Proc. Natl. Acad. Sci. USA
Vol. 92, pp. 12031–12035, December 1995
Genetics

DNA profile match probabilities in a subdivided population: When can subdivision be ignored?

(DNA fingerprints/DNA similarity/population subdivision)

DANIEL E. WEEKS*†, ALAN YOUNG†, AND CHING CHUN LI*‡

*Department of Human Genetics, University of Pittsburgh, Pittsburgh, PA 15261; and †The Wellcome Trust Centre for Human Genetics, University of Oxford, Windmill Road, Oxford OX3 7BN, United Kingdom

Communicated by Newton E. Morton, University of Southampton, Southampton, U.K., August 28, 1995

ABSTRACT **Li and Chakravarti [Li, C. C. & Chakravarti, A. (1994) *Hum. Hered.* 44, 100–109] compared the probability (M_0) of a random match between the two DNA profiles of a pair of individuals drawn from a random-mating population to the probability (M_F) of the match between a pair of random individuals drawn from a subdivided population. The level of heterogeneity in this subdivided population is measured by the parameter F, where there is no subdivision when $F = 0$ and increasing values of F indicate increasing subdivision. Li and Chakravarti concluded that it is conservative to use the match probability M_0, which is derived under the assumption that the two individuals are drawn from a homogeneous random-mating population without subdivision. However, M_0 may not be always greater than M_F, even for biologically reasonable values of F. We explore here those mathematical conditions under which M_0 is less than M_F, and we find that M_0 is not conservative mainly when there is an allele with a much higher frequency than all the other alleles. When empirical data for both variable number of tandem repeat (VNTR) and short tandem repeat (STR) systems are evaluated, we find that in the majority of cases M_0 represents a conservative probability of a match, and so the subdivision of human populations may usually be ignored for a random match, although not, of course, for relatives. Loci for which M_0 is not conservative should be avoided for forensic inference.**

Li and Chakravarti (1) investigated the effects of population subdivision on the probability of a chance match between two DNA profiles. A DNA profile of an individual is generated by typing a sample of DNA at several polymorphic markers. Li and Chakravarti compared the probability of a random match between a pair of individuals drawn from a random-mating population (M_0) to the probability of a match between a pair drawn from a subdivided population (M_F). The level of heterogeneity in this subdivided population is measured by the parameter F, which is defined as an F statistic (2). When $F = 0$, there is no subdivision, and increasing values of F indicate increasing subdivision. Li and Chakravarti concluded that "Since $M_0 > M_F$ for the range of F values in human populations, the use of M_0 is 'conservative'." In other words, they concluded that it is conservative to use a match probability derived under the assumption that the two individuals are drawn from a homogeneous random-mating population without subdivision. However, M_0 may not be always greater than M_F, even for biologically reasonable values of F. We explore here those conditions under which M_0 is less than M_F, when it is inappropriate to ignore population subdivision.

Matching Probabilities

Let $A_1, A_2, \ldots, A_k$ represent the k alleles at an autosomal locus, where the ith allele has frequency p_i, with $\Sigma_{i=1}^{k} p_i = 1$. The probability of a perfect match in a random-mating population is (3)

$$M_0 = \sum_{i=1}^{k} p_i^4 + 4 \sum_{1 \le i < j \le k} p_i^2 p_j^2$$

$$= a_4 + 2(a_2^2 - a_4)$$

$$= 2a_2^2 - a_4,$$

where $a_r = \Sigma_{i=1}^{k} p_i^r$ for integer values of r (= 1, 2, 3, . . .). This equation has been derived by Lange (4) for siblings in the context of an affected-sib-pair method for linkage analysis using identity-by-state relations. Lange's argument can be modified to derive M_0 as follows: Consider the random pair of individuals (**b**, **c**) and (**d**, **e**), where **b**, **c**, **d**, and **e** are four independent genes in a random-mating population. A perfect match occurs whenever (**b**, **c**) = (**d**, **e**)—i.e., when the two individuals have the same genotype. This could happen in two ways: (*i*) **b** = **d** and **c** = **e**, where $P(\mathbf{b} = \mathbf{d}) = a_2$ and $P(\mathbf{c} = \mathbf{e}) = a_2$. The joint probability is thus a_2^2; (*ii*) **b** = **e** and **c** = **d**, which also has a joint probability of a_2^2. However, both possibilities include the special pairs ($\mathbf{A}_i$, $\mathbf{A}_i$) and ($\mathbf{A}_i$, $\mathbf{A}_i$)—i.e., **b** = **c** = **d** = **e**, which has probability a_4. Hence, the probability of a match is

$$M_0 = 2a_2^2 - a_4,$$

as we derived previously by different arguments (3).

The frequency of genotypes in one large population with subdivision may be expressed in terms of two distinct components (ref. 2, pp. 175 and 179):

		Homozygotes		Heterozygotes	
Panmictic	$1 - F$	p_i^2	A_iA_i	$2p_ip_j$	$A_iA_j, i < j$
Fixation	F	p_i	A_iA_i	0	A_iA_j

where F is a measure of heterogeneity ($0 \le F \le 1$) and also the correlation between uniting gametes. The genotype distribution in the panmictic component is the same as that of a random mating population. The fixation component consists of homozygotes only and there are no heterozygotes. Note that for human populations F is typically less than 1% (5).

Let M_F be the probability of a DNA profile match between two random individuals drawn from the heterogeneous population shown above. If both individuals are from the panmictic component, with frequency $(1 - F)^2$, the match probability is simply M_0, the same as that for a random-mating population. If both individuals are drawn from the fixation component, with frequency F^2, the match probability is $\Sigma(p_i \times p_i) = a_2$. If one individual is from the panmictic component and one is

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviations: VNTR, variable number of tandem repeats; STR, short tandem repeat; FBI, Federal Bureau of Investigation.

‡To whom reprint requests should be addressed.

from the fixation component, with frequency $2F(1 - F)$, the match probability is $\Sigma(p_i^2 \times p_i) = a_3$. Hence, the total match probability is (1):

$$M_F = (1 - F)^2 M_0 + 2F(1 - F)a_3 + F^2 a_2,$$

where $F \geq 0$ measures the level of heterogeneity. Note that this equation has recently been independently derived by Collins and Morton (6).

Properties of M_F

Some of the properties of M_F have already been discussed by Li and Chakravarti (1). Here, we shall concentrate on the location of the minimum value of M_F. It was found that the minimum point occurs at the following value of F:

$$F_{\min} = \frac{M_0 - a_3}{M_0 + a_2 - 2a_3},$$

which was derived by setting dM_F/dF equal to zero and solving for F (1). First, we show that the denominator of $F_{\min}$ is always positive. Noting that $\Sigma p_i^2(1 - p_i)^2 = a_2 - 2a_3 + a_4 > 0$, we may write the denominator as

$$\begin{aligned} M_0 + a_2 - 2a_3 &= (2a_2^2 - a_4) + a_2 - 2a_3 + a_4 - a_4 \\ &= 2(a_2^2 - a_4) + (a_2 - 2a_3 + a_4) > 0, \end{aligned}$$

as the quantities in the parentheses are always positive. In the example given by them, $M_0 > a_3$ so that $F_{\min}$ is a positive quantity (Fig. 1*A*). It was on the basis of this example that Li and Chakravarti (1) concluded that the use of M_0 for low values of F in human populations is conservative, because $M_0 > M_F$ in this low-F region.

Now we investigate the conditions under which $M_0 \leq a_3$, with $F_{\min} < 0$. Since, by definition, F must be greater than or equal to zero, then if $M_0 < a_3$, then the minimum value of M_F is M_0, since the equation determining M_F is concave up. Thus, if $M_0 < a_3$, then $M_F \geq M_0$, and therefore it is no longer conservative to use M_0 instead of M_F.

Consider first the case when there are two alleles with frequencies p and q ($p + q = 1$). Then (ignoring the trivial case where either $p = 1$ or $q = 1$)

$$M_0 = a_3 \text{ iff } p^4 + 4p^2q^2 + q^4 = p^3 + q^3.$$

This ultimately reduces to $1 - 6pq = 0$, or $pq = 1/6$, which implies that $p = 0.788675$. Thus, for the two-allele case, $M_0 < a_3$ when $p \geq 0.788675$ or $p \leq 0.211325$, that is, when p is either fairly large or fairly small.

A numerical example at this point may assist in illustrating the situation when $M_0 < a_3$. Consider a population with gene frequencies $(p, q) = (0.9, 0.1)$ or $(0.1, 0.9)$. For this population, $a_2 = 0.82$, $a_3 = 0.73$, $a_4 = 0.6562$, and $M_0 = 2a_2^2 - a_4 = 0.6886 < a_3$. Suppose $F = 0.05$. Then,

$$M_F = (0.95)^2 M_0 + 2(0.05)(0.95)a_3 + (0.05)^2 a_2 = 0.6929,$$

which is greater than $M_0 = 0.6886$. In this situation, M_0 is no longer conservative (Fig. 1*C*). On the other hand, if the gene frequencies are $(p, q) = (0.7, 0.3)$ or $(0.3, 0.7)$, then M_0 is greater than M_F for small values of F and is, therefore, conservative (Fig. 1*A*).

Second, consider the case where there are k alleles of equal frequencies ($p_i = 1/k$). Then $a_2 = 1/k$, $a_3 = 1/k^2$, $a_4 = 1/k^3$, and $M_0 = 2(1/k^2) - 1/k^3 = 1/k^2 + 1/k^2 - 1/k^3 > a_3 = 1/k^2$. Thus, for equally frequent alleles, we always have $M_0 > a_3$ and so $M_0 > M_F$ and it is conservative to use M_0. In fact, this is also true when the allele frequencies are only approximately equal.

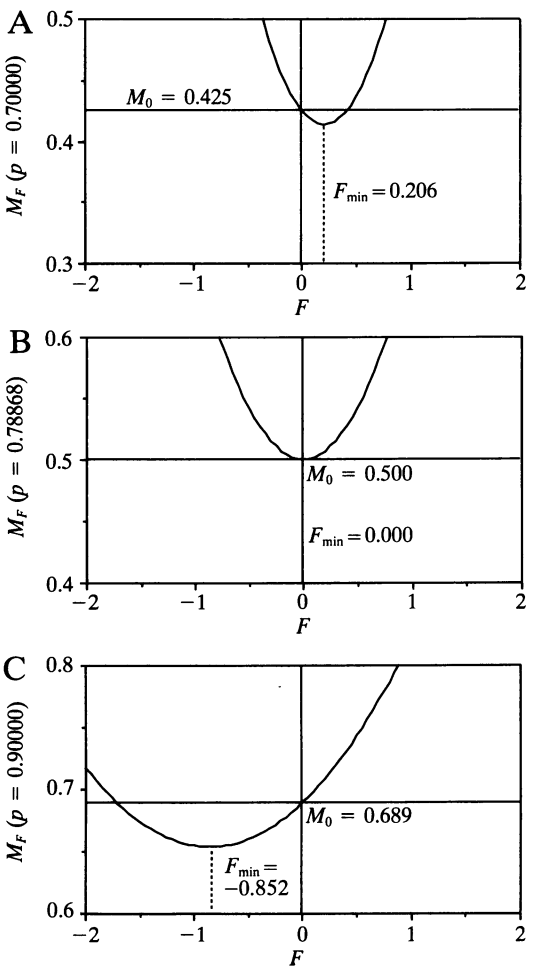

FIG. 1. Graphs of $M_F = (1 - F)^2 M_0 + 2F(1 - F)a_3 + F^2 a_2$ as a quadratic function of F, where a_2, a_3, and M_0 are constants for any given locus. (*A*) The minimum value of M_F occurs at a positive value F. In this case, there is a region of low F values where M_0 is greater than M_F. This is the case with the majority of variable number of tandem repeat (VNTR) loci employed for forensic tests. (*B*) The minimum value of M_F occurs at $F = 0$. In this case, $M_F = M_0$ at $F = 0$. For all other values of F, M_F is greater than M_0. (*C*) The minimum value of M_F occurs at a negative value of F. In this case, all M_F values for positive F are greater than M_0.

To visualize the case where there are three alleles with frequencies p, q, and r ($p + q + r = 1$), we computed those points for which $M_0 - a_3$ was less than zero at a fine grid of points in the (p, q, r) space. The results are displayed in Fig. 2 as a three-dimensional graph, where the equilateral triangle indicates the plane where $p + q + r = 1$. Note that the points for which $M_0 < M_F$ occur in the corners of the triangle, where one of the three allele frequencies is larger than the other two, which have small frequencies.

We now explore mathematically how prevalent the most common allele must be for M_0 to cease to be conservative. Define:

$$\Psi_k = M_0 - a_3 = \sum_{i=1}^{k} p_i^4 + 4 \sum_{1 \leq i < j \leq k} p_i^2 p_j^2 - \sum_{i=1}^{k} p_i^3, \qquad [1]$$

where p_i is the frequency of the ith allele in a k-allele system, subject to

$$\sum_{i=1}^{k} p_i = 1, \quad p_i \geq 0 \text{ for all } i. \qquad [2]$$

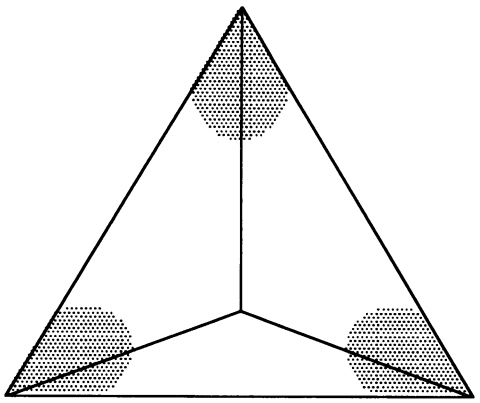

FIG. 2. Graph of the points for which $M_0 < M_F$ in (p, q, r) space. The equilateral triangle with vertices (1, 0, 0), (0, 1, 0), and (0, 0, 1) indicates the plane for which $p + q + r = 1$. Note that the points where M_0 is not conservative (i.e., $M_0 < M_F$) occur in the corners of the triangle, where one allele frequency is larger than the other two frequencies. At all other points in the interior of the triangle, it is conservative to use M_0 instead of M_F.

The problem is to find the region(s) in which $\Psi \leq 0$—i.e., where it is *not* conservative to use M_0 instead of M_F. Suppose without loss of generality that allele 1 is at least as frequent as any other in the population. Let Ω_k be the singly-connected domain (i.e., a region without "holes"), centered on $p_1 = 1$ in which $\Psi_k \leq 0$.

Minimum Bound

CLAIM. *There exists a value* P_c *(greater than* $\frac{1}{2}$*) such that all points with* $p_1 > P_c$ *lie in* Ω_k *for any* $k > 1$.

We provide a proof by induction on k. We have shown earlier that $\Psi_2 < 0$ if p_1 is greater than the value 0.788675.

Lower bound on P_c: Suppose $\Psi_k < 0$ for any arrangement of frequencies such that $p_1 > P_\Lambda$, and consider two arrangements

$$X_n = (p_1, \ldots, p_{n-1}, p_n), \quad X_{n+1} = (p_1, \ldots, p_{n-1}, a, b),$$

where $p_n = a + b$. We claim that $X_n \in \Omega_n \Rightarrow X_{n+1} \in \Omega_{n+1}$. Write $S = \sum_{i<n} p_i^2$ and consider:

$$\begin{aligned}&\Psi_n(X_n) - \Psi_{n+1}(X_{n+1})\\ &= p_n^4 + 4p_n^2S - p_n^3 - a^4 - b^4 - 4a^2S - 4b^2S - 4a^2b^2 + a^3 + b^3\\ &= 4a^3b + 2a^2b^2 + 4ab^3 + 8abS - 3a^2b - 3ab^2\\ &= ab(4a^2 + 2ab + 4b^2 + 8S - 3a - 3b). \qquad [3]\end{aligned}$$

The last term of Eq. **3** can be shown to be positive using the inequalities

$$p_1 > \frac{1}{2} \Rightarrow 8S > 2, \quad a + b < \frac{1}{2} \Rightarrow 3a + 3b < \frac{3}{2}. \qquad [4]$$

Hence $\Psi_n(X_n) > \Psi_{n+1}(X_{n+1})$ as required. Thus P_c must exist and $P_c \geq P_\Lambda$.

Upper bound on P_c: Consider the arrangement

$$X_k(\varepsilon) = (P_\Lambda - \varepsilon, 1 - P_\Lambda + \varepsilon, 0, \ldots, 0).$$

Obviously, $\Psi_k(\varepsilon) = \Psi_2(\varepsilon)$, and, using our results for $k = 2$, we have $\Psi_k(\varepsilon) > 0$ for small $\varepsilon > 0$, so that $P_{\Lambda-\varepsilon} > P_c$. Since ε can be made arbitrarily small this implies $P_c \leq P_\Lambda$.

Combining the upper and lower bounds gives $P_c = P_\Lambda = 0.788675$ for all k.

Maximum Bound

CLAIM. *For* $k \leq 6$, *the smallest value of* p_1 *for which both* $\Psi_k = 0$ *and* $p_1 > p_i$ *for all* $i \neq 1$ *occurs when the other alleles are equifrequent.*

Consider two arrangements:

$$X = (p_1, \ldots, p_{k-2}, a, b), \quad Y = (p_1, \ldots, p_{k-2}, c, c), \qquad [5]$$

where $a + b = 2c > 0$ and $p_1 > p_i$ for all $i \neq 1$. We claim that $\Psi_k(X) \geq \Psi_k(Y)$ if $p_1 > 0.4191$. If this is true, then given any arrangement X of frequencies, we can form a series of arrangements $\{X_1, X_2, \ldots\}$ in which at each stage two of the frequencies (other than p_1) have been replaced by their combined average, and $\Psi_k(X_i) \geq \Psi_k(X_{i+1})$. This series converges to the arrangement

$$Z = (p_1, q, q, \ldots, q), \quad p_1 + (k-1)q = 1. \qquad [6]$$

Hence $\Psi_k(Z) < \Psi_k(X)$, and there must be some critical value P_m (of p_1) such that $\Psi_k(Z(P_m)) = 0$ and $\Psi_k(X(P_m)) > 0$ for any set of non-equifrequent minor alleles.

We prove this by writing $a = c + \varepsilon, b = c - \varepsilon$ and expanding

$$\begin{aligned}\Psi_k(X) - \Psi_k(Y) &= a^4 + b^4 + 4a^2S + 4b^2S + 4a^2b^2 - a^3 - b^3\\ &\quad - 2c^4 - 8c^2S - 4c^4 + 2c^3\\ &= 4c^2\varepsilon^2 + 8\varepsilon^2S + 6\varepsilon^4 - 6c\varepsilon^2\\ &= 2\varepsilon^2(4S + 3\varepsilon^2 + 2c^2 - 3c) \qquad [7]\end{aligned}$$

where $S = \Sigma_{i=1}^{k-2} p_i^2$. Now

$$\begin{aligned}4S + 2c^2 - 3c &\geq 4p_1^2 + 2c^2 - 3c\\ &\geq 4p_1^2 + 2\left(\frac{1-p_1}{2}\right)^2 - 3\left(\frac{1-p_1}{2}\right)\\ &= \frac{9}{2}p_1^2 + \frac{1}{2}p_1 - 1,\end{aligned}$$

since $2c^2 - 3c$ is strictly decreasing on $[0, \frac{3}{4}]$ and $c \leq (1 - p_1)/2 \leq \frac{3}{4}$. Note that

$$\frac{9}{2}p_1^2 + \frac{1}{2}p_1 - 1 > 0 \text{ if } p_1 > \frac{\sqrt{73} - 1}{18} = 0.4191.$$

Thus, $\Psi_k(X) - \Psi_k(Y)$ is positive provided $p_1 > 0.4191$.

Note that the restriction that the number of alleles $k \leq 6$ is a result of the requirement that $p_1 > 0.4191$. For $k > 6$, the critical value of p_1 such that $\Psi_k(p_1, q, q, \ldots, q) = 0$ occurs when $p_1 < 0.4191$ (see Eq. **9** below). Thus, while this value of p_1 may occur when the other alleles are equifrequent, our proof above does not *guarantee* this when $k > 6$. For example, if $k = 7$, then the critical value of p_1 is 0.40342. If we let X = (0.40342, 0.00001, 0.00001, 0.00001, 0.00001, 0.40300, 0.19354), so a = 0.40300 and b = 0.19354, then Y = (0.40342, 0.00001, 0.00001, 0.00001, 0.00001, 0.29827, 0.29827) with c = 0.29827. Then $\Psi_7(X) - \Psi_7(Y) = 0.07036 - 0.07108 = -0.00072$.

Numerical Bounds

We have thus established two values, P_c and P_m (the latter dependent on the dimension) such that

$$0.788675 = P_c < p_1 \leq 1 \Rightarrow \Psi_k < 0 \quad (M_0 \text{ not conservative})$$

$$P_m < p_1 < P_c = 0.788675 \Rightarrow \text{indeterminate}$$

$$p_1 < P_m \Rightarrow \Psi_k > 0 \quad (M_0 \text{ conservative}). \qquad [8]$$

Table 1. Results based on 10,000,000 random probability vectors: Vector with smallest p_1 for which $\Psi < 0$, and percentage of vectors that are nonconservative (i.e., $\Psi < 0$)

Number of alleles	Vector with minimum p_1: Seven largest allele frequencies	% nonconservative
3	0.65556, 0.17258, 0.17186	26.3614
4	0.56374, 0.14778, 0.14474, 0.14373	19.3008
5	0.49692, 0.13400, 0.13213, 0.11855, 0.11839	15.3437
6	0.44572, 0.11679, 0.11196, 0.11127, 0.10774, 0.10653	12.8361
7	0.40548, 0.11436, 0.11205, 0.09733, 0.09246, 0.09041, 0.08791	11.1186
8	0.37211, 0.10346, 0.10069, 0.09225, 0.08596, 0.08396, 0.08345	9.8448
9	0.34763, 0.09622, 0.08916, 0.08552, 0.08096, 0.07791, 0.07647	8.9049

The value of P_m for various dimensions may be computed by finding the smallest positive root of the polynomial

$$M_0 - a_3 = p^4 + 4(k-1)p^2q^2 + (k-1)(2k-3)q^4 - p^3 - (k-1)q^3, \quad [9]$$

where $p + (k-1)q = 1$, giving the results:

k	P_m
2	0.78867
3	0.65554
4	0.56364
5	0.49618
6	0.44444

If the number of alleles is greater than 6, our proof no longer guarantees that the smallest value of p_1 for which both $\Psi_k = 0$ and $p_1 < p_i$ for all $i \neq 1$ occurs when the other alleles are equifrequent. However, we have investigated this by creating 10,000,000 random probability vectors and recording the one with the smallest value of p_1 satisfying these criteria (Table 1). As expected, these values (for $k = 3$ to 6) were very close to the values of P_m displayed above. Table 1 also presents the percentage of vectors that were nonconservative: this percentage starts out at 26% for three alleles and decreases to 8.9% for 9 alleles. However, these percentages may not be accurate, as the relative proportion of the parameter space actually sampled by this Monte Carlo approach decreases rapidly as the number of alleles increases. To check this, we carried out numerical integration for $k = 3$ and 4, obtaining 26.358% and 19.290%, respectively. These match fairly well the results (26.3614, 19.3008) obtained by simulation (Table 1).

Empirical Data

Devlin and Risch (7) estimated allele frequency distributions for two VNTR loci, *D17S79* and *D2S44*, for African American, Caucasian, and Hispanic samples from Federal Bureau of Investigation (FBI) and Lifecodes data bases. On the basis of their allele frequency distributions, we have computed Ψ and found it to be positive at both of these loci for each of the data sets (Table 2), indicating that for these data, M_0 is conservative. Table 2 also displays the frequencies of the five most common alleles.

Hammond *et al.* (8) evaluated several short tandem repeat (STR) loci for forensic use. They present the allele frequencies for 8 STR loci with 5–10 alleles in four populations (Caucasian, Black, Mexican American, and Asian). While most of these systems have at least one common allele (e.g., 81% have an allele with frequency greater than 0.30), all but one of them have a positive Ψ. The exception is a 6-allele system (HUMLIPOL) in Asians, where the most common allele has a frequency of 0.675 (see population III of Table 3).

Discussion

Since the match probability (M_F) for a heterogeneous population is a quadratic function of F, we have to locate the minimum point of M_F in order to plot the graph of M_F. We have found previously that M_F assumes its minimum value at the F value given by $F_{\min} = (M_0 - a_3)/(M_0 + a_2 - 2a_3)$. If $F_{\min}$ is positive (i.e., $M_0 > a_3$), the graph of M_F (Fig. 1*A*) shows $M_0 > M_F$ for low values of F, and we may use M_0 instead of M_F to be conservative (favorable to the defendant).

Now, the question naturally arises: What would be the situation if $M_0 < a_3$ and $F_{\min}$ is negative? This report is essentially dealing with this problem. The situation is shown in Fig. 1*C*. In this case, it is seen in the positive half of F values that M_F is greater than M_0 so that M_0 exaggerates the rarity of the match event and is no longer conservative. A locus which is not conservative is a poor choice for forensic inference.

However, the knowledge described in the paragraph above does not tell us what to do in practical applications. What we need to know are the conditions of the allele frequency distribution that make $M_0 > a_3$ (conservative) or $M_0 < a_3$ (not

Table 2. Observed Ψs and five most common alleles for data from Devlin and Risch (7)

Data set	k	Ψ	Five most common alleles
D17S79			
African Americans, FBI	62	0.00092	0.102, 0.076, 0.075, 0.072, 0.064
Caucasians, FBI	49	0.00283	0.204, 0.161, 0.123, 0.102, 0.044
Southeast Hispanics, FBI	43	0.00279	0.170, 0.153, 0.102, 0.083, 0.079
Southwest Hispanics, FBI	38	0.00111	0.207, 0.140, 0.102, 0.067, 0.058
African Americans, Lifecodes	73	0.00117	0.089, 0.084, 0.079, 0.072, 0.068
Caucasians, Lifecodes	68	0.00280	0.212, 0.144, 0.132, 0.114, 0.064
D2S44			
African Americans, FBI	149	0.00015	0.035, 0.029, 0.027, 0.027, 0.026
Caucasians, FBI	145	0.00013	0.047, 0.030, 0.030, 0.029, 0.029
Southeast Hispanics, FBI	117	0.00017	0.048, 0.040, 0.039, 0.031, 0.029
Southwest Hispanics, FBI	119	0.00018	0.051, 0.050, 0.048, 0.036, 0.034
African Americans, Lifecodes	253	0.00004	0.028, 0.017, 0.017, 0.017, 0.015
Caucasians, Lifecodes	246	0.00005	0.040, 0.024, 0.023, 0.022, 0.019

k is the number of alleles with nonzero frequencies.

Table 3. Examples of allele frequency distributions and the values of a_3 and M_0

Population	Allele frequencies p_i	a_3 M_0	M_0 conservative
I	0.60, 0.20, 0.10, 0.10	$0.2260 > 0.2214$	No
II	0.60, 0.22, 0.10, 0.08	$0.2282 < 0.2288$	Yes
III*	0.675, 0.130, 0.130, 0.045, 0.013, 0.007	$0.3120 > 0.2753$	No
IV†	0.396, 0.370, 0.179, 0.028, 0.019, 0.008	$0.1185 < 0.1695$	Yes

*Example III is based on the HUMLIPOL data for Asians from ref. 8.
†Example IV is based on the HUMCD4 data for Mexican Americans from ref. 8.

conservative). We have explored this above and established that if the frequency of the most common allele, p_1, is greater than $P_c = 0.788675$, then M_0 is not conservative for any number of alleles. In addition, we provide a bound, P_m, for which M_0 is conservative if $p_1 < P_m$; the specific value of P_m depends on the number of alleles. If p_1 lies between P_m and P_c, then M_0 may or may not be conservative. Populations I and II (Table 3) fall into this region of ambiguity, since, for four alleles, we would need $p_1 < P_m = 0.56364$ to be assured that M_0 is conservative. We had hypothesized that M_0 is not conservative whenever there is a single allele whose frequency is much larger than all the others. However, while both populations I and II have an allele that is "much more" frequent than the other alleles, M_0 is not conservative in population I, but is conservative in population II (Table 3). Thus, in regions of ambiguity, it may be necessary to explicitly determine whether or not M_0 is conservative; this involves simple numerical computation of M_0 and a_3 based on the allele frequencies. However, while the exact boundaries of the nonconservative region may be imprecisely known, the points where M_0 is not conservative occur in the corners (Fig. 2), where the most frequent allele is very common, while the points where M_0 is conservative occur internally (the point where $p_i = 1/k$ is always conservative). Populations III and IV (Table 3), which have six alleles, therefore behave as expected, since population III lies in a corner with $p_1 = 0.675$, while in population IV the two most frequent alleles have similar frequencies (0.396 and 0.370).

While we are on the subject of allele frequency distribution, we take the risk of being blamed for belaboring a point which should have been obvious from the beginning. Let us consider a VNTR locus with five alleles but with different distributions (ordered by allele size) in two populations as shown in the following:

Alleles:	A_1	A_2	A_3	A_4	A_5
Population B:	0.1	0.2	0.4	0.2	0.1
Population J:	0.2	0.2	0.1	0.1	0.4

In the B population (say, Brazil) the distribution is bell-shaped (symmetrical). In the J population (say, Japan) the distribution is J-shaped. The two distributions are thus drastically different. However, they have the same values of $a_2 = 0.260$, $a_3 = 0.082$, $a_4 = 0.029$, and $M_0 = 0.1062$. Thus, M_0 is greater than a_3 ($F_{\min}$ is positive) so that M_0 is greater than M_F and is thus conservative. The highest frequency, 0.40, falls below the boundary $P_m = 0.49618$. The allele frequency distribution with respect to size is not important. It is the raw moments of the allele frequencies (a_2, a_3, a_4) that determine the forensic properties of the test locus. A permutation of a set of allele frequencies does not affect its forensic properties. Thus, the allele frequencies may be arranged in descending order from the highest frequency to the lowest (we may call this type of distribution "stair-shaped"). Arranged this way, populations B and J would have the same distribution (0.4, 0.2, 0.2, 0.1, 0.1), revealing that they have the same properties for forensic inference. Thus, as far as forensic applications are concerned, these two frequency distributions (B and J) should be considered identical rather than drastically different.

VNTR loci have been used for a large number of forensic tests, but now there is a shift towards using STR loci. For VNTR loci, the allele frequency distribution tends to be uniform without an allele that is much more frequent than the others, and so M_0 is usually conservative. This conclusion is supported by the fact that VNTR loci have extremely large numbers of alleles: as the number of alleles increases, the percentage of time a random allele frequency vector falls in a nonconservative region decreases (Table 1). In contrast, STR loci have limited numbers of alleles and are much more likely to have a very frequent most common allele; however, M_0 is still usually conservative (at least on the basis of our limited survey of empirical data from ref. 8).

Lewontin (quoted on p. 261 of ref. 9) suggests that the DNA profiling be replaced by idiotyping. Idiotyping is a method based on differences in actual DNA sequences among the repeats in the VNTR loci, so that each person can be recognized by a unique sequence. Then, as for dermal fingerprinting, there would be no need to calculate the probabilities of matching or likelihood ratios. Lewontin (10) says it is a real "DNA fingerprinting" ("DNAprinting" would be a better term). Such methods are now under development. Note, however, that the hypothesis that every person can be identified by a unique sequence requires support from an extremely large data base, and it is violated by siblings. There will be a day when both Lewontin and forensic scientists are happy, but we cannot rush history. Pending DNAprinting (based on sequences rather than fragment size), usually we may disregard the heterogeneity of human populations and use M_0 as a conservative probability of a match.

We would like to thank Dr. B. Devlin for providing the raw VNTR data from his paper, and Dr. Kenneth Lange and Jeffrey O'Connell for helpful comments. This work was supported in part by funds from the University of Pittsburgh, National Institutes of Health Grant HG00932 (D.E.W.), the Wellcome Trust Centre for Human Genetics, the University of Oxford, the Association Française Contre Les Myopathies (AFM), and the W. M. Keck Center for Advanced Training in Computational Biology at the University of Pittsburgh, Carnegie Mellon University, and the Pittsburgh Supercomputing Center.

1. Li, C. C. & Chakravarti, A. (1994) *Hum. Hered.* **44,** 100–109.
2. Wright, S. (1969) *Evolution and the Genetics of Populations: The Theory of Gene Frequencies* (Univ. of Chicago Press, Chicago), Vol. 2.
3. Li, C. C., Weeks, D. E. & Chakravarti, A. (1993) *Hum. Hered.* **43,** 45–52.
4. Lange, K. (1986) *Am. J. Hum. Genet.* **39,** 148–150.
5. Morton, N. E. (1992) *Proc. Natl. Acad. Sci. USA* **89,** 2556–2560.
6. Collins, A. & Morton, N. E. (1994) *Proc. Natl. Acad. Sci. USA* **91,** 6007–6011.
7. Devlin, B. & Risch, N. (1992) *Am. J. Hum. Genet.* **51,** 534–548.
8. Hammond, H. A., Jin, L., Zhong, Y., Caskey, C. T. & Chakraborty, R. (1994) *Am. J. Hum. Genet.* **55,** 175–189.
9. Roeder, K. (1994) *Stat. Sci.* **9,** 222–278.
10. Lewontin, R. C. (1994) *Stat. Sci.* **9,** 259–262.