John M. Hartmann,[1] *B.S., M.B.A.; Bruce T. Houlihan,*[1] *B.S.; Lisa D. Thompson,*[1] *B.S.; Christine Chan,*[1] *B.S.; Russell A. Baldwin,*[1] *B.S.; and Edward L. Buse,*[1] *B.S., M.A.*

The Effect of Sampling Error and Measurement Error and Its Correlation on the Estimation of Multi-Locus Fixed-Bin VNTR RFLP Genotype Probabilities

REFERENCE: Hartmann JM, Houlihan BT, Thompson LD, Chan C, Baldwin RA, Buse EL. The effect of sampling error and measurement error and its correlation on the estimation of multi-locus fixed-bin VNTR RFLP genotype probabilities. J Forensic Sci 1997;42(2):241–245.

ABSTRACT: Bootstrapping was used to examine the effect of sampling error and measurement error and its correlation on fixed-bin genotype probabilities. Bootstrap confidence intervals (CIs) were made relative to the point estimate using the log of the inverse of the probabilities. From databases of 200–250 genotypes, sampling error alone yielded median relative 95% CIs of from one order of magnitude out of five for one locus to one out of ten for four loci. Measurement error of the test genotype fragments increased the latter to about one order of magnitude out of eight. Database measurement error and its correlation had only a slight effect on multi-locus probability uncertainty. Together, these uncertainties are several orders of magnitude greater than error due to population substructuring of a race by its major component ethnic groups.

KEYWORDS: forensic science, DNA typing, sampling error, measurement error, restriction fragment length polymorphism, variable number of tandem repeats, genotype, population genetics

The application of DNA profiling to forensic casework has been opposed by some scientists, who have expressed concern with the impact of population substructuring on the accuracy of genotype probability estimates (1). However, there have been other concerns. Pertinent to this study, Evett (2) and Shapiro (3) noted the correlation of fragment size measurement errors. Evett (4) described the effect, along with several others, as trivial, but did not provide quantitative support for this opinion. Berry et al. (5) and Devlin et al. (6) incorporated measurement error correlation into their treatments of VNTR RFLP patterns, but did not comment on the magnitude of its effect on their likelihood ratios. Chakraborty (7), Evett and Gill (8), and the second National Research Council Committee on DNA Forensic Science (9, NRC II) have discussed the importance of adequate sample size for fixed-bin frequency and likelihood ratio estimates. The National Research Council Committee on DNA Technology in Forensic Science (10, NRC I) recommended using binomial confidence limits for bin frequencies in order to compensate for possible subpopulation derived error (and presumably for sampling error as well). Chakraborty et al. (11), following Goodman (12,13), derived equations for multi-locus confidence intervals and applied them to the actual DNA profiles. The NRC II panel offered a similar method (9). In this paper, we report the implementation of bootstrap confidence intervals to model sampling variance by itself, and in combination with measurement error and error correlation.

Methods

Database Sampling and Analysis

Approximately 275 EDTA-preserved blood samples were obtained from anonymous Southern California Hispanic Red Cross donors within about a one-week period and analyzed according to the method of Budowle and Baechtel (14,15,16), with minor modifications. Buffer washes using Centricon 100s (Amicon: Beverly, MA) were substituted for ethanol precipitation prior to restriction by Hae III, and Ethidium bromide was omitted from the analytical gel and tank buffers. The alkaline-blot membranes were probed with the following plasmid insert probes: MS-1 (Cellmark: Germantown, MD), YNH24 and TBQ7 (Promega: Madison, WI), and pH30 (Genelex: Seattle, WA). The autoradiographs were sized independently by two analysts using the local logarithmic algorithm and equipment described by Monson and Budowle (17). The duplicate fragment sizes were compared and accepted if they were within ±2.5% of the mean size, which was the fragment size used in the analyses described below. All single-band patterns were considered to be homozygotes and the observed band size counted twice. The number of fragments obtained for each locus are listed in Table 1.

Sampling Error—Database sampling error was modeled using independent resampling with replacement (bootstrapping) of individual fragments with uniform probability (18). One thousand

TABLE 1—*Sample sizes.*

Locus	No. Fragments
D1S7	514
D2S44	496
D4S139	486
D10S28	512

[1] Senior forensic scientists and forensic scientists III, respectively, Forensic Science Services, Orange County Sheriff-Coroner Department, Santa Ana, CA.

Received 15 June 1993; and in revised form 15 Nov. 1993, 21 Aug. 1995, and 8 Aug. 1996; accepted 9 Aug. 1996.

bootstrap resamples of the same size as the original sample were drawn for each locus.

Test Genotypes—One thousand four-locus test genotypes were generated by independent uniform resampling with replacement from the original databases.

Measurement Error and its Correlation—Measurement imprecision of both database and test genotype fragments was modeled by adding to each resampled database or test genotype fragment, a normally distributed random error with a fixed 0.8% coefficient of variation. Below, we use the word perturb to describe these slight changes in fragment size. Correlation of measurement errors was modeled using a bivariate correlation of errors (i.e., by locus). In unpublished work, we found our bivariate correlation coefficient to be about 0.6, depending somewhat upon band size.

Single-Band Patterns

VNTR RFLP databases contain substantial proportions of single-band patterns due, in part, to the occurrence of homozygotes. In addition, coalescence of fragments due to limited analytical resolution (19) occurs with all loci. Also, there are alleles, referred to as covert by Chakraborty (20), which are too small to be readily detected. No attempt was made to model covert alleles because estimates for their frequency are not available for all of the databases here. Generation of genotypes by independent resampling will yield fewer exact homozygotes than present in the original samples. Coalescence of bands from autoradiographs of ^{32}P-labeled DNA is due not only to electrophoretic limitations but also autoradiograph bloom, which is a function of many factors, for example, probe activity, exposure length, and film speed. We have found the equation below to describe approximately the edge-to-edge width for our autoradiograph bands. Two resampled bands were considered to be coalesced and their mean substituted for both if the difference between their sizes, D, in base pairs (bp) was less than 0.75 of the width, W, predicted by the following equation, where $\bar{x}$ is the mean fragment size (R^2 .948):

$$W = -46.423 + 0.089\bar{x} - 1.057 \times 10^{-5}\bar{x}^2 + 1.185 \times 10^{-9}\bar{x}^3$$

Bootstrap Confidence Intervals

To model sampling error, 1000 databases equal in size to the originals for each locus were prepared by independent sampling with replacement of individual fragments, which were then randomly paired. To model independent measurement error, each resampled fragment was modified by the addition of an independent, normally distributed random error (perturbed) as described above for the test genotypes. Correlated measurement error was modeled by adding bivariate correlated, normally distributed random errors to each pair of resampled fragments. Each of the three sets of 1000 databases was binned and rebinned as described by Budowle and Monson (21).

Using $2pq$ for heterozygous and $2p(1 - p)$ for homozygous patterns, 1000 one-, two-, three-, and four-locus genotype probabilities were calculated for each test genotype. Each set of 1000 genotype probabilities was then sorted by increasing magnitude and their $\alpha/2$ and $1 - \alpha/2$ quantiles obtained. Where $N = 1000$, for a 90% interval the 50th and 950th ranked genotype probabilities correspond to the .05 and .95 quantiles respectively. In this study, because $(\alpha/2) \cdot 1000$ was an integer in every case, no adjustment was required. This process was performed with the same test genotypes using the unperturbed, and both the independent and correlated perturbed rebin frequencies.

Relative Confidence Intervals—In order to enable straightforward comparisons of confidence intervals and their distributions, the log relative CI ratio was used:

$$R = \frac{\log_{10}(1/P_{1-\alpha/2}) - \log_{10}(1/P_{\alpha/2})}{\log_{10}(1/P_0)}$$

where P_0 is the point estimate. A ratio of 0.1 means the interval was one-tenth the magnitude of the point estimate.

Results

Figure 1 contains plots of the distributions of 1000 four-locus relative 95% confidence intervals for the various perturbation combinations of database and test genotype. In the absence of measurement error, the median confidence interval was about one-tenth the magnitude of the point estimate, and the greatest interval about one-sixth the magnitude of the genotype probability estimate. Perturbation of the database samples resulted in a very slight reduction in the range of observed intervals. This was due to the smoothing effect of measurement error, which reduces somewhat the range and variance of the rebin frequencies. In contrast, perturbation of the test genotype fragment sizes increased the median confidence interval to about one order of magnitude out of eight, and the range of intervals increased to include rare instances exceeding one order of magnitude out of five. Measurement error on occasion results in assignment of an observed fragment to a neighboring bin with a different frequency. The net effect is an increase in the variance of genotype probabilities, and hence a greater relative

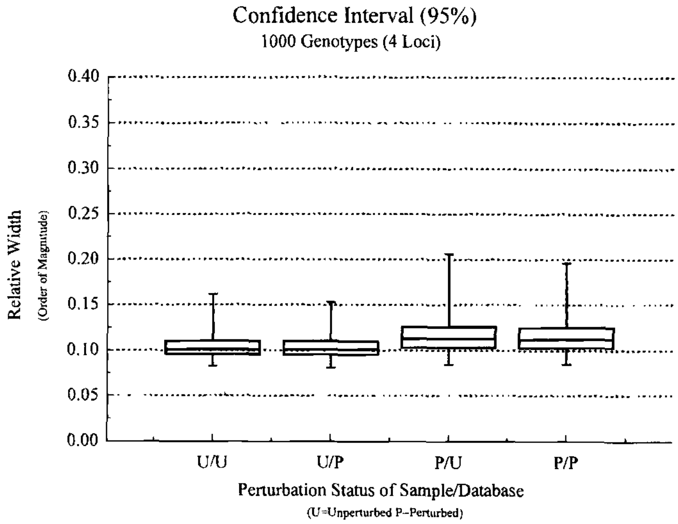

FIG. 1—*Box and whisker plots of 1000 relative 95% confidence intervals for 1000 Hispanic four-locus genotype probabilities.* P = *independent perturbation of fragment sizes.* UU—*test sample and database both unperturbed;* UP—*test sample unperturbed and database perturbed;* PU—*test sample perturbed and database unperturbed;* PP—*test sample and database both perturbed. Each box represents the 25th, 50th, and 75th percentiles, with the whiskers terminating at the extreme values of the distributions. Measurement error of database fragment sizes has almost no effect on the confidence interval width, whereas measurement error of the test fragment sizes only very slightly increase the width.*

confidence interval. Because measurement error of the database samples had such a negligible effect, none of the following analyses included this effect.

Figure 2 contains the distributions of 1000 95% confidence intervals for one, two, three, and four loci. They demonstrate the dramatic reduction in uncertainty achieved by increasing the number of loci. For one locus, the most extreme relative CI was less than one order of magnitude out of two, although the typical ratio was one out of about five. In contrast, for four loci, these figures were one out of five and one out of nine respectively.

Correlation of test genotype measurement errors can be seen in Fig. 3 to reduce slightly the typical relative CI compared with independent errors, possibly due to a bin boundary effect. With or without correlation, the ratios still exceed those observed in the absence of test genotype measurement error. The effect of correlated measurement error is compared with that of independent error for one to four loci in Fig. 4. The one- and two-locus correlated perturbation distributions especially display significantly greater kurtosis but ranges virtually identical with those yielded by independent perturbation. As the number of factors (alleles here) increase, cancellation of errors dominates the more subtle effect of error correlation.

The distribution of 1000 90%, 95%, and 99% four-locus relative confidence intervals using independent and correlated test genotype perturbations can be found in Fig. 5. Among the 99% CIs (the most extreme case), uncertainty reached about one order of magnitude out of four, although three-quarters of the CIs were less than one out of six.

These results are of course dependent upon database size. The NRC I panel called for 15–20 ethnic databases of about 100 persons each (10), whereas the NRC II panel, in contrast, recommended fewer but larger ("at least several hundred") databases (9). The databases used here, are fairly typical of those collected by many forensic laboratories. Hence, these results are pertinent to the typical forensic situation but underestimate the relative intervals that would be obtained using samples of the size recommended by the first NRC panel.

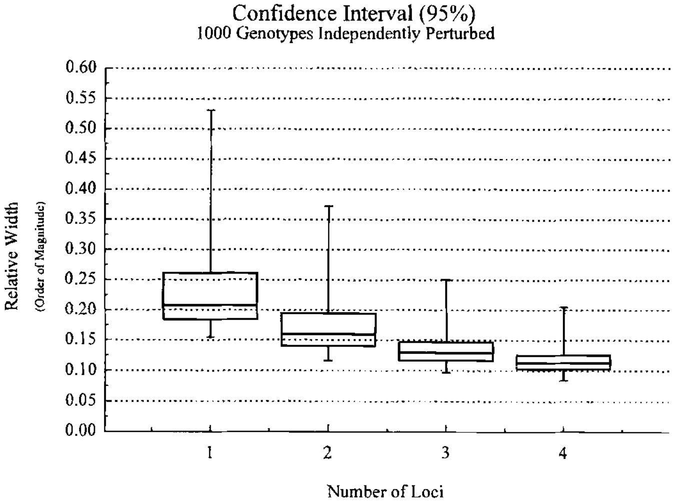

FIG. 2—*Box and whisker plots of 1000 relative 95% confidence intervals for 1000 one-, two-, three-, and four-locus Hispanic genotype probabilities in the order D2S44, D1S7, D4S139, D10S28. Each box represents the 25th, 50th, and 75th percentiles, with the whiskers terminating at the extreme values of the distributions. Increasing the number of loci dramatically reduces the confidence interval widths. For a four-locus genotype probability of 10^{-8}, a typical 95% confidence interval would be about one order of magnitude in width.*

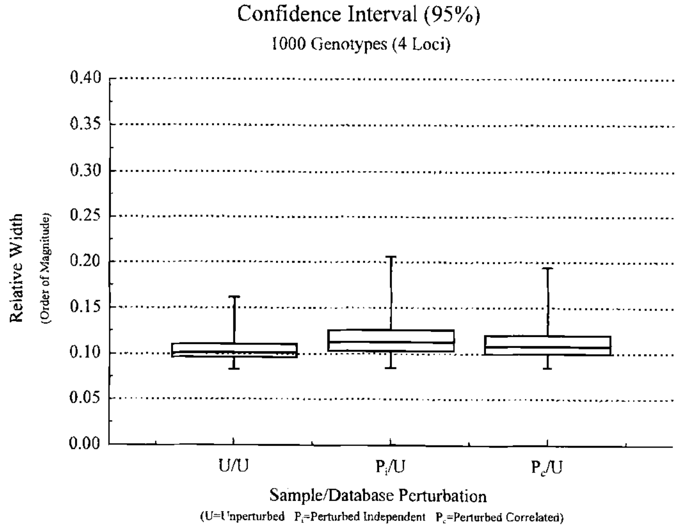

FIG. 3—*Box and whisker plots of 1000 relative 95% confidence intervals for 1000 four-locus Hispanic genotype probabilities. All database fragments unperturbed.* U = *test sample unperturbed;* P_i = *test sample independently perturbed;* P_c = *test sample perturbed with correlation. Each box represents the 25th, 50th, and 75th percentiles, with the whiskers terminating at the extreme values of the distributions. Although for a pair of fragments in a heterozygous pattern, measurement error increases the confidence interval width a small amount, correlation of these errors, which occurs in practice, reduces this effect very slightly.*

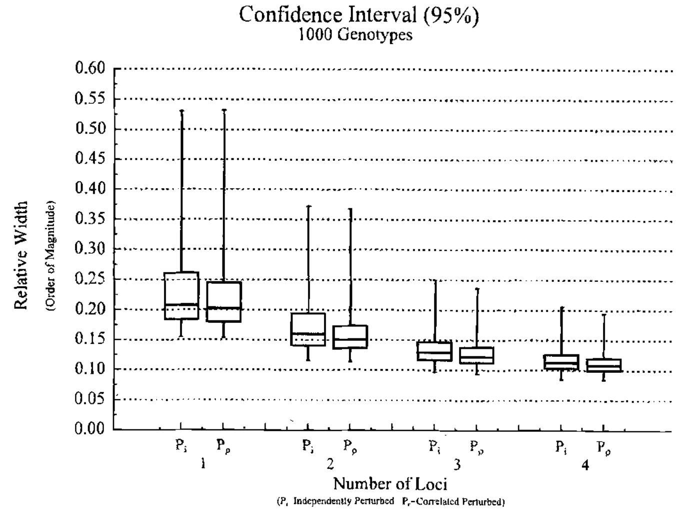

FIG. 4—*Box and whisker plots of 1000 relative 95% confidence intervals for 1000 one-, two-, three-, and four-locus Hispanic genotype probabilities in the order D2S44, D1S7, D4S139, D10S28. All database fragments unperturbed.* P_i = *test sample independently perturbed;* P_c = *test sample perturbed with correlation. Each box represents the 25th, 50th, and 75th percentiles, with the whiskers terminating at the extreme values of the distributions. The number of loci has a much greater effect on the confidence interval width than does the correlation of measurement error.*

Discussion

In this study, we have shown that measurement error of database samples as well as correlation of measurement errors for both database and test samples have little effect on genotype probability

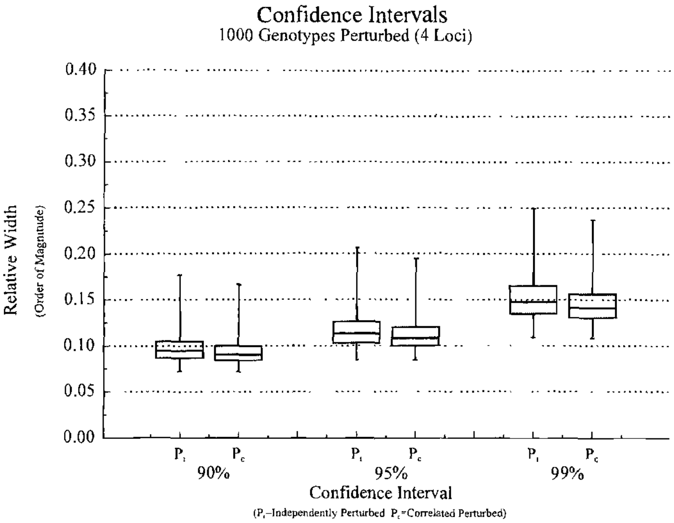

FIG. 5—*Box and whisker plots of 1000 relative 90%, 95%, and 99% confidence intervals for 1000 Hispanic four-locus genotype probabilities. All database fragments unperturbed.* P_i = *test sample independently perturbed;* P_c = *test sample perturbed with correlation. Each box represents the 25th, 50th, and 75th percentiles, with the whiskers terminating at the extreme values of the distributions. Here, the magnitude of the effect of choosing another confidence interval is shown for both independent and correlated measurement errors. Even with a 99% four-locus confidence interval, a figure as large as is customarily encountered in scientific literature, most interval widths are less than one order of magnitude out of six.*

estimates. However, database sampling error and test sample measurement imprecision are significant sources of uncertainty in genotype probability estimation. Elsewhere (22), we have shown the typical error due to the very high degree of substructuring of the East Asian race by major ethnic subpopulations was typically, for four loci, about one order of magnitude out of one hundred, and at most one out of eleven. In this study, the median 95% relative confidence interval was about one order of magnitude out of eight and on rare occasions, individual relative intervals reached one out of five orders of magnitude. Hence, the relative uncertainty due to sampling and test sample measurement imprecision is several orders of magnitude greater than relative error due to substructuring.

Bootstrap confidence intervals as used here are simultaneous. The NRC I panel advocated the use of individual 95% normal approximations to binomial confidence limits in its ceiling allele frequency calculation (10). For a multilocus estimate, this recommendation will result in excessively conservative genotype probability estimates. The error due to substructuring is several orders of magnitude less than the uncertainty due to the actual multinomial sampling uncertainty, so that simultaneous tolerance limits should suffice. These results together with our other studies show that forensic estimates would be more improved by the use of larger representative racial databases to reduce sampling uncertainty, than by more small ethnic databases in an effort to control for substructuring error.

The NRC II panel indicated the calculation of confidence intervals was desirable (10). However, the panel suggested that with the use of databases of at least several hundred persons, the expedient of simply bracketing the point estimate by plus or minus an order of magnitude. (The number of loci was unstated, but is presumably four or five.) Our study demonstrated that this method adequately covers four-locus sampling error using a database of about 200 persons, and so would be more conservative with a greater sample size and/or number of loci.

Here, we examined the effect of two types of error on fixed-bin genotype probability estimates. The levels of confidence intervals are usually interpreted as the proportion of such intervals that would contain the parameter of interest. Fixed-bin genotype probabilities are conservative relative to the matching windows used with them. In our case, we have found in unpublished work that for four loci, 94% of fixed-bin estimates were conservative relative to floating-bin estimates and the remainder underestimated by no more than one order of magnitude out of thirteen. Hence, fixed-bin confidence intervals also will usually be conservative, or almost so, in reference to the floating-bin intervals.

Acknowledgments

We thank Dr. David Gjertson for his assistance in the determination of measurement imprecision and correlation, Dr. Michael Martin for helpful suggestions on bootstrapping, and Dr. Eva Steinberger for suggestions on improving earlier versions of this manuscript.

References

1. Weir B. Population genetics in the forensic DNA debate. Proc Nat Acad Sci USA 1992;89:11654–9.
2. Evett IW, Werret DJ, Gill P, Buckleton JS. DNA fingerprinting on trial. Nature 1989;340:435.
3. Shapiro MM. Imprints on DNA fingerprints. Nature 1991;353: 121–2.
4. Evett IW. Trivial error. Nature 1991;354:114.
5. Berry DA, Evett IW, Pinchin R. Statistical inference in crime investigations using deoxyribonucleic acid profiling. Appl Stat 1992;41(3):499–531.
6. Devlin B, Risch N, Roeder K. Forensic inference from DNA fingerprints. J Am Stat Assn 1992;87:337–50.
7. Chakraborty R. Sample size requirements for addressing the population genetic issues of forensic use of DNA typing. Hum Biol 1992;64(2):141–59.
8. Evett IW, Gill P. A discussion of the robustness of methods for assessing the evidential value of DNA single locus profiles in crime investigations. Theor Appl Electrophoresis 1991;12:226–30.
9. National Research Council Committee on DNA Forensic Science. The evaluation of forensic DNA evidence. Washington (DC): National Academy Press, 1996 (prepublication copy); 4,1–5,41.
10. National Research Council Committee on DNA Technology in Forensic Science. DNA technology in forensic science. Washington (DC): National Academy Press, 1992;74–96.
11. Chakraborty R. Evaluation of standard error and confidence interval of estimated multilocus genotype probabilities, and their implications in DNA forensics. Am J Hum Genet 1993;52:60–70.
12. Goodman LA. On the exact variances of products. J Am Stat Assn 1960;55:708–13.
13. Goodman LA. On the exact product of k random variables. J Am Stat Assn 1962;57:54–60.
14. Budowle B. A protocol for RFLP analysis of forensic biospecimens. Proceedings of the International Symposium on the Forensic Aspects of DNA Analysis; 1989; Quantico (VA). Washington (DC): US Government Printing Office, 1989;57–62.
15. Baechtel FS. The extraction, purification and quantitation of DNA. Proceedings of the International Symposium on the Forensic Aspects of DNA Analysis; 1989; Quantico (VA). Washington (DC): US Government Printing Office, 1989;25–8.
16. Budowle B, Baechtel FS. Modifications to improve the effectiveness of restriction fragment length polymorphism typing. Appl Theor Electrophoresis 1990;1:181–7.
17. Monson KL, Budowle B. A system for semi-automated analysis of DNA autoradiograms. Proceedings of the International Symposium on the Forensic Aspects of DNA Analysis; 1989; Quantico (VA). Washington (DC): US Government Printing Office, 1989;127–32.

18. Efron B. The jackknife, the bootstrap and other resampling plans. Regional Conference Series in Applied Mathematics, No. 38; Philadelphia: Society for Industrial and Applied Mathematics; 1982.
19. Devlin B, Risch N, Roeder K. No excess of homozygosity at loci used for DNA fingerprinting. Science 1990;249:1416–20.
20. Chakraborty R, DeAndrade M, Daiger SP, Budowle B. Apparent heterozygote deficiencies observed in DNA typing data and their implications in forensic applications. Ann Hum Genet 1992;56: 45–57.
21. Budowle B, Monson KL. A statistical approach for VNTR analysis. Proceedings of an International Symposium on the Forensic Aspects of DNA Analysis; 1989; Quantico (VA); Washington (DC): US Government Printing Office, 1989;121–6.
22. Hartmann JM, Houlihan BT, Keister RS, Buse EL. The effect of ethnic and racial population substructuring on the estimation of multi-locus fixed-bin VNTR RFLP genotype probabilities. J Forensic Sci 1997;42(2):231–9.

Additional information and reprint requests:
John Hartmann
Forensic Science Services
Orange County Sheriff-Coroner Department
P.O. Box 449, 320 N. Flower Street
Santa Ana, CA 92703