The research article titled "An Unbiased Estimator of Gene Diversity in Samples Containing Related Individuals" by Michael DeGiorgio and Noah A. Rosenberg discusses the estimation of gene diversity in populations containing related individuals. The authors present an unbiased estimator for gene diversity that corrects for the inclusion of related individuals in samples. This estimator is derived by considering the gene diversity in terms of kinship coefficients, which quantify the relatedness between individuals. The estimator is shown to reduce bias when related individuals are present in the sample, providing a more accurate measure of gene diversity than previous estimators. The article is published in the journal Molecular Biology and Evolution, and the DOI for the article is 10.1093/molbev/msn254.
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, \(\hat{H}\). We begin by computing \(\text{Var}[\hat{p}_i]\) in a sample that may include relatives or inbred individuals. \(\text{Var}[\hat{p}_i]\) 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, \ldots, \(n/2\), an unbiased estimator for \(p_i\), the frequency of allele \(i\), is

\[
\hat{p}_i = \frac{1}{n} \sum_{k=1}^{n/2} X_k.
\]

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

\[
\text{Var}[\hat{p}_i] = \frac{1}{n^2} \sum_{j=1}^{n/2} \sum_{k=1}^{n/2} \text{Cov}[X_j, X_k].
\]

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 \(\Phi_{j,k} = (1/2)(1 + f_k)\), 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 (Jaccard 1974, p. 107). Let \(A_i = \text{Pr}[S = s]\), 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 \(A_1 + (1/2)(A_3 + A_5 + A_7) + (1/4)A_8\) (Jaccard, 1974, p. 108), we obtain

\[
E[X_jX_k] = \sum_{a=0}^{2} \sum_{b=0}^{2} \sum_{s=1}^{9} abA_i \text{Pr}[X_j = a, X_k = b | S = s] = 4\Phi_{j,k}p_i(1 - p_i) + 4p_i^2.
\]

Because \(E[X_j] = E[X_k] = 2p_i\), it follows that

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

Inserting the covariance into equation (5) yields

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

\[
\text{Pr} \left[ X_j, X_k = 0, 0 \right] = 1 - p, \quad \text{Pr} \left[ X_j, X_k = 0, 2 \right] = p(1 - p), \quad \text{Pr} \left[ X_j, X_k = 2, 0 \right] = p(1 - p), \quad \text{Pr} \left[ X_j, X_k = 2, 2 \right] = p^2.
\]

| \(S\) | Condensed Identity State* | \(X_j, X_k\) | \(\text{Pr} \left[ X_j, X_k \right]\) |
|---|---|---|---|
| 1 | | 0, 0 | 1 - \(p\) |
| 2 | | 0, 0 | \((1 - p)^2\) |
| 3 | | 0, 0 | \((1 - p)^2\) |
| 4 | | 0, 0 | \((1 - p)^3\) |
| 5 | | 0, 0 | \((1 - p)^2\) |
| 6 | | 0, 0 | \((1 - p)^3\) |
| 7 | | 0, 0 | \((1 - p)^2\) |
| 8 | | 0, 0 | \((1 - p)^3\) |
| 9 | | 0, 0 | \((1 - p)^4\) |

* 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.

\[
\text{Cov}[X_j, X_k] = E[X_jX_k] - E[X_j]E[X_k] = 4\Phi_{j,k}p_i(1 - p_i).
\]

(6)

(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                             |
| Nazi               | East Asia         | 10                            | 0                                | 1                       | 0                             |
| Orogen             | East Asia         | 10                            | 0                                | 1                       | 0                             |
| Melanesian         | Oceania           | 19                            | 9                                | 3                       | 2                             |
| Colombian          | America           | 13                            | 6                                | 1                       | 0                             |
| Karitana           | America           | 24                            | 6                                | 6                       | 0                             |
| Maya               | America           | 25                            | 2                                | 1                       | 2                             |
| Fima               | America           | 25                            | 15                               | 6                       | 10                            |
| Sarui              | America           | 21                            | 15                               | 14                      | 0                             |

where \( \hat{\phi} = \frac{1}{(n/2)^2} \sum_{j=1}^{n/2} \sum_{k=1}^{n/2} \phi_{j,k} \) 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 \( l \) distinct alleles, allele frequencies \( p_i \in [0, 1] \) and \( \sum i p_i = 1 \). Suppose a sample from a population has \( n/2 \) possibly related and inbred individuals. Then an unbiased estimator for gene diversity is

\[
\hat{H} = \frac{1}{1 - \phi} \left( 1 - \frac{1}{n} \sum_{i=1}^{l} \hat{\phi}^2 \right),
\]

where \( \hat{\phi}_{j,k} \) is the kinship coefficient of individuals \( j \) and \( k \) and \( \hat{\phi} = \frac{1}{(n/2)^2} \sum_{j=1}^{n/2} \sum_{k=1}^{n/2} \phi_{j,k} \) is the average kinship coefficient across pairs of individuals.

**Proof**

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

\[
E[H] = \frac{1}{1 - \phi} \left[ 1 - \sum_{i=1}^{l} \left( \text{Var}[\hat{\phi}_i] + p_i^2 \right) \right]
\]

\[
= \frac{1}{1 - \phi} \left[ 1 - \sum_{i=1}^{l} (p_i(1 - p_i) + p_i^2) \right] = H. \quad \square
\]

**Corollary 2**

Consider a locus with \( l \) distinct alleles, allele frequencies \( p_i \in [0, 1] \) and \( \sum i p_i = 1 \). Suppose a sample from a population has \( n/2 \) possibly related and inbred individuals. Let \( R \) 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 \( R \in R \) and let \( \Phi_R \) be the kinship coefficient for each of these pairs. Then an unbiased estimator for gene diversity is

\[
\hat{H} = \frac{n(n - 1)}{n(n - 1 - f)} - 8 \sum_{R \in R} n_R \Phi_R, \quad (9)
\]

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

**Proof**

Applying the definitions of \( \hat{\phi} \) and \( \phi_{j,k} \) and the fact that \( \phi_{j,k} = 0 \) for a pair of “unrelated” individuals,

\[
\phi = \frac{1}{(n/2)^2} \sum_{j=1}^{n/2} \sum_{k=1}^{n/2} \phi_{j,k} = \frac{4}{n^2} \left( \sum_{k=1}^{n/2} \phi_{j,k} + 2 \sum_{j=1}^{n/2} \sum_{k=j+1}^{n/2} \phi_{j,k} \right)
\]

\[
= \frac{1}{n^2} \left( n + n^2 + 8 \sum_{R \in R} n_R \Phi_R \right).
\]

Inserting this value for \( \hat{\phi} \) into equation (8), we obtain the desired result. \( \square \)
Note that if no related individuals are included in the sample, then \( R \) is the empty set, thus reducing \( \hat{H} \) to \( \hat{H}_{\text{WCR}} \); if additionally no inbred individuals are included, then \( f = 0 \) and \( \hat{H} \) 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

\[
\hat{H} = \frac{n(n-1)}{n(n-1)-2q-2r-s} \hat{H} \tag{10}
\]

Proof

The kinship coefficients are \( \Phi_p = 1/4 \) for parent–offspring pairs, \( \Phi_r = 1/4 \) for full-sib pairs, and \( \Phi_S = 1/8 \) for second-degree pairs. If an individual \( k \) is not inbred, then \( f_k = 0 \). For a sample without inbred individuals, \( f = 0 \). 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 \( R \) 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 \( R \in R \) and let \( \Phi_R \) 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

\[
\text{bias}(\hat{H}) = - \frac{n f}{n(n-1)} + 8 \sum_{R \in R} n_R \Phi_R H, \tag{11}
\]

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

Proof

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

Data from Human Populations

To examine the behavior of \( \hat{H} \) 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
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 $\hat{H}$ and $\tilde{H}$. 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 $\hat{H}$ and $\tilde{H}$ 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
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 \(\hat{H}_{\text{reduced}}\) 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, \(\hat{H}\) is not only unbiased, but it also has MSE comparable with—and sometimes smaller than—that of \(\tilde{H}\).

It is instructive to investigate the influence of specific variables on the MSE, variance, and bias squared of \(\hat{H}\) and \(\tilde{H}\), 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 \(m\)
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. 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 $[\text{bias}(\hat{H}_{\text{full}})]^2 \approx \frac{8 \times 15 \times (1/4)}{60 \times 59} H^2 \approx 7.182 \times 10^{-5} H^2$, and a close match to this prediction was observed. The
regression displayed in figure 4 has regression model
\[
\text{bias}(\hat{H}_{\text{full}})^2 = (7.187 \times 10^{-5})H^2.
\]

Three main results can be observed in our simulations. First, \( \hat{H} \) is unbiased and has comparable bias in samples containing relatives to that obtained by applying \( \tilde{H} \) to samples with relatives removed. Using \( \hat{H} \), or excluding relatives and using \( \tilde{H} \), 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 \( \hat{H} \) and \( \tilde{H} \) have lower variance in full samples of individuals than that of \( \hat{H} \) in reduced samples that exclude relatives. Third, because \( \hat{H} \) has less bias than \( \tilde{H} \) in samples containing relatives, \( \hat{H} \) has comparable, and sometimes smaller, MSE to \( \tilde{H} \) (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 full sample to that of \( \hat{H} \) applied to a reduced sample excluding relatives, with a considerable reduction of variance. Thus, \( \tilde{H} \) combines in to a single estimator the desirable properties possessed by \( \hat{H} \) applied to samples with relatives and by \( \tilde{H} \) 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 \( \hat{H}_{502} \) and \( \hat{H}_{1048} \) for application of \( H \) to these samples. Note that because the H952 data set contains no relative pairs, \( \hat{H}_{502} = \hat{H}_{952} \), and there is no need to consider \( \hat{H}_{502} \) separately. We also use the notation \( \hat{H}_{507} \), \( \hat{H}_{603} \), and \( \hat{H}_{603} \) 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 \( \hat{H} \) produce identical values.

Mean of the Estimator

For investigating the properties of \( \hat{H} \) and \( \hat{H} \) applied to the H1048 data set, because the true value of \( H \) is unknown for the actual data, we treated the value of \( \hat{H}_{502} \) for each locus as a substitute “true” value. Because \( \hat{H} \) is unbiased when applied to data not containing relatives, \( \hat{H}_{502} \) 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 \( \hat{H}_{502} \), \( \hat{H}_{1048} \), and \( \hat{H}_{1048} \) 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 \( \hat{H}_{502} \) across loci and the mean of \( \hat{H}_{1048} \) across loci minus the mean of \( \hat{H}_{502} \) across loci.

Figure 5 shows comparisons of the mean of \( \hat{H}_{1048} - \hat{H}_{502} \) across loci and the mean of \( \hat{H}_{1048} - \hat{H}_{502} \) across loci. In general, the three estimators produce similar estimates in a given population. However, notice that in figure 5A, \( \hat{H}_{1048} \) is reduced compared with \( \hat{H}_{502} \), a likely consequence of the bias of \( \hat{H} \) when applied to samples containing relatives. When \( \hat{H}_{1048} \) is used in place of \( \hat{H}_{1048} \), because \( \hat{H}_{1048} \) 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 \( \hat{H}_{1048} \)) across loci and the mean of \( \hat{H}_{502} \) across loci (fig. 5B). 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 \( \hat{H}_{502} \) was 8.804 \( \times \) \( 10^{-6} \), suggesting that inclusion of relatives in a sample has a statistically significant impact on \( H \). In contrast, \( \hat{H}_{1048} \) and \( \hat{H}_{502} \) 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 \( \hat{H}_{502} - \hat{H}_{1048} \) \( 3.262 \times 10^{-4} \) was smaller than for \( \hat{H}_{502} - \hat{H}_{1048} \) \( 2.387 \times 10^{-3} \); the same was true for the mean of \( \hat{H}_{502} - \hat{H}_{1048} \) \( 6.660 \times 10^{-4} \) compared with the mean of \( \hat{H}_{502} - \hat{H}_{1048} \) \( 2.387 \times 10^{-3} \).

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 \( \times \) \( 10^{-8} \) for \( \hat{H}_{603} \) compared with \( \hat{H}_{507} \) and a nonsignificant \( P \) value of 0.708 when comparing \( \hat{H}_{603} \) with \( \hat{H}_{507} \). The mean across populations of \( \hat{H}_{507} - \hat{H}_{603} \) \( 6.649 \times 10^{-4} \) was smaller than for \( \hat{H}_{507} - \hat{H}_{603} \) \( 4.866 \times 10^{-3} \), as was the mean of \( \hat{H}_{507} - \hat{H}_{603} \) \( 1.358 \times 10^{-3} \) relative to that of \( \hat{H}_{507} - \hat{H}_{603} \) \( 4.866 \times 10^{-3} \). In addition, similar numbers of populations had \( \hat{H}_{603} > \hat{H}_{507} \) (12) and \( \hat{H}_{603} < \hat{H}_{507} \) (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 \( \hat{H}_{603} \) and \( \hat{H}_{507} \) and the standard deviations of \( \hat{H}_{603} \) and \( \hat{H}_{603} \) across loci. We observed that compared with \( \hat{H}_{603} \), \( \hat{H}_{603} \) 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 \( \hat{H} \) 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...
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—\( \hat{H}_{507} \), \( \hat{H}_{1048} \), and \( \hat{H}_{1048} \)—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 \( \hat{H}_{507} \), \( \hat{H}_{1048} \), and \( \hat{H}_{1048} \) (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. 4A). 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. 7B). When excluding the large number of relatives present in the Surui sample (\( \hat{H}_{507} \)) or correcting for their inclusion (\( \hat{H}_{1048} \)), the Surui population is not as extreme an outlier (fig. 7A and C).

### 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, \( \tilde{H} \) 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 \( \hat{H}_{507} \). The most noticeable of these populations are the Surui, Karitiana, and Pima populations from the Americas (fig. 5B); 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

---

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

| Estimator | P value for Wilcoxon Signed Rank Test | Mean of \( \hat{H}_{\text{reduced}} - \hat{H}_{\text{full}} \) across Populations | Mean of | \( \hat{H}_{\text{reduced}} - \hat{H}_{\text{full}} \) | Fraction of Populations with \( \hat{H}_{\text{full}} > \hat{H}_{\text{reduced}} \) |
|---|---|---|---|---|---|
| \( \hat{H}_{507} \) versus \( \hat{H}_{1048} \) | 8.804 \times 10^{-6} | 2.387 \times 10^{-3} | 2.387 \times 10^{-3} | 0 |
| \( \hat{H}_{507} \) versus \( \hat{H}_{1048} \) | 0.703 | 3.262 \times 10^{-4} | 6.660 \times 10^{-4} | 0.226 |
| \( \hat{H}_{507} \) versus \( \hat{H}_{603} \) | 2.980 \times 10^{-8} | 4.866 \times 10^{-4} | 4.866 \times 10^{-4} | 0 |
| \( \hat{H}_{603} \) versus \( \hat{H}_{603} \) | 0.708 | 6.649 \times 10^{-4} | 1.358 \times 10^{-3} | 0.462 |

In the header line, \( \hat{H}_{\text{reduced}} \) refers to \( \hat{H}_{507} \) or \( \hat{H}_{507} \) depending on which estimator is being considered; similarly, \( \hat{H}_{\text{full}} \) refers to \( \hat{H}_{1048} \), \( \hat{H}_{1048} \), \( \hat{H}_{507} \), or \( \hat{H}_{507} \).
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 $\hat{H}$ would not necessarily produce unbiased estimates in samples containing inbred individuals. $\hat{H}$ 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 $\hat{H}$ 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 $\hat{H}_{A,B} = 1 - \sum_{i=1}^{L} \hat{p}_i \hat{q}_i$, 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), $E[\sum_{i=1}^{L} \hat{p}_i^2 \hat{q}_i] = \sum_{i=1}^{L} \hat{p}_i \hat{q}_i$ (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, Hofshan 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, Veullie 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. Variazibilita 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