PATERNITY INDEX CALCULATIONS IN SINGLE LOCUS HYPERVARIABLE DNA PROBES:

VALIDATION AND OTHER STUDIES

Charles Brenner[1] and Jeffrey W. Morris[2]

from Proceedings for
The International Symposium on Human Identification
1989, Editor (none), Copyright 1990 Promega Corporation, pp 21-53.

Table of Contents

[1] 2486 Hilgard Ave., Berkeley, California 94709

[2] Long Beach Genetics, 2384 Pacifica Place, Rancho Domingues, CA 90220

List of Figures

List of Tables

Section I. Introduction

Validation of calculations of PI in single locus hypervariable DNA probe systems require answers to the following:

1. Do computed values of PI properly assess the genetic evidence? (Section V)

2. Are systems independent? (Can PI's be multiplied together)? (Section VII)

The following question, while not essential for computation of PI, is of interest and importance for validation studies.

3. Are populations in Hardy-Weinberg equilibrium? (Section VI)

We begin by discussing computation of paternity index (section II), data collection (section III) and evaluation of experimental parameters (section IV).

Section II. Computation of PI

Consider a typical paternity case pattern wherein the mother and child are both heterozygous and share a single band, and an alleged father, also heterozygous, appears by coelectrophoresis to share the child's remaining band.

We will define the paternity index in this case to be PI_0. The paternity index in other situations, such as the homozygosity of some of the parties, or a motherless case, can usually easily be formulated in terms of PI_0.

$$PI_0 = 1 / 2 \cdot \Pr\{\text{random match}\} \qquad \text{(II.1)}$$

and the problem of assigning a paternity index therefore boils down to the question of computing Pr{random match} — the chance that a random allele would match (by coelectrophoresis) the present one.

The simplest approach, and a common one, is to reason as follows. Let

δ = coelectrophoresis resolution threshold;

y = molecular weight of the shared band Y;

f(x) = the probability distribution of the probe.

Then a random allele of weight x will co-migrate with Y provided that

$$y-\delta < x < y+\delta,$$

so define

$$S(y) = \int_{y-\delta}^{y+\delta} f(x)\,dx,$$

which is illustrated by the shaded area, of width 2δ, in the figure at right, and to a first approximation we have

$$\Pr\{\text{random match}\} = S(y). \qquad \text{(II.2)}$$

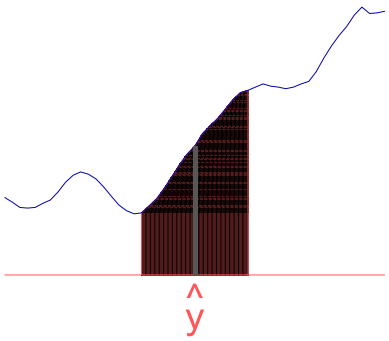

Figure 1 simple method

The simple method is too simple. It assumes that y is the size of the allele. In practice, y is only a measurement. The nor-

mal curve at the left is suggestive of the uncertainty of the measurement. If we make the assumption that y is normally distributed with

σ = standard deviation, and

N(t, σ) = normal distribution

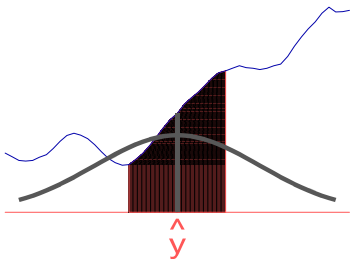

Figure 2 measurement error

then as suggested by Figure 2 we consider a weighted average of the shaded areas as the center of the area is wiggled back and forth, weighted according to the normal curve. Thus we are led to the formula

$$\Pr\{\text{random match}\} = \int_{x=0}^{\infty} N(x-y,\sigma)S(x)dx. \qquad \text{(II.3)}$$

However, the reasoning is still not complete. The normal distribution represents the probability that an allele whose true size is y would be measured somewhere else, whereas the experimental situation is the converse: y is the measured position, it is the true size that may be elsewhere.

Refer to the figure at the right. An allele of size y is equally likely to be mis-measured as L or as R. But a measurement at Y is much more likely to be a mis-measurement of R than of L, simply because, and precisely in the proportion that, R is the more common allele. By an application of Bayes' Theorem, then, the correct weighting function, illustrated here, is proportional to N(t−x, σ)f(x), and so we have

$$\Pr\{\text{random match}\} = \int_{x=0}^{\infty} (1/k)N(x-y,\sigma)f(x)S(x)dx, \qquad \text{(II.4)}$$

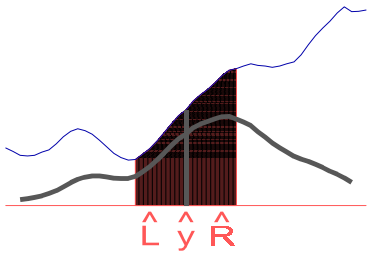

Figure 3 refinement

where k is the normalizing factor

$$k = \int_{x=0}^{\infty} N(x-y,\sigma)f(x)dx.$$

The error in using (II.3) instead of (II.4) is always in the direction of calculating too large an index of paternity. The error is particularly significant where large PI' s are concerned, or where the measurement is in a sparse region of f(x) but quite near a dense region.

Section III. Collection of Data

In this chapter we discuss several issues pertaining to reliability of molecular weight determination.

A. Reading Membranes

Our membranes include a ladder of standards every tenth lane or so. This amounts to three or four ladders.

We use a digitizing tablet to enter data from DNA sizing membranes into a computer. This method is low-tech, but simple and reliable. The membrane is taped to the surface of the tablet, the appropriate computer program is invoked, and then the operator digitizes the DNA bands one at a time using a hand-held object called a "cursor" which contains cross-hairs and buttons.

Intraassay variation — that is, the difference between two different digitizations of the same

membrane — averages 17 (±14) base pairs at 6.7Kb, and 9 (±7) at 3.3Kb. These differences amount to 0.29%·MW and 0.26%·MW.

As a first step to minimizing errors, a schematic image of the membrane is built on the computer display as the membrane is being digitized. This was the first of several reasons that led us to include color graphics capability as part of the computer requirements. The screen image shows colored lanes corresponding to the lanes of the membrane. As each point is digitized by virtue of the operator pressing a cursor button, a blip appears at a corresponding position on the screen. The lanes containing standards — every tenth lane or so — are in a brightly contrasting color. These lanes usually have 10 bands, as opposed to the one to four (in case of co-electrophoresis) bands in data lanes. A glance at the screen will quickly tell whether all bands are present and in the correct relative positions.

Digitization is very quick —2-3 minutes for a 30 lane membrane.

B. Determining DNA Fragment Lengths

Converting digitized positions to molecular weight is a matter of interpolation. Since fragment mobility will always vary from membrane to membrane, the molecular weight of the DNA fragments in each data band can only be assigned by comparing the position of the band to the positions of standard markers (bands of known molecular weight) on the same membrane.

1. Interpolation in the migration direction

Imagine for the moment that there is no variation from lane to lane. Then interpolation consists of

a. Choosing a curve L(m), where m is migration distance, that gives the molecular weight L of the standard bands when applied to their corresponding m.

b. Evaluating L(m) for the migration distances m of the data.

In Elder & Southern [Measurement of DNA Length by Gel Electrophoresis, Analytical Biochem 128:227-231, 1983] a number of such curves are considered. We experimented with them, used the method of cubic splines for a while, then reevaluated and now use a hyperbola.

Nominal size	Interpolation method: hyperbola	Interpolation method: spline
3.057 kb	−0.024 ←	−0.011
4.076	−0.042	−0.046
5.095	−0.023	−0.067 ←
6.114	−0.037	−0.023
7.113	−0.052	−0.054
8.152	−0.064 ←	−0.033
9.171	−0.020	−0.028
10.190	−0.007	−0.169 ←←
11.209	−0.018	−0.242 ←←
12.228	−0.043	+0.167 ←←

Table I Comparison of errors (←) in interpolation methods

The method of cubic splines is a standard technique in curve fitting but has nothing particular to do with DNA. It is an adequate method of interpolating, as judged by demoting one of the standard markers to a data band, and seeing if we can correctly guess it's length by using the cubic through its neighbors. And it may be seen as a virtue that the spline passes exactly through each of the standard markers, and so "predicts" their lengths exactly.

On the other hand, when on occasion it's necessary to extrapolate — because a data band is too close to the edge of the membrane to have 2 standards on either side —the cubic method may fail miserably, even ludicrously. It these circumstances sometimes it gives molecular weights that are negative, or 20Kb too high.

The hyperbola method, being a single simple formula with only three parameters L_0, m_0, and c:

$$L(m) = L_0 + c \,/\, (m - m_0)$$

doesn' t even necessarily pass through the standard markers. However:

- According to Elder & Southern, such a formula is suggested by a simple model of migration. And since it somewhat mirrors reality, it is less fragile. When extrapolation is necessary the results are still poor, but they' re not awful.

- Passing a curve exactly through all the standard markers may be a poor idea. If some of them are out of the way, so you have to meander to get to them, perhaps they were digitized, or they migrated, inaccurately. To put it another way, rather than judging a data band by comparing it only with nearby standards, farther away standards should have some influence also, though only a modest one.

- It is quicker and more convenient to compute.

Lately we have been running a standard ladder from BRL as an additional control. Interpreting our readings of this ladder under both methods of interpolation (**Table I**) gives an additional evaluation of method.

2. Interpolation in the transverse direction

Up to now we' ve pretended that the membrane is one-dimensional. In fact, rates of migration might differ from lane to lane, or the membrane might be warped (e.g. through handling). Thus, if lanes 1 and 10 are standard lanes, migration distances in lane 5 should probably be interpreted by applying interpolation in the "lane" dimension as well. Our algorithm is to fit a polynomial through all the standard lanes, so that if, say, 30 is also a standard lane, it too would have some influence, albeit a small one, on the interpretation of lane 5. This method is recommended by the following arguments:

a. Suppose the following sort of deformation occurs: Lie the gel flat on the table. Suppose lane 30 is slid upwards while lanes 1 and 20 are held in position. To the extent that the gel is springy, lane 5 will bow downward; the lower and upper edges of the gel bend into a rough parabola.

b. Suppose the gel is not springy. Hold lanes 20 and 30 fast, and deform by sliding lane 1 up. Perhaps the soft gel will absorb all the deformation at the few leftmost lanes, lane 5 not moving at all. Then in interpreting lane 5 it would be a mistake to take too much account of lane 1, while ignoring lane 30 which is in fact more representative.

c. By visual examination some membranes appear bowed —judging both from the relative positions of the standard lanes with respect to one another, and from the way the bands slant to make a happy or sad face, it appears that the distortion is smoothly distributed across the width of the membrane.

However, experimental data is not yet available to test these ideas.

Section IV. Evaluation of Parameters

A fundamental problem for computation of PI in single locus hypervariable DNA probe systems is selection of experimental parameters. For coelectrophoresis experiments, the key parameter is delta (δ), the discrimination power of the experimental technique. Of secondary importance is the standard deviation of measurement variation, which is of primary importance for calculation of PI based only on allele measurements [Gjertson, et al Am J. Hum. Genet. 43 (1988) 860-869]. There are two separate standard deviations: Interassay (between run), appropriate for accounting for measurement error in searching the data base for computation of matching frequency, and intraassay (within run), appropriate for assessing the significance of measured paternal and tested man alleles. Repeat experiments will yield

these parameters.

A. Determination of σ

σ is defined as the interassay standard error of measurement. It is not as important as δ if paternity index is evaluated by the co-electrophoresis method (page 1). Still, it plays a role in the formula, and is pivotal if one uses Gjertson' s method.

We were interested in answering two questions:

Mean allele size (kb)	relative range	σ (%)
2.789	.104	0.86
3.384	.116	0.89
5.106	.11	0.72
8.554	.152	0.58

Table II Relative error of measurement, based on 21 runs of a control subject

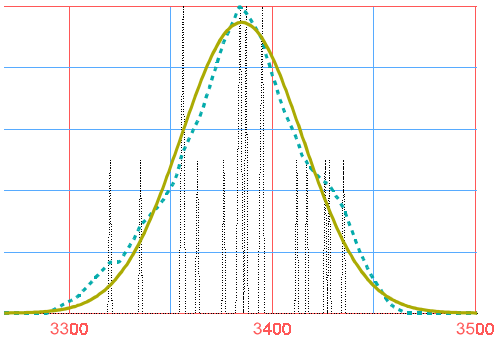

Figure 4 Interassay variation among 21 runs of a control subject. Raw data, smoothed data, and a normal curve for comparison are pictured.

1. How much is σ? **Table II** gives our data on our genomic control. These values of σ are higher than observed with the BRL ladder (typically 0.3 - 0.5%·MW) and with repeat runs on case material (typically 0.6%·MW). Part of the discrepancy may be due to our standard placement of the genomic control (lane #30), where it may be subject to edge effects.

2. Are errors really normally distributed? shows the various results of running a control subject many times. The raw data is indicated by the dotted lines. The bold dashed line is a computer-smoothed version of the same data, and it does seem similar to the normal curve included for comparison.

B. Determination of δ

1. Lower limits for δ can be obtained in two ways:

a. δ cannot be less than the width of a band. As the bands in a "near heterozygote" move closer together, they will begin to merge when the difference in their size (measured center-to-center) is equal to the band width. One cannot expect reliably to distinguish two such close bands. One cannot measure width of bands with great precision, but measurement of a few dozen randomly selected band widths yields a good approximation for the lower limit of δ, which, for our experimental conditions and pS194 and pL336, is about 1%·MW.

b. δ cannot be smaller than the distance between the measured bands of the closest heterozygote. Figure 6 is a scattergram showing the relative distance between band measurements for "near heterozygotes" (plotted as a function of the smaller allele), and distribution of homozygotes.

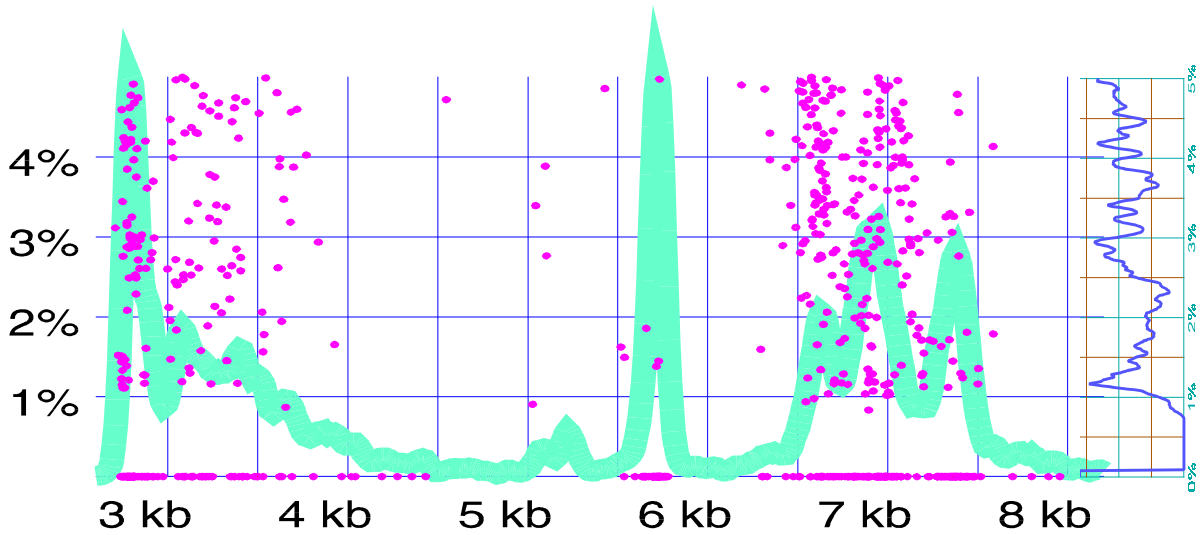

Figure 6 Dots indicate homozygotes and near homozygotes. The combined allele distribution for all populations and probes is graphed as background. The sideways strip on the right indicates dot density as a function of vertical scale.

From the computer files, there are few heterozygous individuals with allelic difference < 1%·MW, and many < 2%·MW. These observations suggest that δ is between 1%·MW and 2%·MW.

Note that, in practice, we do not distinguish heterozygotes with bands closer than 1%·MW, although repeat runs with greater resolution occasionally will resolve a "fat band" as heterozygous (see above). Thus, the limiting factor in resolution of heterozygotes appears to be delta, rather than nonexistence of heterozygotes with alleles closer than 1%·MW. Again, the lower limit of δ for our experimental conditions is about 1%·MW.

2. Upper limits for δ may be obtained in two ways

In this section we mention two methods for estimating δ from experimental data by assuming Hardy-Weinberg equilibrium. These methods are useful not because Hardy-Weinberg is a plausible assumption (which it may or may not be — see Section VI), but for a slightly better reason. Namely, we feel comfortable in assuming that any deviation from equilibrium will be in the direction of excess homozygosity. In that case, the true value of δ will be if anything lower than estimated by these methods.

a. Comparing observed with calculated heterozygosity (h) affords a method for estimating δ.

The calculation is to choose pairs of alleles at random from the data base, calculating the frequency with which pairs of alleles differ by more than δ, and choosing δ so that calculated heterozygosity agrees with observed heterozygosity. This calculation assumes Hardy-Weinberg equilibrium. When applied to our Hispanic data bases:

probe							
pS194	δ(%·MW)	1.0%	1.2%	1.4%	1.6%	1.8%	2.0%
	h(calc)	0.887	0.868	0.851	0.835	0.82	0.805
	observed h = 539/635 = .848						
pL336	δ(%·MW)	1.0%	1.2%	1.4%	1.6%	1.8%	2.0%
	h(calc)	0.934	0.922	0.911	0.901	0.891	0.882
	observed h = 535/586 = .913						

Values of δ determined by this method are summarized in . Calculated heterozygosity matches observed heterozogosity when δ ≈ 1.4%·MW for each probe, comfortably larger than the minimum estimates (1.0%·MW) obtained above. When this experiment was repeated with smaller Caucasian and Black data bases, anomalous results were obtained — observed heterozygosity agreed with calculated heterozygosity for Caucasians when δ ≈ 1.9%·MW, and for Blacks when δ ≈ 2.4%·MW. Observed band widths and close heterozygotes were not different among races (data not shown), so the source of the anomalous results suggested either limitations of equation V.B.2, sampling variation, or deviation from Hardy-Weinberg equilibrium. Most of the data used for the above were obtained from blots which are limited to deletion of alleles less than 12.4 kb (pS194) and 8.8 kb (pL336). Less than 1% of alleles for Caucasians and Blacks are expected to be missed by this technique [D. Dykes, personal communication], which is insufficient to account for the observed differences.

Probe	Race	heterozygosity	δ
pS194			
	Caucasian	82.7%	1.893%
	Black	84.8	2.366
	Black+Caucasian	83.5	2.297
	Hispanic	84.9	1.423
pL336			
	Caucasian	87.7	1.88
	Black	87.6	2.46
	Black+Caucasian	87.6	1.82
	Hispanic	91.3	1.366

Table III Values of δ that explain the observed heterozygosity, by population

b. Comparing observed with calculated rates of exclusion (A) gives another method of estimating δ.

> Frequently, mean exclusion probability is determined from case material by shifting tested men by one or more positions, thus matching mother-child pairs with non-fathers. Men for whom both alleles differ from the paternal allele(s) by more than δ are considered to be excluded. This method yields a calculated, rather than experimental, exclusion probability, as it depends on the value chosen for δ. Such methods have yielded illogical results —calculated A larger than observed h. Such results are equivalent to a claim that ones chance to win the lottery is greater with a single ticket than with two. If the limitations for equation V.B.2 and Hardy-Weinberg equilibrium are satisfied, the inevitable conclusion for such anomalous results is that the chosen value for δ is too small. If such a δ is used for calculations of PI, the resulting PI' s will be too

large. However, if experimental A is available, agreement with calculated A provides validation of the chosen value of δ. In the absence of several hundred exclusion cases, A may be calculated from equation V.B.2, and also calculated from case material by displacing tested men from mother child pairs, and choosing δ so that the two estimates of A are in agreement. This method is subject to the limitations (above) for equation V.B.2.

3. A model for δ

How does δ vary with varying molecular weight?

Particularly in view of the difficulty of determining δ through measurement, it seems worthwhile to consider what relationship should be expected theoretically. In quoting δ as a percent of molecular weight we are implicitly assuming that the same percentage applies at various weights. Is this a reasonable assumption?

We assume that discrimination ability is limited by band thickness. Let

m = migration distance (cm),

L = molecular weight (kb),

β = thickness of a band (cm),

d = discrimination threshold (kb), and

$\delta = 100 \cdot d/L$ = % discrimination threshold.

Then if we know how L and β behave as functions of m, it should be possible to compute how δ behaves.

We know that L and m are very nearly related by a hyperbolic relation of the form

$$L(m) - L_0 = c / (m - m_0).$$

$L_0 = -0.88$ kb, $m_0 = -0.965$ cm, and $c = 64$ kb-cm are typical values for the parameters, which depend on running conditions. We have found that such an equation gives an excellent fit for molecular weights from 3-12kb. As for the behavior of β with respect to m, many assumptions are possible but we' ll consider only the possibilities

$$\beta(m) = \beta_0 + k \cdot m^{\alpha}$$

for $\beta_0=0.05$cm and various choices of the parameter α. $\alpha=0$ is the naive assumption that β is a constant; $\alpha=1$ represents the pessimistic assumption that a band grows twice as thick if it migrates twice as far; and $\alpha=½$ corresponds to the notion that band diffusion is some sort of random walk process.

Then

$$\delta_{\alpha}(L) = (100/cL)(L-L_0)^2(\beta_0+k \cdot m^{\alpha}),$$

(Appendix A) and k can be determined by assuming $\delta=1.4\% \cdot MW$ at L=7 kb. The result is Figure 7. Since at least the dotted line, corresponding to $\alpha=1$, is close to constant, this analysis lends plausibility to the practice of taking δ to be a constant percentage amount.

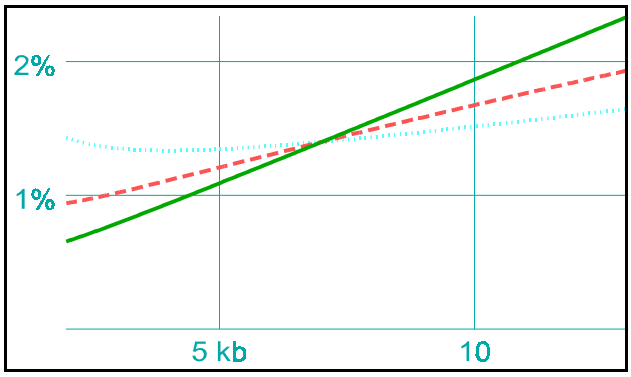

Figure 7 δ as function of MW assuming
———— constant band size
---- band grows like $\sqrt{m}$
···· band grows linearly with m

Section V. Validation of PI

Suppose that PI's have been computed for a series of paternity cases. In this section we discuss methods for testing whether the collection of values, as a whole, are credible.

Some tests are given that depend on checking some formulas relating exclusion rates, expected heterozygosity rates, and simple functions of PI.

We begin with some identities relating paternity index to probability of exclusion (subsection V.A). Then we discuss means to estimate probability of exclusion (subsection V.B). We use these results (subsection V.C) to compare the results of calculating average paternity index in different ways,

which provides a check on our parameters and our methods.

Finally (subsection V.D) we present a forensic application similar in spirit to V.B.

A. Average Values of PI —relationships between mean exclusion probability (A) and paternity index (PI)

The overbar (⁻) notation indicates an average value taken over a large number of cases.

For fathers:	$\overline{1/PI} = 1-A$	(V.A.1)
For non-excluded non-fathers:	$\overline{PI} = 1/(1-A)$	(V.A.2)
W (50% prior probability, non-exclusion)	$= 1/(2-A)$	(V.A.3)

The above relationships are independent of assumptions of Hardy-Weinberg equilibrium. Equations (V.A.1) and (V.A.2) may be found in [Nijenhuis, LE. pp. 103-114, Inclusion Probabilities in Parentage Testing. Ed R. Walker AABB 1983 Arlington, VA]. Equation (V.A.3) may be found in [Morris, JW. p 267-276, of the same volume].

In order to make use of equations (V.A.1), (V.A.2), and (V.A.3), we shall require values for the mean exclusion probability, A. It can be determined experimentally, by coelectrophoresis of specimens from child and non-father. This, obviously, is the optimal method. However, estimation of mean exclusion probability by this method requires several hundred cases of non-paternity. We are aware of only one such published study (see **Table XII**, Appendix B). In this case agreement of calculated A from h (equation V.B.2 below) with experimental findings is excellent.

B. The Heterozygosity-Exclusion Analogy — relationship between heterozygosity (h) and mean exclusion probability (A)

The determination that an individual is homozygous or heterozygous at a given locus may be regarded as a coelectrophoresis experiment. The maternal and paternal alleles are coelectrophoresed in the same lane and a visual determination is made as to whether or not two alleles can be distinguished. For a given population, heterozygosity (h) may be regarded as the mean exclusion probability for the hypothesis that maternal and paternal alleles are identical. This biostatistic might be useful if one is investigating the possibility that the phenotypes of a given individual arose from incest.

It is standard practice to determine exclusion/non-exclusion by coelectrophoresis in the same lane of specimens from child and tested man. If the alleles of the tested man are clearly distinguished from the paternal allele(s), an exclusion is recorded. This procedure is highly analogous to determination of homozygosity/heterozygosity - the difference is that for heterozygosity/homozygosity the maternal allele has

h	A calculated by (V.B.1)	(V.B.2)
.80	.640	.599
.85	.723	.695
.90	.810	.797
.95	.903	.898

Table IV A predictions

one chance to match the paternal allele, while for exclusion/non-exclusion, there are two chances for a match between the tested man and paternal allele. It is perhaps not surprising that the mean exclusion probability (A) can be approximated by a simple function of heterozygosity (h):

$$A \approx h^2. \qquad \text{(V.B.1)}$$

A more accurate approximation is:

$$A \approx h^2(1-2hH^2) \qquad \text{(V.B.2)}$$

where

$$H = \text{homozygosity}$$
$$= 1-h.$$

Derivations of equations (V.B.1) and (V.B.2) are given in Appendix B. Predicted values for A as a function of h are shown in **Table IV**.

Limitations of equations (V.B.1) & (V.B.2)

a. It is essential that criteria for heterozygosity/homozygosity be identical to those used for exclusion/non-exclusion. In our laboratory heterozygosity depends on identifying two discrete bands in the phenotype of an individual - the same criteria (alleles of the tested man must be discretely different from the paternal allele(s)) is used for exclusion. Thus, "fat bands" are recorded as homozygous in the data base, and as non-exclusions in case material (for purposes of analysis) although we have found in both experimental situations that retesting under conditions of increased electrophoretic resolution will sometimes permit splitting of "fat bands" into two discrete alleles.

b. Experimental conditions for heterozygosity/homozygosity and exclusion/non-exclusion must be identical. In practice, this is easy to achieve, as data bases are often constructed from unrelated adults in disputed paternity cases, and thus heterozygosity/homozygosity is determined in the same gel (often in adjacent lanes) as exclusion/non-exclusion.

c. Equations (V.B.1) and (V.B.2) do not require Hardy-Weinberg equilibrium. However, if the population making up the data base for computation of homozygosity/heterozygosity is heterogeneous, then equations (V.B.1) and (V.B.2) assume that the offspring for whom A is computed or observed is heterogeneous in the same way. In practice, one must observe the same homozygosity rate for offspring in paternity matters as for the data base.

d. Equations (V.B.1) and (V.B.2) assume that silent alleles ("blanks") occur with negligible frequency. While this appears to be the case with most single locus hypervariable DNA probe systems, failure to identify alleles because of experimental conditions will cause significant distortion. If 2% of bands are missed because they are too large or too small for the conditions of electrophoresis or blotting, observed heterozygosity will be low by about four percentage points.

C. Some Consistency Checks

However PI's are computed, relationships (V.A.1), (V.A.2), and (V.A.3) must hold. Optimally, a large series of case material should be used, so that A and PI distribution may be determined experimentally. Our case material is limited, so we have made use of simulation methods.

A representative collection of PI's for fathers were calculated from the data bases in the following fashion. Each observed allele was considered in turn to be a paternal allele, and PI_0 was computed, taking $\delta = 1.4\% \cdot MW$ and $\sigma = 0.6\% \cdot MW$, as:

$$PI_0 = 1 / 2 \cdot \Pr\{\text{random match}\} \qquad \text{(II.1)}$$

This equation, introduced on page 1, assumes that a single paternal allele is identified and the non-excluded alleged father is heterozygous. Since we are assuming that the alleged father is the biological father, PI_0 must be corrected for possible homozygosity of the father and for the possibility of two possible paternal alleles. We do this by the following approximation, whose derivation is given in Appendix C:

$$PI_f \approx PI_0 \cdot [1 + h\,(1-h)\,(2-h)/2] \qquad \text{(V.C.1)}$$

Apply equation V.A.1 to the collection of simulated PI_f' s, and we must have

$$\text{1}/\overline{PI}_f = 1 - A. \qquad \text{(V.C.2)}$$

Equation V.C.2 can be used to predict A, given the allele distribution and assuming a value for δ.

The reciprocal of the mean reciprocal PI for fathers is in some sense a typical PI for fathers, and so we define

$$\text{"typical PI for fathers"} = \hat{PI}_f = 1 / \text{1}/\overline{PI}_f.$$

Among other things, $\hat{PI}_f$ can be regarded as a measure of the power of a test. In this respect it is worth considering other statistics that one might consider in this role.

(a) An arithmetic mean, $\overline{PI}$ or $\overline{PI}_f$, is not a good measure of typical performance because it is too much influenced by occasional large values.

(b) The defect of averaging PI' s can be avoided by averaging in the domain of probabilities rather than likelihood ratios. That is, for each PI compute a probability W, find $\overline{W}$, and then backtrack to the corresponding PI. This method has the theoretical blemish of depending on a choice of prior probability, but experiment shows that the choice of prior doesn' t matter very much.

(c) The geometric mean,

$$(PI_1 \cdot PI_2 \cdot \ldots \cdot PI_N)^{1/N}$$

is also a sensible statistic in that likelihood ratios are meant to be multiplied. It also has the appeal of being the only candidate that naturally combines systems: The geometric mean PI for a combination of independent systems is the product of the geometric means per system.

(d) The median PI.

All methods except (a) give about the same number.

From equation V.A.2, a corresponding statistic for non-fathers is

$$\overline{PI}\text{ (non-paternity, non-exclusion)} = 1/(1-A)$$

This will be called "mean PI for random non-excluded men" = $\overline{PI}_r$.

Note that

$$\hat{PI}_f = \overline{PI}_r.$$

Since $A \approx h^2$ (equation V.B.1),

$$\overline{PI}_r \approx 1/(1-h^2)$$
$$= 1/(1+h)(1-h).$$

Since $1+h \approx 2$ and $1-h = H$,

$$\hat{PI}_f = \overline{PI}_r \approx 1/2H. \qquad \text{(V.C.3)}$$

This equation yields a "quick and dirty" estimate of typical PI for fathers from observed or calculated heterozygosity. The similarity of equations V.C.3 and II.1 should be noted.

Table V summarizes results obtained with our data bases.

	Calculated using δ (=1.4%·MW)				Calc from obs h		
	calc h	1/PI_f (V.C.1)	calc A (V.C.2)	$\hat{PI}_f$	obs h	calc A (V.B.2)	PI_r
pS194							
Hispanic (N=638)	.851	.276	.725	3.63	.848	.691	3.24
Caucasian (N=472)	.858	.262	.738	3.81	.822	.641	2.79
Black (N=304)	.898	.192	.808	5.21	.842	.679	3.16
pL336							
Hispanic (N=586)	.911	.167	.833	6.00	.913	.822	5.62
Caucasian (N=221)	.903	.180	.820	5.56	.873	.741	3.86
Black (N=226)	.925	.143	.857	7.00	.876	.747	3.95

Table V Validation through comparing methods of calculation

For both probes, the Hispanic population shows good agreement between estimates of A calculated from $\hat{PI}_f$ (equation V.C.2) and calculated from observed heterozygosity (equation V.B.2).

The agreement is less good for Caucasian and Black populations because observed heterozygosity is significantly less than calculated for δ = 1.4%·MW. Our limited case material yields the following results ("fathers" are men non-excluded in conventional and DNA probe systems) for δ = 1.4%·MW, σ = 0.6%·MW:

	Calculated from case material				Calc from observed h
	1/PI_f (V.C.1)	A (V.C.2)	$\hat{PI}_f$	median PI	calc A (V.B.2)
PS194					
Hispanic (N=26)	.339	.661	2.95	2.98	.691
Caucasian (N=19)	.309	.691	3.24	3.16	.641
pL336					
Hispanic (N=26)	.190	.810	5.26	6.24	.822
Caucasian (N=19)	.156	.844	6.41	9.32	.741

Table VI Validation by comparing observed and calculated values

Our analysis yields the conclusion that for our experimental conditions typical and median PI' s should be about 3 for pS194, and about 4-6 for pL336. These values are substantially less than those claimed for single locus probes of equivalent polymorphism [Balazs, et al Am. J. Hum Genet. 44 (1989) 182-190]. See Endean, D., these Proceedings, for a further evaluation of these probes.

Additional strategies for validation of PI

2. For case material consisting of roughly equal numbers of fathers and non-fathers, equation (V.A.3) can be used. For cases in which non-exclusion is observed for a given probe, W (50% prior) for that system can be computed and averaged, yielding estimates for mean exclusion probability for comparison to observed or calculated A.

3. For unselected case material tested in an extensive test battery, realistic prior probability can be closely estimated from exclusion rate. For each probe system, W (realistic prior) can be calculated for each case. Sum of W's (including exclusion cases) equal the expected number of fathers (realistic prior probability · sample size).

4. For very large collections of unselected case material with extensive testing, PI's in each probe system may be grouped by magnitude. By definition of PI, the frequency of PI's among fathers with **a** < PI < **b** should be more than **a** but less than **b** times greater than the frequency among non-fathers. See Morris, J.W. [Transfusion 29 (1989) p281] for application to conventional systems.

Demonstration of Additional Strategy V.C.2

Suppose we have 100 cases and know (through thorough testing with conventional methods) that 70 of them represent fathers.

Test the 100 cases with the new test. Compute the paternity indices using the method that we are hoping to validate. Convert each paternity index to W, probability of paternity, using the laboratory's experienced prior probability of 70%.

Thus for example a man with paternity index of 2 has a posterior probability of paternity of

$$W = \frac{0.7\times 2}{0.7\times 2 + 0.3}$$

$$= 0.82,$$

which means that 100 such men supposedly include 82 fathers, or that this one man is (probabilistically) 0.82 fathers.

Therefore, if we add up all the W's computed this way, we will be computing the total expected number of fathers. If that total is not close to 70, the paternity indices must be wrong.

However, this method is only useful when the test is not too powerful, as in the example. Suppose to the contrary that the new test shows the following results:

- 29 of the 30 non-fathers are also excluded by the new test.
- The paternity index under the new test for the other non-father and for the 70 fathers is computed to be 100.

Then we get

$$\sum_{\substack{29 \text{ excluded men} \\ 71 \text{ men with } PI = 100}} W = 29\times 0 + 71\times 0.996$$

$$= 70.7$$

whereas if another proposed procedure gives an index of only 10 instead of 100, we have

$$\sum_{\substack{29\ excluded\ men \\ 71\ men\ with\ PI=10}} W = 29\times 0 + 71\times 0.959$$

$$= 68.1,$$

and the difference is probably not significant with only 100 cases.

D. A Forensic Application—estimation of mean probability of phenotype match (R) from homozygosity (h)

If phenotype matching is determined by coelectrophoresis, the mean probability of matching is:

$$R \approx (1-h)^2(1+h)$$
$$\approx 2H^2.$$

Derivation of these approximations, which assume Hardy-Weinberg equilibrium, can be found in Appendix D. If coelectrophoresis is not performed, it seems to us that the method of Gjertson is easily adaptable to phenotype matching, and can be modified, as suggested by Gjertson, to account for "non-random" measurement error. [see Evett, I, these proceedings].

Section VI. Hardy-Weinberg Equilibrium

A. Quest for a Test

A population is defined to be in Hardy-Weinberg equilibrium with respect to a given polymorphic genetic marker provided that the various alleles assort randomly. The condition of random assortment is commonly stated as a formula:

> If two alleles A and B occur with frequencies a and b in the population, then the phenotype AA should occur with frequency a^2 and the phenotype AB should occur with frequency 2ab.

That the population should be in Hardy-Weinberg equilibrium with respect to the tested systems is an invariably stated assumption for the evaluation of blood-stain evidence in paternity and in forensics. Therefore attention is beginning to be focused on the question of equilibrium with respect to DNA probes.

For example, Lander [pp 30-31, Expert's Report in People v. Castro, 1989] expressed doubt that populations and probes used by Lifecodes in the Castro case are in Hardy-Weinberg equilibrium, basing his conclusion partly on Wahlund's test applied to Lifecodes reported data, and partly, concerning rare alleles, by analogy with conditions such as Tay-Sachs. (Populations are not expected to by in Hardy-Weinberg equilibrium with respect to rare alleles.)

We shall return to consideration of Wahlund's test.

	Probe		
	pS194	pL336	both
Caucasian	468	219	191
Hispanic	635	586	517
Black	297	225	209

Table VII numbers of people typed in each DNA probe

Up to this point, little data seems to have presented to justify the assumption of Hardy-Weinberg equilibrium for any DNA probes. Having on our computers a considerable number of allele measurements we decided to make a quick test of equilibrium. It turned out to be harder than expected.

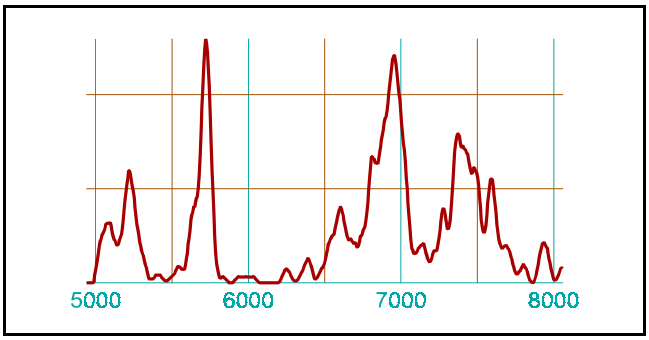

Figure 8 Distribution of pS194 Black alleles

The obvious problem is that unlike in traditional systems, the alleles are not classified, but only measured. As everyone well knows, the measurements are not precise. The imprecision, with a standard deviation on the order of 0.6% MW, is the source of considerable consternation for this kind of analysis. For it may well be —in fact probably is — the case that the underlying data consists of discrete alleles. But the alleles do not remain discrete when we measure them. Assuming that many of the peaks and shoulders in Figure 8 represent alleles, it is clear that measurements of nearby alleles overlap to a considerable extent. How, then, can we determine allele frequencies? How can we test for Hardy-Weinberg equilibrium? Can the equations of equilibrium be recast to be a condition on the measurements, which we have, rather than on the state of nature itself, which eludes accurate description?

Many researchers deal with DNA probe alleles by "binning" — such as rounding measurements off to the nearest 0.1kb, or classifying alleles according to the standards rungs between which they fall. One disadvantage — among several — with any binning method this: If there are bins, inevitably there

must be boundaries. Those alleles near boundaries are subject to misclassification, and if the bins are large, the misclassification will be by a large amount, since a mis-binned allele is in effect categorized with other alleles whose typical size is the middle of the (wrong) bin.

1. Hardy-Weinberg test by binning

a. Nonetheless, for lack of anything better, in an attempt to test Hardy-Weinberg equilibrium we decided to try binning the data into allelic classes. As Max Baur has pointed out, the amount of misclassification is minimized by placing the boundaries at where allele measurements are sparse — that is, at natural troughs in the allele distribution spectrum as shown in Figure 9. By choosing such boundaries we were able to partition the measurements into three to five more or less natural allelic classes. These give rise to six to fifteen phenotypic classes, for each of which we make a prediction, by assuming Hardy-Weinberg equilibrium, and compare against the population data. At first the results seemed quite hopeful. According to the χ^2 statistic, none of the six combinations of race and probe are in disequilibrium.

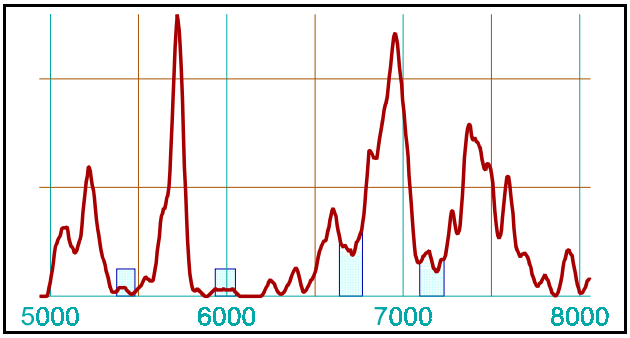

Figure 9 Bins. Hatching indicates regions prone to misclassification, ±δ from the boundaries. (pS194 Black alleles)

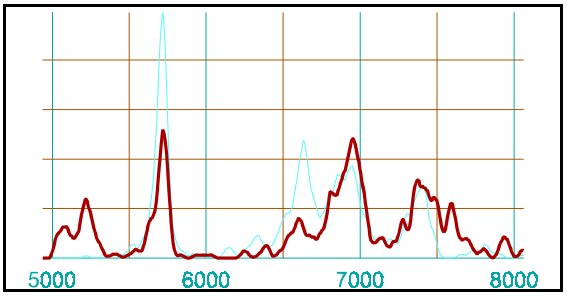

Figure 10 Caucasian and Black pS194 allele frequency distributions

Hoping to go further, and in an effort to establish the sensitivity of the method, we next applied the same test to a mixed population of pS194 Caucasians and Blacks. It is visually apparent (Figure 10) that if each race separately is in equilibrium, then the combination must have an excess of homozygosity and cannot be in equilibrium. Therefore it was a little disappointing to find that the mixed population wasn't far from equilibrium by our test. A further experiment using judiciously chosen bin boundaries turned something up, but such a **post hoc** procedure tends to be less than convincing.

b. More bins

Perhaps using more bins would make a more sensitive test. Accordingly we raised the number to ten (at the expense of using unnatural boundaries) and promptly demonstrated disequilibrium — at 98% significance and more —for most of the mixed population combinations.

Unfortunately, the Caucasians and Blacks taken separately seemed to be just as far out of whack. On the face of it this is not astonishing, because the conclusions are really statements about collections of measurements, not about collections of people. However it does raise questions:

(1)(a) Are we stumbling over an artifact in the data?

(b) If so, how can we avoid it?

(2) Since the binned measurements are not in Hardy-Weinberg equilibrium, does this limit our ability to infer PI's from binned measurements?

c. Very many bins

Moving up to 45 equally populated bins, the χ^2's loomed enormous. Even the heretofore reliable

pS194 Hispanic data manifested incontrovertible disequilibrium.

A simple calculation shows why this is to be expected when the number of bins is large. With each bin containing 1/45 of the alleles, each of the 45 phenotypic combinations binned as homozygous are predicted (by Hardy-Weinberg) to occur with frequency $1/45^2$, for a total of 1/45 of the population predicted to be homozygous according to this binning. This happens to be seven times smaller than the rate of homozygosity actually observed, so we have artifactual excess homozygosity. In retrospect, it is clear that even ten bins was guaranteed to be too many.

Another way to understand the difficulty, and to see how it already occurs to a degree even with a small number of boundaries, is to realize that misclassification of alleles is biased in favor of misclassifying alleles that participate in an apparently homozygous phenotype. The reason is that the standard error of measurement, σ, is only about 0.6%·MW, whereas δ, the minimum difference reliably distinguishable by coelectrophoresis, is at least 1%·MW. Suppose two alleles' actual physical size is such that they share one of the hatched strips of Figure 9 and are within δ of one another. If they come from different people, then they may by chance be measured as being in the same bin, or in different bins. But if they come from one person, then we are bound to classify them both into the same bin.

2. Matching phenotypes

a. With binning schemes failing on every front, we next devised an experiment that doesn't depend on trying to force the DNA probe data into the traditional mold of allelic categories.

We have 635 Hispanic pS194 phenotypes, which are presumably representative of the whole population. We imagine coelectrophoresing each of them against every other, and for each phenotype we associate an "observed" score which is the number of phenotypes it would match. Estimating these scores depends on δ.

Now we take the 1270 alleles and match every pair of them, thus constructing a hypothetical population of 1270^2 phenotypes, which is in Hardy-Weinberg equilibrium since by construction every allele assorts equally with every other allele. For each of the 635 phenotypes, we associate a "Hardy-Weinberg score" by calculating what fraction of the hypothetical phenotypes it would appear to match. This too depends on δ.

Finally, we do a χ^2 check to see if the Hardy-Weinberg score predicts the observed score.

In brief the predictions were not even believable.

The first problems to come to light were pitfalls in the computations. Guided by the clue that the observed scores tended to be too high (e.g. an average of 26 matches observed, with 25 predicted), we realized that our initial rule for evaluating a match between phenotypes — both alleles within δ — exaggerates matches by overlooking several subtle points.

- Heterozygotes should never match homozygotes.
- For the same reason, a heterozygote from the real population should only match a hypothetical phenotype that has alleles at least δ apart.
- A homozygote from the real population should only match a hypothetical phenotype if all three genes are within δ.

These changes lowered the "observed" a bit more than the "predicted", but the difference was still too large. More thought turned up the idea that our way of picking data creates an ascertainment bias. In scoring "observed" matches, each phenotype is scored as matching itself; perhaps it shouldn't be. A similar modification can to be made in the method of computing the "prediction" score.

Making these further changes again depressed both scores, but successfully closed the gap

between them.

Too much. The χ^2 test now reported that for pS194 Hispanics the fit between prediction and observation was so good as to be suspicious. For some other populations the fit was hopelessly bad. In no case was it moderate. Further rethinking was in order.

Part of the computer output from the experiment was a list of the phenotypes whose scores made the bulk of the contribution to χ^2. Examining this list turned up a clue — four of the five leading contributors were so close to one another as to match substantially the same sets of phenotypes! In effect, we were counting the same data over and over. If the fit were good, it would seem to be very very good; if bad, horrid.

But how to count it correctly? In technical terms, the problem was that we interpreted χ^2 as if each "observed-predicted" pair represented a degree of freedom. If we think of each phenotype as representing a "sphere of influence" consisting of those phenotypes near enough to match it by coelectrophoresis, since the spheres overlap severely each one really represents only a fraction of a degree of freedom.

b. Test phenotypes

Not being statistician enough to evaluate and compensate for the overlap, we formulated another approach motivated simultaneously by trying to get rid of the overlap, and coming to terms with the heretofore confusing issue of ascertainment bias.

In the preceding experiment each of the 635 phenotypes plays two roles.

(i) On the one hand, it is a test phenotype which is matched with two populations:

(a) the real population consisting of the remaining 634 phenotypes, and

(b) the hypothetical Hardy-Weinberg population of 1269^2 phenotypes.

(ii) On the other hand, it is a data point in the set mentioned in (i.a).

Viewed in this light, the overlap problem occurs because the elements of set (i) are packed close together, whereas the closeness in their alternate role, set (ii), is not a problem. Set (ii) has to be our data, but there is no reason why (i) has to be the same set; (i) can be purely artificial.

Accordingly, we modified the previous experiment by choosing a different set (i) of test phenotypes according to these criteria:

(1) To overcome the overlap problem, they should be far enough from one another that the spheres of influence are disjoint.

(2) To avoid considerations of ascertainment bias, they should not be selected from among the real phenotypes.

(3) To test the hypothesis of equilibrium as thoroughly as possible, the spheres of influence of the test set should cover as much as possible of (i.a).

(1) Tiling with test phenotypes

A simple rectilinear approach was adopted. Ten test alleles, spaced 2δ apart, pretty much cover the observed allele spectrum of pS194. Considering every pair of these gives rise to 55 test phenotypes, each of which corresponds to a degree of freedom.

Expected and observed numbers of matches were calculated as described above. Our statistical consultant advised against including cells with an expectation less than 5, or perhaps less than 3, in a χ^2 computation, so such cells were lumped into a "tail" category.

The results were a mixed bag — believable for the most part, and tending to imply that

equilibrium is more or less present for the pure populations and probably absent for the mixed ones, but suspiciously sensitive to even small changes in δ.

Evidently, choices for δ which resulted in splitting of apparently homogeneous peaks resulted in factitious increases in χ^2, just as with binning. Our overall impression from these studies is that the individual populations may well be in Hardy-Weinberg equilibrium, and the mixed populations may not be, and it would help a lot to know what δ really is.

(2) Natural test phenotypes

In order to minimize the boundary misclassifications, we modified the arrangement of the test phenotypes. Instead of choosing test alleles "butted together" (2δ apart), we selected them to lie at the largest allelic peaks —but still at least 2δ apart, of course. This resulted in lots of large gaps and rather fewer test phenotypes than before —usually 10 to 21 —but they still cover most of the dense regions.

	δ (%·MW)					
Probe, race	1.4		1.75		2.2	
pS194						
Caucasian	½	(13)	0.12	(14)	0.2	(13)
Black	0.11	(10)	0.3	(12)	½	(12)
Hispanic	0.3	(9)	½	(9)	½	(12)
Cauc/Black	0.018	(19)	0.3	(2)	½	(21)
pL336						
Caucasian	½	(6)	0.4	(4)	½	(5)
Black	0.4	(2)	0.4	(5)	0.3	(7)
Hispanic	0.4	(15)	½	(16)	0.2	(11)
Cauc/Black	½	(12)	½	(15)	½	(9)

Table VIII p values, χ^2 test for Hardy-Weinberg equilibrium. (degrees of freedom=number of test phenotypes, shown in parentheses)

The results are shown in **Table VIII**.

This test turned out to be no better than natural binning. The individual populations were not significantly different from Hardy-Weinberg equilibrium over a wide range of δ's. Only the Caucasian/Black admixture for pS194 significantly deviated from Hardy-Weinberg, and only at a value of δ for which individual populations deviated from Hardy-Weinberg equilibrium by Wahlund's test (**Table IX**). Moreover, sample size for Caucasians and Hispanics in pL336 proved to be a limiting factor.

B. The Wahlund Test

δ (%·MW) =	1.6	1.65	1.7	1.75	1.8	1.85	1.9
chi^2 =	12.4	11.4	10.9	10.7	11.1	11.6	12.4
significant at p =	.05	.08	.09	.1	.09	.07	.05
(with 6 d.f.)							

Table IX Wahlund checks — various δ — homozygotes vs. expected

For a given value of δ (assumed to be the same for each combination probe and population) a check for excess homozygosity taken across all six populations (3 races and 2 probes) gives the results of **Table IX**.

The best fit is at δ = 1.75%·MW; significant deviation from Hardy-Weinberg equilibrium is not

present. Other values of δ between 1.6%·MW and 1.9%·MW fit less well, and give an indication of the robustness of the method.

In this respect it is well worth noting that Lander's computation by this test (**Table IX**), that Lifecodes' data were not in equilibrium could just has well have been interpreted to suggest that the wrong δ was being applied.

We don't have Lifecodes' data available to make computations with various δ's but on the average the expected number of homozygotes is proportional to δ. The observed number exceeds the expected by a factor of 3.65±0.01 in both cases. Lander used δ = 0.4%·MW; a more appropriate δ for these probes and experimental conditions appears to be 1.8%·MW [D. Endean, these Proceedings].

Our actual data for δ = 1.75%·MW are shown as **Table X**.

	Locus	
Proportion of Homozygotes	D2S44	D17S79
Expected by Hardy-Weinberg	4.7%	3.5%
Seen in Sample	17.1%	12.8%
ratio	3.64	3.66

Table X Lander's calculations, and ratios

A test of the power of this method is needed. To this end, we note from **Table X** that for Caucasian pS194, χ^2=0.2 with 1df, which is not significant, and for Black pS194, χ^2=2.4 with 1df, which is not statistically significant (p=.12) either. However, combining the Caucasian and Black pS194 data, we found significance at p=0.03, as shown in **Table X**.

Thus significant deviation from equilibrium was obtained with this probe and mixture. However

		homozygotes			heterozygotes			TOTALS		
		obs	exp	χ^2	obs	exp	χ^2	obs	exp	χ^2
Cauc+ Black	pS194	126	105.1	4.1	639	659.9	0.7	765	765	4.8

χ^2 = 4.8 with 1 degree of freedom is significant at p=0.03.

Table XI Wahlund check — mixture exhibits disequilibrium although individual populations do not

we could not consistently demonstrate significance with mixtures by this method. For example, even though Hispanic and Black pL336 populations are clearly different (Figure 11), the mixture showed no greater statistical deviation than the Black population alone (data not shown).

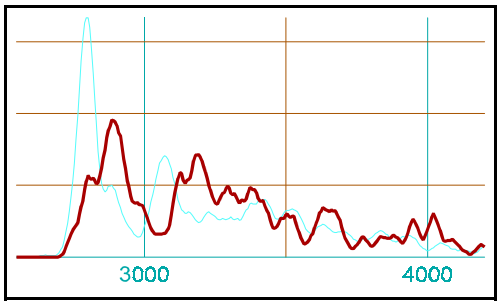

Figure 11 Black (thick line) and Hispanic pL336 allele frequencies

Section VII. Independence of Loci

Independence of loci is required to multiply PI' s or phenotype frequencies. We have used a binning technique to address the following question:

Is allele a_i in pS194 associated with allele b_k in pL336 (Is the frequency with which a_i and b_k are found in the same phenotype equal to the product of a_i occurrence and b_j occurrence)? To this end, we made use of allele bins of equal size. This allowed us to make maximal use of the data, as expectations for each cell were equal. While "edge effects" i.e. misclassification of alleles, remains an issue, factitious increased homozygosity is not present as a confounding factor. Moreover, the method is independent of δ. However, our data is limited to those individuals phenotyped in both systems (see Table I). Using this method we were able to subdivide the alleles into as many as 15 equal bins, yielding as many as 225 a_i, b_k, with the following results:

The above tests compared the number of observed phenotypes containing a_i, b_k with the expected frequency of such phenotypes, multiplying the observed frequency of phenotypes containing a_i with the observed frequency of phenotypes containing b_k. (Test based on "phenotype count"). We also tested, for the same arrays, observed frequency of phenotypes containing a_i, b_k against expectation based on allele count. This is a simultaneous test of both independence and (weakly) of Hardy-Weinberg equilibrium. Significant association was not observed (data not shown). Independence as defined in this manner implies that cumulative exclusion probabilities may be obtained by multiplication.

The level of discrimination tested by 15x15 arrays corresponds to validating cumulative exclusion probability at levels of 0.98 +, compared to cumulative mean exclusion probabilities for each race of about 0.95, calculated from individual calculated mean exclusion probabilities. We were disappointed in not observing significant deviation from independence with mixed populations. However, we would expect significant deviation from independence for a mixed population only if the allele frequency distributions were significantly different at <u>both</u> loci, and this does not seem to be the case with pS194 and pL336.

<u>A. Independence of phenotypes</u>

This question is of importance in blood stain work, and can be tested by the following question: is phenotype a_ia_j in pS194 associated with phenotype b_lb_k in pL336? If alleles are sorted by bins, three alleles for each locus yields six phenotypes for each, and 6x6 cells (25df). Five alleles yields fifteen phenotypes for each, and 15x15 cells (196df). No significant deviations from independence of phenotypes was observed (data not shown) at these levels. Further testing was limited by data base size and by problems of misclassification of phenotypes and factitious homozygosity, as noted above for check of Hardy-Weinberg equilibrium.

<u>B. Further validation strategies</u>

Independence of haplotypes a_i, b_k can be tested independent of delta by the binning procedure described above. This test, which validates multiplication of PI, requires experimental haplotypes, which are obtainable from disputed paternity case work.

Powerful tests of Hardy-Weinberg equilibrium require independent determination of δ. This can be obtained by observation of A, and fitting $\hat{PI}_r$ and/or $\hat{PI}_f$ to expected values.

Section VIII. Discussion

Our studies have led us to the following conclusions:

1. Independence of loci may be easily tested by binning from data bases for the presence of a_i and b_k in the same phenotype. Independence at this level implies that cumulative mean exclusion probabilities may be computed in the time honored way. Tests of independence of haplotypes (implying that PI's may be multiplied) requires experimentally obtained haplotype frequencies, but should be easy to perform. Independence of phenotypes cannot be strongly tested directly, because of binning artifacts. However, independence of haplotypes strongly suggests independence of phenotypes.

2. Powerful tests of Hardy-Weinberg equilibrium require independently determined values of δ. Even so, Hardy-Weinberg tests are fraught with artifactual pitfalls. Thus, we would view with skepticism claims that a population is or is not in Hardy-Weinberg Equilibrium.

3. For paternity testing, the experimentally determined mean exclusion probability (A) is a key biostatistic, as it provides on one hand a method to evaluate the appropriateness of PI calculations and, on the other, a method to determine δ. Once δ has been selected in this manner, observed heterozygosity (h) provides a test of Hardy-Weinberg equilibrium (Wahlund's rule).

4. Observed heterozygosity provides a direct way to evaluate the appropriateness of phenotype match calculations.

IX. Acknowledgments

We thank Lynne De La Pointe for her careful and unstinting experimental work and help in software evaluation, and Wendy Dorchester for her advice on statistical measures. Collaborative Research provided us with a generous gift of probes used in this study. This study was supported by the DNA·VIEW Users' Group.

Appendix A — Solution of the Model for δ

The migration model assumes that migration distance, m, and molecular weight, L=L(m), are related by

$$L(m) - L_0 = c / (m - m_0).$$

Differentiating with respect to m:

$$dL/dm = -c / (m - m_0)^2.$$

By a slight abuse of notation we consider dL and dm to be quantities. Assuming that band thickness determines the coelectrophoresis threshold,

for a given

$$\text{band thickness} = \beta = -dm$$

there corresponds a relative percentage discrimination threshold

$$\begin{aligned} \delta &= 100 \cdot dL/L \\ &= (100/cL)(L-L_0)^2\beta. \end{aligned}$$

Using the model

$$\beta(m) = \beta_0 + k \cdot m^{\alpha},$$

we have

$$\delta_{\alpha}(L) = (100/cL)(L-L_0)^2(\beta_0+k \cdot m^{\alpha}).$$

Appendix B — Formula for A from h

System	h	A (eqn V.A.1)	A (eqn V.A.2)	A (actual)
n allele codominant system of equal allele frequencies				
n=4	.750	.563	.510	.504
n=5	.800	.640	.599	.595
n=6	.833	.694	.662	.660
n=7	.857	.734	.708	.707
n=8	.875	.766	.744	.743
n=9	.889	.790	.773	.772
n=10	.900	.810	.795	.795
single locus DNA probe pS194 (Pst I)	.85	.723	.695	.686

For n allele codominant system of equal allele frequencies, A (actual) were computed from an exact equation [Garber RA and Morris JW, Inclusion Probabilities in Parentage Testing. Ed R. Walker AABB, Arlington VA 1983, table 22-1, p278]). For pS194, heterozygosity and A (actual) are values obtained by Dykes, et al [Electrophoresis 9(1988)359-368] and Polesky, et al [DNA for Parentage Testing, Leesburg, VA, April 17-18, 1989, AABB]

Table XII Validation of Estimates for Mean Exclusion Probability (A) from heterozygosity (h)

In this Appendix, we derive equations (V.B.1) and (V.B.2) of page 9.

Suppose h, the observed rate of heterozygosity, has been determined from population data on individuals. From this a theoretical prediction can be made of the expected value for A — the mean exclusion probability —for paternity casework.

We consider all possible mother-child pairs, and categorize them according to the matching band patterns. For each category, we calculate the frequency (as a fraction of all possible mother-child pairs), and the A for all such cases.

Note that h is the chance that two alleles selected at random will fail to match.

Case 1. —Single paternal allele. This occurs either

a. when the mother is homozygous, or frequency = $1-h$

b. when the mother is heterozygous, and

the maternal and paternal alleles

are different frequency = h^2

total frequency = $1-h+h^2$

$A = h^2$ for this case — a tested man is excluded if both of his alleles fail to match the unique paternal allele.

Case 2. —Apparently two paternal alleles. This occurs

when the mother is heterozygous and the paternal allele

matches the non-contributed allele from the mother. frequency = $h(1-h)$

$A \approx h^4$ for this case — to be excluded a tested man's two alleles must each mis-match twice. (It is only an approximation because if the man's alleles mis-match the child's first allele, the chance of missing the second one is diminished.)

The expected value for A is therefore

$$\begin{aligned} A &\approx h^2 \cdot [1-h+h^2] + h^4 \cdot h(1-h) \\ &= h^2 \cdot [1 - h \cdot (1-h-h^2+h^3)] \\ &= h^2 \cdot [1 - h \cdot (1-h)(1-h)(1+h)]. \end{aligned}$$

Writing H=homozygosity for $1-h$, and using $1+h \approx 2$, this becomes

$$A \approx h^2 \cdot [1 - 2hH^2]. \qquad \text{(V.B.2)}$$

For a rougher approximation, use the fact that $H^2 \approx 0$, whence

$$A \approx h^2. \qquad \text{(V.B.1)}$$

Appendix C — Correction Factor to PI_0

When we sum over PI, or over W, calculating

$$PI_0 = 1 / 2 \cdot Pr\{\text{random match}\}$$

for each data point, we are ignoring cases in which AF is homozygous and cases with two paternal alleles.

What effect does this have?

For fathers

1. single paternal allele frequency = h
   a. single match
      (1) AF heterozygous frequency = h
      $PI = 1 / 2Pr\{\text{random match}\}$
      $= PI_0$.
   b. double match
      (1) AF homozygous frequency = 1−h
      $PI = 2 \cdot PI_0$.
2. two paternal alleles frequency = 1−h
   a. single match frequency = h^2
      $PI = ½ \cdot PI_0$.
   b. double match
      (1) AF homozygous frequency = 1−h
      $PI = PI_0$.
      (2) AF heterozygous frequency = h
         (a) 2 matches frequency = 2(1−h)
         $PI = PI_0$.

Combining the above, we get

$$\begin{aligned}\text{corrected PI} &= PI_0 [\, Pr\{PI=PI_0\} + 2 \cdot Pr\{PI=2PI_0\} + ½ \cdot Pr\{PI=½PI_0\} \,] \\ &= PI_0 [\, 1-h(1-h)-h^2(1-h) + 2 \cdot h(1-h) + ½ \cdot h^2/(1-h) \,] \\ &= PI_0 [\, 1 + h(1-h) - ½h^2(1-h) \,] \\ &= PI_0 [\, 1 + ½h(1-h)(2-h) \,] \\ &= PI_0 (1 + \epsilon).\end{aligned}$$

h=.95	ϵ=.0249
.90	.0495
.85	.0733
.80	.0960

Table XIII ϵ values

Appendix D — An Approximate Formula for R

Estimation of mean probability of phenotype match (R) from heterozygosity (h)

Case I —match to heterozygote frequency = h

matching frequency $\approx 2(1\text{-}h)^2$

Case II —match to homozygote frequency = 1−h

matching frequency $= (1-h)^2$

Hence $R \approx h \cdot 2(1-h)^2 + (1-h) \cdot (1-h)^2$

$\approx (1-h)^2(1+h).$

Since $1+h \approx 2$ and $1-h=H$,

$R \approx 2H^2.$

Appendix E — Materials and Methods

Genomic DNA was extracted by salting out according to the method of Dykes, and restricted with Pst I. 5 µg quantities were separated on 0.7% agarose gels in Tris/Borate/EDTA, pH8.2. 30 slot gels (20x24 cm) were run (constant voltage) at 35v (20 milliamps) for 62-65 h, until the 2.3 kb visible marker (lambda phage restricted with Hind III) had run 19-21 cm. DNA was blotted into nylon membranes (Oncor) and hybridized at 42C to biotinylated (Oncor) probes pS194 and pL336 (Collaborative Research). Stringency conditions were 52-60C (pS194) and 60-62C (pL336), 0.16% SSC, 30 min. Hybridization and detection of bands was performed with Oncor reagents according to protocols supplied by the manufacturer. Band sizes were determined by digitizing pad, making use of MW ladders made up from lambda phage restricted with BstE II, Hind III, Sph I, and augmented for detection of alleles greater than 12kb with lambda phage restricted with Xho I and Nco I. A genomic control was run on every gel. Each band was digitized twice. Averaged values were used for analysis. Data analysis and statistical studies were performed with DNA·VIEW, an integrated software package. PI' swere computed by the double integral equation as modified by Brenner [Morris, JW et al. J. Forensic Sci 34 (1989) 1311-1317], [Equation II.4 herein].