DOI: 10.1534/genetics.110.121756

Unbiased Estimation of Gene Diversity in Samples Containing Related Individuals: Exact Variance and Arbitrary Ploidy

Michael DeGiorgio,*,[1,2] Ivana Jankovic*,[1,3] and Noah A. Rosenberg*,[†]

**Center for Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, Michigan 48109 and [†]Department of Human Genetics and the Life Sciences Institute, University of Michigan, Ann Arbor, Michigan 48109*

Manuscript received August 3, 2010
Accepted for publication September 21, 2010

ABSTRACT

Gene diversity, a commonly used measure of genetic variation, evaluates the proportion of heterozygous individuals expected at a locus in a population, under the assumption of Hardy–Weinberg equilibrium. When using the standard estimator of gene diversity, the inclusion of related or inbred individuals in a sample produces a downward bias. Here, we extend a recently developed estimator shown to be unbiased in a diploid autosomal sample that includes known related or inbred individuals to the general case of arbitrary ploidy. We derive an exact formula for the variance of the new estimator, $\tilde{H}$, and present an approximation to facilitate evaluation of the variance when each individual is related to at most one other individual in a sample. When examining samples from the human X chromosome, which represent a mixture of haploid and diploid individuals, we find that $\tilde{H}$ performs favorably compared to the standard estimator, both in theoretical computations of mean squared error and in data analysis. We thus propose that $\tilde{H}$ is a useful tool in characterizing gene diversity in samples of arbitrary ploidy that contain related or inbred individuals.

FOR a given locus, gene diversity, also known as expected heterozygosity, characterizes the proportion of heterozygous genotypes expected in a population under Hardy–Weinberg equilibrium (Nei 1973). Nei and Roychoudhury (1974) devised an estimator of gene diversity that is unbiased for random samples of unrelated, noninbred individuals. When inbred individuals or close relatives are included in a sample, however, this estimator has a downward bias (Weir 1989; DeGiorgio and Rosenberg 2009). To account for the effects of inbreeding in a sample of diploid individuals, Weir (1989, 1996) derived the expected value of gene diversity, producing an unbiased estimator of gene diversity that makes use of the mean inbreeding coefficient across sampled individuals, where the inbreeding coefficient of an individual is defined as the probability for a randomly chosen locus that the two alleles of the individual are inherited identically by descent from a common ancestor. Using the mean kinship coefficient across pairs of sampled individuals, DeGiorgio and Rosenberg (2009) extended this estimator to account for the bias produced in samples containing close relatives, where the kinship coefficient between two individuals, j and k, is defined as the probability that an allele randomly selected from individual j at a random locus and an allele randomly selected from individual k at the same locus are identical by descent (IBD).

The DeGiorgio and Rosenberg (2009) estimator is useful for autosomal markers in samples from diploid organisms that contain related or inbred individuals. However, in studying gene diversity among related individuals in nondiploid cases (*e.g.*, Buteler *et al.* 1999) or in cases of mixed ploidy, such as in the analysis of sex chromosomes (*e.g.*, Reiland *et al.* 2002), unbiasedness for this estimator has not been demonstrated. Here, we extend the DeGiorgio and Rosenberg (2009) estimator of gene diversity to account for situations in which known related and inbred individuals are included in a sample and in which the sample contains an arbitrary mixture of individuals of different ploidy. We use a more general method to obtain the estimator than the method used for diploids by DeGiorgio and Rosenberg (2009), and we show that the general estimator reduces to the DeGiorgio and Rosenberg (2009) estimator in the diploid case. We also derive a formula for the variance of our estimator, $\tilde{H}$, to facilitate evaluation of the statistical properties of the estimator. This variance formula, which is a function of identity states among individuals, includes terms that involve identity-by-descent among two, three, and four individuals and among pairs of pairs of individuals. Our variance function is convenient because extensive work on IBD probabilities among individuals (*e.g.*, Cotterman

[1]These authors contributed equally to this work.

[2]*Corresponding author:* Center for Computational Medicine and Bioinformatics, University of Michigan, 2017 Palmer Commons, 100 Washtenaw Ave., Ann Arbor, MI 48109-2218.
E-mail: degiormi@umich.edu

[3]*Present address:* David Geffen School of Medicine, University of California, Los Angeles, CA 90095.

1940; Harris 1964; Gillois 1965; Cockerham 1971; Jacquard 1974; Thompson 1974; Lange 2002) has provided a framework for calculating the quantities incorporated in the formula.

Using the variance formula, we examine the performance of our estimator in scenarios involving the human X chromosome, for which males and females, who might both be included in a typical sample, differ in ploidy. In our evaluations, we first show that the exact theoretical values of the variance, which are obtained from a quite complex formula, are closely matched by simulations. We also validate that when each sampled individual is related to at most one other individual in the sample, the exact theoretical variance can be approximated well by a simpler formula. Using the variance approximation and simulations, we compare the behavior of our estimator to that of the Nei and Roychoudhury (1974) estimator, which does not account for relatives. We then analyze human SNPs from the X chromosome and find that $\tilde{H}$ also performs well in practice.

THEORY

Consider a sample of g groups, each with different ploidy (*e.g.*, haploid males and diploid females on the human X chromosome). Suppose that the sample from group b contains n_b m_b-ploid individuals, $b = 1, 2, \ldots, g$. Further, let (b, k), $k = 1, 2, \ldots, n_b$, denote individual k from group b. The number of copies of allelic type i in individual k from group b is

$$X^{(i)}_{(b,k)} = \sum_{\ell=1}^{m_b} A^{(i)}_{(b,k),\ell}, \tag{1}$$

where $A^{(i)}_{(b,k),\ell}$ is an indicator random variable that takes on the value 1 if the ℓth allele in individual (b, k) has type i and that equals 0 otherwise.

Note that $\mathbb{E}[A^{(i)}_{(b,k),\ell}] = p_i$, where p_i is the frequency of allelic type i in the population. We can then define an unbiased estimator for the frequency of allele i as

$$\hat{p}_i = \frac{1}{\sum_{b=1}^{g} n_b m_b} \sum_{b=1}^{g} \sum_{k=1}^{n_b} X^{(i)}_{(b,k)}. \tag{2}$$

Rewriting the estimator of Nei and Roychoudhury (1974) for the mixed-ploidy case, if no inbred or related individuals are included in the sample, then an unbiased estimator of gene diversity is

$$\hat{H} = \frac{\sum_{b=1}^{g} n_b m_b}{(\sum_{b=1}^{g} n_b m_b) - 1} \left(1 - \sum_{i=1}^{I} \hat{p}_i^2\right). \tag{3}$$

If inbred or related individuals are included in the sample, then $\hat{H}$ is a biased estimator of $H = 1 - \sum_{i=1}^{I} p_i^2$. We follow the approach of DeGiorgio and Rosenberg (2009), correcting for this bias by first obtaining the variance of sample allele frequencies. However, we use a different method here for obtaining the variance of sample allele frequencies, determining the bias correction for diploids as a special case of a more general computation.

An unbiased estimator: Suppose we have four possibly, but not necessarily, distinct individuals (a, j), (b, k), (a', j'), and (b', k'). Define $\Phi_{(a,j)(b,k)}$ as the probability that two alleles randomly chosen, one from individual (a, j) and the other from individual (b, k), are IBD. Similarly, define $\Phi_{(a,j)(b,k)(a',j')}$ as the probability that three alleles randomly chosen, one from (a, j), one from (b, k), and one from (a', j'), are IBD. Define $\Phi_{(a,j)(b,k)(a',j')(b',k')}$ as the probability that four alleles randomly chosen, one from (a, j), one from (b, k), one from (a', j'), and one from (b', k'), are IBD. Finally, define $\Phi_{(a,j)(b,k),(a',j')(b',k')}$ as the joint probability that two alleles randomly chosen, one from (a, j) and the other from (b, k), are IBD and two alleles randomly chosen, one from (a', j') and the other from (b', k'), are IBD. These four types of probability of identity-by-descent are identical to the θ, γ, δ, and Δ coefficients of Cockerham (1971), respectively. We can then define

$$\overline{\Phi}_2 = \sum_{a=1}^{g} \sum_{b=1}^{g} \sum_{j=1}^{n_a} \sum_{k=1}^{n_b} w_a w_b \Phi_{(a,j)(b,k)} \tag{4}$$

$$\overline{\Phi}_3 = \sum_{a=1}^{g} \sum_{b=1}^{g} \sum_{a'=1}^{g} \sum_{j=1}^{n_a} \sum_{k=1}^{n_b} \sum_{j'=1}^{n_{a'}} w_a w_b w_{a'} \Phi_{(a,j)(b,k)(a',j')} \tag{5}$$

$$\overline{\Phi}_4 = \sum_{a=1}^{g} \sum_{b=1}^{g} \sum_{a'=1}^{g} \sum_{b'=1}^{g} \sum_{j=1}^{n_a} \sum_{k=1}^{n_b} \sum_{j'=1}^{n_{a'}} \sum_{k'=1}^{n_{b'}} w_a w_b w_{a'} w_{b'} \Phi_{(a,j)(b,k)(a',j')(b',k')} \tag{6}$$

$$\overline{\Phi}_{2,2} = \sum_{a=1}^{g} \sum_{b=1}^{g} \sum_{a'=1}^{g} \sum_{b'=1}^{g} \sum_{j=1}^{n_a} \sum_{k=1}^{n_b} \sum_{j'=1}^{n_{a'}} \sum_{k'=1}^{n_{b'}} w_a w_b w_{a'} w_{b'} \Phi_{(a,j)(b,k),(a',j')(b',k')} \tag{7}$$

as weighted mean kinship coefficients across all sets of pairs, triplets, quartets, and pairs of pairs of individuals. The weight associated with an individual in group x, $w_x = m_x / \sum_{b=1}^{g} n_b m_b$, is proportional to the ploidy associated with the group. Define the inbreeding coefficient for individual (b, k), denoted by $f_{(b,k)}$, as the probability that two alleles randomly chosen without replacement from individual (b, k) are IBD and let $\overline{f}_b = (1/n_b) \sum_{k=1}^{n_b} f_{(b,k)}$ be the mean inbreeding coefficient across individuals in group b. This definition reduces to the standard definition for the diploid case.

In this section we first present two equations (Equations 8 and 9) that aid in the development of a generalized estimator of gene diversity (Theorem 1). This general estimator, the main result of the section, corrects the bias created by the inclusion of related and inbred individuals in a sample consisting of individuals with any mixture of ploidy. Using this estimator, we provide generalizations of results presented by DeGiorgio

and Rosenberg (2009) for diploids to the case of arbitrary ploidy (Equations 13 and 14) and we show how these generalizations can be reduced to the diploid case.

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 g groups, each with different ploidy, and n_b m_b-ploid individuals in group b, $b = 1, 2, \ldots, g$, each of whom is possibly inbred and related to other individuals in the sample. Consider the ℓth allele of individual (a, j) and the tth allele of individual (b, k). By definition of expected value, we have

$$\begin{aligned}\mathbb{E}\left[A^{(i)}_{(a,j),\ell}A^{(i)}_{(b,k),t}\right] &= \mathbb{P}\left[A^{(i)}_{(a,j),\ell} = 1, A^{(i)}_{(b,k),t} = 1\right] \\ &= \Phi_{(a,j)(b,k)}p_i + (1 - \Phi_{(a,j)(b,k)})p_i^2 \\ &= \Phi_{(a,j)(b,k)}p_i(1 - p_i) + p_i^2. \end{aligned} \tag{8}$$

In taking the expected value of our estimator of gene diversity, we need to evaluate the quantity $\mathbb{E}[\hat{p}_i^2]$. Using Equation 8, we show in appendix a that

$$\mathbb{E}[\hat{p}_i^2] = \overline{\Phi}_2 p_i(1 - p_i) + p_i^2. \tag{9}$$

Plugging Equations 8 and 9 into $\mathrm{Var}[\hat{p}_i] = \mathbb{E}[\hat{p}_i^2] - (\mathbb{E}[\hat{p}_i])^2$ yields $\mathrm{Var}[\hat{p}_i] = \overline{\Phi}_2 p_i(1 - p_i)$, which reduces to the result presented for the diploid case in Equation 7 of DeGiorgio and Rosenberg (2009), by reduction of the definition of $\overline{\Phi}_2$ for the diploid case. The following theorem provides a generalized unbiased estimator of gene diversity when a sample with any mixture of ploidy contains related or inbred individuals.

Theorem 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 g groups, each with different ploidy, and n_b m_b-ploid individuals in group b, $b = 1, 2, \ldots, g$, each of whom is possibly inbred and related to other individuals in the sample. Then*

$$\tilde{H} = \frac{1}{1 - \overline{\Phi}_2}\left(1 - \sum_{i=1}^{I} \hat{p}_i^2\right) \tag{10}$$

is an unbiased estimator for gene diversity.

The proof that $\tilde{H}$ is unbiased follows that of Proposition 1 in DeGiorgio and Rosenberg (2009), substituting the more general $\overline{\Phi}_2$ in place of the corresponding mean kinship coefficient in the earlier proof.

When reducing the definition of $\overline{\Phi}_2$ for the diploid case studied by DeGiorgio and Rosenberg (2009), the result in Theorem 1 is identical to the result presented for this case in Proposition 1 of DeGiorgio and Rosenberg (2009). One interesting consequence of Theorem 1 is that $\tilde{H}$ has a simple representation in terms of the sample proportion of identity-by-state and the probability of identity-by-descent computed on the basis of assumed levels of inbreeding and relationship. This representation is

$$\tilde{H} = \frac{1 - \hat{\mathbb{P}}[\mathrm{IBS}]}{1 - \mathbb{P}[\mathrm{IBD}]}, \tag{11}$$

where $\hat{\mathbb{P}}[\mathrm{IBS}]$ is the probability that two alleles in the sample, chosen uniformly at random with replacement, are identical by state, and $\mathbb{P}[\mathrm{IBD}]$ is the probability that two alleles in the sample, chosen uniformly at random with replacement, are identical by descent. A proof that Equation 11 is a consequence of Equation 10 is provided in appendix a. Note that Equations 10 and 11 have a connection to estimators of relatedness in a context in which relatedness is unknown. Such estimators essentially invert equations similar to Equation 11 to get estimators of $\overline{\Phi}_2$ (Ritland 1996; Rousset 2002).

We next seek to transform the estimator in Equation 10 into one that is more convenient for data analysis. Let $\mathcal{G}_{a,b}$, $a, b = 1, 2, \ldots, g$, be the set of distinct types of relative pairs for pairs of distinct individuals in a sample, one from group a and one from group b. Let η_R be the number of pairs of individuals with relationship type R in $\mathcal{G}_{a,b}$, and let Φ_R be the kinship coefficient for each of these pairs. Then, as shown in appendix a, we can write $\overline{\Phi}_2$ as

$$\begin{aligned}\overline{\Phi}_2 = &\frac{1}{\left(\sum_{b=1}^{g} n_b m_b\right)^2} \\ &\times\left[\sum_{b=1}^{g} n_b m_b + \sum_{b=1}^{g} n_b m_b(m_b - 1)\overline{f}_b + 2\sum_{b=1}^{g}\sum_{R\in\mathcal{G}_{b,b}} m_b^2\eta_R\Phi_R\right. \\ &\left. + 2\sum_{a=1}^{g-1}\sum_{b=a+1}^{g}\sum_{R\in\mathcal{G}_{a,b}} m_a m_b\eta_R\Phi_R\right].\end{aligned} \tag{12}$$

This version of $\overline{\Phi}_2$ is convenient for computation. To obtain a formula for $\tilde{H}$ that is convenient for computation and that is a generalized version of an analogous quantity for the diploid case in Equation 9 of DeGiorgio and Rosenberg (2009), we can substitute Equations 3 and 12 into Equation 10 to get

$$\tilde{H} = \frac{\left(\sum_{b=1}^{g} n_b m_b\right)\left(\sum_{b=1}^{g} n_b m_b - 1\right)}{D}\hat{H}, \tag{13}$$

where

$$\begin{aligned}D = &\left(\sum_{b=1}^{g} n_b m_b\right)\left(\sum_{b=1}^{g} n_b m_b - 1\right) - \sum_{b=1}^{g} n_b m_b(m_b - 1)\overline{f}_b \\ &- 2\sum_{b=1}^{g}\sum_{R\in\mathcal{G}_{b,b}} m_b^2\eta_R\Phi_R - 2\sum_{a=1}^{g-1}\sum_{b=a+1}^{g}\sum_{R\in\mathcal{G}_{a,b}} m_a m_b\eta_R\Phi_R.\end{aligned}$$

A proof of Equation 13 is provided in appendix a. We note that by using $g = 1$, $n_1 = n$, and $m_1 = 2$ in Equation 13, we obtain Equation 9 of DeGiorgio and Rosenberg (2009).

Note that $\tilde{H} = c\hat{H}$, where

$$c = \frac{(\sum_{b=1}^{g} n_b m_b)(\sum_{b=1}^{g} n_b m_b - 1)}{D}.$$

By rearranging and taking the expected value, we get $\mathbb{E}[\hat{H}] = \mathbb{E}[\tilde{H}]/c = H/c$. Therefore,

$$\begin{aligned}\text{bias}(\hat{H}) &= \frac{1-c}{c} H \\ &= -\frac{1}{(\sum_{b=1}^{g} n_b m_b)(\sum_{b=1}^{g} n_b m_b - 1)} \\ &\times \left[\sum_{b=1}^{g} n_b m_b (m_b - 1)\overline{f}_b + 2\sum_{b=1}^{g} \sum_{R \in \mathcal{G}_{b,b}} m_b^2 \eta_R \Phi_R \right. \\ &\left. + 2\sum_{a=1}^{g-1} \sum_{b=a+1}^{g} \sum_{R \in \mathcal{G}_{a,b}} m_a m_b \eta_R \Phi_R \right] H. \end{aligned} \tag{14}$$

Equation 14 is a generalized version of the bias formula in the diploid case, in Equation 11 of DeGiorgio and Rosenberg (2009). The bias is always negative and it has a magnitude that increases linearly with respect to H. Using $g = 1$, $n_1 = n$, and $m_1 = 2$ in Equation 14, we obtain Equation 11 of DeGiorgio and Rosenberg (2009).

Variance of the estimator: In the previous section, we derived an unbiased estimator $\tilde{H}$ of gene diversity in a sample of arbitrary ploidy. It is useful to determine the variance of the estimator, a quantity that in the diploid case DeGiorgio and Rosenberg (2009) obtained only by simulation. The following theorem provides a formula for the variance of the generalized estimator of gene diversity in samples with any mixture of ploidy.

Theorem 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 g groups, each with different ploidy, and n_b m_b-ploid individuals in group b, $b = 1, 2, \ldots, g$, each of whom is possibly inbred and related to other individuals in the sample. Then the variances of the $\tilde{H}$ and $\hat{H}$ estimators of gene diversity are*

$$\text{Var}[\tilde{H}] = \frac{1}{(1-\overline{\Phi}_2)^2} \text{Var}\left[1 - \sum_{i=1}^{I} \hat{p}_i^2\right] \tag{15}$$

and

$$\text{Var}[\hat{H}] = \left[\frac{\sum_{b=1}^{g} n_b m_b}{(\sum_{b=1}^{g} n_b m_b) - 1}\right]^2 \text{Var}\left[1 - \sum_{i=1}^{I} \hat{p}_i^2\right], \tag{16}$$

where

$$\begin{aligned}&\text{Var}\left[1 - \sum_{i=1}^{I} \hat{p}_i^2\right] \\ &= \overline{\Phi}_{2,2} - \overline{\Phi}_2^2 + 2\left[\overline{\Phi}_2^2 - \overline{\Phi}_4\right] \sum_{i=1}^{I} p_i^2 \\ &+ 4\left[2\overline{\Phi}_4 + \overline{\Phi}_2 - 2\overline{\Phi}_3 - \overline{\Phi}_{2,2}\right] \sum_{i=1}^{I} p_i^3 \\ &+ \left[3\overline{\Phi}_{2,2} + 8\overline{\Phi}_3 - 6\overline{\Phi}_4 - 4\overline{\Phi}_2 - \overline{\Phi}_2^2\right] \left(\sum_{i=1}^{I} p_i^2\right)^2. \end{aligned} \tag{17}$$

The proof of Theorem 2 is long and is provided in appendix b.

We next derive an approximate formula that in our calculations below we use in place of Equation 17 inside of Equations 15 and 16. The approximation is based only on pairwise kinship coefficients and is useful in cases in which the number of relatives in a sample is small enough that no individual is related to more than one other sampled individual. In such cases, the only nonzero terms included in $\overline{\Phi}_3$, $\overline{\Phi}_4$, and $\overline{\Phi}_{2,2}$ all involve sampling the same individual or pairs of individuals more than once. Thus, the $\overline{\Phi}_3$, $\overline{\Phi}_4$, and $\overline{\Phi}_{2,2}$ terms, along with $\overline{\Phi}_2^2$, are ignored, as they are likely to be much smaller than $\overline{\Phi}_2$ in cases in which the number of relationships in the sample is small.

In addition to the assumptions listed in Theorem 2, suppose that each individual in the sample is related to no more than one other individual in the sample. If we ignore terms involving $(\sum_{b=1}^{g} m_b n_b)^{-k}$, $k > 1$, then terms involving $\overline{\Phi}_2^2$, $\overline{\Phi}_3$, $\overline{\Phi}_4$, and $\overline{\Phi}_{2,2}$ in Equation 17 can be ignored. The only terms in Equation 17 that we retain are those of order $(\sum_{b=1}^{g} m_b n_b)^0$ and $(\sum_{b=1}^{g} m_b n_b)^{-1}$. $\overline{\Phi}_2$ is of order $(\sum_{b=1}^{g} m_b n_b)^{-1}$. Therefore, reducing Equation 17 leads to

$$\text{Var}\left[1 - \sum_{i=1}^{I} \hat{p}_i^2\right] \approx 4\overline{\Phi}_2 \left[\sum_{i=1}^{I} p_i^3 - \left(\sum_{i=1}^{I} p_i^2\right)^2\right]. \tag{18}$$

This formula is an approximation to Equation 17 when the number of relatives in a sample is small enough that no individual is related to more than one other sampled individual.

We now show that when no related individuals are included in a sample of diploids, the variance in Equation 18 is exactly the formula given by Weir (1989). Suppose a sample from a diploid population consists of n unrelated, but possibly inbred, individuals. Further suppose that we ignore terms involving n^{-k}, $k > 1$. Then $\Phi_{kk} = (1/2)(1 + f_k)$, where f_k is the inbreeding coefficient for individual k. We can write the mean pairwise kinship coefficient as

$$\overline{\Phi}_2 = \frac{1}{n^2} \sum_{k=1}^{n} \Phi_{kk} = \frac{1}{n^2} \sum_{k=1}^{n} \frac{1}{2}(1 + f_k) = \frac{1}{2n}(1 + \overline{f}),$$

where $\overline{f} = (1/n) \sum_{k=1}^{n} f_k$ is the mean inbreeding coefficient across individuals. Plugging $\overline{\Phi}_2 = (1 + \overline{f})/(2n)$ into Equation 18, we get

TABLE 1

Relationship types with corresponding X-linked kinship coefficients

Relationship no. (*k*)	Relationship type	Symbol for relationship class	Sexes of the pair	Φ
1	Full-sibs	t_1	Male–male	$\frac{1}{2}$
2	Half-sibs (female parent)	t_1	Male–male	$\frac{1}{2}$
3	Uncle–nephew (female parent)	t_2	Male–male	$\frac{1}{4}$
4	Grandfather–grandson (female parent)	t_1	Male–male	$\frac{1}{2}$
5	Parent–offspring	v_1	Female–female	$\frac{1}{4}$
6	Full-sibs	v_2	Female–female	$\frac{3}{8}$
7	Half-sibs (male parent)	v_1	Female–female	$\frac{1}{4}$
8	Half-sibs (female parent)	v_3	Female–female	$\frac{1}{8}$
9	Aunt–niece (male parent)	v_3	Female–female	$\frac{1}{8}$
10	Aunt–niece (female parent)	v_4	Female–female	$\frac{3}{16}$
11	Grandmother–granddaughter (male parent)	v_1	Female–female	$\frac{1}{4}$
12	Grandmother–granddaughter (female parent)	v_3	Female–female	$\frac{1}{8}$
13	Parent–offspring	u_1	Male–female	$\frac{1}{2}$
14	Full-sibs	u_2	Male–female	$\frac{1}{4}$
15	Half-sibs (female parent)	u_2	Male–female	$\frac{1}{4}$
16	Uncle–niece (male parent)	u_2	Male–female	$\frac{1}{4}$
17	Uncle–niece (female parent)	u_3	Male–female	$\frac{1}{8}$
18	Aunt–nephew (female parent)	u_4	Male–female	$\frac{3}{8}$
19	Grandfather–granddaughter (female parent)	u_2	Male–female	$\frac{1}{4}$
20	Grandmother–grandson (female parent)	u_2	Male–female	$\frac{1}{4}$

Relationship types with X-linked kinship coefficient of zero are not shown. These include the male–male relationships of parent–offspring, half-sibs (through male), uncle–nephew (through male), and grandfather–grandson (though male) as well as the male–female relationships of half-sibs (through male), aunt–nephew (through male), grandfather–granddaughter (through male), and grandmother–grandson (through male).

$$\mathrm{Var}\left[1-\sum_{i=1}^{I}\hat{p}_i^2\right]\approx\frac{2}{n}(1+\bar{f})\left[\sum_{i=1}^{I}p_i^3-\left(\sum_{i=1}^{I}p_i^2\right)^2\right]. \tag{19}$$

The X chromosome case: A common situation in which data of mixed ploidy arise is on sex chromosomes, for which members of one sex have two copies of a specific sex chromosome and members of the other sex have one copy. Later, we examine data on the human X chromosome, for which females have two copies and males have one. Thus, we now utilize Equation 13 to derive an unbiased estimator of gene diversity in samples from the X chromosome.

Consider an X-linked 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_F females and n_M males, each of whom is possibly inbred and related to other sampled individuals. Let $\mathcal{M}$, $\mathcal{F}$, and $\mathcal{U}$ be the sets of distinct types of male–male, female–female, and male–female relative pairs in the sample, respectively. Further, let η_R be the number of pairs of individuals with relationship type R and let Φ_R be the kinship coefficient for each of these pairs. Let males be group 1 and let females be group 2. Plugging $g = 2$, $n_1 = n_\mathrm{M}$, $n_2 = n_\mathrm{F}$, $m_1 = 1$, and $m_2 = 2$ into Equation 13, we obtain an unbiased estimator for gene diversity at an X-linked locus as

$$\hat{H}=\frac{(n_\mathrm{M}+2n_\mathrm{F})(n_\mathrm{M}+2n_\mathrm{F}-1)}{(n_\mathrm{M}+2n_\mathrm{F})(n_\mathrm{M}+2n_\mathrm{F}-1)-2n_\mathrm{F}\bar{f}_\mathrm{F}-2\sum_{R\in\mathcal{M}}\eta_R\Phi_R-8\sum_{R\in\mathcal{F}}\eta_R\Phi_R-4\sum_{R\in\mathcal{U}}\eta_R\Phi_R}\hat{H}, \tag{20}$$

where $\bar{f}_\mathrm{F} = (1/n_\mathrm{F})\sum_{k=1}^{n_\mathrm{F}} f_k$ is the mean inbreeding coefficient across female individuals and f_k is the inbreeding coefficient for female k.

The following special case of Equation 20 is useful for the examples we consider in subsequent sections. It

TABLE 2

Symbols used for relative pair types

Relationship class symbol	Υ_0	Υ_1	Υ_2	Φ	Sex	Relative types
t_1	$\frac{1}{2}$	$\frac{1}{2}$	0	$\frac{1}{2}$	Male–male	Full-sib, half-sib (through female parent), grandfather–grandson (through female)
t_2	$\frac{3}{4}$	$\frac{1}{4}$	0	$\frac{1}{4}$	Male–male	Uncle–nephew (through female)
v_1	0	1	0	$\frac{1}{4}$	Female–female	Parent–offspring, half-sib (through male parent), grandmother–granddaughter (through male)
v_2	0	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{3}{8}$	Female–female	Full-sib
v_3	$\frac{1}{2}$	$\frac{1}{2}$	0	$\frac{1}{8}$	Female–female	Half-sib (through female parent), aunt–niece (through male), grandmother–granddaughter (through female)
v_4	$\frac{1}{4}$	$\frac{3}{4}$	0	$\frac{3}{16}$	Female–female	Aunt–niece (through female)
u_1	0	1	0	$\frac{1}{2}$	Male–female	Parent–offspring
u_2	$\frac{1}{2}$	$\frac{1}{2}$	0	$\frac{1}{4}$	Male–female	Full-sib, half-sib (through female parent), uncle–niece (through male), grandfather–granddaughter (through female), grandmother–grandson (through female)
u_3	$\frac{3}{4}$	$\frac{1}{4}$	0	$\frac{1}{8}$	Male–female	Uncle–niece (through female)
u_4	$\frac{1}{4}$	$\frac{3}{4}$	0	$\frac{3}{8}$	Male–female	Aunt–nephew (through female)
t_3	—	—	—	$\frac{5}{24}$	Male–male	Uncertain second-degree relative
v_5	—	—	—	$\frac{17}{96}$	Female–female	Uncertain second-degree relative
u_5	—	—	—	$\frac{3}{20}$	Male–female	Uncertain second-degree relative

Υ_0, Υ_1, and Υ_2 designate the probabilities that individuals share 0, 1, and 2 alleles IBD at an X-linked locus, respectively. All types of relative pairs denoted by the same symbol have the same kinship coefficient, sexes, and probabilities of sharing 0, 1, and 2 alleles IBD. Φ can be calculated from Υ_1 and Υ_2 using $\Phi_{ij} = \Upsilon_1$ if i and j are both male, $\Phi_{ij} = \frac{1}{4}\Upsilon_1 + \frac{1}{2}\Upsilon_2$ if i and j are both female, and $\Phi_{ij} = \frac{1}{2}\Upsilon_1 + \Upsilon_2$ if i is male and j is female. For each possible pair of sexes (male–male, female–female, and male–female), the kinship coefficient for second-degree relatives of an uncertain type was found by averaging the kinship coefficients for all second-degree relationships in Table 1 with that pair of sexes, assuming that all were equally likely. Second-degree relationships include half-sib, grandparent–grandchild, and avuncular pairs. For male–male pairs, $t_3 = \left(2 \times \frac{1}{2} + 1 \times \frac{1}{4} + 3 \times 0\right)/6 = \frac{5}{24}$. For female–female pairs, $v_5 = \left(2 \times \frac{1}{4} + 3 \times \frac{1}{8} + 1 \times \frac{3}{16}\right)/6 = \frac{17}{96}$. For male–female pairs, $u_5 = \left(4 \times \frac{1}{4} + 1 \times \frac{1}{8} + 1 \times \frac{3}{8} + 4 \times 0\right)/10 = \frac{3}{20}$. The divisor in each of the previous equations describes the total number of possible second-degree relatives for that sex pair (*e.g.*, grandmother–grandson, aunt–nephew, etc., for the male–female case). This number includes second-degree relatives that are not related on the X chromosome, because the assignment of relationships in the data set was based on autosomal data. The kinship coefficients for t_3, v_5, and u_5 were used only for analysis of population data, and they were not used in our investigations of the effects on the estimators of varying the parameters.

makes use of Table 1, which shows the various types of relationships possible for the X chromosome in pairs of individuals. Suppose a noninbred sample from a population has n_F females and n_M males, among which η_k pairs of relationship type k are included. Let Φ_k be the kinship coefficient for each of these pairs. Because the sample is not inbred, the mean inbreeding coefficient across female individuals is $\overline{f}_F = 0$. Plugging $\overline{f}_F$ as well as η_k and Φ_k for each relationship type k (Table 1) into Equation 20, we obtain

$$\tilde{H} = \frac{(n_M + 2n_F)(n_M + 2n_F - 1)}{(n_M + 2n_F)(n_M + 2n_F - 1) - 2\sum_{k=1}^{4}\eta_k\Phi_k - 8\sum_{k=5}^{12}\eta_k\Phi_k - 4\sum_{k=13}^{20}\eta_k\Phi_k}\hat{H}. \quad (21)$$

DATA ANALYSIS

Data: We investigated the properties of $\tilde{H}$ on mixed-ploidy data using analytical computations of bias, variance, and mean squared error; simulations; and analysis of data from human populations. Our choices for simulation parameters were designed on the basis of values in the data. In our analytical computations and simulations, we based our assumed true allele frequencies on sample allele frequencies at 36 X-chromosomal loci typed in 950 unrelated individuals, 624 males and 326 females, from the Human Genome Diversity Panel (HGDP-CEPH) microsatellite data set of 1048 individuals (Ramachandran *et al.* 2008). Individuals 127 and 139 from the Ramachandran *et al.* (2008) data set were not included in our analyses. The 950 individuals were assumed to have no first- or second-degree relationships, on the basis of the Rosenberg (2006) analysis of the full HGDP-CEPH panel.

Our data analysis was performed on a data set of 13,052 X-chromosomal single-nucleotide polymorphism (SNP) loci genotyped in 485 individuals from 29 populations in the HGDP-CEPH panel (Jakobsson

TABLE 3

Comparison of exact, approximate, and simulation variances

Estimator	Locus	Relative pairs	Exact variance	Approximate variance	Simulation variance	Relative difference of approximation (%)
$\tilde{H}$	ATCT003	$10t_1$, $10u_2$, $10v_2$	7.55×10^{-4}	6.82×10^{-4}	7.48×10^{-4}	9.59
	DXS1068	$10t_1$, $10u_2$, $10v_2$	1.54×10^{-3}	1.47×10^{-3}	1.50×10^{-3}	4.82
	GATA48H04	$10t_1$, $10u_2$, $10v_2$	1.50×10^{-3}	1.48×10^{-3}	1.54×10^{-3}	1.52
	DXS1068	$20t_1$	3.62×10^{-3}	3.31×10^{-3}	3.49×10^{-3}	8.55
	DXS1068	$80t_1$	8.08×10^{-4}	7.82×10^{-4}	7.93×10^{-4}	3.16
	DXS1068	$20u_2$	1.97×10^{-3}	1.90×10^{-3}	1.97×10^{-3}	3.29
	DXS1068	$80u_2$	4.68×10^{-4}	4.60×10^{-4}	4.61×10^{-4}	1.71
	DXS1068	$20v_2$	1.99×10^{-3}	1.87×10^{-3}	1.95×10^{-3}	5.87
	DXS1068	$80v_2$	4.64×10^{-4}	4.53×10^{-4}	4.56×10^{-4}	2.37
$\hat{H}$	ATCT003	$10t_1$, $10u_2$, $10v_2$	7.45×10^{-4}	6.74×10^{-4}	7.38×10^{-4}	9.59
	DXS1068	$10t_1$, $10u_2$, $10v_2$	1.52×10^{-3}	1.45×10^{-3}	1.49×10^{-3}	4.82
	GATA48H04	$10t_1$, $10u_2$, $10v_2$	1.49×10^{-3}	1.46×10^{-3}	1.53×10^{-3}	1.52
	DXS1068	$20t_1$	3.53×10^{-3}	3.23×10^{-3}	3.40×10^{-3}	8.55
	DXS1068	$80t_1$	8.03×10^{-4}	7.77×10^{-4}	7.88×10^{-4}	3.16
	DXS1068	$20u_2$	1.95×10^{-3}	1.88×10^{-3}	1.94×10^{-3}	3.29
	DXS1068	$80u_2$	4.67×10^{-4}	4.59×10^{-4}	4.60×10^{-4}	1.71
	DXS1068	$20v_2$	1.95×10^{-3}	1.84×10^{-3}	1.91×10^{-3}	5.87
	DXS1068	$80v_2$	4.61×10^{-4}	4.51×10^{-4}	4.54×10^{-4}	2.37

The exact (Equations 15 and 16), approximate (Equation 18 inserted into Equations 15 and 16), and simulation variances were calculated for the combination of 10 male–male (t_1), 10 male–female (u_2), and 10 female–female (v_2) full-sib pairs at the ATCT003 ($H = 0.7794$), DXS1068 ($H = 0.7344$), and GATA48H04 ($H = 0.6476$) loci as well as for sets of 20 and 80 pairs of each full-sib pair type at DXS1068. Simulation variances were calculated over 100,000 replicates. The relative difference of the approximation was computed as $100 \times |\text{approximate variance} - \text{exact variance}|/(\text{exact variance})$.

et al. 2008). We also removed individuals related through the X chromosome, yielding a data set of 446 unrelated individuals. Unlike the Jakobsson *et al.* (2008) data set of 443 unrelated individuals, our set of 446 individuals did not retain individuals 866, 1046, or 1049, which are not in the H952 subset of the HGDP-CEPH panel. However, individuals 292, 451, 477, 983, 988, and 1089 were included in the data set of nonrelatives because they were all involved exclusively in male–male parent–offspring relationships and were therefore unrelated through the X chromosome to other sampled individuals.

Data analysis methods: We used simulations and analytical calculations to evaluate the behavior of the estimator $\tilde{H}$ for X-chromosomal loci under conditions of varying heterozygosities, sample sizes, and relationships of sampled individuals. We compared the relative performance of $\tilde{H}$ and $\hat{H}$ by applying $\tilde{H}$ and $\hat{H}$ to samples containing related individuals and $\hat{H}$ to samples in which relatives were removed so that no relative pairs remained. True allele frequencies were based on microsatellite sample allele frequencies (see *Data*). In the simulations, individuals of a relative pair were generated by randomly choosing the allele(s) of the first individual on the basis of the empirical allele frequency distribution from the data set. For a given type of relative pair, we then simulated the allele(s) of the second individual by copying alleles from the first individual using the probabilities of sharing zero, one, and two alleles IBD for that type of pair. Table 2 depicts these probabilities, as well as the symbols used here to denote the various classes of relative pairs. If only one allele was shared, then it was copied in the second individual from the first allele of the first (independently generated) individual. In cases of male–female relative pairs, the male was generated first and the second allele of the female was always chosen independently from the allele frequency distribution.

To create a reduced data set of unrelated individuals, the second (possibly dependent) individual was not included for same-sex pairs, whereas for male–female pairs, the male relative was removed. Thus, because each individual in our simulation was included in exactly one relative pair, the number of individuals used to calculate $\hat{H}$ for the unrelated sample was always half of that used for the other two estimators. Removing the male in male–female pairs results in the loss of one-third of the alleles, compared to a loss of one-half of the alleles for removal of an individual from a same-sex pair. Thus, compared to removing females, removing males from male–female pairs generates a larger sample of alleles while still ensuring that no individuals are related.

The value assumed for the true heterozygosity, H, of a specific locus, was calculated from the assumed true allele frequencies on the basis of genotypic data of the 950 unrelated individuals. In each simulated scenario, for each of the three estimators, this true heterozygosity

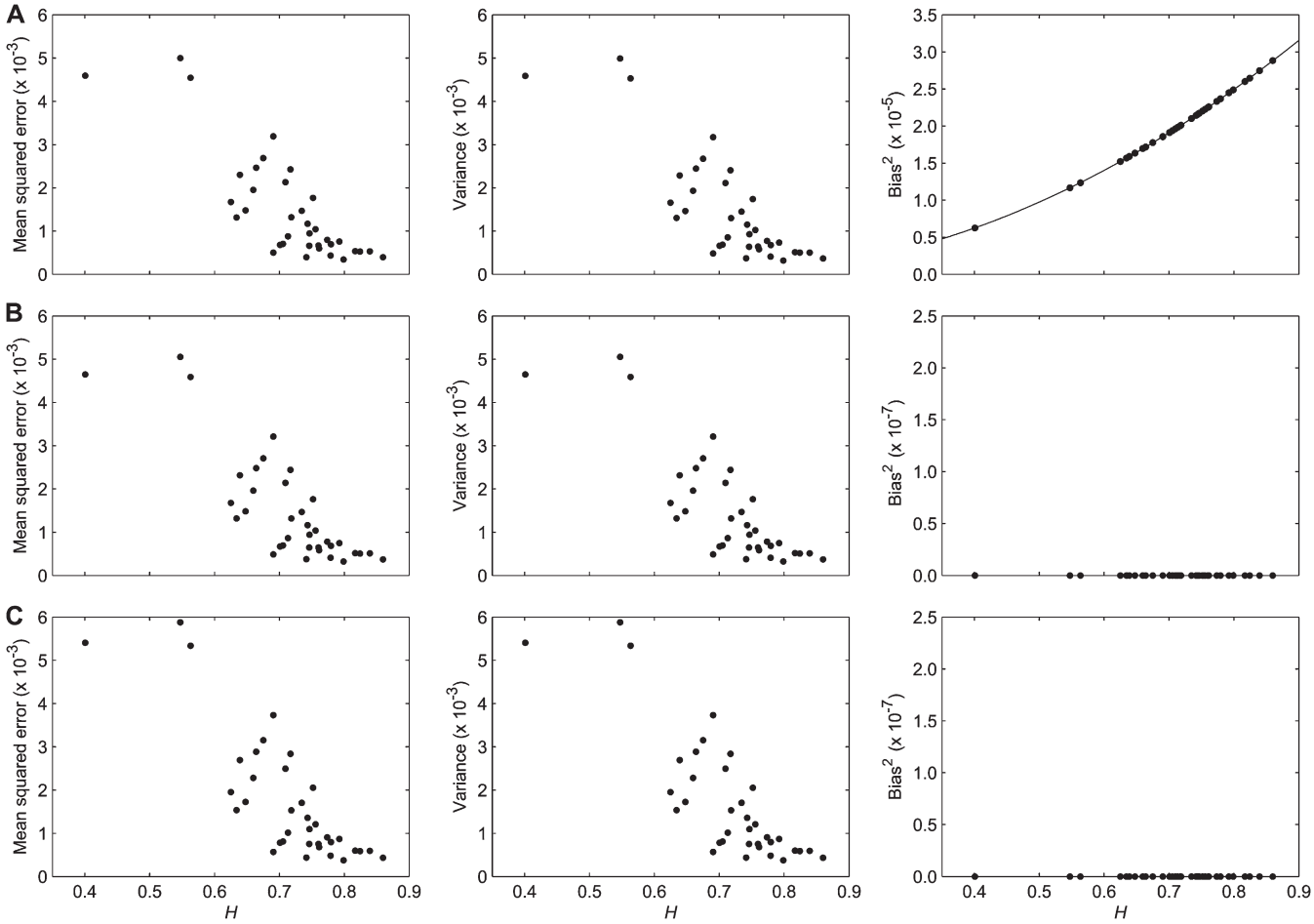

Figure 1.—Mean squared error, variance, and bias squared for each estimator, obtained analytically using the variance approximation (Equation 18 inserted into Equations 15 and 16), as a function of heterozygosity for 36 loci. The scheme considered included 60 individuals in 10 t_1 pairs ($\Phi = \frac{1}{2}$), 10 u_2 pairs ($\Phi = \frac{1}{4}$), and 10 v_2 pairs ($\Phi = \frac{3}{8}$). (A) $\hat{H}_{\text{full}}$. The curve through the points in the third column is described by Equation 14. (B) $\tilde{H}_{\text{full}}$. (C) $\hat{H}_{\text{reduced}}$.

was compared to the mean of the estimates produced by the estimator in 100,000 replicate simulations. The subscript *full* is used to denote cases in which an estimator was applied to the entire sample, whereas the subscript *reduced* indicates that relatives were removed from the sample. For a given scenario, the bias of each estimator was found by subtracting H from the mean value of the estimates for that estimator. Variance was calculated as the squared mean of the estimates across simulations subtracted from the mean across simulations of the squares of the estimates. Mean squared error (MSE) was then calculated as the sum of bias squared and variance.

Approximate variance: Because each of our analyses was performed on samples that contained only pairs of related individuals, the assumptions that underlie the derivation of the approximate variance (Equation 18) apply. We compared the exact, the approximate, and the simulated variance for $\tilde{H}$ and $\hat{H}$ in a series of cases that included only full-sib pairs. We chose nine representative cases of the various parameters that can affect estimator performance. Three of these cases considered an equal mix of male–male, female–female, and male–female full-sib pairs at the ATCT003 ($H = 0.7794$), DXS1068 ($H = 0.7344$), and GATA48H04 ($H = 0.6476$) loci, chosen to represent high, intermediate, and low heterozygosity, respectively. Additionally, we considered cases at the intermediate-heterozygosity locus involving 20 male–male, 80 male–male, 20 female–female, 80 female–female, 20 male–female, and 80 male–female pairs, to examine the effects of sample size and the sexes of the individuals. In each of our evaluations, we calculated the exact variances (Equations 15 and 16), approximate variances (Equation 18 plugged into Equations 15 and 16), and simulation variances obtained from 100,000 replicate simulations.

As Table 3 shows, in all cases examined, the exact, approximate, and simulated variances are similar, with the approximate variance slightly underestimating the exact variance. Because of the complexity of the formula for the exact variance, the difference between the approximate variance and the exact variance does not have a simple dependence on heterozygosity or sample size. However, it can be observed in Table 3 that for both

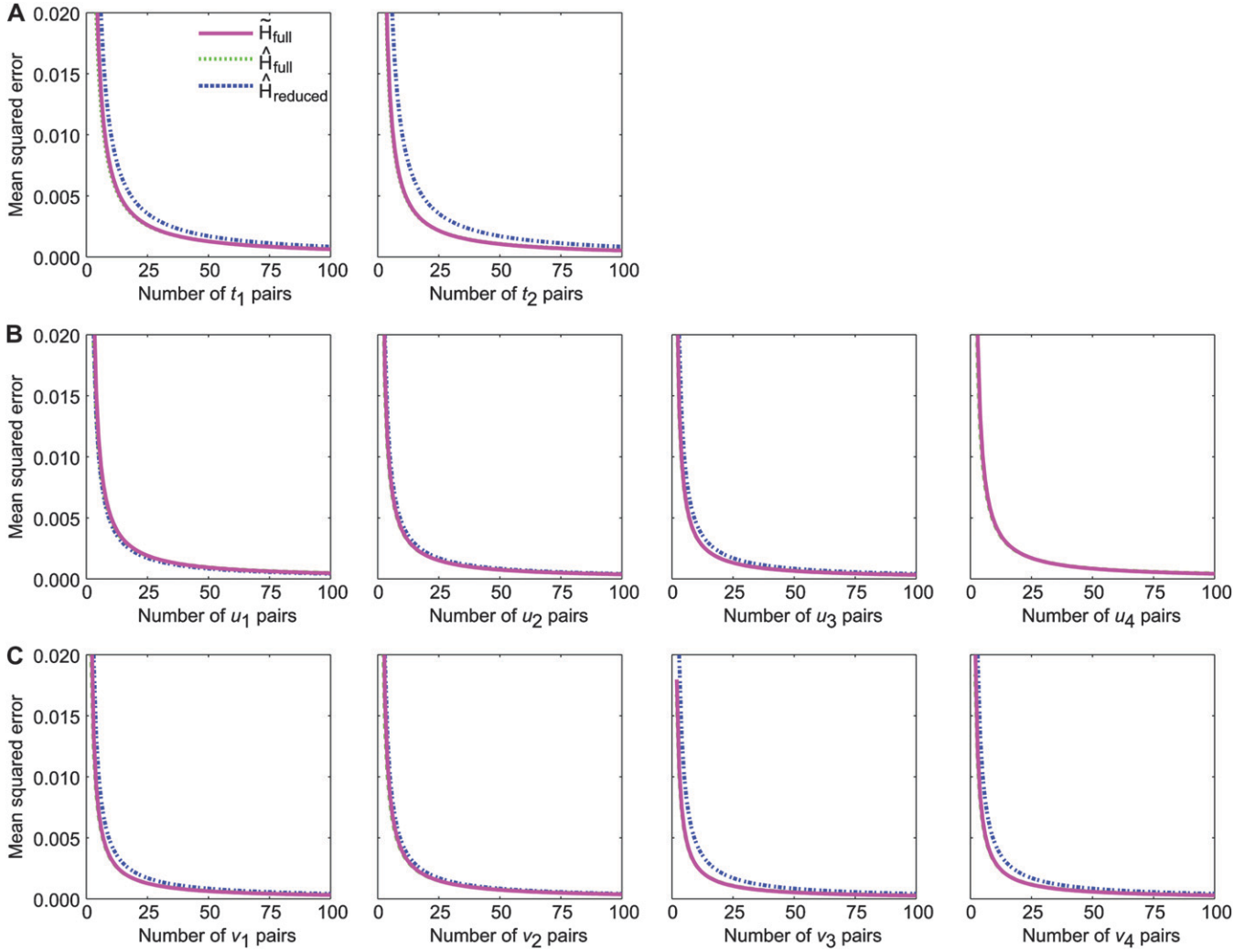

FIGURE 2.—Mean squared error as a function of sample size (number of pairs = number of individuals/2), calculated analytically using the variance approximation (Equation 18 inserted into Equations 15 and 16), on the basis of allele frequencies at the DXS1068 locus ($H = 0.7344$). Each plot considers different sample sizes for one type of relative pair (Table 2). The range of each plot is truncated at 0.020 and the graph of $\tilde{H}_{\text{full}}$ covers that of $\hat{H}_{\text{full}}$. (A) Male–male relative pairs. (B) Male–female relative pairs. (C) Female–female relative pairs. Note that the $\hat{H}_{\text{reduced}}$ line in the graph of mean squared error as a function of the number of u_4 pairs is behind the other two lines.

$\tilde{H}$ and $\hat{H}$, the relative difference between the approximate variance and exact variances is smallest at low heterozygosity and large sample size, typically near ~2%. In cases of high heterozygosity and small sample size, the relative difference remains at most ~10%. We note that the same approximation to the variance of $1 - \sum_{i=1}^{I} \hat{p}_i^2$ in Equation 18 is applied in obtaining the approximate variances of both $\tilde{H}$ and $\hat{H}$. Thus, because the approximation is generally reasonably accurate and because it treats $\tilde{H}$ and $\hat{H}$ in the same way, our use of the approximation is sensible in our subsequent comparisons of the mean squared errors of $\tilde{H}$ and $\hat{H}$.

Effect of parameters on the estimators: Several factors can potentially affect the performance of the estimators. These factors include the true value of heterozygosity itself, the sample size, the type of relative pair represented in the sample, and, if multiple types of relative pairs are included, the combination of particular types of relative pairs. We now examine each of these factors in sequence.

Varying heterozygosity: To investigate the influence of varying heterozygosity on the estimator, we evaluated the scenario of 60 related individuals in 10 t_1 pairs, 10 u_2 pairs, and 10 v_2 pairs (see Table 2) for each of the 36 X-linked microsatellite loci. This scheme incorporates 30 full-sib pairs, considering equally many males and females and utilizing three distinct kinship coefficients: $\frac{1}{2}$ for male–male pairs (t_1), $\frac{1}{4}$ for male–female pairs (u_2), and $\frac{3}{8}$ for female–female pairs (v_2). The 36 loci represent a spread of assumed true heterozygosities ranging from 0.4008 to 0.8599. For each locus, we calculated $\tilde{H}_{\text{full}}$ (Equation 21), as well as $\hat{H}_{\text{full}}$ and $\hat{H}_{\text{reduced}}$ (NEI and ROYCHOUDHURY 1974).

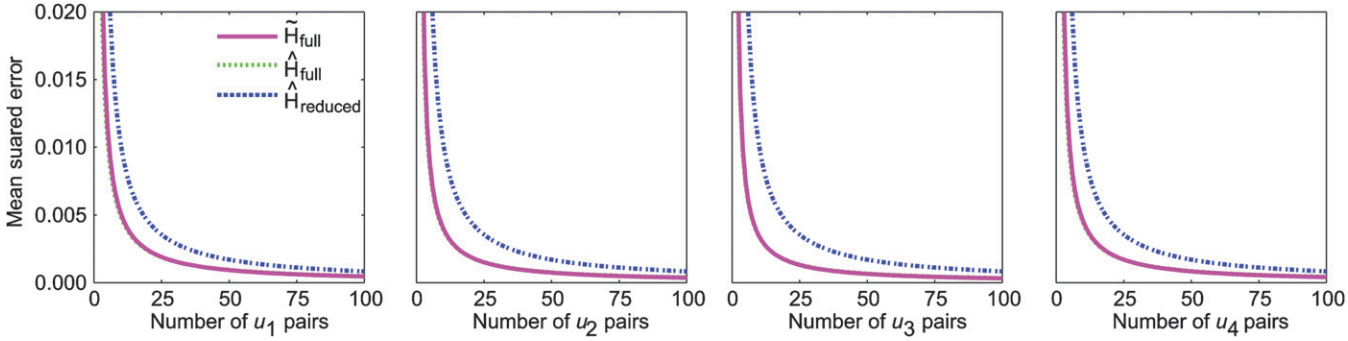

FIGURE 3.—Mean squared error as a function of sample size (number of pairs = number of individuals/2), calculated analytically using the variance approximation (Equation 18 inserted into Equations 15 and 16), on the basis of allele frequencies at the DXS1068 locus ($H = 0.7344$) for male–female relative pairs in which the females were removed to evaluate $\hat{H}_{reduced}$. The range of each plot is truncated at 0.020. The graph of $\tilde{H}_{full}$ covers that of $\hat{H}_{full}$.

Figure 1 displays the properties of the three estimators, $\tilde{H}_{full}$, $\hat{H}_{full}$, and $\hat{H}_{reduced}$, based on application of analytical computations of bias (Equation 14 for $\hat{H}_{full}$) and the variance approximation (Equation 18 plugged into Equations 15 and 16) to each of the 36 loci. $\tilde{H}_{full}$ and $\hat{H}_{reduced}$ are unbiased estimators and therefore have zero bias, whereas $\hat{H}_{full}$ exhibits increasing bias squared as heterozygosity increases. The bias squared for $\hat{H}_{full}$ as a function of heterozygosity is plotted using the theoretical prediction based on Equation 14: $\left[\text{bias}\left(\hat{H}\right)\right]^2 = \left(-\left(\left(2\left(10 \times \frac{1}{2}\right) + 8\left(10 \times \frac{3}{8}\right) + 4\left(10 \times \frac{1}{4}\right)\right)/\left((30 + 2 \times 30) \times (30 + 2 \times 30 - 1)\right)\right)H\right)^2 = (3.897 \times 10^{-5})H^2$. Generally, over the space of heterozygosities defined by the 36 microsatellite loci, the MSE and variance of all three estimators decrease with increasing heterozygosity.

Varying sample size and type of relative pair: We next applied the estimators to scenarios of varying sample size. The ATCT003 ($H = 0.7794$), DXS1068 ($H = 0.7344$), and GATA48H04 ($H = 0.6476$) loci were chosen from the data set to represent high, intermediate, and low heterozygosities, respectively. Only the data for the intermediate heterozygosity locus DXS1068 are shown; the other two loci yield similar results. For each locus and for each of the 10 types of relative pairs in Table 2, we varied the sample size from 2 to 100 pairs. We considered a sample size of at least 2 pairs, as no information is available for the computation of $\hat{H}_{reduced}$ from a single pair of male–male relatives. For all three loci, analytical calculations were performed using the variance approximation (Equation 18 plugged into Equations 15 and 16).

Figure 2 shows that as sample size increases, MSE decreases for all three estimators, and it is always comparable for $\tilde{H}_{full}$ and $\hat{H}_{full}$ ($\tilde{H}_{full}$ mostly overlaps $\hat{H}_{full}$ in Figure 2). Usually, we expect MSE in a reduced sample to be highest due to greater variance. However, although the results conformed to this prediction for most types of relative pairs, for male–female relative pairs for which there was probability $\geq \frac{3}{4}$ for sharing exactly one allele IBD (types u_1 and u_4), the MSE of $\hat{H}_{reduced}$ was actually lower than the MSE for $\tilde{H}_{full}$ and $\hat{H}_{full}$. The same result was also detected in our simulations (data not shown). Investigating further, we found that in male–male and female–female pairs, cases with high probabilities for sharing one or two alleles IBD had MSEs for $\tilde{H}_{full}$ and $\hat{H}_{full}$ that were closer to the $\hat{H}_{reduced}$ MSE values, compared with the higher MSE for $\hat{H}_{reduced}$ observed in other cases. The MSE of $\hat{H}_{reduced}$ is smaller relative to that of the other estimators for u_1 and u_4 male–female pairs because when only one-third of the sample is removed in creating the unrelated set of individuals (removal of males), the increase in variance due to the relatively small decrease in sample size in $\hat{H}_{reduced}$ is comparable to the increased variance caused by the high IBD probabilities for u_1 and u_4 pairs in $\tilde{H}_{full}$ and $\hat{H}_{full}$, unlike in other cases. When females, instead of males, are removed from male–female pairs, decreasing the sample by two-thirds rather than one-third, the estimators behave more intuitively (Figure 3), with $\hat{H}_{reduced}$ yielding the highest MSE.

Varying combinations of relative pairs: Finally, we studied the effect of relative pair combinations in a sample, using allele frequencies at the ATCT003, DXS1068, and GATA48H04 loci. Only the results for the highest heterozygosity locus, ATCT003, are shown; as was true in the previous section, each locus yielded similar results. For each locus, we examined each of the 231 possible divisions of exactly 20 full-sib pairs into male–male (t_1), male–female (u_2), and female–female (v_2) pairs. Figure 4 displays the MSE, variance, and bias squared of the three estimators, calculated analytically using the variance approximation (Equation 18), for various combinations of t_1, u_2, and v_2 pairs for the ATCT003 locus. Variance was highest for $\hat{H}_{reduced}$, because it had the smallest sample of alleles. For all estimators, variance was highest where the configuration of full-sibs had mostly male–male pairs, again due to the smaller sample of alleles. $\tilde{H}_{full}$ and $\hat{H}_{reduced}$ were unbiased across the space of possible combinations. $\hat{H}_{full}$ showed a trend in bias squared in which configurations with a greater proportion of males had higher bias squared, as is predicted analytically from the smaller sample size

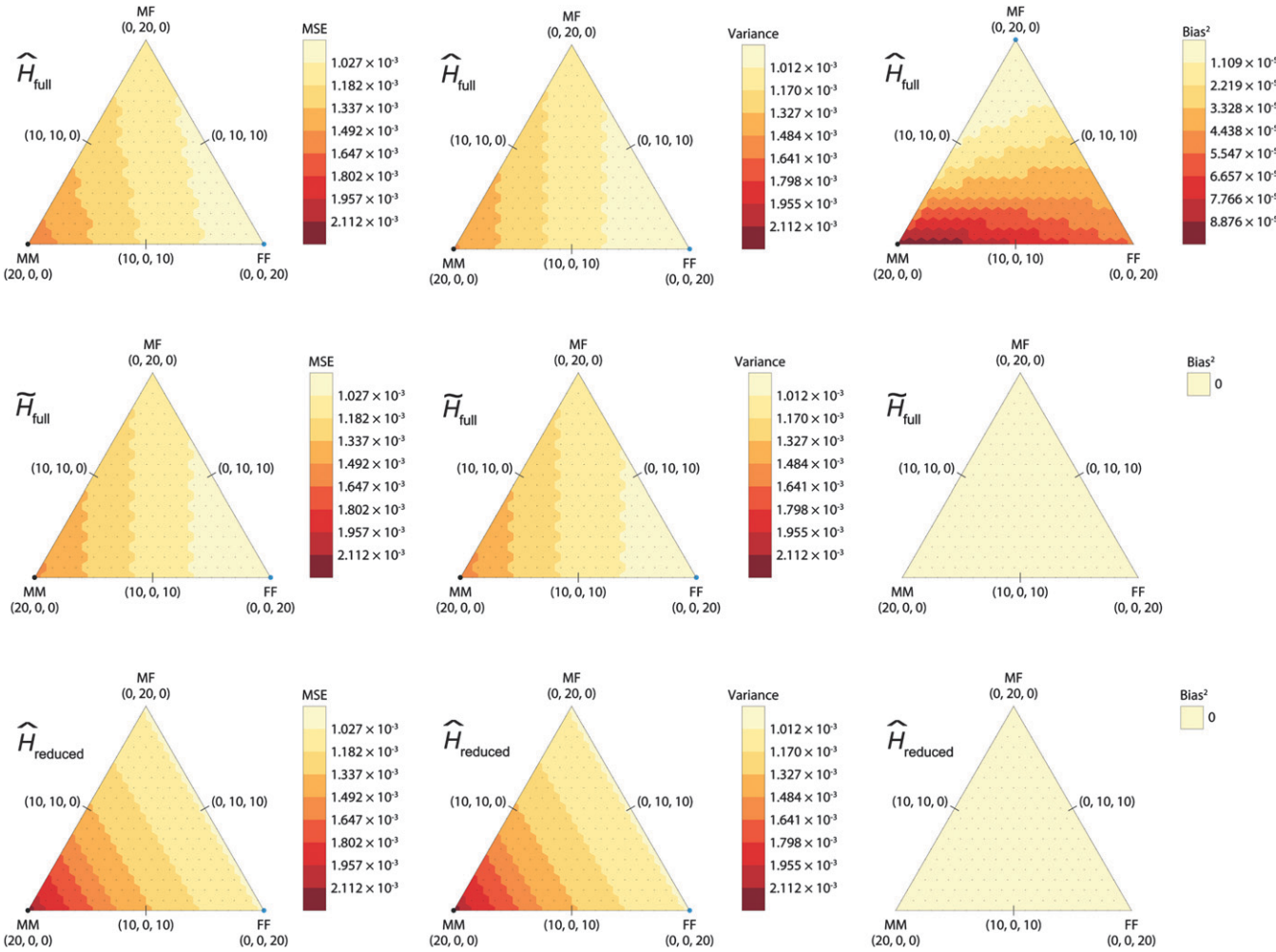

FIGURE 4.—Mean squared error (MSE), variance, and bias squared of $\hat{H}_{\text{full}}$, $\tilde{H}_{\text{full}}$, and $\hat{H}_{\text{reduced}}$, calculated analytically using the variance approximation (Equation 18 inserted into Equations 15 and 16), as functions of the configuration of t_1 male–male ($\Phi = \frac{1}{2}$), u_1 male–female ($\Phi = \frac{1}{2}$), and v_2 female–female ($\Phi = \frac{3}{8}$) pairs in 20 total relative pairs, on the basis of allele frequencies at the ATCT003 locus ($H = 0.7794$). Each row displays a different estimator and each column displays a different statistic. The three vertices of each triangle represent 20 male–male, 20 male–female, and 20 female–female full-sib pairs. The numbers on the scale indicate the cutoff values for colors. Note that unlike for the other two estimators, the scale for bias squared of $\hat{H}_{\text{full}}$ includes nonzero values. The black dot on each graph (except the bias squared graphs for $\tilde{H}_{\text{full}}$ and $\hat{H}_{\text{reduced}}$) represents the largest value in that triangle, and the blue dot represents the smallest value.

(Equation 14). For all configurations, the bias squared of $\hat{H}_{\text{full}}$ was greater than that for the other estimators. Among the three estimators, MSE was highest for $\hat{H}_{\text{reduced}}$. Similarly to the observation for variance, MSE was greatest for configurations with a high proportion of male–male pairs. Although $\tilde{H}_{\text{full}}$ performed slightly worse in having a greater variance compared to $\hat{H}_{\text{full}}$, it had a slightly lower MSE due to its lower bias. More generally, although $\tilde{H}_{\text{full}}$ performed better in the setting of Figure 4, the exact formula can be used to determine which estimator has lowest MSE for a given scenario.

Application to data: We next investigated the behavior of our estimator using X-chromosomal SNP data sets of 485 individuals and 446 unrelated individuals (see *Data*). Table 4 displays the relative pairs in the sample of 485 individuals. Because we analyzed the estimators separately by population, the subscripts of 485 and 446 refer to whether or not relatives were included in a calculation, not to the actual numbers of individuals in that calculation. In the same manner as in DEGIORGIO and ROSENBERG (2009), we took $\hat{H}_{446}$ for each population to be a proxy for true heterozygosity, because this quantity provided an unbiased estimate when no relatives were included in the sample. Note that removed individuals belonged only to pairs related through the X chromosome; individuals related only autosomally (such as male–male parent–offspring pairs) were included in the reduced sample. In our analysis, we compared the means of $\tilde{H}_{485}$ and $\hat{H}_{485}$ across the 13,052 loci to the corresponding mean of $\hat{H}_{446}$.

Figure 5 compares the difference between the mean of $\hat{H}_{485}$ across loci ($\overline{\hat{H}_{485}}$) and the mean of $\hat{H}_{446}$ ($\overline{\hat{H}_{446}}$) with the difference between the mean of $\tilde{H}_{485}$ ($\overline{\tilde{H}_{485}}$) and the mean of $\hat{H}_{446}$ ($\overline{\hat{H}_{446}}$). As Figure 5A shows, $\hat{H}_{485}$

TABLE 4

Types of relative pairs in populations from the data set of 485 individuals reported by Jakobsson *et al.* (2008)

	t_1	t_2	t_3	u_1	u_2	u_3	u_4	u_5	v_1	v_2	v_3	v_4	v_5
Bantu (Kenya)	1												
Bedouin				1				1					
Biaka Pygmy	1		2		1			2					
Druze					2				1				2
Kalash			1										
Mandenka			1										1
Maya					1				1				2
Mbuti Pygmy			1	1									
Melanesian	1			3					2	2			1
Mozabite										1			
Palestinian										1			1
Pima	1	1		1	1	1						1	
Yoruba				1	1				1	1			
Total	4	1	5	7	6	1	0	3	5	5	0	1	7

Symbols for the types of relative pairs appear in Table 2.

generally yields a lower heterozygosity estimate than $\hat{H}_{446}$ due to the downward bias caused by related individuals. Applying $\bar{H}_{485}$ reduces the magnitude of the difference between the estimate of heterozygosity in sets with and without relatives (Figure 5B), and $\bar{H}_{485}$ yields values that are not consistently lower than those of $\hat{H}_{446}$. It is important to note that because 15 of 45 of the relative pairs in the data have an uncertain second-degree relationship (t_3, u_5, or v_5), $\tilde{H}_{485}$ might have overcorrected bias in cases in which the individuals were not related via the X chromosome and undercorrected bias in cases in which the individuals actually were related on the X chromosome.

A Wilcoxon signed-rank test was used to evaluate the differences between $\bar{H}_{485}$ and $\hat{H}_{446}$ applied to the 13 populations that contained relatives (see Table 4). This test yielded a *P*-value of 0.0024, indicating that the inclusion of relatives had a significant impact on the estimation of heterozygosity using $\hat{H}$. In contrast, the Wilcoxon signed-rank comparison of $\tilde{H}_{485}$ and $\hat{H}_{446}$ yielded a *P*-value of 0.6355, indicating that the inclusion of relatives did not significantly alter the estimation of heterozygosity when $\tilde{H}$ was used. The mean difference $\bar{H}_{485} - \hat{H}_{446}$ (-8.0493×10^{-5}) and the mean absolute difference $|\bar{H}_{485} - \hat{H}_{446}|$ (6.3159×10^{-4}) were smaller across the 13 populations than the mean difference (-1.9393×10^{-3}) and the mean absolute difference (1.9849×10^{-3}), $\hat{H}_{446} - \bar{H}_{485}$ and $|\hat{H}_{446} - \bar{H}_{485}|$, respectively.

We also investigated the behavior of $\tilde{H}$ and $\hat{H}$ with regard to variance for the 13 populations that contained relatives. We compared $\hat{H}_{485} - \hat{H}_{446}$ and $\bar{H}_{485} - \hat{H}_{446}$, which we used as proxies for bias, following the methods of DeGiorgio and Rosenberg (2009), and the standard deviations of the two estimators applied with relatives included. From Figure 6, we observe that while there was a sizeable difference in the bias proxy between $\bar{H}_{485}$ and $\hat{H}_{485}$, there was only a small difference in standard deviation. This result is compatible with the results from our analytical computations, which suggest that $\tilde{H}$ corrects bias without substantially increasing variance.

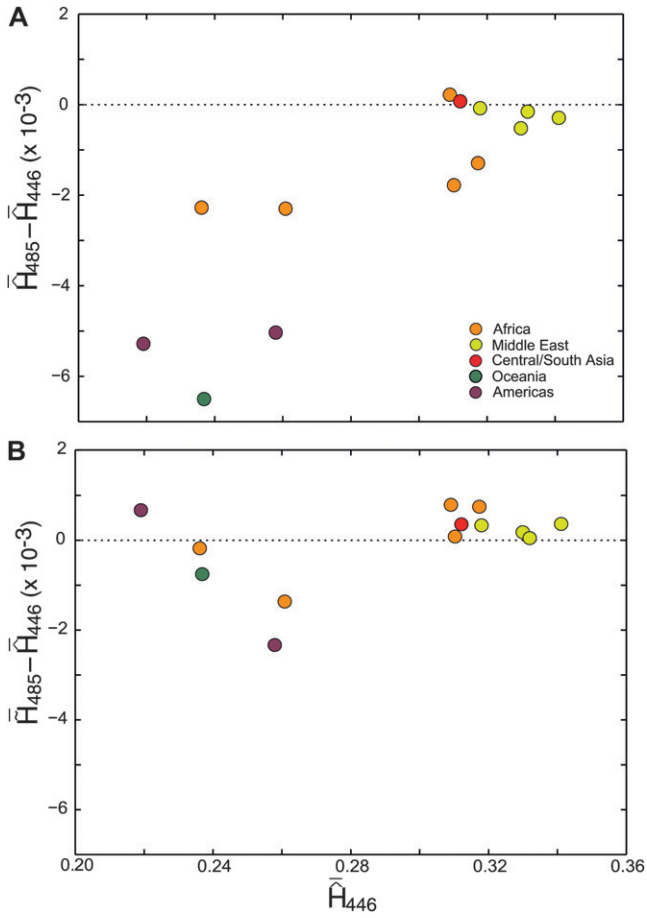

Figure 5.—Comparison of the difference between the mean of $\hat{H}_{485}$ across loci and the mean of $\hat{H}_{446}$ with the difference between the mean of $\tilde{H}_{485}$ and the mean of $\hat{H}_{446}$. (A) The difference between the mean of $\hat{H}_{485}$ and the mean of $\hat{H}_{446}$ for each of the 13 populations containing relatives (Table 4). (B) The difference between the mean of $\tilde{H}_{485}$ and the mean of $\hat{H}_{446}$ for each of the 13 populations. The estimators were applied to a data set of 13,052 SNP loci with 485 individuals belonging to 29 populations, and the results for the 13 populations with relatives are shown. Included in the set of 485 individuals was a subset of 446 individuals that contained no relatives. The subscripts of 485 and 446 refer to whether or not relatives were included, not to the actual number of individuals in the calculation. Each data point represents one population, with color indicating the geographic region of that population. The dotted line indicates a difference of zero.

DISCUSSION

Our estimator, $\tilde{H}$, is an effective tool for assessing the gene diversity of a sample of arbitrary ploidy containing related or inbred individuals. It can be used to provide unbiased estimates of expected heterozygosity when the inbreeding and kinship coefficients of sampled individ-

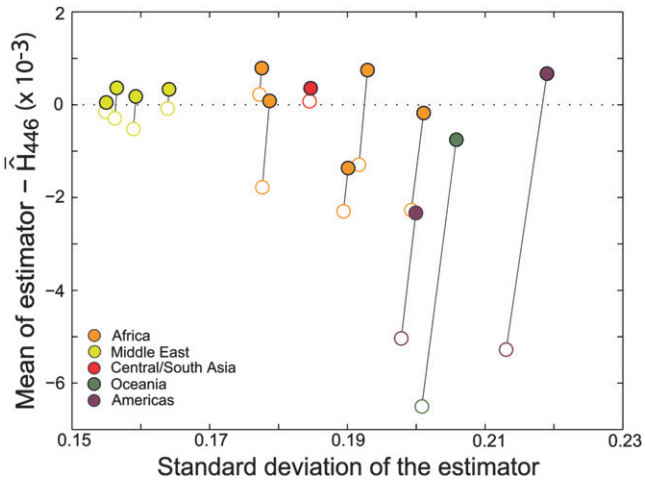

FIGURE 6.—Comparison of the difference between the mean of the estimator and the mean of $\hat{H}_{446}$ and standard deviation of the estimator, for the estimators $\tilde{H}_{485}$ and $\hat{H}_{485}$. These estimators were applied to a full data set of 13,052 X chromosome SNP loci with 485 individuals belonging to 29 populations, whereas 446 individuals were included in the reduced data set that contained no relatives. Only the 13 populations containing relatives are shown. The subscripts 485 and 446 refer to whether or not relatives were included, not to the actual number of individuals in the calculation. Open and solid points represent the estimates for $\hat{H}_{485}$ and $\tilde{H}_{485}$, respectively. The dotted line indicates a difference of zero. Lines connect data points representing the same population, with each population colored by geographic region.

uals are known. We found that the unbiasedness of the diploid estimator of DeGiorgio and Rosenberg (2009) extends to a much more general set of scenarios, provided that kinship coefficients are appropriately weighted by ploidy in the computation.

Here, we evaluated the properties of $\tilde{H}$ in the specific case of the human X chromosome. Through our analytical calculations, we have shown that, similarly to the DeGiorgio and Rosenberg (2009) estimator in the diploid case, the performance of $\tilde{H}$ is generally superior to that of $\hat{H}$ when the sample to which the estimators are applied contains relatives. $\tilde{H}$ accounts for the bias introduced by relatedness while simultaneously maintaining comparable MSE and variance to $\hat{H}$. Our estimator also performs well compared to $\hat{H}$ when applied to data from human populations. While the true heterozygosity of each population is not known, when we compared $\tilde{H}$ and $\hat{H}$ to an approximation of true heterozygosity, with $\hat{H}$ applied to the data set with no related individuals, we found that the difference between the estimate when relatives were included and when relatives were not included was significantly smaller for $\tilde{H}$. Because the reduction in this proxy for bias is accompanied by only a small increase in standard deviation, we argue that $\tilde{H}$ should often be preferred over $\hat{H}$ in the estimation of gene diversity in a sample containing relatives.

In addition to developing the $\tilde{H}$ estimator for gene diversity, we also determined the analytical variance of our estimator, allowing us to theoretically evaluate the properties of $\tilde{H}$. We also developed an approximation for variance (Equation 18) that is simpler to compute and that is applicable when each individual has at most one relative in the sample. Knowledge of the theoretical variance can further allow investigators to evaluate the circumstances under which $\tilde{H}$ applied to a full sample, including relatives, is superior to using $\hat{H}$ with a reduced sample in which members of relative pairs have been removed. For example, Figure 2 indicates that removing relatives will provide a lower MSE of the heterozygosity estimate in some cases. However, Figure 4 suggests that $\tilde{H}_{\text{full}}$ yields a lower MSE than $\hat{H}_{\text{reduced}}$ except in the small fraction of relative–pair combinations that contain large numbers of u_1 pairs. Thus, we propose that in most cases the use of $\tilde{H}$ on a sample set that includes related individuals affords a better estimate of gene diversity than applying $\hat{H}$ on a sample that contains no relatives and that investigators can use the theoretical variance of $\tilde{H}$ to determine whether a given situation is likely to be among the exceptions.

We thank Laurent Excoffier and three anonymous reviewers for their valuable comments. This work was supported by National Institutes of Health (NIH) grant R01 GM081441, NIH training grant T32 GM070449, a University of Michigan Rackham Merit Fellowship, and grants from the Burroughs Wellcome Fund and the Alfred P. Sloan Foundation.

LITERATURE CITED

Buteler, M. I., R. L. Jarret and D. R. LaBonte, 1999 Sequence characterization of microsatellites in diploid and polyploid *Ipomoea*. Theor. Appl. Genet. **99:** 123–132.

Cockerham, C. C., 1971 Higher order probability functions of identity of alleles by descent. Genetics **69:** 235–246.

Cotterman, C., 1940 A calculus for statistico-genetics. Ph.D. Thesis, Ohio State University, Columbus, OH. Reprinted in *Genetics and Social Structure*, pp. 157–272, edited by P. Ballonoff. Dowden, Hutchinson & Ross, Stroudsburg, PA, 1974.

DeGiorgio, M., and N. A. Rosenberg, 2009 An unbiased estimator of gene diversity in samples containing related individuals. Mol. Biol. Evol. **26:** 501–512.

Gillois, M., 1965 Relation d'identité en génétique. Ann. Inst. H. Poincaré Sect. B **2:** 1–94 (in French).

Harris, D. L., 1964 Genotypic covariances between inbred relatives. Genetics **50:** 1319–1348.

Jacquard, A., 1974 *The Genetic Structure of Populations*. Springer, New York.

Jakobsson, M., S. W. Scholz, P. Scheet, J. R. Gibbs, J. M. VanLiere *et al.*, 2008 Genotype, haplotype and copy-number variation in worldwide human populations. Nature **451:** 998–1003.

Lange, K., 2002 *Mathematical and Statistical Methods for Genetic Analysis*, Ed. 2. Springer, New York.

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

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

Ramachandran, S., N. A. Rosenberg, M. W. Feldman and J. Wakeley, 2008 Population differentiation and migration: coalescence times in a two-sex island model for autosomal and X-linked loci. Theor. Popul. Biol. **74:** 291–301.

Reiland, J., S. Hodge and M. A. F. Noor, 2002 Strong founder effect in *Drosophila pseudoobscura* colonizing New Zealand from North America. J. Hered. **93:** 415–420.

Ritland, K., 1996 Estimators for pairwise relatedness and individual inbreeding coefficients. Genet. Res. Camb. **67:** 175–185.

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

Rousset, F., 2002 Inbreeding and relatedness coefficients: What do they measure? Heredity **88:** 371–380.

Thompson, E. A., 1974 Gene identities and multiple relationships. Biometrics **30:** 667–680.

Weir, B. S., 1989 Sampling properties of gene diversity, pp. 23–42 in *Plant Population Genetics, Breeding and Genetic Resources*, edited by A. H. D. Brown, M. T. Clegg, A. L. Kahler and B. S. Weir Sinauer Associates, Sunderland, MA.

Weir, B. S., 1996 *Genetic Data Analysis II*. Sinauer Associates, Sunderland, MA.

Communicating editor: L. Excoffier

APPENDIX A

In this section, we present proofs for Equations 9, 11, 12, and 13.

Proof of Equation 9. Applying the definition of $\hat{p}_i$ and using Equation 8, we have

$$\begin{aligned}
\mathbb{E}\left[\hat{p}_i^2\right] &= \frac{1}{\left(\sum_{b=1}^{g} n_b m_b\right)^2} \sum_{a=1}^{g}\sum_{b=1}^{g}\sum_{j=1}^{n_a}\sum_{k=1}^{n_b} \mathbb{E}\left[X_{(a,j)}^{(i)} X_{(b,k)}^{(i)}\right] \\
&= \frac{1}{\left(\sum_{b=1}^{g} n_b m_b\right)^2} \sum_{a=1}^{g}\sum_{b=1}^{g}\sum_{j=1}^{n_a}\sum_{k=1}^{n_b}\sum_{\ell=1}^{m_a}\sum_{t=1}^{m_b} \mathbb{E}\left[A_{(a,j),\ell}^{(i)} A_{(b,k),t}^{(i)}\right] \\
&= \frac{1}{\left(\sum_{b=1}^{g} n_b m_b\right)^2} \sum_{a=1}^{g}\sum_{b=1}^{g}\sum_{j=1}^{n_a}\sum_{k=1}^{n_b}\sum_{\ell=1}^{m_a}\sum_{t=1}^{m_b} (\Phi_{(a,j)(b,k)} p_i(1-p_i) + p_i^2) \\
&= \frac{1}{\left(\sum_{b=1}^{g} n_b m_b\right)^2} \sum_{a=1}^{g}\sum_{b=1}^{g}\sum_{j=1}^{n_a}\sum_{k=1}^{n_b} m_a m_b (\Phi_{(a,j)(b,k)} p_i(1-p_i) + p_i^2) \\
&= \frac{\left(\sum_{b=1}^{g} n_b m_b\right)^2}{\left(\sum_{b=1}^{g} n_b m_b\right)^2} \overline{\Phi}_2 p_i(1-p_i) + \frac{\left(\sum_{b=1}^{g} n_b m_b\right)^2}{\left(\sum_{b=1}^{g} n_b m_b\right)^2} p_i^2 \\
&= \bar{\Phi}_2 p_i(1-p_i) + p_i^2.
\end{aligned}$$

■

Proof of Equation 11. $\hat{\mathbb{P}}[\text{IBS}] = \sum_{i=1}^{I} \hat{p}_i^2$. We only need to show that $\mathbb{P}[\text{IBD}] = \overline{\Phi}_2$. Note that while we write $\hat{\mathbb{P}}[\text{IBS}]$ as an estimate, $\mathbb{P}[\text{IBD}]$ depends only on quantities that are treated as known with certainty and we write it as a known quantity itself. Consider two alleles from the sample (that are not necessarily distinct). Let $C_{(a,j)(b,k)}$ denote the event that the first of the two alleles is from individual (a, j) and the second is from individual (b, k), where (a, j) and (b, k) are not necessarily distinct. Supposing that the two alleles are drawn uniformly at random from the sample, with replacement, let $\mathbb{P}\left[C_{(a,j)(b,k)}\right]$ denote the probability of event $C_{(a,j)(b,k)}$. Let $\mathbb{P}\left[\text{IBD} | C_{(a,j)(b,k)}\right]$ be the probability that two alleles are IBD given that the first allele is chosen from individual (a, j) and the second is chosen from individual (b, k). Then

$$\begin{aligned}
\mathbb{P}[\text{IBD}] = \sum_{b=1}^{g} \Bigg\{ &\sum_{k=1}^{n_b} \mathbb{P}\left[\text{IBD} \mid C_{(b,k)(b,k)}\right] \mathbb{P}\left[C_{(b,k)(b,k)}\right] + \sum_{j=1}^{n_b}\sum_{\substack{k=1 \\ k\neq j}}^{n_b} \mathbb{P}\left[\text{IBD} \mid C_{(b,j)(b,k)}\right] \mathbb{P}\left[C_{(b,j)(b,k)}\right] \Bigg\} \\
&+ \sum_{a=1}^{g}\sum_{\substack{b=1 \\ b\neq a}}^{g}\sum_{j=1}^{n_a}\sum_{k=1}^{n_b} \mathbb{P}\left[\text{IBD} \mid C_{(a,j)(b,k)}\right] \mathbb{P}\left[C_{(a,j)(b,k)}\right].
\end{aligned}$$

Note that, for individuals (a, j) and (b, k), which are not necessarily distinct,

$$\mathbb{P}\left[C_{(a,j)(b,k)}\right] = \left(\frac{m_a}{\sum_{c=1}^{g} n_c m_c}\right)\left(\frac{m_b}{\sum_{c=1}^{g} n_c m_c}\right) = \frac{m_a m_b}{\left(\sum_{c=1}^{g} n_c m_c\right)^2}$$
$$\mathbb{P}\left[\text{IBD} \mid C_{(a,j)(b,k)}\right] = \Phi_{(a,j)(b,k)}.$$

It follows that

$$\mathbb{P}[\text{IBD}] = \sum_{b=1}^{g}\left\{\sum_{k=1}^{n_b}\Phi_{(b,k)(b,k)}\frac{m_b^2}{\left(\sum_{c=1}^{g} n_c m_c\right)^2} + \sum_{j=1}^{n_b}\sum_{\substack{k=1\\k\neq j}}^{n_b}\Phi_{(b,j)(b,k)}\frac{m_b^2}{\left(\sum_{c=1}^{g} n_c m_c\right)^2}\right\}$$
$$+ \sum_{a=1}^{g}\sum_{\substack{b=1\\b\neq a}}^{g}\sum_{j=1}^{n_a}\sum_{k=1}^{n_b}\Phi_{(a,j)(b,k)}\frac{m_a m_b}{\left(\sum_{c=1}^{g} n_c m_c\right)^2}$$
$$= \frac{1}{\left(\sum_{b=1}^{g} n_b m_b\right)^2}\sum_{a=1}^{g}\sum_{b=1}^{g}\sum_{j=1}^{n_a}\sum_{k=1}^{n_b} m_a m_b \Phi_{(a,j)(b,k)}$$
$$= \overline{\Phi}_2. \qquad \blacksquare$$

Proof of Equation 12. For an m_b-ploid individual k, $\Phi_{(b,k)(b,k)} = 1/m_b + (1 - 1/m_b)f_{(b,k)} = (1/m_b)[1+(m_b-1)f_{(b,k)}]$. Note that $\Phi_{(a,j)(b,k)} = 0$ if individuals (a, j) and (b, k) are unrelated. We can then break $\overline{\Phi}_2$ into three components, considering three different types of pairs of individuals: same group–same individual, same group–different individual, and different group. Therefore

$$\overline{\Phi}_2 = \frac{1}{\left(\sum_{b=1}^{g} n_b m_b\right)^2}\sum_{a=1}^{g}\sum_{b=1}^{g}\sum_{j=1}^{n_a}\sum_{k=1}^{n_b} m_a m_b \Phi_{(a,j)(b,k)}$$
$$= \frac{1}{\left(\sum_{b=1}^{g} n_b m_b\right)^2}\left[\sum_{b=1}^{g}\sum_{k=1}^{n_b} m_b^2\Phi_{(b,k)(b,k)} + 2\sum_{b=1}^{g}\sum_{j=1}^{n_b-1}\sum_{k=j+1}^{n_b} m_b^2\Phi_{(b,j)(b,k)} + 2\sum_{a=1}^{g-1}\sum_{b=a+1}^{g}\sum_{j=1}^{n_a}\sum_{k=1}^{n_b} m_a m_b\Phi_{(a,j)(b,k)}\right]$$
$$= \frac{1}{\left(\sum_{b=1}^{g} n_b m_b\right)^2}\left[\sum_{b=1}^{g}\sum_{k=1}^{n_b} m_b^2\frac{1}{m_b}\left[1 + (m_b - 1)f_{(b,k)}\right] + 2\sum_{b=1}^{g}\sum_{R\in\mathcal{G}_{b,b}} m_b^2\eta_R\Phi_R + 2\sum_{a=1}^{g-1}\sum_{b=a+1}^{g}\sum_{R\in\mathcal{G}_{a,b}} m_a m_b\eta_R\Phi_R\right]$$
$$= \frac{1}{\left(\sum_{b=1}^{g} n_b m_b\right)^2}\left[\sum_{b=1}^{g} n_b m_b + \sum_{b=1}^{g} n_b m_b(m_b - 1)\overline{f}_b + 2\sum_{b=1}^{g}\sum_{R\in\mathcal{G}_{b,b}} m_b^2\eta_R\Phi_R + 2\sum_{a=1}^{g-1}\sum_{b=a+1}^{g}\sum_{R\in\mathcal{G}_{a,b}} m_a m_b\eta_R\Phi_R\right]. \qquad \blacksquare$$

Proof of Equation 13. First we note that

$$1 - \overline{\Phi}_2 = \frac{D}{\left(\sum_{b=1}^{g} n_b m_b\right)^2}.$$

Substituting $1 - \overline{\Phi}_2$ into $\tilde{H}$ (Equation 10) gives

$$\tilde{H} = \frac{\left(\sum_{b=1}^{g} n_b m_b\right)^2}{D}\left(1 - \sum_{i=1}^{I} \hat{p}_i^2\right).$$

Rearranging Equation 3 we get

$$1 - \sum_{i=1}^{I} \hat{p}_i^2 = \frac{\sum_{b=1}^{g} n_b m_b - 1}{\sum_{b=1}^{g} n_b m_b} \hat{H},$$

from which

$$\begin{aligned}\tilde{H} &= \frac{\left(\sum_{b=1}^{g} n_b m_b\right)^2}{D}\left(\frac{\sum_{b=1}^{g} n_b m_b - 1}{\sum_{b=1}^{g} n_b m_b}\hat{H}\right) \\ &= \frac{\left(\sum_{b=1}^{g} n_b m_b\right)\left(\sum_{b=1}^{g} n_b m_b - 1\right)}{D}\hat{H}.\end{aligned}$$

■

APPENDIX B

In this section, we present results that aid in the derivation of the variance of our gene diversity estimator. Lemma 3 derives certain expectations involving four alleles. These expectations are used to calculate the variance and covariance of squared allele frequency estimates in Lemma 4. Lemma 4 is then used to prove the variance formula in Theorem 2 when related and inbred individuals are included in a sample.

LEMMA 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 g groups, each with different ploidy, and n_b m_b-ploid individuals in group b, $b = 1, 2, \ldots, g$, each of whom is possibly inbred and related to other individuals in the sample. Consider the ℓth allele of individual (a, j), the tth allele of individual (b, k), the ℓ'th allele of individual (a', j'), and the t'th allele of individual (b', k'). For clarity, let $w = (a, j)$, $x = (b, k)$, $y = (a', j')$, and $z = (b', k')$. Then for allelic types i and $i' \neq i$,*

$$\begin{aligned}\mathbb{E}\left[A_{w,\ell}^{(i)} A_{x,t}^{(i)} A_{y,\ell'}^{(i)} A_{z,t'}^{(i)}\right] &= \Phi_{wxyz} p_i \\ &\quad + \left[\Phi_{wxy} + \Phi_{wxz} + \Phi_{wyz} + \Phi_{xyz} + \Phi_{wx,yz} + \Phi_{wy,xz} + \Phi_{wz,xy} - 7\Phi_{wxyz}\right] p_i^2 \\ &\quad + [12\Phi_{wxyz} + (\Phi_{wx} + \Phi_{wy} + \Phi_{wz} + \Phi_{xy} + \Phi_{xz} + \Phi_{yz}) \\ &\qquad - 3(\Phi_{wxy} + \Phi_{wxz} + \Phi_{wyz} + \Phi_{xyz}) \\ &\qquad - 2(\Phi_{wx,yz} + \Phi_{wy,xz} + \Phi_{wz,xy})] p_i^3 \\ &\quad + [1 + (\Phi_{wx,yz} + \Phi_{wy,xz} + \Phi_{wz,xy}) + 2(\Phi_{wxy} + \Phi_{wxz} + \Phi_{wyz} + \Phi_{xyz}) \\ &\qquad - 6\Phi_{wxyz} - (\Phi_{wx} + \Phi_{wy} + \Phi_{wz} + \Phi_{xy} + \Phi_{xz} + \Phi_{yz})] p_i^4 \qquad \text{(B1)}\end{aligned}$$

$$\begin{aligned}\mathbb{E}\left[A_{w,\ell}^{(i)} A_{x,t}^{(i)} A_{y,\ell'}^{(i')} A_{z,t'}^{(i')}\right] &= \left[\Phi_{wx,yz} - \Phi_{wxyz}\right] p_i p_{i'} + \left[2\Phi_{wxyz} + \Phi_{wx} - (\Phi_{wxy} + \Phi_{wxz}) - \Phi_{wx,yz}\right] p_i p_{i'}^2 \\ &\quad + \left[2\Phi_{wxyz} + \Phi_{yz} - (\Phi_{wyz} + \Phi_{xyz}) - \Phi_{wx,yz}\right] p_i^2 p_{i'} \\ &\quad + [1 + \Phi_{wx,yz} + \Phi_{wy,xz} + \Phi_{wz,xy} + 2(\Phi_{wxy} + \Phi_{wxz} + \Phi_{wyz} + \Phi_{xyz}) \\ &\qquad - 6\Phi_{wxyz} - (\Phi_{wx} + \Phi_{wy} + \Phi_{wz} + \Phi_{xy} + \Phi_{xz} + \Phi_{yz})] p_i^2 p_{i'}^2. \qquad \text{(B2)}\end{aligned}$$

Proof. We need to evaluate

$$\mathbb{E}\left[A_{w,\ell}^{(i)} A_{x,t}^{(i)} A_{y,\ell'}^{(i')} A_{z,t'}^{(i')}\right] = \sum_{s=1}^{15} \Delta_s \mathbb{P}\left[A_{w,\ell}^{(i)} = 1, A_{x,t}^{(i)} = 1, A_{y,\ell'}^{(i')} = 1, A_{z,t'}^{(i')} = 1 \mid S = s\right],$$

where S represents one of the 15 identity states in Figure B1 for four alleles—one from w, one from x, one from y, and one from z—and Δ_s is the identity coefficient, the probability of observing state $S = s$ for four alleles randomly chosen, one from w, one from x, one from y, and one from z. We can rewrite the identity coefficients in terms of kinship coefficients by using the following relationships:

$$\begin{aligned}
\sum_{s=1}^{15}\Delta_s &= 1\\
\Phi_{wxyz} &= \Delta_1\\
\Phi_{wxy} &= \Delta_1+\Delta_2\\
\Phi_{wxz} &= \Delta_1+\Delta_3\\
\Phi_{wyz} &= \Delta_1+\Delta_4\\
\Phi_{xyz} &= \Delta_1+\Delta_5\\
\Phi_{wx,yz} &= \Delta_1+\Delta_6\\
\Phi_{wy,xz} &= \Delta_1+\Delta_9\\
\Phi_{wz,xy} &= \Delta_1+\Delta_{12}\\
\Phi_{wx} &= \Delta_1+\Delta_2+\Delta_3+\Delta_6+\Delta_7\\
\Phi_{wy} &= \Delta_1+\Delta_2+\Delta_4+\Delta_9+\Delta_{10}\\
\Phi_{wz} &= \Delta_1+\Delta_3+\Delta_4+\Delta_{12}+\Delta_{13}\\
\Phi_{xy} &= \Delta_1+\Delta_2+\Delta_5+\Delta_{12}+\Delta_{14}\\
\Phi_{xz} &= \Delta_1+\Delta_3+\Delta_5+\Delta_9+\Delta_{11}\\
\Phi_{yz} &= \Delta_1+\Delta_4+\Delta_5+\Delta_6+\Delta_8. \qquad \text{(B3)}
\end{aligned}$$

Note that the Δ-coefficients above are identical to the δ-coefficients in Cockerham (1971). Also, the Φ-coefficients involving two individuals, three individuals, and pairs of pairs of individuals are identical to Cockerham's θ-, γ-, and Δ-coefficients, respectively (Cockerham 1971). If $i' = i$, we get

$$\mathbb{E}\left[A^{(i)}_{w,\ell}A^{(i)}_{x,t}A^{(i)}_{y,\ell'}A^{(i)}_{z,t'}\right] = \Delta_1 p_i + (\Delta_2+\Delta_3+\Delta_4+\Delta_5+\Delta_6+\Delta_9+\Delta_{12})p_i^2. \qquad \text{(B4)}$$

If $i \neq i'$, we get

$$\mathbb{E}\left[A^{(i)}_{w,\ell}A^{(i)}_{x,t}A^{(i')}_{y,\ell'}A^{(i')}_{z,t'}\right] = \Delta_6 p_i p_{i'} + \Delta_7 p_i p_{i'}^2 + \Delta_8 p_i^2 p_{i'} + \Delta_{15} p_i^2 p_{i'}^2. \qquad \text{(B5)}$$

The desired result follows by substituting Equation B3 into Equations B4 and B5. ■

Note that expressions mathematically identical to Equations B1 and B2 except with different notation appear in Table 1 of Cockerham (1971). However, a slight conceptual difference is that our formulas involve an expectation of a product among four arbitrary alleles, not necessarily four alleles in two pairs of diploid genotypes. We now use Lemma 3 to derive $\mathrm{Var}\left[\hat{p}_i^2\right]$ and $\mathrm{Cov}\left(\hat{p}_i^2,\ \hat{p}_{i'}^2\right)$.

Lemma 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 g groups, each with different ploidy, and* n_b m_b*-ploid individuals in group b,* $b = 1, 2, \ldots, g$, *each of whom is possibly inbred and related to other individuals in the sample. Then for allelic types i and* $i' \neq i$,

$$\begin{aligned}
\mathbb{E}\left[\hat{p}_i^4\right] &= \overline{\Phi}_4 p_i + \left[4\overline{\Phi}_3 + 3\overline{\Phi}_{2,2} - 7\overline{\Phi}_4\right]p_i^2 + \left[12\overline{\Phi}_4 + 6\overline{\Phi}_2 - 12\overline{\Phi}_3 - 6\overline{\Phi}_{2,2}\right]p_i^3\\
&\quad + \left[1 + 3\overline{\Phi}_{2,2} + 8\overline{\Phi}_3 - 6\overline{\Phi}_4 - 6\overline{\Phi}_2\right]p_i^4 \qquad \text{(B6)}\\
\mathbb{E}\left[\hat{p}_i^2\hat{p}_{i'}^2\right] &= \left[\overline{\Phi}_{2,2} - \overline{\Phi}_4\right]p_i p_{i'} + \left[2\overline{\Phi}_4 + \overline{\Phi}_2 - 2\overline{\Phi}_3 - \overline{\Phi}_{2,2}\right]p_i p_{i'}^2 + \left[2\overline{\Phi}_4 + \overline{\Phi}_2 - 2\overline{\Phi}_3 - \overline{\Phi}_{2,2}\right]p_i^2 p_{i'}\\
&\quad + \left[1 + 3\overline{\Phi}_{2,2} + 8\overline{\Phi}_3 - 6\overline{\Phi}_4 - 6\overline{\Phi}_2\right]p_i^2 p_{i'}^2 \qquad \text{(B7)}
\end{aligned}$$

and therefore

$$\begin{aligned}
\mathrm{Var}\left[\hat{p}_i^2\right] &= \overline{\Phi}_4 p_i + \left[4\overline{\Phi}_3 + 3\overline{\Phi}_{2,2} - 7\overline{\Phi}_4 - \overline{\Phi}_2^2\right]p_i^2 + \left[12\overline{\Phi}_4 + 4\overline{\Phi}_2 + 2\overline{\Phi}_2^2 - 12\overline{\Phi}_3 - 6\overline{\Phi}_{2,2}\right]p_i^3\\
&\quad + \left[3\overline{\Phi}_{2,2} + 8\overline{\Phi}_3 - 6\overline{\Phi}_4 - 4\overline{\Phi}_2 - \overline{\Phi}_2^2\right]p_i^4 \qquad \text{(B8)}\\
\mathrm{Cov}(\hat{p}_i^2,\ \hat{p}_{i'}^2) &= \left[\overline{\Phi}_{2,2} - \overline{\Phi}_4 - \overline{\Phi}_2^2\right]p_i p_{i'} + \left[2\overline{\Phi}_4 + \overline{\Phi}_2^2 - 2\overline{\Phi}_3 - \overline{\Phi}_{2,2}\right]p_i p_{i'}^2\\
&\quad + \left[2\overline{\Phi}_4 + \overline{\Phi}_2^2 - 2\overline{\Phi}_3 - \overline{\Phi}_{2,2}\right]p_i^2 p_{i'} + \left[3\overline{\Phi}_{2,2} + 8\overline{\Phi}_3 - 6\overline{\Phi}_4 - 4\overline{\Phi}_2 - \overline{\Phi}_2^2\right]p_i^2 p_{i'}^2. \qquad \text{(B9)}
\end{aligned}$$

Proof. Applying the definition of $\hat{p}_i$, we have

$$
\begin{aligned}
\mathbb{E}\big[\hat{p}_i^4\big] &= \frac{1}{\left(\sum_{b=1}^{g} n_b m_b\right)^4} \sum_{a=1}^{g}\sum_{b=1}^{g}\sum_{a'=1}^{g}\sum_{b'=1}^{g}\sum_{j=1}^{n_a}\sum_{k=1}^{n_b}\sum_{j'=1}^{n_{a'}}\sum_{k'=1}^{n_{b'}}\sum_{l=1}^{m_a}\sum_{t=1}^{m_b}\sum_{l'=1}^{m_{a'}}\sum_{t'=1}^{m_{b'}} \\
&\quad \times \mathbb{E}\Big[A^{(i)}_{(a,j),\ell} A^{(i)}_{(b,k),t} A^{(i)}_{(a',j'),\ell'} A^{(i)}_{(b',k'),t'}\Big] \\
&= \frac{1}{\left(\sum_{b=1}^{g} n_b m_b\right)^4} \sum_{a=1}^{g}\sum_{b=1}^{g}\sum_{a'=1}^{g}\sum_{b'=1}^{g}\sum_{j=1}^{n_a}\sum_{k=1}^{n_b}\sum_{j'=1}^{n_{a'}}\sum_{k'=1}^{n_{b'}} m_a m_b m_{a'} m_{b'} \\
&\quad \times \{\Phi_{(a,j)(b,k)(a',j')(b',k')} p_i \\
&\qquad + \big[\Phi_{(a,j)(b,k)(a',j')} + \Phi_{(a,j)(b,k)(b',k')} + \Phi_{(a,j)(a',j')(b',k')} + \Phi_{(b,k)(a',j')(b',k')} \\
&\qquad\quad + \Phi_{(a,j)(b,k),(a',j')(b',k')} + \Phi_{(a,j)(a',j'),(b,k)(b',k')} + \Phi_{(a,j)(b',k'),(b,k)(a',j')} \\
&\qquad\quad - 7\Phi_{(a,j)(b,k)(a',j')(b',k')}\big] p_i^2 \\
&\qquad + \big[12\Phi_{(a,j)(b,k)(a',j')(b',k')} \\
&\qquad\quad + \Phi_{(a,j)(b,k)} + \Phi_{(a,j)(a',j')} + \Phi_{(a,j)(b',k')} + \Phi_{(b,k)(a',j')} + \Phi_{(b,k)(b',k')} + \Phi_{(a',j')(b',k')} \\
&\qquad\quad - 3\big(\Phi_{(a,j)(b,k)(a',j')} + \Phi_{(a,j)(b,k)(b',k')} + \Phi_{(a,j)(a',j')(b',k')} + \Phi_{(b,k)(a',j')(b',k')}\big) \\
&\qquad\quad - 2\big(\Phi_{(a,j)(b,k),(a',j')(b',k')} + \Phi_{(a,j)(a',j'),(b,k)(b',k')} + \Phi_{(a,j)(b'k'),(b,k)(a',j')}\big)\big] p_i^3 \\
&\qquad + [1 + \Phi_{(a,j)(b,k),(a',j')(b',k')} + \Phi_{(a,j)(a',j'),(b,k)(b',k')} + \Phi_{(a,j)(b',k'),(b,k)(a',j')} \\
&\qquad\quad + 2\big(\Phi_{(a,j)(b,k)(a',j')} + \Phi_{(a,j)(b,k)(b',k')} + \Phi_{(a,j)(a',j')(b',k')} + \Phi_{(b,k)(a',j')(b',k')}\big) \\
&\qquad\quad - 6\Phi_{(a,j)(b,k)(a',j')(b',k')} \\
&\qquad\quad - (\Phi_{(a,j)(b,k)} + \Phi_{(a,j)(a',j')} + \Phi_{(a,j)(b',k')} + \Phi_{(b,k)(a',j')} + \Phi_{(b,k)(b',k')} \\
&\qquad\quad + \Phi_{(a',j')(b',k')})] p_i^4\} \\
&= \overline{\Phi}_4 p_i + \big[4\overline{\Phi}_3 + 3\overline{\Phi}_{2,2} - 7\overline{\Phi}_4\big] p_i^2 + \big[12\overline{\Phi}_4 + 6\overline{\Phi}_2 - 12\overline{\Phi}_3 - 6\overline{\Phi}_{2,2}\big] p_i^3 \\
&\qquad + \big[1 + 3\overline{\Phi}_{2,2} + 8\overline{\Phi}_3 - 6\overline{\Phi}_4 - 6\overline{\Phi}_2\big] p_i^4.
\end{aligned}
$$

For the case with alleles i and $i' \neq i$, we have

$$
\begin{aligned}
\mathbb{E}\big[\hat{p}_i^2 \hat{p}_{i'}^2\big] &= \frac{1}{\left(\sum_{b=1}^{g} n_b m_b\right)^4} \sum_{a=1}^{g}\sum_{b=1}^{g}\sum_{a'=1}^{g}\sum_{b'=1}^{g}\sum_{j=1}^{n_a}\sum_{k=1}^{n_b}\sum_{j'=1}^{n_{a'}}\sum_{k'=1}^{n_{b'}}\sum_{\ell=1}^{m_a}\sum_{t=1}^{m_b}\sum_{\ell'=1}^{m_{a'}}\sum_{t'=1}^{m_{b'}} \\
&\quad \times \mathbb{E}\Big[A^{(i)}_{(a,j),\ell} A^{(i)}_{(b,k),t} A^{(i')}_{(a',j'),\ell'} A^{(i')}_{(b',k'),t'}\Big] \\
&= \frac{1}{\left(\sum_{b=1}^{g} n_b m_b\right)^4} \sum_{a=1}^{g}\sum_{b=1}^{g}\sum_{a'=1}^{g}\sum_{b'=1}^{g}\sum_{j=1}^{n_a}\sum_{k=1}^{n_b}\sum_{j'=1}^{n_{a'}}\sum_{k'=1}^{n_{b'}} m_a m_b m_{a'} m_{b'} \\
&\quad \times \big\{\big[\Phi_{(a,j)(b,k),(a',j')(b',k')} - \Phi_{(a,j)(b,k)(a',j')(b',k')}\big] p_i p_{i'} \\
&\qquad + \big[2\Phi_{(a,j)(b,k)(a',j')(b',k')} + \Phi_{(a,j)(b,k)} \\
&\qquad\quad - \big(\Phi_{(a,j)(b,k)(a',j')} + \Phi_{(a,j)(b,k)(b',k')}\big) - \Phi_{(a,j)(b,k),(a',j')(b',k')}\big] p_i p_{i'}^2 \\
&\qquad + \big[2\Phi_{(a,j)(b,k)(a',j')(b',k')} + \Phi_{(a',j')(b',k')} \\
&\qquad - \big(\Phi_{(a,j)(a',j')(b',k')} + \Phi_{(b,k)(a',j')(b',k')}\big) - \Phi_{(a,j)(b,k),(a',j')(b',k')}\big] p_i^2 p_{i'} \\
&\qquad + \big[1 + \Phi_{(a,j)(b,k),(a',j')(b',k')} + \Phi_{(a,j)(a',j'),(b,k)(b',k')} + \Phi_{(a,j)(b',k'),(b,k)(a',j')} \\
&\qquad\quad + 2\big(\Phi_{(a,j)(b,k)(a',j')} + \Phi_{(a,j)(b,k)(b',k')} + \Phi_{(a,j)(a',j')(b',k')} + \Phi_{(b,k)(a',j')(b',k')}\big) \\
&\qquad\quad - 6\Phi_{(a,j)(b,k)(a',j')(b',k')} \\
&\qquad\quad - \big(\Phi_{(a,j)(b,k)} + \Phi_{(a,j)(a',j')} + \Phi_{(a,j)(b',k')} + \Phi_{(b,k)(a',j')} + \Phi_{(b,k)(b',k')} \\
&\qquad\quad + \Phi_{(a',j')(b',k')}\big)\big] p_i^2 p_{i'}^2\big\} \\
&= \big[\overline{\Phi}_{2,2} - \overline{\Phi}_4\big] p_i p_{i'} + \big[2\overline{\Phi}_4 + \overline{\Phi}_2 - 2\overline{\Phi}_3 - \overline{\Phi}_{2,2}\big] p_i p_{i'}^2 + \big[2\overline{\Phi}_4 + \overline{\Phi}_2 - 2\overline{\Phi}_3 - \overline{\Phi}_{2,2}\big] p_i^2 p_{i'} \\
&\quad + \big[1 + 3\overline{\Phi}_{2,2} + 8\overline{\Phi}_3 - 6\overline{\Phi}_4 - 6\overline{\Phi}_2\big] p_i^2 p_{i'}^2.
\end{aligned}
$$

Applying the definition of variance, we have

$$\begin{aligned}
\operatorname{Var}\left[\hat{p}_i^2\right] &= \mathbb{E}\left[\hat{p}_i^4\right] - \left(\mathbb{E}\left[\hat{p}_i^2\right]\right)^2 \\
&= \overline{\Phi}_4 p_i + \left[4\overline{\Phi}_3 + 3\overline{\Phi}_{2,2} - 7\overline{\Phi}_4\right] p_i^2 + \left[12\overline{\Phi}_4 + 6\overline{\Phi}_2 - 12\overline{\Phi}_3 - 6\overline{\Phi}_{2,2}\right] p_i^3 \\
&\quad + \left[1 + 3\overline{\Phi}_{2,2} + 8\overline{\Phi}_3 - 6\overline{\Phi}_4 - 6\overline{\Phi}_2\right] p_i^4 - \left[\overline{\Phi}_2 p_i(1-p_i) + p_i^2\right]^2 \\
&= \overline{\Phi}_4 p_i + \left[4\overline{\Phi}_3 + 3\overline{\Phi}_{2,2} - 7\overline{\Phi}_4 - \overline{\Phi}_2^2\right] p_i^2 + \left[12\overline{\Phi}_4 + 4\overline{\Phi}_2 + 2\overline{\Phi}_2^2 - 12\overline{\Phi}_3 - 6\overline{\Phi}_{2,2}\right] p_i^3 \\
&\quad + \left[3\overline{\Phi}_{2,2} + 8\overline{\Phi}_3 - 6\overline{\Phi}_4 - 4\overline{\Phi}_2 - \overline{\Phi}_2^2\right] p_i^4.
\end{aligned}$$

Applying the definition of covariance, we have

$$\begin{aligned}
\operatorname{Cov}(\hat{p}_i^2, \hat{p}_{i'}^2) &= \mathbb{E}\left[\hat{p}_i^2 \hat{p}_{i'}^2\right] - \mathbb{E}\left[\hat{p}_i^2\right]\mathbb{E}\left[\hat{p}_{i'}^2\right] \\
&= \left[\overline{\Phi}_{2,2} - \overline{\Phi}_4\right] p_i p_{i'} + \left[2\overline{\Phi}_4 + \overline{\Phi}_2 - 2\overline{\Phi}_3 - \overline{\Phi}_{2,2}\right] p_i p_{i'}^2 + \left[2\overline{\Phi}_4 + \overline{\Phi}_2 - 2\overline{\Phi}_3 - \overline{\Phi}_{2,2}\right] p_i^2 p_{i'} \\
&\quad + \left[1 + 3\overline{\Phi}_{2,2} + 8\overline{\Phi}_3 - 6\overline{\Phi}_4 - 6\overline{\Phi}_2\right] p_i^2 p_{i'}^2 - \left[\overline{\Phi}_2 p_i(1-p_i) + p_i^2\right]\left[\overline{\Phi}_2 p_{i'}(1-p_{i'}) + p_{i'}^2\right] \\
&= \left[\overline{\Phi}_{2,2} - \overline{\Phi}_4 - \overline{\Phi}_2^2\right] p_i p_{i'} + \left[2\overline{\Phi}_4 + \overline{\Phi}_2^2 - 2\overline{\Phi}_3 - \overline{\Phi}_{2,2}\right] p_i p_{i'}^2 \\
&\quad + \left[2\overline{\Phi}_4 + \overline{\Phi}_2^2 - 2\overline{\Phi}_3 - \overline{\Phi}_{2,2}\right] p_i^2 p_{i'} + \left[3\overline{\Phi}_{2,2} + 8\overline{\Phi}_3 - 6\overline{\Phi}_4 - 4\overline{\Phi}_2 - \overline{\Phi}_2^2\right] p_i^2 p_{i'}^2. \qquad \blacksquare
\end{aligned}$$

We now utilize Lemma 4 to prove Theorem 2.

Proof of Theorem 2. Applying the definition of variance, we have

$$\begin{aligned}
\operatorname{Var}\left[1 - \sum_{i=1}^{I} \hat{p}_i^2\right] &= \sum_{i=1}^{I}\sum_{i'=1}^{I} \operatorname{Cov}(\hat{p}_i^2, \hat{p}_{i'}^2) \\
&= \sum_{i=1}^{I} \operatorname{Var}\left[\hat{p}_i^2\right] + 2\sum_{i=1}^{I-1}\sum_{i'=i+1}^{I} \operatorname{Cov}(\hat{p}_i^2, \hat{p}_{i'}^2) \\
&= \sum_{i=1}^{I} \Big\{ \overline{\Phi}_4 p_i + \left[4\overline{\Phi}_3 + 3\overline{\Phi}_{2,2} - 7\overline{\Phi}_4 - \overline{\Phi}_2^2\right] p_i^2 \\
&\qquad + \left[12\overline{\Phi}_4 + 4\overline{\Phi}_2 + 2\overline{\Phi}_2^2 - 12\overline{\Phi}_3 - 6\overline{\Phi}_{2,2}\right] p_i^3 \\
&\qquad + \left[3\overline{\Phi}_{2,2} + 8\overline{\Phi}_3 - 6\overline{\Phi}_4 - 4\overline{\Phi}_2 - \overline{\Phi}_2^2\right] p_i^4 \Big\} \\
&\quad + 2\sum_{i=1}^{I-1}\sum_{i'=i+1}^{I} \Big\{ \left[\overline{\Phi}_{2,2} - \overline{\Phi}_4 - \overline{\Phi}_2^2\right] p_i p_{i'} + \left[2\overline{\Phi}_4 + \overline{\Phi}_2^2 - 2\overline{\Phi}_3 - \overline{\Phi}_{2,2}\right] p_i p_{i'}^2 \\
&\qquad + \left[2\overline{\Phi}_4 + \overline{\Phi}_2^2 - 2\overline{\Phi}_3 - \overline{\Phi}_{2,2}\right] p_i^2 p_{i'} \\
&\qquad + \left[3\overline{\Phi}_{2,2} + 8\overline{\Phi}_3 - 6\overline{\Phi}_4 - 4\overline{\Phi}_2 - \overline{\Phi}_2^2\right] p_i^2 p_{i'}^2 \Big\} \\
&= \overline{\Phi}_4 + \left[4\overline{\Phi}_3 + 3\overline{\Phi}_{2,2} - 7\overline{\Phi}_4 - \overline{\Phi}_2^2\right] \sum_{i=1}^{I} p_i^2 \\
&\quad + \left[12\overline{\Phi}_4 + 4\overline{\Phi}_2 + 2\overline{\Phi}_2^2 - 12\overline{\Phi}_3 - 6\overline{\Phi}_{2,2}\right] \sum_{i=1}^{I} p_i^3 \\
&\quad + \left[3\overline{\Phi}_{2,2} + 8\overline{\Phi}_3 - 6\overline{\Phi}_4 - 4\overline{\Phi}_2 - \overline{\Phi}_2^2\right] \sum_{i=1}^{I} p_i^4 \\
&\quad + 2\left[\overline{\Phi}_{2,2} - \overline{\Phi}_4 - \overline{\Phi}_2^2\right] \sum_{i=1}^{I-1}\sum_{i'=i+1}^{I} p_i p_{i'} \\
&\quad + 2\left[2\overline{\Phi}_4 + \overline{\Phi}_2^2 - 2\overline{\Phi}_3 - \overline{\Phi}_{2,2}\right] \sum_{i=1}^{I-1}\sum_{i'=i+1}^{I} p_i p_{i'}^2 \\
&\quad + 2\left[2\overline{\Phi}_4 + \overline{\Phi}_2^2 - 2\overline{\Phi}_3 - \overline{\Phi}_{2,2}\right] \sum_{i=1}^{I-1}\sum_{i'=i+1}^{I} p_i^2 p_{i'} \\
&\quad + 2\left[3\overline{\Phi}_{2,2} + 8\overline{\Phi}_3 - 6\overline{\Phi}_4 - 4\overline{\Phi}_2 - \overline{\Phi}_2^2\right] \sum_{i=1}^{I-1}\sum_{i'=i+1}^{I} p_i^2 p_{i'}^2.
\end{aligned}$$

Simplifying, we get

$$
\begin{aligned}
\mathrm{Var}\left[1-\sum_{i=1}^{I}\hat{p}_i^2\right] &= \overline{\Phi}_4 + 2\left[2\overline{\Phi}_3 + \overline{\Phi}_{2,2} - 3\overline{\Phi}_4\right]\sum_{i=1}^{I}p_i^2 + 4\left[2\overline{\Phi}_4 + \overline{\Phi}_2 - 2\overline{\Phi}_3 - \overline{\Phi}_{2,2}\right]\sum_{i=1}^{I}p_i^3 \\
&\quad + \left[\overline{\Phi}_{2,2} - \overline{\Phi}_4 - \overline{\Phi}_2^2\right]\left(\sum_{i=1}^{I}p_i^2 + 2\sum_{i=1}^{I-1}\sum_{i'=i+1}^{I}p_i p_{i'}\right) \\
&\quad + \left[2\overline{\Phi}_4 + \overline{\Phi}_2^2 - 2\overline{\Phi}_3 - \overline{\Phi}_{2,2}\right]\left(2\sum_{i=1}^{I}p_i^3 + 2\sum_{i=1}^{I-1}\sum_{i'=i+1}^{I}\left[p_i^2 p_{i'} + p_i p_{i'}^2\right]\right) \\
&\quad + \left[3\overline{\Phi}_{2,2} + 8\overline{\Phi}_3 - 6\overline{\Phi}_4 - 4\overline{\Phi}_2 - \overline{\Phi}_2^2\right]\left(\sum_{i=1}^{I}p_i^4 + 2\sum_{i=1}^{I-1}\sum_{i'=i+1}^{I}p_i^2 p_{i'}^2\right) \\
&= \overline{\Phi}_4 + 2\left[2\overline{\Phi}_3 + \overline{\Phi}_{2,2} - 3\overline{\Phi}_4\right]\sum_{i=1}^{I}p_i^2 + 4\left[2\overline{\Phi}_4 + \overline{\Phi}_2 - 2\overline{\Phi}_3 - \overline{\Phi}_{2,2}\right]\sum_{i=1}^{I}p_i^3 \\
&\quad + \left[\overline{\Phi}_{2,2} - \overline{\Phi}_4 - \overline{\Phi}_2^2\right]\sum_{i=1}^{I}\sum_{i'=1}^{I}p_i p_{i'} \\
&\quad + 2\left[2\overline{\Phi}_4 + \overline{\Phi}_2^2 - 2\overline{\Phi}_3 - \overline{\Phi}_{2,2}\right]\sum_{i=1}^{I}\sum_{i'=1}^{I}p_i^2 p_{i'} \\
&\quad + \left[3\overline{\Phi}_{2,2} + 8\overline{\Phi}_3 - 6\overline{\Phi}_4 - 4\overline{\Phi}_2 - \overline{\Phi}_2^2\right]\sum_{i=1}^{I}\sum_{i'=1}^{I}p_i^2 p_{i'}^2 \\
&= \overline{\Phi}_4 + 2\left[2\overline{\Phi}_3 + \overline{\Phi}_{2,2} - 3\overline{\Phi}_4\right]\sum_{i=1}^{I}p_i^2 + 4\left[2\overline{\Phi}_4 + \overline{\Phi}_2 - 2\overline{\Phi}_3 - \overline{\Phi}_{2,2}\right]\sum_{i=1}^{I}p_i^3 \\
&\quad + \left[\overline{\Phi}_{2,2} - \overline{\Phi}_4 - \overline{\Phi}_2^2\right]\left(\sum_{i=1}^{I}p_i\right)^2 \\
&\quad + 2\left[2\overline{\Phi}_4 + \overline{\Phi}_2^2 - 2\overline{\Phi}_3 - \overline{\Phi}_{2,2}\right]\left(\sum_{i=1}^{I}p_i^2\right)\left(\sum_{i=1}^{I}p_i\right) \\
&\quad + \left[3\overline{\Phi}_{2,2} + 8\overline{\Phi}_3 - 6\overline{\Phi}_4 - 4\overline{\Phi}_2 - \overline{\Phi}_2^2\right]\left(\sum_{i=1}^{I}p_i^2\right)^2 \\
&= \overline{\Phi}_{2,2} - \overline{\Phi}_2^2 + 2\left[\overline{\Phi}_2^2 - \overline{\Phi}_4\right]\sum_{i=1}^{I}p_i^2 + 4\left[2\overline{\Phi}_4 + \overline{\Phi}_2 - 2\overline{\Phi}_3 - \overline{\Phi}_{2,2}\right]\sum_{i=1}^{I}p_i^3 \\
&\quad + \left[3\overline{\Phi}_{2,2} + 8\overline{\Phi}_3 - 6\overline{\Phi}_4 - 4\overline{\Phi}_2 - \overline{\Phi}_2^2\right]\left(\sum_{i=1}^{I}p_i^2\right)^2.
\end{aligned}
$$

Applying the identity $\mathrm{Var}\left[\left(1-\sum_{i=1}^{I}\hat{p}_i^2\right)/\left(1-\overline{\Phi}_2\right)\right] = \mathrm{Var}\left[1-\sum_{i=1}^{I}\hat{p}_i^2\right]/\left(1-\overline{\Phi}_2\right)^2$ gives Equation 15. ■

It is interesting (and convenient) that although the derivation requires the use of all 15 Δ-coefficients, the only coefficients required in the variance formula are $\overline{\Phi}_2$, $\overline{\Phi}_3$, $\overline{\Phi}_4$, and $\overline{\Phi}_{2,2}$. The 15 Δ-coefficients in Figure B1 completely specify the 14 Φ-coefficients in Equation B3 (along with the 15th Φ-coefficient equal to Δ_{15}). Through symmetry of the 6 Φ-coefficients involving two individuals, symmetry of the 4 Φ-coefficients involving three individuals, and symmetry of the 3 Φ-coefficients involving pairs of pairs of individuals, by averaging over sets of individuals, the variance of gene diversity becomes a function of only 4 average Φ-coefficients.

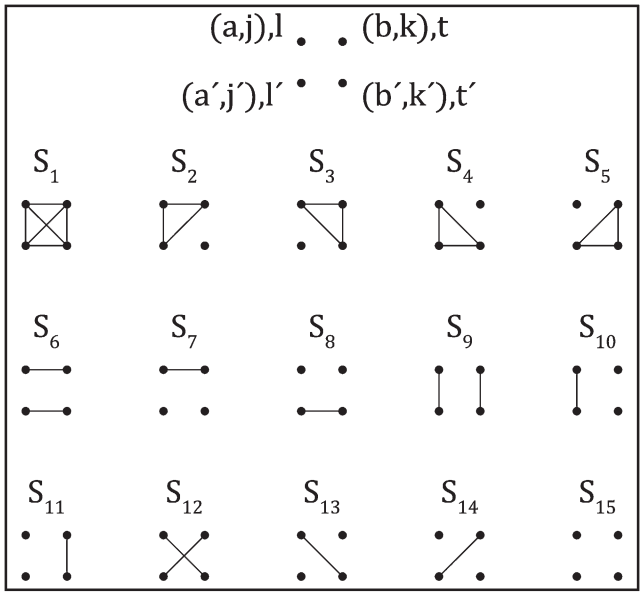

FIGURE B1.—Identity states. Two alleles (dots) are identical by descent if and only if there is a line connecting them. This figure is similar to Figure 6.2 of JACQUARD (1974) and is reproduced here for convenience.