Keith L. Monson,[1] *Ph.D. and Bruce Budowle,*[1] *Ph.D.*

A Comparison of the Fixed Bin Method with the Floating Bin and Direct Count Methods: Effect of VNTR Profile Frequency Estimation and Reference Population

REFERENCE: Monson, K. L. and Budowle, B., **"A Comparison of the Fixed Bin Method with the Floating Bin and Direct Count Methods: Effect of VNTR Profile Frequency Estimation and Reference Population,"** *Journal of Forensic Sciences*, JFSCA, Vol. 38, No. 5, September 1993, pp. 1037–1050.

ABSTRACT: When the results of a forensic comparison of highly polymorphic variable number tandem repeat (VNTR) loci fail to exclude a suspect as a possible contributor of biological evidence, it is desirable to convey to the trier of fact the significance of the match. Furthermore, in a forensic context, it is desirable that the estimated frequency of occurrence be conservative, that is, that any uncertainty in the estimate will favor the accused. Using an empirical approach with a data base of 2046 individuals belonging to one of four population groups, this study examined the effect of the method used to estimate frequency of occurrence of a VNTR profile from a reference data base, and the consequences of using a data base that may not represent the circumstances of the crime. The fixed bin method was at least as conservative as the floating bin and genotype counting (direct counting) methods. Secondly, for forensic purposes, profile frequency estimates from different reference populations do not deviate greatly. VNTR profiles are rare in any of the data bases.

KEYWORDS: pathology and biology, VNTR, allele frequency, population databases

In a forensic analysis, DNA profiles from a suspect or victim are compared with those derived from evidentiary material recovered from a crime scene. If the profiles are operationally similar, so that they can not be excluded as originating from the same source, they are deemed a "match." The most polymorphic and therefore highly individualizing genetic markers currently available for these kinds of comparisons are variable number tandem repeat (VNTR) loci identified by single-locus probes (SLP). However, a single SLP-VNTR profile obtained by restriction fragment length polymorphism (RFLP) analysis is not unique. To determine the significance of a match, the appropriate hypothesis is to assume the suspect is *not* the contributor of the sample and then to determine what portion of the population of potential perpetrators might be responsible for the sample. In other words, we assume the suspect is innocent and ask what the chance is of anyone else leaving the material. The probability that a similar profile could occur in another person is therefore calculated from a relevant population data base, consisting

Received for publication 1 Sept. 1992; revised manuscript received 31 Dec. 1992 and 12 March 1993; accepted for publication 26 March 1993.

[1]Forensic Science Research and Training Center, FBI Academy, Quantico, VA 22135.

This is publication number 92-08 of the Laboratory Division of the Federal Bureau of Investigation. Names of commercial manufacturers are provided for identification only and inclusion does not imply endorsement by the Federal Bureau of Investigation.

of VNTR profiles of a group of unrelated individuals. In a forensic context it is desirable that the estimate be conservative, that is, that any uncertainty in the estimate will favor the accused.

Due to the nature of analysis, fragment lengths of VNTR alleles form a quasi-continuous distribution [*1*]. Since it is not feasible to measure the length of an RFLP fragment exactly, several operationally defined methods have been proposed to estimate the frequency of occurrence of each of the alleles observed at several VNTR loci. Then, the probability of observing the same profile at random is determined by assuming Hardy-Weinberg frequencies and independence between loci and calculating the expected phenotype/genotype frequency, that is, frequencies are multiplied across loci. Two approaches involving binning of population data have been described to accommodate statistical analysis, the fixed bin method [*1*] and the floating bin method [*2*]. (A variation of a binning approach, the "ceiling principle," has been reported recently [*3*].) A third approach, called genotype (or direct) counting, consists of counting the number of times an observed profile has been seen in the data base. Each method is summarized in the following.

The fixed bin method [*1*] is the predominant method for RFLP profile frequency assessment in North American crime laboratories. It categorizes fragment lengths measured in a population sample according to a set of arbitrarily defined fixed boundaries, which must be wider than the measurement error of the analytical system. The observed measurement error in the FBI fixed-bin system is $\pm 2.5\%$ of the estimated base pair size of a DNA fragment, or a total of 5% [*1*]. The number of DNA fragment lengths that fall into each bin, when divided by the total number of fragments measured in a sample population, determines the frequency of a bin. Bins with fewer than five counts are merged with contiguous bins. The frequency of an observed allele is estimated by calculating the size range in which the allele may fall ($\pm 2.5\%$ of the measured range) and determining in which bin(s) the fragment could reside. If the size range spans a bin boundary, the allele is assigned to the higher frequency bin. Advantages of the fixed bin method include dynamic compensation for rare alleles in a data base of limited size, provision for assigning frequency estimates to alleles not yet observed, overestimation of allele frequencies, and convenient distribution of concise frequency tables rather than raw population data.

In the floating bin method [*2*], a window is calculated around the observed fragment size, and alleles occurring within this window in the reference data base are summed and divided by the total number of chromosomes in a sample population to arrive at the estimated frequency of occurrence. The size of the window is based on observed measurement error. The floating bin approach has been shown to provide valid estimates for the frequency of occurrence of a VNTR band [*4*].

Direct count methods tally the number of times a band pattern has been seen previously in a data base. This approach can be applied either on a per locus basis or across an entire profile. Direct count methods are related to the floating bin method. In the floating bin method, counts are summed within a window established around each allele. In direct count methods the profiles of individuals in a data base that overlap windows around each allele making up a profile are summed for the purpose of estimating frequency of occurrence.

Direct counting of the number of times a *composite* profile (consisting of results across several loci) has been seen produces a value of n_{obs} in a data base of n_{tot} individuals. The estimated frequency would be n_{obs}/n_{tot}. Since there is no multiplication of frequencies across loci, fewer population genetic principles (such as Hardy-Weinberg and gametic phase equilibria) are invoked, but this technique depends primarily on the size of the data base and is relatively insensitive to the number of loci typed. Thus, this method ignores Mendelian principles and fails to consider that for highly polymorphic VNTR loci more genotypes exist than are revealed even in a reasonably large data base. For

example, in a data base of 359 profiles for the single VNTR locus D1S7, when 351 possible genotypes were conservatively defined by 26 fixed bin boundaries, over 190 genotypes were observed with only seven genotypes observed at least five times [5]. Additionally, direct counting across an entire profile (consisting of results from several loci) fails to adequately convey the rarity of a given profile. Our experience with SLP-VNTR suggests that profiles consisting of results from several loci are so rare that it would be highly unusual to find a match, and that typically, we would obtain a frequency of $1/n_{tot}$. Although it is sometimes suggested that the observed profile be added to the data base, resulting in a frequency of $1/(n_{tot} + 1)$, such a procedure is statistically weak, since the goal is to estimate the chance that another person might share the same profile. It is more defensible to use an upper bound on zero frequency, which can be estimated by confidence limits. Finally, direct counting across an entire profile is illogical because it fails to account for increasing statistical dimensions. If no matching profile was found across three loci, for example, the calculated probability would be unchanged if a fourth locus were probed and still no match were found.

A second formulation of the direct count method counts the number of times the pattern from *each* of the loci probed occurs in the data base. In this approach, traditionally used for blood group markers, the frequencies from each locus are then multiplied together to estimate the frequency of the composite profile. The advantage over the first method of direct counting is that we have a better expectation of observing a few single-locus matches in a data base of reasonable size, but genetic principles must still be invoked to justify multiplication across loci. Since neither direct count method nor the floating bin method inherently compensate for the low level of statistical confidence in small data bases and infrequently observed events, use of a minimum frequency may be advisable.

Although a binning approach overestimates the frequency of any single allele by grouping alleles together [9], some critics charge that the extent of conservatism is undemonstrated, particularly after multiplication of frequencies across loci [7,8]. This paper follows an empirical approach to address two pragmatic issues relevant to forensic applications: (1) the performance of the fixed bin method, which is used by the majority of North American forensic laboratories to estimate the frequency of occurrence of VNTR profiles from a sample population data base, compared with the floating bin and direct counting methods, and (2) what are the forensic consequences of using an inappropriately assigned general reference data base, which may not represent the circumstances of the crime, for calculations.

Materials and Methods

Sampling and Measurement

DNA samples were collected from 2046 unrelated individuals who were identified as either African-American, Caucasian, or Hispanic by surnames and/or by self-identification. There were 828 Caucasians and 579 African-Americans, 328 Hispanics from Florida, and 311 Hispanics from California and Texas [10]. RFLP analysis of the loci D2S44 [11], D1S7 [12], D17S79 [2], and D4S139 [13] was performed according to the methods of Budowle and Baechtel [14] following digestion by the restriction endonuclease HaeIII. DNA fragment lengths of the VNTR profiles were determined by comparison with a ladder of digested viral DNA consisting of 30 size standards ranging from 640 to 23,410 bp (Lifecodes, Stamford, CT) using a computer-assisted image analysis system [15].

Calculation of Probabilities

In these experiments, several conventions were followed for both the fixed and floating bin methods in order to facilitate fair and realistic comparisons. The conventions are

generally the same used by the FBI in case work evaluation [*1*]. The single-locus frequency of a two-band pattern was calculated using 2pq, where p and q are the estimated binned allele frequencies where each of the two VNTR bands reside (including effects of measurement error). Due to an inability to differentiate a true homozygote from a pseudo-homozygote (a single band pattern which may have resulted from one fragment being below the level of detectability of an autoradiogram, or the existence of a very small allele which has run off the gel, or the inability to resolve closely spaced alleles), the frequency of occurrence of a one-band pattern was calculated using 2p [*1*]. The frequency of occurrence of a profile made up of several single-locus profiles was calculated by multiplication across all loci. Issues regarding assumption of Hardy-Weinberg equilibrium and multiplying across loci have been addressed elsewhere [*5,16–19*], and will not be reconsidered here. The single-locus types of the very few DNA samples exhibiting three-band patterns at the D4S139 or D17S79 locus were not included in this study. If no assay results were available for a specimen at a particular locus, a frequency of 1.00 was assigned to that locus. Since measurement error increases markedly for fragments above 10 kB, any profile at a particular locus that contained an allele greater than 10 kB also was assigned a locus frequency of 1.00. Additionally, since the size of fragments less than 640 base pairs can not be ascertained, any profile containing such a fragment was assigned a frequency of 1.00.

In the application of the floating bin and direct counting approaches, individuals of one group sometimes exhibit alleles not observed in other groups. (This is not an issue with the fixed bin approach, which automatically provides for previously unobserved alleles). Therefore, a minimum frequency was used for cross-group calculations with these methods.

To be consistent with the procedures used in the fixed bin method, a minimum frequency of five counts divided by the total number of alleles was used in the floating bin method. Windows of 5% and 10% ($\pm 2.5\%$ and $\pm 5.0\%$, respectively) around measured fragment lengths were compared with the fixed bin calculations.

Preliminary studies showed that the second method of direct counting, where matches at each of the loci making up an observed profile are counted and frequencies are multiplied across loci, would lead to total probabilities much less conservative than either the fixed or floating bin method. This observation is expected, because the data bases are not large enough to observe all possible genotypes. In order to create a more conservative comparison with the other methods, a minimum frequency was calculated. For a data base of size n which is available at a given VNTR locus, an upper bound on the genotype frequency at the $100(1 - \alpha)\%$ confidence limit is given by $1 - \alpha^{1/n}$ [*5*].

Results and Discussion

Fixed Bin v. Floating Bin

Figures 1 a to d and 2 a to d compare the probability of occurrence of RFLP profiles (from 1–4 locus profiles) from various reference populations, estimated by the fixed bin and floating bin methods. Each data point represents the profile for each of the 2046 individuals, and each figure represents the use of either an African-American, Caucasian, southeastern Hispanic, or southwestern Hispanic reference population. Therefore, all individuals, regardless of racial/ethnic background, have been evaluated in each reference data base. This permitted evaluation of some individuals from the appropriate data base, while others served for evaluation of the forensic consequences of using an inappropriate data base. Due to the operational constraints for VNTR alleles whose sizes fall outside the range of 640 to 10090 base pairs, described in Materials and Methods, and since some profiles contain information from fewer than four loci, the scatter plots indicate

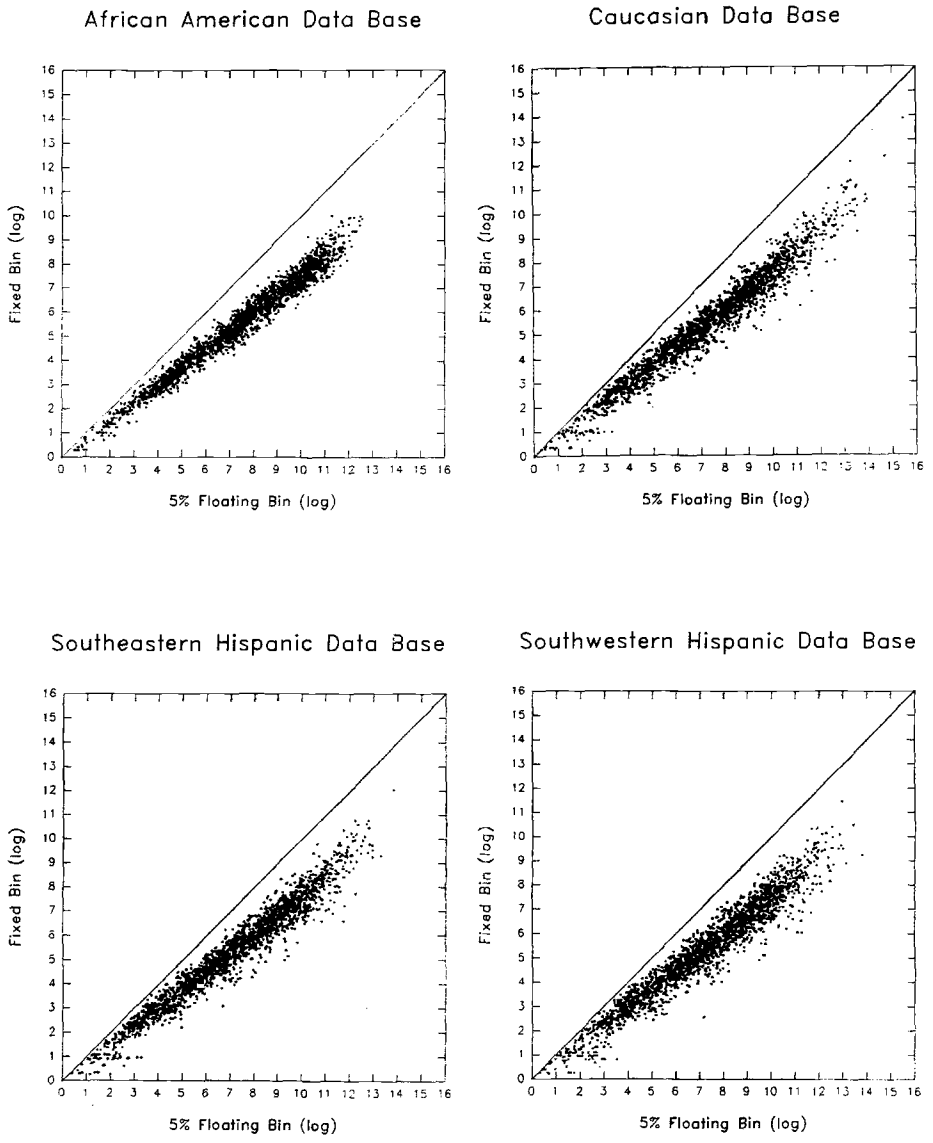

FIG. 1—*Comparison of probabilities estimated by the fixed bin and 5% floating bin methods for 2046 individuals from four data bases, using the subset of those individuals belonging to each of the following populations as reference data base: (a) African-American; (b) Caucasian; (c) Southeastern Hispanic; (d) Southwestern Hispanic. The diagonal indicates the line where each method would produce the same estimate. Refer to text for conventions followed.*

relative rarity of VNTR profiles for situations encountered in case work. Thus, points which fall nearer the origin tend to be profiles from a single locus, while those furthest from the origin would be four locus profiles. The diagonal on each plot indicates the theoretical line where both methods would produce the same result. Generally, when the data points fall close to the diagonal line, or cluster 50% above and 50% below line, the fixed and floating bin methods may be said to yield similar results.

The allelic frequency when using the floating bin method can be estimated within a

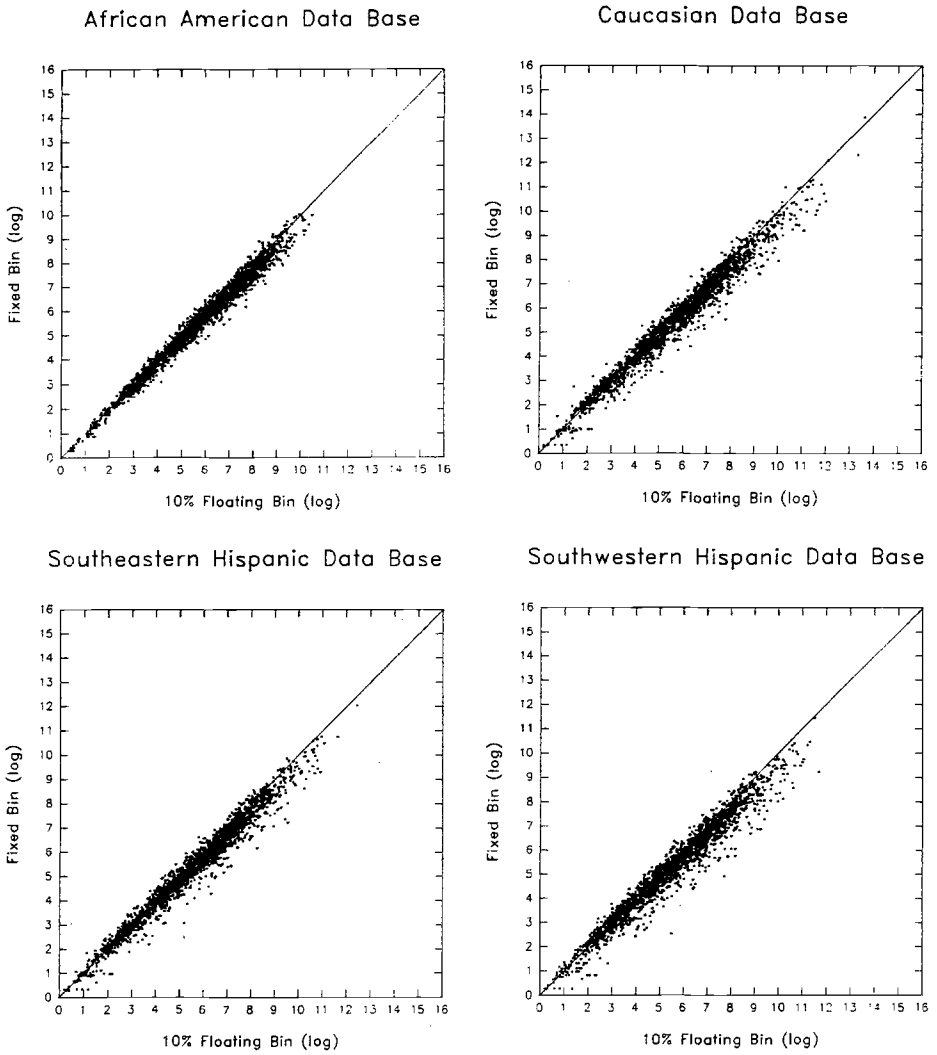

FIG. 2—*The same comparisons described in Fig. 1, but using a 10% floating bin. Instances where the 10% floating bin method is more conservative than the fixed bin method (consistently less than a factor of ten) are tabulated in Table 1.*

window of 5% (±2.5%). In Figs. 1 a to d, regardless of reference population, the fixed bin approach was more conservative in all but one case (of 2046 tests), compared with the 5.0% (±2.5%) floating window. This single exception involved a specimen from an African-American individual tested against the Caucasian data base, an individual for whom results were available only for the locus D17S79. In this instance, both methods gave the same result, about 1 chance in 2. (The fact that this individual happens to have a single-band pattern at locus D17S79, which is less polymorphic than the other loci probed, and the conservative manner in which the probability is calculated for single-band patterns account for this very nondiscriminating value.)

Figures 2 a to d compare the fixed bin method with a 10% floating bin method. Figures 2 a to d indicate that the fixed bin and floating bin methods tend to yield similar estimates for a floating bin window of 10%. However, the fixed bin method is usually more conservative. The number of specimens for which the 10% floating bin method results in an overall probability which is more conservative than that calculated by the fixed bin method are tabulated in Table 1. Even when the 10% floating bin is more conservative, all but one of the data points differ from the fixed bin estimation by less than a factor of ten. Variations of one order of magnitude are forensically insignificant when the estimated inclusion values are extremely small. In the single exception, again a specimen from an African-American individual referenced to the Caucasian data base, the probability estimates were 1 in 558 for the fixed bin method and 1 in 28 using a 10% floating bin (a factor of 19.7).

It has been suggested that bin boundaries could bisect a true allele, resulting in an underestimate of the true allele frequency and a bias against a defendant [*1,20,21*]. While this may occur with either a fixed or floating bin approach, it is of greater concern with the fixed bin method. An artificial boundary is not based on the underlying biological distribution of the alleles in a data base and, therefore, by chance could bisect a true allele and possibly result in an underestimate. The large size of the fixed bins relative to measurement error (and thus the overall large allele frequency estimate), in addition to deferral to the larger bin after measurement error range determination would argue against such an occurrence having an appreciable effect. Notwithstanding, it has been suggested that bins could be combined when a boundary bisects a peak in the population distribution [*3,21*]. The data within Figs. 1 and 2 show no significant impact of this phenomenon when a fixed bin method is employed, whether 1, 2, 3, or 4 loci are analyzed, and regardless of the reference data base. These results indicate that even if this effect were to occur in one bin, the conservative estimation at the remaining bins (large bins relative to measurement error, deferral to the higher bin in case of overlap and use of 2p for single-band patterns) would compensate to produce a reasonable estimate for forensic purposes, and that the proposed combination of bisected bins is unnecessary. Additionally, Chakraborty et al. [*9*] have shown that for the FBI population data bases, such summing of bins is not required.

TABLE 1—*Number of specimens from each of four population groups (and percentage of the number of individuals in each population group) for which the 10% floating bin method results in an overall probability that is more conservative than the fixed bin method. Figures in parentheses indicate the total number of specimens in each reference population. In only one case did the two differ by more than a factor of ten (where the factor was 19.7).*

Test Population	Reference Population			
	African-American	Caucasian	Southeast Hispanic	Southwest Hispanic
African-American	154	123	122	115
(579)	26.6%	21.2%	21.0%	19.9%
Caucasian	194	357	309	258
(828)	23.4%	43.1%	37.3%	31.2%
Southeast Hispanic	82	103	90	79
(328)	25.0%	31.4%	27.4%	24.1%
Southwest Hispanic	67	101	80	107
(311)	21.5%	32.5%	25.7%	34.4%
Combined	497	684	601	559
(2046)	24.3%	33.4%	29.4%	27.3%

Fixed Bin v. Composite Direct Count

Because of the limitations described in the opening paragraphs, and since each profile consisting of three or four loci matched only itself, the method of directly counting the number of times a profile (consisting of the results from several loci) was observed was uninformative: in all cases the estimated frequency was clamped by the size of the data base at $1/n_{tot}$.

Fixed Bin v. Direct Count (by locus)

It is strongly reflective of the polymorphism of VNTR loci that for each of the 2046 individuals studied, and using any of the four reference data bases, the frequency of the pattern revealed at each locus (calculated by direct counting before imposition of a minimum frequency) never exceeded the minimum frequency (the 95% upper confidence level, calculated as a function of the number of samples for which results were available at that locus). Figure 3 is typical of comparisons between the fixed bin and direct count methods (by locus, using the 95% upper confidence level). The direct count method applied in this way produces the same estimate for probability for every specimen probed at the same number of loci. Thus in Table 2, the direct count probability estimate for each specimen generally takes on one of four values (depending on whether the specimen was probed at one, two, three, or four loci), which are dependent only on the size of the data bases in this study.

Use of a 95% upper confidence level notwithstanding, the fixed bin method is at least as conservative as the direct count method. A comparison of the number of times that the direct count method (by locus) is more conservative than the fixed bin method for each of the four data bases is tabulated in Table 3. In the African-American and Caucasian data bases, the direct count method is more conservative in 7.5% of the 2046 samples, while it is about 19.5% in the two Hispanic data bases. Of those samples, most are within one order of magnitude. The estimates which differ by more than that are of little forensic significance, as they are typically in the 1 in 10^6 to 10^{13} range. Data bases of comparable size produce similar results, due to the dependence of the direct count method on the size of the data base. The trend is clear, however, that the fixed bin method tends to be more conservative, becoming more so as the size of the data base increases (within practical limits of data base size).

Cross-Group Comparisons

The forensic consequences of using a data base which may not precisely represent the demographics of the locale in which a crime was committed, or the racial/ethnic background of the accused, were examined empirically. Estimated probability of occurrence of the RFLP profile determined for every one of the 2046 individuals was calculated by the fixed bin and direct count (by locus, using the 95% upper confidence level) methods, using as a reference data base each of four constituent groups (African-American, Caucasian, southeastern Hispanic, and southwestern Hispanic). The results of every possible cross-group comparison of probabilities calculated by each method are presented as log-log scatter plots in Figs. 4 a to f. The diagonal indicates the theoretical line where the data base of both groups would lead to the same estimated probability.

Figures 4 a to f show that for all cross-group pairs the data points (each of which represents one of the 2046 specimens) fall near the diagonal. Table 4 presents fixed bin probability comparisons for six reference data base pairs. Cross-group deviations tend to be small: in 87.3 to 98.7% of cases, regardless of reference data base used, the estimated probabilities were within one order of magnitude, while 98.6 to 100% were within two orders of magnitude and 99.9 to 100% were within three orders of magnitude. When

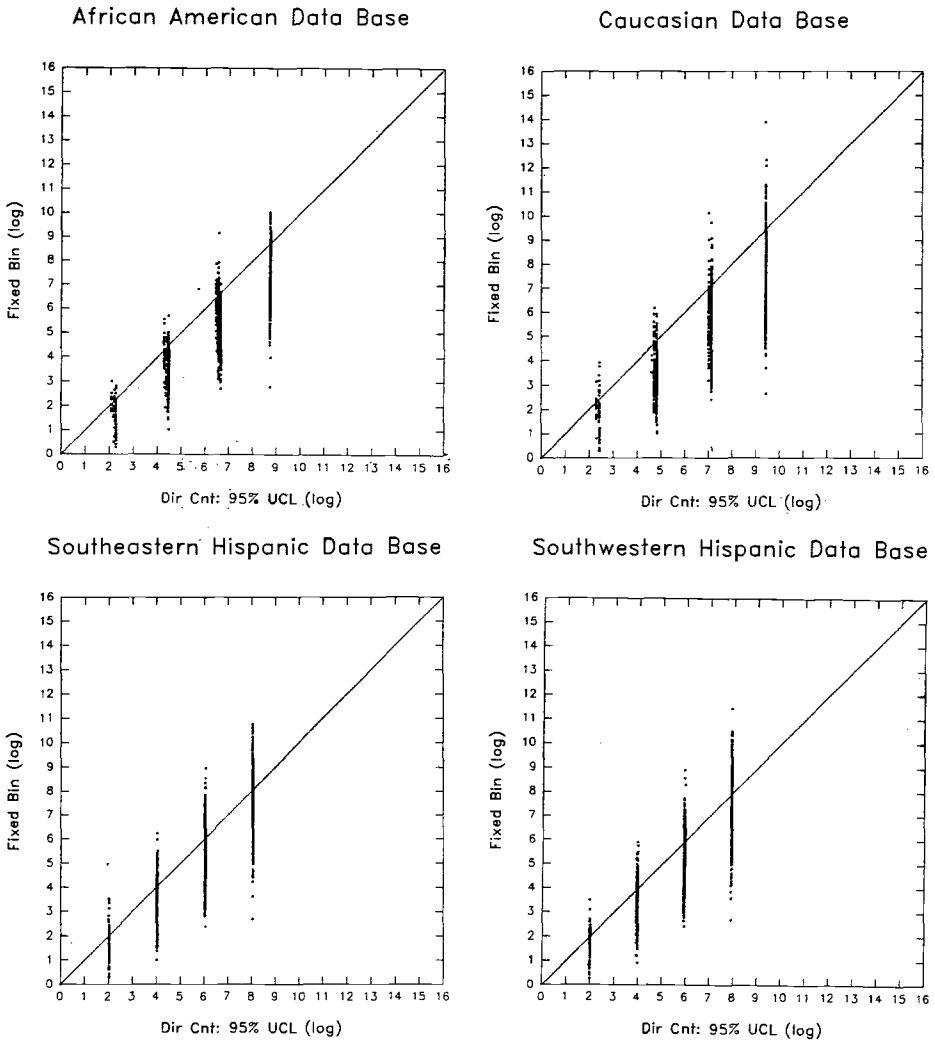

FIG. 3—*Comparison of probability estimates by the fixed bin and direct count (by locus, using 95% upper confidence limit as a minimum frequency) methods. The diagonal indicates the line where each method would produce the same estimate. Cases where the direct count method is more conservative, and within what order of magnitude, are tabulated in Table 3 and discussed in the text.*

differences of two or three orders of magnitude exist, they typically are in probability estimates of 1 in 10^6 or rarer.

The finding that the estimated probability of most profiles falls within one order of magnitude regardless of reference data base used should not be construed as a recommendation that any individual probability estimate be reduced by a factor of ten to allow for potential misassignment of reference data base. It has already been established that the fixed bin method is on average more conservative than a 10% floating window (which is twice the width required to include the system measurement error). The maximum allowable matching window (for the FBI) for determining whether or not DNA profiles are operationally similar is approximately 5% of the estimated fragment length of the observed allele. The fixed bin method reduces potential effects of substructure within a

TABLE 2—*Minimum frequency (95% upper confidence limit) used in the direct count method (by locus) calculated according to the formula given in the text. Figures in parentheses indicate the number of specimens in each reference population for which results were available at the given locus.*

Locus	Reference Population			
	African-American	Caucasian	Southeast Hispanic	Southwest Hispanic
D2S44	0.0063 (475)	0.0038 (792)	0.0099 (300)	0.0105 (284)
D1S7	0.0083 (359)	0.0050 (595)	0.0098 (305)	0.0103 (288)
D17S79	0.0054 (550)	0.0039 (776)	0.0095 (314)	0.0102 (293)
D4S139	0.0067 (448)	0.0050 (594)	0.0096 (311)	0.0112 (265)
Composite Freq. (four loci)	1.89×10^{-9}	3.68×10^{-10}	8.85×10^{-9}	1.24×10^{-8}
Composite Prob. (four loci)	5.29×10^{8}	2.72×10^{9}	1.13×10^{8}	8.05×10^{7}

TABLE 3—*Number of specimens (and percentage of the total of 2046 specimens) for which the direct count method (using a minimum frequency equal to the 95% upper confidence level when a profile has not yet been observed in the data base and multiplication of frequencies across loci) results in an overall probability value* (P) *more conservative than the fixed bin method, using each of four populations as reference data base.*

Test Population	Probability					
	$P \leq 10^1$	$P > 10^1$ $P \leq 10^2$	$P > 10^2$ $P \leq 10^3$	$P > 10^3$ $P \leq 10^4$	$P > 10^4$ $P \leq 10^5$	Sub-Total
African-American	136	15	1	0	0	152
	6.6%	0.7%	<0.1%	0	0	7.4%
Caucasian	120	29	3	1	1	154
	5.9%	1.4%	0.1%	<0.1%	<0.1%	7.5%
Southeast	295	85	18	1	0	399
Hispanic	14.4%	4.2%	0.9%	<0.1%	0	19.5%
Southwest	315	75	11	1	0	402
Hispanic	15.4%	3.7%	0.5%	<0.1%	0	19.6%

general population grouping. Any possibility of misassignment of a general population group may be addressed by estimating the probability of the individual profile in all appropriate data bases.

Similar conclusions may be drawn from cross-group comparisons using the direct count method (Table 5): probabilities estimated using the African-American data base when compared with the Caucasian or with either Hispanic data base are all within an order of magnitude, as are those calculated using either of the Hispanic data bases. Roughly two-thirds of the comparisons using the Caucasians and either of the Hispanic data bases, as reference, result in direct count probability estimates which fall within one and two orders of magnitude of one another. This observation contrasts with comparisons of the four other data base pairs by the direct count method and seems to contradict the results in Table 4 using the fixed bin method. However, the source of the contradiction is an artifact of the direct count method rather than a previously unappreciated difference

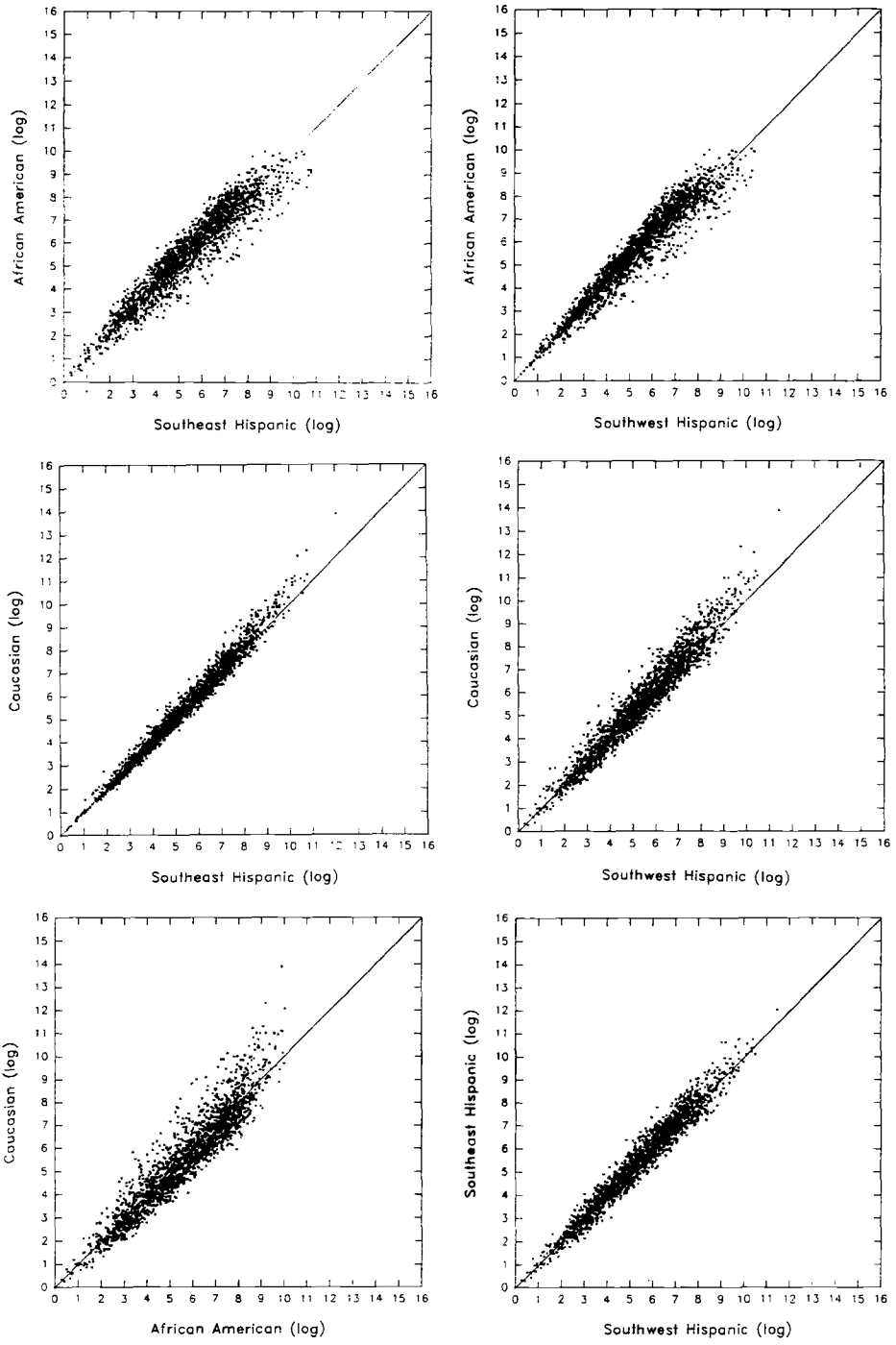

FIG. 4—*Effect of using different reference data bases to calculate probability estimates for 2046 individuals from four data bases by the fixed bin method: (a) African-American v. Southeastern Hispanic; (b) African-American v. Southwestern Hispanic; (c) Caucasian v. Southeastern Hispanic; (d) Caucasian v. Southwestern Hispanic; (e) Caucasian v. African-American; (f) Southeastern Hispanic v. Southwestern Hispanic. The diagonal indicates the line where either data base would produce the same estimate. As shown in Table 4, estimates differ by less than a factor of ten for 87 to 98% of the profiles, regardless of data base used.*

TABLE 4—*Comparison of reference data bases used to calculate probability of random occurrence* (P) *of an RFLP profile for 2046 individuals drawn from four populations by the fixed bin method. The number of samples and the percentage of the total for which the absolute value of probabilities falls within the indicated order of magnitude are shown for six reference data base pairs.*

Population Comparison	Probability $\lvert P\rvert \le 10^1$	$\lvert P\rvert > 10^1$ $\lvert P\rvert \le 10^2$	$\lvert P\rvert > 10^2$ $\lvert P\rvert \le 10^3$	$\lvert P\rvert > 10^3$ $\lvert P\rvert \le 10^4$
African-American v.	1853	186	7	0
Southeast Hispanic	91.6%	9.1%	0.3%	0
African-American v.	1883	159	4	0
Southwest Hispanic	92.0%	7.8%	0.2%	0
Caucasian v.	2019	27	0	0
Southeast Hispanic	98.7%	1.3%	0	0
Caucasian v.	1909	132	5	0
Southwest Hispanic	93.3%	6.5%	0.2%	0
Caucasian v. African-	1786	231	27	2
American	87.3%	11.3%	1.3%	0.1%
Southeast Hispanic v.	2007	39	0	0
Southwest Hispanic	98.1%	1.9%	0	0

TABLE 5—*Comparison of reference data bases used to calculate probability of random occurrence* (P) *of an RFLP profile for 2046 individuals drawn from four populations, by the direct count method (using multiplication of frequencies across loci and a minimum frequency equal to the 95% upper confidence level when a profile has not yet been observed in the data base). The number of samples and the percentage of the total for which the absolute value of probabilities falls within the indicated order of magnitude are shown for six reference data base pairs.*

Population Comparison	Probability $\lvert P\rvert \le 10^1$	$\lvert P\rvert > 10^1$ $\lvert P\rvert \le 10^2$
African-American v. Southeast Hispanic	2046	0
	100.0%	0
African-American v. Southwest Hispanic	2046	0
	100.%	0
Caucasian v. Southeast Hispanic	633	1413
	30.9%	69.1%
Caucasian v. Southwest Hispanic	562	1484
	27.5%	72.5%
Caucasian v. African-American	2046	0
	100.0%	0
Southeast Hispanic v. Southwest Hispanic	2046	0
	100.0%	0

between the Caucasian and each of the Hispanic data bases. As described earlier, the rarity of each individual's RFLP profile causes the minimum frequency, dependent solely on the size of the reference data base, to be invoked at each locus. Using a four-locus profile as an example, probabilities estimated using the Caucasian data base as reference exceed those estimated using the southeastern and southwestern Hispanic data bases by factors of 24 and 33.8, respectively (Table 2). Other reference pairs produce estimates within an order of magnitude of one another. The Caucasian v. Hispanic estimates in Table 5 which fall within one order of magnitude arise from samples for which results were available from fewer than four loci. Due to the strict dependence on data base size and rarity of individual profiles the direct count method is insensitive to population frequency differences.

The variation in breadth of the fixed bin plots in Figure 4 a to f likely is related to the degree of genetic differences between the groups being compared. The largest spread is between Caucasians v. African Americans, while the narrowest spreads are between Caucasians v. southeastern Hispanics and southeastern Hispanics v. southwestern Hispanics. The close agreement between probability estimates over multiple loci in Fig. 4f, using southeastern Hispanic or southwestern Hispanic reference data bases would suggest that, although the groups differ both in racial admixture and in their VNTR allele frequency distributions, the additional conservative buffers placed on data for Hispanics suggested by Budowle et al. [*10*] may be unnecessary. This additional level of conservatism has been used by the FBI for the past three years in case work. The larger frequency between the two data bases which occurs in each bin is used to create a combined frequency table for each locus. Application of this approach was advocated to avoid underestimation of frequency arising from latent or undetermined substructure. Although conservative, this approach creates an abstract frequency table that represents no single known population. The comparisons of southeastern and southwestern Hispanics would suggest that the use of this approach in conjunction with an inherently conservative method such as the fixed bin method is unwarranted. Regardless, frequency estimations of an entire profile should be made in one or more relevant data bases which are then compared, rather than as a mosaic of single allele frequencies across several data bases.

Conclusions

Comparisons of the estimated probability of occurrence of 2046 one- to four-locus RFLP profiles by the fixed bin and floating bin methods (with a minimum count of five for the latter), using a floating bin window width of $\pm 2.5\%$, show that in all cases the fixed bin method is more conservative by one to four orders of magnitude. It is not until a floating bin window of 10% ($\pm 5.0\%$) is used that the results begin to correspond. These results are consistent with those of Chakraborty et al. [*9*] indicating that the fixed bin method overestimates each allele frequency, on average, by at least a factor of 2.

Due to the rarity of RFLP profiles, the direct count method is insensitive to the derivation of the data base. Data bases much larger than those used in this study would be required to detect any frequency differences among populations. Frequency estimates by the fixed bin method are at least as conservative as those estimated using the direct count method (by locus, with a 95% upper confidence limit as minimum frequency and multiplied across loci). The fixed bin estimate will tend to become even more conservative than that of the direct count method as the size of the reference data base increases, because the calculated minimum frequency of the latter will decrease. A direct count method (by locus) with no minimum frequency should result in an estimate which is less conservative than the fixed bin method. Directly counting the number of times an entire profile (consisting of results at several loci) has been seen is uninformative because it depends chiefly on the size of the data base and tends to misrepresent the rarity of any profile.

The data presented show that although the distributions of bin frequencies are different among the reference populations [*10*], all four populations are highly polymorphic, i.e., they exhibit high gene diversity for all loci. All VNTR profiles in the various sample populations are rare, and for forensic purposes, profile frequency estimates from different reference populations do not deviate much. Nonetheless, if there is any reasonable doubt about which reference data base should be employed, it is recommended that probability be calculated and reported using any and all relevant data bases.

The fixed bin method has been shown to provide a robust and conservative estimate for frequency of VNTR profiles relative to the floating bin and direct count methods. It is relatively insensitive to misassignment of reference data base and to population substructure. In addition, it compensates for small sample size and previously unobserved alleles.

References

[1] Budowle, B., Giusti, A. M., Waye, J. S., Baechtel, F. S., Fourney, R. M., Adams, D. E., Presley, L., Deadman, H. A., and Monson, K. L., "Fixed-Bin Analysis for Statistical Evaluation of Continuous Distributions of Allelic Data from VNTR Loci," *American Journal of Human Genetics*, Vol. 48, 1991, pp. 841–855.

[2] Balazs, I., Baird, M., Clyne, M., and Meade, E., "Human Population Genetic Studies of Five Hypervariable DNA Loci," *American Journal of Human Genetics*, Vol. 44, 1989, pp. 182–190.

[3] National Research Council, *DNA Technology in Forensic Science*, 1992.

[4] Devlin, B., Risch, N., and Roeder, K., "Forensic Inference from DNA Fingerprints," *Journal of the American Statistical Association*, Vol. 87, 1992, pp. 337–350.

[5] Weir, B., "Independence of VNTR Alleles Defined as Fixed Bins," *Genetics*, Vol. 130, 1992, pp. 873–887.

[6] Budowle, B. and Monson, K. L., "Perspectives on the Fixed Bin Method and the Floor Approach/Ceiling Principle," *Proceedings from The International Symposium on Human Identification 1992*, Promega Corp., Madison, WI, pp. 391–406.

[7] Lander, E., "Lander Reply," *American Journal of Human Genetics*, Vol. 49, 1991, pp. 899–903.

[8] Lewontin, R. C. and Hartl, D. L., "Population Genetics in Forensic DNA Typing," *Science*, Vol. 254, 1991, pp. 1745–1750.

[9] Chakraborty, R., Jin, L., Zhong, Y., Srinivasan, M. R., and Budowle, B., "On Allele Frequency Computation for DNA Typing Data," *International Journal of Legal Medicine* (in press).

[10] Budowle, B., Monson, K. L., Anoe, K., Baechtel, F. S., Bergman, D. L., Buel, E., Campbell, P. A., et al., "A Preliminary Report on Binned General Population Data on Six VNTR Loci in Caucasians, Blacks and Hispanics from the United States," *Crime Laboratory Digest*, Vol. 18, No. 1, 1991, pp. 9–26.

[11] Nakamura, Y., Gillilan, S., O'Connell, P., Leppert, M., Lathrop, G. M., Lalouel, J. M., and White, R., "Isolation and Mapping of a Polymorphic DNA Sequence pYNH24 on Chromosome 2 (D2S44)," *Nucleic Acids Research*, Vol. 15, 1987, p. 10073.

[12] Wong, Z., Wilson, V., Patel, I., Povey, S., and Jeffreys, A. J., "Characterization of a Panel of Highly Variable Minisatellites Cloned from Human DNA," *Annals of Human Genetics*, Vol. 51, 1987, pp. 269–288.

[13] Milner, E. C., Latshaw, C. L., Willems van Dijk, K., Charmley, P., Concannon, P., and Schroeder, H. W., "Isolation and Mapping of a Polymorphic DNA Sequence (pH30) from Human Chromosome 4 (D4S139)," *Nucleic Acids Research*, Vol. 17, 1989, p. 4002.

[14] Budowle, B. and Baechtel, F. S., "Modifications to Improve the Effectiveness of Restriction Fragment Length Polymorphism Typing," *Applied and Theoretical Electrophoresis*, Vol. 1, 1990, pp. 181–187.

[15] Monson, K. L. and Budowle, B., "A System for Semi-Automated Analysis of DNA Autoradiograms," *Proceedings of An International Symposium on the Forensic Aspects of DNA Analysis*, U.S. Government Printing Office, Washington, D.C., 1990, pp. 127–132.

[16] Chakraborty, R. and Kidd, K., "The Utility of DNA Typing in Forensic Work," *Science*, Vol. 254, 1991, pp. 1735–1739.

[17] Chakraborty, R., de Andrade, M., and Daiger, S. P., "Apparent Heterozygote Deficiencies Observed in DNA Typing Data and Their Implications in Forensic Applications," *American Journal of Human Genetics*, Vol. 56, 1992, pp. 45–57.

[18] Chakraborty, R. and Jin, L., "Heterozygote Deficiency, Population Substructure and their Implications in DNA Fingerprinting," *Human Genetics*, Vol. 88, 1992, pp. 267–272.

[19] Risch, N. J. and Devlin, B., "On the Probability of Matching DNA Fingerprints," *Science*, Vol. 255, 1992, pp. 717–720.

[20] Brenner, C. and Morris, J., "Paternity Index Calculations in Single Locus Hypervariable DNA Probes: Validation and Other Studies," *Proceedings of The International Symposium on Human Identification*, Promega Corp., Madison, WI, 1989, pp. 21–53.

[21] Budowle, B., Baechtel, F. S., Giusti, A. M., and Monson, K. L., "Data for Forensic Matching Criteria for VNTR Profiles," *Proceedings of The International Symposium on Human Identification*, Promega Corp., Madison, WI, 1989, pp. 103–115.

Address requests for reprints or additional information to
Keith L. Monson, Ph.D.
Forensic Science Research and Training Center
FBI Academy
Quantico, VA 22135