Estimation of Heterozygosity for Single-Probe Multilocus DNA Fingerprints[1]

J. Claiborne Stephens, Dennis A. Gilbert,† Naoya Yuhki,* and Stephen J. O'Brien**

*Laboratory of Viral Carcinogenesis and †Biological Carcinogenesis and Development Program, Frederick Cancer Research and Development Center

In spite of the increasing application of DNA fingerprinting to natural populations and to the genetic identification of humans, explicit methods for estimation of basic population genetic parameters from DNA fingerprinting data have not been developed. Contributing to this omission is the inability to determine, for multilocus fingerprinting probes, relatively important genetic information, such as the number of loci, the number of alleles, and the distribution of these alleles into specific loci. One of the most useful genetic parameters that could be derived from such data would be the average heterozygosity, which has traditionally been employed to measure the level of genetic variation within populations and to compare genetic variation among different loci. We derive here explicit formulas for both the estimation of average heterozygosity at multiple hypervariable loci and a maximum value for this estimate. These estimates are based upon the DNA restriction-pattern matrices that are typical for fingerprinting studies of humans and natural populations. For several empirical data sets from our laboratory, estimates of average and maximal heterozygosity are shown to be relatively close to each other. Furthermore, variances of these statistics based on simulation studies are relatively small. These observations, as well as consideration of the effect of missing alleles and alternate numbers of loci, suggest that the average heterozygosity can be accurately estimated using phenotypic DNA fingerprint patterns, because this parameter is relatively insensitive to the lack of certain genetic information.

Introduction

Hypervariable minisatellite or VNTR (variable number of tandem repeat) probes have proved to be quite powerful for resolving genetic identity or relationship in humans (Gill et al. 1985; Jeffreys et al. 1985*b*, 1985*c*; Baird et al. 1986; Gilbert et al. 1990*b*) and other vertebrate species (Burke and Bruford 1987; Jeffreys and Morton 1987; Jeffreys et al. 1987; Wetton et al. 1987; Burke et al. 1989; Gilbert et al. 1990*a*, 1991; Kuhnlein et al. 1990; Reeve et al. 1990). When used in conventional Southern blotting experiments, these probes produce "DNA fingerprints," so-called because of the demonstrable individual specificity in outbred populations (Jeffreys et al. 1985*a*, 1985*b*, 1985*c*; Jeffreys 1987; Gilbert et al. 1990*b*). The genetic basis for this specificity is that a single VNTR probe is homologous to a moderate or large number of chromosomally dispersed genomic loci, many of which exhibit considerable genetic poly-

1. Key words: hypervariable minisatellites, genetic relatedness, polymorphism, gene diversity, population genetics.

Address for correspondence and reprints: J. Claiborne Stephens, Laboratory of Viral Carcinogenesis, National Cancer Institute, Frederick Cancer Research and Development Center, Frederick, Maryland 21702-1201.

Mol. Biol. Evol. 9(4):729–743. 1992.

0737-4038/92/0904-0013$02.00

morphism (Jeffreys et al. 1985*a*, 1990; Wong et al. 1986, 1987; Jeffreys 1987; Nakamura et al. 1987). The simultaneous screening of many highly polymorphic loci is, in many ways, ideal for resolving questions of genetic individualization.

To date, published population studies involving minisatellite probes have ignored several important genetic parameters. In these studies, band sharing (or difference) has been used as a simple phenotypic metric to compare populations (Jeffreys et al. 1985*a*, 1985*b*, 1985*c*; Burke and Bruford 1987; Wetton et al. 1987; Gilbert et al. 1990*a*; Reeve et al. 1990). Previous theoretical approaches (Lynch 1988, 1990) to DNA fingerprint data have concentrated on the relationship of phenotypic band sharing to other population genetic parameters (measures of relatedness, identity in state, and population homozygosity) but have not attempted to estimate genomic heterozygosity, in theory or in practice. Our own laboratory has attempted to relate band sharing to heterozygosity; however, the present paper represents a new, explicit estimation for average heterozygosity, an estimation that we believe has marked advantages over previous attempts (Gilbert et al. 1990*b*; Yuhki and O'Brien 1990).

Our estimate is intended for the following empirical conditions encountered frequently with multilocus DNA fingerprint probes: (1) multiple loci are screened simultaneously with a single probe; (2) these loci are generally unlinked and otherwise genetically independent; (3) many (often more than 10) bands of unknown genetic relationship to each other are observed in each individual; and (4) a pair of bands may be alternative alleles at a single locus or may be alleles at two different loci, but they are generally not both fragments of the same allele at a locus. The lack of information about relationship among the various alleles makes each allele effectively dominant: we cannot generally discern whether an individual is heterozygous or homozygous for each allele.

Condition 2 is important for computing heterozygosity of multiple minisatellite loci and, indeed, for any collection of loci presumed to be a random sample of the genome. This condition will be true for large population samples, in the absence of tight linkage, epistatic selection, and other forces that generate linkage disequilibrium among loci. In DNA fingerprint data sets, one can test for allelic independence by tracking the departure of pairwise fragment cooccurrences from expectation (Gilbert et al. 1990*b*); however, with available sample sizes the power of this test is not very high. Nonetheless, linkage disequilibrium is rare between tested linked loci that are more than a few centimorgans apart (Hill 1974; Bodmer and Cavalli-Sforza 1976; Hartl 1980). In our estimates of heterozygosity, we assume that alleles of minisatellite loci are independent, although we remain aware of the caveats implicit in this assumption.

The first three conditions listed above are often true for minisatellite probes, because of the homology between the probe's minisatellite core sequence and multiple polymorphic loci interspersed throughout the genome. In principle, these conditions also apply to any data set obtained from a probe that has homology to multiple genomic loci; examples of such probes are those corresponding to loci with many dispersed pseudogenes or those obtained from multigene families such as the major histocompatibility complex (MHC) (Yuhki and O'Brien 1990), although one needs to be more careful about linkage effects in the latter situation.

We note that a large number of probes to specific, individual loci have been developed from minisatellite systems by screening clone libraries with a panel of oligonucleotides based on the core sequences (Wong et al. 1986, 1987; Nakamura et al. 1987, 1988; Gilbert et al. 1991). Our present concern is with the estimation of het-

erozygosity when detailed knowledge of the component loci underlying a DNA fingerprint is not available or would be impractical to obtain. This is often the case for DNA studies of natural populations, especially when probes from heterologous species are used.

The fourth condition listed above is usually appropriate for minisatellite probes, since polymorphism is generally due to variable lengths of the tandem repeat—potential nucleotide sequence variation within the flanking restriction sites or within the tandem repeat itself is generally not monitored (but see Jeffreys et al. 1990). For probes detecting either variation within multigene families or other nontandem repeat variation, the fourth condition may not hold, in which case the presence of two bands as a single allele would be detectable as an absolute concordance in the occurrence of the two bands (observation of only ++ and −− individuals).

Several factors determine the average heterozygosity of systems of loci such as those under consideration here. One would like to know the number of polymorphic loci, the number and frequency of alleles, the distribution of alleles into specific loci, and the dominance relationships among the alleles at each locus. These factors are seldom known in experimental studies, but the principal factors relevant to the calculation of heterozygosity can be estimated, as we show below. The reliability of our estimates has been tested by using simulation studies, which show that the estimation of heterozygosity is not overly sensitive to the absence of the aforementioned genetic information.

Calculation of Heterozygosity

The standard unbiased estimate of heterozygosity (h) at a given locus is

$$h = \left(1 - \sum_{j=1}^{A} x_j^2\right) \frac{2n}{2n-1}, \qquad (1)$$

(Nei and Roychoudhury 1974), where n is the number of individuals sampled and A is the number of alleles at the locus, each with estimated frequency x_j. The quantity calculated in equation (1) has also been called the "gene diversity" (Nei 1973), because it has a meaningful genetic interpretation in all types of organisms (e.g., viruses, bacteria, and polyploid organisms), not just in diploids. (In the case of nondiploid organisms, the $2n$'s should be replaced by n times the ploidy.) This value is the observed heterozygosity for diploid populations if the genotypes are in Hardy-Weinberg equilibrium. In field studies of natural populations, it is customary to test whether genotypes are in Hardy-Weinberg equilibrium. Equilibrium is typically observed unless there is strong population subdivision or unless directional forces such as natural selection are in operation.

An unbiased estimate of average heterozygosity (H) for a system of L loci is

$$H = \frac{\sum_{i=1}^{L} \left(1 - \sum_{j=1}^{A_i} x_{ij}^2\right) \frac{2n}{2n-1}}{L} = \frac{2n}{2n-1}\left(1 - \frac{\sum_{i=1}^{L}\sum_{j=1}^{A_i} x_{ij}^2}{L}\right) = \frac{2n}{2n-1}\left(1 - \frac{\sum_{k=1}^{A} x_k^2}{L}\right), \qquad (2)$$

where A_i is the number of alleles at the ith locus and x_{ij} is the estimated frequency of the jth allele at the ith locus. Since the sum of A_i equals A (the total number of alleles

observed over all polymorphic and monomorphic loci), and since we can determine A by observation, the double summation in the middle term has been reduced to the single summation at the end of equation (2). In words, the specific distribution of alleles into loci is immaterial to the estimation of heterozygosity.

Because each band or allele in a DNA fingerprint is effectively dominant, we must estimate each allele frequency (x_k) from the frequency of occurrence of the kth band (s_k). If we make the assumption that genotypes are in Hardy-Weinberg equilibrium, then

$$x_k = 1-\sqrt{1-s_k}\,, \tag{3}$$

as in the papers by Gilbert et al. (1990*b*) and Yuhki and O'Brien (1990). The estimates of the individual x_k's can be totaled to provide an estimate of L, as in the paper of Gilbert et al. (1990*b*):

$$L = \sum_{k=1}^{A} x_k\,. \tag{4}$$

Substitution into equation (2) gives

$$H = \frac{2n}{2n-1}\left(1-\frac{\sum_{k=1}^{A} x_k^2}{\sum_{k=1}^{A} x_k}\right) \tag{5}$$

as our estimate of average heterozygosity. The formula corresponding directly to the observable quantities (A and s_k) is

$$H = \frac{2n}{2n-1}\left(\frac{\sum_{k=1}^{A} s_k}{A-\sum_{k=1}^{A}\sqrt{1-s_k}}-1\right). \tag{6}$$

We have estimated A across all loci as being the number of different scorable bands on the gel, a procedure which may miss additional alleles and additional loci. We discuss the effect of this possibility below.

Maximum Heterozygosity for an Observed Number of Alleles

The primary observations in the types of data sets considered here are A and s_k. Above we have estimated x_k's and the number of loci from these data. What number of polymorphic loci and distribution of alleles into loci will maximize the heterozygosity estimate for a given data set? According to Yuhki and O'Brien (1990), each band for which $s_k = 1$ is treated as a monomorphic locus (x_k=1). Let L_M be the number of such loci and let $A_P = A-L_M$ be the number of polymorphic bands, i.e., the number of alleles at the variable loci. We derive an estimate of the number of polymorphic loci (L_P) that maximizes the average heterozygosity for a given L_M and A_P. In doing this we will ignore the observed s_k for the polymorphic bands, since our objective is

to ask, What distribution of A_P alleles into L_P polymorphic loci will maximize H for a given L_M?

We note that for $L_M = 0$, $A_P = A$, and H is maximized trivially by assuming both that all of the alleles occur at a single locus ($L=L_P=1$) and that each of the alleles is present in frequency $1/A$. In this case, $H=[2n/(2n-1)][1-(1/A)]$, from equation (1). For $L_M > 0$, there is an optimum L_P that maximizes H, which reflects the trade-off between many slightly heterozygous loci and a few highly heterozygous loci.

For a fixed number of alleles at a locus, an even distribution of x_k's maximizes heterozygosity, as above for $L_P = 1$. Likewise, for a fixed number of loci, an even distribution of alleles into loci will also maximize heterozygosity. In algebraic terms, heterozygosity is maximized when $A_i = A_P/L_P$ and when each $x_{ij} = 1/A_i = L_P/A_P$. Recognizing that the L_M monomorphic loci contribute nothing to heterozygosity, we have, from equation (2),

$$H_{\max} = \frac{2n}{2n-1}\left\{\frac{\sum_{i=1}^{L_P}\left[1-\sum_{j=1}^{A_P/L_P}\left(\frac{L_P}{A_P}\right)^2\right]}{L_M+L_P}\right\} = \frac{2n}{2n-1}\left[\frac{L_P(1-L_P/A_P)}{L_M+L_P}\right]. \quad (7)$$

Because L_M and A_P are known, it remains for us to choose L_P such that $H_{\max}$ is, in fact, the maximum. From elementary calculus, it can be shown that

$$L_P = \sqrt{L_M^2+L_MA_P}-L_M = \sqrt{L_MA}-L_M\,. \quad (8)$$

Statistical Considerations

If the x_k in our estimate had been estimated from genotypic (rather than phenotypic) observations, our estimate L would normally be quite accurate and generally could be treated as a constant. The variance of L in our simulation studies is generally quite small, with coefficients of variation that are on the order of 2%–3%. If L were an unbiased estimate with a symmetric distribution, the general effect would be to increase both the mean and the variance of H, because L occurs in the denominator of equation (2). An additional concern is that the allele frequency estimates were derived by assuming that genotype frequencies are in Hardy-Weinberg equilibrium. This is the case for the vast majority of loci and organisms, but one should still be cautious in any particular study. Overall, it does not appear that use of phenotypic observations (i.e., the s_k) would greatly distort H from the value that would be obtained were gene frequencies estimated more accurately.

Nei and Roychoudhury (1974) have shown that the sampling variance of average heterozygosity is composed of an intralocus variance and an interlocus variance. The interlocus variance is due to heterozygosity differences among the constituent loci in the sample, whereas the intralocus variance is based on sampling fluctuations of x_k's at each locus. Although we are unable to resolve the x_k's of a DNA fingerprint in a locus-by-locus fashion, it is still feasible to obtain a conservative estimate of the variance. According to Nei and Roychoudhury, if average heterozygosity is estimated as

$$H = \frac{\sum_{i=1}^{L} h_i}{L}, \quad (9)$$

Band	Individual abcdefghijklmno	s_k	x_k	Band	Individual abcdefghijklmno	s_k	x_k
1	---+-+++-+-----	0.33	0.18	35	-+--+--------+	0.20	0.11
2	+-++--+-+++--+-	0.53	0.32	36	--+----------	0.07	0.03
3	----------+++-+	0.27	0.14	37	+-+--+--------	0.20	0.11
4	+-++--+-------	0.27	0.14	38	-++-+------+-++	0.40	0.23
5	--++-+---+-----	0.27	0.14	39	----------+-+--	0.13	0.07
6	----------++-++	0.27	0.14	40	----+----+-----	0.13	0.07
7	-+--+--++-+-++-	0.47	0.27	41	----------+-	0.07	0.03
8	+++++++++++++++	1.00	1.00	42	-+--+-+-------	0.20	0.11
9	++++++++++----	0.73	0.48	43	----------+--	0.07	0.03
10	+--+-++++------	0.40	0.23	44	----+--------+-	0.13	0.07
11	+++++++++++++++	1.00	1.00	45	----------+-+--	0.13	0.07
12	+++++++++++++++	1.00	1.00	46	++++++++++-++++	0.93	0.74
13	----+----------	0.07	0.03	47	-+--+----++++++	0.53	0.32
14	+-+--++++-+-++-	0.60	0.37	48	--+----------	0.07	0.03
15	----------+--+	0.13	0.07	49	+--+-+-------+-	0.27	0.14
16	----------++-+	0.20	0.11	50	-----+--------+	0.13	0.07
17	----------+-++-	0.20	0.11	51	------+-+------	0.13	0.07
18	+-+---+++-+-+--	0.47	0.27	52	----------+--	0.07	0.03
19	++-+-+-+-+++--+	0.60	0.37	53	----------+-++-	0.20	0.11
20	----------+-+--	0.13	0.07	54	-+--+-+-++-+--+	0.47	0.27
21	----------+++++	0.33	0.18	55	-+--+----------	0.13	0.07
22	---+--+-+------	0.20	0.11	56	----------++-++	0.27	0.14
23	+-++--++-------	0.33	0.18	57	-------+-------	0.07	0.03
24	+++++++++++++++	1.00	1.00	58	+++++---++++-+-	0.67	0.42
25	++--++++-------	0.40	0.23	59	-----+---+----+	0.20	0.11
26	+----+--+-+++++	0.53	0.32	60	----------++-+-	0.20	0.11
27	--+----------	0.07	0.03	61	----------+----	0.07	0.03
28	+-++-+---------	0.27	0.14	62	---+-+-+--+----	0.27	0.14
29	-+-----+--+++-+	0.40	0.23	63	----+----+-----	0.13	0.07
30	----------+--+-	0.13	0.07	64	----------+----+	0.13	0.07
31	----------+	0.07	0.03	65	+++++++++++++++	1.00	1.00
32	---+-+---------	0.13	0.07	66	+++++++++++++++	1.00	1.00
33	-++----++-+----	0.33	0.18	67	+++++++++++++++	1.00	1.00
34	------+-++-----	0.20	0.11				

FIG. 1.—DNA fingerprinting data set of 15 unrelated domestic cats (labeled a–o). Genomic DNA was digested with *Msp*I and was probed with feline minisatellite probe FCZ9 (Gilbert et al. 1991). Sixty-seven distinct scorable bands were seen; a plus sign (+) denotes presence in an individual, and a minus sign (−) indicates absence in an individual. s_k estimates and x_k estimates are indicated; other estimated parameters are $L = \sum_{k=1}^{67} x_k = 16.22$ and $H = 0.4180$; $L_M = 7$ and $A_P = 60$, from which $L_P = 14.66$ and $H_{max} = 0.5115$.

where h_i is the estimate of heterozygosity of the ith locus, the sampling variance of H may be estimated as

$$V(H) = V(h)/L\,, \tag{10}$$

in which $V(h)$ is the expected variance of h, estimated by

$$V(h) = \frac{\sum_{i=1}^{L} (h_i - H)^2}{L-1}\,, \tag{11}$$

under the assumption that the heterozygosities among loci are not correlated. In the present situation, the individual h_i's are not precisely estimable. However, in practice, the observed band frequencies are often extremely bimodal; that is, some number of loci (i.e., L_M) are monomorphic, with the rest (i.e., L_P) being highly heterozygous. In the extreme, then, each locus has $h = 1$ with probability $H = L_P/L$, or $h = 0$ with probability $L_M/L = 1-H$. This situation suggests that heterozygosity at each locus is

Table 1
Estimates of Average and Maximal Heterozygosity by Using Multilocus Minisatellite Probes in Natural Populations

Species	Population	Probe	Restriction Enzyme	No. of Individuals	No. of Bands	L_M	$L-L_M$	L_p	P	$P_{(max)}$	H	H_{max}	APD	Reference
Human		33.6	*Hae*III	33	121	0	9.41	1.00	1.00	1.00	0.86	0.99	76.9	Gilbert et al. (1990b)
		33.6	*Hin*fI	33	114	0	8.34	1.00	1.00	1.00	0.88	0.99	81.1	Gilbert et al. (1990b)
Fox	SCA	33.6	*Hin*fI	16	37	7	8.73	9.09	0.56	0.57	0.31	0.39	25.3	Gilbert et al. (1990a)
	SRO	33.6	*Hin*fI	16	30	9	6.68	7.43	0.43	0.45	0.26	0.29	23.7	Gilbert et al. (1990a)
	SCR	33.6	*Hin*fI	11	25	16	2.56	4.00	0.14	0.20	0.09	0.11	10.0	Gilbert et al. (1990a)
	SCL	33.6	*Hin*fI	16	33	20	2.90	5.69	0.13	0.22	0.08	0.13	8.5	Gilbert et al. (1990a)
	SMI	33.6	*Hin*fI	10	31	25	1.22	2.84	0.05	0.10	0.04	0.05	4.7	Gilbert et al. (1990a)
	SNI	33.6	*Hin*fI	14	19	19	0.00	0.00	0.00	0.00	0.00	0.00	0	Gilbert et al. (1990a)
Domestic cat ...		33.6	*Hin*fI	16	29	4	3.31	6.77	0.45	0.63	0.36	0.46	40.5	Gilbert et al. (1991)
		33.15	*Hin*fI	18	27	0	7.16	1.00	1.00	1.00	0.58	0.96	42.5	Gilbert et al. (1991)
		FCZ8	*Msp*I	17	76	6	10.31	15.35	0.63	0.72	0.48	0.56	47.1	Gilbert et al. (1991)
		FCZ9	*Msp*I	15	67	7	9.22	14.66	0.57	0.68	0.43	0.51	44.5	Gilbert et al. (1991)
Lion	SRI	FCZ8	*Msp*I	15	71	5	9.27	13.84	0.65	0.74	0.50	0.58	49.7	Gilbert et al. (1991)
	NGC	FCZ8	*Msp*I	23	89	8	10.96	18.68	0.58	0.70	0.45	0.54	45.8	Gilbert et al. (1991)
	GIR	FCZ8	*Msp*I	16	23	20	1.35	1.45	0.06	0.07	0.03	0.04	2.6	Gilbert et al. (1991)

Table 2
Estimation of Average and Maximal Heterozygosity by Using MHC Probes in Natural Populations

Species and Restriction Enzyme	No. of Individuals	No. of Bands	H	H_{max}	APD	MAPD[a]
Human:						
*Pst*I	8	28	0.077	0.110	8.08	
*Bam*HI	6	15	0.106	0.156	13.20	
*Eco*RI	8	8	0.108	0.117	8.94	
						10.07
Domestic cat (sample 1):						
*Pst*I	16	16	0.107	0.143	8.28	
*Bam*HI	16	8	0.138	0.172	10.72	
*Eco*RI	18	10	0.092	0.127	8.65	
*Eco*RV	16	10	0.194	0.292	13.01	
						10.17
Domestic cat (sample 2):						
*Pst*I	11	18	0.119	0.146	7.94	
*Bam*HI	11	15	0.067	0.077	6.26	
*Eco*RI	18	10	0.092	0.127	8.65	
*Eco*RV	22	9	0.042	0.063	3.53	
						6.60
Cheetah (eastern):						
*Pst*I	13	18	0.000	0.000	0.00	
*Bam*HI	13	14	0.027	0.039	2.32	
*Eco*RI	13	18	0.042	0.063	3.53	
*Eco*RV	13	13	0.033	0.042	3.05	
						2.23
Cheetah (southern, sample 1):						
*Pst*I	8	20	0.000	0.000	0.00	
*Bam*HI	8	14	0.025	0.039	1.92	
*Eco*RI	8	17	0.041	0.049	3.85	
*Eco*RV	8	13	0.042	0.042	2.74	
						2.13
Cheetah (southern, sample 2):						
*Pst*I	8	20	0.034	0.056	4.37	
*Bam*HI	9	15	0.024	0.036	2.21	
*Eco*RI	9	18	0.065	0.081	6.12	
*Eco*RV	9	13	0.005	0.020	0.89	
						3.40
Lion (Serengeti):						
*Pst*I	18	19	0.084	0.095	7.80	
*Bam*HI	18	18	0.148	0.172	9.68	
*Eco*RI	18	26	0.081	0.092	8.35	
*Eco*RV	18	12	0.080	0.134	5.25	
*Hin*dIII	14	22	0.133	0.172	12.94	
						8.80
Lion (Ngorongoro Crater):						
*Pst*I	15	19	0.047	0.076	4.69	
*Bam*HI	15	21	0.060	0.068	6.08	
*Eco*RI	15	24	0.051	0.072	5.46	
*Eco*RV	15	11	0.016	0.024	2.27	
*Hin*dIII	15	17	0.034	0.049	4.26	
						4.55

Table 2 (*Continued*)

Species and Restriction Enzyme	No. of Individuals	No. of Bands	H	H_{max}	APD	MAPD[a]
Lion (Gir Forest):						
*Pst*I	18	25	0	0	0.00	
*Bam*HI	18	14	0	0	0.00	
*Eco*RI	18	12	0	0	0.00	
*Eco*RV	18	9	0	0	0.00	
*Hin*dIII	18	11	0	0	0.00	
						.00

NOTE.—These values replace those in the paper by Yuhki and O'Brien (1990). Probes used were pFLA24 (for felids) and HLAB7 (for humans).

[a] Average of APD among all enzymes used.

a Bernoulli trial, in which case the average heterozygosity has binomial variance of approximately

$$V(H) = H(1-H)/L\,, \tag{12}$$

which is similar to the result obtained by using equations (10) and (11).

Of more concern, statistically, is $V_s(H)$, the variance that describes the distribution of H on repeated sampling of a given population. Nei (1987) has given an estimate of this variance, for heterozygosity of the ith locus:

$$V_{si}(h) = \frac{2n}{2n(2n-1)}\left\{2(2n-2)\left[\sum_{j=1}^{A_i} x_j^3 - \left(\sum_{j=1}^{A_i} x_j^2\right)^2\right] + \sum_{j=1}^{A_i} x_j^2 - \left(\sum_{j=1}^{A_i} x_j^2\right)^2\right\}. \tag{13}$$

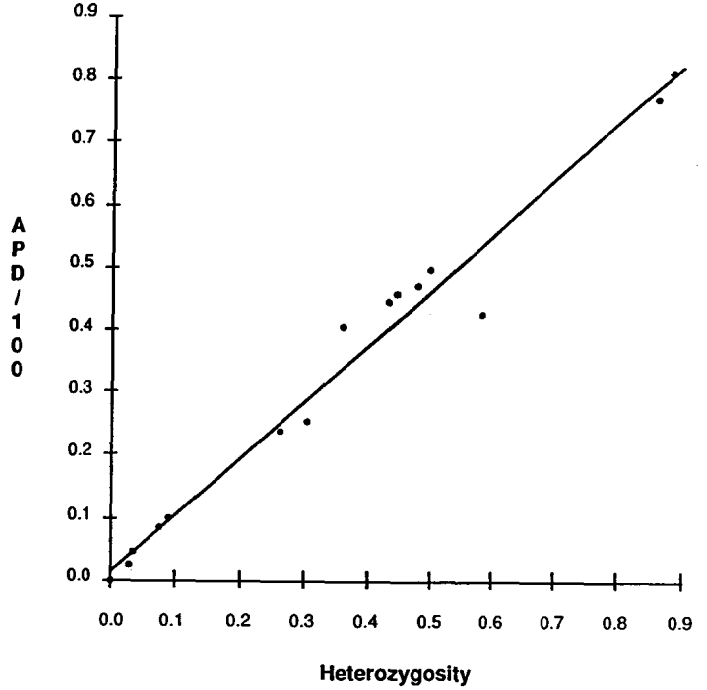
Unfortunately, this expression does not readily lead to the desired variance in the current situation, because the alleles would need to be partitioned among the appropriate loci. Below, we estimate $V_s(H)$ by computer simulation.

Sample Calculations

We have developed several population genetic data sets that involve DNA fingerprinting data (Gilbert et al. 1990*a*, 1990*b*, 1991; Yuhki and O'Brien 1990). Figure 1 shows both how each data set is coded and the relevant parameters and their estimates from this data set. Table 1 summarizes our estimates of the genetic parameters as calculated above for several DNA fingerprinting data sets, along with the phenotypic metric average percent difference (APD) in bandsharing. APD is simply the average among all pairwise comparisons between individuals of the quantity

$$\mathrm{PD} = \frac{(b_x - b_{xy}) + (b_y - b_{xy})}{b_x + b_y} \times 100 = 100\left(1 - \frac{2b_{xy}}{b_x + b_y}\right) = 100(1 - S_{xy})\,, \tag{14}$$

where $S_{xy} = 2b_{xy}/(b_x + b_y)$. The quantity PD is the percent difference in band sharing, defined by Yuhki and O'Brien (1990), where b_x is the number of bands observed in individual x, b_y is the number of bands observed in individual y, and b_{xy} is the number of bands shared between x and y. Note that PD has a simple mathematical relationship to S_{xy}, the similarity index defined by Lynch (1990).



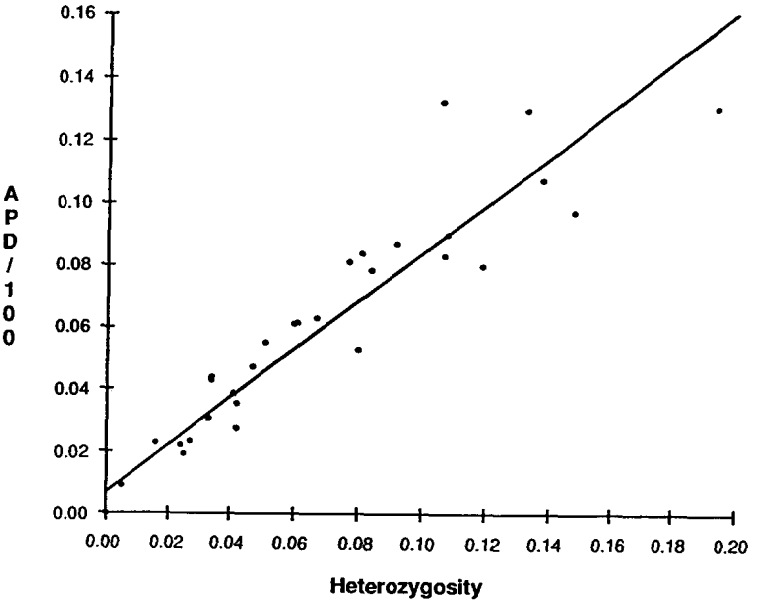


FIG. 2.—Relationship of average percent difference (APD/100) in bandsharing values when H is calculated from eq. (5) in the text. *Left,* DNA fingerprinting data sets (table 1). Sample correlation coefficient 0.986; $P < 0.01$. *Right,* MHC fingerprinting data sets (table 2). Sample correlation coefficient 0.949; $P < 0.01$.

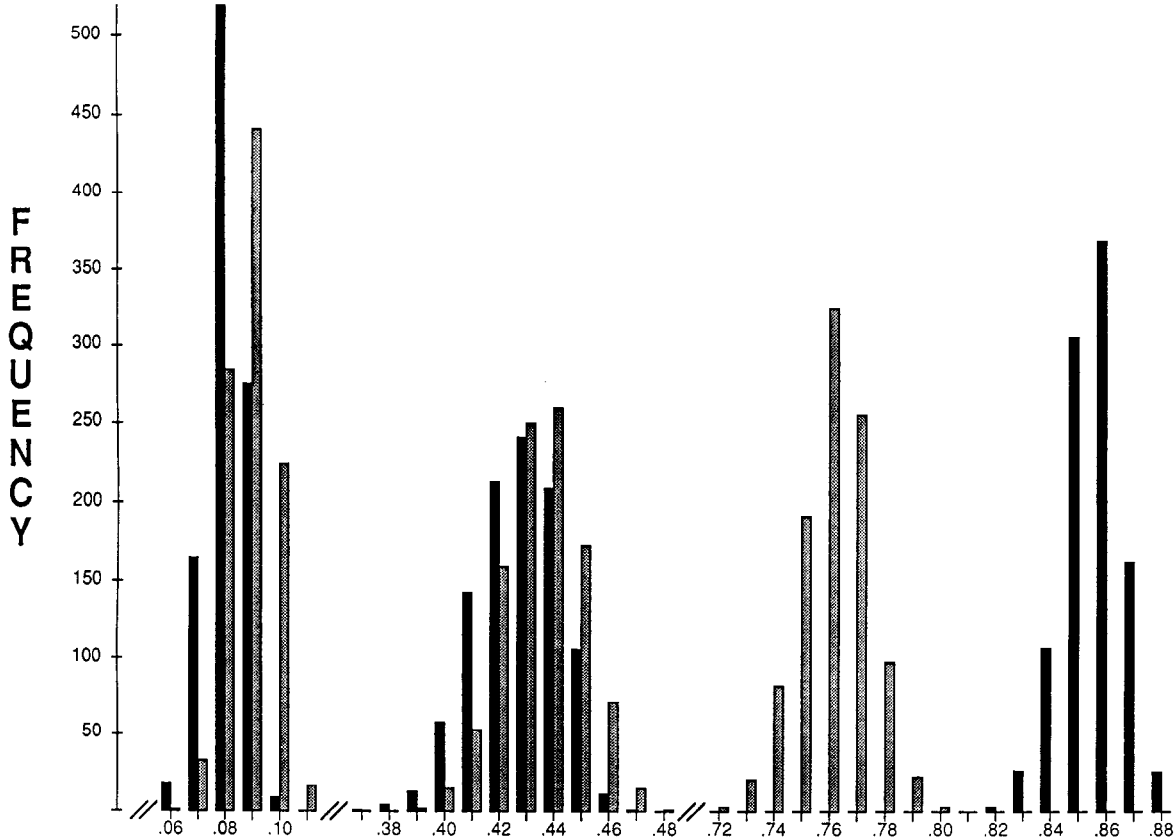


FIG. 3.—Distribution of H [■; eq. (5)] and APD/100 [▩; eq. (14)] in simulation studies of each of three empirical population samples. The leftmost distributions (.06–.11) are for 1,000 replicate samples based upon the Santa Clara Island fox population (SCL in table 1; Gilbert et al. 1990*a*). The intermediate distributions (.37–.48) are for 1,000 replicate samples based upon a study of domestic cats (FCZ9 in table 1; Gilbert et al. 1991). The rightmost distributions (.72–.88) are for 1,000 replicate samples based upon a study of human cell lines (probe 33.6 restriction enzyme *Hae*II in table 1; Gilbert et al. 1990*b*); respective means and standard deviations of H are 0.076 ± 0.0068, 0.422 ± 0.0148, and 0.851 ± 0.0102, and respective means and standard deviations of APD/100 are 0.084 ± 0.0079, 0.431 ± 0.0141, and 0.756 ± 0.0125.

For convenience in comparing H with H_{max}, we decomposed the estimated L from equation (4) into L_M and $L-L_M$. Table 1 shows the estimated fraction of loci that are polymorphic [$P = 1-(L_M/L)$], corresponding to our standard estimate [eq. (4)], and corresponding to our maximum estimate [$P_{(max)} = L_P/(L_P+L_M)$; eq. (8)]. For many cases, even those of heterozygosity extremes, H is fairly close to H_{max}, which supports our claim that the heterozygosity estimate is reasonably insensitive to the missing genetic information.

Table 2 shows estimates of heterozygosity based on restriction-fragment pattern when cDNA probes from MHC class I genes in several species are used (Yuhki and O'Brien 1990). As with the heterozygosity estimates in table 1, H is generally close to H_{max}. In both tables we have also calculated the APD, which has been shown previously to correlate well with heterozygosities obtained from other genetic systems, such as allozymes (Yuhki and O'Brien 1990). Figure 2 shows the relationship of APD to our estimate of heterozygosity [eq. (5)] from the same data sets. In both cases, the correlation is quite good and highly significant, as predicted from theory (Lynch 1990).

We have run computer simulations to examine the distribution of H [eq. (5)], L [eq. (4)], H_{max} [eq. (7)], and APD [from eq. (14)] among replicate samples taken from the same underlying population. We chose three of our empirical data sets as models for the simulated populations. In these simulations, observed band frequencies were taken as population band frequencies, and all bands were assumed to be independent. Sample sizes of individuals in the simulations were taken as the sample sizes from the empirical data sets. The number of bands observed in a simulation varied because of sampling errors, in accordance with the following considerations:

Consider a band whose frequency is only $1/n$ in the population. The probability of not observing this band in a sample of n individuals is $[1-(1/n)]^n$, which is 0.3–0.4 for any sample size greater than 3. If the true number of bands with population frequency $1/n$ is $f_T(1)$, the observed number of such bands is expected to be $f_0(1) = \{1-[1-(1/n)]^n\} f_T(1)$, which is substantially less than $f_T(1)$. Generalizing this notion, our simulations set the number of bands with population frequency j/n to $f_T(j) = f_0(j)/\{1-[1-(j/n)]^n\}$, to allow for "loss" of such bands. The $f_0(j)$ were determined from each empirical data set being modeled in a simulation. The distributions of H and APD/100, for 1,000 replicates for each of the three model data sets, are shown in figure 3. In each case, the variance of H is equivalent to or perhaps even smaller than that for APD/100. It is interesting that H and APD/100 are not statistically significantly different for the first two samples but are clearly different for the sample of human cell lines. In the latter case, H values are substantially greater than APD/100 values, which is somewhat counterintuitive in light of the fact that the limiting (observable) situation of each band in frequency $1/n$ would give APD/100 = 1, but $H = 1-(1/n)$. However, Lynch's (1990) theoretical results indicate that the similarity index is an upwardly biased estimator of population homozygosity, which means that APD/100 could be expected to be a biased underestimate of average heterozygosity [see eq. (14)]. It is interesting that his results suggested that bias would be greatest when most alleles are at intermediate frequency, whereas our results suggest that this bias is greatest when most alleles are rare (in this case, 121 alleles are distributed among 9.4 loci).

Discussion

DNA fingerprints yield an extraordinarily rich picture of the genetic diversity in humans and in natural populations—hence their particular importance for demographic and other studies where close genetic relationship and even individual identity is being evaluated (Gill et al. 1985; Jeffreys et al. 1985*a*, 1985*b*, 1985*c*; Burke and Bruford 1987; Jeffreys and Morton 1987; Wetton et al. 1987; Burke et al. 1989; Gilbert et al. 1990*a*, 1990*b*, 1991; Kuhnlein et al. 1990; Reeve et al. 1990). An interpretive drawback to such studies stems from their complexity—the typically large number of alleles at an unknown number of loci, all envisioned on a single Southern blot, precludes rigorous ascertainment of relevant genetic information. As shown above, this lack of information need not undermine the estimation of at least one important parameter, the average heterozygosity of the component loci in the DNA fingerprint.

Comigration of distinct alleles as a single band will tend to make our heterozygosity estimates underestimates, since a band of moderate or high frequency would generally be replaced by bands of lower frequency. That other alleles at the loci in question may be missed—e.g., if they are above or below the scorable region of the gel—also looms as a potential reservation with regard to our estimation of heterozygosity. In an evaluation of DNA fingerprints of human cell lines, Gilbert et al. (1990*b*) addressed the importance of the missing alleles, by using an acrylamide gel that allowed resolution of the smaller band sizes. They found that the inclusion of about a dozen new bands of low molecular weight that were resolved on polyacrylamide sequencing gels made very little difference in the overall population genetic parameters (i.e., heterozygosity and probability of individual identification) calculated from this data set.

We can calculate that an additional band will increase the heterozygosity estimated by equation (6) only if its frequency is less than $1-H^2$ [equivalent to the x_k being less than $1-H$ in eq. (5)]. As in the paper by Gilbert et al. (1990*b*), additional bands can be gleaned from the smaller size range of bands (conventionally, <1–2 kb). Typically, this range is ignored or even run off the gel, since the smaller bands tend to be more monomorphic. One may challenge this protocol, under the contention that exclusion of such bands biases the heterozygosity upward. However, ascertainment of heterozygosity of every locus homologous to a specific probe would seem to be a rather idealistic goal. After all, many homologous loci will fail to be detected as a function of stringency, so the heterozygosity estimate must be viewed as an operational one. Thus, the estimated heterozygosity of a DNA fingerprint may be seen as a characteristic of the specific probe/enzyme combination but should not be separated from technical factors such as the stringency of the gel or the size range of resolvable bands. Rigorous comparisons among labs will thus necessitate comparable technical criteria such as the range of band sizes and stringency. In this perspective, DNA fingerprints are analogous to conventional types of molecular data (e.g., protein electrophoresis), in that they provide estimates of genomic diversity that are directly comparable to equivalently collected data, and are correlatable with other estimates of diversity (e.g., allozyme electrophoresis, RFLP typing, and DNA sequence variation) used to quantify genetic variation in populations.

We have presented here both an estimate of average heterozygosity, H, and maximum heterozygosity, H_{max}, by using multilocus hypervariable gene families. The estimate was applied to data sets from four species (table 1) by using minisatellite probes, as well as to several mammal species typed with class I MHC probes (table

2). In many cases H approaches H_{max} and correlates well with other molecular estimates of genomic diversity that have been reported elsewhere (Gilbert et al. 1990*a*, 1990*b*, 1991; Yuhki and O'Brien 1990).

Acknowledgments

We would like to thank Drs. M. Gail, Des Cooper, and Marilyn Raymond for their comments on an earlier version of the manuscript. We would also like to thank Matt Fivash for advice and interest in the statistical aspects of our estimates. This project has been funded at least in part with Federal funds from the Department of Health and Human Services under contract number N01-CO-74102 with Program Resources, Inc. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

LITERATURE CITED

BAIRD, M., I. BALAZS, A. GIUSTI, L. MIYAZAKI, L. NICHOLAS, K. WEXLER, E. KANTER, J. GLASSBERG, F. ALLEN, P. RUBINSTEIN, and L. SUSSMAN. 1986. Allele frequency distribution of two highly polymorphic DNA sequences in three ethnic groups and its application to the determination of paternity. Am. J. Hum. Genet. **39**:489–501.

BODMER, W., and L. L. CAVALLI-SFORZA. 1976. Genetics, evolution, and man. W. H. Freeman, San Francisco.

BURKE, T., and M. W. BRUFORD. 1987. DNA fingerprinting in birds. Nature **327**:149–152.

BURKE, T., N. B. DAVIES, M. W. BRUFORD, and B. J. HATCHWELL. 1989. Parental care and mating behaviour of polyandrous dunnocks *Prunella modularis* related to paternity by DNA fingerprinting. Nature **338**:249–251.

GILBERT, D. A., N. LEHMAN, S. J. O'BRIEN, and R. K. WAYNE. 1990*a*. Genetic fingerprinting reflects population differentiation in the California Channel Island fox. Nature **344**:764–767.

GILBERT, D. A., C. PACKER, A. E. PUSEY, J. C. STEPHENS, and S. J. O'BRIEN. 1991. Analytical DNA fingerprinting in lions: parentage, genetic diversity, and kinship. J. Hered. **82**:378–386.

GILBERT, D. A., Y. A. REID, M. H. GAIL, D. PEE, C. WHITE, R. J. HAY, and S. J. O'BRIEN. 1990*b*. Application of DNA fingerprints for cell-line individualization. Am. J. Hum. Genet. **47**:499–514.

GILL, P., A. J. JEFFREYS, and D. J. WERRETT. 1985. Forensic application of DNA "fingerprints." Nature **318**:577–579.

HARTL, D. L. 1980. Principles of population genetics. Sinauer, Sunderland, Mass.

HILL, W. G. 1974. Estimation of linkage disequilibrium in randomly mating populations. Heredity **33**:229–239.

JEFFREYS, A. J. 1987. Highly variable minisatellites and DNA fingerprints. Biochem. Soc. Trans. **15**:309–317.

JEFFREYS, A. J., J. F. Y. BROOKFIELD, and R. SEMEONOFF. 1985*c*. Positive identification of an immigration test-case using human DNA fingerprints. Nature **317**:818–819.

JEFFREYS, A. J., and D. B. MORTON. 1987. DNA fingerprints of dogs and cats. Anim. Genet. **18**:1–15.

JEFFREYS, A. J., R. NEUMANN, and V. WILSON. 1990. Repeat unit sequence variation in minisatellites: a novel source of DNA polymorphism for studying variation and mutation by single molecule analysis. Cell **60**:473–485.

JEFFREYS, A. J., V. WILSON, R. KELLY, B. A. TAYLOR, and G. BULFIELD. 1987. Mouse DNA

'fingerprints': analysis of chromosome localization and germ-line stability of hypervariable loci in recombinant inbred strains. Nucleic Acids Res. **15**:2823–2836.

JEFFREYS, A. J., V. WILSON, and S. L. THEIN. 1985*a*. Hypervariable minisatellite regions in human DNA. Nature **314**:67–73.

———. 1985*b*. Individual-specific fingerprints of human DNA. Nature **316**:76–79.

KUHNLEIN, U., D. ZADWORNY, Y. DAWE, R. W. FAIRFULL, and J. S. GAVORA. 1990. Assessment of inbreeding by DNA fingerprinting: development of a calibration curve using defined strains of chickens. Genetics **125**:161–165.

LYNCH, M. 1988. Estimation of relatedness by DNA fingerprinting. Mol. Biol. Evol. **5**:584–599.

———. 1990. The similarity index and DNA fingerprinting. Mol. Biol. Evol. **7**:478–484.

NAKAMURA, Y., M. CARLSON, K. KRAPCHO, M. KANAMORI, and R. WHITE. 1988. New approach for isolation of VNTR markers. Am. J. Hum. Genet. **43**:854–859.

NAKAMURA, Y., M. LEPPERT, P. O'CONNELL, R. WOLFF, T. HOLM, M. CULVER, C. MARTIN, E. FUJIMOTO, M. HOFF, E. KUMLIN, and R. WHITE. 1987. Variable number of tandem repeat (VNTR) markers for human gene mapping. Science **235**:1616–1622.

NEI, M. 1973. Analysis of gene diversity in subdivided populations. Proc. Natl. Acad. Sci. USA **70**:3321–3323.

———. 1987. Molecular evolutionary genetics. Columbia University Press, New York.

NEI, M., and A. K. ROYCHOUDHURY. 1974. Sampling variances of heterozygosity and genetic distance. Genetics **76**:379–390.

REEVE, H. K., D. F. WESTNEAT, W. A. NOON, P. W. SHERMAN, and C. F. AQUADRO. 1990. DNA "fingerprinting" reveals high levels of inbreeding in colonies of the eusocial naked mole-rat. Proc. Natl. Acad. Sci. USA **87**:2496–2500.

WETTON, J. H., R. E. CARTER, D. T. PARKIN, and D. WALTERS. 1987. Demographic study of a wild house sparrow population by DNA fingerprinting. Nature **327**:147–149.

WONG, Z., V. WILSON, A. JEFFREYS, and S. THEIN. 1986. Cloning a selected fragment from a human DNA "fingerprint": isolation of an extremely polymorphic minisatellite. Nucleic Acids Res. **14**:4605–4616.

WONG, Z., V. WILSON, I. PATEL, S. POVEY, and A. J. JEFFREYS. 1987. Characterization of a panel of highly variable minisatellites cloned from human DNA. Ann. Hum. Genet. **51**: 269–288.

YUHKI, N., and S. J. O'BRIEN. 1990. DNA variation of the mammalian major histocompatibility complex reflects genomic diversity and population history. Proc. Natl. Acad. Sci. USA **87**: 836–840.

MASATOSHI NEI, reviewing editor

Received July 25, 1991; revision received January 8, 1992

Accepted January 8, 1992