The paradigm shift in DNA profile interpretation

Jo-Anne Bright[1], Duncan Taylor[2,3], Simone Gittelson[4], and John Buckleton[1,5]

[1] *Institute of Environmental Science and Research Limited, Private Bag 92021 Auckland 1142, New Zealand*

[2] *Forensic Science South Australia, 21 Divett Place, Adelaide, SA 5000, Australia*

[3] *School of Biological Sciences, Flinders University, GPO Box 2100, Adelaide, SA 5001, Australia*

[4] *National Institute of Standards and Technology, 100 Bureau Drive, Gaithersburg, MD 20899-8980, United States*

[5] *University of Washington, Department of Biostatistics, Seattle, WA 98195, United States*

*Corresponding author at: Institute of Environmental Science and Research Limited, Private Bag 92021, Auckland, 1142, New Zealand. Email address: Jo.bright@esr.cri.nz

Abstract

The interpretation of mixtures is moving away from the binary method towards probabilistic genotyping. Previous scholarship noted limitations of the binary method for very low template profiles or profiles with a high ratio for the contributors (high ratio mixtures). By modelling stochastic effects, probabilistic genotyping has overcome many of these limitations. Probabilistic genotyping makes it possible to make statements about the probability of the observed peaks, given different propositions, for mixtures that used to be considered uninterpretable using the binary method. A study of the lowest possible trace contribution to a mixture produces the expected result using probabilistic genotyping: an equal distribution of genotype weights and an average likelihood ratio of one for non-donors. A discussion of validation studies highlights that validation is not about testing every possible combination of variables, but about testing for the expected trends in likelihood ratio values in scenarios with a predefined expectation for these values.

Keywords

Interpretation; Mixtures; Likelihood Ratio; PCAST Report; Probabilistic Genotyping

1. Introduction

DNA mixtures occur when two or more individuals contribute to a sample. Mixtures can vary greatly in complexity. Key variables are the number of donors, the template amount of each donor's DNA, and the level of DNA degradation of each donor. There is little published material that we can find about behaviours relating to the limits of DNA mixture interpretation in the 90s and 00s. From our own personal knowledge, we recall that two-person mixtures were regularly examined in the 90s and that a ratio of 10:1 was considered a reasonable limit beyond which the profile was considered too complex to interpret. Three-person and higher order mixtures were seldom attempted.

It is difficult to find a clear statement of the logic for the 10:1 rule of thumb. Clayton and Buckleton ([1] citing Clayton et al. [2]) state that the alleles of the minor must be above

background noise: "*the threshold represents approximately a 1:10 mixture ratio…*". This would suggest that the 10:1 rule of thumb is consequential on the analytical threshold (AT) and not causal. For a modern 3500 capillary electrophoresis (CE) machine the AT is often of the order of 50-200 rfu. Peaks can be up to 30,000 rfu in height. This suggests that a major contributor's peak heights around 20,000 rfu and a trace contributor's peak heights around 100 rfu would both be in the analysable range. This is a ratio of 200:1.

We cannot find a mention in Clayton et al. [2] of a 10:1 rule. What we did find was "*When the minor component of a mixture falls to about 5:1 or less, there are additional complications in interpretation due to problems in discerning true alleles—present as minor bands—from other artefacts of the system. Conversely, as the ratio of a mixture increases interpreting the major component becomes less complicated…. when the minor component of a high-ratio mixture is under consideration, the evidential significance is often lower ...*".

The binary method for interpreting mixtures was formalised for two-person mixtures [3, 4] into a set of rules for their interpretation. The ISFG DNA Commission papers in 2006 and 2012 also focussed on two-person mixtures [5, 6] but we cannot locate a mention of the 10:1 rule of thumb in them. With hindsight, a ratio higher than 10:1 should not have been a problem even with the manual *LR* based interpretation technologies available at the time. Such mixtures would, however, have been a challenge for combined probability of inclusion (CPI) based methods [7].

By 2011 we were aware of three- and four-person mixtures being interpreted in casework within our own laboratories. Again we can find no reference from a guideline producing body or other authoritative source either sanctioning or discouraging an extension to higher order mixtures.

The recent report and addendum [8] of the President's Council of Advisors on Science and Technology (PCAST), sanctioned the interpretation of two person mixtures at any ratio and three-person mixtures with the minor (as referred to in the report) or person of interest (POI) (as referred to in the addendum) no less than 20%. The difference between the guideline when using the words "minor" or "person of interest" is considerable.

The PCAST report limited itself to validation studies published in the peer reviewed literature, reports in newspapers, and the National Institute of Standards and Technology (NIST) and Innocence Project presentations. PCAST did not explicitly state concern about the use of computer programs to interpret higher order or higher ratio mixtures. In fact, they do appear to countenance the interpretation of higher order and high ratio mixtures once studies are published defining their limits. The requirement for publication in the peer reviewed literature is certainly based on valid considerations but is probably unimplementable, unnecessary and unproductive. We will discuss this in a separate section (section 3).

We speculate that the 10:1 rule slowly embedded itself in the collective consciousness of forensic biologists without the concern being formally articulated or investigated empirically. Taken collectively the 10:1 rule and the PCAST report represent an unformalised unease with mixtures where the component of interest is in very low template or low proportion. This is a concern we also recall from our own pasts.

It is worthwhile exploring the possible causes of the unease with high ratio mixtures. Again this is hampered by a lack of explicit published statements to this effect although John Butler

gives some insight [9] at page 176. For simplicity consider a two-person mixture with one contributor, the major, in vast excess to the other, the trace. The major will make large allelic peaks as well as back stutter peaks. It may also make forward, double back, and some exotic stutter peaks, such as minus 2 bp stutters at SE33. With enhanced sensitivity methods there may also be drop-in peaks. If the trace contributor is of the order of 1% of the total intact DNA then it would be impossible to detect the presence of its alleles were they to fall on the alleles and back stutters of the major. Where these peaks fall on the forward, double back or exotic stutter positions the presence of a trace contributor may be suspected due to higher than expected peak heights but it would be difficult to conclude the presence of a trace contributor with confidence. It is always impossible to determine with absolute certainty if a small peak is from a trace contributor or a drop-in event.

There is a reasonable negative reaction when faced with such data. When considering a proposed set of contributors, say the complainant and person of interest, some peaks that are expected are not present (e.g. drop-out), some peaks that are present are not expected (e.g. drop-in), some that are present and should be are of an unexpected height (e.g. due to stochastic effects).

We will need to use the term very low template or high ratio often so, to ease the flow of this text we will term these 'high ratio'. For the avoidance of uncertainty, we interpret the smallest component where the other components represent most of the intact DNA in the extract.

In the last few years interpretation software have become available that implement a method termed probabilistic genotyping (for example see [10, 11]). These new methods have a remarkable ability to extend the range of mixtures reliably interpreted. This new ability abuts awkwardly with the orthodoxy that there is something dangerous with interpreting high order mixtures and that limits must be set.

In section 2 we will explore the risk of interpreting high ratio mixtures using a probabilistic genotyping software. Section 3 discusses recent issues related to validation, including the PCAST report, four-person mixtures, the effect of replication and known contributors, and the magnitude of validation tests. The discussion and conclusion follow in sections 4 and 5, respectively.

2. High ratio mixtures

2.1. Variability at low height

The nature of forensic work is that the questioned sample is often suboptimal. There is a number of ways that the sample may be compromised: First, there may be very little intact DNA available or second, there may be plenty of DNA but the component of interest is a very small fraction of the total. In both of these situations, the profiles can fairly be termed 'low template'. It is valuable to define two other terms: Low Copy Number and enhanced sensitivity methods. Low Copy Number is a term from the UK used to describe a particular technique based on 34 cycles of polymerase chain reaction (PCR)[1] [4]. The use of the term 'Low Copy Number' is likely to create confusion and it is better not used. Enhanced sensitivity methods include protocols with increased PCR cycles, increased capillary electrophoresis injection time and/or voltage.

[1] The 'standard' number of cycles is usually 28 to 30 for autosomal multiplexes

Low template profiles show increased stochastic effects. This means that allelic and stutter peaks may be bigger or smaller than expected. They may be so small that they cannot be detected, this is termed drop-out. Using enhanced sensitivity methods in essence scales up the expected heights of the peaks but also scales up the differences between the expected and observed peak heights (commonly referred to as peak height variability, or stochastic imbalance). An additional complexity is the appearance of peaks not associated with the DNA extract. For these, Buckleton coined the term drop-in by analogy to drop-out. There is the potential for one or two of these in an enhanced sensitivity profile (e.g., [12]). Drop-in is unobserved or very rare at standard cycle numbers with standard ATs (e.g., [4]).

The absence of peaks expected to be present, the appearance of unexpected peaks, and the variability in the height of those that are there, understandably raise concern. We have heard comments such as "*Low template profiles are unreliable*", "*I would not look at that*" and "*The limit is* 10:1".

One train of thought leading to the incorrect conclusion that probabilistic genotyping analyses of low level mixtures are unreliable might be:

1. low level DNA leads to variability in peak heights and hence non-reproducible profiles,
2. non-reproduciable profiles represent unreliable data,
3. unreliable data analysed by probabilistic genotyping produces unreliable results.

It is worthwhile defining what is meant by an unreliable DNA profile. Vital to this definition is the purpose intended. For example, is wood a reliable fuel? Yes, it is reliable for home heating but it is unreliable for fuelling fighter jets. We will return to this point in the discussion.

Consider another example that illustrates the strength of the inferences that can be made from what may seem like very low quality, or even corrupted, data. Below are the first 42 letters[2] from a book where we have dropped out letters with probability 0.4, dropped in letters with probability 0.005, and changed capitalisation with probability 0.5.

when N O D h.m S. b A, As N ur LiSt

It might be pretty hard to determine the book from which this text originated. You could reasonably say that you could not reliably determine the book from this evidence. But now we ask you could this text come from the beginning of the Origin of Species, the first 42 letters of which are: "When on board H.M.S. Beagle, as naturalist…."? Yes, there is a non-zero probability of observing this configuration of letters if it comes from the beginning of the Origin of Species. In fact, there is a probability of 4×10^{-17} of observing this text if it comes from the beginning of the Origin of Species.

Could this evidence come from the first 42 letters of the Bible which, in the version to hand, are: "In the beginning God created the heavens a(nd the earth)"? Technically, yes, but the amount of corruption required would tend to support the conclusion that it wasn't the source of the text (there is a probability of only 8×10^{-96} of observing this text if it comes from the beginning of the Bible). We could repeat this exercise with hundreds of other first lines but that would rapidly become tedious. Our point is that this highly corrupted evidence still gives a very strong inference in support of the book being the Origin of Species rather than another book. This analogy is good but inexact. There are 26 letters in the alphabet and if we add in punctuation this is more than the number of alleles typically present in the population at any STR locus. In addition we could have mimicked a mixture by taking letters from different books. However we hope the message survives. It is twofold: We need to think of the probability of the evidence if it comes from the person of interest (in our example The Origin

[2] we include punctuation and spaces

of Species) and the probability of the evidence if it comes from someone else. If we do this, we can reliably draw inferences from partially corrupted evidence.

Table 1. A simulated questioned sample from a victim's breast. The alleles corresponding with the victim are bolded. The alleles "not from the victim" have been simulated from a person of interest (POI). With probability 0.4 the allele is dropped out. With probability 0.005 an allele is dropped in. Alleles of the simulated trace that appeared in allelic or back stutter positions of the major profile were removed from the list. Peaks in forward stutter positions of the major are italicised.

Locus	Profile from breast			
	Major profile, corresponds with victim		Trace alleles not corresponding to the victim or back stutter. Peaks in forward stutter positions are italicised	
D8	**11**	**12**	*13*	
D21	**28**	**28**	*29*	30
D7	**11**	**12**		
CSF	**10**	**11**		
D3	**15**	**15**	*16*	
TH01	**7**	**9.3**		
D13	**10**	**13**	*11*	
D16	**9**	**11**	*12*	
D2	**19**	**23**		
D19	**13**	**13**	17	
vWA	**17**	**19**		
TPOX	**9**	**11**	*12*	
D18	**14**	**16**	12	19
D5	**11**	**13**		
FGA	**21**	**22**	25	

Table 1 gives a simulated Identifiler profile with a major and a trace contribution. In making this profile an allele of the trace contributor is dropped out with probability 0.4. With probability 0.005 an allele is dropped in. We ask the question: What is the genotype of the donor of the trace? It would be impossible to answer this question. The alleles labelled in the column "trace alleles not corresponding to the victim or back stutter" may be from the donor or may be drop-in. Because of drop-out the true donor may have many alleles not seen in the profile.

Next look at Table 2. This has two proposed genotypes added for POIs. Could POI 1 be the donor of the trace alleles in the breast profile? One can readily see that this proposal can easily make the trace alleles seen in the mixture.

Table 2. The same profile as shown in Table 1 with two proposed genotypes for persons of interest (POI) given.

Locus	Profile from breast				POI 1: proposed genotype		POI 2: Proposed genotype	
	Major profile, corresponds with victim		trace alleles neither the victim alleles nor back stutter					
D8	**11**	**12**	*13*		13	13	14	15
D21	**28**	**28**	*29*	30	29	30	30	30
D7	**11**	**12**			10	11	10	10
CSF	**10**	**11**			9	13	10	10
D3	**15**	**15**	*16*		14	16	16	16
TH01	**7**	**9.3**			9	9.3	9.3	9.3
D13	**10**	**13**	*11*		11	13	11	11
D16	**9**	**11**	*12*		12	12	11	12
D2	**19**	**23**			17	23	17	18
D19	**13**	**13**	17		13	17	14	16
vWA	**17**	**19**			17	18	14	18
TPOX	**9**	**11**	*12*		9	12	8	8
D18	**14**	**16**	12	19	12	19	12	18
D5	**11**	**13**			11	11	11	12
FGA	**21**	**22**	25		20	25	21	22

Next look at POI 2 in Table 2. This is another proposed genotype. Could POI 2 be the donor of the trace alleles in the breast profile? The answer is yes, just. We would speculate that five of the peaks are large forward stutters and that there are five drop-in events. But realistically it is much easier to make the trace profile from proposed genotype 1 than proposed genotype 2.

This is how probabilistic genotyping works. It assigns a probability for the evidence profile if it comes from various proposed genotypes. Most of the software programs test all realistic genotypes[3] without knowledge of the genotype of the POI. This is true of the probabilistic genotyping software STRmix™, which was used to obtain the results presented in sections 2.2, 3.2 and 3.3 of this paper.

We are not advocating laissez faire. There are definitely limits. There are in fact two types of limits that laboratories using probabilistic genotyping systems will face, one is the limit of the functioning of the software, which will be bound by the models being used[4], the other is a cost-benefit limit whereby the question is not whether an analysis can be done, but rather whether it should be done (see section 1.4 in [13] and section 2.3 in [14]). But we also need to keep up to date. The new probabilistic genotyping software products used to interpret complex DNA

[3] Many software start, at least theoretically, with a list of all possible genotypes. These are sometimes culled to those that are plausible, based on an initial analysis of the profile data or an analysis of those that were visited at least once during burn-in.

[4] For example, a probabilistic system might not have a model for dealing with saturated data, and so the limit would be that any profiles with peaks present at an intensity above a predetermined capillary electrophoresis saturation threshold would be beyond the limits of the software.

profiles are much better at drawing source inferences than previous methods, which were largely ineffective with low template profiles [11, 15, 16].

In this paper we will be dealing with likelihood ratios as the expression of evidential weight and hence the term false inclusion will require some reassessment. Taylor et al. has suggested the term misleading *LR*s [17].

It may be worthwhile discussing what we should expect from DNA interpretation in an ideal world. We need to describe the profile as having information. For example, good template, low contributor number profiles will be described as having high information content. Low template multi-contributor profiles will be described as having low information content. The latter is considered low information content as, even though it likely has more points of data, the information they collectively provide about the genotypes of the donors (compared with a random genotype assignment) is lower than the former. We suggest that we would want a strong indication of inclusion for true donors and a strong indication of exclusion for false contributors. As the information content diminishes we expect the strength of the inference either way to diminish until such a time as the result of the analysis is described as uninformative.

2.2. Lowest possible trace

It is possible to look at mixtures where one contributor is "not there." This describes the situation where we have, say, a ground truth profile known to be single source but we treat it as a two-person mixture. One contributor, therefore, is not there. This is one of the standard tests prescribed by the SWGDAM guidelines for the validation of probabilistic genotyping systems [18].

Before we describe the outcome of this experiment we discuss what the fears might be since this clearly violates the 10:1 rule of thumb. The only realistic cause of concern would be that we would produce a very large inclusionary *LR* for a false donor.

We compared a real single source profile with one manufactured to have all heights at expectation (all stutters at perfect ratio and all heterozygotes in perfect balance). In Table 3 we give the layout of the D8S1179 locus as an example. At this locus the true contributor is a 12,12 homozygote.

Table 3. Peak heights and stutter ratio for a real and manufactured single source sample.

		peak height (rfu)	
		Real	Manufactured
allele	11	269	221
	12	4212	4401
Stutter ratio		6.39%	5.02%

The "perfect" stutter ratio for this allele is estimated by empirical studies is 5.02%. The real sample has a stutter at position 11 that is slightly too large but certainly not so large that it would draw any attention. However, when we treat this as a two-person profile, it is easier to explain the 11 height if there is an allele of the imaginary trace at 11.

In Table 4 are the genotype weights from STRmix™ assuming two contributors.

Table 4. The genotype combinations of the major and imaginary trace, and the weights produced by STRmix™ for the real and manufactured samples. Q refers to any allele other than 11 or 12.

Genotype of the		Real	Manufactured
major	imaginary trace	Weights	
12,12	11,11	0.357	0.152
	11,12	0.219	0.166
	Q,11	0.207	0.167
	Q,Q	0.074	0.175
	12,12	0.073	0.168
	Q,12	0.070	0.172

For the manufactured sample there is no sensible place to put the trace, and hence the software distributes the weights equally. However, for the real sample there is some slight advantage in having a trace with the 11 allele. We see that genotypes for the imaginary trace carrying one or two copies of the 11 allele are preferred (with over 78% of total weight).

Figure 1 shows the $\log_{10}(LR)$ (henceforth we will refer to this as the log(*LR*))for 10,000 simulated false donors tested as the trace against these two samples. The *y*-axis gives the log(*LR*). Positive log(*LR*)s give support for the inclusion of the donor and negative log(*LR*)s give support for the exclusion of the donor. We give the maximum and average *LR* for the 10,000 false donors. The average *LR* is expected to be about one if the system is operating properly [17].

Note that the values in Figure 1 show the power of even very small perturbations from expected peak heights to add inclusionary or exclusionary power to either real or imagined genotypes. This phenomenon has been shown to be true even for highly complex and low level profiles [19]. It is this same ability of probabilistic genotype software to make use of small amounts of fluorescence that also means it can be used for probabilistic analysis of real trace contributors in high ratio mixtures.

Recall that the trace in this experiment is not there. This experiment shows *LR*s for the imaginary trace above 1 and, in the case examined, up to 33. This explains the vertical height of the distribution of false donor tests for very low template contributors that has been observed in all specificity tests done, including [20]. For any real profile, mixed or otherwise, there are some stochastic effects. There is an advantage in having the very low level trace with alleles in those positions. Hence some false donors help explain the profile whilst others actually make it harder to explain. Those that help get a positive log likelihood ratio and those that hinder get a negative one.

In theory it should be possible to calculate the maximum *LR* for any given non-existent contributor but we are aware of no application that does this. We can see from the mathematics that the *LR* for a non-existent contributor is nearest 1 when the mixed or unmixed profile is near perfect balance with regard to peak heights. When there are consistent stochastic effects

at all loci at the extreme edge of expectation then the *LR* for some non-existent contributors would be large.

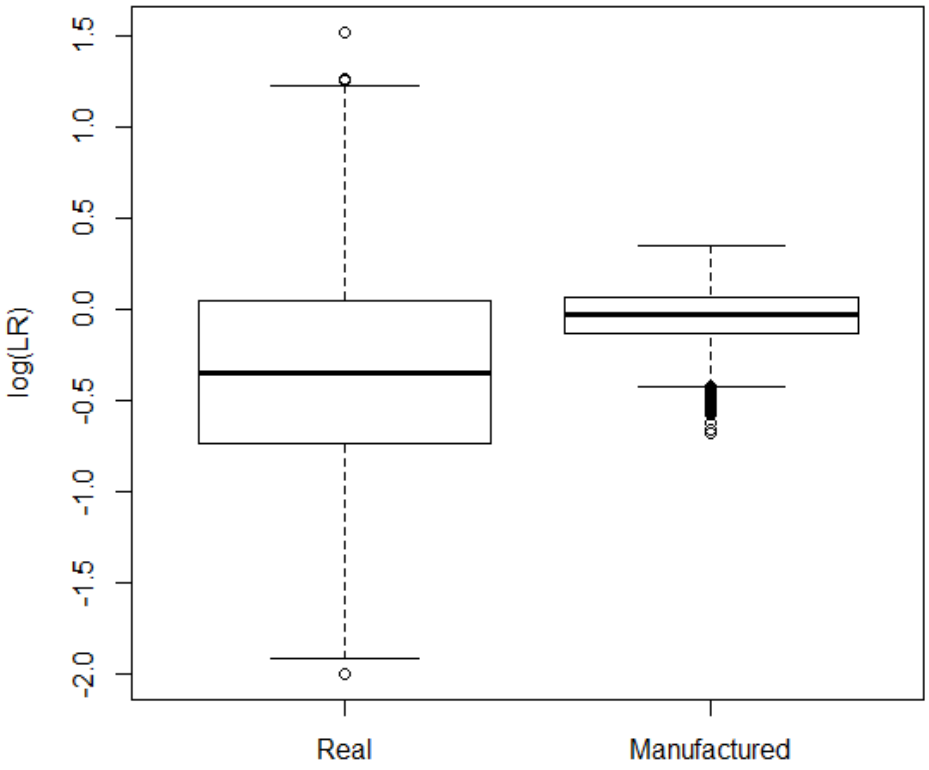


Figure 1. The $\log_{10}(LR)$ for 10,000 simulated false donors tested as the trace against the real and manufactured samples. Maximum values of the *LR* were 33.2 and 2.2 for the real and manufactured respectively. Average values of the *LR* were 0.95 and 0.97 for the real and manufactured respectively.

This would tend to focus attention for one sensible search for limits on high template profiles with poor PCR. We suggest that a targeted approach to searching for limits, such as this, based on key analyses and understanding of the mathematics is likely to be a powerful supplement or replacement for the blind shotgun approach of trying many combinations of genotypes and templates recommended by PCAST [8].

To prove this, we "adjust" a good template profile of a three-person mixture and treat it as a four-person mixture. The mixture in question was produced using GlobalFiler® under manufacturer's instructions. The resulting profile was run on a 3500xl capillary electrophoresis instrument (ThermoFisher). The total amount of DNA was 200 pg, with the contributors added in proportions 3:2:1. When we analysed this as a four-person mixture the resulting probabilistic genotyping analysis gave mixture proportions of 0.54:0.25:0.21:0.00 (the STRmix™ estimate for the proportion of the non-existent trace was 0.0004). The most trace component is not required (in fact we would say it is not there) and the software recognises that fact by giving it a very low proportion. We then added in some imbalance by randomly changing the peak heights of all peaks in the profile by some value uniformly chosen between -50% and 50%. The result was mixture proportions 0.54:0.28:0.18:0.00 (again, where the estimate for the most trace component was 0.0005), still the additional contributor is not required. The original

profile could be described better with the DNA models (by approximately 16 orders of magnitude) and in the imbalanced profile both stutter and allelic peak height variability had to be pushed into the higher tail of their prior distributions. The reason the non-existent contributor was not assigned a higher mixture proportion in the probabilistic genotyping analysis was because the imbalances were added in a stochastic manner, i.e. there was no consistent pattern of contribution across loci that could be explained by a fourth individual. However, we see the same pattern of *LR*s from H_d true testing as was seen in Figure 1, that is the increased imbalances (although not yielding much additional mass to the non-existent contributor i.e. from 0.0004 to 0.0005) utilised the tiny amount of mass they did provide to 'fill in' some fluorescence gaps where it was able and the result was a wider distribution of *LR*s (Figure 2).

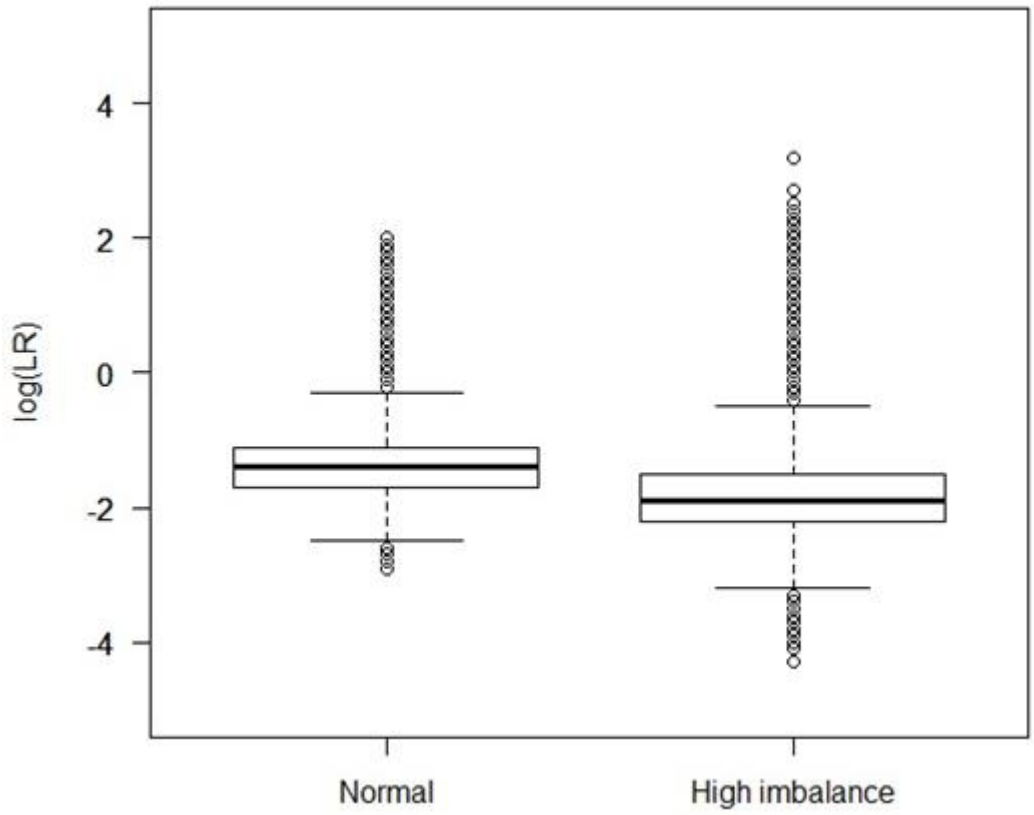


Figure 2. The $\log_{10}(LR)$s for 100,000 simulated false donors tested as the trace against the unaltered profile ("Normal") and the profile with heightened imbalances ("High imbalance").

3. Proof of validity

3.1. PCAST report

PCAST is rightly strongly positive about probabilistic genotyping and sees it as a large improvement over previous methods. They note perceived limits to the current proof of validity. Particularly they highlight gaps regarding high ratio and high contributor number mixtures. This arises because:

1. They have limited themselves for proof of validity to material in the peer reviewed literature[5]. They discount the work presented in a book [22] which, of course, is not peer reviewed. A broader range of sources for matters other than validity is used.
2. They disregard any validation studies using casework samples[6],
3. They consider the only proof to be running a sample of the same type as the questioned sample and observing the result. Subsequent discussions never progressed the definition of how similar the test needed to be to the analysis in question or how the test was scored as either successful or not.

PCAST appears concerned about low template, high ratio, and allelic masking. As discussed above these are key issues from the past. What we have not done is prove to these respected authorities that any concern about this is largely ameliorated by the advent of probabilistic genotyping. This is the paradigm shift in DNA interpretation referenced in the title.

There is actually a very considerable amount of validation data available in the internal validations of laboratories in addition to the developmental software validations. Internal validation summaries are close to unpublishable in journals, as they are considered not novel.[7]

It has never been a requirement to publish the results of internal validation studies in the past. Publication does not ensure validity. There are many examples of published material being subsequently refuted[8].

Equally, the absence of publication does not indicate invalidity. There are a great many ways that validity can be demonstrated. We do not reject the expectation that validity be demonstrated; we simply ask for some achievable criterion.

The Scientific Working Group on DNA Analysis Methods has published comprehensive guidelines for the validation of probabilistic genotyping software [18]. These guidelines accept

[5] Subsequent to the appearance of the PCAST report and addendum the FBI internal validation has been published. PCAST do signal that: "*The range in which foundational validity has been established is likely to grow as adequate evidence for more complex mixtures is obtained and published*." However no mechanism appears to exist to update the recommendations. Emails and phone calls to Eric Lander and PCAST are unanswered at writing. [21] 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.

[6] In a meeting with PCAST members on 18th November 2016, Lander expressed the view that casework samples were not suitable for the empirical work called for by PCAST. We do not discuss the correctness or otherwise of that view here.

[7] At a session of the National Commission on Forensic Science in Washington D.C. on 10th January 2017 chaired by Dr John Butler (see https://www.nist.gov/topics/forensic-science/ncfs-meeting-12-webcast for the webcast, last accessed on May 19, 2017), we heard from Dr Michael Peat the editor of the Journal of Forensic Sciences. Peat stated that the journal's policy was not to publish the results of internal validation studies. Shortly after, at the same session, we heard from Dr Eric Lander, the co-chair of PCAST. Lander, who had not been present for Peat's presentation, described PCAST's requirement for publication. The friction between these forces is producing much heat but little light.

[8] Vulcan was supposed to be a planet somewhere between Mercury and the Sun. Its existence was proposed based on certain peculiarities of Mercury's orbit [23] Lettre de M. Le Verrier à M. Faye sur la théorie de Mercure et sur le mouvement du périhélie de cette planète. Comptes rendus hebdomadaires des séances de l'Académie des sciences (Paris). 1859; vol. 49 379-83.. Einstein's theory of general relativity (1915) explained once and for all why Mercury orbited the Sun in such a strange fashion. This inspired the name of the home planet of the character Spock from *Star Trek*.

the reality that publication is difficult. However, validations are routinely made available for inspection and in some cases have been placed in the public domain. These unpublished internal validations are better than something generic but published. This is because they are specific to the exact methods employed at the laboratory.

3.2. Example of empirical studies

In Figures 3 and 4 we show a plot from a fraction of the data used as part of one validation of STRmix™. These are profiled using GlobalFiler as per manufacturer's instructions and run on two different 3500xl CE machines using a 50 rfu AT. Mixtures were constructed in varying proportions and amplified with varying amounts of template DNA as described in Table 5. Each experimental setup was amplified in duplicate. These contributors differ from those reported in Taylor [20].

Table 5: Mixture setup

Ratio of contributors				Total DNA (pg)
C_1	C_2	C_3	C_4	
10	5	2	1	100, 200, 400, 1000
20	10	10	1	
50	25	10	1	

Profiles were analysed using software STRmix™ V2.4.05 and V2.5.02. In all analyses the non-autosomal loci i.e. the Y-indel locus and DYS391 were ignored. For all calculations, the product rule was used (i.e. no co-ancestry coefficient) and the point estimate has been given. *LR* calculations considered each person within a database of 194 individuals - as a potential contributor, or POI, to the mixed DNA profiles. In doing so there are comparisons to all individuals who are known to have contributed to the DNA profile (when H_p is true) and the remainder, who are known not to have contributed (when H_d is true).

The true donors have been separated into those with a proportion above 10% and those with a proportion below 10%. In Fig. 3, the *x*-axis is template in picograms (that is one millionth of one millionth of a gram). There is a courtroom trend to compare this with the template in a diploid human cell which is about six picograms. The questioning might go:

"This profile is estimated to contain 5 picograms."

"That is less than one cell isn't it?"

It is worth mentioning that any cell membrane has long been disrupted and we are not talking about cells at all by this stage. In the example mentioned of 5 picograms this means that there is about a 43% chance that any allele is not sampled at all, and about a 57% chance that it is sampled at least once.

As the template diminishes the likelihood ratios for both the true and false contributors tends towards one. This is a log(*LR*) of zero (marked with a central horizontal line in the graph). They do not trend to exactly zero. Many graphs have been produced like the one shown.

These exhibit an overlap of *LR*s for donors and non-donors around the range log(*LR*) ±3. This is a result of the small stochastic imbalances described in the section "the lowest possible trace." This is the correct result. It is not an error of the analytical pipeline, but a consequence of natural variation in peak height behavior when the specimen only weakly informs the

underlying genotype. A software that does not show this effect is not using peak height properly.

As the log(*LR*) approaches zero the analyses are losing their informativeness. The overlap in this range is a consequence of the number of false donor tests carried out. Typically, this number is in the hundreds or thousands, and so overlap in the *LR*s is expected to be observed up to a similar level.

As the log(*LR*) gets higher there is still a chance that a false donor could produce such a result. Should we have the computing power to run billions of tests, we would see overlaps approaching log(*LR*) values of nine.

This should not be news to the reader. There has always been a chance that a non-donor could match a profile or mixture by chance. In fact, an *LR* of at least one million will occur from no more than 1 in a million false donors [17]. This is the false inclusion rate of the DNA itself. STRmix™, as a software, has never added to this false inclusion rate (e.g., [17, 21, 24]).

The publication of these data fulfils PCAST's requirement for validation of extremely low template or high ratio mixtures.

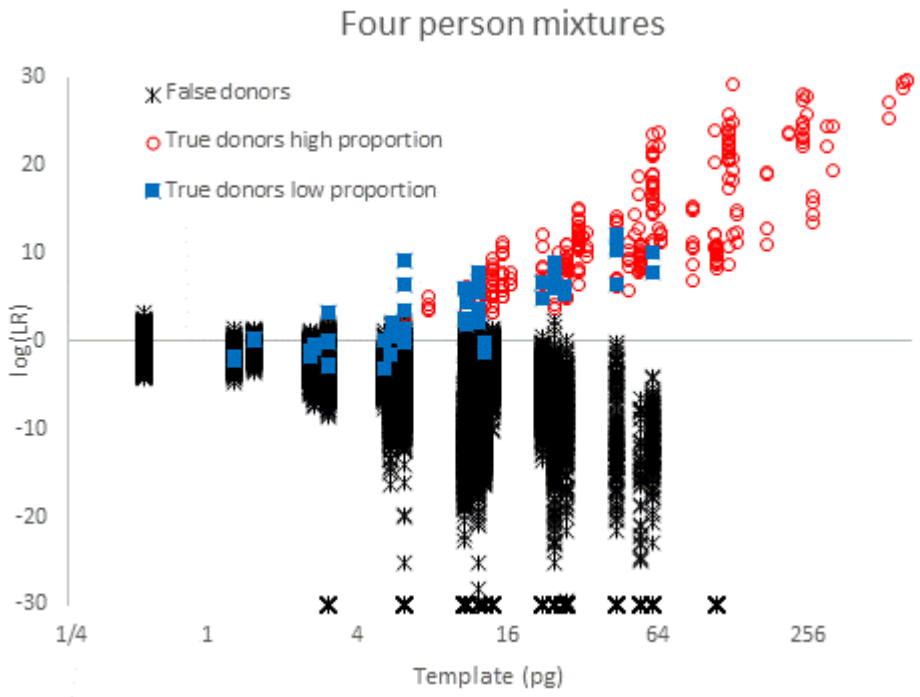


Figure 3. A plot of $\log_{10}(LR)$ vs template for each donor in a four-person mixture. For the false donor tests the template is assigned as the smallest template of the four true donors. For those samples with template above 1 pg, the 194 false donors have been tested against the profile. Due to plotting limitations, samples with a template of 0 are represented in this plot at template 0.5 pg. There are 100,000 false donors tested against a profile where the smallest of 4 donors is not present at all (template = 0 pg). For the true donor tests the data have been divided into proportion above 10% (high) and those with proportion below 10% (low).

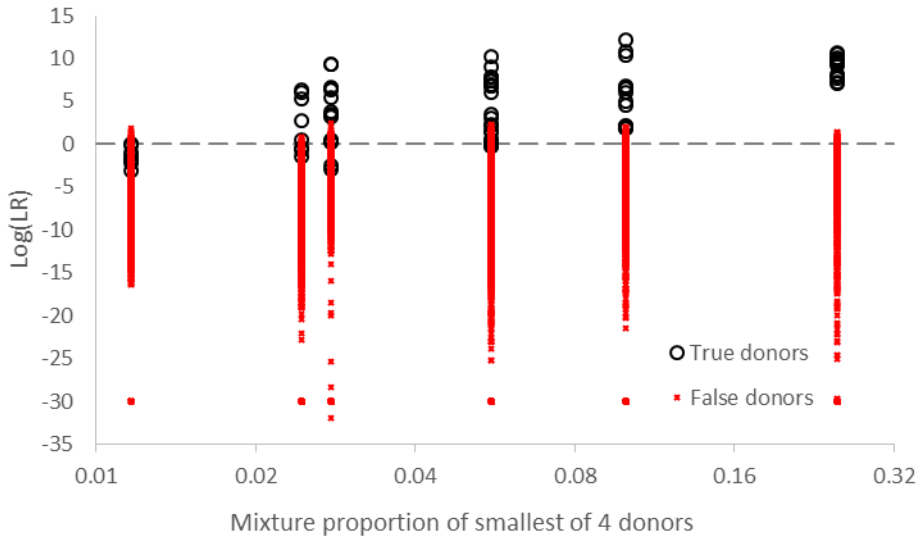


Figure 4. A plot of $\log_{10}(LR)$ vs mixture proportion for the smallest of a set of four person donors. For those with samples with mixture proportion above zero the 194 false donors have been tested against the profile. For the data series plotted at template of zero there are 100,000 false donors tested against a profile where the smallest of four donors is not there at all.

3.3. Power of replication and known contributors

There is a very significant effect of replication and known contributors. This was shown by Taylor [20]. We reproduce a key graphic here (Figure 5.1…5.3).

This demonstrates that replicates or known contributors significantly improve the ability to discriminate true from false donors. This has been known for quite some time although we have trouble finding an early published statement to this effect. The magnitude of the effect came as a pleasant surprise. This would suggest that any attempt to assess the performance of the software should include replication and known contributors as variables along with number of contributors, template and degradation of each contributor.

In Figure 3 it is worth noting that the points that represent the minor contributor in high ratio mixtures (the square points) do not all necessarily sit around $\log(LR) = 0$. In fact, towards the centre and right hand side of the graph (as template increases) the *LR*s produced by these contributors reach and exceed 10 orders of magnitude.

3.4. Magnitude of validation testing

Factors already known to affect the performance of probabilistic genotyping software include template of the POI, number of contributors, replication and the number of known contributors. Mixture proportion of the POI has been highlighted by PCAST and should be included in a validation. The PCAST addendum quotes Butler as noting that it is important to consider samples with different extents of allelic overlap among the contributors[9]. The size of what would be considered an acceptable experimental set-up with regards to the number of samples, the degrees of overlap and the mixture proportions has not been defined. We note that modern multiplexes produce 21 or 24 different combinations because they have that number of loci.

Let us consider four-person mixtures amplified in triplicate at five different template levels. We use only four persons A,B,C and D but we vary the mixture proportion of each contributor in five steps between 0.5 and 0.0. This comes to 1,875 samples.

Let us say we next run these samples for four H_p true analyses and 200 H_d true tests each. This tallies 7500 H_p true and 375,000 H_d true calculations. We have not, as yet varied the genotypes of the contributors. We have received a suggestion from PCAST of 100 different genotype combinations and separately a suggestion that 200 H_d true tests is insufficient.

At this point we still have no criterion that suggests whether these tests have succeeded or failed. This is especially difficult for the H_p true tests. Recall that the output is an *LR*. Imagine we run one of these and obtain an *LR* of 10^8. Is that correct? There are very few circumstances where we can predict the correct *LR* (given the models) [26]. For the remainder we cannot obtain a correct *LR*, or in other valid views, there is no correct *LR* (e.g., [27]). This is a fairly dry venture. We appear to be asked to run thousands of tests taking many thousands of hours and simply record and then publish the result. However dry this is to do it will be worse to read in the unlikely event that it is published.

None of these published tests will be the same as the case in question although, with luck, they may straddle it.

[9] This plausibly harks back to NIST 13 mixture exercise example 5. NIST's experimental design was four people selected from a database of 259. Within the Identifiler profile these four people collectively showed at most 4 alleles per locus. These were mixed in the ratio 1:1:1:1 without degradation resulting in a mixture that is a perfect fit for 3 people (1:2:1) or 4 people (1:1:1:1 or 1:2:1:trace). The highest likelihood solution is 1:2:1:trace. It is impossible to differentiate these solutions without reference to the genotype of the POI. Use of the POI was discouraged by the then extant SWGDAM guidelines [25] Scientific Working Group on DNA Analysis Methods (SWGDAM). SWGDAM Interpretation Guidelines for Autosomal STR Typing by Forensic DNA Testing Laboratories. 2010. The use of the POI is only acceptable in systems that treat the analysis under H_p and H_d completely separately.

The NIST 13 experimental design does not take cognisance that manual assignment of the number of contributors relies on a random selection of genotypes. There are, for example, over 183 million combinations of 4 people in a set of 259. It is from amongst these that this one has been selected. Whilst this tests the limits of the software, it is important that any subsequent error rate is not misrepresented as typical. It is very unlikely that NIST is encouraging use of the genotype of the POI which leaves open the question of the focus of example 5.

Figure 5.1 *LR*s produced for four-person mixtures, with LOWESS lines and polygons showing coverage of scatterplot points

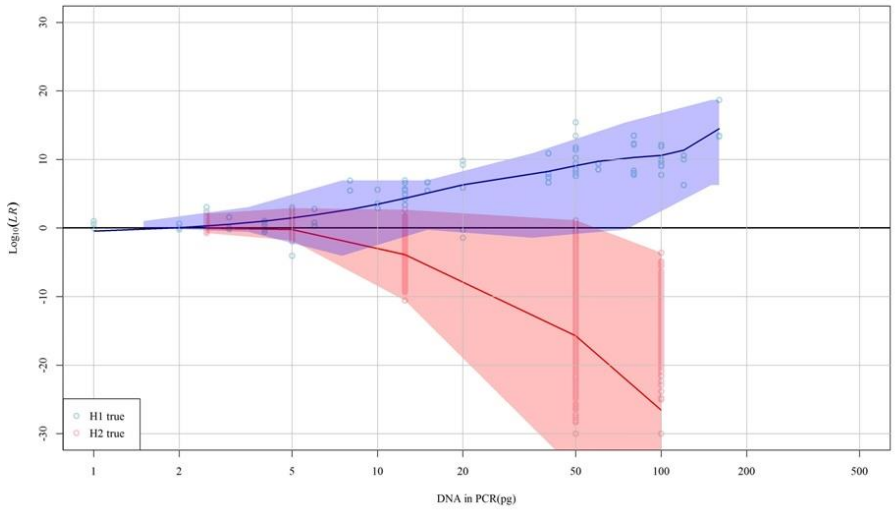


Figure 5.2 *LR*s produced for four-person mixtures using three replicate amplifications, with LOWESS lines and polygons showing coverage of scatterplot points

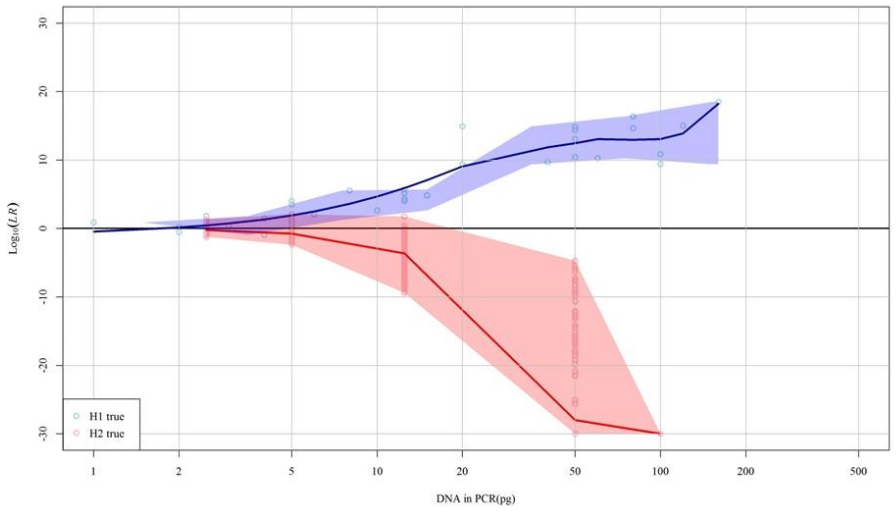
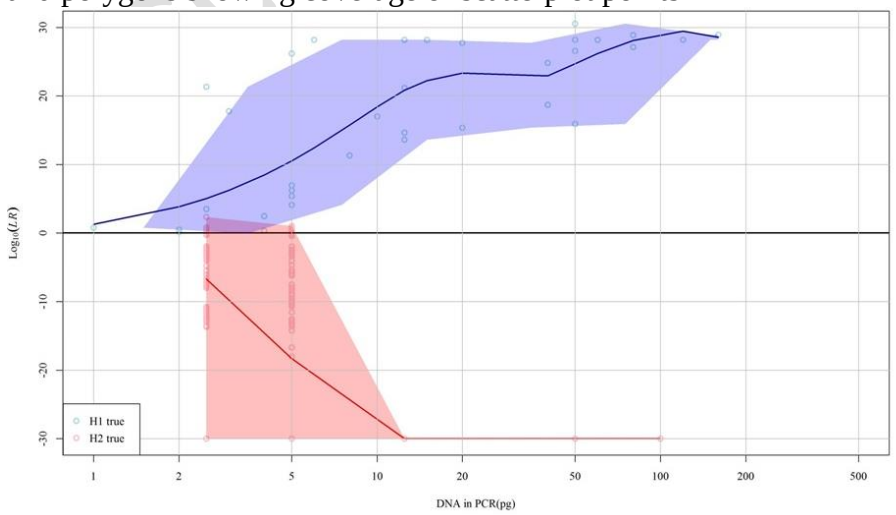


Figure 5.3 *LR*s produced for four-person mixtures using three replicate amplifications and assuming three out of the four known contributors in each analysis, with LOWESS lines and polygons showing coverage of scatterplot points



New interpretation software challenge the way we need to view and discuss validation. Many of the variables are continuous and there are many variables. If we attempt coverage we are calling for very extensive testing. This will need to be repeated for every major change in equipment, chemistry, or a major instrument repair. It is important that we do not erect an unproductive barrier to progress.

The output, whether it is an *LR* or match statistic, is difficult to score for accuracy. This is because of the difficulty or impossibility in determining the correct answer. We can make general comments about *LR* expectations such as:

1. We desire largely *LR*s greater than 1 for H_p true and less than 1 for H_d true.
2. We believe that the average *LR* for a large number of H_d true tests should be 1.
3. We expect the *LR* to tend upwards with increasing template of the POI for H_p true tests,
   3a. We expect poor PCR (extremes within the modelling) to produce low *LR*s even for H_p true tests.
4. We expect the *LR* to tend downwards with increasing template for H_d true tests.
   4a. We expect any individuals producing high *LR*s for H_d true tests to have appropriate alleles for the mixture.
5. We expect *LR*s for H_p and H_d true tests to be nearer 1 for increased contributor number,
6. We expect *LR*s for H_p and H_d true tests to be farther away from 1 for increased known contributors or replicates.

4. Discussion

Earlier we mentioned that we would return to the discussion of what constitutes a reliable DNA profile in the context of probabilistic genotyping systems. Consider the three steps we gave in the train of thought that would lead to an individual believing that complex and low level DNA profiles would not produce reliable results. We agree with the first and last of these statements. Low levels of DNA do produce profiles with a higher level of peak height variability than samples with ample DNA. We also agree that unreliable data analysed by probabilistic genotyping produces unreliable results. It is the crucial middle step with which we disagree. Let us define what is meant by an unreliable DNA profile; an unreliable profile is one that has some features that simply cannot be explained from any of the knowledge we have about DNA profile behaviour. The same definition can be used in probabilistic genotyping; unreliable data would be a DNA profile that has features that are not characterised at analysis or modelled by the software. Most continuous probabilistic genotyping methods have models for saturation, back and forward stutter, DNA template's relation to peak heights, degradation's effect on peak height, locus amplification efficiencies, PCR replicate amplification efficiencies, drop-in, drop-out and peak height variability. The last one is the most relevant to the train of thought above. The peak height variability model should consider that high peaks will have low variability relative to their intensity and low peaks will have relative high variability. If data are produced that have originated from low levels of DNA, and hence produced low peak heights, the software should be able to handle this and consider it appropriately. In other words, as peak heights become smaller, the peak height variability model in the probabilistic genotyping system will consider more possible explanatory genotypes, spreading the probability of the observed data given these genotypes (commonly called weights) across them. This is analogous to what a human would do prior to probabilistic genotyping capabilities, i.e.

they would consider that many explanatory genotypes were possible. The difference with the human interpretation is that upon reaching this conclusion the only course of action in many instances was then to deem the profile too complex for further interpretation. This was not a comment on the reliability of the data produced, but rather the inability of a human to enumerate the complex formulae required to evaluate it.

5. Conclusion

The advent of probabilistic genotyping has greatly extended the range of samples that can be reliably interpreted. It has also changed the key factors limiting interpretation. Old concerns about low template or low mixture proportion seem to be ameliorated by systems that reliably report an *LR* near one when the profile is uninformative. These changes challenge the now outdated, but still prevalent, thinking. These old concerns still appear in courts and guidance producing bodies. The obvious solution is communication. Publication of additional material, much of which already exists, in the peer reviewed literature is constrained by the fair policy that such material is not novel. New methods of communication must be facilitated.

Acknowledgements
This work was supported in part by grant 2011-DN-BX-K541 from the US National Institute of Justice. Points of view in this document are those of the authors and do not necessarily represent the official position or policies of their organisations. The authors would like to thank Johanna Veth, Catherine McGovern, Michael Coble and Steve Lund whose helpful comments improved this paper.

Disclaimer
Certain commercial equipment, instruments, and suppliers are identified in this paper to foster understanding. Such identification does not imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the materials or equipment identified are necessarily the best available for the purpose.

References

[1] Clayton TM, Buckleton JS. Mixtures. Forensic DNA Evidence Interpretation. Boca Raton: CRC Press; 2004. p. 217-74.
[2] Clayton T, Whitaker JP, Sparkes RL, Gill P. Analysis and interpretation of mixed forensic stains using DNA STR profiling. Forensic Science International. 1998;91:55 - 70.
[3] Bill M, Gill P, Curran J, Clayton T, Pinchin R, Healy M, et al. PENDULUM - A guideline based approach to the interpretation of STR mixtures. *For Sci Int*. 2005;148:181-9.
[4] Buckleton JS, Bright J-A, Taylor D. Forensic DNA Evidence Interpretation. 2nd ed. Boca Raton: CRC Press; 2016.
[5] Gill P, Brenner CH, Buckleton JS, Carracedo A, Krawczak M, Mayr WR, et al. DNA commission of the International Society of Forensic Genetics: Recommendations on the interpretation of mixtures. Forensic Science International. 2006;160:90-101.
[6] Gill P, Gusmão L, Haned H, Mayr WR, Morling N, Parson W, et al. DNA commission of the International Society of Forensic Genetics: Recommendations on the evaluation of STR typing results that may include drop-out and/or drop-in using probabilistic methods. Forensic Science International: Genetics. 2012;6:678-88.

[7] Bieber FR, Buckleton JS, Budowle B, Butler JM, Coble MD. Evaluation of forensic DNA mixture evidence: protocol for evaluation, interpretation, and statistical calculations using the combined probability of inclusion. BMC Genetics. 2016;17:125.
[8] President's Council of Advisors on Science and Technology. Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods. 2016.
[9] Butler JM. Advanced topics in forensic DNA typing: Interpretation: Elsevier; 2014.
[10] Taylor D, Bright J-A, Buckleton J. The interpretation of single source and mixed DNA profiles. Forensic Science International: Genetics. 2013;7:516-28.
[11] Perlin MW, Legler MM, Spencer CE, Smith JL, Allan WP, Belrose JL, et al. Validating TrueAllele® DNA mixture interpretation. . Journal of Forensic Sciences. 2011;56:1430-47.
[12] Petricevic S, Whitaker J, Buckleton J, Vintiner S, Patel J, Simon P, et al. Validation and development of interpretation guidelines for low copy number (LCN) DNA profiling in New Zealand using the AmpFlSTR® SGM Plus(TM) multiplex. Forensic Science International: Genetics. 2010;4:305-10.
[13] Taylor D, Buckleton J, Bright J-A. Does the use of probabilistic genotyping change the way we should view sub-threshold data? Australian Journal of Forensic Science. 2017;49:78-92.
[14] Gittelson S, Steffen CR, Coble MD. Low-template DNA: a single DNA analysis or two replicates? Forensic Science International. 2016;264:139-45.
[15] Bille TW, Weitz SM, Coble MD, Buckleton J, Bright J-A. Comparison of the performance of different models for the interpretation of low level mixed DNA profiles. Electrophoresis. 2014;35:3125--33.
[16] Kelly H, Bright J-A, Buckleton J, Curran JM. A comparison of statistical models for the analysis of complex forensic DNA profiles. Science & Justice. 2014;54:66-70.
[17] Taylor D, Buckleton J, Evett I. Testing likelihood ratios produced from complex {DNA} profiles. Forensic Science International: Genetics. 2015;16:165--71.
[18] Scientific Working Group on DNA Analysis Methods (SWGDAM). Guidelines for the Validation of Probabilistic Genotyping Systems. 2015.
[19] Taylor D, Buckleton J. Do low template DNA profiles have useful quantitative data? Forensic Science International: Genetics. 2015;16:13-6.
[20] Taylor D. Using continuous DNA interpretation methods to revisit likelihood ratio behaviour. Forensic Science International: Genetics. 2014;11:144-53.
[21] 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.
[22] Buckleton J, Bright JA, Taylor D. Forensic DNA evidence interpretation. 2nd ed. Florida, USA: CRC Press; 2016.
[23] Lettre de M. Le Verrier à M. Faye sur la théorie de Mercure et sur le mouvement du périhélie de cette planète. Comptes rendus hebdomadaires des séances de l'Académie des sciences (Paris). 1859; vol. 49 379-83.
[24] Bright J-A, Taylor D, McGovern C, Cooper S, Russell L, Abarno D, et al. Developmental validation of STRmix™, expert software for the interpretation of forensic DNA profiles. Forensic Science International: Genetics. 2016;23:226-39.
[25] Scientific Working Group on DNA Analysis Methods (SWGDAM). SWGDAM Interpretation Guidelines for Autosomal STR Typing by Forensic DNA Testing Laboratories. 2010.
[26] 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.14:125-31.
[27] Berger CEH, Slooten K. The LR does not exist. Science & Justice. 2016;56:388-91.