The discriminatory power of STRmix illustrated by ROC curves

Maarten Kruijver[a], Jo-Anne Bright[a], Hannah Kelly[a], Catherine McGovern [a], James Curran [b], Rebecca Richards [a], Sibéal Waldron [c], Andrew McWhorter [d], Anne Ciecko [e], Brian Peck [f], Chase Baumgartner [g], Christina Buettner [h], Scott McWilliams [h], Claire McKenna [i], Colin Gallacher [j], Ben Mallinder[j], Darren Wright[k], Deven Johnson[l], Dorothy Catella[m], Eugene Lien[n], Craig O'Connor[n], Arlene Petrosky[o], Jason Bundy[p], Jillian Echard[q], John Lowe[r], Joshua Stewart[s], Kathleen Corrado[t], Sheila Gentile[t], Marla Kaplan[u], Michelle Hassler[v], Naomi McDonald[w], Paul Stafford Allen[x], Rachel H. Oefelein[y], Shawn Montpetit[z], Melissa Strong[z], Sarah Noël[aa], Simon Malsom[ab], Kyle Duke[ac], Jessica Skillman[ad], Tamyra Moretti[ae], Teresa McMahon[af], Thomas Grill[ag], Tim Kalafut[ah], MaryMargaret Greer-Ritzheimer[ai], Vickie Beamer[aj], Duncan A. Taylor[ak,al] , John S. Buckleton[a,b]

a. Institute of Environmental Science and Research Limited, Private Bag 92021, Auckland, 1142 New Zealand
b. University of Auckland, Department of Statistics, Auckland, New Zealand
c. Forensic Science, Ireland
d. Texas Department of Public Safety, Houston Laboratory
e. Midwest Regional Forensic Laboratory, Andover, Minnesota
f. Centre of Forensic Sciences, Toronto, Canada
g. Texas Department of Public Safety, Austin Laboratory
h. Wyoming State Crime Laboratory
i. Austin Police Department, City of Austin, Texas
j. Scottish Police Authority (SPA)
k. Idaho State Police Forensic Services
l. Sacramento District Attorney's Office Laboratory of Forensic Services, California
m. Oakland County Sheriff's Office, Michigan
n. New York City Office of Chief Medical Examiner (OCME)
o. Broward Sheriff's Office Crime Laboratory, Florida
p. Florida Department of Law Enforcement
q. Connecticut DESPP Division of Scientific Services
r. Key Forensic Services Ltd., UK, Warrington Laboratory
s. Texas Department of Public Safety, Corpus Christi Laboratory
t. Onondaga County Center for Forensic Sciences, New York
u. Oregon State Police Laboratory (OSP)
v. San Diego County Sheriff's Regional Crime Laboratory
w. Texas Department of Public Safety, Lubbock Laboratory
x. Cellmark Forensic Services, UK
y. DNA Labs International
z. San Diego Police Department Crime Laboratory, California
aa. Laboratoire de sciences judiciaires et de médecine légale (LSJML) Montréal, Canada
ab. Key Forensic Services Ltd., UK, Norwich Laboratory
ac. California Department of Justice Bureau of Forensic Services
ad. Department of Forensic Sciences Laboratory, Washington DC (DFS)
ae. Federal Bureau of Investigation (FBI)
af. Forensic Science Northern Ireland
ag. Erie County Central Services Laboratory, Buffalo, New York
ah. US Army Criminal Investigation Laboratory (USACIL)

ai. DuPage County Sheriff's Crime Laboratory, Illinois
aj. Scottsdale Police Department Crime Laboratory, Arizona
ak. Forensic Science South Australia, GPO Box 2790, Adelaide, South Australia 5001
al. School of Biological Sciences, Flinders University, GPO Box 2100, Adelaide, SA 5001, Australia

Abstract

The use of probabilistic genotyping methods has seen a significant uptake in recent years throughout the world. There is a continuing need for empirical validation of such methods in contexts involving different wet chemistry conditions and various types of (mixed) samples. We have published a large scale empirical validation study in response to the 2016 PCAST report addressing, specifically, the issue that some sample categories were perceived to have received little, to no, attention in the empirical validation literature. More recently, the use of receiver operating characteristic (ROC) analysis was suggested as an important part of such validation exercises. We present the results of ROC analysis for a previously published study. The ROC curves demonstrate the great discriminatory power of DNA demonstrated using the probabilistic genotyping software STRmix™.

Keywords: Forensic DNA analysis; validation; ROC; verbal scale; NIST; probabilistic genotyping; STRmix

Introduction

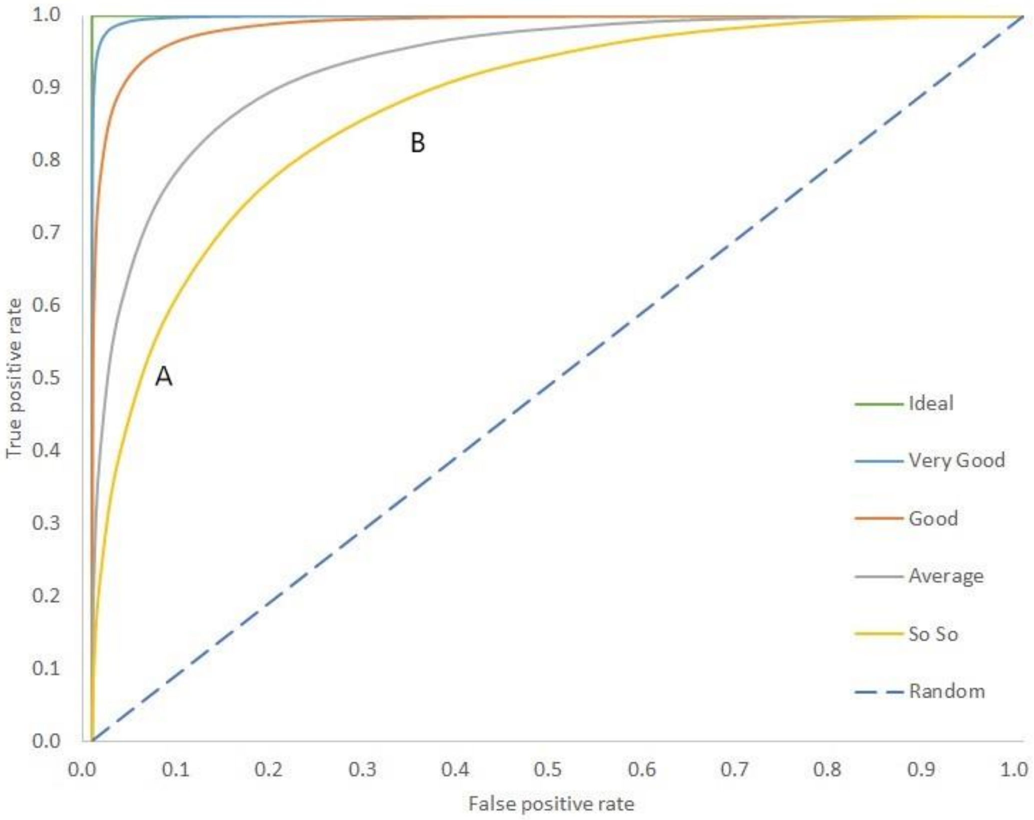
In 2016, the President's Council of Advisors on Science and Technology (PCAST) issued a report [1] and subsequently an addendum [2]. This report discussed a number of forensic disciplines including the interpretation of complex DNA mixtures, defined as any profile with three or more donors. The report noted perceived limits to the proof of validity of the use of probabilistic genotyping (PG) in some situations at the time of publication. PCAST had limited themselves for proof of validity to empirical studies published in the peer reviewed literature. It is often difficult to publish internal validation studies as they are rightly seen as not novel [3]. Two significant publications have been published since the release of the PCAST report. These are:

(1) The Federal Bureau of Investigation Laboratory, Quantico, has published its STRmix™ internal validation in the peer reviewed literature [4]. This publication reports 277 mixtures with two to five donors and a range of mixture ratios and templates. This is in accordance with the SWGDAM guidelines for the validation of probabilistic genotyping systems [5].

(2) Thirty one (31) laboratories who were either using or planning to use STRmix™ published an analysis of mixtures of three, four, five, and six contributors [3].

Subsequent to the PCAST report NIST launched a study which it termed a scientific foundation review (hereafter NIST study) [6]. Butler et al. recently presented "DNA Mixture Interpretation Principles: Insights from the NIST Scientific Foundation Review" [7]. They state that "*We should approach validation of DNA mixture interpretation methods from a performance basis rather than a list of tasks and tests to conduct* …". In the section on performance based validation they call for the use of receiver operating characteristic (ROC) curves and give an example (see Figure 1).
ROC curves are a technique for visualizing the performance of (binary) classifiers. They have been used in, for example, signal detection theory and investigating the behaviour of diagnostic systems. Bleka et al. [8] introduce the use of ROC curves for the assessment of PG methods.

Figure 1. Different ROC curves for different conditions remade following Butler et al. [7]. We have added the points A and B which do not appear in the original.



A method exhibiting an ROC curve indicated by the dashed line in Figure 1 indicates that the method is no better at classifying between two states than a random guess. For example, if the classifier randomly selects one class $p \times 100\%$ of the time it will get $p \times 100\%$ of the true positives correct but will also have a false positive rate of $p \times 100\%$. The point (0, 0) represents the strategy of never issuing a positive classification. Such a classifier commits no false positive errors but also makes no true positive classifications. The opposite strategy, of unconditionally issuing positive classifications, is represented by the upper right point (1, 1). The point (0, 1) represents perfect classification. A point on the ROC graph is better than another if it is above and to the left of it. The point A in Figure 1 is said to be a conservative classifier since it makes few false positive classifications at the expense of fewer true positives. Point B is a liberal classifier since it makes more false positives but obtains more true positives.

The area under each of the ROC curves (AUC) is a measure of the performance of the model. The higher the AUC, the better the model. An AUC of 1 represents a perfect model ("ideal" in Figure 1) and an AUC of 0.5 represents a random performance. The output of a classifier is the assigned class.

In DNA interpretation forensic scientists should not make categorical statements of inclusion [9], but will exclude. In line with the principles of evidence evaluation [10, 11], the forensic scientist should restrict their attention to questions of the kind; "What is the probability of the findings given the propositions?".

Before proceeding we will need to rename the terminology in common use for ROC plots. An *LR* does not give a categorical positive (or negative) but rather is a continuous representation of support for or against a proposition. We can talk about a "true positive" only by assigning some decision threshold, *t*. A true positive would then denote the situation where the person of interest (POI) is indeed a donor to the mixture, and the associated *LR* (contrasting the probability of the evidence under this hypothesis with the probability of the evidence under the hypothesis the POI is not a donor) is greater than some threshold *t*. Therefore, we will rename the true positives as correct support (CS) for H_p where H_p is the event that a certain person is a donor. We rename the false positives as false support (FS) for H_p.

A plot of empirical FS and CS for *LR* thresholds visually summarises the power of discrimination. Such a plot is constructed by computing the FS and CS for all unique *LR* values in the dataset after which the ROC curve follows these values for the ordered *LR*s. Several ROC curves can be compared visually, e.g. for different average peak heights of contributors or to compare the results of interpretation of single source with mixed DNA profiles.

Great care must be taken when applying this approach to avoid giving any impression that thresholds should be applied to *LR*s, and that above some threshold some decision should be taken and below that threshold some different decision should be taken. The *LR* is intended to be combined with other evidence, notably in the form of the prior odds, before any decision is made. The assignment of priors and this combination is ideally done by the fact finders.

Methods

Data description

The data used in this study are described in Bright et al. [3]. Briefly, the data come from 2825 profiles comprising 1591 apparent three, 1136 apparent four, and 98 apparent five person mixtures. These are termed apparent because that is the number of contributors (*NoC*) assigned by a human operator. These mixtures came from 31 different laboratories generated using eight different STR multiplexes and analysed on two different types of capillary electrophoresis (CE) instruments. Each of the 2825 mixtures was deconvoluted and compared to 10,000 random non-donors (sampled according to FBI extended Caucasian allele frequencies [12]) and the true donors. This resulted in 10,297 true donor *LR*s and 28,250,000 non-donor *LR*s.

In order to give an indication of the number of H_d true tests supporting inclusion, the fraction of the 28,250,000 false donor tests that fall in each of the SWGDAM verbal qualifier categories [13] is given in Table 1.

Table 1. Fraction of the 28,250,000 false donor tests from Bright et al. [3] that fall in each of the SWGDAM verbal qualifier categories

LR range	Fraction of false donor *LR*s in this range (N = 28,250,000)
$1 \leq LR < 2$	0.003197
$2 \leq LR < 100$	0.003143
$100 \leq LR < 10^4$	5.53×10^{-5}
$10^4 \leq LR < 10^6$	7.08×10^{-7}
$10^6 \leq LR$	0

ROC

The value for the *LR* is determined by a large number of factors but two variables that explain much of the variance appear to be the number of contributors, *NoC*, and the average peak height, APH. Accordingly, data were divided into groups based on the apparent *NoC* and the APH per contributor. The APH value was assigned to one of four bins: [0, 100), [100, 200), [200, 500), and [500, ∞) where the endpoints of the bins are in relative fluorescence units (rfu). APH was calculated for each contributor by averaging the peak heights of the unmasked alleles; those not shared between contributors and not in back stutter positions of any other contributor alleles. Alleles that had dropped out were assigned a height of half the laboratory's analytical threshold. ROC plots were created in R [14] using the `pROC` package [15].

Results

Three summaries of the data are given in Figures 2 through 4 for the assigned three, four, and five person mixtures. These are scatter plots, violin plots, and ROC plots. In the violin and ROC plots, the subset of APH for each contributor is plotted separately. Scatter plots show individual log(*LR*)s for both H_p and H_d comparisons plotted against APH for the known contributor (or smallest contributor in the case of H_d true comparison). Exclusions ($LR = 0$) are plotted as $\log(LR) = -40$. Within the violin plots, the width of the shaded area represents the proportion of the data located there. Exclusions for non-contributors are not plotted and are represented at the bottom of each plot as the percentage of data. The scatter plots and violin plots are reproduced from Bright et al. [3]. Note to aid in comprehension, the same plotting symbol has been used for all experimental *NoC* in contrast to Bright et al. Within the scatter plots and violin plots it can be seen that *LR*s for true contributors increase as an individual contributor's APH increases. Conversely, generally low *LR*s are obtained for non-contributors trending towards $LR = 1$ as the APH decreases. Within the ROC plots, it can be seen that for each given value of *NoC*, the curves trend to the top left (northwest) with increasing APH. This reflects the fact that as average peak height increases, the evidential value of the stain increases, and this is demonstrated by a reduction in FS and an increase in CS. The decrease in smoothness of the ROC curves with respect to the increase in *NoC* is due to the decreasing number of mixtures involved in each set of comparisons.

Figure 2. Scatter, violin, and ROC plot for apparent three person mixtures. The scatter and violin plots are reproduced (in amended form for the scatter plot) from Bright et al. [3] with permission from Elsevier.

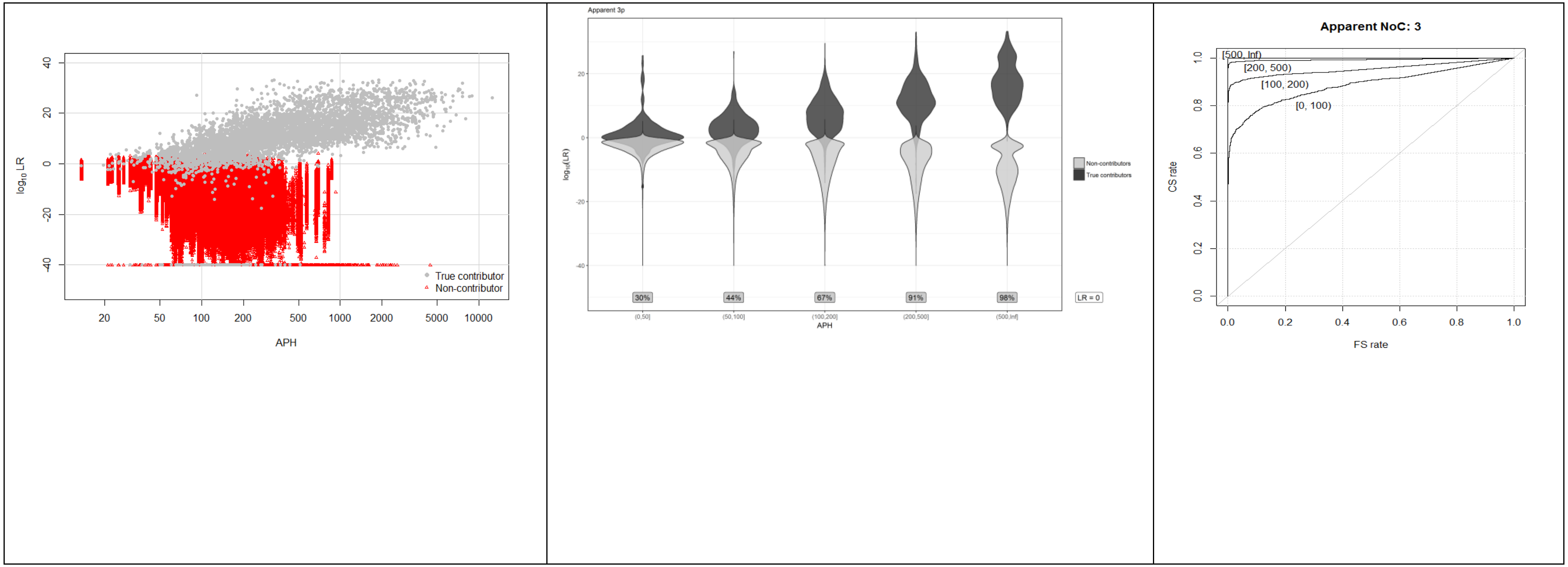


Figure 3. Scatter, violin, and ROC plot for apparent four person mixtures. The scatter and violin plots are reproduced (in amended form for the scatter plot) from Bright et al. [3] with permission from Elsevier.

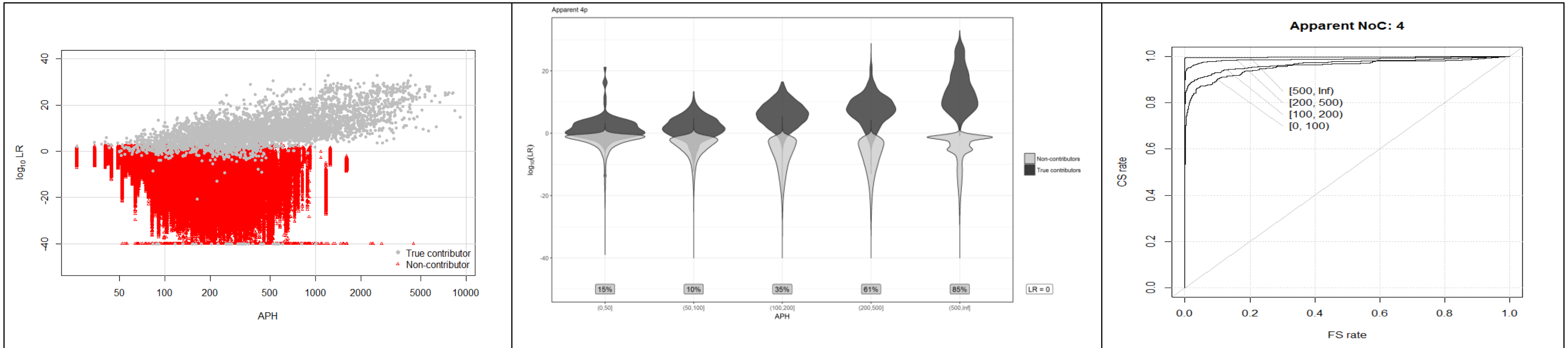
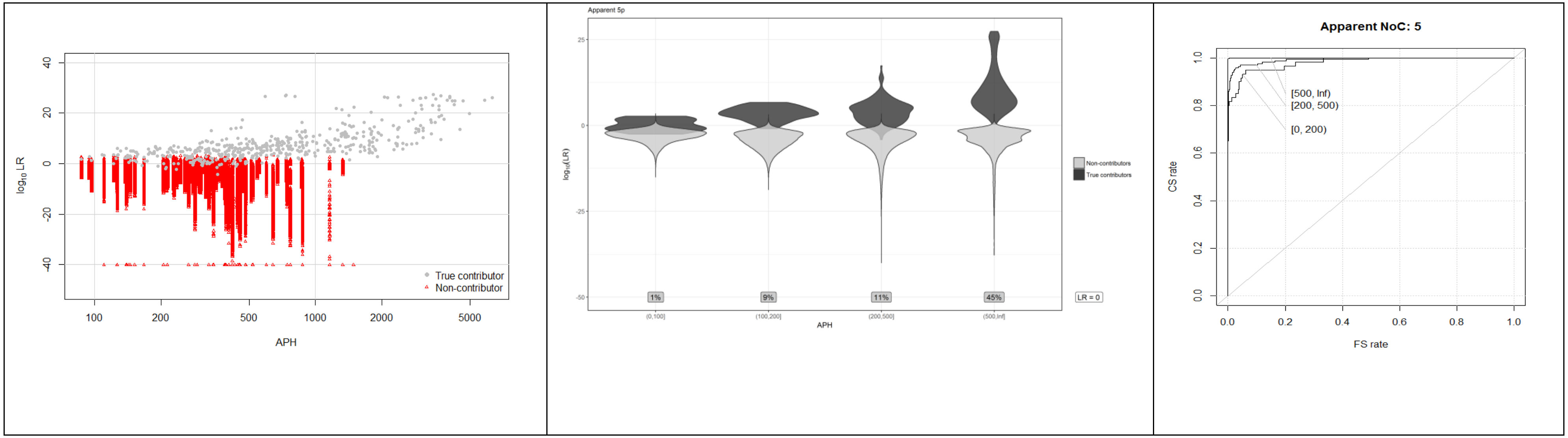


Figure 4. Scatter, violin, and ROC plot for apparent five person mixtures. The scatter and violin plots are reproduced (in amended form for the scatter plot) from Bright et al. [3] with permission from Elsevier.



The area under each of the ROC curves (AUC) is given in Table 2. The standard error [16] is given in parentheses.

Table 2. AUC for each of the ROC curves, standard error in parentheses

	Assigned number of contributors		
APH range	3	4	5
[0,100)	0.8869 (0.0041)	0.9554 (0.0027)	0.9819 (0.0018)
[100,200)	0.9539 (0.0028)	0.9685(0.0023)	
[200,500)	0.9935 (0.0011)	0.9883 (0.0014)	0.9918 (0.0012)
[500, ∞]	0.9997 (0.0002)	0.9974 (0.0007)	0.9999 (0.0001)

Discussion

PCAST argued that the foundational validity of methods that involve at least some subjectivity can only be established through empirical validation. Validation studies published in response to the PCAST report [3, 4] have established foundational validity in the sense that PCAST suggested.

Prior to this work, we published two other styles of graphical summary of the data. These were the violin plots and scatter plots shown in Figures 2 through 4. All summaries, whether as a statistic or a graph, represent a loss of information. The value of a good summary is that it portrays an important aspect of the data with clarity but without undue loss of information. We can confirm from presentations that the violin plots represent a significant challenge to cognition. The scatter plots do, however, appear to be a readily assimilated summary. The work here seeks to consider whether ROC plots add to these summaries.

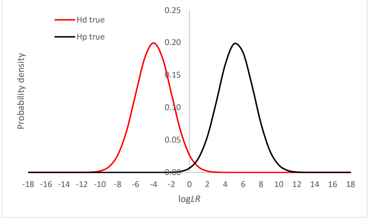
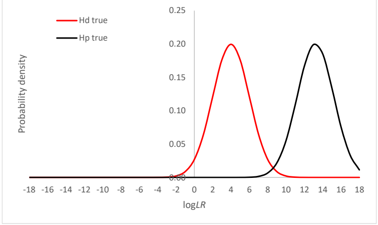
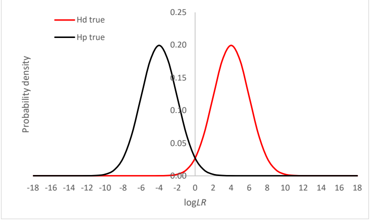
Before proceeding to discuss the ROC plots we reprise the use of the terms discrimination and accuracy when applied to *LR*s. Discrimination in this regard would refer to having different distributions of the *LR* for H_p and H_d true situations. These distributions need only to be different to be discriminating. In Figure 5 we show three pairs of distributions. All of these have the same discrimination. In fact, they would have the same discrimination even if the two distributions in Figure 5a swapped position and hence have low *LR*s for H_p true and high *LR*s for H_d true (the reversed pair Figure 5c) or if any monotone transformation was applied, for example multiplying all *LR*s by 10 or squaring the *LR*.

An alternate way of describing ROC curves is a plot of sensitivity versus 1 – specificity. This interpretation could also be applied to the diagonal and the points (0,0), (0,1), (1,0), (1,1), A, and B in Figure 1. We obviously want to avoid false support for H_p but this comes at the expense of false support for H_d. Figure 2c tells us, for example, that for an apparent three person mixture with APH in the range 0-100 rfu, a negligible risk of false support for H_d (high specificity) requires that we accept a sensitivity of less than about 60%.

Accuracy of an *LR* would be a very different concept from discrimination. Evett warns us, and we follow, not to think of a 'true' *LR*. Accuracy is usually defined as the difference between the assigned value and the true value, and hence we have great difficulty talking about the accuracy of a single *LR* value. We can however make some very strong statements about expected performance. First, no single *LR* for a mixture should exceed the single source *LR* for the person of interest. For this statement to be applied it is necessary to check that the conditioning information when applying the coancestry correction is the same. Then, taken as a group, the average *LR* for the false donors should be 1, or less than 1 if there is deliberate conservatism in the assignments [17, 18]. We expect certain patterns to groups of *LR*s. Both the distribution for the true and false donors should tend to 1 as the information inherent in the profiles is reduced.

Ramos championed calibration of the *LR* to improve performance [19]. Ramos makes some insightful observations, one of which we reprise in Figure 5a through c which show a well calibrated, a poorly calibrated, and a reversed *LR*, respectively. Ramos makes the point, which we have echoed above, that calibration refers to a group of *LR*s not a single *LR*.

The *LR*s shown in Figures 5a through c would have the same ROC plot and AUC (the reversed plot requires inversion of the classification parameter). ROC plots therefore do not inform on accuracy but do inform on discrimination. Where they inform on discrimination only two of the curves have an interpretation in an absolute sense.

<table>
<tr>
<td> </td>
<td> </td>
<td> </td>
</tr>
<tr>
<td>Figure 5a. A well calibrated <i>LR</i></td>
<td>Figure 5b. A poorly calibrated <i>LR</i></td>
<td>Figure 5c. A “reversed” set of <i>LR</i>s</td>
</tr>
</table>

The dashed line in Figure 1 labelled ‘random’ is the situation where the system has no discrimination. The curve labelled ‘ideal’ shows complete discrimination, there is no overlap at all between the two distributions. This is impossible to achieve in DNA profiling analysis. For any curve in between, one curve can be described as more discriminating than another if it lies closer to the 1,1 point. Any of these ‘in between’ curves can be converted to a discrimination if we know the value of the classification parameter to be used. Alternatively, the AUC can be interpreted as the probability that a classifier will rank a randomly chosen true donor higher than a randomly chosen non-donor. However, we neither seek to treat the *LR* as a score nor as a classifier. Rather we seek to use it as an assignment of the weight of evidence. We want that weight of evidence to be meaningful in its own right.

We note that a highly discriminating but poorly calibrated (inaccurate) system would be considered invalid. Whereas a less discriminating but well calibrated system would be considered valid. The ROC therefore cannot inform on validity, albeit that it may supplement other methods for exploratory data analysis.

We summarize the information content in these three styles of graphical summary in Table 3.

Table 3. Summary of information content for three styles of data summaries for *LR*s

Scatter plot	Violin plot	ROC curve
Actual values of the *LR* vs Actual values of APH for H_p and H_d true	Smoothed visualization of the probability density function of the *LR* values for H_p and H_d true	Lose the actual *LR* values but retain the relative rate of values for H_p and H_d true
	APH reduced to categories	APH reduced to categories
Can visualize calibration		Does not inform calibration, focussed on discrimination

One cannot determine the actual *LR* values from inspection of the ROC plots in isolation of the data that sits behind them. Despite the loss of information inherent in the ROC plots there is a clarity to them. Inspection of the ROC plots and AUC values in Table 2 suggests that the discrimination of the four and five person mixtures is greater than the three person mixtures. This would be an incorrect conclusion drawn from the ROC plot and is due to the low APH range of the apparent 3 person mixtures being dominated by false exclusions caused by under assignment of *NoC*. Approximately one quarter of the apparent three person mixtures were actually higher order mixtures where at least one of the known contributors had dropped out. This means comparatively more false exclusions would be obtained for the known contributors than for the apparent four and five person mixtures.

This work is restricted to apparent three, four, and five person mixtures reported by Bright et al. [3]. Further work investigating the effect for two person mixtures would be of interest. We would expect that the same situation would apply; that is, if ground truth *NoC*= 3 and the profile was assigned *NoC*=2 given one contributor being low level then the AUC would be determined almost exclusively from that result. This again would be artifactual and in many ways is the correct result.

We have produced ROC curves, as suggested by [7], as an aid to assessing foundational validity. These curves demonstrate the great discrimination power of STRmix™. However, the ROC curves do obscure valuable information that can be found in other visualisations. In particular, the scatter plots and violin plots give an indication of the expected range of *LR*s given the APH of a contributor, and they can be used by casework analysts when interpreting mixed DNA profiles in order to check the intuitiveness of a result.

Acknowledgements

This work was supported in part by grant 2017-DN-BX-K541 from the US National Institute of Justice. Points of view in this document are those of the authors and are not to be construed as official or as reflecting the views of the US Department of Justice, or the Department of the Army, or the Department of Defense. Names of commercial manufacturers or products included are incidental only, and inclusion does not imply endorsement by the authors, DFSC, OPMG, DA or DOD.

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] 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.
[4] Moretti TR, Just RS, Kehl SC, Willis LE, Buckleton JS, Bright J-A, et al. Internal validation of STRmix for the interpretation of single source and mixed DNA profiles. Forensic Science International: Genetics. 2017;29:126-44.
[5] Scientific Working Group on DNA Analysis Methods (SWGDAM). Guidelines for the Validation of Probabilistic Genotyping Systems. 2015. http://media.wix.com/ugd/4344b0_22776006b67c4a32a5ffc04fe3b56515.pdf Accessed 27 August 2019.
[6] NIST to Assess the Reliability of Forensic Methods for Analyzing DNA Mixtures. 2017. https://www.nist.gov/news-events/news/2017/10/nist-assess-reliability-forensic-methods-analyzing-dna-mixtures Accessed March 21, 2019.
[7] Butler JM, Iyer H, Press R, Taylor M, Vallone PM, Willis S. DNA Mixture Interpretation Principles: Insights from the NIST Scientific Foundation Review. 2018. https://vb6ykw2twb15uf9341ls5n11-wpengine.netdna-ssl.com/wp-content/uploads/2018/07/5.-John-Butler-ISHI-29-Presentation.pdf Accessed 7 November 2018.
[8] Bleka Ø, Benschop CCG, Storvik G, Gill P. A comparative study of qualitative and quantitative models used to interpret complex STR DNA profiles. Forensic Science International: Genetics. 2016;25:85-96.
[9] Scientific Working Group on DNA Analysis Methods (SWGDAM). Interpretation Guidelines for Autosomal STR Typing by Forensic DNA Testing Laboratories. 2017. https://media.wix.com/ugd/4344b0_2a08f65be531488caa8037ed55baf23d.pdf Accessed 17 March 2017.
[10] Evett IW, Weir BS. Interpreting DNA Evidence – Statistical Genetics for Forensic Scientists. Sunderland: Sinauer Associates, Inc.; 1998.
[11] Cook R, Evett IW, Jackson G, Jones PJ, Lambert JA. A model for case assessment and interpretation. Science and Justice. 1998;38:151-6.
[12] Moretti TR, Moreno LI, Smerick JB, Pignone ML, Hizon R, Buckleton JS, 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;25:175-81.
[13] Scientific Working Group on DNA Analysis Methods. Recommendations of the SWGDAM Ad Hoc Working Group on Genotyping Results Reported as Likelihood Ratios. 2018. https://docs.wixstatic.com/ugd/4344b0_dd5221694d1448588dcd0937738c9e46.pdf Accessed 8 November 2018.
[14] R Development Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing, http://www.R-project.org/; 2013.

[15] Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J-C, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics. 2011;12:77.
[16] Hanley JA, McNeil BJ. The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve. Radiology. 1982;143:29-36.
[17] Taylor D, Curran JM, Buckleton J. Importance sampling allows Hd true tests of highly discriminating DNA profiles. Forensic Science International: Genetics. 2017;27:74-81.
[18] Taylor D, Buckleton J, Evett I. Testing likelihood ratios produced from complex DNA profiles. Forensic Science International: Genetics. 2015;16:165-71.
[19] 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.