Estimation of the number of contributors of theoretical mixture profiles based on allele counting: Does increasing the number of loci increase success rate of estimates?

Authors: Gina M. Dembinski[a], Carl Sobieralski[b], Christine J. Picard[a,c]

[a] Department of Biology, Indiana University-Purdue Un Indianapolis, 723 West Michigan Street, Indianapolis, IN, 46202, U.S.A.

[b] Indiana State Police Laboratory, 550 West 16th Street, Suite C, Indianapolis, IN, 46202, U.S.A.

[c] Forensic and Investigative Sciences, Indiana University Purdue University Indianapolis, 402 N. Blackford Street, Indianapolis, IN, 46202, U.S.A.

gdembins@iupui.edu, CSobieralski@isp.in.gov, cpicard@iupui.edu

Corresponding Author: Christine J. Picard, Department of Biology, 723 W. Michigan Street, Indiana University-Purdue University Indianapolis (IUPUI), Indianapolis, Indiana 46202, cpicard@iupui.edu

Highlights

- This study evaluated maximum allele count of mixtures with expanded U.S. core loci
- There was no significant improvement as compared to previous U.S. core loci panel
- Maximum allele count was accurate for two person male mixtures based on 3 Y-STRs
- The maximum allele count method is not reliable beyond three person mixtures

Abstract

DNA mixtures are more frequently encountered in casework due to increased kit sensitivity, protocols with increased cycle number, and requests for low copy number DNA samples to be tested. Generally, the first step in mixture interpretation is determining the

number of contributors, with the most common approach of maximum allele count. Although there are previous studies regarding the accuracy of this approach, none have evaluated the accuracy with the newly expanded U.S. core STR loci. In this work, 4,976,355 theoretical mixture combinations were generated with the PowerPlex® Fusion 6C system which includes 23 autosomal STR loci and three Y-STR loci. The number of contributors could be correctly assumed for 100% two-person and 99.99% three-person mixtures, whereas, four-, five-, and six-person mixtures were correctly assumed in 89.7%, 57.3%, and 7.8% of mixtures, respectively. Y-STR analysis showed the 3 Y-STR markers are only accurate for two-person male mixtures (96.7%). This work demonstrates that maximum allele count using the expanded U.S. core loci is not much improved from previous smaller panels, reiterating that this method is not as accurate beyond three contributors.

Key words: forensic science, DNA, mixtures, PowerPlex® Fusion 6C, allele counting

Background

Mixtures are a common challenge in DNA profile interpretation. DNA mixtures are more frequently encountered in forensic casework than in the earlier years of STR typing. This is mainly because of increased sensitivity of the commercially available genotyping kits and the opportunity to optimize PCR reactions for 'touch' or low copy number (LCN) DNA. One published study retroactively reviewed 1547 cases over four years (1997-2000). Of the 2424 samples from those cases, 163 (6.7%) showed a mixture profile, and only eight of the 163 (0.3%) samples were mixtures of more than two contributors [1]. A decade later, a survey study initiated by SWGDAM in 2008 collected case data from 14 laboratories on 4541 samples, where 45.2% showed a mixture profile,

and 526 (11.6%) samples were mixtures of more than two contributors [2]. This survey was the basis of the 2010 SWGDAM DNA interpretation guidelines to focus on single source and two person mixture samples, although updated SWGDAM guidelines do include criteria for more than two contributors [3].

The first step in interpreting a DNA profile is identifying the presence of a mixture, or, a profile with more than one contributor. This is typically determined by analyzing the number of allelic peaks and peak height ratios, while considering stochastic effects, including stutter. According to SWGDAM guidelines, if one or more loci have 3 or more alleles present, excluding tri-allelic loci, then the sample is assumed to be a mixture [3]. The next logical step is determining the number of contributors in that mixture. This is a key step to the deconvolution of the mixture to assign genotypes to each individual present for providing statistical weight to the evidence. The most common approach for estimating the number of contributors is maximum allele count [4]. Maximum allele count is used to estimate the number of minimum contributors to the mixed sample by evaluating the locus that has the greatest number of allelic peaks [3], because a single individual should only have a maximum of two alleles at a locus. For example, if a locus has five allele peaks, there has to be a minimum of three contributors because for a two-person mixture, the expected maximum number of alleles is four.

There have been some previous studies to characterize the number of contributors according to maximum allele count. Paoletti et al. [5] generated conceptual three- and four- person mixtures from an FBI database which contained genotypes from the 13 core CODIS STR loci from 959 individuals. Based on maximum allele count, they found 3% of the 146,536,159 three-person mixtures could be mischaracterized as two-person mixtures, and that 76% of the 57,211,376 four-person mixtures could be mischaracterized

as two- or three-person mixtures [5]. Haned et al. [4] conducted simulations from published genotypes of individuals with 15 STR loci (including 13 CODIS loci) by generating 1000 mixtures comprised of between two to five contributors to compare maximum allele count with maximum likelihood, another method for determining number of contributors [4]. They concluded that mixtures of two or three contributors was greater than 90% for both methods, but with mixtures of 4 or 5 contributors, maximum likelihood yielded greater success rates. For example, correct mixtures for Caucasians with four and five contributors with maximum allele count was 34% and 2%, respectively, but with maximum likelihood, 77% and 64%, respectively [4].

There are three different mathematical models that can be used for mixture interpretation: binary, semi-continous, or continuous [6]. Binary statistical models (i.e., random match probability (RMP), likelihood ratio (LR), and combined probability of exclusion/inclusion (CPE/CPI)) are still very common in practice, however, they are limited to cases where at least one contributor can be deconvoluted from the mixture (RMP), or, the need for all alleles to be present for the loci to be used (CPI). They also require the analyst to assume the number of contributors in order to perform the statistic (with the exception of CPE/CPI) [3]. Furthermore, according to the recently released PCAST report, the CPE/CPI statistic was deemed inadequate and subjective [7]. Probabilistic genotyping methods and software (semi-continuous and continuous models) have been developed as an improved alternative to simpler binary practices for mixture interpretation, examples of programs include TrueAllele® [8], STRmix™[9], Lab Retriever [10], *forensim* R package [11], and NOCit, a program that estimates number of contributors as part of the PROVEDIt initiative [12]. However, the PCAST report points to the fact that although probabilistic genotyping methods are an improvement, further

testing should be done to ensure the scientific validity on reliability and on the algorithms being implemented [7]. This need for further testing could be seen demonstrated in *New York v. Oral Hillary*, where two different probabilistic programs were used to evaluate the same mixture profile and two different conclusions resulted (TrueAllele® did not find a link to the defendant whereas STRMix[TM] could not rule out the defendant) [13]. Many laboratories are still implementing binary methods, including maximum allele count to determine number of contributors.

To date, there have not been any published studies to evaluate the maximum allele count method on the expanded U.S. core STR loci. The PowerPlex® Fusion 6C system (Promega Corp., Madison, WI) incorporates 27 loci which includes the expanded 20 U.S. CODIS core loci [14]. It was the objective of this work to evaluate how the maximum allele count method would determine number of contributors for theoretically generated combined two-, three-, four-, five-, and six-person mixtures (4,976,355 total mixture profiles) based on 236 unrelated genotypes using the PowerPlex® Fusion 6C kit (Promega Corp.).

Material and Methods

Single source reference DNA profiles (N = 236) were amplified from non-related anonymous volunteers collected by the Indiana State Police Laboratory. Genotypes were generated using the PowerPlex® Fusion 6C System (Promega Corp.) using the BioMek NX[P] and BioMek 3000 Automated Workstations under standard casework operating procedures of the Indiana State Police Laboratory [15]. The genotypes, each designated with a random number identifier, were entered into an electronic database using Microsoft Excel for theoretical mixture generation and analysis. There were 4,976,355 total mixture combinations generated.

A macro using Visual Basic in Microsoft Excel was used to generate all possible combinations of two- and three-person mixtures. The macro was also used to generate combinations for the four-, five-, and six-person mixtures, however, due to the large number of possible combinations and the limitation in number of rows possible in Excel (1,048,576 rows), only a random subset (generated using a random number function in Excel) of all possible combinations were analyzed (see Table 1). Two separate random sample sets were generated for the four-person combinations to ensure allele count distributions were representative of the whole set (Table 2). Statistics were performed in Microsoft Excel. Each set of mixtures were analyzed with the following defined parameters: (i) the minimum allele count is the count across all loci per profile that had the lowest number of alleles observed in at least one locus; (ii) the maximum allele count is the count across all loci per profile that had the highest number of alleles observed in at least one locus; and, (iii) the overall count is the frequency distribution of all allele counts across all loci of all possible n person profile combinations. As the kit also contains three Y-STR markers, a separate analysis of the Y-STRs was performed whereas the generated mixture combinations were filtered to analyze those between male mixtures only.

Results and Discussion

For all mixture combinations, the profiles were considered under ideal conditions (equal ratios, no stutter or artifacts, and all alleles were above the stochastic threshold). Therefore, allele count analysis was calculated based on the assumed presence of all possible allelic peaks from all individuals in the mixture. This does not reflect the possibilities of mixed ratios, stutter, or allele-dropout, which are not unexpected in casework mixtures [16, 17].

Two-Person Mixtures

For the 27,730 two-person mixtures, the minimum allele count was two in 70% of mixtures, and one in the remaining 30% of profiles (Figure 1a). Although it was possible to see loci with only one allele, which is more typical in single source profiles, a maximum allele count of four was 99.99% (Figure 1b). There were four (0.01%) profiles that had a maximum of three allele in at least one locus. In either case, a minimum of two contributors would still be indicated (Figure 1b). Based on maximum allele count, two-person mixtures could accurately be determined as having two contributors in all cases. SE33, D1S1656, and Penta E are among the loci that have the highest frequency of the highest allele count of four (Figure 2).

Three-Person Mixtures

For the 2,162,940 three-person mixtures, the majority (80%) exhibited at least one locus with a minimum of two alleles (Figure 3a). There were 100,831 (4.7%) profiles where there was at least one locus with a minimum allele count of one allele (Figure 3a). One three-person mixture (0.0005%) was shown to have a minimum of four alleles, meaning all loci exhibited at least 4 alleles. More than two million profiles were generated and only one profile exhibited this pattern indicating how rare it is for three individuals to have at least four unique alleles between them at every locus. The maximum allele count method is accurate for three-person mixtures (99.99%: 21.5% with five alleles, and 78.5% with six alleles) (Figure 3b). There were 277 (0.01%) profiles that may have been interpreted as a mixture with only two contributors. SE33, D1S1656, and Penta E remain the loci with the highest frequencies of the highest allele count of six, whereas TPOX had the highest frequency of only two alleles (Figure 4).

Four-Person Mixtures

Minimum allele counts in the 916, 895 four-person mixture profiles were nearly evenly distributed between two and three alleles (52% and 46.7%, respectively, Figure 5a). There were 11,174 (1.2%) profiles that showed at least one locus with only one allele within the profile (Figure 5a). At least one locus exhibited seven or eight alleles in approximately 90% (61% and 28.7%, respectively, Figure 5b). However, there were 94,880 profiles (10.3%) that would be interpreted as three-person mixtures as there was only a maximum count of five or six alleles. Therefore there is a 10.3% chance it may be misinterpreted as a three-person mixture, based on maximum allele count alone. The loci that have higher allele count frequencies for the highest possible allele count of eight remain SE33, D1S1656, and Penta E as seen for the two- and three-person mixtures, although the frequencies are lower here, indicating there are fewer profiles for four-person mixtures that exhibit the maximum number of unique alleles at these loci (Figure 6).

Five-Person Mixtures

It was shown to be more difficult to discriminate a minimum of five contributors than mixtures with fewer contributors. The majority (72.5%) of the 962,598 profiles had at least one locus with a minimum of three alleles, followed by 24.5% with two alleles (Figure 7a). There were 3216 profiles (0.3%) that had at least one locus with only one allele. The maximum allele count was nine in 42.9% of profiles, which is indicative of a five-person mixture (Figure 7b). However, the next highest distribution of profiles (37.4%) were observed to have a maximum allele count of eight, which would indicate a four-person mixture (Figure 7b), and again an unknown mixture could be misinterpreted 37.4% of the time as a four-person mixture based on maximum allele count alone. Approximately 57% of observed five-person mixtures had at least one locus with nine or 10 alleles, thereby

making estimations of the number of contributors inaccurate for nearly half of the profiles generated here (Figure 7b). Per locus, SE33 had the highest distribution exhibiting a maximum of 10 alleles with 204,446 (21%) profiles (Figure 8). The second and third highest loci with a maximum allele count of 10 were at much lower distributions: D12S391 with 26,974 (2.8%) profiles, and Penta E with 16,964 (1.8%) profiles (Figure 8).

Six-Person Mixtures

　　　Six-person mixtures, as expected, were the most difficult to discriminate by allele count as compared to the two-, three-, four-, and five-person mixtures. The minimum allele count was similar to the five-person mixtures with only approximately a 5% increase in profiles with at least one locus with a minimum of three alleles (72.5% to 77.3%, Figure 9a). There were 98 (0.01%) profiles where at least one locus exhibited one allele (Figure 9a); conversely, there were 6 profiles (0.001%) that had a minimum allele count of 5.

The maximum count that would distinguish a six-person mixture from a five-person mixture would be 11 or 12 alleles, and only 71,176 (7.8%) of profiles had at least one locus with either 11 or 12 alleles (7.1% and 0.7%, respectively, Figure 9b). The most common maximum allele count was nine with 411,324 (45.4%) profiles (Figure 9b). This could incorrectly indicate five contributors almost 50% of the time for unknown samples based on maximum allele count alone. In terms of specific loci, only three were observed to produce profiles with 12 alleles, all with low frequency: SE33 with 6,332 profiles (0.07%), D12S91 with 68 (0.0008%) profiles, and D18S51 with 22 (0.0002%) profiles (Figure 10).

The two loci with the highest profile counts with 11 alleles were SE33 with 60,676 (0.7%) profiles and D12S391 with 2,464 (0.03%) profiles (Figure 10).

Overall by Locus

As previously discussed, SE33 was the most polymorphic locus, with the greatest number of profiles containing the maximum allele counts for mixtures with all number of contributors (Figure 11). This is not surprising considering SE33 has 58 distinguishable alleles, which is twice the number of unique alleles compared to FGA, the next most variable STR [18]. If SE33 is considered as the only locus, 97.7% of profiles have three or four alleles for two-person mixtures, and with three-person mixtures, 85.3% of profiles have five or six alleles which would indicate minimally two and three contributors, respectively (Table 3). When considering allele count distribution across the number of contributors, 70% of profiles with four alleles is observed at SE33 is a two-person mixture, 44% of profiles with five alleles is a three-person mixture, with much smaller percentages of five alleles seen in other mixtures (Table 3). The locus is less informative with the higher count of alleles, where 59% of profiles with eight alleles could be a five- or six-person mixture, and if this is the maximum allele count, it would be incorrectly interpreted as a four-person mixture (Table 3). For five-person mixtures, 58% of profiles showed nine or 10 alleles. For six-person mixtures, only 7% of profiles at SE33 had 11 or 12 alleles observed. The majority of allele distribution for six-person mixtures was eight or nine alleles (Table 3). It is interesting to note that SE33 was not included as one of the required expanded U.S. core loci, it is listed as a possible locus to include by preference, because although it is highly polymorphic, it also has a high mutation rate [19]. In fact, it is not included in many of the current amplification kits [20]. An issue also arises in successfully genotyping SE33, as it is susceptible to allele drop-out and mobility shifts due to mutations

in the primer binding sites, resulting in discordant genotyping results between different kits
[21].

When considering other loci, D12S391 (an expanded U.S. CODIS core locus), and D18S51
(one of the 13 original CODIS core loci) [22] were the only other loci to have a maximum of 12
alleles for the six-person mixtures (Figure 11). Conversely, TPOX had the lowest allele maximum
counts especially for four-, five-, and six-person mixtures; there were only a maximum count of
five (Figure 11). This is not surprising as TPOX is considered the least polymorphic of the
commonly used STR loci [23] and has one of the highest allele frequencies of the autosomal loci
(0.54 for the '8' allele in the Caucasian population) [24]. All loci exhibited the possibility of
maximum allele count of four for two-person mixtures, and all loci but D3S1358 exhibited the
maximum possibility of six alleles for three-person mixtures (Figure 11). The maximum possible
counts for four-, five-, and six- person mixtures were more variable between all loci, with fewer
loci exhibiting the higher possible counts (Figure 11). This is not a surprising trend as it is
inherently difficult to deconvolute higher count mixtures.

Total Allele Count Distributions

A DNA profile is considered as a whole and not just by a single locus. Another overall
observation to make with all these mixture profiles is the average number of alleles seen across all
23 autosomal loci in the profile. Maximum allele count is important for assuming the number of
contributors, but there may be trends between the number of overall allele counts and number of
contributors. For two-person mixtures across all loci in the profile, 49% of all loci had three alleles,
and 27.3% had four alleles (Figure 12a). When comparing three-person mixtures to the two-person
mixture distributions, there is a decrease in the number of loci that exhibit three alleles, and an
increase in those that exhibit four alleles (Figure 12a,b). Additionally, 7.2% and 5.8% of loci had

five or six alleles, respectively (Figure 12b). When considering total allele count across a four-person mixture compared to a three-person mixture, there is a decrease in loci that exhibit four alleles and increase in those exhibiting five and six alleles (Figure 12b,c). However, when comparing between four-, five-, and six-person mixtures, the distribution of loci, other than the increase of the highest two allele counts between each, is very similar demonstrating that discriminating between four-, five-, and six-person mixtures based on allele counts from this set of 23 autosomal loci would be very difficult (Figure 12 c,d,e).

Y-STR Analysis

The PowerPlex® Fusion 6C system (Promega Corp.) includes three Y-STR markers, two of which were an addition since the PowerPlex® Fusion system (Promega Corp.). The Y-STR data was analyzed separately and only for mixtures that had all male contributors (e.g., two-person mixtures with two males, three-person mixtures with three males, etc.). The purpose was to determine how often a two-, three-, four-, five-, and six-person male mixtures would be correctly identified based solely on the Y-STR loci. The trends are similar with the Y-STRs as with the autosomal loci for the two-person mixture, where approximately 97% (N= 5,901) of two-person male mixtures exhibited at least one Y-STR locus with two alleles (Table 4). For three-person male mixtures, approximately two-thirds of all profiles (67%, N= 148,387) exhibited at least one locus with three alleles (Table 4). Similarly to three-person male mixtures, most four-person male mixtures (64%, N= 26,246) exhibited three or less alleles. Considering five- and six-person male mixtures, approximately 50% for both sets exhibited four or less alleles (Table 4). Therefore the three Y-STR markers were informative for number of contributors in the two-person male mixtures, and for the majority of three-person male mixtures, but are not very discriminating for

number of contributors in the four-, five-, and six-person male mixtures. DYS576 was found to be the most diverse of the three YSTRs analyzed.

Conclusions

This study shows that using maximum allele count with profiles generated with an increased number of STR loci which includes the expanded U.S. core loci, is accurate for estimating number of contributors for two- and three-person mixtures, with improvement when 13 STR loci are used(3% of three-person mixtures mischaracterized with 13 loci, whereas only 0.01% mischaracterized with 23 loci and 70% of four-person mixtures mischaracterized with 13 loci, only 10.3% with 23 loci) [5]. However, the general trend remains the same, as the number of contributors increases, the accuracy of the estimates suffers, even with the expanded panel of standard loci.

Probabilistic genotyping methods have been developed to improve capabilities and standardization for mixture interpretation using continuous models (see Background). However, one study found that when the same samples were analyzed within and between laboratories, using STRmix™, consistent and concordant results were only demonstrated when the number of contributors was not ambiguous, and that the assignment of the number of contributors is essential to effectively interpret DNA profiles [25].

While the future of mixture interpretation is heading towards these computer-based probabilistic genotyping methods, of which the limits and scope of application are still being determined, current practices in many laboratories still implement maximum allele count in their standard operating procedures. This study highlights that caution still need be given when assuming the number of contributors in suspected mixtures of greater than three individuals as even with the expanded U.S. core STR loci, discrimination for four-, five-, and six-person mixtures is complex and difficult. Highly variable STRs, such as SE33, can be more useful for these higher

14

order mixtures as seen here, however, there are still limitations in the successful typing of this locus and it is not available in all genotyping kits. Other genotype factors are taken into consideration when interpreting a profile besides the presence or absence of alleles, such as the peak height ratios, stutter, relatedness etc. The analyses reported here were based on theoretical 1:1 mixtures of unrelated individuals under ideal conditions, and therefore further empirical testing with different ratios of contributors, or even related individuals, would be valuable to perform as it would simulate more realistic conditions of casework mixture samples.

REFERENCES

[1] Y. Torres, I. Flores, V. Prieto, M. López-Soto, M.J. Farfán, A. Carracedo, P. Sanz, DNA mixtures in forensic casework: a 4-year retrospective study. Forensic Sci Int. 134 (2003) 180-186.

[2] A.M. Gross, Numbers and Types of Casework Mixtures. DNA Mixture Interpretation: Principles and Practice in Component Deconvolution and Statistical Analysis. Washington, D.C.: AAFS Annual Meeting Workshop #16; 2008. http://strbase.nist.gov/training/AAFS2008_1_CaseworkSurvey.pdf (accessed 6 November 2017).

[3] SWGDAM. SWGDAM Interpretation Guidelines for Autosomal STR Typing by Forensic DNA Testing Laboratories. 2010. http://www.forensicdna.com/assets/swgdam_2010.pdf (accessed 6 November 2017).

[4] H. Haned, L. Pene, J.R. Lobry, A.B. Dufour, D. Pontier, Estimating the number of contributors to forensic DNA mixtures: does maximum likelihood perform better than maximum allele count? J Forensic Sci. 56 (2011) 23-28.

[5] D.R. Paoletti, T.E. Doom, C.M. Krane, M.L. Raymer, D.E. Krane, Empirical Analysis of the STR Profiles Resulting from Conceptual Mixtures. J Forensic Sci. 50 (2005) 1361-1366.

[6] H. Kelly, J.A. Bright, J.S. Buckleton, J.M. Curran, A comparison of statistical models for the analysis of complex forensic DNA profiles. Sci Justice 54 (2014) 66-70.

[7] PCAST. Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods. President's Council of Advisors on Science and Technology; 2016. https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/PCAST/pcast_forensic_science_report_final.pdf (accessed 6 November 2017).

[8] M.W. Perlin, M.M. Legler, C.E. Spencer, J.L. Smith, W.P. Allan, J.L. Belrose, B.W. Duceman, Validating TrueAllele DNA mixture interpretation. J Forensic Sci. 56 (2011) 1430-1447.

[9] J.A. Bright, D. Taylor, C. McGovern, S. Cooper, L. Russell, D. Abarno, J.S. Buckleton, Developmental validation of STRmix, expert software for the interpretation of forensic DNA profiles. Forensic Sci Int Genet. 23 (2016) 226-239.

[10] K. Inman, N. Rudin, K. Cheng, C. Robinson, A. Kirschner, L. Inman-Semerau, K.E. Lohmueller, Lab Retriever: a software tool for calculating likelihood ratios incorporating a probability of drop-out for forensic DNA profiles. BMC Bioinformatics 16 (2015) 1-10.

[11] H. Haned, Forensim: an open-source initiative for the evaluation of statistical methods in forensic genetics. Forensic Sci Int Genet. 5 (2011) 265-268.

[12] H. Swaminathan, C.M. Grgicak, M. Medard, D.S. Lun, NOCIt: a computational method to infer the number of contributors to DNA samples analyzed by STR genotyping. Forensic Sci Int Genet. 16 (2015) 172-180.

[13] S. Augenstein, Prosecutor's DNA Evidence Tossed from Upstate N.Y. Murder Trial. Forensic Magazine 2016. https://www.forensicmag.com/news/2016/08/prosecutors-dna-evidence-tossed-upstate-ny-murder-trial (accessed 6 November 2017).

[14] M.G. Ensenberger, K.A. Lenz, L.K. Matthies, G.M. Hadinoto, J.E. Schienman, A.J. Przech, M.W. Morganti, D.T. Renstrom, V.M. Baker, K.M. Gawrys, M. Hoogendoorn, C.R. Steffen, P. Martin, A. Alonso, H.R. Olson, C.J. Sprecher, D.R. Storts, Developmental validation of the PowerPlex Fusion 6C System. Forensic Sci Int Genet. 21 (2016) 134-144.

[15] Forensic Biology Section Casework Test Methods. Indiana State Police. http://www.in.gov/isp/labs/files/Biology_Casework_Test_Method_Version22.pdf (accessed 6 November 2017).

[16] P. Gill, C.H. Brenner, J.S. Buckleton, A. Carracedo, M. Krawczak, W.R. Mayr, N. Morling, M. Prinz, P.M. Schneider, B.S. Weir, DNA commission of the International Society of Forensic Genetics: Recommendations on the interpretation of mixtures. Forensic Sci Int. 160 (2006) 90-101.

[17] B. Budowle, A.J. Onorato, T.F. Callaghan, A. Della Manna, A.M. Gross, R.A. Guerrieri, J.C. Luttman, D.L. McClure, Mixture interpretation: defining the relevant features for guidelines for the assessment of mixed DNA profiles in forensic casework. J Forensic Sci. 54 (2009) 810-821.

[18] V. Colonna, Q. Ayub, Y. Chen, L. Pagani, P. Luisi, M. Pybus, E. Garrison, Y. Xue, C. Tyler-Smith, the 1000 Genomes Project Consortium, Human genomic regions with exceptionally high levels of population differentiation identified from 911 whole-genome sequences. Genome Biol. 15 (2014) 6: R88, doi: 10.1186/gb-2014-15-6-r88.

[19] J. Ge, A. Eisenberg, B. Budowle, Developing criteria and data to determine best options for expanding the core CODIS loci. Investig Genet. 3 (2012) 1, doi: 10.1186/2014-2223-3-1.

[20] J.M. Butler, C.R. Hill, M.D. Coble, Variability of New STR Loci and Kits in US Population Groups.   http://strbase.nist.gov/pub_pres/Profiles-in-DNA_Variability-of-New-STR-Loci.pdf (accessed 6 November 2017).

[21] J.M. Butler, C.R. Hill, M.C. Kline, I. Bastisch, V. Weirich, R.S. McLaren, D.R. Storts, SE33 variant alleles: Sequences and implications. Forensic Sci Int Genet Supp. 3 (2011) e502-e503.

[22] D.R. Hares, Selection and implementation of expanded CODIS core loci in the United States. Forensic Sci Int Genet. 17 (2015) 33-34.

[23] J.M. Butler, C.R. Hill, Biology and Genetics of New Autosomal STR Loci Useful for Forensic DNA Analysis. Forensic Sci Rev. 24 (2012) 15-26.

[24] T.R. Moretti, L.I. Moreno, J.B. Smerick, M.L. Pignone, R. Hizon, J.S. Buckleton, J.A. Bright, A.J. Onorato, Population data on the expanded CODIS core STR loci for eleven populations of significance for forensic DNA analyses in the United States. Forensic Sci Int Genet. 25 (2016) 175-181.

[25] S. Cooper, C. McGovern, J.A. Bright, D. Taylor, J. Buckleton, Investigating a common approach to DNA profile interpretation using probabilistic software. Forensic Sci Int Genet. 16 (2015) 121-131.
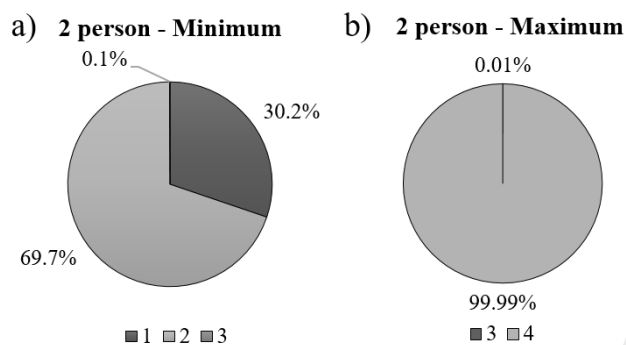
Figure 1. Two-person mixture allele counts. a) The minimum allele count distribution; and b) the maximum allele count distribution.
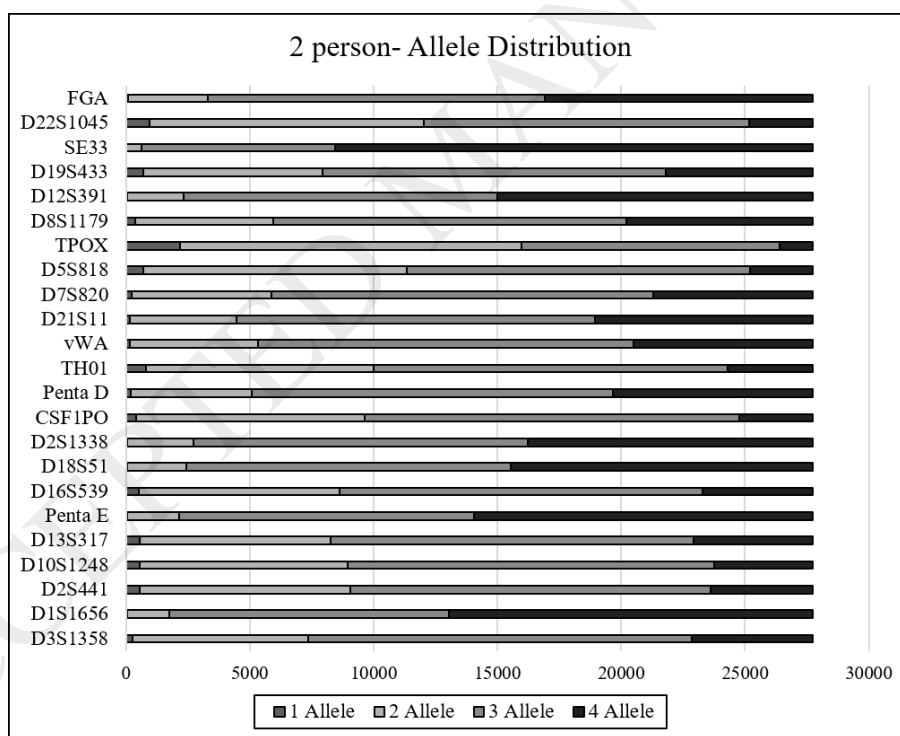

Figure 2. Frequency of allele counts by locus of the 2-person mixtures (N= 27,730).
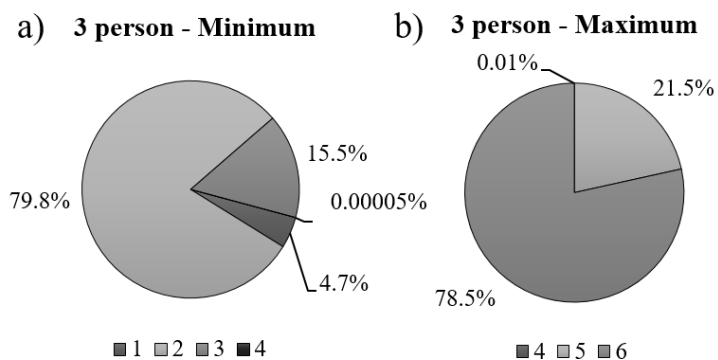
a) **3 person - Minimum**

b) **3 person - Maximum**

Figure 3. Three-person mixture allele counts (N=2,162,940). a) The minimum allele count distribution; and b) the maximum allele count distribution.
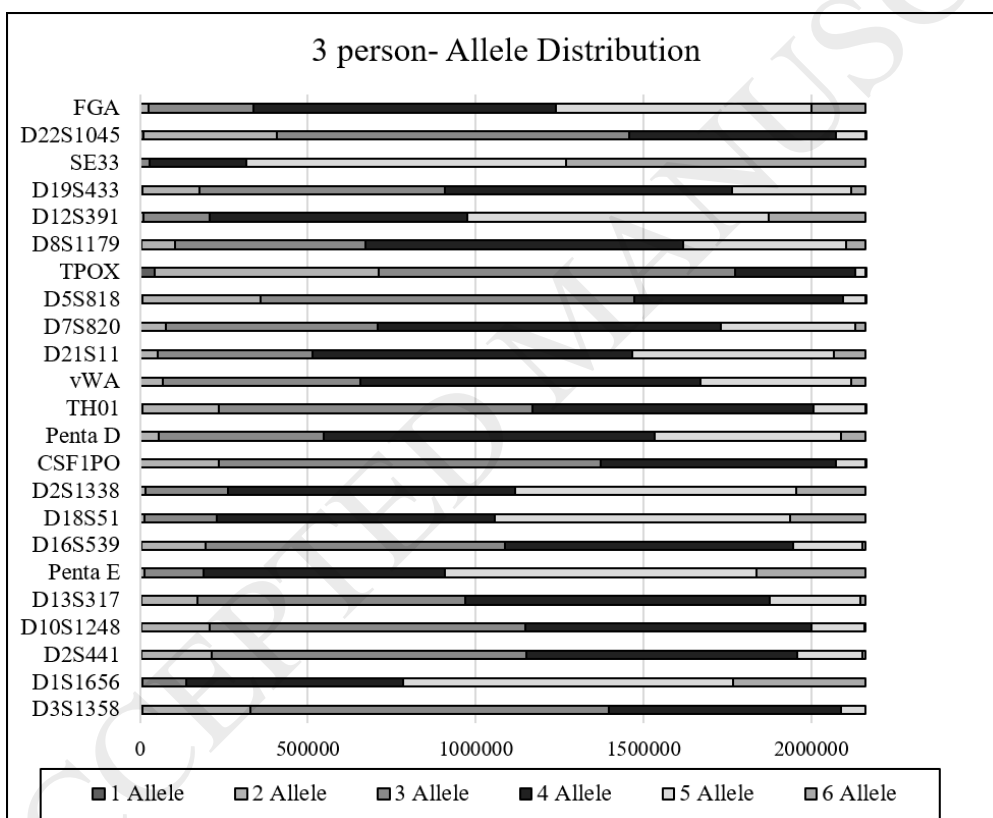

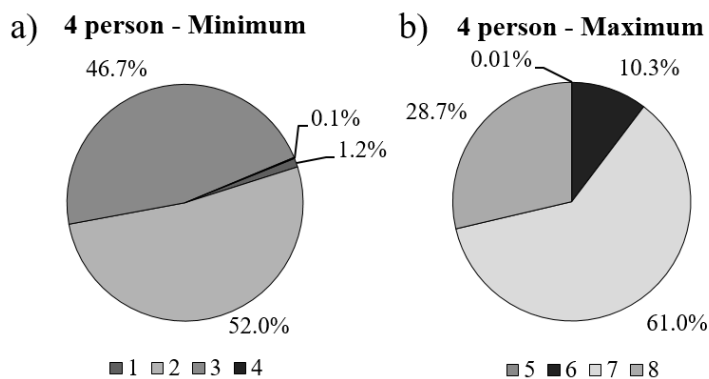**Figure 4. Frequency of allele counts by locus of 3-person mixtures (N= 2,162,940).**

a) **4 person - Minimum**

b) **4 person - Maximum**

Figure 5. Four-person allele counts. a) The minimum allele count distribution; and b) maximum allele count distribution.
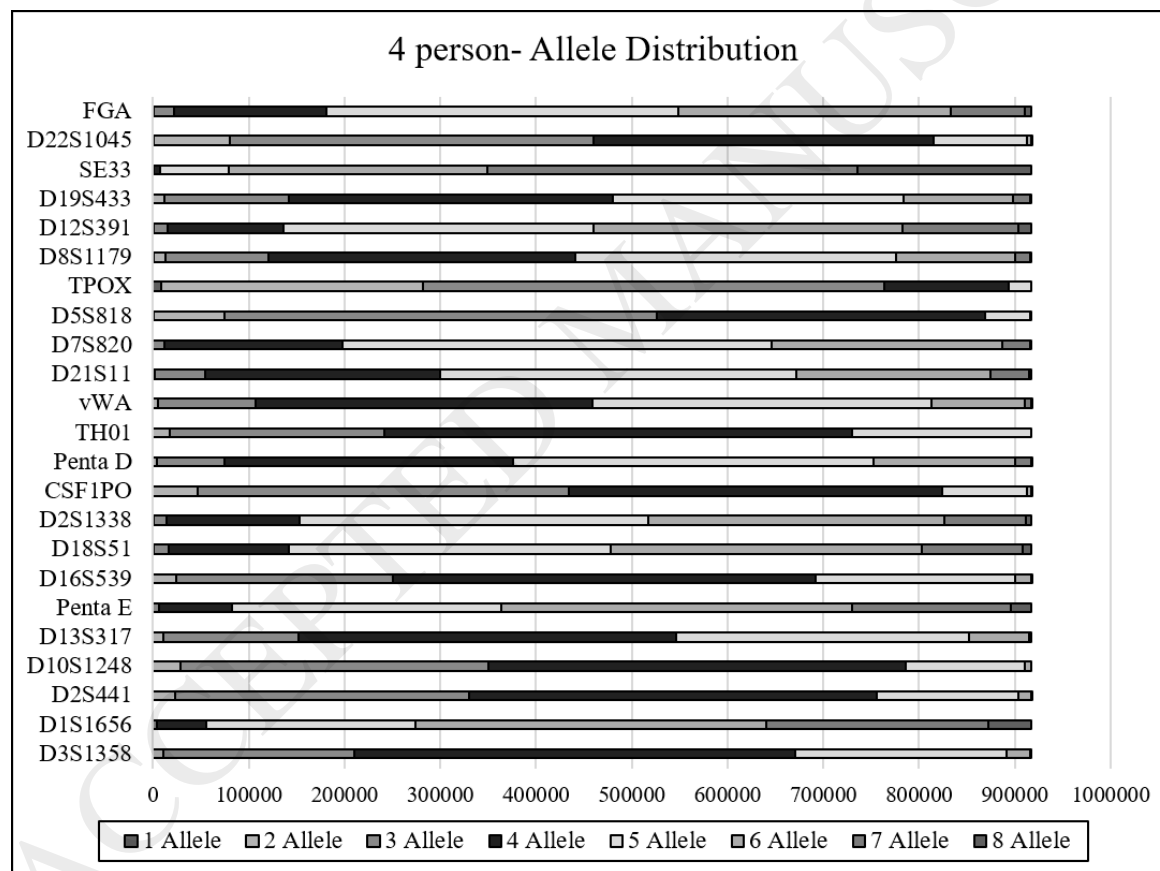
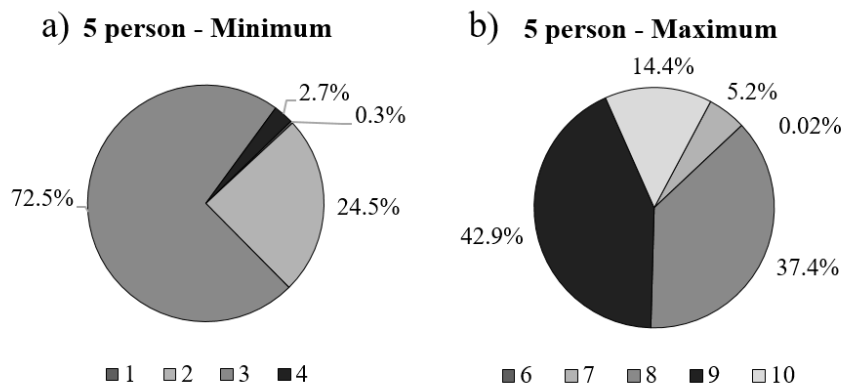Figure 6. Frequency of allele counts by locus of 4-person mixtures (N= 916,895)

a) **5 person - Minimum**

b) **5 person - Maximum**

Figure 7. Five-person allele counts. a) The minimum allele count
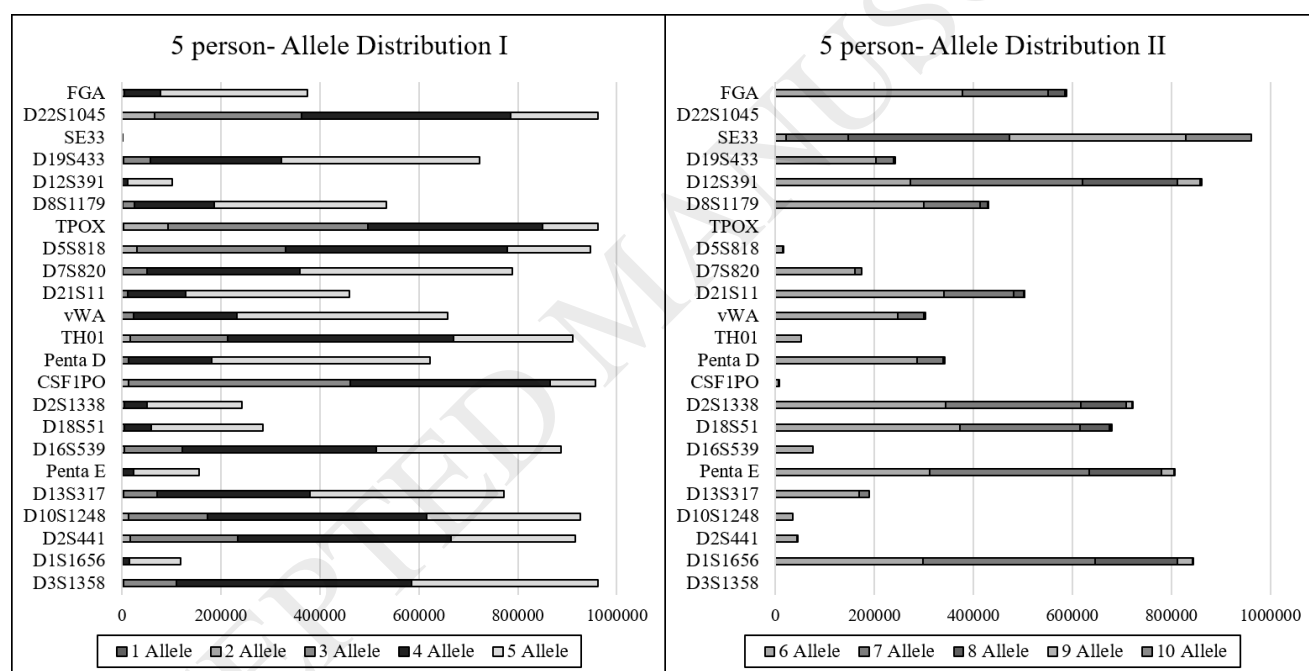distribution; and b) maximum allele count distribution.

Figure 8. Frequency of allele counts by locus of 5-person mixtures (N= 962,598).
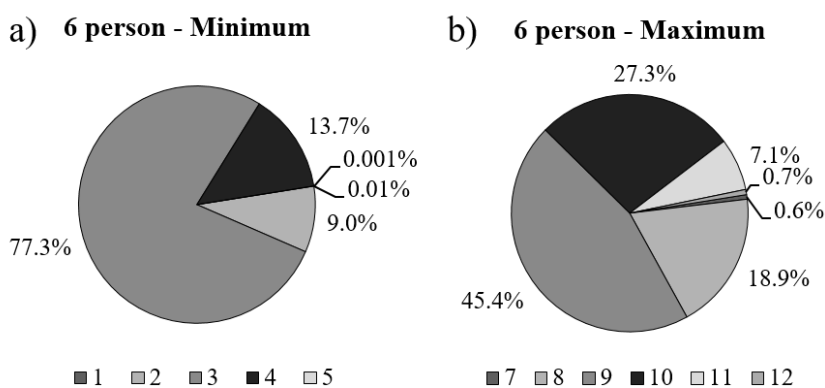
a) **6 person - Minimum**

b) **6 person - Maximum**

Figure 9. Six-person allele counts. a) The minimum allele count distribution; and b) maximum allele count distribution.
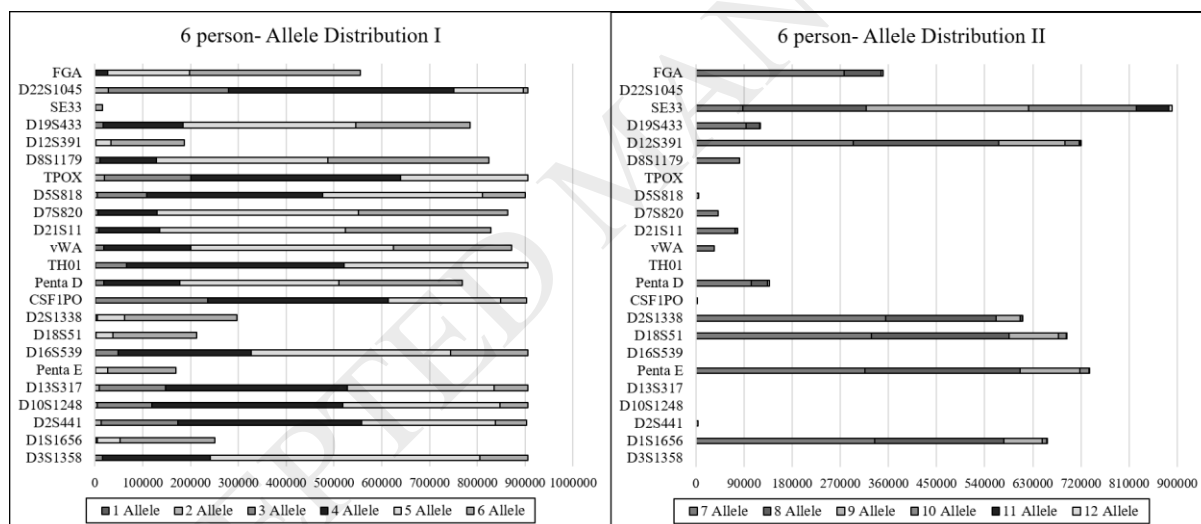

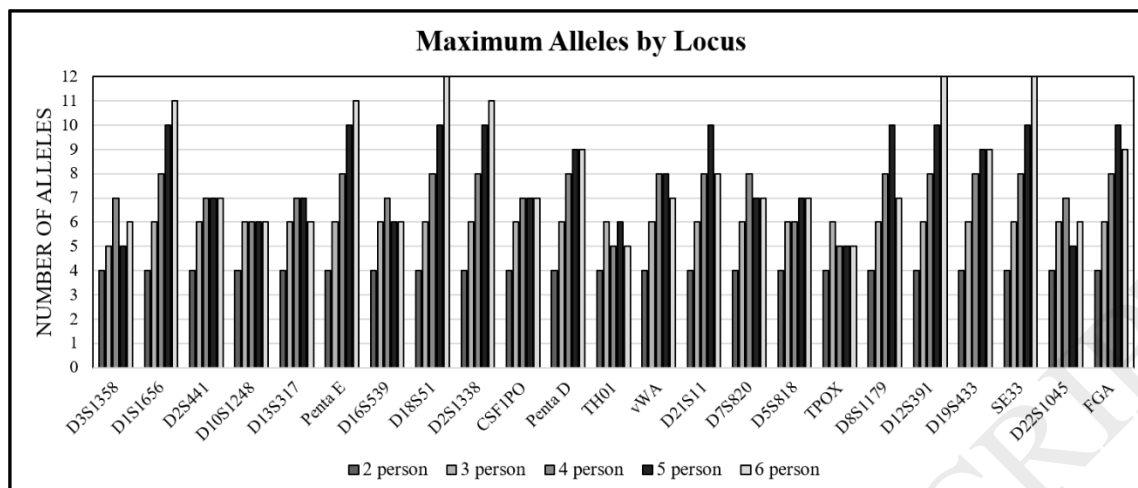**Figure 10. Frequency of allele counts by locus of the 6-person mixtures (N= 906,192).**

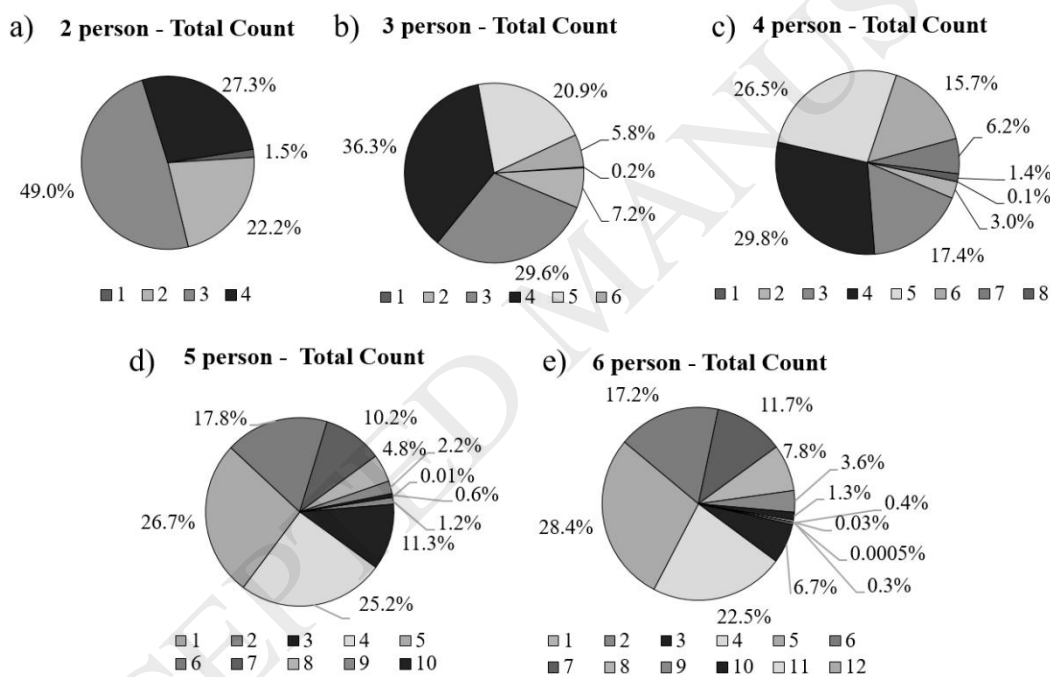**Figure 11. Maximum alleles per loci for all mixtures.**

Figure 12. Total allele count distributions across all the autosomal loci across all observed profiles for a) 2-person mixtures, b) 3-person mixtures, c) 4-person mixtures, d) 5-person mixtures, and e) 6 person mixtures.

Table 1. Number of mixture combinations.

Number of Contributors	Number of mixtures generated (number of database samples used)
2 person	27,730 (236)
3 person	2,162,940 (236)
4 person	916,895 (70)*
5 person	962,598 (43)*
6 person	906,192 (32)*
TOTAL	4,976,355

*= subset of total possible combinations

Table 2. Comparison of maximum allele count between two separately generated combinations of four-person mixtures (N=916,895 each, $p > 0.5$)

4-person Mixture Combinations: Allelic Distribution		
Number of Alleles	**Group 1**	**Group 2**
1	5555	11204
2	533410	623641
3	3467181	3671196
4	6601424	6289826
5	5805377	5596867
6	3236480	3307998
7	1194533	1302849
8	244625	285004
TOTAL	21088585	21088585

Table 3. Frequency of 2-, 3-, 4-, 5-, and 6-person mixtures based on number of alleles seen at the SE33 locus among all observed mixture profiles.

Alleles	2 person (%)	3 person (%)	4 person (%)	5 person (%)	6 person (%)
1	0.03	0.00005			
2	2.3	0.03			
3	28.2	1.3	0.02		

4	**69.5**	13.4	0.75	0.002	0.001
5		**44.0**	7.8	0.14	0.1
6		41.3	29.4	2.2	1.7
7			**42.2**	13.1	9.7
8			19.8	33.8	25.4
9				**37.0**	**33.4**
10				20.9	12.4
11					6.7
12					0.7

Bold = highest distribution of alleles per *n*-person mixture

Table 4. Y-STR Distributions per *n*-person Male Mixtures

Mixture	Number of Mixtures	Max Alleles	Frequency (%)
2 males	6,105	1	3.3
		At least one locus with 2	96.7
3 males	221,815	1	0.1
		2 or less	32.9
		At least one locus with 3	67.0
4 males	40,919	2 or less	11.0
		3 or less	64.0
		At least one locus with 4	25.0
5 males	4,278	2 or less	1.5
		3 or less	40.0
		4 or less	52.5
		At least one locus with 5	6.0
6 males	5,005	2 or less	0.2
		3 or less	10.8
		4 or less	51.0
		5 or less	35
		At least one locus with 6	3.0