. Why should a deduced profile ever have a lower LR than a non-contributor sibling profile? Software that uses peak height information may have produced a more reliable outcome in that situation.

The minor true contributors had consistently a lower percentage of simulated siblings with LRs above 1. However, it was those same minor contributors whose siblings produced outlier LRs higher than the true contributor – sometimes two, three or four orders of magnitude higher, as seen with D68 in Case 7.

Non-contributors who share high levels of alleles with true contributors are not unicorns. They represent an everyday factual scenario in crime scene investigations. To not account for this scenario in probabilistic genotyping is to invite wrongful convictions of innocent individuals.

For more information or to request the underlying data, please contact:

Clinton Hughes
Forensic DNA Attorney
Brooklyn Defender Services
chughes@bds.org