RESEARCH ARTICLES

An Unbiased Estimator of Gene Diversity in Samples Containing Related Individuals

Michael DeGiorgio and Noah A. Rosenberg*†*

*Center for Computational Medicine and Biology, University of Michigan; and †Department of Human Genetics and the Life Sciences Institute, University of Michigan

Gene diversity is sometimes estimated from samples that contain inbred or related individuals. If inbred or related individuals are included in a sample, then the standard estimator for gene diversity produces a downward bias caused by an inflation of the variance of estimated allele frequencies. We develop an unbiased estimator for gene diversity that relies on kinship coefficients for pairs of individuals with known relationship and that reduces to the standard estimator when all individuals are noninbred and unrelated. Applying our estimator to data simulated based on allele frequencies observed for microsatellite loci in human populations, we find that the new estimator performs favorably compared with the standard estimator in terms of bias and similarly in terms of mean squared error. For human population-genetic data, we find that a close linear relationship previously seen between gene diversity and distance from East Africa is preserved when adjusting for the inclusion of close relatives.

Introduction

Gene diversity, or expected heterozygosity, is a frequently used measure of genetic variation applied in diverse areas of population genetics. Together with its counterpart, gene identity or expected homozygosity, it has been used to quantify genetic variation in populations (Driscoll et al. 2002; Hoelzel et al. 2002), evaluate genetic divergence and population relationships (Nei 1973; Ramachandran et al. 2005), detect inbreeding (Li and Horvitz 1953), measure linkage disequilibrium (Ohta 1980; Sabatti and Risch 2002), and test for the influence of natural selection (Watterson 1978; Depaulis and Veuille 1998; Sabeti et al. 2002).

Consider a polymorphic locus with I distinct alleles and a population with parametric allele frequencies $p_1, p_2, \ldots, p_I$, where $p_i \in [0, 1]$ and $\sum_{i=1}^{I} p_i = 1$. The term "gene diversity," which is defined as

$$H = 1 - \sum_{i=1}^{I} p_i^2, \qquad (1)$$

was proposed by Nei (1973), though the use of equation (1) as a measure of diversity dates to considerably earlier (Gini 1912; Simpson 1949; Gibbs and Martin 1962).

Now consider a sample of n observations of alleles, in which the number of observations of allelic type i is n_i. The count estimate of p_i is $\hat{p}_i = n_i/n$. If no inbred or related individuals are included in the sample, then an unbiased estimator of gene diversity is (Nei and Roychoudhury 1974)

$$\hat{H} = \frac{n}{n-1}\left(1 - \sum_{i=1}^{I} \hat{p}_i^2\right). \qquad (2)$$

If relatives or inbred individuals are included in the sample, then $\hat{H}$ is no longer an unbiased estimator of H. To understand why this statement is true, suppose that a sample contains a pair of close relatives. Because these individuals are related, they may share one or two alleles identically by descent (IBD) at a locus (compared with zero alleles shared IBD in unrelated individuals). As a result, estimation of p_i is based on fewer independent observations than for a sample not containing any relatives. Although $E[\hat{p}_i] = p_i$ when relatives are included, $\mathrm{Var}[\hat{p}_i]$ is greater than it would be had no relatives been included. Observe that the computation of $E[\hat{H}]$ involves a negative coefficient for $E[\hat{p}_i^2]$. Because $E[\hat{p}_i^2] = \mathrm{Var}[\hat{p}_i] + E[\hat{p}_i]^2$, $E[\hat{H}]$ decreases as $\mathrm{Var}[\hat{p}_i]$ increases. Thus, the inclusion of relatives results in a downward bias, so that $E[\hat{H}] < H$. For the case in which inbred unrelated individuals with known inbreeding coefficients are included in a sample, Weir (1989, 1996) provided the expectation of $1 - \sum_{i=1}^{I} \hat{p}_i^2$, producing an unbiased estimator of gene diversity

$$\hat{H}_{\mathrm{Weir}} = \frac{n}{n-1-\bar{f}}\left(1 - \sum_{i=1}^{I} \hat{p}_i^2\right), \qquad (3)$$

where $\bar{f}$ is the average inbreeding coefficient across individuals (see also Shete 2003). When inbred individuals are included, $\bar{f} \neq 0$, and it follows that $E[\hat{H}] < E[\hat{H}_{\mathrm{Weir}}] = H$.

In this article, we conduct a detailed investigation of the case in which a sample includes related individuals. We derive an unbiased estimator of H for samples containing related individuals with known levels of relationship. Our derivation makes use of a formula of Bourgain et al. (2003) and McPeek et al. (2004) for the variance of count estimates of allele frequencies in samples containing inbred and related individuals. The resulting estimator incorporates kinship coefficients, the same quantitative descriptors of pairwise relationships that have been used in diverse problems involving relatives—such as evaluation of phenotypic covariances in families (Lange 2002), estimation of relatedness parameters (Weir et al. 2006), and quantitative-trait linkage analysis (Almasy and Blangero 1998). When a sample consists only of unrelated noninbred individuals, our new estimator $\tilde{H}$ reduces to the standard estimator $\hat{H}$, and it reduces to $\hat{H}_{\mathrm{Weir}}$ if inbred but not related individuals are included. Using data simulated based on allele frequencies from human populations, we find that the new estimator $\tilde{H}$ corrects for bias generated by inclusion of related individuals and that it attains a mean squared error (MSE) comparable with that of $\hat{H}$. We apply this new estimator to microsatellite data from human population samples containing relatives and show that, compared with the

Key words: heterozygosity, identity by descent, kinship coefficient.

E-mail: degiormi@umich.edu.

Mol. Biol. Evol. 26(3):501–512. 2009
doi:10.1093/molbev/msn254
Advance Access publication November 6, 2008

standard estimator, it produces estimates closer to those obtained when excluding relatives from the analysis.

Theory

We assume that gene diversity is estimated from $n/2$ diploid individuals. Our aim is to obtain a bias-correction factor that can be incorporated into a new estimator of gene diversity, $\tilde{H}$. We begin by computing $\mathrm{Var}[\hat{p}_i]$ in a sample that may include relatives or inbred individuals. $\mathrm{Var}[\hat{p}_i]$ was reported by Bourgain et al. (2003) and McPeek et al. (2004); we provide an alternative derivation that uses a generalization of the simpler method of Broman (2001). This approach was originally applied in a setting that did not consider inbreeding, and we generalize the computation to include inbreeding. Note that the variances of other estimators of allele frequencies have previously been derived in fairly general settings (McPeek et al. 2004) and that the estimator $\hat{p}_i$ is not a maximum likelihood estimator when related individuals are included in a sample (Boehnke 1991). However, our interest here is specifically on the count-based estimator of allele frequencies, as it is this estimator that is used in the standard estimator of gene diversity in equation (2).

Define X_k to be the number of alleles of type i that are carried by individual k at a particular locus. X_k can equal 0, 1, or 2, and $E[X_k] = 2p_i$. Regardless of the relationships among individuals 1, 2, …, $n/2$, an unbiased estimator for p_i, the frequency of allele i, is

$$\hat{p}_i = \frac{1}{n}\sum_{k=1}^{n/2} X_k. \tag{4}$$

The variance of $\hat{p}_i$ is given by

$$\mathrm{Var}[\hat{p}_i] = \frac{1}{n^2}\sum_{j=1}^{n/2}\sum_{k=1}^{n/2}\mathrm{Cov}\left[X_j, X_k\right]. \tag{5}$$

Suppose that individuals j and k are related. The coefficient of kinship between individuals j and k, $\Phi_{j,k}$, is the probability that two alleles chosen at the locus—one from individual j and the other from individual k—are identical by descent. In the special case of $j = k$, the kinship coefficient is $\Phi_{k,k} = (1/2)(1 + f_k)$, where f_k is the inbreeding coefficient for individual k (Lange 2002, p. 81).

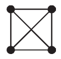
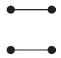
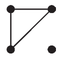
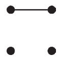
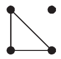
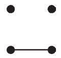
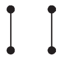
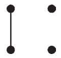
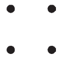
Conditional on the nature of the relationship between individuals j and k and on their inbreeding coefficients, the four alleles in the two individuals can take on one of nine condensed identity states (Jacquard 1974, p. 107). Let $\Delta_s = \Pr[S = s]$, where the condensed identity state S ranges from 1 to 9 and the probability is conditional on the type of the relationship. Using table 1 and the fact that the kinship coefficient for the pair of individuals equals $\Delta_1 + (1/2)(\Delta_3 + \Delta_5 + \Delta_7) + (1/4)\Delta_8$ (Jacquard, 1974, p. 108), we obtain

$$E[X_jX_k] = \sum_{a=0}^{2}\sum_{b=0}^{2}\sum_{s=1}^{9} ab\Delta_s \Pr[X_j = a, X_k = b | S = s]$$

$$= 4\Phi_{j,k}p_i(1 - p_i) + 4p_i^2.$$

Because $E[X_j] = E[X_k] = 2p_i$, it follows that

Table 1
Joint Distribution of the Numbers of *i* Alleles Carried by Individuals *j* and *k* Given Their Descent Configuration *S*, Assuming Allele *i* Has Frequency *p*

S	Condensed Identity State[a]	X_j, X_k	$\Pr[X_j, X_k \mid S]$
1		0, 0	$1 - p$
		2, 2	p
2		0, 0	$(1 - p)^2$
		0, 2	$p(1 - p)$
		2, 0	$p(1 - p)$
		2, 2	p^2
3		0, 0	$(1 - p)^2$
		0, 1	$p(1 - p)$
		2, 1	$p(1 - p)$
		2, 2	p^2
4		0, 0	$(1 - p)^3$
		0, 1	$2p(1 - p)^2$
		0, 2	$p^2(1 - p)$
		2, 0	$p(1 - p)^2$
		2, 1	$2p^2(1 - p)$
		2, 2	p^3
5		0, 0	$(1 - p)^2$
		1, 0	$p(1 - p)$
		1, 2	$p(1 - p)$
		2, 2	p^2
6		0, 0	$(1 - p)^3$
		0, 2	$p(1 - p)^2$
		1, 0	$2p(1 - p)^2$
		1, 2	$2p^2(1 - p)$
		2, 0	$p^2(1 - p)$
		2, 2	p^3
7		0, 0	$(1 - p)^2$
		1, 1	$2p(1 - p)$
		2, 2	p^2
8		0, 0	$(1 - p)^3$
		0, 1	$p(1 - p)^2$
		1, 0	$p(1 - p)^2$
		1, 1	$p(1 - p)$
		1, 2	$p^2(1 - p)$
		2, 1	$p^2(1 - p)$
		2, 2	p^3
9		0, 0	$(1 - p)^4$
		0, 1	$2p(1 - p)^3$
		0, 2	$p^2(1 - p)^2$
		1, 0	$2p(1 - p)^3$
		1, 1	$4p^2(1 - p)^2$
		1, 2	$2p^3(1 - p)$
		2, 0	$p^2(1 - p)^2$
		2, 1	$2p^3(1 - p)$
		2, 2	p^4

[a] The first row of dots represents the two alleles for individual j, and the second row represents the two alleles for individual k. Two alleles are identical by descent if there is a line connecting them.

$$\mathrm{Cov}\left[X_j, X_k\right] = E\left[X_jX_k\right] - E\left[X_j\right]E\left[X_k\right] = 4\Phi_{j,k}p_i\left(1 - p_i\right). \tag{6}$$

Inserting the covariance into equation (5) yields

$$\mathrm{Var}[\hat{p}_i] = \frac{4p_i(1 - p_i)}{n^2}\sum_{j=1}^{n/2}\sum_{k=1}^{n/2}\Phi_{j,k} = \bar{\Phi}p_i(1 - p_i), \tag{7}$$

Table 2
The 26 Populations Containing Relatives in the H1048 Data Set (Modified from Rosenberg 2006, Supplementary tables 16 and 19)

Population	Geographic Region	Number of Sampled Individuals	Number of Parent–Offspring Pairs	Number of Full-Sib Pairs	Number of Second-Degree Pairs
Bantu (Kenya)	Africa	12	0	1	0
Biaka Pygmy	Africa	32	4	2	7
Mandenka	Africa	24	0	0	2
Mbuti Pygmy	Africa	15	2	0	1
San	Africa	7	1	0	0
Yoruba	Africa	25	2	2	0
French	Europe	29	0	1	0
Orcadian	Europe	16	1	0	0
Bedouin	Middle East	48	1	0	1
Druze	Middle East	47	1	2	2
Mozabite	Middle East	30	0	1	0
Palestinian	Middle East	51	0	1	5
Balochi	Central/South Asia	25	0	1	0
Hazara	Central/South Asia	24	0	1	1
Kalash	Central/South Asia	25	1	0	1
Sindhi	Central/South Asia	25	1	0	0
Cambodian	East Asia	11	1	0	0
Lahu	East Asia	10	1	1	0
Naxi	East Asia	10	0	1	0
Oroqen	East Asia	10	0	1	0
Melanesian	Oceania	19	9	3	2
Colombian	America	13	6	1	0
Karitiana	America	24	6	6	0
Maya	America	25	2	1	2
Pima	America	25	15	6	10
Surui	America	21	15	14	0

where $\bar{\Phi}=\frac{1}{(n/2)^2}\sum_{j=1}^{n/2}\sum_{k=1}^{n/2}\Phi_{j,k}$ is the average kinship coefficient across pairs of individuals (including comparisons of individuals with themselves). This result can be seen to be equivalent to the variance reported by McPeek et al. (2004, p. 361).

Proposition 1

Consider a locus with I distinct alleles, allele frequencies $p_i \in [0, 1]$ and $\sum_{i=1}^{I} p_i=1$. Suppose a sample from a population has $n/2$ possibly related and inbred individuals. Then an unbiased estimator for gene diversity is

$$\tilde{H} = \frac{1}{1 - \bar{\Phi}}\left(1 - \sum_{i=1}^{I}\hat{p}_i^2\right), \tag{8}$$

where $\Phi_{j,k}$ is the kinship coefficient of individuals j and k and $\bar{\Phi}=\frac{1}{(n/2)^2}\sum_{j=1}^{n/2}\sum_{k=1}^{n/2}\Phi_{j,k}$ is the average kinship coefficient across pairs of individuals.

Proof

We need to show that $E[\tilde{H}]=H$. Observing that $E[\hat{p}_i^2]=\mathrm{Var}[\hat{p}_i]+E[\hat{p}_i]^2$ and $E[\hat{p}_i]=p_i$, we apply equation (4) and then the variance of $\hat{p}_i$ in equation (7) to get

$$\begin{aligned} E[\tilde{H}] &= \frac{1}{1-\bar{\Phi}}\left[1 - \sum_{i=1}^{I}\left(\mathrm{Var}[\hat{p}_i] + p_i^2\right)\right] \\ &= \frac{1}{1-\bar{\Phi}}\left[1 - \sum_{i=1}^{I}\left(\bar{\Phi}p_i(1-p_i) + p_i^2\right)\right] = H. \quad \square \end{aligned}$$

Corollary 2

Consider a locus with I distinct alleles, allele frequencies $p_i \in [0, 1]$ and $\sum_{i=1}^{I} p_i=1$. Suppose a sample from a population has $n/2$ possibly related and inbred individuals. Let $\mathcal{R}$ be the set of distinct types of relative pairs in the sample. Further, let n_R be the number of pairs of individuals with relationship type $R \in \mathcal{R}$ and let Φ_R be the kinship coefficient for each of these pairs. Then an unbiased estimator for gene diversity is

$$\tilde{H} = \frac{n(n-1)}{n\left(n-1-\bar{f}\right) - 8\sum_{R\in\mathcal{R}} n_R\Phi_R}\hat{H}, \tag{9}$$

where $\bar{f}=\frac{1}{n/2}\sum_{k=1}^{n/2} f_k$ is the average inbreeding coefficient across individuals and f_k is the inbreeding coefficient for individual k.

Proof

Applying the definitions of $\bar{\Phi}$ and $\Phi_{k,k}$ and the fact that $\Phi_{j,k} = 0$ for a pair of "unrelated" individuals,

$$\begin{aligned} \bar{\Phi} &= \frac{1}{(n/2)^2}\sum_{j=1}^{n/2}\sum_{k=1}^{n/2}\Phi_{j,k} = \frac{4}{n^2}\left(\sum_{k=1}^{n/2}\Phi_{k,k} + 2\sum_{j=1}^{n/2}\sum_{k=j+1}^{n/2}\Phi_{j,k}\right) \\ &= \frac{1}{n^2}\left(n + n\bar{f} + 8\sum_{R\in\mathcal{R}} n_R\Phi_R\right). \end{aligned}$$

Inserting this value for $\bar{\Phi}$ into equation (8), we obtain the desired result. □

Table 3
MSE, Variance, and Bias Squared of Estimates for Data Simulated Based on Allele Frequencies at Two Loci (AAT263P and ACT3F12)

			AAT263P ($H = 0.6778$)			ACT3F12 ($H = 0.8263$)		
m	(q, r, s)	Estimator	MSE	Variance	Bias^2	MSE	Variance	Bias^2
10	(2, 0, 0)	$\hat{H}_{\text{full}}$	**9.196×10^{-3}**	**9.141×10^{-3}**	5.454×10^{-5}	3.774×10^{-3}	**3.694×10^{-3}**	7.923×10^{-5}
		$\tilde{H}_{\text{full}}$	9.337×10^{-3}	9.337×10^{-3}	**6.387×10^{-8}**	**3.773×10^{-3}**	3.773×10^{-3}	**4.249×10^{-8}**
		$\hat{H}_{\text{reduced}}$	9.911×10^{-3}	9.911×10^{-3}	9.160×10^{-8}	4.034×10^{-3}	4.034×10^{-3}	1.110×10^{-7}
	(2, 2, 0)	$\hat{H}_{\text{full}}$	**1.084×10^{-2}**	**1.064×10^{-2}**	2.047×10^{-4}	4.692×10^{-3}	**4.390×10^{-3}**	3.020×10^{-4}
		$\tilde{H}_{\text{full}}$	1.110×10^{-2}	1.110×10^{-2}	**1.542×10^{-9}**	**4.581×10^{-3}**	4.581×10^{-3}	2.562×10^{-10}
		$\hat{H}_{\text{reduced}}$	1.385×10^{-2}	1.385×10^{-2}	6.957×10^{-9}	5.899×10^{-3}	5.899×10^{-3}	**1.595×10^{-10}**
	(2, 0, 2)	$\hat{H}_{\text{full}}$	**9.885×10^{-3}**	**9.777×10^{-3}**	1.078×10^{-4}	4.236×10^{-3}	**4.066×10^{-3}**	1.706×10^{-4}
		$\tilde{H}_{\text{full}}$	1.009×10^{-2}	1.009×10^{-2}	1.048×10^{-7}	**4.197×10^{-3}**	4.197×10^{-3}	**1.855×10^{-10}**
		$\hat{H}_{\text{reduced}}$	1.363×10^{-2}	1.363×10^{-2}	**5.363×10^{-8}**	5.839×10^{-3}	5.839×10^{-3}	3.014×10^{-9}
20	(5, 2, 2)	$\hat{H}_{\text{full}}$	**5.107×10^{-3}**	**5.054×10^{-3}**	5.273×10^{-5}	2.030×10^{-3}	**1.959×10^{-3}**	7.060×10^{-5}
		$\tilde{H}_{\text{full}}$	5.160×10^{-3}	5.160×10^{-3}	**9.794×10^{-8}**	**2.000×10^{-3}**	2.000×10^{-3}	**5.322×10^{-9}**
		$\hat{H}_{\text{reduced}}$	6.929×10^{-3}	6.929×10^{-3}	6.236×10^{-7}	2.736×10^{-3}	2.736×10^{-3}	6.622×10^{-9}
	(5, 0, 0)	$\hat{H}_{\text{full}}$	**4.553×10^{-3}**	**4.535×10^{-3}**	1.788×10^{-5}	1.768×10^{-3}	**1.739×10^{-3}**	2.916×10^{-5}
		$\tilde{H}_{\text{full}}$	4.593×10^{-3}	4.593×10^{-3}	**1.365×10^{-8}**	**1.762×10^{-3}**	1.762×10^{-3}	1.086×10^{-8}
		$\hat{H}_{\text{reduced}}$	4.941×10^{-3}	4.941×10^{-3}	4.670×10^{-8}	1.913×10^{-3}	1.913×10^{-3}	**3.941×10^{-9}**
	(2, 5, 2)	$\hat{H}_{\text{full}}$	**5.092×10^{-3}**	**5.043×10^{-3}**	4.935×10^{-5}	2.048×10^{-3}	**1.975×10^{-3}**	7.219×10^{-5}
		$\tilde{H}_{\text{full}}$	5.148×10^{-3}	5.148×10^{-3}	**5.843×10^{-9}**	**2.016×10^{-3}**	2.016×10^{-3}	5.047×10^{-10}
		$\hat{H}_{\text{reduced}}$	6.948×10^{-3}	6.948×10^{-3}	5.923×10^{-9}	2.755×10^{-3}	2.755×10^{-3}	**1.884×10^{-11}**
30	(15, 0, 0)	$\hat{H}_{\text{full}}$	**3.580×10^{-3}**	**3.548×10^{-3}**	3.233×10^{-5}	1.396×10^{-3}	**1.346×10^{-3}**	4.973×10^{-5}
		$\tilde{H}_{\text{full}}$	3.609×10^{-3}	3.609×10^{-3}	3.411×10^{-9}	**1.370×10^{-3}**	1.370×10^{-3}	2.490×10^{-9}
		$\hat{H}_{\text{reduced}}$	4.924×10^{-3}	4.924×10^{-3}	**2.990×10^{-10}**	1.903×10^{-3}	1.903×10^{-3}	**2.346×10^{-9}**
	(5, 5, 5)	$\hat{H}_{\text{full}}$	**3.370×10^{-3}**	**3.345×10^{-3}**	2.464×10^{-5}	1.294×10^{-3}	**1.260×10^{-3}**	3.525×10^{-5}
		$\tilde{H}_{\text{full}}$	3.393×10^{-3}	3.393×10^{-3}	3.169×10^{-8}	**1.278×10^{-3}**	1.278×10^{-3}	**2.062×10^{-9}**
		$\hat{H}_{\text{reduced}}$	4.930×10^{-3}	4.930×10^{-3}	**1.154×10^{-8}**	1.890×10^{-3}	1.890×10^{-3}	2.397×10^{-8}
	(0, 5, 5)	$\hat{H}_{\text{full}}$	**2.970×10^{-3}**	**2.962×10^{-3}**	7.105×10^{-6}	1.122×10^{-3}	**1.110×10^{-3}**	1.181×10^{-5}
		$\tilde{H}_{\text{full}}$	2.988×10^{-3}	2.988×10^{-3}	**4.302×10^{-8}**	**1.119×10^{-3}**	1.119×10^{-3}	4.230×10^{-9}
		$\hat{H}_{\text{reduced}}$	3.623×10^{-3}	3.623×10^{-3}	4.632×10^{-8}	1.369×10^{-3}	1.369×10^{-3}	**2.294×10^{-9}**

Sample size is indicated by m, and q, r, and s represent the numbers of parent–offspring, full-sib, and second-degree pairs, respectively. Each value is based on 100,000 simulated data sets, and the same simulated data sets were used for all estimators and for all three quantities (MSE, variance, bias squared). We use $\hat{H}_{full}$ and $\tilde{H}_{full}$ to denote $\hat{H}$ and $\tilde{H}$ applied to a sample of m individuals. For $\hat{H}$ applied to a sample of m individuals in which $q + r + s$ related individuals are removed to create a sample of $m - q - r - s$ individuals, we use the notation $\hat{H}_{\text{reduced}}$. Boldface type indicates the estimator with the smallest MSE, variance, or bias squared.

Note that if no related individuals are included in the sample, then $\mathcal{R}$ is the empty set, thus reducing $\tilde{H}$ to $\hat{H}_{\text{Weir}}$; if additionally no inbred individuals are included, then $\bar{f}=0$ and $\tilde{H}$ reduces to $\hat{H}$.

Corollary 3

Consider a locus with I distinct alleles, allele frequencies $p_i \in [0, 1]$ and $\sum_{i=1}^{I} p_i=1$. Suppose a sample from a population has $n/2$ noninbred individuals, among which q parent–offspring pairs, r full-sib pairs, and s second-degree (avuncular, grandparent–grandchild, and half-sib) relative pairs are included. Assuming the sample has no other relative pairs, an unbiased estimator for gene diversity is

$$\tilde{H} = \frac{n(n-1)}{n(n-1) - 2q - 2r - s}\hat{H}. \qquad (10)$$

Proof

The kinship coefficients are $\Phi_{\text{P}} = 1/4$ for parent–offspring pairs, $\Phi_{\text{F}} = 1/4$ for full-sib pairs, and $\Phi_{\text{S}} = 1/8$ for second-degree pairs. If an individual k is not inbred, then $f_k = 0$. For a sample without inbred individuals, $\bar{f}=0$. Inserting the quantity and kinship coefficient for each of the three types of relative pairs into equation (9), we obtain equation (10). □

Corollary 4

Consider a locus with I distinct alleles, allele frequencies $p_i \in [0, 1]$ and $\sum_{i=1}^{I} p_i=1$. Suppose a sample from a population has $n/2$ possibly related and inbred individuals. Let $\mathcal{R}$ be the set of distinct types of relative pairs in the sample. Further, let n_R be the number of pairs of individuals with relationship type $R \in \mathcal{R}$ and let Φ_R be the kinship coefficient for each of these pairs. Then the bias of $\hat{H}$ is always negative, increases in magnitude as H increases, and is given by

$$\text{bias}(\hat{H}) = -\frac{n\bar{f} + 8\sum_{R\in\mathcal{R}} n_R \Phi_R}{n(n-1)}H, \qquad (11)$$

where $\bar{f}=\frac{1}{n/2}\sum_{k=1}^{n/2} f_k$ is the average inbreeding coefficient across individuals and f_k is the inbreeding coefficient for individual k.

Proof

As shown in Corollary 2, $\tilde{H}=c\hat{H}$, where $c=n(n-1)/[n(n-1-\bar{f}) - 8\sum_{R\in\mathcal{R}} n_R\Phi_R]$. Rearranging and taking the expected value gives $E[\hat{H}]=E[\tilde{H}]/c=H/c$. The desired result follows from simplifying the expression for bias$(\hat{H})$, or $(1-c)H/c$. □

Data from Human Populations

To examine the behavior of $\tilde{H}$ in a realistic setting, we performed simulations and data analysis using microsatellite loci from the H1048 and H952 subsets (Rosenberg 2006) of the Human Genome Diversity Project–Centre d'Etude du Polymorphisme Humain (HGDP–CEPH) Cell

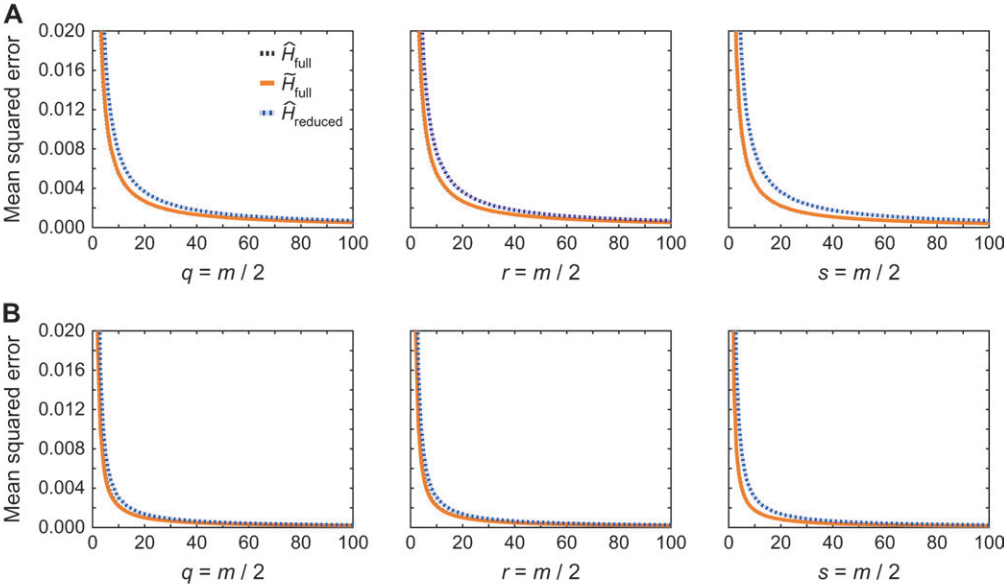


FIG. 1.—MSE as a function of sample size m for three different estimators. Each plot in a given row represents samples with a different type of relative pair. The numbers of parent–offspring, full-sib, and second-degree pairs are denoted by q, r, and s, respectively. The full and reduced samples contain m and $m/2$ individuals, respectively. The $\tilde{H}_{\text{full}}$ curve is almost directly on top of the $\hat{H}_{\text{full}}$ curve. (*A*) Allele frequencies simulated based on observed frequencies at locus AAT263P ($H = 0.6778$). (*B*) Allele frequencies simulated based on observed frequencies at locus ACT3F12 ($H = 0.8263$). The range of the plots is truncated at 0.02, so that the MSE for small sample sizes is not shown. Each point in the graphs is based on 100,000 simulated data sets, and the same simulated data were used for all three estimators.

Line Panel (Cann et al. 2002; Cavalli-Sforza 2005). The H1048 subset consists of 1,048 individuals in 53 populations. Among the 53 populations, the samples from 26 of them contain at least one pair of closely related individuals with either a first-degree (parent–offspring, full-sib) or second-degree (avuncular, grandparent–grandchild, and half-sib) relationship (table 2). The H952 subset is a collection of 952 individuals included in the larger H1048 subset. No two of the 952 individuals are believed to have a first- or second-degree relationship. Levels of relationship in H1048, as estimated previously from microsatellite genotypes (Rosenberg 2006), were treated here as known with certainty. Because no cycles were observed in pedigrees from the HGDP–CEPH panel (Rosenberg 2006), we assumed that none of the panel members were inbred. Genotypes at 783 autosomal microsatellite loci (Ramachandran et al. 2005; Rosenberg et al. 2005) were investigated in the H1048 and H952 data sets.

Simulations

Simulation Procedure

Simulations based on the microsatellite loci were used to examine the properties of $\tilde{H}$ and $\hat{H}$. For each of the 783 loci, we treated allele frequencies estimated from the H952 subset of individuals as true allele frequencies. The parametric gene diversity H was obtained for a locus as one minus the sum of the squares of these allele frequencies. All of our simulations assumed no inbreeding.

For a given locus, individual genotypes were simulated by sampling two alleles independently from the allele frequency distribution. To simulate a related individual with a given level of relationship to another individual, the number of alleles shared IBD with its relative was drawn under the appropriate probability distribution for the specified type of relative pair (parent–offspring, full-sib, or second-degree). This number of shared alleles (0, 1, or 2) was copied from a random individual that had already been generated and that had not yet been paired with a relative; if the number of alleles copied was 1, then an allele was chosen at random from the previously generated individual. The rest of the alleles, if any, were sampled independently from the allele frequency distribution. Gene diversity was estimated using $\tilde{H}$ and $\hat{H}$ for samples with and without related individuals. We applied $\hat{H}$ both to entire samples as well to samples in which the "second" member of each relative pair was discarded. For each locus, simulated sets of individuals were obtained 100,000 times, and $\hat{H}$, $\tilde{H}$, $\hat{H}^2$, and $\tilde{H}^2$ were averaged across all replicates. The true value for gene diversity, H, was then subtracted from the mean of $\hat{H}$ and $\tilde{H}$ to calculate bias for each estimator (and the result was squared to give bias squared). Variance of $\hat{H}$ was calculated by subtracting the square of the mean of $\hat{H}$ from the mean of $\hat{H}^2$ (variance of $\tilde{H}$ was calculated analogously). MSE was then calculated by adding variance and bias squared. Note that in our simulations, relative pairs were all disjoint, so that no individual was contained in multiple relative pairs; however, in our derivations, it is not required for relative pairs to be disjoint for $\tilde{H}$ to be unbiased.

Simulation Results

To illustrate the performance of the estimators across the span of gene diversities present in the human microsatellite data set, loci were placed in increasing order by assumed parametric gene diversity, and six equally spaced loci—with the 112th, 224th, 336th, 448th, 560th, and 672nd highest values of gene diversity—were chosen for

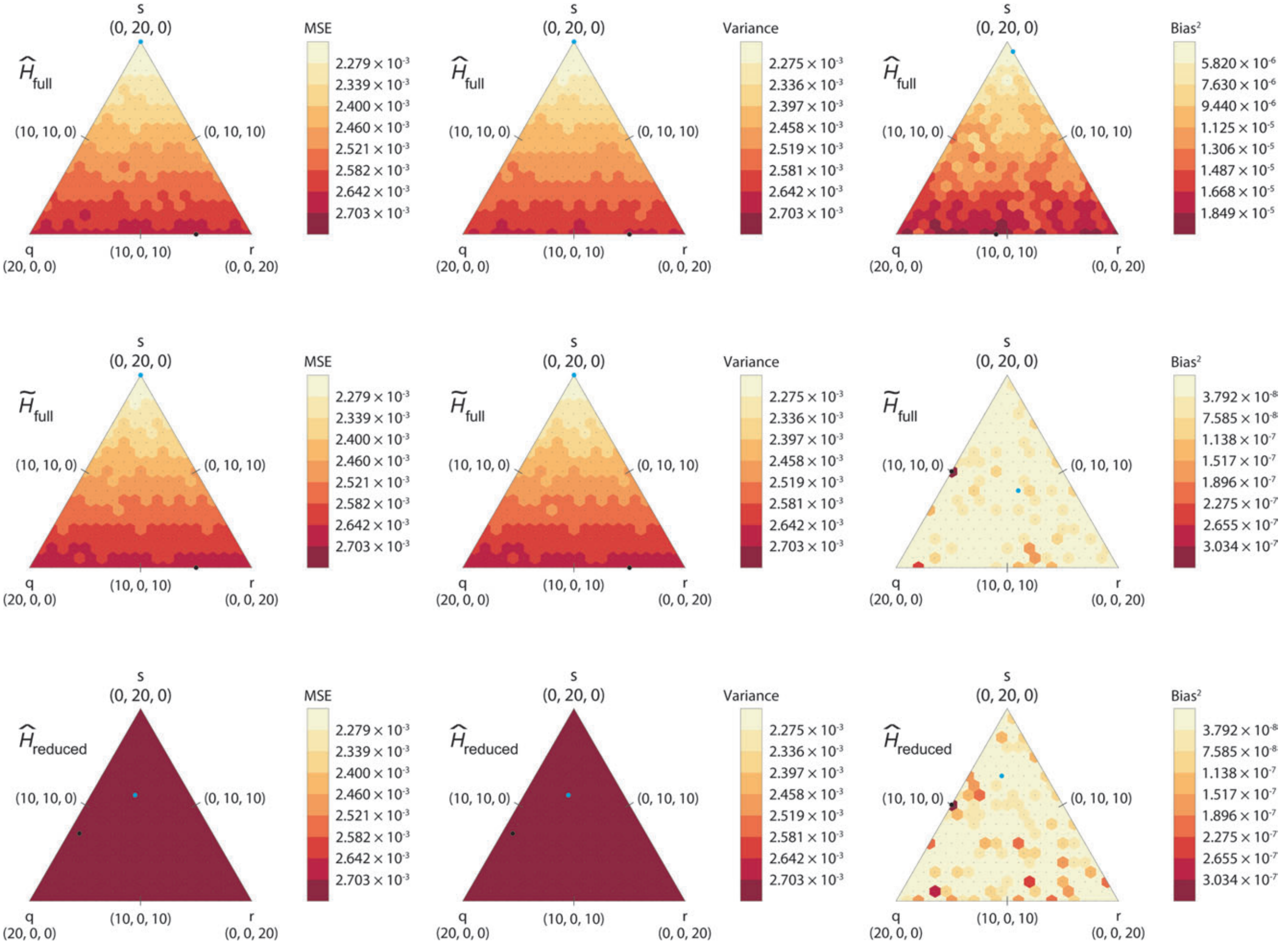


FIG. 2.—Heat maps of simulated MSE, variance, and bias squared for each estimator applied to a full sample of 40 and a reduced sample of 20 individuals, as functions of the mixture of types of relative pairs included in the sample. The simulation was based on allele frequencies at the AAT263P locus ($H = 0.6778$). The sample of 40 individuals includes q parent–offspring, r full-sib, and s second-degree pairs. The three vertices correspond to samples that contain either all parent–offspring, all full-sib, or all second-degree pairs. Moving horizontally along the triangle changes the numbers of parent–offspring and full-sib pairs in the sample and moving vertically changes the number of second-degree pairs. The numbers indicated on the scale are the cutoff values for each color. Each row of triangles represents a different estimator, and each column represents a different statistic. Blue and black dots represent the points at which the smallest and largest values occur in each triangle, respectively. Each point in the graphs is based on 100,000 simulated data sets, and the same simulations were used for all three estimators.

analysis. Similar results were obtained with all six loci (data not shown), and therefore, among the six loci only the locus with the lowest gene diversity (AAT263P, $H = 0.6778$) and the locus with the highest gene diversity (ACT3F12, $H = 0.8263$) were chosen for display. For both loci, table 3 shows the simulated MSE, variance, and bias squared for the different estimators, considering three different sample sizes and three combinations of the number of related individuals for each sample size. Because the simulation results are based on 100,000 replicate data sets, each of the quantities presented is small. However, it is possible to observe differences in the properties of the three estimators. Among the three estimators, $\hat{H}$ applied to full samples gives the lowest variance, $\tilde{H}$ produces slightly higher variance, and $\hat{H}$ applied to samples with related individuals removed produces the highest variance. Bias squared is very close to zero for $\hat{H}$ applied to samples with related individuals removed, as well as for $\tilde{H}$, but it is noticeably higher for $\hat{H}$ applied to full samples containing relatives. For the locus with the lower value of H (0.6778), $\hat{H}$ applied to full samples has the smallest MSE in all cases tested, although $\tilde{H}$ has MSE very close to that of $\hat{H}$. However, for the locus with the higher value of H (0.8263), MSE is always smallest for $\tilde{H}$. Therefore, $\tilde{H}$ is not only unbiased, but it also has MSE comparable with—and sometimes smaller than—that of $\hat{H}$.

It is instructive to investigate the influence of specific variables on the MSE, variance, and bias squared of $\tilde{H}$ and $\hat{H}$, by varying the simulation parameters over the space of gene diversities, sample sizes, and possible sets of relative pairs, and calculating MSE, variance, and bias squared for each scenario. We use $\hat{H}_{\text{full}}$ and $\tilde{H}_{\text{full}}$ to denote $\hat{H}$ and $\tilde{H}$ applied to a sample of individuals. For $\hat{H}$ applied to a sample in which related individuals are removed, we use the notation $\hat{H}_{\text{reduced}}$.

Figure 1 displays the effect of sample size on MSE for each of the estimators, for scenarios in which all simulated individuals belong to relative pairs of a particular type. Here, the full and reduced samples consist of m

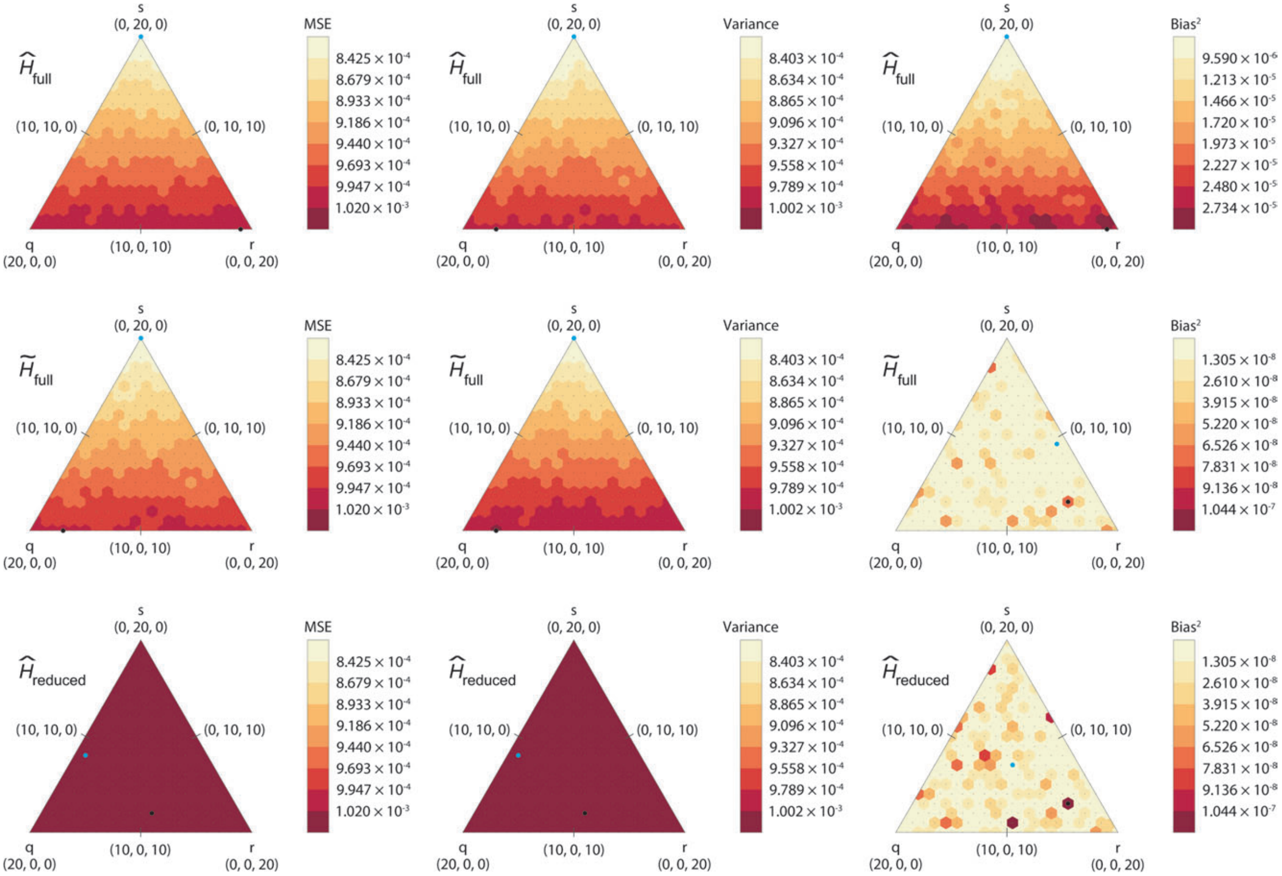


FIG. 3.—Heat maps of simulated MSE, variance, and bias squared for each estimator applied to a full sample of 40 and a reduced sample of 20 individuals, as functions of the mixture of types of relative pairs included in the sample. The simulation was based on allele frequencies at the ACT3F12 locus ($H = 0.8263$). See figure 2 caption for additional details.

and $m/2$ individuals, respectively. When $q = m/2$, $r = m/2$, or $s = m/2$, MSE is consistently lower for $\hat{H}_{\text{full}}$ and $\tilde{H}_{\text{full}}$ (which have virtually identical MSE and therefore have overlapping lines in the graph) than for $\hat{H}_{\text{reduced}}$. As the sample size increases, the MSEs of all estimators approach zero.

We next examined how the three estimators performed in simulated samples containing the same sample size and total number of relative pairs but with different combinations involving different numbers of parent–offspring, full-sib, and second-degree pairs. The same two loci that were analyzed in table 3 and figure 1 were investigated to show the effect of the combination of relative pairs at differing degrees of gene diversity. Figures 2 and 3 illustrate MSE, variance, and bias squared for each estimator as functions of the combination of types of relative pairs in a full sample of size 40 and a reduced sample of size 20 individuals. Each point in a triangle represents the number of parent–offspring, full-sib, and second-degree relative pairs in a sample; the sum of these quantities is equal to half the sample size. MSE and variance are always lower for $\hat{H}_{\text{full}}$ and $\tilde{H}_{\text{full}}$ than for $\hat{H}_{\text{reduced}}$, which relies on a smaller sample size, and $\hat{H}_{\text{full}}$ and $\tilde{H}_{\text{full}}$ show similar trends. Bias squared for the unbiased $\tilde{H}_{\text{full}}$ is similar to that for $\hat{H}_{\text{reduced}}$, which eliminates relatives from the sample, whereas it is much larger for $\hat{H}_{\text{full}}$. As the number of first-degree pairs is increased (decreasing the number of second-degree pairs), both variance and MSE increase. For $\hat{H}_{\text{full}}$, as can be predicted from equation (11), bias squared also increases with an increase in the number of first-degree pairs. Because they are both unbiased estimators, $\tilde{H}_{\text{full}}$ and $\hat{H}_{\text{reduced}}$ display no particular pattern for bias squared.

Finally, we studied the trends in MSE, variance, and bias squared for the estimators over the space of gene diversities, holding the full sample size fixed at 30 individuals and the reduced sample size fixed at 15. Unlike the analyses in table 3 and figures 1–3, which show results based on two representative loci, this analysis used simulations based on all 783 microsatellites. We considered a scenario in which the sample of 30 individuals consisted of 15 parent–offspring pairs. Figure 4 illustrates that for all three estimators, MSE and variance tend to decrease as gene diversity increases. Because $\tilde{H}_{\text{full}}$ and $\hat{H}_{\text{reduced}}$ are both unbiased, bias squared shows no trend for these estimators. However, because bias for $\hat{H}_{\text{full}}$ is linear with respect to gene diversity (eq. 11), bias squared is quadratic. On the basis of equation (11), we predict $\left[\text{bias}\left(\hat{H}_{\text{full}}\right)\right]^2 = \left(-\frac{8\times15\times(1/4)}{60\times59}H\right)^2 \approx \left(7.182 \times 10^{-5}\right)H^2$, and a close match to this prediction was observed. The

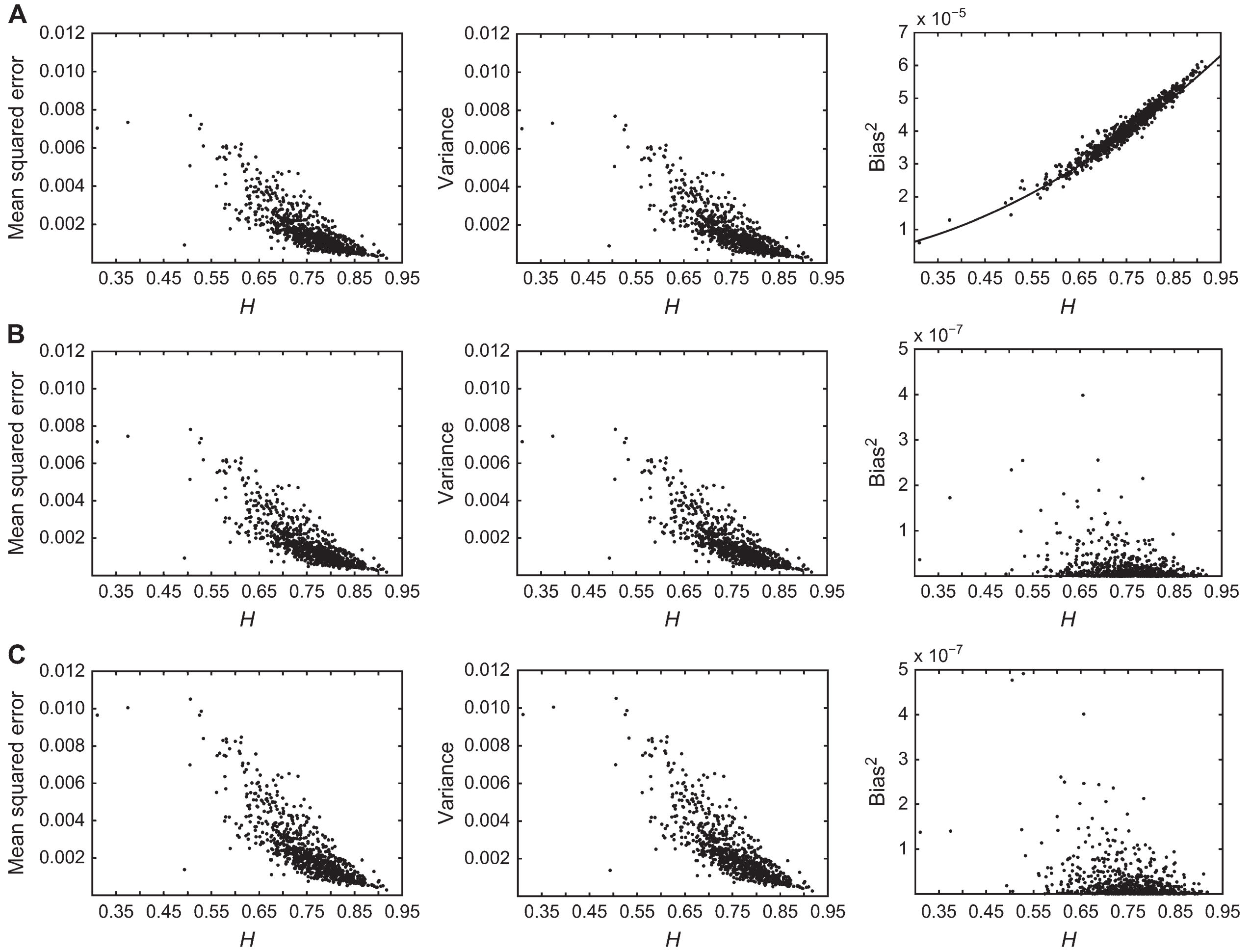


FIG. 4.—MSE, variance, and bias squared for each estimator applied to a full sample of 30 and a reduced sample of 15 individuals, as functions of parametric gene diversity, considering simulated values based on each of the 783 loci. The simulations incorporated 30 individuals in 15 parent–offspring pairs. (*A*) $\hat{H}_{\text{full}}$. A quadratic regression of bias squared on H (with the constant and linear terms forced to be 0) is given by $(7.187 \times 10^{-5})H^2$, with $R^2 = 0.959$. The Spearman correlation coefficient is -0.8364 for H and MSE and -0.8394 for H and variance. (*B*) $\tilde{H}_{\text{full}}$. The Spearman correlation coefficient is -0.8394 for H and MSE and -0.8394 for H and variance. (*C*) $\hat{H}_{\text{reduced}}$. The Spearman correlation coefficient is -0.8447 for H and MSE and -0.8447 for H and variance. Each point in the graphs is based on 100,000 simulated data sets, and the same simulations were used for all three estimators.

regression displayed in figure 4 has regression model $\left[\text{bias}\left(\hat{H}_{\text{full}}\right)\right]^2 = \left(7.187 \times 10^{-5}\right)H^2$.

Three main results can be observed in our simulations. First, $\tilde{H}$ is unbiased and has comparable bias in samples containing relatives to that obtained by applying $\hat{H}$ to samples with relatives removed. Using $\tilde{H}$, or excluding relatives and using $\hat{H}$, reduces the bias compared with using $\hat{H}$ without excluding relatives. Second, $\tilde{H}$ has comparable (but consistently slightly higher) variance to the values obtained with $\hat{H}$ in samples containing relatives. Both $\tilde{H}$ and $\hat{H}$ have lower variance in full samples of individuals than that of $\hat{H}$ in reduced samples that exclude relatives. Third, because $\tilde{H}$ has less bias than $\hat{H}$ in samples containing relatives, $\tilde{H}$ has comparable, and sometimes smaller, MSE to $\hat{H}$ (although its variance is larger). Both estimators have lower MSE than $\hat{H}$ applied to subsets that exclude relatives.

The properties of the estimators depend on a number of parameters. All estimators have lower MSE as sample size increases. In addition, the MSEs of $\hat{H}$ and $\tilde{H}$ are smaller when second-degree relative pairs are investigated, in comparison to scenarios that include an equivalent number of first-degree pairs. Furthermore, the MSEs of $\hat{H}$ and $\tilde{H}$ are generally smaller for loci with larger gene diversities, with the magnitude of the bias of $\hat{H}$ increasing linearly with increasing gene diversity.

We can conclude that for samples containing relatives, $\tilde{H}$ has comparable variance to $\hat{H}$, with a considerable reduction of bias. $\tilde{H}$ has comparable bias in a full sample to that of $\hat{H}$ applied to a reduced sample excluding relatives, with a considerable reduction of variance. Thus, $\tilde{H}$ combines into a single estimator the desirable properties possessed by $\hat{H}$ applied to samples with relatives and by $\hat{H}$ applied to samples without relatives.

Application to Data

Notation

For convenience, we use the following notation: $\hat{H}_{952}$ and $\hat{H}_{1048}$ for application of $\hat{H}$ to the samples of 952 and

1,048 individuals, respectively, and $\tilde{H}_{952}$ and $\tilde{H}_{1048}$ for application of $\tilde{H}$ to these samples. Note that because the H952 data set contains no relative pairs, $\tilde{H}_{952}=\hat{H}_{952}$, and there is no need to consider $\tilde{H}_{952}$ separately. We also use the notation $\hat{H}_{507}$, $\hat{H}_{603}$, and $\tilde{H}_{603}$ when restricting our analysis to the 26 populations containing at least one relative pair; for each of the 27 remaining populations, the estimators $\hat{H}$ and $\tilde{H}$ produce identical values.

Mean of the Estimator

For investigating the properties of $\hat{H}$ and $\tilde{H}$ applied to the H1048 data set, because the true value of H is unknown for the actual data, we treated the value of $\hat{H}_{952}$ for each locus as a substitute "true" value. Because $\hat{H}$ is unbiased when applied to data not containing relatives, $\hat{H}_{952}$ provides a sensible proxy for the unknown true gene diversity. This approach enabled us to consider how estimates of H from data including relatives might differ from estimates based on the same data excluding all relatives. For each of the 53 populations, we computed the means of $\hat{H}_{952}$, $\hat{H}_{1048}$, and $\tilde{H}_{1048}$ across the 783 microsatellite loci. Because the true H is unknown and bias cannot be calculated, we instead examine the mean of $\hat{H}_{1048}$ across loci minus the mean of $\hat{H}_{952}$ across loci and the mean of $\tilde{H}_{1048}$ across loci minus the mean of $\hat{H}_{952}$ across loci.

Figure 5 shows comparisons of the mean of $\hat{H}_{1048}-\hat{H}_{952}$ across loci and the mean of $\tilde{H}_{1048}-\hat{H}_{952}$ across loci. In general, the three estimators produce similar estimates in a given population. However, notice that in figure 5*A*, $\hat{H}_{1048}$ is reduced compared with $\hat{H}_{952}$, a likely consequence of the bias of $\hat{H}$ when applied to samples containing relatives. When $\tilde{H}_{1048}$ is used in place of $\hat{H}_{1048}$, because $\tilde{H}_{1048}$ corrects for the inclusion of known related individuals, there is a considerable reduction in the magnitude of the difference between the mean of the estimator ($\hat{H}_{1048}$ or $\tilde{H}_{1048}$) across loci and the mean of $\hat{H}_{952}$ across loci (fig. 5*B*). These observations are reflected in Wilcoxon signed rank tests that compare paired lists of mean heterozygosities across loci for the 53 populations (table 4). The P value for a comparison of $\hat{H}_{1048}$ with $\hat{H}_{952}$ was 8.804×10^{-6}, suggesting that inclusion of relatives in a sample has a statistically significant impact on $\hat{H}$. In contrast, $\tilde{H}_{1048}$ and $\hat{H}_{952}$ showed no significant difference, with a P value of 0.703 for the Wilcoxon signed rank test. Similar results were obtained for other comparisons of the three estimators. The mean across populations of $\hat{H}_{952}-\tilde{H}_{1048}\left(3.262 \times 10^{-4}\right)$ was smaller than for $\hat{H}_{952}-\hat{H}_{1048}\left(2.387 \times 10^{-3}\right)$; the same was true for the mean of $\left|\hat{H}_{952}-\tilde{H}_{1048}\right|\left(6.660 \times 10^{-4}\right)$ compared with the mean of $\left|\hat{H}_{952}-\hat{H}_{1048}\right|\left(2.387 \times 10^{-3}\right)$.

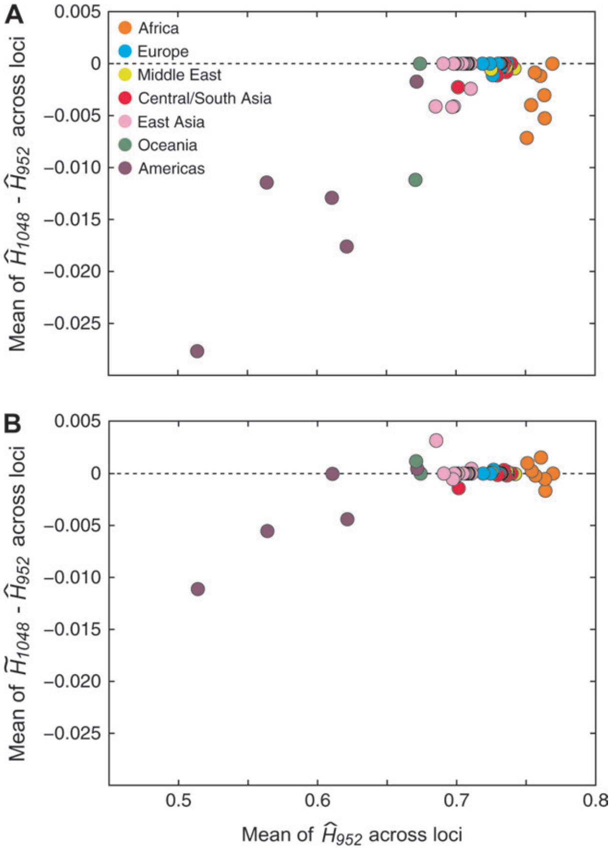


FIG. 5.—Comparison of the mean of $\hat{H}_{1048}-\hat{H}_{952}$ and the mean of $\tilde{H}_{1048}-\hat{H}_{952}$. Each population is represented by a point colored based on the geographic location of the population, and the dotted line represents zero difference between the full-data estimator and $\hat{H}_{952}$. Because 27 of the 53 populations do not contain related individuals, the gene diversities given by $\hat{H}_{1048}$ and $\tilde{H}_{1048}$ are the same for these populations. (*A*) The mean of $\hat{H}_{1048}-\hat{H}_{952}$, displaying a reduction of $\hat{H}$ when applied to samples containing related individuals. (*B*) The mean of $\tilde{H}_{1048}-\hat{H}_{952}$, displaying a decrease in the magnitude of the difference between the full-data estimator and $\hat{H}_{952}$.

Comparable results were obtained when using only the 26 populations that contained relative pairs. The Wilcoxon signed rank test produced a statistically significant P value of 2.980×10^{-8} for $\hat{H}_{603}$ compared with $\hat{H}_{507}$ and a nonsignificant P value of 0.708 when comparing $\tilde{H}_{603}$ with $\hat{H}_{507}$. The mean across populations of $\hat{H}_{507}-\tilde{H}_{603}\left(6.649 \times 10^{-4}\right)$ was smaller than for $\hat{H}_{507}-\hat{H}_{603}\left(4.866 \times 10^{-3}\right)$, as was the mean of $\left|\hat{H}_{507}-\tilde{H}_{603}\right|\left(1.358 \times 10^{-3}\right)$ relative to that of $\left|\hat{H}_{507}-\hat{H}_{603}\right|\left(4.866 \times 10^{-3}\right)$. In addition, similar numbers of populations had $\tilde{H}_{603}>\hat{H}_{507}$ (12) and $\tilde{H}_{603}<\hat{H}_{507}$ (14); by contrast, there were no populations with $\hat{H}_{603}>\hat{H}_{507}$.

Because estimators often have a trade-off between bias and variance, we investigated the relationship between the mean values across loci of $\hat{H}_{603}-\hat{H}_{507}$ and $\tilde{H}_{603}-\hat{H}_{507}$ and the standard deviations of $\hat{H}_{603}$ and $\tilde{H}_{603}$ across loci. We observed that compared with $\hat{H}_{603}$, $\tilde{H}_{603}$ produces a noticeable decrease in the mean difference from $\hat{H}_{507}$ with only a slight increase in the standard deviation (fig. 6). This result is somewhat analogous to the simulation-based result that $\tilde{H}$ has less bias than $\hat{H}$ and comparable variance.

Gene Diversity versus Distance from Africa

Based on an observed decline of gene diversity estimates with geographic distance from East Africa, Ramachandran et al. (2005) argued that the geographic expansion of modern humans can be described by a series of founder events originating in Africa. This analysis utilized

Table 4
Statistical Tests Applied to the Mean Gene Diversity across Loci

	P value for Wilcoxon Signed Rank Test	Mean of $H_{\text{reduced}} - H_{\text{full}}$ across Populations	Mean of $\lvert H_{\text{reduced}} - H_{\text{full}}\rvert$ across Populations	Fraction of Populations with $H_{\text{full}} > H_{\text{reduced}}$
$\hat{H}_{952}$ versus $\hat{H}_{1048}$	8.804×10^{-6}	2.387×10^{-3}	2.387×10^{-3}	0
$\hat{H}_{952}$ versus $\tilde{H}_{1048}$	0.703	3.262×10^{-4}	6.660×10^{-4}	0.226
$\hat{H}_{507}$ versus $\hat{H}_{603}$	2.980×10^{-8}	4.866×10^{-3}	4.866×10^{-3}	0
$\hat{H}_{507}$ versus $\tilde{H}_{603}$	0.708	6.649×10^{-4}	1.358×10^{-3}	0.462

In the header line, H_{reduced} refers to H_{952} or H_{507} depending on which estimator is being considered; similarly, H_{full} refers to $\hat{H}_{1048}$, $\tilde{H}_{1048}$, $\hat{H}_{603}$, or $\tilde{H}_{603}$.

the $\hat{H}$ estimator applied to the 783 microsatellites typed in the H1048 subset of individuals, excluding the Surui population. To evaluate how the results of Ramachandran et al. (2005) were affected by the bias of $\hat{H}$ in samples with close relatives, we analyzed the relationships of the three estimators of gene diversity—$\hat{H}_{952}$, $\hat{H}_{1048}$, and $\tilde{H}_{1048}$—with geographic distance from East Africa (fig. 7). Distance from Addis Ababa was measured in kilometers via waypoint routes and was based on the values from Rosenberg et al. (2005).

The three estimators produced relatively similar regressions (fig. 7), demonstrating that the close linear relationship of gene diversity and distance from Africa is not greatly affected by inclusion of relatives in the analysis. We observed very similar values for the coefficients of determination (R^2) of linear regressions when using $\hat{H}_{952}$, $\hat{H}_{1048}$, and $\tilde{H}_{1048}$ (note that all three R^2 values are higher than that reported by Ramachandran et al. (2005), whose lower value resulted from an error in the calculation of their fig. 4*A*). The Surui population, which has the smallest gene diversity and is the farthest population from Addis Ababa, deviates considerably from the regression line when using $\hat{H}_{1048}$ to measure gene diversity (fig. 7*B*). When excluding the large number of relatives present in the Surui sample ($\hat{H}_{952}$) or correcting for their inclusion ($\tilde{H}_{1048}$), the Surui population is not as extreme an outlier (fig. 7*A* and *C*).

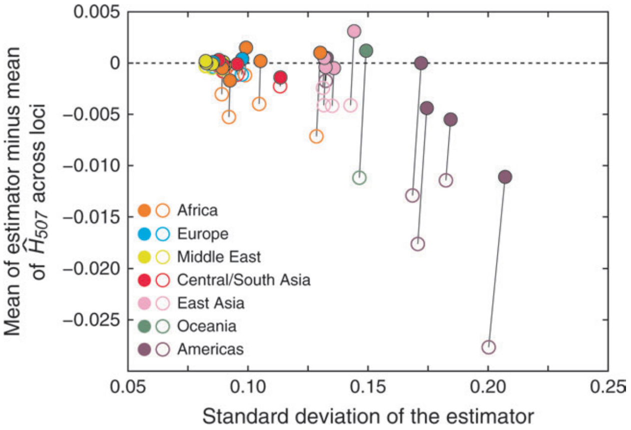


FIG. 6.—Comparison of the mean difference of an estimator ($\hat{H}_{603}$ or $\tilde{H}_{603}$) from $\hat{H}_{507}$ with the standard deviation of the estimator. Each population is represented by a point colored based on the geographic location of the population. Open and filled circles represent the estimates for $\hat{H}_{603}$ and $\tilde{H}_{603}$, respectively.

Discussion

In this article, we have developed an unbiased estimator $\tilde{H}$ for gene diversity in samples containing related and inbred individuals. The bias-correction factor in this estimator, which we derived from the variance of allele frequency estimates, depends only on the average kinship coefficient between pairs of sampled individuals. Using data simulated based on allele frequency distributions from human populations, we found that $\tilde{H}$ performs well with regard to both bias and MSE. The bias generated by $\tilde{H}$ applied to data including relatives is approximately the same as the bias generated by the standard estimator $\hat{H}$ applied to data containing only unrelated individuals. The MSE for $\tilde{H}$ is comparable to—and often smaller than—the MSE of $\hat{H}$ when related individuals are included. Calculation of $\tilde{H}$ relies only on sample allele frequencies and on the average kinship coefficient and is therefore easy to perform when relationships among individuals are known. Thus, the new estimator $\tilde{H}$ offers a combination of unbiasedness, low MSE, and ease of computation, providing an improved approach to the estimation of gene diversity in samples containing relatives.

Using data from human populations, we found that $\tilde{H}$ largely corrected a reduction in the standard estimator $\hat{H}$, producing estimates that were not significantly different from those obtained if we instead removed relatives from the data set and applied $\hat{H}$. This shift toward the values obtained in data without relatives occurred together with only a slight increase in standard deviation across loci relative to $\hat{H}$. However, by treating dependent observations as independent, $\hat{H}$ perhaps produces a smaller variance than is appropriate in samples with relatives. Thus, we conclude that as an alternative to removing relatives from samples containing relative pairs, $\tilde{H}$ can be applied to obtain suitable gene diversity estimates.

When we applied $\tilde{H}$ to the human data, a few populations still produced a "bias," in that $\tilde{H}_{1048}$ remained considerably lower than $\hat{H}_{952}$. The most noticeable of these populations are the Surui, Karitiana, and Pima populations from the Americas (fig. 5*B*); the "bias" was larger for these low-diversity populations, whereas theory predicts less bias when diversity is lower (eq. 11). It should first be noted that unlike for the other populations, inferences about second-degree relationships obtained by Rosenberg (2006) were somewhat uncertain for the Surui and Karitiana populations. Thus, table 2 and our analysis did not include inferred second-degree relationships in those populations, when in fact many are likely to be

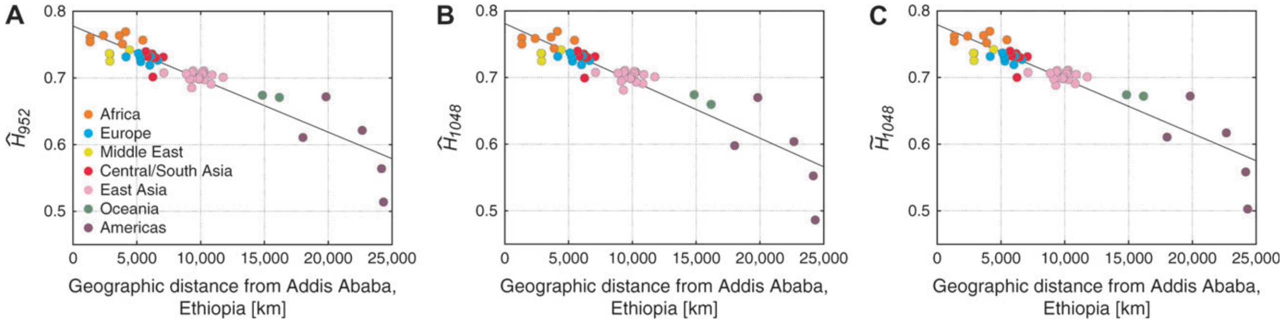


FIG. 7.—Gene diversity versus geographic distance from Addis Ababa, Ethiopia. (*A*) $\hat{H}_{952}$ versus distance from Addis Ababa. The linear regression is given by $H = 0.7778 - (7.955 \times 10^{-6}) \times$ distance, with $R^2 = 0.856$. (*B*) $\hat{H}_{1048}$ versus distance from Addis Ababa. The linear regression is given by $H = 0.7809 - (8.595 \times 10^{-6}) \times$ distance, with $R^2 = 0.844$. (*C*) $\tilde{H}_{1048}$ versus distance from Addis Ababa. The linear regression is given by $H = 0.7792 - (8.161 \times 10^{-6}) \times$ distance, with $R^2 = 0.849$.

present. This is a likely reason why the "bias" in the Surui and Karitiana populations was only partially eliminated. For the Pima population, a likely explanation is that the sample contains many related individuals in extended families (Rosenberg, 2006), and our computation only adjusted for first- and second-degree relative pairs. If these higher order relationships had been fully known, however, it would have been possible to use our estimator to adjust for them.

Our estimator adjusts for inbreeding by averaging over inbreeding coefficients for sampled individuals. It is important to note that the inbreeding coefficients that we have included are exact values obtained from pedigrees. If an estimated inbreeding coefficient was used in place of the exact value, then $\tilde{H}$ would not necessarily produce unbiased estimates in samples containing inbred individuals. $\tilde{H}$ would also lead to a bias if relationships were misspecified. In our data example, relationships were assumed to be known, and for a data set of the size used for inferring the relationships (Rosenberg 2006) this assumption is generally sensible. However, for small data sets in which relationship inferences are uncertain, caution must be used when interpreting the bias of $\tilde{H}$ applied to the same data from which relationships are estimated.

The estimators we have considered relate to within-population gene diversity. What if we consider the gene diversity between populations? Suppose we have samples from two populations, *A* and *B*, each containing related inbred individuals. The between-population analog of gene diversity is $\hat{H}_{A,B} = 1 - \sum_{i=1}^{I} \hat{p}_i \hat{q}_i$, where $\hat{p}_i$ and $\hat{q}_i$ are estimates of the frequency of allele *i* at a given locus in populations *A* and *B*, respectively (Nei 1987). Because the bias in within-population gene diversity estimates only arises from the quadratic $\hat{p}_i^2$ term in equation (1), $E\left[\sum_{i=1}^{I} \hat{p}_i \hat{q}_i\right] = \sum_{i=1}^{I} p_i q_i$ (Nei 1987, p. 222), and $\hat{H}_{A,B}$ continues to be an unbiased estimator for between-population gene diversity in samples containing relatives.

Acknowledgments

We thank Ivana Jankovic, Yi-Ju Li, and two anonymous reviewers for helpful comments. This work was supported by NIH training grant T32 GM070449, NIH grant R01 GM081441, and grants from the Burroughs Welcome Fund and the Alfred P. Sloan Foundation.

Literature Cited

Almasy L, Blangero J. 1998. Multipoint quantitative-trait linkage analysis in general pedigrees. Am J Hum Genet. 62:1198–1211.

Boehnke M. 1991. Allele frequency estimation from data on relatives. Am J Hum Genet. 48:22–25.

Bourgain C, Hoffjan S, Nicolae R, Newman D, Steiner L, Walker K, Reynolds R, Ober C, McPeek MS. 2003. Novel case–control test in a founder population identifies P-selectin as an atopy-susceptibility locus. Am J Hum Genet. 73:612–626.

Broman KW. 2001. Estimation of allele frequencies with data on sibships. Genet Epidemiol. 20:307–315.

Cann HM, de Toma C, Cazes L, et al. (41 co-authors). 2002. A human genome diversity cell line panel. Science. 296:261–262.

Cavalli-Sforza LL. 2005. The Human Genome Diversity Project: past, present and future. Nat Rev Genet. 6:333–340.

Depaulis F, Veuille M. 1998. Neutrality tests based on the distribution of haplotypes under an infinite-site model. Mol Biol Evol. 15:1788–1790.

Driscoll CA, Menotti-Raymond M, Nelson G, Goldstein D, O'Brien SJ. 2002. Genomic microsatellites as evolutionary chronometers: a test in wild cats. Genome Res. 12:414–423.

Gibbs JP, Martin WT. 1962. Urbanization, technology, and the division of labor: international patterns. Am Sociol Rev. 27:667–677.

Gini CW. 1912. Variabilita e mutabilita. Studi Economico-Giuridici della R. Universita di Cagliari 3.

Hoelzel AR, Fleischer RC, Campagna C, Le Boeuf BJ, Alvord G. 2002. Impact of a population bottleneck on symmetry and genetic diversity in the northern elephant seal. J Evol Biol. 15:567–575.

Jacquard A. 1974. The genetic structure of populations. New York: Springer.

Lange K. 2002. Mathematical and statistical methods for genetic analysis. 2nd ed. New York: Springer.

Li CC, Horvitz DG. 1953. Some methods of estimating the inbreeding coefficient. Am J Hum Genet. 5:107–117.

McPeek MS, Wu X, Ober C. 2004. Best linear unbiased allele-frequency estimation in complex pedigrees. Biometrics. 60:359–367.

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

Nei M. 1987. Molecular evolutionary genetics. New York: Columbia University Press.

Nei M, Roychoudhury AK. 1974. Sampling variances of heterozygosity and genetic distance. Genetics. 76: 379–390.

Ohta T. 1980. Linkage disequilibrium between amino acid sites in immunoglobulin genes and other multigene families. Genet Res. 36:181–197.

Ramachandran S, Deshpande O, Roseman CC, Rosenberg NA, Feldman MW, Cavalli-Sforza LL. 2005. Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. Proc Natl Acad Sci USA. 102:15942–15947.

Rosenberg NA. 2006. Standardized subsets of the HGDP-CEPH Human Genome Diversity Cell Line Panel, accounting for atypical and duplicated samples and pairs of close relatives. Ann Hum Genet. 70:841–847.

Rosenberg NA, Mahajan S, Ramachandran S, Zhao C, Pritchard JK, Feldman MW. 2005. Clines, clusters, and the effect of study design on the inference of human population structure. PLoS Genet. 1:660–671.

Sabatti C, Risch N. 2002. Homozygosity and linkage disequilibrium. Genetics. 160:1707–1719.

Sabeti PC, Reich DE, Higgins JM, et al. (17 co-authors). 2002. Detecting recent positive selection in the human genome from haplotype structure. Nature. 419:832–837.

Shete S. 2003. Uniformly minimum variance unbiased estimation of gene diversity. J Hered. 94:421–424.

Simpson EH. 1949. Measurement of diversity. Nature. 163:688–688.

Watterson GA. 1978. The homozygosity test of neutrality. Genetics. 88:405–417.

Weir BS. 1989. Sampling properties of gene diversity. In: Brown AHD, Clegg MT, Kahler AL, Weir BS, editors. Plant population genetics, breeding and genetic resources. Sinauer Associates. p. 23–42.

Weir BS. 1996. Genetic data analysis II. Sunderland (MA): Sinauer Associates.

Weir BS, Anderson AD, Hepler AB. 2006. Genetic relatedness analysis: modern data and new challenges. Nat Rev Genet. 7:771–780.

Yi-Ju Li, Associate Editor

Accepted October 23, 2008