ESR
Science for Communities

Article:

Coble, M.D., Bright, J.A., Buckleton, J.S., & Curran, J.M. (2015). **Uncertainty in the number of contributors in the proposed new CODIS set**. *Forensic Science International: Genetics*, *19*, 207–211.

This is the **Accepted Manuscript** (final version of the article which included reviewers' comments) of the above article published by **Elsevier** at
https://doi.org/10.1016/j.fsigen.2015.07.005

INSTITUTE OF ENVIRONMENTAL SCIENCE AND RESEARCH LIMITED
Kenepuru Science Centre: 34 Kenepuru Drive, Kenepuru, Porirua 5022 | PO Box 50348, Porirua 5240, New Zealand
T: +64 4 914 0700 **F:** +64 4 914 0770
www.esr.cri.nz

Uncertainty in the number of contributors in the proposed new CODIS set

Michael D. Coble[1*], Jo-Anne Bright[2,3], John Buckleton[1,2] and James M. Curran[3]

[1] *National Institute of Standards and Technology, 100 Bureau Drive, Gaithersburg, Maryland, USA 20899*

[2] *ESR, Private Bag 92021, Auckland 1142, New Zealand*

[3] *University of Auckland Department of Statistics, Private Bag 92019, Auckland 1142, New Zealand*

* Corresponding author at: National Institute of Standards and Technology, 100 Bureau Drive MS 8314, Gaithersburg, Maryland, USA 20899-8314 Tel.: +1 (301) 975-4330; fax: +1 (301) 975-8550. E-mail address: mcoble@nist.gov (M.D. Coble).

The probability that multiple contributors are detected within a forensic DNA profile improves as more highly polymorphic loci are analysed. The assignment of the correct number of contributors to a profile is important when interpreting the DNA profiles. In this work we investigate the probability of a mixed DNA profile appearing as having originated from a fewer number of contributors for the African American, Asian, Caucasian and Hispanic US populations. We investigate a range of locus configurations from the proposed new CODIS set. These theoretical calculations are based on allele frequencies only and ignore peak heights. We show that the probability of a higher order mixture (five or six contributors) appearing as having originated from one less individual is high. This probability decreases as the number of loci tested increases.

Keywords: Forensic DNA; mixed DNA profiles; interpretation

Accepted Manuscript

Introduction

DNA profiles are presented as electropherograms (epgs). Each distinct peak at a locus may correspond with an allele. The height of peaks within the epg are measured in relative fluorescent units (RFU) and are approximately proportional to the amount of DNA added to

the PCR reaction. The relative height of peaks from an individual is approximately constant across the profile but tends to decrease in height as the size of the alleles increases [1,2].

Mixed profiles arise when DNA from two or more individuals is present in a DNA extract. One of the first crucial steps of DNA mixture analysis is to not only identify the presence of a mixture, but to identify the number of contributors within the mixture [3]. The probability that a mixture will be detected improves as more highly polymorphic loci are typed. Loci that are highly polymorphic have more allele types and therefore there is an increased chance of seeing DNA from the different contributors to a mixed profile.

Previously, Paoletti et al. [4] investigated the amount of allele sharing between simulated mixtures of four or fewer contributors made from individuals taken from a database typed using the 13 CODIS loci. They reported that within that dataset approximately 3% of three contributor mixtures would have appeared as having originated from two contributors and more than 70% of four contributor mixtures as from two or three contributors. Buckleton et al. [5] undertook a similar exercise for loci within the SGMPlus and ProfilerPlus multiplexes. They reported that 3.3% of three-person mixtures would appear as two contributor profiles based on allele count and 6.2% for ProfilerPlus profiles. Neither work considered peak imbalance which might indicate the presence of more contributors than solely the number of alleles per locus.

In 2010, the CODIS Core Loci Working Group was formed to investigate the expansion of the minimum number of core STR markers tested in the U.S. from the current 13 STR loci. One of the aims was to balance the total number of loci recommended with the level of discrimination offered in order to reduce the likelihood of adventitious matches and in anticipation of more transnational sharing of DNA profile information [6]. The sex test Amelogenin, 18 autosomal STRs and one Y STR are the new minimum recommended STR set with another three autosomal STRs strongly recommended (Table 1) [6,7]. A final list of the 20 autosomal STR loci (excluding SE33 from the set of strongly recommended loci), Amelogenin, and DYS391 that encompass the new core CODIS loci was published by Hares [8].

Table 1: Summary of loci combinations examined

Loci combination	Loci
Existing CODIS	CSF1PO, D13S317, D16S539, D18S51, D21S11, D3S1358, D5S818, D7S820, D8S1179, FGA, TH01, TPOX, vWA
Proposed new CODIS	CSF1PO, D10S1248, D12S391, D13S317, D16S539, D18S51, D19S433, D1S1656, D21S11, D22S1045, D2S1338, D2S441, D3S1358, D5S818, D7S820, D8S1179, FGA, TH01, TPOX, vWA
GlobalFiler	CSF1PO, D10S1248, D12S391, D13S317, D16S539, D18S51, D19S433, D1S1656, D21S11, D22S1045, D2S1338, D2S441, D3S1358, D5S818, D7S820, D8S1179, FGA, SE33, TH01, TPOX, vWA
Fusion	CSF1PO, D10S1248, D12S391, D13S317, D16S539, D18S51, D19S433, D1S1656, D21S11, D22S1045, D2S1338, D2S441, D3S1358, D5S818, D7S820, D8S1179, FGA, Penta D, Penta E, TH01, TPOX, vWA

In response to these efforts, commercial STR multiplex providers have designed larger multiplexes with increased discrimination power. The Applied Biosystems GlobalFiler multi-plex (Life Technologies, Carlsbad CA) amplifies all 22 recommended STRs plus SE33

Accepted Manuscript

and an additional Y-indel locus [9]. The PowerPlex Fusion multiplex (Promega Corp, Madison WI) amplifies all 22 recommended loci plus two highly polymorphic loci Penta D and Penta E [10].

The relative proportion of DNA from the different individuals can be used to resolve their individual profiles from the mixture. However, peak height information is not always used to assist with profile interpretation. Different methods of interpretation use different amounts of profile information.

The combined probability of inclusion (CPI) is a method used to assign the weight of evidence where a probative profile is obtained from an evidentiary sample. The CPI calculation does not make direct use of peak heights, the assumed number of contributors, the genotype of known contributors or the genotype of suspects [11]. Likelihood ratios are the leading alternative to the CPI [12]. Interpretation methods employing likelihood ratios include semi-continuous models (also described as discrete [13] or drop models [14]) and continuous models. Semi continuous models incorporate the probability of drop-out but ignore peak height information whereas continuous models take into account peak heights. A more detailed discussion of the merits of the different types of models is given in Kelly et al. [14] and Steele and Balding [13]. LR methods require the assignment of the number of contributors to a profile. In this paper we investigate the probability of an N contributor mixture appearing as originating from a fewer number of contributors for the African American, Asian, Caucasian and Hispanic US populations across a range of locus configurations. We investigate the probability of allele sharing across the existing CODIS and proposed new CODIS loci, Life Technologies GlobalFiler, and Promega's Fusion multiplex. A similar study has recently been published by Curran and Buckleton for the European standard set (ESS) [15]. These theoretical calculations are based on allele frequencies only and ignore peak heights.

Methods

The probability of an N contributor mixture masking as a k person mixture, where $k = 1, \dots, N-1$ for the four major US subpopulations (African American, Asian, Caucasian and Hispanic) was calculated for values of N ranging from 2 up to 6. For example, in a 6 person mixture, the probability of masking as a 5, 4, 3, 2, and 1 (single source) was calculated. The method followed is the same as that described in Buckleton et al. [5]. It is possible, when N=2, to calculate these probabilities exactly at a single locus, and then the multi-locus probability can be calculated under the assumption of linkage equilibrium (LE) by simple multiplication of the individual locus probabilities. However this task rapidly becomes impossible for N>2 because the number of genotype combinations increases significantly. Therefore we take a Monte Carlo approach whereby we randomly simulate a large number (up to 10 billion iterations) of sets of N genotypes and count the number of unique alleles at each locus. The multilocus probabilities are then computed by assuming LE and taking the product as before. The allele frequencies used in the simulation were published by Hill et al. [16] and are available at http://www.cstl.nist.gov/div831/strbase/NISTpopdata/NIST-US1036-AlleleFrequencies.xlsx.

This method lets us assess the probability of an N person mixture masking as a k person mixture, where $k = 1, \dots, N-1$. DYS391 was ignored for all locus combinations.

Tvedebrink has demonstrated a method to calculate the exact distribution of the number of alleles for any number of contributors [17]. The method was attempted for this work however it was not successful due to the large numbers of contributors and loci being investigated.

Accepted Manuscript

It is important to note that these simulations have ignored peak height information again because it is difficult to construct realistic scenarios from which generalizations may be drawn. We appeal, again, to the argument made in Curran and Buckleton [15], that a set of *N* people all contributing equally is, in some sense, the worst- case scenario. The simulations also ignore drop-out which is not realistic with higher order mixtures. However, ignoring heights would only serve to increase the probability of a higher order mixture presenting as one with fewer contributors. The loci combinations analysed are provided in Table 1.

Results

Tables 2–5 present the cumulative probability of an N person mixture appearing as a k or fewer person mixture where $k = 1, \dots, N-1$ for the four major U.S. subpopulations: African American, Caucasian, Hispanic, and Asian. Within these tables the number of loci per configuration is in parentheses. The probability of a higher order mixture (five or six contributors) appearing as having originated from one or two individuals based on the allele count alone is very small for any of the locus configurations. Values less than 1.0×10^{-4} are presented in scientific notation. For example, within the African American population the probability of a six contributor profile appearing as a two contributor profile is 1.81×10^{-9} using the existing CODIS loci and 4.51×10^{-21} using the proposed new CODIS set plus (GlobalFiler). The probability decreases as the number of loci tested increase which is the expected result. The probability of a higher order mixture appearing as having originated from one fewer individual is high. For example, within the Caucasian population the probability of a six contributor profile appearing as five or fewer contributors based on allele count is 0.9999 using the existing CODIS set dropping to 0.8599 with the new proposed CODIS set plus SE33.

Discussion

The results of Tables 2–5 show that the additional loci available in the new 'megaplex' kit configurations reduce the probability of incorrectly identifying the true number of contributors compared with the existing CODIS 13 set of loci. We observed no major departures in the probabilities among the 4 major U.S. population groups tested in this study. When the complexity of the mixture reaches 5 and 6 contributors, the benefits of additional loci are apparent. It should be acknowledged however that the probative value of a 5 or 6 contributor profile is not likely to be high [18].

The inclusion of SE33 in the proposed new CODIS locus set significantly decreases the risk of allele masking compared with when it is absent. SE33 is known to be a very discriminatory locus [19].

These simulations look at the presence of alleles only, ignoring peak height. We also ignore the possibility of allele masking by stutter peaks when a low-level minor contributor's alleles are below the stutter threshold of the major contributor alleles. Masking by stutter can also lead to an underestimation of the number of contributors in complex mixtures of greater than two individuals. Peaks heights can reliably be used for assigning the number of contributors to a profile [20,21] even at low template [22] and therefore we expect these probabilities will be lower when this information is taken into account. We have demonstrated when assigning the number of contributors to a mixed DNA profile there is a risk of understating that number if only relying on allele count. This risk decreases when using more loci but increases for higher order mixtures. This supports the previous findings by Curran and Buckleton [15] who also reported that the risk is increased if allelic drop-out is a possibility.

Accepted Manuscript

Table 2: Cumulative probability of an *N* person mixture appearing as a *k* or fewer person mixture, where $k = 1, \ldots, N-1$ for the African American allele frequencies

Configuration (# loci)	*N* contributor mixture	Appearing as *k* or fewer				
		1	2	3	4	5
Existing CODIS (13)	6	1.53E-41	1.81E-09	0.0822	0.8740	0.9993
GlobalFiler (21)		1.95E-76	4.51E-21	0.0001	0.3301	0.9384
Proposed CODIS (20)		1.61E-69	7.75E-18	0.0022	0.6962	0.9982
Fusion (22)		9.26E-81	3.81E-22	0.0002	0.5220	0.9937
Existing CODIS (13)	5	1.16E-33	7.74E-07	0.2872	0.9736	
GlobalFiler (21)		6.91E-62	4.09E-15	0.0085	0.743	
Proposed CODIS (20)		2.79E-56	9.47E-13	0.0466	0.9307	
Fusion (22)		2.34E-65	7.94E-16	0.0115	0.8724	
Existing CODIS (13)	4	1.33E-25	0.0003	0.6969		
GlobalFiler (21)		5.41E-47	2.43E-09	0.2156		
Proposed CODIS (20)		9.91E-43	7.58E-08	0.4079		
Fusion (22)		1.37E-49	1.02E-09	0.2622		
Existing CODIS (13)	3	2.62E-17	0.0430			
GlobalFiler (21)		1.20E-31	0.0004			
Proposed CODIS (20)		8.97E-29	0.0017			
Fusion (22)		2.38E-33	0.0003			
Existing CODIS (13)	2	8.63E-09				
GlobalFiler (21)		7.29E-16				
Proposed CODIS (20)		2.01E-14				
Fusion (22)		1.17E-16				

Accepted Manuscript

Table 3: Cumulative probability of an *N* person mixture appearing as a *k* or fewer person mixture, where $k = 1, \ldots, N-1$ for the Caucasian allele frequencies

Configuration (# loci)	*N* contributor mixture	Appearing as *k* or fewer				
		1	2	3	4	5
Existing CODIS (13)	6	2.13E-40	6.30E-09	0.1612	0.9456	0.9999
GlobalFiler (21)		5.40E-75	9.11E-21	5.31E-05	0.1882	0.8599
Proposed CODIS (20)		1.18E-66	1.66E-16	0.0038	0.6798	0.9969
Fusion (22)		1.51E-75	2.11E-19	0.0009	0.5758	0.9941
Existing CODIS (13)	5	9.66E-33	2.10E-06	0.4141	0.9897	
GlobalFiler (21)		8.21E-61	7.10E-15	0.0048	0.6099	
Proposed CODIS (20)		5.82E-54	9.10E-12	0.0592	0.9228	
Fusion (22)		3.57E-61	7.80E-14	0.027	0.8885	
Existing CODIS (13)	4	6.82E-25	0.0005	0.7856		
GlobalFiler (21)		4.15E-46	3.50E-09	0.1653		
Proposed CODIS (20)		6.13E-41	3.10E-07	0.4323		
Fusion (22)		2.05E-46	1.80E-08	0.3369		
Existing CODIS (13)	3	8.40E-17	0.0595			
GlobalFiler (21)		5.83E-31	0.0004			
Proposed CODIS (20)		1.66E-27	0.0031			
Fusion (22)		3.52E-31	0.001			
Existing CODIS (13)	2	1.70E-08				
GlobalFiler (21)		2.10E-15				
Proposed CODIS (20)		1.00E-13				
Fusion (22)		1.60E-15				

Accepted Manuscript

Table 4: Cumulative probability of an *N* person mixture appearing as a *k* or fewer person mixture, where $k = 1, \ldots, N-1$ for the Hispanic allele frequencies

Configuration (# loci)	*N* contributor mixture	Appearing as *k* or fewer				
		1	2	3	4	5
Existing CODIS (13)	6	3.38E-42	7.10E-10	0.0833	0.9013	0.9997
GlobalFiler (21)		3.96E-74	2.18E-20	0.0001	0.2726	0.9074
Proposed CODIS (20)		2.55E-66	1.27E-16	0.004	0.7294	0.9982
Fusion (22)		1.61E-76	2.90E-20	0.0004	0.5275	0.9881
Existing CODIS (13)	5	2.89E-34	4.39E-07	0.2949	0.9804	
GlobalFiler (21)		4.35E-60	1.34E-14	0.0077	0.6911	
Proposed CODIS (20)		8.54E-54	7.37E-12	0.0621	0.939	
Fusion (22)		5.14E-62	1.85E-14	0.0191	0.8657	
Existing CODIS (13)	4	4.34E-26	0.0002	0.7093		
GlobalFiler (21)		1.19E-45	5.16E-09	0.2026		
Proposed CODIS (20)		6.76E-41	2.74E-07	0.4451		
Fusion (22)		4.18E-47	7.18E-09	0.3008		
Existing CODIS (13)	3	1.31E-17	0.0405			
GlobalFiler (21)		1.02E-30	0.0005			
Proposed CODIS (20)		1.56E-27	0.003			
Fusion (22)		1.11E-31	0.0007			
Existing CODIS (13)	2	6.91E-09				
GlobalFiler (21)		2.47E-15				
Proposed CODIS (20)		9.28E-14				
Fusion (22)		8.62E-16				

Accepted Manuscript

Table 5: Cumulative probability of an N person mixture appearing as a k or fewer person mixture, where $k = 1, \ldots, N-1$ for the Asian allele frequencies

Configuration (# loci)	N contributor mixture	Appearing as k or fewer				
		1	2	3	4	5
Existing CODIS (13)	6	4.41E-38	7.17E-08	0.2499	0.97	0.9999
GlobalFiler (21)		2.99E-69	4.97E-17	0.0012	0.3611	0.9278
Proposed CODIS (20)		2.62E-61	3.38E-13	0.0436	0.9063	0.9998
Fusion (22)		7.62E-71	1.51E-16	0.0043	0.5838	0.9813
Existing CODIS (13)	5	7.08E-31	1.14E-05	0.5195	0.9946	
GlobalFiler (21)		4.35E-56	3.38E-12	0.0286	0.7468	
Proposed CODIS (20)		1.18E-49	2.02E-09	0.2208	0.9823	
Fusion (22)		2.39E-57	8.01E-12	0.0607	0.8745	
Existing CODIS (13)	4	1.74E-23	0.0014	0.8403		
GlobalFiler (21)		1.51E-42	1.54E-07	0.3113		
Proposed CODIS (20)		1.06E-37	8.30E-06	0.6607		
Fusion (22)		1.61E-43	2.80E-07	0.4218		
Existing CODIS (13)	3	7.19E-16	0.0874			
GlobalFiler (21)		1.34E-28	0.0021			
Proposed CODIS (20)		2.27E-25	0.0119			
Fusion (22)		2.91E-29	0.0029			
Existing CODIS (13)	2	4.64E-08				
GlobalFiler (21)		2.80E-14				
Proposed CODIS (20)		1.06E-12				
Fusion (22)		1.30E-14				

Accepted Manuscript

Acknowledgements

The authors would like to acknowledge Dr. John Butler (NIST) for helpful discussions and the suggestions from two anonymous reviewers. This work was supported in part by grant 2011-DN-BX-K541 from the U.S. 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 the U.S. Department of Justice or Department of Commerce. Certain commercial equipment, instruments and materials are identified in order to specify experimental procedures as completely as possible. In no case does such identification imply a recommendation or endorsement by the National Institute of Standards and Technology nor does it imply that any of the materials, instruments or equipment identified are necessarily the best available for the purpose.

References

1. Bright, J.-A., D. Taylor, J. Curran, and J. Buckleton, Degradation of forensic DNA profiles. Australian Journal of Forensic Sciences, (2013). 45(4): 445-449.
2. Tvedebrink, T., P.S. Eriksen, H.S. Mogensen, and N. Morling, Statistical model for degraded DNA samples and adjusted probabilities for allelic drop-out. Forensic Science International: Genetics, (2012). 6(1): 97-101.
3. Clayton, T., J.P. Whitaker, R.L. Sparkes, and P. Gill, Analysis and interpretation of mixed forensic stains using DNA STR profiling. Forensic Science International, (1998). 91: 55 - 70.
4. Paoletti, D.R., T.E. Doom, C.M. Krane, M.L. Raymer, and D.E. Krane, Empirical analysis of the STR profiles resulting from conceptual mixtures. Journal of Forensic Sciences . (2005). 50: 1361-1366.
5. Buckleton, J.S., J.M. Curran, and P. Gill, Towards understanding the effect of uncertainty in the number of contributors to DNA stains. Forensic Science International: Genetics, (2007). 1(1): 20-28.
6. Hares, D.R., Expanding the CODIS core loci in the United States. Forensic Science International: Genetics, (2012). 6(1): e52-e54.
7. Hares, D.R., Addendum to expanding the CODIS core loci in the United States. Forensic Science International: Genetics, (2012). 6(5).
8. Schellberg, T., N. Oldroyd, and L.L. Schade, Maximizing the power of forensic DNA databases with next generation STR technology. Forensic magazine, (2012). http://www.forensicmag.com/articles/2012/10/maximizing-power-forensic-dna-databases-next-generation-str-technology.
9. Promega Corporation. PowerPlex® Fusion System. 2013 [cited 27 September 2013]; Available from: http://worldwide.promega.com/products/pm/genetic-identity/powerplex-fusion/?origUrl=http%3a%2f%2fwww.promega.com%2fproducts%2fpm%2fgenetic-identity%2fpowerplex-fusion%2f.
10. Bille, T.W., J.A. Bright, and J.S. Buckleton, Application of random match probability calculations to mixed STR profiles. Journal of Forensic Sciences, (2013). 52(2): 474-485.
11. Gill, P., C.H. Brenner, J.S. Buckleton, A. Carracedo, M. Krawczak, W.R. Mayr, et al., DNA commission of the International Society of Forensic Genetics: Recommendations on the interpretation of mixtures. Forensic Science International, (2006). 160(2–3): 90-101.
12. Steele, C.D. and D.J. Balding, Statistical evaluation of forensic DNA profile evidence. Annual Review of Statistics and Its Application, (2014). 1: 20.1-20.24.
13. Kelly, H., J.-A. Bright, J.S. Buckleton, and J.M. Curran, A comparison of statistical models for the analysis of complex forensic DNA profiles. Science & Justice, (2014). 54(1): 66-70.
14. Curran, J.M. and J. Buckleton, Uncertainty in the number of contributors for the European Standard Set of loci. Forensic Science International: Genetics, (2014). 11: 205-206.

Accepted Manuscript

15. Hill, C.R., D.L. Duewer, M.C. Kline, M.D. Coble, and J.M. Butler, U.S. population data for 29 autosomal STR loci. Forensic Science International: Genetics, (2013). 7(3): e82-e83.
16. Tvedebrink, T., On the exact distribution of the numbers of alleles in DNA mixtures. Forensic Science International: Genetics Supplement Series, (2013). 4(1): e278-e279.
17. Taylor, D., Using continuous DNA interpretation methods to revisit likelihood ratio behaviour. Forensic Science International: Genetics, (2014). 11: 144-153.
18. Reid, T.M. and e. al., Distribution of HUMACTBP2 (SE33) alleles in three North American populations. Journal of Forensic Science, (2003). 48(6): 1422-1423.
19. Puch-Solis, R., L. Rodgers, A. Mazumder, S. Pope, I. Evett, J. Curran, et al., Evaluating forensic DNA profiles using peak heights, allowing for multiple donors, allelic dropout and stutters. Forensic Science International: Genetics, (2013). 7(5): 555-563.
20. Bright, J.-A., D. Taylor, J.M. Curran, and J.S. Buckleton, Developing allelic and stutter peak height models for a continuous method of DNA interpretation. Forensic Science International: Genetics, (2013). 7(2): 296-304.
21. Taylor, D. and J.S. Buckleton, Do low template DNA profiles have useful quantitative data? Forensic Science International: Genetics, (2015). 16: 13-16.

Accepted Manuscript