Kenneth T. Gary,[1] *David L. Duewer,*[2] *Ph.D.; and Dennis J. Reeder,*[1] *Ph.D.*

Graphical Tools for RFLP Measurement Quality Assurance: Laboratory Performance Charts*

REFERENCE: Gary KT, Duewer DL, Reeder DJ. Graphical tools for RFLP measurement quality assurance: laboratory performance charts. J Forensic Sci 1999;44(5):978–982.

ABSTRACT: Forensic restriction fragment length polymorphism analyses typically provide two band size results at each genetic locus for each sample. In collaboration with the member laboratories of the Technical Working Group for DNA Analysis Methods, we have developed graphical techniques that compactly summarize even large numbers of such paired measurements. This paper provides a detailed description of the basic tool, a modified bivariate control chart for data from multiple samples and/or multiple loci. We demonstrate how various modifications and combinations of these "laboratory performance charts" can be used for quality control, quality assurance, and quality demonstration.

KEYWORDS: forensic science, DNA profiling, graphical analysis, interlaboratory comparison, proficiency testing, quality control, TWGDAM

The National Institute of Standards and Technology collaborates with member laboratories of the Technical Working Group for DNA Analysis Methods (TWGDAM) in the study and documentation of DNA measurement technologies for human identification (1–6). While newer technologies are of increasing importance, restriction fragment length polymorphism (RFLP) techniques continue to be used in many forensic and parentage testing laboratories.

We previously described a graphical tool, termed the single-locus chart (SLC), for displaying large amounts of RFLP data for a given sample at a given genetic locus (7). With two modifications, the basic components of the SLC can be used to create a laboratory performance chart (LPC) for each participant in an among-laboratory comparison exercise. Each LPC efficiently displays one laboratory's measurements for all samples at all loci, enabling identification of a number of laboratory-specific measurement characteristics.

We present here our procedures for preparing RFLP measurement LPCs. Given the routine summary statistics (the mean or median band size for each allele of each sample for all loci), the essential components of the LPC can be constructed with pen and graph paper. All aspects are more easily accomplished with simple spreadsheet programs. However constructed, this graphical technique can be a useful tool for RFLP quality assurance as well as providing a readily understood demonstration of laboratory measurement quality.

[1] Biotechnology Division, Chemical Science and Technology Laboratory, National Institute of Standards and Technology, Gaithersburg, MD.

[2] Analytical Chemistry Division, Chemical Science and Technology Laboratory, National Institute of Standards and Technology, Gaithersburg, MD.

* This work was supported in part by the Federal Bureau of Investigation and by the National Institute of Justice, U.S. Department of Justice.

Received 6 Oct. 1998; accepted 25 Jan. 1999.

Methods and Materials

Demonstration Data

The data presented here were taken from one or more reports of the following: the Forensic Laboratory Proficiency Testing Program, Collaborative Testing Services, Inc., Herndon, VA; the International Quality Assessment Scheme, Cellmark Diagnostics, Oxon, UK; and the Forensic Identity Program, College of American Pathologists, Chicago, IL.

Computation

All graphics were generated using commercial spreadsheet software.

Design and Construction

The design and construction of the single-locus chart (SLC), typically displaying data from many laboratories for a single sample at a single locus, has been presented previously (7). The laboratory performance chart (LPC), displaying data for one laboratory for multiple samples at multiple loci, can be regarded as an overlay of many SLCs with each overlay displaying only data for one laboratory. Figure 1 is an example LPC; its major components are similar to those of the SLC with modified axis labels and standardized measurement correlation.

Axis Labels

Two RFLP band size measurements are typically made for each sample at a given locus. These measurements are represented as the ordered pair $\{x_1, x_2\}$, where $x_1 \geq x_2$. A set of these ordered pairs can be displayed as an SLC scattergram, plotting the smaller measurement as a function of the larger. The scattergram is centered on the expected size of the two bands, $(\overline{X}_1, \overline{X}_2)$. Each axis of the scattergram is scaled to a fixed number of expected standard deviations (S) about its center; S for a band of size $\overline{X}$ is

$$S(\overline{X}) = 7.5\left(1 + \frac{\overline{X}}{19{,}500}\right)^{7.1} \quad (1)$$

for $\overline{X}$ from about 1000 base pairs (bp) to 22,000 bp (4). While SLC axes are labeled in bp to simplify interpretation, the tic-mark interval for every axis is $S(\overline{X})$.

LPCs present the data in the same manner, but with the axes labeled as number of $S(\overline{X})$ about the expected value. This is equiva-

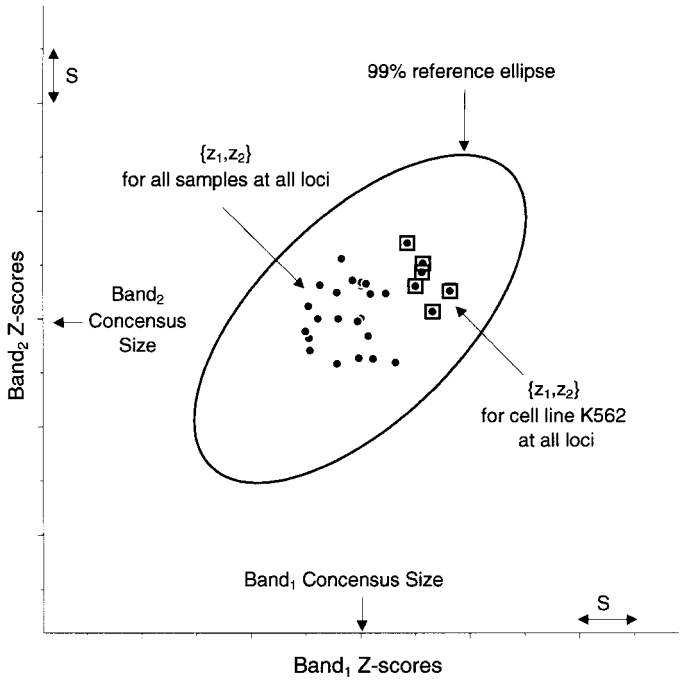

FIG. 1—*Example Laboratory Performance Chart (LPC). The* Z*-score data are denoted by solid circles, the cell line K562 control data are highlighted by open squares, and the 99% reference ellipse is represented as a solid line.*

lent to a "*Z*-score" transformation of all measurement pairs (8)

$$z_1 = \left(\frac{x_1 - \overline{X}_1}{S(\overline{X}_1)}\right), \quad z_2 = \left(\frac{x_2 - \overline{X}_2}{S(\overline{X}_2)}\right) \tag{2}$$

Each *z*-transformed measurement has zero mean and unit *S*, allowing multiple SLCs to be plotted on top of one another with common center {0,0} and scale (unit *S*). *Z*-scoring refocuses data interpretation from absolute bp measurements to the relative magnitude of displacement from the expected values.

Reference Ellipse

A bivariate ellipse can be defined to enclose a given percentage of valid data for any given single-locus, single-sample data set (7). Since *Z*-score data have mean zero and unit *S*, the generating equation for a 99% tolerance ellipse about *Z*-score data is:

$$\begin{aligned} Z_2 &= \phi \\ Z_1 &= \phi R \pm \sqrt{(1 - R^2)(K_{bi_{99}} - \phi^2)} \end{aligned} \tag{3}$$

where the value for $K_{bi_{99}}$ is approximately 9.21 for bivariate normal data, ϕ has domain $-\sqrt{K_{bi_{99}}} \le \phi \le +\sqrt{K_{bi_{99}}}$, and R represents the inherent correlation between replicate measurements of the particular pair of RFLP bands.

The observed correlation between $\{x_1, x_2\}$ measurement pairs is a complex function of the relative location of sample and sizing ladder bands. We currently approximate R as

$$R = 0.72 - 0.65 \times \log_{10}(x_2/x_1) \tag{4}$$

Z-scoring does not change the numerical value of *R* (8). Thus, 99% tolerance ellipses for different data sets have different shapes even though they share the same origin and scale. In order to display a common reference ellipse, each data set is further transformed to have a constant reference *R* value (R_{ref})

$$z_1' = z_1 \times Q, \quad z_2' = z_2 \times Q \tag{5}$$

where

$$Q = +\sqrt{\frac{(z_1^2 - z_2^2 - 2 \times z_1 \times z_2 \times R)(R_{ref}^2 - 1)}{(z_1^2 - z_2^2 - 2 \times z_1 \times z_2 \times R_{ref})(R^2 - 1)}} \tag{6}$$

Figure 2 illustrates the effect of this transformation on the 99% tolerance ellipses, where R_{ref} is defined to be 0.60. Figure 3 illustrates the effect of the transformation on one particular data set. The transformation has little effect on data close to the origin or when *R* has a value close to R_{ref}. Because of the distortion among data sets and the variable number of data in each data set, the ellipse is more for visual reference than any rigorous statistical evaluation.

Examples and Discussion

Snapshot Evaluations

The LPCs in Figs. 4 through 6 display exemplar proficiency data taken from recent interlaboratory proficiency tests. Figure 4 displays laboratories that are in good concordance with the consensus values (the *Z*-scores average to about zero), with overall measurement precision (the scatter of the data about its center) ranging from excellent (Fig. 4A) to poor (Fig. 4C). We note that the data displayed in Fig. 4C were from a laboratory that used a manual digitizing pad rather than computerized image processing.

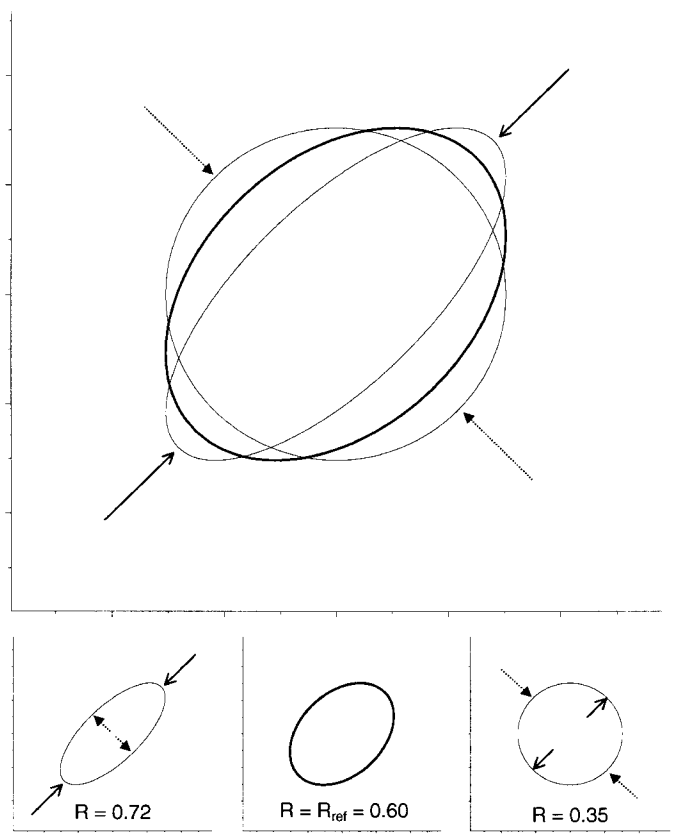

FIG. 2—*Overlay of LPC ellipses with different* R *values. The darker ellipse is the "target" ellipse to which the other two must conform.*

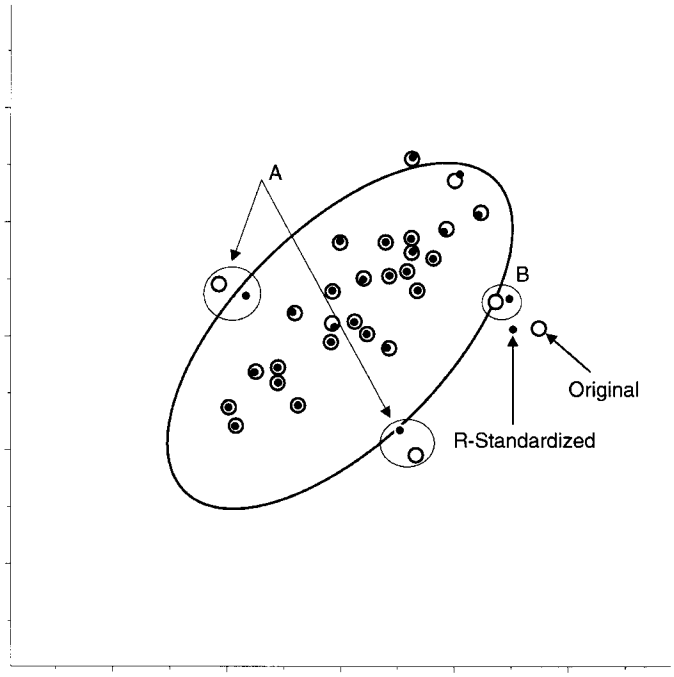

FIG. 3—*Overlay of original and R-standardized data. The measurement pairs labeled "A" are shown to become more comparable with R-standardization; the pair labeled "B" is revealed as less comparable.*

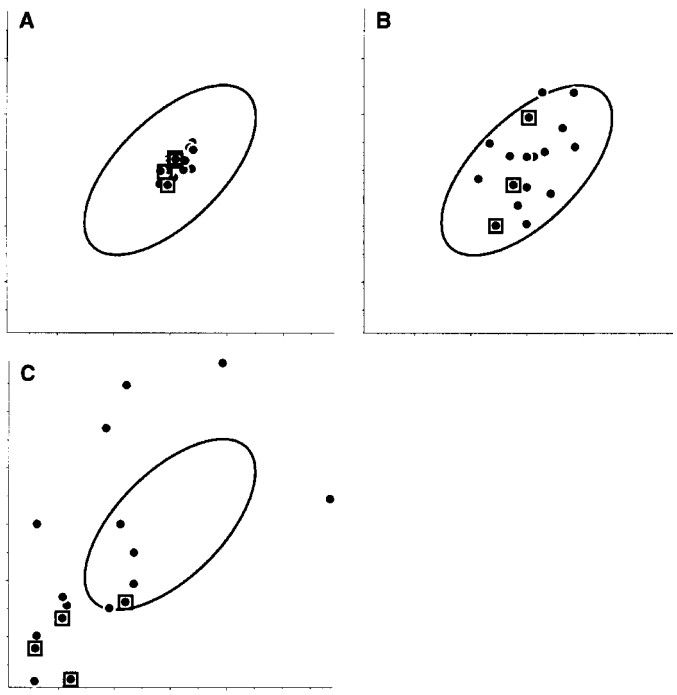

FIG. 4—*Measurement concordance and apparent precision. LPC 4A displays one set of proficiency test data for a laboratory with excellent among-laboratory concordance and within-laboratory precision: all points are near the center of the ellipse. LPC 4B displays data for a laboratory with adequate concordance but marginal apparent precision: the data are symmetrically scattered just within the reference ellipse. LPC 4C displays data for a laboratory with inadequate apparent precision: while fairly symmetrically scattered, many of the data are outside the reference ellipse.*

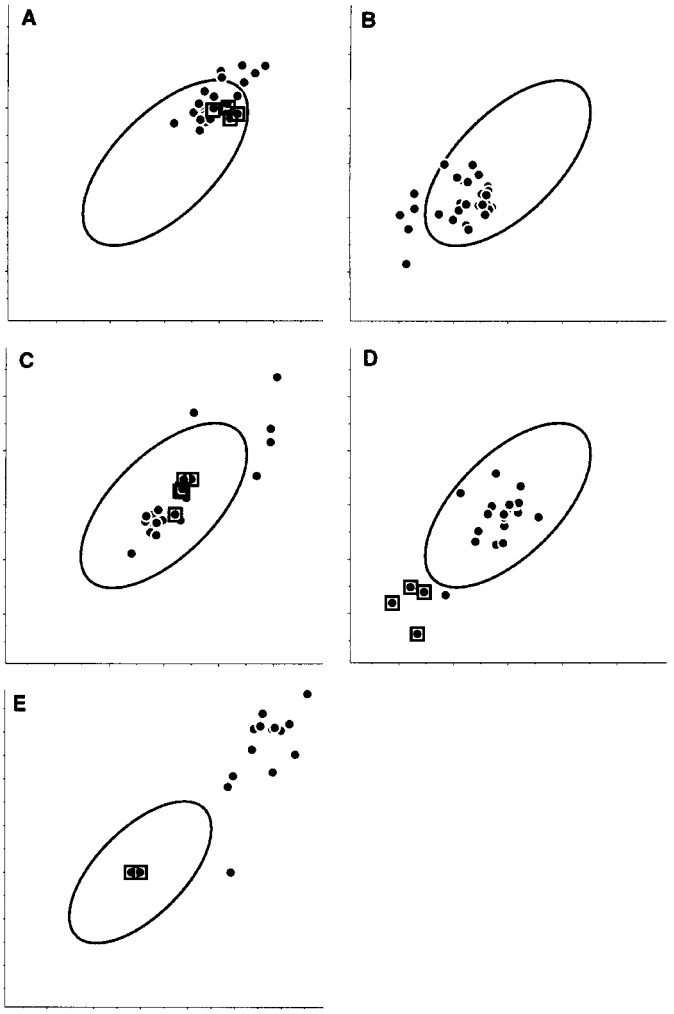

FIG. 5—*Patterns of measurement discordance. LPCs 5A and 5B display sets of proficiency test data for laboratories with good within-laboratory precision but which are consistently discordant from the consensus values. LPCs 5C and 5D display data sets for laboratories making sample-specific measurement errors: all data outside the reference ellipse in 5C was reported for one unknown sample at different loci, nearly all data outside the ellipse in 5D were reported for the participant's cell line K562 control. LPC 5E displays a data set with highly discordant data for all (unknown) tests samples but highly concordant (known) K562 control values (the axis range is here adjusted to* $\pm 10\ S(\bar{X})$ *to enable viewing of all data).*

Figures 5A and 5B show data with good overall measurement precision but poor measurement concordance. This pattern of discordance may indicate that these particular laboratories use protocols that consistently differ in some manner from those of the majority of participants.

Figures 5C and 5D show sample-specific errors. In Fig. 5C, all data for one particular test sample are discordant, while the data for the other samples are in good concordance. In Fig. 5D, only the laboratory-supplied cell line K562 control data are discordant. We have observed these patterns when the loading buffer for a particular sample differs from that used for the other samples and/or when too much DNA is loaded on the gel.

Figure 5E is a more alarming instance of sample-specific discordance: all data for the test samples (all band sizes unknown) are outside the reference ellipse while all data for the K562 control (all band sizes known) are well within the ellipse. While this may also

reflect differences in the loading buffer, it is possible that the reported K562 values are not measured values—at least not control measurements for the gel used for the test samples.

The data patterns seen in Figs. 5C through 5E are uncommon but have been observed in many different proficiency tests. This challenges the adequacy of using a single cell line control as a gatekeeper for among-laboratory exchange of DNA profiling information. Use of Blind Internal Standard (BIS) bloodstain samples such as those adopted by the Royal Canadian Mounted Police forensic laboratory system could increase the confidence of such exchange (9).

Tracking Performance over Time

The LPC can also be used to document measurement performance over time by plotting multiple charts in chronological order. Figure 6 tracks the performance of one particular laboratory in six consecutive forensic proficiency tests, "9503" through "9715," conducted by Collaborative Testing Services, Inc. This laboratory made minor modifications to its protocol throughout these studies. Four significant modifications were made between tests "9603" and "9615": from autoradiography to luminography, from one manufacturer's sizing ladder to another, from medium electroendoosmotic (ME) agarose to low electroendoosmotic (LE) agarose, from use of ethidium bromide (EB) in the running buffer during electrophoresis to use of EB only after electrophoresis is completed. These changes somewhat improved this laboratory's among-laboratory measurement concordance and dramatically improved its within-laboratory measurement precision.

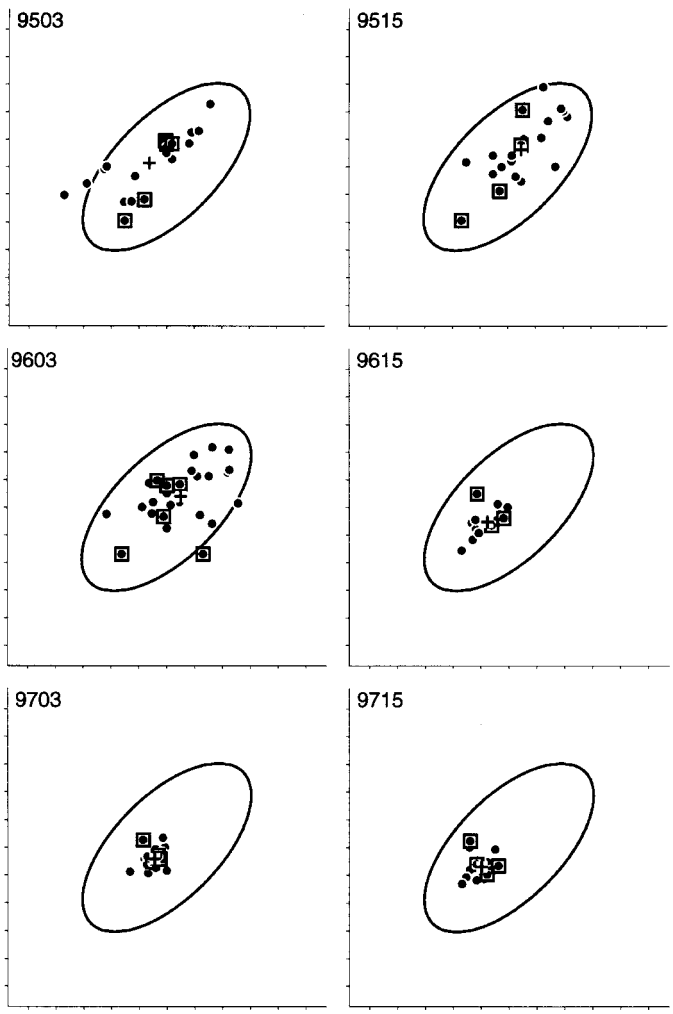

FIG. 6—*Tracking measurement performance over time. These six LPCs display results for one laboratory in six consecutive forensic proficiency tests conducted by Collaborative Testing Service, Inc. The average concordance in each test is denoted "+".*

TABLE 1—*RFLP protocol variants reported in Collaborative Testing Services, Inc., Forensic Proficiency Tests from 1995 through 1997.*

	Proficiency Test					
Modification	9503	9515	9603	9615	9703	9715
% with LE agarose, no EB in running buffer	na	69	72	78	78	84
% using luminography	15	29	27	44	47	63
% using a "BRL" sizing ladder	74	73	75	84	87	96
Number of participants	74	96	118	88	99	93

Apparent Precision

Of the four major changes that the laboratory in Fig. 6 made to its protocol, we believe changes in the use of EB may account for the pattern of improvement that was observed. EB is known to change the relative location of sizing ladder and sample bands (4,8). The effect of intercalation of EB into ladder and sample fragments is dependent on the size of the fragment, its composition and sequence, and the relative concentration of EB in the running buffer (4). When EB is included in the running buffer, a slightly different set of unknown fragments is measured than is the case if EB is not included. The offset inferred is on the order of $\pm 0.5\, S(\bar{X})$ for each band size measurement, but this offset is not predictable and cannot be "mathematically" eliminated to determine the "true precision" (estimated from within-laboratory replicate measurements) of any given measurement. We term this pattern of fragment-specific discordance from consensus values "apparent precision."

As shown in Table 1, an increasing number of participants have switched from ME to LE agarose over successive years in combination with discontinuance of EB in the running buffer. It appears that the forensic community has converged on a *de facto* standard LE agarose-based protocol that does not include EB in the running buffer. Had the insights obtainable from SLC and LPC graphical displays been available in the early stages of forensic RFLP measurements, perhaps this convergence could have been accelerated.

Acknowledgments

We thank Cellmark Diagnostics, Collaborative Testing Services, Inc., and the College of American Pathologists for making complete sets of their forensic DNA proficiency testing program reports available to us.

References

1. Mudd JL, Baechtel FS, Duewer DL, Currie LA, Reeder DJ, Leigh SD, et al. Interlaboratory comparison of autoradiographic DNA profiling measurements. 1. Data and summary statistics. Anal Chem 1994;66:3303–17.

2. Kline MC, Duewer DL, Newall MA, Redman JW, Reeder DJ, Richard M. Interlaboratory evaluation of short tandem repeat triplex CTT*. J Forensic Sci 1997;42:897–906.
3. Duewer DL, Currie LA, Reeder DJ, Leigh SD, Liu H-K, Mudd JL. Interlaboratory comparison of autoradiographic DNA profiling measurements. 2. Measurement uncertainty and its propagation. Anal Chem 1995;67:1220– 31.
4. Stolorow AM, Duewer DL, Reeder DJ, Buel E, Herrin G. Interlaboratory comparison of autoradiographic DNA profiling measurements. 3. Repeatability and reproducibility of RFLP band sizing, particularly bands of molecular size >10k base pairs. Anal Chem 1996;68:1941–7.
5. Duewer DL, Currie LA, Reeder DJ, Leigh SD, Liu H-K, Mudd JL. Interlaboratory comparison of autoradiographic DNA profiling measurements. 4. Protocol effects. Anal Chem 1997;69:1882–92.
6. Duewer DL, Benzinger EA. Products of partial digestion with *Hae* III. Part 2. Quantification. J Forensic Sci 1997;42:850–63.
7. Duewer DL, Liu H-K, Reeder DJ. Graphical tools for RFLP measurement quality assurance: single-locus charts. Submitted to J Forensic Sci (5/97).
8. Mendenhall W. Introduction to probability and statistics, Wadsworth Publishing Co., Belmont, California 1967.
9. Duewer DL, Lalonde SA, Aubin RA, Fourney RM, Reeder DJ. Interlaboratory comparison of autoradiographic DNA profiling measurements: precision and concordance. 1997;43:465–71.

Additional information and reprint requests:
Dennis J. Reeder
NIST
100 Bureau Drive, STOP 8311
Gaithersburg, MD 20899-8311
email: dennis.reeder@nist.gov