TECHNICAL NOTE

George Herrin, Jr.,[1] *Ph.D.*

A Comparison of Models Used for Calculation of RFLP Pattern Frequencies

REFERENCE: Herrin, G., **"A Comparison of Models Used for Calculation of RFLP Pattern Frequencies,"** *Journal of Forensic Sciences*, JFSCA, Vol. 37, No. 6, November 1992, pp. 1640–1651.

ABSTRACT: In recent years the application of DNA typing information to criminal investigations has gained widespread acceptance. The primary method currently in use relies on length variation of DNA restriction fragments between individuals. These variations are identified using variable number tandem repeat (VNTR) DNA probes. As this technology becomes more widely used, it is crucial that scientifically valid methods of interpreting the significance of a DNA typing result be adopted. The method chosen should not only give a reliable approximation of the statistical likelihood of a particular RFLP pattern occurring, but should also be easy to present and for the court to understand. In this manuscript five methods of calculating a frequency of occurrence of a RFLP pattern will be presented (fixed bin genotype, floating bin phenotype, floating bin genotype, National Research Council (NRC) method using fixed bins and the NRC method using floating bins). The calculations discussed here demonstrated that the fixed bin genotype method produces a frequency very similar to that obtained from floating bin phenotypes. In addition, regardless of the method chosen or the database size, the frequency of any particular banding pattern in the population over several loci was found to be very rare.

KEYWORDS: criminalistics, RFLP statistics, DNA typing

The use of RFLP typing data in criminal investigations has introduced new challenges in the interpretation of the significance of matching patterns. Biological markers previously used in forensic settings all possess discrete alleles or phenotypes which could be readily characterized. This is not the case with VNTR data. With the loci currently used, the length of the repeat unit at the locus is small compared to the resolving power of the analytical system used to distinguish the DNA fragments. This results in an essentially continuous size distribution of the DNA fragments possible from each locus. Because it is not possible to obtain an absolute length of the fragments without sequencing, grouping of fragments into 'bins' is performed by most laboratories performing analysis of forensic evidence in the United States.

Two approaches to binning the data have been developed. The first approach uses fixed bins based on a series of DNA fragments of defined length which are included with

Received for publication 15 July 1991; revised manuscript received 18 June 1992; accepted for publication 29 June 1992.

[1]Georgia Bureau of Investigation-Division of Forensic Sciences, Decatur, GA.

each analysis [*1*]. The second approach uses floating bins which are based on the calculated size of the fragment in question and a match criteria empirically determined within the laboratory [*2,3*]. These floating bins are centered on the fragment sizes calculated for a particular sample.

Once a match between two or more samples has been declared, some measure of the likelihood of observing a particular pattern (that is, the frequency of the pattern in a reference population) is calculated. Again there are two approaches to determine how the frequency of any particular pattern can be calculated from RFLP analysis:

1) genotypes—each DNA fragment (band) from a specific locus is treated as an individual allele.
2) phenotypes—the banding pattern from a specific locus is treated in its entirety.

The genotype approaches are based on the Hardy-Weiberg model of population genetics [*4,5*] which states the following:

$$p^2 + 2pq + q^2 = 1 \tag{1}$$

p,q are allele frequencies at the locus in question.

Several of the assumptions used by the Hardy-Weinberg model have been contested with regard to RFLP data (for example, random mating, population substructure) [*6*]. Current methods based on the Hardy-Weinberg model disregard the homozygote terms (p^2/q^2) in the calculations [*1–3*]. These terms are not used because it is very difficult to identify an individual as truly homozygous at a VNTR locus with available analytical techniques [*7*]. The genotype for an individual who exhibits a one banded pattern for a locus is calculated using the heterozygous term ($2pq$) from the Hardy-Weinberg model, where $q = 1$ and $p =$ allele frequency of the observed fragment (that is, $2p$). For the purposes of this discussion the term genotype will be used to refer to a banding pattern frequency which has been calculated using the individual band (allele) frequencies.

The phenotype method does not rely on the Hardy-Weinberg model, but is based upon observed data within the population database. Use of this method requires counting the number of occurrences of a particular RFLP banding pattern at each locus within the appropriate database and dividing the number of occurrences by the number of samples in the database. If the test pattern contains a single band, then only those samples from the database with a corresponding single banded pattern would be counted. This is directly analogous to the method used with blood groups from which only a phenotype can be measured. For the purposes of this discussion the term phenotype will be used to refer to a banding pattern frequency which has been calculated using the observed number of occurrences of a particular pattern frequency within the database. This pattern could consist of either one or two bands.

This paper investigates the relationship between the frequencies calculated by each of these methods as well as how the number of individuals collected in the reference database may affect that frequency. It also demonstrates that regardless of the method chosen, the frequency of occurrence of a pattern is rare.

Methods

The frequencies were calculated using each of five methods: (1) floating bin phenotype, (2) fixed bin genotype, (3) floating bin genotype, (4) NRC floating bin, and (5) NRC fixed bin. The RFLP patterns for these samples were determined using probes for five loci [MS1 (D1S7), YNH24 (D2S44), TBQ7 (D10S28), pH30 (D4S139), and V1 (D17S79)] and the restriction enzyme *HaeIII* (8). The first 100 samples from the Caucasian and Black databases for which data from all 5 loci was available were analysed with each of

the five methods above using the corresponding database. The size (N) of the Caucasian database for each of the five loci is approximately 280 individuals, for the Black database approximately 475.

Floating Bin Phenotype—This method has been used in legal proceedings (Georgia vs Caldwell, Dr. W. Anderson personal communication). This method calculates the frequency of a pattern using Eq 2 with the following variable definitions.

$$\frac{1}{F} = \prod_{i=1}^{n} \left(\frac{(\sum_{j=1}^{N} I_j) + 1}{N + 1} \right)_i \tag{2}$$

where $1/F$ is the estimated frequency of the phenotype in the database from all loci examined

n is the number of loci used in the analysis

N is the number of individuals in the population database for each locus

1 is added to N to include the pattern from the test sample in the database

I_j is a pattern from the database which satisfies Condition 1 below

1 is added to the total number of patterns from the database which satisfy Condition 1 in order to include the pattern from the test sample.

Cond. 1: $(B_1 - w_1) \leq B_{1_j} \leq (B_1 + w_1)$ and $(B_2 - w_2) \leq B_{2_j} \leq (B_2 + w_2)$.

where B_1, B_2 are the calculated fragment sizes for the pattern for which a frequency is being determined.

B_{1_j}, B_{2_j} are the calculated fragment sizes of samples from the database being tested in Condition 1.

w_1, w_2 are parameters that define the width of the frequency bin and are a percentage of B_1, B_2 respectively. The percentage used to calculate w_1, w_2 is related to the match criteria of the laboratory.

The value of w_1, w_2 was set at 5% of the calculated fragment size for fragments <10 kb and 10% for fragments >10 kb based on the match criteria developed within the laboratory. The match criteria used by GBI-DOFS laboratory declares two fragment sizes must lie within 4% of each other for fragments <10 kb and within 8% for fragments >10 kb [*9*]. This results in a bin width for frequency calculations which is 2.5 times that used for declaring a match between two samples. This expansion of the bin used for frequency calculations relative to the window used for matching ensures that all fragments in the database which would be declared a match with the questioned fragment will be counted.

Because the loci utilized for forensic analysis produce a large number of alleles, there are a very large number of possible phenotypes. With the databases currently available, the possibility exists that the phenotype observed in a test sample may not have been previously observed in the database. To avoid this possibility, the banding pattern from the test sample is temporarily included in the database. This results in a minimum pattern frequency of $1/(N + 1)$ at an individual locus for patterns not previously observed within the database.

Fixed Bin Genotype—This is the method described in Budowle et al. [*1*] and is currently used by most crime laboratories performing RFLP analysis. The bins were regrouped in order to provide a minimum of 5 fragments per bin as described in [*1*]. For purposes of simplification of this comparison, one parameter discussed by those authors has been omitted. Fragments within 2.5% of a bin boundary were not placed into the bin containing the highest number of events in the comparisons presented here. This method calculates pattern frequencies using Eq 3.

$$\frac{1}{F} = \prod_{i=1}^{n} \left(2 \left(\frac{\sum_{j=1}^{N} B_j}{2N} \right) \left(\frac{\sum_{j=1}^{N} B_j}{2N} \right) \right)_i \tag{3}$$

here $1/F$ is the estimated frequency of the genotype in the database from all loci examined

n is the number of loci used in the analysis

N is the number of individuals in the population database for each locus

B_j is a fragment from the database which satisfies Condition 2 or Condition 3

$$\text{Cond. 2: } (\text{Bin } 1)_L \leq B_j \leq (\text{Bin } 1)_U$$

$$\text{Cond. 3: } (\text{Bin } 2)_L \leq B_j \leq (\text{Bin } 2)_U$$

where $(\text{Bin } 1)_L$, $(\text{Bin } 2)_L$ are the lower boundaries of the fixed bins which contain fragments of the sizes calculated for the pattern for which a frequency is being determined. It is not necessary that Bin 1 be different from Bin 2.

$(\text{Bin } 1)_U$, $(\text{Bin } 2)_U$ are the upper boundaries of the bins.

B_j are the calculated fragment sizes of the samples in the database.

Floating Bin Genotype—This method is described in [*2,3*]. It is similar to the fixed bin genotype method with the exception of using floating bins as described in Condition 1. The frequency of pattern is calculated by this method as shown in Eq 3 with the following alteration in the term definitions.

B_j is a fragment from the database which satisfies Condition 4 or Condition 5

$$\text{Cond. 4: } (B_1 - w_1) \leq B_j \leq (B_1 + w_1)$$

$$\text{Cond. 5: } (B_2 - w_2) \leq B_j \leq (B_2 + w_2).$$

where B_1, B_2 are the calculated fragment sizes of the pattern for which a frequency is being determined.

B_j is the calculated fragment size from sample in the database being tested in Conditions 4 and 5.

w_1, w_2 are parameters that define the width of the frequency bin and are a percentage of B_1, B_2 respectively. The percentage used to calculate w_2, w_2 is related to the match criteria of the laboratory.

NRC Method—This method is described in [*10*]. This method is essentially a genotype calculation (Eq 3) with restrictions placed on the lower limit of the allele frequencies. There were two ceiling limits suggested within the NRC report. An interim ceiling of 0.1 for allele frequencies less than the ceiling was suggested until additional ethnically defined databases could be collected at which time the ceiling may drop to 0.05. For allele frequencies which are greater than the imposed ceiling, a 95% upper confidence limit was calculated for that allele frequency as shown in Eq 4. For the calculations which were performed in this study, both methods of binning were examined with each of the ceiling limits proposed by the NRC.

$$p = p + 1.96 \sqrt{\frac{p(1 - p)}{N}} \tag{4}$$

where p is the observed allele frequency

N is the number of chromosomes studied.

Results

The average frequency of the first 100 patterns from the databases using the five loci is shown in Table 1 for each of the methods studied. Subsets of the total databases (approximately half) were also examined in order to determine the effects of smaller

TABLE 1—*Comparison of the described methods of calculating the frequency of occurrence of a RFLP banding pattern. The average frequency of the first 100 samples for which data from the 5 loci studied is shown here using databases of size* N. *For the database subsets the first* N *samples from the database were used. For the NRC methods the average frequency of the first 100 samples from the indicated databases were calculated. For the database subsets using the NRC methods the first* N *samples from the Black and Caucasian (250 and 150 respectively) databases were used.*

Calculation Method	Black (N = 250)	Black (N = 475)	Caucasian (N = 150)	Caucasian (N = 280)
Floating bin phenotype	3.0×10^{10}	3.0×10^{11}	1.4×10^{9}	6.8×10^{9}
Fixed bin genotype	2.9×10^{11}	9.1×10^{11}	1.4×10^{10}	2.6×10^{10}
Floating bin genotype	7.7×10^{14}	8.2×10^{14}	6.1×10^{12}	8.2×10^{12}
Fixed bin NRC (0.1 ceiling)	4.2×10^{7}	5.5×10^{7}	1.8×10^{7}	2.6×10^{7}
Floating bin NRC (0.1 ceiling)	1.3×10^{8}	1.4×10^{8}	6.9×10^{7}	8.3×10^{7}
Fixed bin NRC (0.05 ceiling)	6.6×10^{8}	2.0×10^{9}	1.3×10^{8}	2.8×10^{8}
Floating bin NRC (0.05 ceiling)	2.0×10^{10}	2.7×10^{10}	3.6×10^{9}	8.1×10^{9}

data sets on the observed frequency. The results are presented as 1 in x, where x is the calculated number of occurrences of the banding pattern at this combination of five VNTR loci. In each method, as the size of the database (N) used to calculate the frequency of the pattern was increased, the probability of finding another individual with the same DNA banding pattern remained essentially constant or decreased by one order of magnitude or less.

Another comparison of the methods is shown in Figs. 1 through 4. In this set of charts

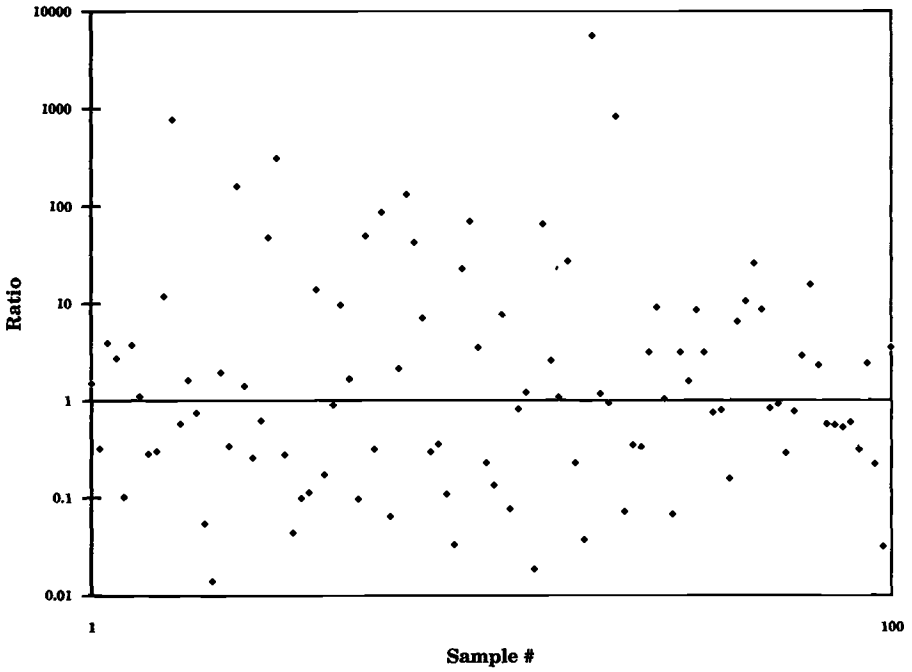

FIG. 1—*Ratio of the VNTR pattern frequency obtained using the floating bin phenotype method to that obtained using the fixed bin genotype method. Each point represents the data from one individual within the Black database for which data from five loci were available. Approximately 65% of the values for the two methods are within one order of magnitude of each other.*

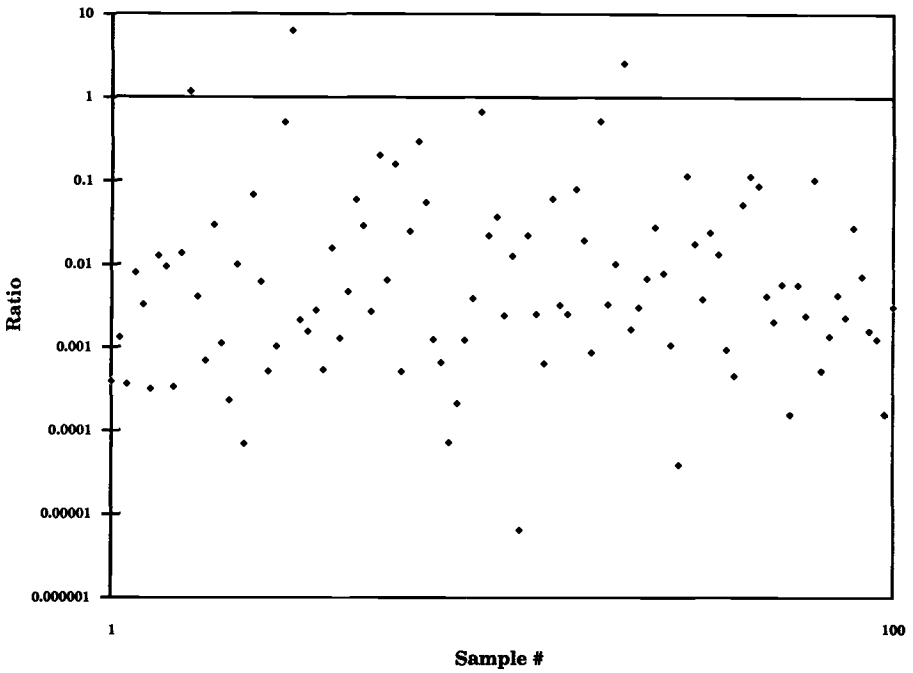

FIG. 2—*Ratio of the VNTR pattern frequency obtained using the floating bin phenotype method to that obtained using the floating bin genotype method. Each point represents the data from one individual within the Black database for which data from five loci were available. This chart demonstrates that the floating bin genotype method consistently yields the least conservative result.*

the methods are compared to determine which method yields a consistently higher frequency of occurrence. Each of these charts was prepared from the same 100 Black individuals used to develop the averages shown in Table 1. Similar results were obtained from the Caucasian database (not shown). Those points that lie below the equivalence line (Ratio = 1) in Figs. 1 and 2 represent samples for which the floating bin phenotype results in a greater frequency of occurrence (that is, more conservative estimate) than the fixed bin genotype or the floating bin genotype. Figures 3 and 4 illustrate that either of the binning methods using the NRC calculations produces more conservative frequency estimates than those produced from either floating bin phenotypes or fixed bin genotypes. From the data presented, the following ranking of methods may be constructed in order of most to least conservative. When the ceiling for the NRC method was set at 0.05, the results between each of the methods become more closely equivalent (data not shown).

NRC ceiling > Fixed Genotype ≈ Floating Phenotype > Floating Genotype

Those data points in Fig. 1 which represent large departures from equivalence (≥1000 fold difference) between the two methods usually represent samples in which the genotype method is more conservative due to the $2p$ calculation used for one banded patterns, rather than the homozygosity terms from the Hardy-Weinberg equation. The data from the Caucasian database was very similar (not shown).

The data presented in Fig. 5 demonstrate the binning method chosen has only a small effect on the overall pattern frequency calculated using the NRC method with a 0.1 ceiling. Only 1 of the 100 samples examined showed a difference of more than one order of magnitude. When the ceiling was changed to 0.05 the ratios decreased by approximately

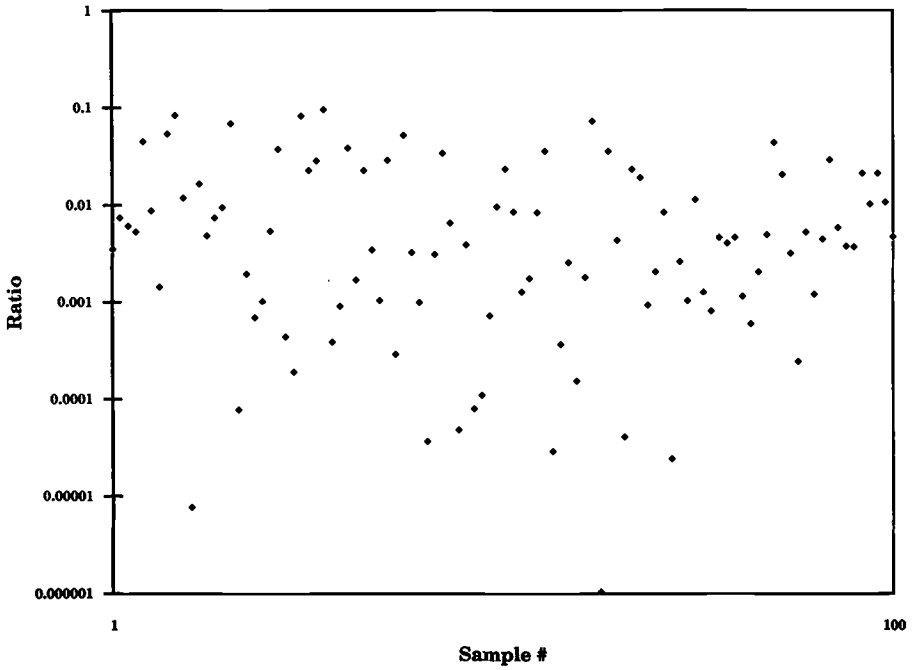

FIG. 3—*Ratio of the VNTR pattern frequency obtained using the NRC floating bin method with a 0.1 ceiling to that obtained using the floating bin phenotype method. Each point represents the data from one individual within the Black database for which data from five loci were available. This chart demonstrates that the floating bin phenotype method consistently yields a less conservative result than the NRC method with this ceiling.*

one order of magnitude (data not shown). The floating bin method resulted in overall pattern frequencies 2 orders of magnitude higher than the fixed bin method in 74% of the samples studied.

Because differences between the methods may have been masked by multiplying across several loci, graphs corresponding to Figs. 1 to 4 were prepared using individual loci. D1S7 pattern frequencies using these five methods are compared in Figs. 6 to 9. The ratio between the methods at the individual locus is very similar to that obtained using all five loci together. Graphs prepared from data at the other four loci were essentially the same (not shown).

Average allele frequencies for each loci using the interim 0.1 ceiling of the NRC method were calculated and are shown in Table 2. It is interesting to note that with the exception of D17S79 no differences were observed in the average frequency between the two racial groups tested.

Conclusion

The data presented compares five methods of calculating the frequency of occurrence of a DNA banding pattern obtained from VNTR loci. Each of these methods has ad-

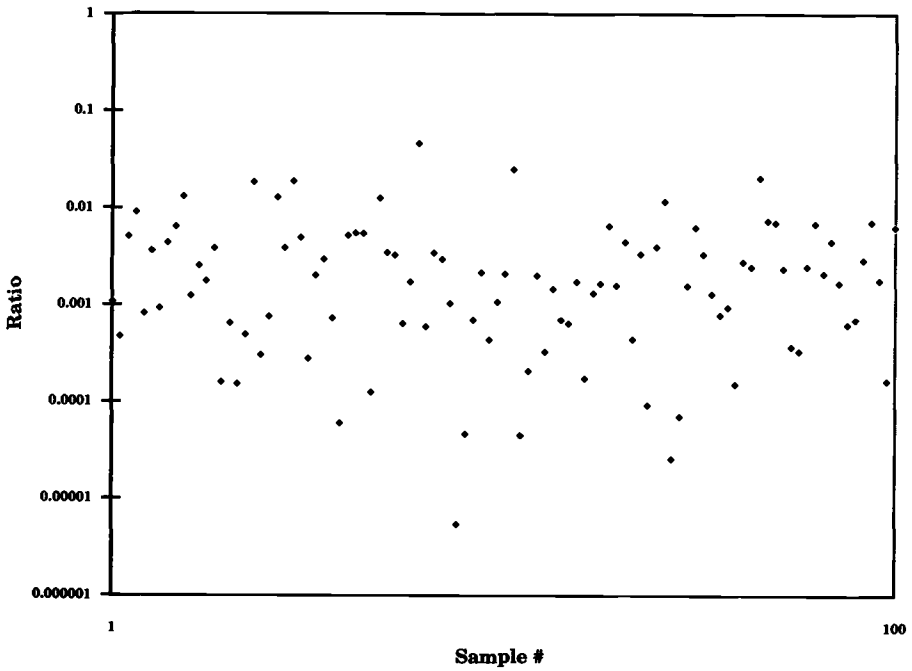

FIG. 4—*Ratio of the VNTR pattern frequency obtained using the NRC fixed bin method with a 0.1 ceiling to that obtained using the fixed bin genotype method. Each point represents the data from one individual within the Black database for which data from five loci were available. This chart demonstrates that the fixed bin genotype method consistently yields a less conservative result than the NRC method with this ceiling.*

vantages. The fixed bin method has the advantages of easily classifying the data obtained, allowing quick comparisons between laboratories, and is more amenable to evaluation by statistical methods for testing adherence to certain assumptions. The floating bin methods have the advantage of using fragments within the database which are more likely to be the same biologically as the fragments from the test sample. The NRC method has the advantage of compensating for any potential population subgrouping.

The decision to use a phenotype method over a genotype method hinges in large part on the desire of the laboratory personnel to deal with Hardy-Weinberg issues in court. Some of the Hardy-Weinberg assumptions are avoided by the phenotype method. Both the phenotype and genotype methods assume independence between loci and a lack of significant population subgrouping. These assumptions are consistent with the lack of data demonstrating linkage between these VNTR loci for heterozygous patterns [*11*] or the presence of population subgrouping within the Caucasian and Black populations which would significantly change the final calculated frequency.

The method proposed by the NRC in their report assumes population subgrouping may be present within the databases currently used. To counteract the effects of such subgrouping, a ceiling limit is imposed on the allele frequencies used to calculate the pattern frequency. It is interesting to note that the report does imply that with the use

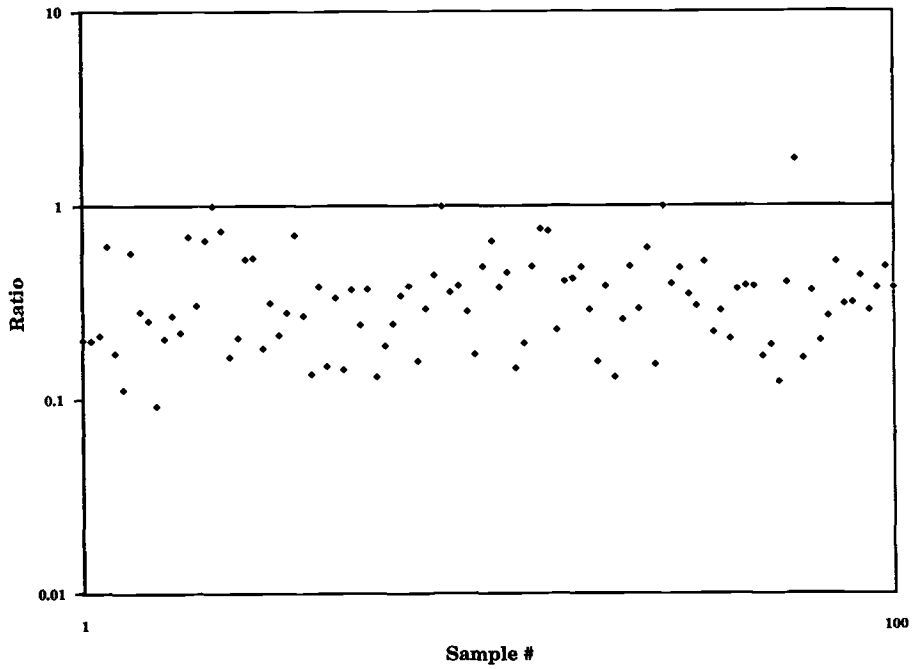

FIG. 5—*Ratio of the VNTR pattern frequency obtained using the floating bin NRC method to that obtained using the fixed bin NRC method with a 0.1 ceiling in each case. Each point represents the data from one individual within the Black database for which data from five loci were available. This chart demonstrates that the binning method will result in a change in pattern frequency of less than one order of magnitude.*

of the ceiling principle, the loci may be assumed to be in linkage equilibrium and Hardy-Weinberg equilibrium. One offshoot of the ceiling principle is the possibility of calculating a very rough approximation of any banding pattern without using a database at all by using the average allele frequencies observed for each locus as shown in Table 2.

Currently used methods for calculating a pattern frequency produce results which are similar (within two orders of magnitude in most cases) to those produced by the NRC method with a 0.05 ceiling. This suggests that gross errors are not being made in the approximations of pattern frequencies.

For database sizes sufficiently large the floating bin phenotype method should give approximately the same result as that yielded by the floating bin genotype method which is based on the Hardy-Weinberg model of population genetics. The results (Fig. 2) indicate this is not yet the case, most probably because the database sizes would have to be much larger in order to accurately represent those individuals with rare phenotypes. These individuals are currently assigned a minimum frequency of $1/(N + 1)$.

In summary, each of the five methods detailed here, floating bin phenotypes, fixed and floating bin genotypes, fixed and floating bin genotypes with ceilings (NRC) produce frequencies which are very similar in practical terms. The choice of which method to use within a forensic laboratory should be based on several factors including scientific validity and ease of presentation in court.

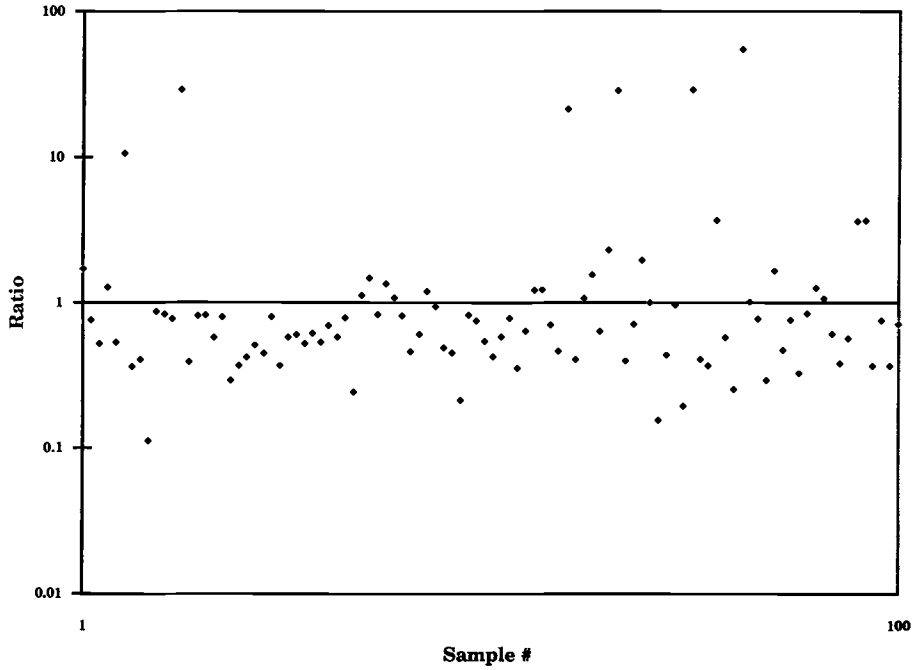

FIG. 6—*Ratio of the VNTR pattern frequency obtained using the floating bin phenotype method to that obtained using the fixed bin genotype method. Each point represents the data from one individual within the Black database for the locus D1S7.*

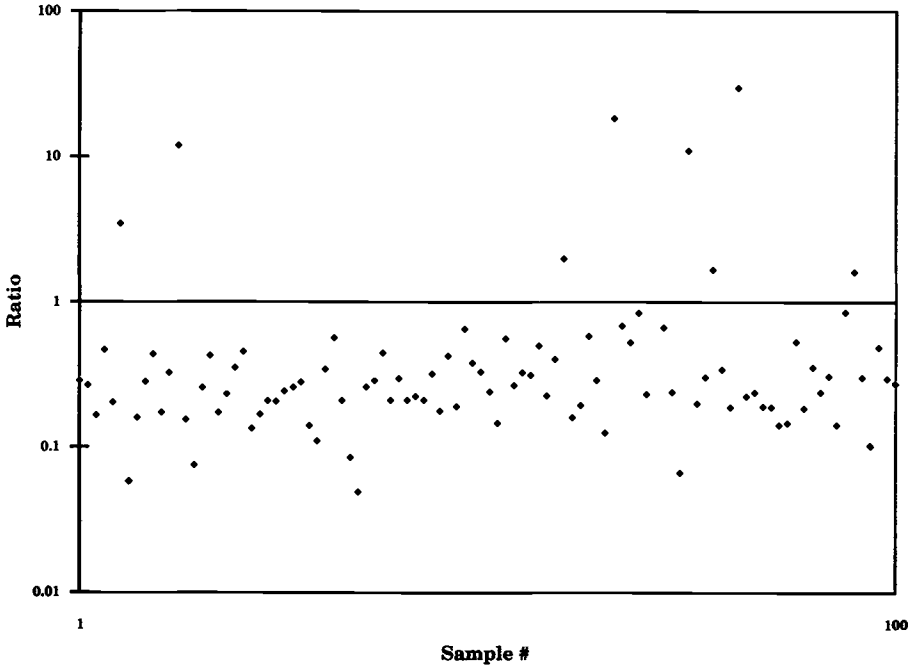

FIG. 7—*Ratio of the VNTR pattern frequency obtained using the floating bin phenotype method to that obtained using the floating bin genotype method. Each point represents the data from one individual within the Black database for the locus D1S7.*

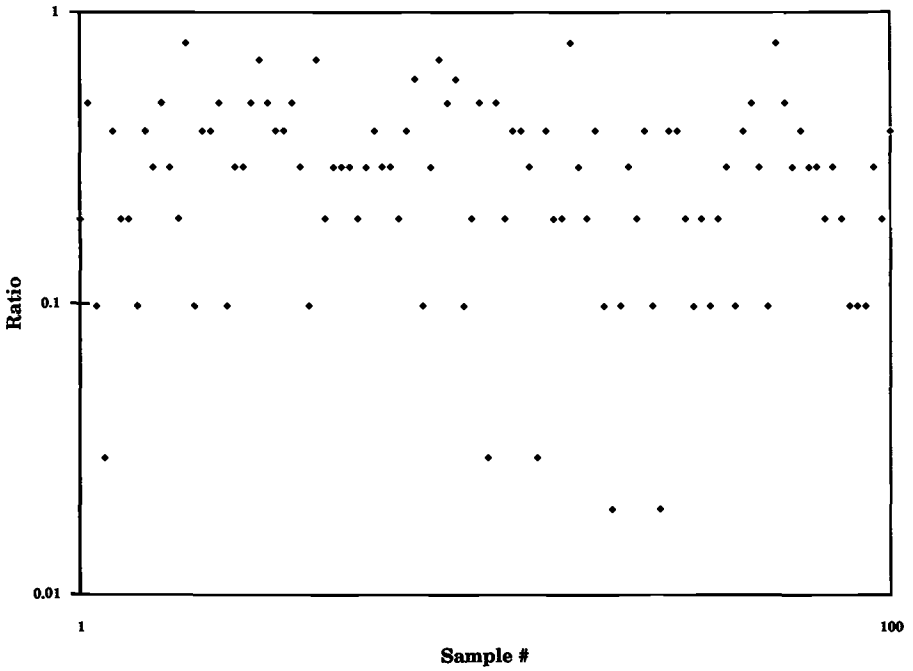

FIG. 8—*Ratio of the VNTR pattern frequency obtained using the floating bin NRC method with a 0.1 ceiling to that obtained using the floating bin phenotype method. Each point represents the data from one individual within the Black database for the locus D1S7.*

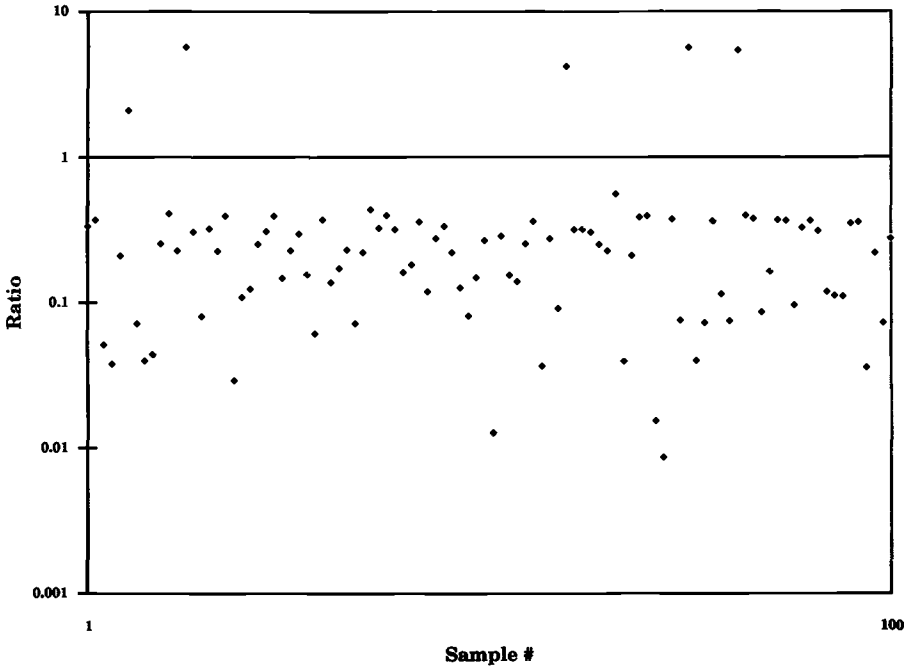

FIG. 9—*Ratio of the VNTR pattern frequency obtained using the fixed bin NRC method with a 0.1 ceiling to that obtained using the fixed bin genotype method. Each point represents the data from one individual within the Black database for the locus D1S7.*

TABLE 2—*Comparison of average allele frequency from each locus using each of the binning methods discussed. The allele frequencies were calculated using the NRC method with a ceiling of 0.1.*

Binning method	Race	D1S7	D2S44	D10S28	D4S139	D17S79
Fixed	Black	0.10	0.11	0.10	0.13	0.20
Floating	Black	0.10	0.10	0.10	0.10	0.14
Fixed	Caucasian	0.10	0.11	0.10	0.13	0.24
Floating	Caucasian	0.10	0.11	0.10	0.10	0.17

References

[*1*] Budowle, B., Baechtel, F. S., Giusti, A. M., and Monson, K. L., "Applying Highly Polymorphic Variable Number of Tandem Repeats Loci Genetic Markers to Identity Testing," *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*] Herrin, G., Jr., Forman, L., and Garner, D. D., "The Use of Jeffrey's Multilocus and Single Locus DNA Probes in Forensic Analysis," In *DNA and Other Polymorphisms in Forensic Science*, Year Book Medical, Chicago, IL, 1990.

[*4*] Hardy, G. H., "Mendelian Proportions in a Mixed Population," *Science*, Vol. 28, 1908, pp. 41–50.

[*5*] Weinberg, W., "On the Demonstration of Heredity in Man," In *Papers on Human Genetics*, Prentice-Hall, Englewood Cliffs, NJ, 1908.

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

[*7*] Devlin, B., Risch, N., and Roeder, K., "No Excess of Homozygosity at Loci Used for DNA Fingerprinting," *Science*, Vol. 249, 1990, pp. 1416–1420.

[*8*] Budowle, B. and Baechtel, F. S., "Modifications to Improve the Effectiveness of Restriction Fragment Length Polymorphism Typing," *App and Theor. Electro.* Vol. 1, 1990, pp. 181–187.

[*9*] Staples, T., Goff, C. K., Wegel, Jr., J. G., and Herrin, Jr., G., "RFLP Match Criteria Determination from Data Analysed on a Bioimage System," *Proceedings from the Second International Symposium on Human Identification*. 1991, p. 316.

[*10*] McKusick, V. A., Ferrara, P. B., Kazazian, H. H., King, M. C., Lander, E. S., Lee, H. C., Lempert, R. O., Macklin, R., Marr, T. G., Reilly, P. R., Sensabaugh, Jr., G. F., and Weinstein, J. B., in *DNA Technology in Forensic Science*, National Academy Press, 1992.

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