11-2005

Empirical Analysis of the STR Profiles Resulting from Conceptual Mixtures

David R. Paoletti
Wright State University

Travis E. Doom
Wright State University

Carissa M. Krane
University of Dayton, ckrane1@udayton.edu

Michael L. Raymer
Wright State University

Dan E. Krane
Wright State University

J Forensic Sci, Nov. 2005, Vol. 50, No. 6
Paper ID JFS2004475
Available online at: www.astm.org

David R. Paoletti,[1] *M.S.; Travis E. Doom,*[1,2] *Ph.D.; Carissa M. Krane,*[3] *Ph.D.; Michael L. Raymer,*[1,2] *Ph.D.; and Dan E. Krane,*[4] *Ph.D.*

Empirical Analysis of the STR Profiles Resulting from Conceptual Mixtures

ABSTRACT: Samples containing DNA from two or more individuals can be difficult to interpret. Even ascertaining the number of contributors can be challenging and associated uncertainties can have dramatic effects on the interpretation of testing results. Using an FBI genotypes dataset, containing complete genotype information from the 13 Combined DNA Index System (CODIS) loci for 959 individuals, all possible mixtures of three individuals were exhaustively and empirically computed. Allele sharing between pairs of individuals in the original dataset, a randomized dataset and datasets of generated cousins and siblings was evaluated as were the number of loci that were necessary to reliably deduce the number of contributors present in simulated mixtures of four or less contributors. The relatively small number of alleles detectable at most CODIS loci and the fact that some alleles are likely to be shared between individuals within a population can make the maximum number of different alleles observed at any tested loci an unreliable indicator of the maximum number of contributors to a mixed DNA sample. This analysis does not use other data available from the electropherograms (such as peak height or peak area) to estimate the number of contributors to each mixture. As a result, the study represents a worst case analysis of mixture characterization. Within this dataset, approximately 3% of three-person mixtures would be mischaracterized as two-person mixtures and more than 70% of four-person mixtures would be mischaracterized as two- or three-person mixtures using only the maximum number of alleles observed at any tested locus.

KEYWORDS: forensic science, DNA typing, DNA mixtures, short tandem repeats, Combined DNA Index System, allele sharing, bioinformatics

PCR-based amplification of STR loci has become the method of choice for the purpose of human identification in forensic investigations (1,2). While alternatives exist (3,4), most DNA-typing laboratories use commercially available kits to amplify and label STR alleles associated with evidence and reference samples that are then size fractionated with capillary electrophoresis systems such as the ABI 310 or 3100 Genetic Analyzers (5,6). Software such as GeneScan® and Genotyper® are then used to determine the presence or absence of STR alleles associated with a sample.

Interpreting evidence samples containing mixed DNA profiles is more complicated than the analysis of single source samples. Programs exist that can aid analysts attempting to "deconvolve" the contributors of mixed samples on the basis of associations between peak heights or areas (7,8) and, in some instances, using the genotype information from individuals presumed to have been contributors (7). Analysis of a multiple contributor sample is particularly challenging when potential contributors have several alleles in common (such as is often the case with close relatives), when stochastic variations in peak heights occur, or when technical artifacts such as stutter, allelic dropout, and degradation/inhibition occur.

With only rare exceptions, an individual should possess exactly two actual alleles for every locus. These alleles may differ from each other (heterozygous) or be effectively indistinguishable (homozygous). For example, at one STR locus an individual may be found to have alleles 11 and 12, or 12 and 13 (Fig. 1*A*, 1*B*). When more than two actual alleles are observed in the testing results from any single locus, it can be reasonably assumed that the presence of DNA from more than one contributor is the most likely explanation. The absence of a fifth or sixth actual allele is often interpreted as support of there being only two contributors to mixtures (Fig. 1*C*) even though it is formally possible for the number of contributors to be greater than two. However, if three actual alleles are observed, the sample may arise from a mixture of two individuals, a mixture of three individuals with overlapping alleles, or even a mixture of four or more individuals. Likewise, observing five or more actual alleles at one locus is an indication of three or more contributors. However, it becomes increasingly difficult to determine the exact number of contributors as the number of observed alleles increases. Although previous research (9) has analyzed how often a two-person mixture will present 1, 2, 3, or 4 alleles at six individual loci, no published studies address the issues of how often a three-person mixture will present no more than four different alleles at any tested locus, or how often a four-person mixture will present no more than six different alleles at any tested locus.

Exhaustive analysis of 959 complete 13-locus STR genotypes from a population dataset, as well as of randomized sets of comparable genotypes in this study could assist DNA analysts by formally addressing the relative statistical confidences of declarations regarding the number of contributors to a DNA mixture on the basis of alleles observed at typed STR loci. Peak heights and areas sometimes provide additional data that is useful for the purpose of mixture deconvolution but this information is not utilized in the analysis presented here. Instead, this study examines the interpretation of STR data in cases where this information is unreliable (i.e. when degradation has occurred and/or stutter complicates interpretation), unavailable (i.e., only a laboratory's summary report is provided for review) or uninformative (i.e., the relative contributions by two or more contributors are similar).

[1] Computer Science Department, Wright State University, Dayton, OH 45435.
[2] Forensic Bioinformatics, Inc., 2850 Presidential Drive, Suite 150, Fairborn, OH 45324.
[3] Biology Department, University of Dayton, Dayton, OH 45969.
[4] Department of Biological Sciences, Wright State University, Dayton, OH 45435.

Received 6 Nov. 2004; and in revised form 28 April 2005; accepted 9 June 2005; published 14 Sept. 2005.

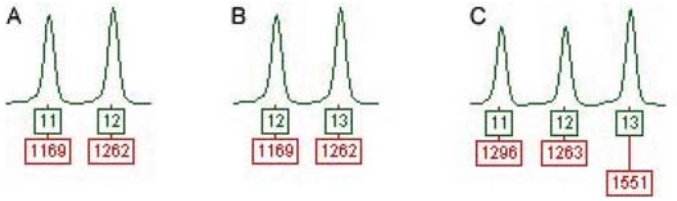

FIG. 1—*Sample electropherograms from possible single-source and mixed samples. A single STR locus with alleles 11 and 12 (A); the same STR locus but now with alleles 12 and 13 (B); and the same STR locus but now with three distinct alleles: 11, 12, and 13 (C). Allele number designations for each peak appear immediately below it with corresponding peak height information (in relative fluorescence units) immediately below.*

Materials and Methods

In order to conduct this study, a dataset containing complete STR-DNA profiles of several hundred individuals was needed. Such a dataset from the FBI, used for the determination of allele frequencies, has already been analyzed for Hardy-Weinberg equilibrium (10), and is publicly available (11). This FBI dataset consists of complete typing information for the 13 commonly used CODIS STR loci for 959 individuals from six different racial groups: Bahamian (153 individuals), Jamaican (157 individuals), Southwest Hispanic (202 individuals), Trinidadian (76 individuals), US African American (177 individuals), and US Caucasian (194 individuals). The original dataset (11) contains typing information for a larger number of individuals, but any with incomplete information, i.e., allele "0," were discarded for this study.

Published analyses of the FBI dataset make no specific mention of the extent of the effort made to assure that close relatives were not included in the population sampling (10). Thus, to guarantee that that there are absolutely no relation-based linkages among individuals for some aspects of this study, a dataset of "randomized individuals" was also generated. In this randomization, the actual alleles observed in the FBI dataset were distributed randomly to produce a new set of genotypes equal in number to the original dataset. Allele frequencies in this randomized dataset are the same as in the original dataset but individuals are unequivocally unrelated by descent (alleles are not the same because they have been faithfully passed from a common ancestor). Instead, any allele sharing can arise only through identity by state (alleles are the same because there is a finite number of different alleles that can be detected). Each locus was considered independently during the production of randomized genotypes. For each locus, the alleles of all individuals in the original dataset (without respect to racial classification) were randomly redistributed among the same number of synthetic individuals; an example of one possible redistribution amongst three individuals is shown in Table 1. This redistribution occurs without replacement, thus each locus in a randomized dataset has the same allele frequencies as the corresponding locus in the original dataset. Source code for all of the analyses can be found at (12).

All individuals in the original dataset are assumed to have two and only two alleles per locus (rare conditions resulting in unusual allele counts such as null alleles, triploidy or chimerism are beyond the scope of this study). Similarly, all simulated mixtures of genotypes are considered to be free of any typing errors that might further complicate the interpretation.

TABLE 1—*Example of alleles being redistributed amongst three individuals.*

	vWA	
Individual	Original	Redistributed
A	18, 19	15, 17
B	17, 18	18, 18
C	14, 15	14, 19

Shared Allele Counts

Homozygotes were deemed to share two alleles with other homozygotes with the same genotype (e.g. an individual who was 12, 12 was determined to share two alleles with another 12, 12 individual but none with a 10, 10 individual). Homozygotes could share either one or no alleles with heterozygotes (e.g., a 12, 12 homozygote would share one allele with an 11, 12 individual and none with a 10, 11 individual). Average shared allele counts were the average pair wise total (with a maximum of 26 arising from two alleles across 13 loci) number of shared alleles observed between all possible pairs of individuals in a dataset.

Shared Allele Counts with Related Individuals

The greater the number of shared alleles between pairs or clusters of individuals within a population, the greater the chance that maximum numbers of alleles observed per locus may suggest an incorrect minimum number of contributors. In order to determine upper bounds on the number of shared alleles observed between pairs of individuals, we consider a "worst case" situation. What would be observed if there were relatives of every individual in the dataset? To answer this question, virtual families of individuals were created using genotypes from the randomized dataset.

Each virtual family consists of two "cousins" (C1–C2), their four "parents" (P1–P4, of which P2 and P3 are siblings), and their six "grandparents" (G1–G6) (Fig. 2). The grandparents of each family are drawn from datasets of randomized individuals to preclude complications from the possible presence of related individuals already being present in the FBI dataset. Thus, each set of six randomly selected virtual individuals produces one family.

In this simulation, each of two parents contributes one of its alleles (chosen randomly) for each locus in the production of their virtual offspring. For each dataset of 959 individuals, 159 sets of six individuals were chosen to be "grandparents" to produce 159 3-generation families, each containing a cousin and sibling pair. From these synthetic families two datasets containing related individuals were created: one populated with pairs of siblings and one populated with pairs of cousins. In order to maintain similarity in the scope of the study, roughly the same number of two-person combinations in the virtual family datasets were considered as in the original dataset (459,361). This is done by choosing the grandparents randomly over the course of 2,889 runs (459,361 two-person combinations/159 virtual families).

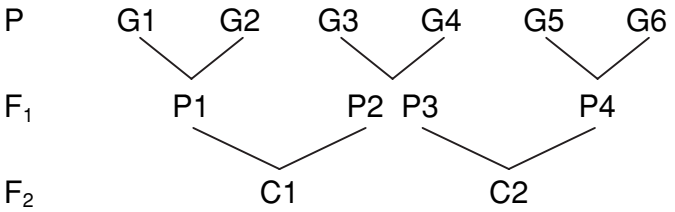

FIG. 2—*Creation of virtual families from the dataset. The individuals in the "P" line (G1–G6) are 6 profiles chosen randomly from the dataset of synthetic individuals, representing grandparents. They produce offspring shown in line F1 (parents P1–P4) with siblings P2/P3 as shown. Line F2 shows the grandchildren (C1, C2) of the original profiles, who are first cousins.*

Three-person Mixture Analysis

The number of three-person combinations of a set of n individuals is determined as:

$$N = \frac{n!}{(n-3)!3!}$$

where n is the number of individuals, and N the number of combinations. For instance, for 4 individuals (A, B, C, and D) there are $4!/(1! \cdot 3!) = 24/6 = 4$ different three-person combinations (ABC, ABD, ACD, and BCD). Given 959 individuals, each of the 146,536,159 different possible three-person 13 locus genotype combinations were considered and the number of different alleles represented at each of the 13 loci was determined for each mixture. The same process was also performed using the genotypes within each of five randomized datasets, and the results averaged.

If no more than two different alleles were observed at all of the 13 loci considered, then the three-person mixture was considered to be potentially mischaracterized as the profile of a single individual. Likewise, if no more than four alleles were observed over all loci, then the three-person mixture has the potential to be mischaracterized as a mixture of only two individuals. Potential differences in peak heights (i.e. due to additivity associated with shared alleles) were not considered in this study.

We have observed that the standard operating procedures of forensic DNA testing laboratories sometimes allow analysts to discard information from loci that they determine to be anomalous based on their training and experience. One factor that could conceivably cause a locus to appear anomalous would be the observation of five or six alleles (suggesting a minimum of three contributors to a mixture), while the other 12 loci possess only four or fewer different actual alleles (consistent with a mixture of two contributors). To assess the ramifications of invoking analyst discretion to discard such a locus, we analyzed the number of three-person genotype mixtures where discarding a single locus with the highest number of different observed alleles produces results consistent with mischaracterization of the mixture as a single source sample or as a two-person mixture.

Four-person Mixture Analysis

Computing all possible four-person mixtures of a set of 959 individuals is impractical (there are 35,022,142,001 such mixtures). Consequently, analyses of four-person mixtures was restricted to a subset of the FBI dataset (specifically, the 194 Caucasians) resulting in 57,211,376 different four-person mixtures. For this analysis, we assume that mixed genotypes that allow the observation of 7 or 8 different alleles at even one locus will be correctly identified as a four-person mixture. In the same way, mixed genotypes where the locus or loci with the greatest number of different alleles observed have either 5 or 6 alleles will be considered to be mistakenly characterized as a three-person mixture. Mixed genotypes where the locus or loci with the greatest number of different alleles have either 3 or 4 alleles observed will be considered to be mistakenly characterized as a two-person mixture.

The number of loci that need to be considered for at least 95% of the simulated four-person mixtures to be correctly characterized as having originated from at least four contributors (e.g., had at least one locus with 7 or 8 alleles) was empirically determined. Since 13 loci proves to be insufficient to reach 95% confidence, new virtual loci were introduced by randomly selecting one of the original 13 loci and creating a simulated locus with equivalent discriminating power by randomly redistributing alleles. This process was repeated in five parallel simulations until a level of 95% correct characterization was independently achieved in each simulation.

Results

Shared Allele Counts

The 959 individuals of the FBI dataset can also be combined in $n = 459{,}361$ different pairings. The distribution of the number of alleles shared between pairs of unrelated individuals shows expected similarity ($p = 1.00$ by a two-tailed t-test) between the original dataset ($\overline{x} = 8.59$, $\sigma = 2.16$) and the randomized dataset ($\overline{x} = 8.59$, $\sigma = 2.15$). No pairs of individuals were found to share more than 25 of 26 alleles in the original dataset, or 20 of 26 in the randomized datasets. The distribution of the number of alleles shared between virtual cousins ($\overline{x} = 10.95$, $\sigma = 2.27$) and virtual siblings ($\overline{x} = 16.94$, $\sigma = 2.30$) was significantly different ($p = 0.00$) as were the distributions for shared alleles between virtual siblings and unrelated individuals ($p = 0.00$) but not for virtual cousins and unrelated individuals ($p = 0.29$). The distributions for the number of shared alleles in pairings of randomized individuals, virtual siblings, and virtual cousins are roughly Gaussian (Fig. 3) as was the distribution of pair wise allele sharing in the original dataset.

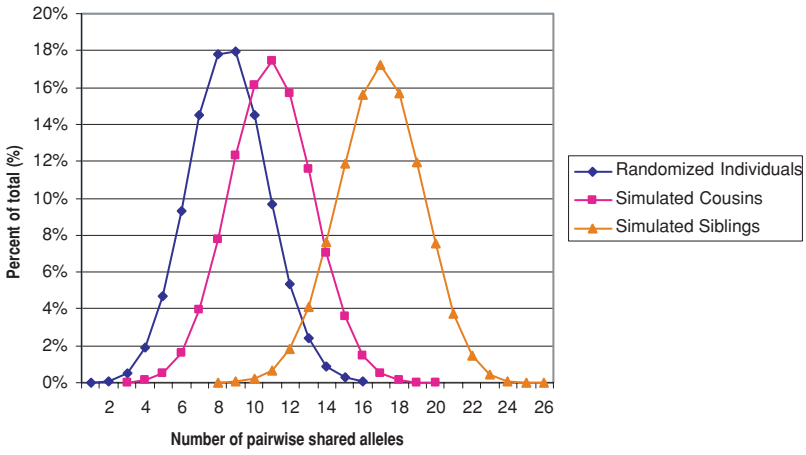

FIG. 3—*The distributions for the number of shared alleles amongst all possible pairs of synthetic individuals, pairs of synthetic siblings, and pairs of synthetic cousins. A total of 459,361 pairs each of randomized individuals, simulated cousins, and simulated siblings were considered.*

Shared Allele Counts with Related Individuals

The original dataset's mischaracterization rate of 3.39% (appearing to be a two-person mixture rather than a three-person mixture on the basis of maximum allele count per locus) is more than two standard errors of the mean above the average observed in the five datasets of randomized individuals. The higher mischaracterization level in the original dataset is attributable at least in part to a greater number of pair wise shared allele counts at or above 19 out of a possible 26 in the original dataset relative to each of the five randomized datasets (3 vs. an average of 1.4). Although neither of these numbers are statistically significant, given the averages and standard deviations, these unusual results are most easily explained by hypothesizing that one or more pairs of individuals in the original dataset may be related. This possibility was the motivation for producing and analyzing the random datasets against the original.

Analyses of virtual families suggest that such high levels of allele sharing are likely to be due to identity by descent, not by state (Fig. 3). It is worth noting that while it is uncommon for virtual siblings in this simulation to have indistinguishable genotypes (matching at all 26 alleles) an average of 3.0 such perfectly matching sibling pairs were generated in the five repetitions of this simulation (459,361 sibling pairs each).

Three-person Mixture Analysis

All possible different three-person mixtures of the 959 individuals in the FBI dataset were considered. Of those 146,536,159 three-person mixtures, 4,967,112 (3.39%) had contributors that possessed overlapping alleles such that none of the 13 loci exhibited more than four alleles (Table 2). Consequently, each of these 4,967,112 mixtures could be mischaracterized as a mixture of two individuals using information from maximum allele count alone. A smaller fraction ($\overline{x} = 3.18\%$, $\sigma = 0.2131\%$, and standard error of the mean $= 0.0953\%$) of similarly mischaracterized mixed profiles was found with the five randomized datasets. None of these three-person mixtures can be misinterpreted as a single contributor, as nowhere are there only one or two alleles observed across all loci.

Many of the simulated three-person mixtures were found to possess just one locus that contained more than four alleles. When these single loci were not considered (e.g., because they were "inconsistent with" or "anomalous relative to" the majority of loci) the mischaracterizations increased dramatically (Table 3) both in the original and five randomized datasets. In this case, 26,791,938 (18.28%) of the three-person mixtures from the original dataset could be mischaracterized as a mixture of just two individuals. An average of 25,591,392.6 or 17.46% ($\sigma = 0.5444\%$) were similarly mischaracterized in the five randomized datasets. As in the previous analysis, none of these three-person mixtures can be misinterpreted as a single contributor.

When this locus with the largest number of observed alleles is discarded, the category can change at most once. For example, if a mixture has only one locus having 6 alleles, one locus having 5 alleles, and the remaining loci with 4 or fewer alleles, it would be counted as changing from a maximum of 6 observed alleles to a maximum of 5 observed alleles (Table 4). Further, if the two highest maximum observed alleles counts are the same, no change occurred. Other useful information, such as per-locus information similar to Table 3, is available on-line (12).

TABLE 2—*Count and percent of three-person mixtures in which a particular number of unique alleles was the maximum observed across all loci, both for the original and randomized individuals**.

Unique Alleles	Count	Percent (%)
2	0	0.00%
3	78	0.00%
4	4,967,034	3.39%
5	93,037,010	63.49%
6	48,532,037	33.12%

A—Original dataset.

Unique Alleles	Count	Percent (%)
2	0.0	0.00%
3	115.8	0.00%
4	4,653,064.2	3.18%
5	92,019,609.6	62.80%
6	49,863,369.4	34.03%

B–Average over five randomized datasets.

* The unique allele column reports the maximum number of different alleles that were observed across all loci for these cenceptual three-person mixtures.

TABLE 3—*Count and percent of three-person mixtures in which a particular number of unique alleles was the second highest observed across all loci, both for the original and randomized individuals, after removal of the locus with the maximum number of unique alleles**.

Unique Alleles	Count	Percent
2	0	0.00%
3	3,398	0.00%
4	26,788,540	18.28%
5	112,469,398	76.75%
6	7,274,823	4.96%

A—Original dataset.

Unique Alleles	Count	Percent
2	0.0	0.00%
3	3,872.0	0.00%
4	25,587,520.6	17.46%
5	113,412,323.0	77.40%
6	7,532,443.4	5.14%

B—Average over five randomized datasets.

* The unique allele column reports the maximum number of different alleles that were observed across all loci for these conceptual three-person mixtures but only after a single locus with the highest count has been discarded.

Four-person Mixture Analysis

A large majority (43,667,840 or 76.34%) of the 57,211,376 possible four-person conceptual mixtures of the 194 Caucasians in the FBI dataset can be mischaracterized as two- or three-person mixtures when maximum allele count observed across all 13 loci was used as the only basis for characterization (Table 5). As expected, the mischaracterization rate decreases as the number of loci considered is increased (Fig. 4). The simulation was run five times, each run halting independently when the mischaracterization rate dropped below 5%. The fewest number of additional loci required to fall below a 5% level of mischaracterization in these simulations was 171 while the most was 177. The addition of information from an average of 27 simulated loci resulted in at least half of the four-person mixtures having at least one locus with more than six different alleles (and thus correctly classified). As with the three-person mixtures, per-locus data similar to Table 5A is available on-line (12).

Discussion

The extent of allele sharing observed between pairs of individuals is clearly influenced by the degree to which the individuals being

TABLE 4—*Count and percent of three-person mixtures in which a particular number of unique alleles was the second highest observed across all loci, both for the original and randomized individuals, after removal of the locus with the maximum number of unique alleles**.

Unique Alleles			Percent
From	To	Count	of Total
6	5	37,751,585	25.76%
6	4	3,505,340	2.39%
6	3	289	0.00%
5	4	18,317,945	12.50%
5	3	1,252	0.00%
4	3	1,779	0.00%

A—Original dataset.

Unique Alleles			Percent
From	To	Count	of Total
6	5	38,850,621.8	26.51%
6	4	3,479,995.2	2.37%
6	3	309.0	0.00%
5	4	17,456,523.4	11.91%
5	3	1,385.0	0.00%
4	3	2,062.2	0.00%

B—Average over five randomized datasets.

* The unique allele column reports the maximum number of different alleles that were observed across all loci for these conceptual three-person mixtures before and after a single locus with the highest count has been discarded. The middle four rows represent those cases where a change of interpretation has occurred, from three contributors to two.

TABLE 5—*Count and percent of four-person mixtures (from the FBI Caucasian dataset) in which a particular number of unique alleles was the highest observed across all loci, both for the original 13 loci and after the addition of 182 new loci derived from the original 13 loci**.

Unique Alleles	Count	Percent
4	13,480	0.02%
5	8,596,320	15.03%
6	35,068,040	61.30%
7	12,637,101	22.09%
8	896,435	1.57%

A—Original 13 loci.

Unique Alleles	Count	Percent
4	0	0.00%
5	0	0.00%
6	2,542,148	4.44%
7	45,542,753	79.60%
8	9,126,475	15.95%

B—After the addition of 182 new loci.

* The unique allele column reports the maximum number of different alleles that were observed across all loci for these conceptual four-person mixtures. Three or less alleles are never observed, and thus omitted.

compared are related to each other. Simulations confirm that first-degree relatives (siblings) are more likely to have alleles in common than second degree relatives (cousins) or unrelated individuals (13) (Fig. 3). The larger amount of pair wise allele sharing observed between individuals in the original FBI dataset relative to the five datasets of randomized individuals suggests that the FBI dataset may contain some pairs of closely related individuals. The extent of allele sharing between siblings in large-scale simulations also suggests that perfect 13 locus matches (26 out of 26 possible alleles) occur at a frequency (an average of 3.0 per 459,361). This frequency suggests that some are likely to exist in large populations such as the general population of the United States and even eventually in DNA profile datasets that contain large numbers of close relatives.

A large number (4,967,112) of the possible three-person combinations of the 959 actual individuals included in the FBI population dataset have sufficient allelic overlap between individuals to allow none of 13 STR CODIS loci to exhibit more than four alleles in a mixed sample (Table 2). This observation has important implications for the interpretation of forensic DNA testing results given that mixtures where both the number of contributors and the genotypes

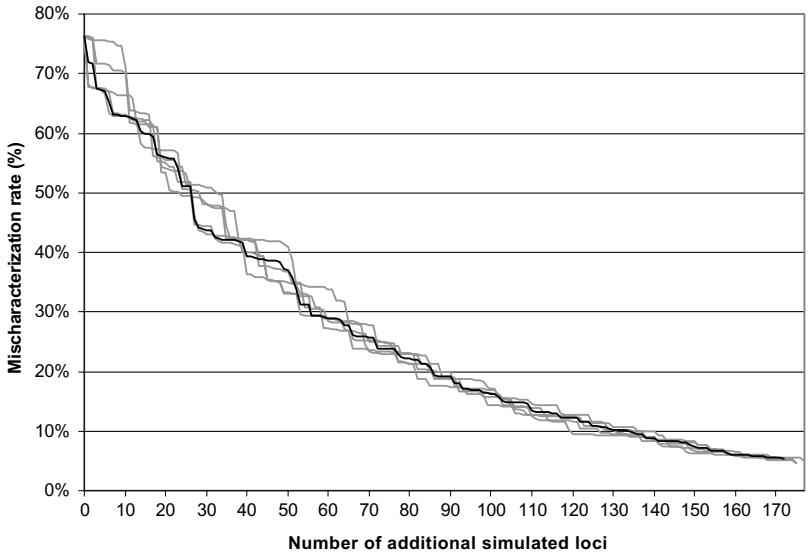

FIG. 4—*The percent of four-person mixtures that could be mischaracterized as arising from just two or three contributors solely on the basis of the maximum number of different alleles observed at any locus considered. Mischaracterization occurred at a rate of 76.34% when only the thirteen CODIS loci in the original FBI dataset were considered. A black line shows the average obtained from five simulations, each of which is shown individually in gray. Each simulation was halted when the mischaracterization rate dropped below 5%. A total of 57,211,376 conceptual four-person mixtures were made for each data point in each simulation as new loci were added.*

of one or more likely contributors are often disputed. Inclusions in such mixtures should always be accompanied by statistics that convey the strength of such a finding since allele sharing and large numbers of observed alleles both diminish the possibility of exclusion. Simulations with equivalent datasets containing randomized individuals yield similar results suggesting that the mischaracterization level is not attributable solely to artifacts or population substructure within the original FBI dataset (data not shown).

The mischaracterization of conceptual three-person mixtures increased more than five-fold when a single locus with the largest number of different alleles was eliminated from consideration. Reasoning along the lines of "a single locus with more than four alleles is likely to be attributable to technical artifacts when all other tested loci are consistent with there being only two contributors" is intuitively appealing. However, the observed dramatic increase in mischaracterization rate in both the original and five randomized datasets suggests that such rationalization is not well-founded.

Much higher rates (76.34%) of potential underestimates of the number of contributors to mixed samples were observed when four-person mixtures were considered. Difficulties in inferring the correct number of contributors to both three- and four-person mixtures are ultimately due to overlapping alleles between individuals. Allele sharing between two individuals is attributable to only two conditions: 1) identity by descent or, 2) identity by state. Allele sharing due to identity by descent can make mixtures involving related individuals particularly problematic. Allele sharing due to identity by state is a potential problem in all mixtures and arises from two related characteristics of the commonly employed STR CODIS loci. First, many of these loci possess a small number of detectable alleles (e.g., in this dataset, there are only six observed alleles for the TPOX locus and only seven for the D13, D5, D3, and TH01 loci, therefore mixtures that display more than six alleles at these loci must be rare). Second, some alleles at some loci are relatively common and therefore likely to overlap between contributors to a mixture. The key factor is that the addition of more individuals (and thus more alleles) into the mixture causes the mixture to become more likely to hide any indications of subsequent individuals, as the relative proportion of present versus absent alleles at each locus increases with each new contributor.

Conclusions

Maximum allele count by itself is not very reliable in terms of predicting the number of contributors to mixed forensic DNA samples, particularly when: the number of loci considered is small, the number of contributors may be large (4 or more), and/or a single point of seemingly inconsistent/anomalous information can be disregarded. However, maximum allele count still results in mistaken inference of the number of contributors at a rate of over 3% even in the best of circumstances (e.g., the 13 STR CODIS loci are considered, there are only three contributors, and no seemingly inconsistent/anomalous information is disregarded). In light of these observations, the practice of many testing laboratories to simply report that a sample arises from "two or more individuals" when more than two alleles are observed at one or more loci during the course of testing is both reasonable and appropriate.

Acknowledgments

We gratefully acknowledge both Jason Gilder and Carolyn Rowland for their assistance and critical reviews of this manuscript. The comments of two anonymous reviewers were also extremely helpful and very much appreciated.

References

1. Edwards A, Hammond HA, Lin J, Caskey CT, Chakraborty R. Genetic variation at five trimeric and tetrameric tandem repeat loci in four human population groups. Genomics 1992;12:241–53. [PubMed]
2. Frégeau CJ, Fourney RM. DNA typing with fluorescently tagged short tandem repeats: a sensitive and accurate approach to human identification. BioTechniques 1993;15:100–19. [PubMed]
3. Krenke BE, Tereba A, Anderson SJ, Buel E, Culhane S, Finis CJ, Tomsey CS, Zachetti JM, Sprecher CJ. Validation of a 16-locus fluorescent multiplex system. J Forensic Sci 2002;47(4):773–85. [PubMed]
4. Moretti TR, Baumstark AL, Defenbaugh DA, Keys KM, Smerick JB, Budowle B. Validation of short tandem repeats (STRs) for forensic usage: performance testing of fluorescent multiplex STR systems and analysis of authentic and simulated forensic samples. J Forensic Sci 2001;46(3):647–60. [PubMed]
5. Moretti TR, Baumstark AL, Defenbaugh DA, Keys KM, Brown AL, Budowle B. Validation of STR typing by capillary electrophoresis. J Forensic Sci 2001;46(3):661–76. [PubMed]
6. Frégeau CJ, Bowen KL, Fourney RM. Validations of highly polymorphic fluorescent multiplex short tandem repeat systems using two generations of DNA sequencers. J Forensic Sci 1999;44(1):133–66. [PubMed]
7. Perlin MW, Szabady B. Linear mixture analysis: a mathematical approach to resolving mixed DNA samples. J Forensic Sci 2001;46(6):1372–7. [PubMed]
8. Wang T, Xue N, Rader M, Birdwell JD. Least square deconvolution (LSD) of STR/DNA mixtures. 7th CODIS User's Conference. Available at http://www.lit.net/presentations2/wang_lest_square%20STRDNA_2001_3.pdf. 2001.
9. Gill P, Sparkes R, Klimpton C. Development of guidelines to designate alleles using an STR multiplex system. Forensic Sci Int 1997;89(3):185–97. [PubMed]
10. Budowle B, Moretti TR, Baumstark AL, Defenbaugh DA, Keys KM. Population data on the thirteen CODIS core short tandem repeat loci in African Americans, U.S. Caucasians, Hispanics, Bahamians, Jamaicans, and Trinidadians. J Forensic Sci 1999;44(6):1277–86.
11. http://www.fbi.gov/hq/lab/fsc/backissu/july1999/dnaloci.txt.
12. birg.cs.wright.edu/~dpaolett/JFS05-materials.zip.
13. Presciuttini S, Ciampini F, Alu M, Cerri N, Dobosz M, Domenici R, Peloso G, Pelotti S, Piccinini A, Ponzano E, Ricci U, Tagliabracci A, Baley-Wilson JE, De Stefano F, Pascali V. Allele sharing in first-degree and unrelated pairs of individuals in the Ge.F.I. AmpFlSTR® Profiler Plus database. Forensic Sci Int 2003;131(2–3):85–9.

Additional information and reprint requests:
Dan E. Krane, Ph.D.
Biological Sciences
3640 Colonel Glenn Highway
Dayton, OH 45435-0001
E-mail: Dan.Krane@wright.edu