1.1 Biostatistics

Measurement Error in Determination of Band Size for Highly Polymorphic Single-Locus DNA Markers

D.W. Gjertson, J. Hopfield, P.A. Lachenbruch*, M.R. Mickey, T. Sublett, C. Yuge, and P.I. Terasaki

UCLA Tissue Typing Laboratory, 1000 Veteran Avenue, Los Angeles, California, 90024

*Department of Biostatistics, UCLA School of Public Health, Los Angeles, California, 90024

INTRODUCTION

When assessing single-locus DNA information in reference to forensic and parentage problems, most proposed statistical methods (Baird et al, 1986; Gjertson et al, 1988; Morris et al, to be published; and Berry, submitted for publication) incorporate continuous measurement errors (ϵ) into their calculations. Errors arise while distinguishing length of enzyme-cleaved fragments from mobility in gel electrophoresis. Statistically, actual true gene size plus measurement error constitutes the underlying model for an observed allele (sometimes on the log scale) where errors are usually assumed to be distributed normally with mean zero and variance σ^2.

The following is a report of an experiment conducted to determine the standard error in measuring RFLP alleles (log scale). The DNA sequences of the D14S1 region (Wyman and White, 1980) provided the practical example.

Background

Differences among individuals in the lengths of their DNA fragments (or, equivalently, their size measured in number of base pairs) are inherited characteristics that can be recognized by altered mobility of bands in agarose gel electrophoresis (Botstein et al, 1980). To estimate the length of unknown DNA fragments (L) from mobility of bands (m) in gel electrophoresis, a relationship is established between the mobilities and lengths of standard fragments whose lengths are known. Figure 1 depicts three plots of band mobility versus length for 11 standard fragments. The log model (Fisher and Dingman, 1971) and inverse model (Southern, 1979) are the two basic relationships proposed to estimate L from m. Even in their initial studies, Fisher and Dingman (1971) recognized the inadequacy of the log model over a wide range of fragments where plots of log L versus m showed marked departures from linearity for fragments with high molecular weights. Likewise, Southern (1979) observed curvature in plots of the inverse model (i.e., L = k/m) when electrophoresis was carried out at

Advances in Forensic Haemogenetics 3
Edited by H. F. Polesky and W. R. Mayr

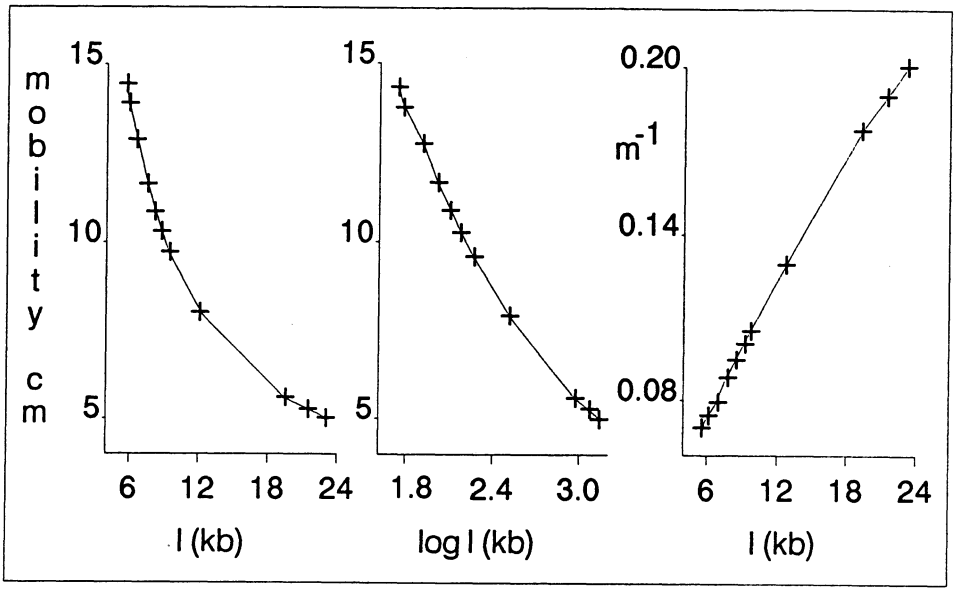

Fig. 1. Three plots of gel band mobility versus length of 11 standard DNA fragments. Ruled mobilities (cm) are 5.0, 5.3, 5.6, 8.0, 9.7, 10.4, 10.8, 11.7, 12.8, 13.9 and 14.4. The respective DNA fragment sizes (kb) are 23.130, 21.226, 19.397, 12.220, 9.416, 8.614, 8.271, 7.421, 6.682, 5.804 and 5.643.

high voltage gradients. He proposed a correcting factor (m_o) to give the best fit to a line of the form: $L = k_1/(m-m_o) + k_2$. Schaffer and Sederoff (1981) justified the correction as a means to account for factors which may affect the apparent location of the origin and further derived a "mechanism of migration", based on kinetic principles, which is consistent with the inverse model. Elder and Southern (1983) compared several methods (all variations of the log or inverse models) for relating mobility to fragment length. They concluded that the reciprocal relationship was the most accurate (maximum error < 0.1% of fragment size). The accuracy in mobility measurement, under any of the above methods, can be further enhanced by densitometers and computer processing techniques (Agard et al, 1981 and Elder et al, 1983).

MATERIAL AND METHODS

Figure 2 illustrates the general procedure for producing D14S1 RFLPs used in the subsequent empirical study. The extraction of high-molecular-weight genomic DNA and digestion with restriction enzyme EcoRI were performed by the Maniatis et al method (1982) with slight modification (Gjertson et al, 1988). Hybridization followed for 36 hours in the presence of ^{32}P

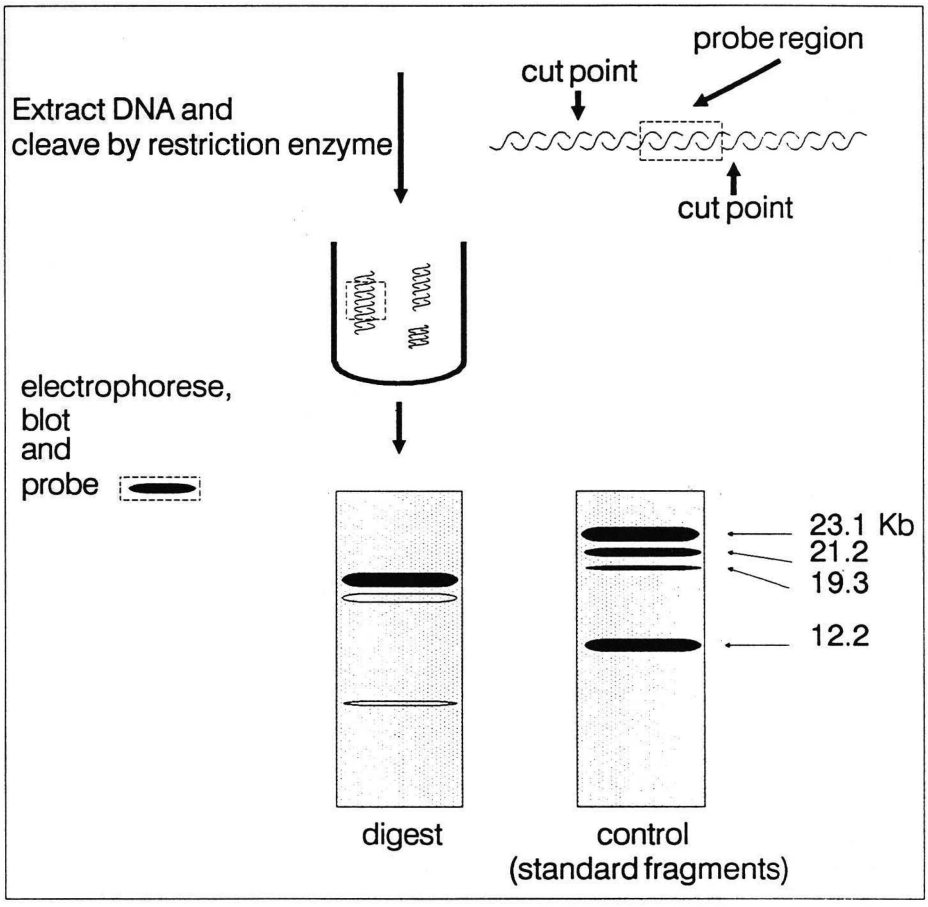

Fig. 2. D14S1 RFLP production. Extracted DNA is cleaved with restriction enzyme EcoRI, electrophoresed on agarose gel, hybridized with ^{32}P-labeled D14S1 probe, pAW101, and visualized by autoradiography. D14S1 fragment sizes are determined by calibrating the gel using known standards.

labeled D14S1 probe, pAW101 (kindly provided by Dr. Ray White, University of Utah). Dried membranes were exposed overnight on X-Omat-AR film (Kodak) between two intensifying screens (Li+) at -80C.

The mobility of D14S1 fragments and standard bands was measured from a common baseline to the nearest mm by two technicians.

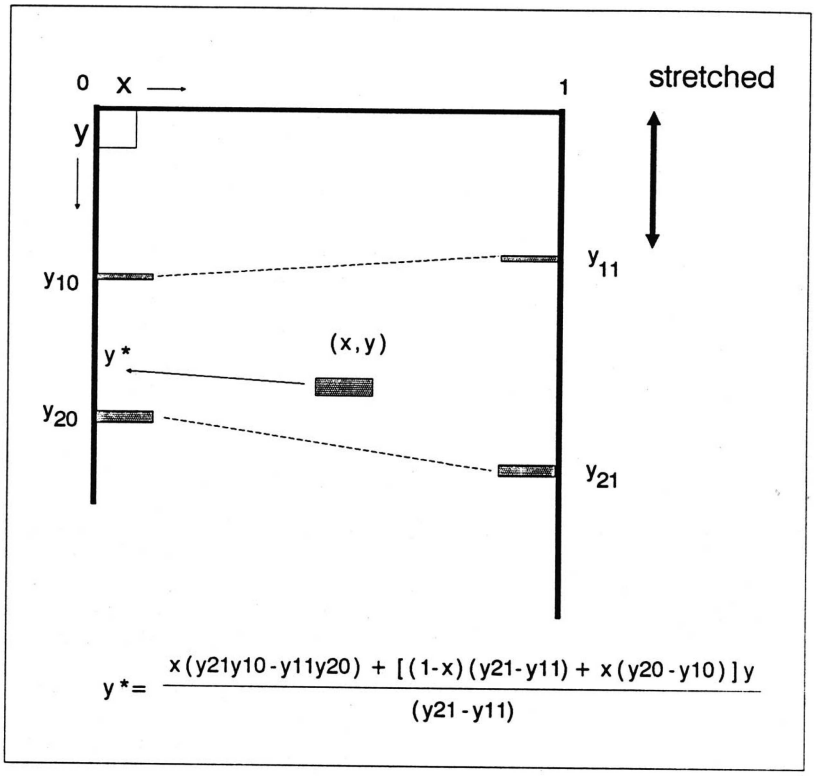

Fig. 3. Translation of RFLP gel bands. X,Y are the coordinates of an untranslated band bracketed by standards y_{10},y_{20} at x=0 and y_{11},y_{21} at x=1. Y^* is the value of y translated to x=0.

Before converting mobility to fragment size, all measurements were translated to a standard lane (usually the left-most lane), accounting for gel irregularities such as membrane stretching during Southern (1975) transfer. As diagrammed in Fig. 3, let (x,y) be the coordinates of an untranslated band bracketed by standards y_{10},y_{20} at x=0 and y_{11},y_{21} at x=1. Here x represents distance (normalized to one) across lanes of the gel and y represents band migration distances. Y^* is the value of y at x=0. If x=0, then, obviously, y^*=y. Next, assume gel irregularities cause linear disturbances across gels (i.e., gels are stretched from one edge in a technique analogous to pulling on a sheet of rubber). Thus, when x=1, y^*=c+dy. So, with c and d increasing linearly in x on [0,1],

$$y^* = xc + [1 + x(d-1)]y. \qquad [1]$$

Solving for c and d in terms of y_{10}, y_{20}, y_{11} and y_{21} gives the desired result for translating mobilities,

$$y^* = \frac{x(y_{21}y_{10}-y_{11}y_{20}) + [(1-x)(y_{21}-y_{11})+(x)(y_{20}-y_{10})]y}{(y_{21}-y_{11})} \quad [2]$$

Analyzing the mobilities of four standard fragments showed that, in the size region 14 kb to 24 kb where D14S1 markers are generally found, log of fragment size and mobility are nearly linear. (The average R^2 of 24 sets of measurements equaled .995 with range .986 to .999.) Thus, following translation, unknown band mobilities were converted to fragment size from the functional log relationship of standards on the same gel.

DNA extracted from the white blood cells of 72 unrelated putative fathers and mothers was split into containers and enzyme-digested with EcoRI via the methods listed above. Each enzyme-digested aliquot was replicated once (i.e., a total of four samples on each subject was generated). The 288 samples were systematically allocated into 12 gels (24 lanes each). The allocation scheme regarded the 12 gels as 6 replicates of 2 gels (A and B), 12 subjects in each. Four of the 12 subjects had all replicates placed within the same gel (e.g., 2 subjects in gel A and two in gel B). Four subjects had their replicates placed across gels with one aliquot-replicate pair in gel A and the other in gel B. The remaining four subjects had their replicates placed across gels with each aliquot-replicate pair also split across gels. In addition, the four subjects at each level of gel assignment were assigned to lanes such that one of the following conditions held: (1) replicates were adjacent near edge of gel; (2) replicates were adjacent in middle of gel; (3) replicates were apart near edge; or (4) replicates were apart dispersed in middle. In total, the allocation scheme placed 6 subjects in each of the 12 cells formed by the 3 levels of gel and the 4 levels of lane position effects.

Next, gels were electrophoresed, DNA-probed with pAW101, and developed onto film according to the techniques mentioned above. Each exposed band on the finished film was read by two independent readers who were blinded to the allocation scheme. Band migration was assigned by measuring the distance (mm) from baseline to the center of the band. Finally, migration distance was converted to fragment size.

RESULTS

A total of 396 exposed bands were translated. Seven (10%) subjects exhibited no bands in any of their replicates (i.e., the measured mobility was 0.0, 0.0). Three of these subjects were retested under more stringent conditions whereby double enzyme-digested samples exhibited bands. The most probable explanation for the initial missing bands is incomplete digestion with enzyme (Rittner et al, 1989). The seven individuals were dropped from the analysis. In the remaining

65 individuals, all replicates measured consistently with 34 (52.3%) "heterozygotes" (i.e., two resolvable bands) and 31 (47.7%) "homozygotes" (i.e., only one distinguishable band). The 396 exposed bands represented 99 independent observations, replicated 4 times each.

The absolute difference (D) between the two technicians' converted log fragment size determinations of the 396 recorded bands was used as the experimental unit in the analysis of gel and lane effects. Before testing for these effects, diagnostic plots (results not shown) from BMDP2V (Dixon, 1983) suggested that observations be transformed since cell means and cell standard deviations appeared related (R^2 = .6290). The square root transformation [i.e., $D^* = D^{\frac{1}{2}} + (D+1)^{\frac{1}{2}}$] alleviated ($R^2$ = .0229) the observed linear relationship. The analysis of variance for gel and lane effects following the square root transformation performed by BMDP2V (Dixon, 1983) is summarized in Table 1. Table 1 shows cell counts and transformed cell means of the absolute difference between the technicians' fragment size determinations for the 12 different combinations of gel and lane effects. Included at the bottom of Table 1 are ANOVA tail probabilities for gel (p=.34), lane (p=.12), and replicate (p=.89) effects. No strong evidence exists to refute claims that gels were manufactured uniformly (i.e., possible gel impurities affecting migration are spread throughout and across gels).

Table 1. Results of analysis of gel and lane effects: mean and number per cell of square root transformed absolute differences in log fragment size measurements

	reps adjacent		reps apart	
	edge	middle	edge	middle
same gel	1.112 (6)	1.083 (7)	1.072 (9)	1.072 (7)
aliquots different gels	1.109 (10)	1.103 (9)	1.088 (11)	1.095 (4)
replicates different gels	1.101 (10)	1.079 (11)	1.086 (8)	1.071 (7)

p-value for gel=0.34, lane=0.12, and reps=0.89.

Table 2. Estimates of variance components for D14S1 fragments (log scale), N = 99

$\hat{\sigma}^2_{subj}$	= .00625	
$\hat{\sigma}^2_{tech}$	= .00000	
$\hat{\sigma}^2_{aliq}$	= .00004	
$\hat{\sigma}^2_{reps}$	= .00031	s.e. ≈ .02
$\hat{\sigma}^2_{\epsilon}$	= .00009	

Actual true D14S1 allele size plus error constituted the underlying model for an observed fragment size y (log scale). Further, the model was decomposed into overall (μ), subject (S), aliquot (A), replicate (R), technician (T) and error (ϵ) effects,

$$y_{ijkl} = \mu + S_i + A_{j(i)} + R_{k(ij)} + T_l + \epsilon_{ijkl}, \qquad [3]$$

where i=1,...,99, and j,k,l=1,2. The assumptions of the model included (1) Y_{ijkl} are observable random variables; (2) μ is an unobservable constant; and (3) $S_1,...,S_I$ (subjects), $A_{1(1)},...,A_{J(I)}$ (aliquots within subjects), $R_{1(11)},...,R_{K(IJ)}$ (replicates within aliquots), $T_1,...,T_L$ (technician), and $\epsilon_{1111},...,\epsilon_{IJKL}$ (error) are unobservable random variables that are pairwise uncorrelated, have zero means, and var[S]=σ_S^2, var[A]=σ_A^2, var[R]=σ_R^2, var[T]=σ_T^2 and var[ϵ]=σ_ϵ^2. Also, the unobservable random variables are distributed jointly normally on the log scale. The results of the analysis of variance components performed by BMDP8V (Dixon, 1983) are summarized in Table 2. As expected, the polymorphic nature of D14S1 was demonstrated by the relatively large subject variance estimate ($\hat{\sigma}^2_S$=0.00625). Also, variability due to technicians ($\hat{\sigma}^2_T$=.000002), and interactions (e.g., between subjects and technicians, $\hat{\sigma}^2_{ST}$=.00003, and between technicians and aliquots, $\hat{\sigma}^2_{TA(S)}$=.00001) were quite small. Thus, an estimate of total fragment size measurement error ($\hat{\sigma}^2$) was obtained by summing the remaining components:

$$\hat{\sigma}^2 = \hat{\sigma}^2_{A(S)} + \hat{\sigma}^2_{R(SA)} + \hat{\sigma}^2_{TR(SA)},$$

$$= .00004 + .00031 + .00009 = .00044.$$

Therefore, $\hat{\sigma} \approx 0.02$ or D14S1 gel measurement error is approximately 2% of fragment size with the aforementioned methodology. Noticing that replication was the largest factor contributing to σ^2, the analysis was rerun using a more parsimonious model, $y_{ij} = \mu + S_i + \epsilon_{j(i)}$, where i=1,...,99 and j=1,...,8. Under this simple repeated measures model, σ^2 is

estimated by σ^2_ϵ, and $\hat{\sigma}^2_\epsilon$ = .00041 (again, ≈2% of fragment size) with a 95% CI of (.00037,.00045).

CONCLUSIONS

Using D14S1 replicates, we have illustrated a method of determining the size of random measurement error for an RFLP system. Our rather crude (ruling) measuring technique for DNA bands produced an error rate of approximately 2% of band size. Further refinements in measuring will reduce these errors, but determining their size is necessary for assessing probabilities of identification and parentage in criminal and civil disputes.

Systematical variability between gel lanes was controlled by translating all measurements to the left-most standard lane via 2-dimensional linear regression with the right-most standard lane. Based on our same-subject replicate data, a linear function appears appropriate when both standard lanes are proximal to unknowns. Recently, for reasons of possible degradation of heterozygotes and observed isolated lane irregularities, Landers (1989) and Morris (1989) have suggested that positive controls be placed in each lane bracketing the region of mobility to improve quality control. With such controls in place, translating measurements has even simpler form (set x=1 in the formula [2] for y^*).

REFERENCES

Agard DA, Steinberg RA, Stroud RM (1981) Quantitative analysis of electrophoretograms: a mathematical approach to super-resolution. Anal Biochem 111:257-268

Baird M, Balazs I, Giusti A, Miyazaki L, Nicholas L, Wexler K, Kanter E, Glassberg J, Allen F, Rubinstein P, Sussman L (1986) Allele frequency distribution of two highly polymorphic DNA sequences in three ethnic groups and its application to the determination of paternity. Am J Hum Genet 39:489-501

Botstein D, White RL, Skolnick M, Davis RW (1980) Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am J Hum Genet 32:314-331

Berry DA (submitted for publication) Inferences in forensic identification and paternity cases using highly polymorphic DNA sequences.

Dixon WJ (ed) (1983) BMDP Statistical Software, 1983 printing with additions. University of California Press, Berkeley, California

Elder JK, Amos A, Southern EM, Shippey GA (1983) Measurement of DNA length by gel electrophoresis. I. Improved accuracy of mobility measurements using a digital microdensitometer and computer processing. Anal Biochem 128:223-226

Elder JK, Southern EM (1983) Measurement of DNA length by gel electrophoresis. II: Comparison of methods for relating mobility to fragment length. Anal Biochem 128:227-231

Fisher MP, Dingman CW (1971) Role of molecular conformation in determining the electrophoretic properties of polynucleotides in agarose-acrylamide composite gels. Biochemistry 10:1895-1899

Gjertson DW, Mickey MR, Hopfield J, Takenouchi T, Terasaki PI (1988) Calculation of probability of paternity using DNA sequences. Am J Hum Genet 43:860-869

Landers ES (1989) DNA fingerprinting on trial. Nature 339:501-505

Maniatis T, Fritsch EF, Sambrook J (1982) Molecular cloning: a laboratory manual. Cold Spring Harbor Laboratory, Cold Spring Harbor, New York

Morris JW (1989) Personal communication.

Morris JW, Sanda AI, Glassberg J (to be published) Biostatistical evaluation of evidence from continuous alleles frequency distribution DNA probes in reference to disputed paternity and disputed identity. J Forensic Sci

Rittner C, Schacker U, Rittner G, Schneider PM (1989) DNA polymorphisms in paternity testing: chances, risks, and strategies. Biotest Bulletin 4:27-33

Schaffer HE, Sederoff RR (1981) Improved estimation of DNA fragment lengths from agarose gels. Anal Biochem 115:113-122

Southern EM (1975) Detection of specific sequences among DNA fragments separated by gel electrophoresis. J Mol Biol 98:503-517

Southern EM (1979) Measurement of DNA length by gel electrophoresis. Anal Biochem 100:319-323

Wyman AR, White R (1980) A highly polymorphic locus in human DNA. Proc Natl Acad Sci 77:6754-758