# Letters to the Editor

## Multiple Alleles and Estimation of Genetic Parameters: Computational Equations Showing Involvement of All Alleles

Genetic loci that exhibit multiple (more than two) segregating alleles are generally more useful than bi-allelic ones for population genetic studies simply because they offer greater potential for variation in observed number of alleles as well as allele frequency differences across populations. Since allele frequencies at a locus in a population are structurally constrained (they always add to one), a matrix treatment of allele frequency data at a multi-allelic locus requires deleting one allele from the analysis. Hence the resultant estimator may be construed as dependent on which allele is being eliminated in the process of estimation (BALAKRISHNAN 1973). Such situations have been faced by BALAKRISHNAN and SANGHVI (1968) and SMOUSE and SPIELMAN (1977) when they attempted to estimate genetic distances between populations by statistics parallel to Mahalanobis-$D^2$ (MAHALANOBIS 1936) for multivariate data. ROBERTS and HIORNS (1962) also suggested a method of estimating genetic admixture in a hybrid population using allele frequency data that requires elimination of one allele of a multiallelic locus. Recently, this issue has resurfaced in the least-square estimation of admixture components in a hybrid population (LONG 1991). Since these investigators generally presented their estimating equations in terms of "shortened" vectors of allele frequencies (by deleting one allele from each locus) and the variance-covariance matrix of such "shortened" vectors of sampled allele frequencies, in general it is not obvious whether or not the resultant estimators depend upon the allele that is eliminated from the analysis. As a result, such methods are criticized on the ground of the subjectivity involved in selecting the allele to be eliminated (BALAKRISHNAN 1973) although in some applications algebraic verifications are given to show that any allele can be dropped without affecting the estimate (LONG 1991). The purpose of this communication is to show that by exploiting a well-known property of the variance-covariance matrix of the cell frequencies of a multinomial distribution (KURCZYNSKI 1970) a simple translation of the matrix estimators can be obtained, which indicates that even though the formal representation requires deleting one allele, the computational equation truly needs the frequencies of all alleles. Therefore, such estimators are functions of the full array of allele frequencies.

This simple exercise has at least three implications. First, it shows that the resultant estimators can be computed by algebraic operations involving all allele frequencies (which consequently results in numerically more accurate estimates, because matrix inversions generally introduce round off errors, which can be substantial particularly when the array size is large). Second, analytical relationships between different estimators of genetic parameters (e.g., distance, fixation indices, or admixture components) can be studied with greater ease with such representations (see *e.g.*, CHAKRABORTY and RAO 1991). Finally, genetic polymorphisms detected by DNA markers such as the variable number of tandem repeat (VNTR) loci often involve allele numbers (per locus) exceeding several dozen, and treating them with matrix operations requires a large array size, and even with that numerical inaccuracies cannot be avoided. On the contrary, algebraic expressions such as the ones presented here should make the analysis of such allele frequency data easier and certainly numerically more accurate.

Although the technique suggested here has wider applications, I consider only two specific estimation problems.

**Genetic distance with multiple alleles:** Denoting $p_{ijk}$ as the frequency of the $k$th allele ($k = 1, 2, \ldots, s_j + 1$) of the $j$th locus ($j = 1, 2, \ldots, r$) in the $i$th population ($i = 1, 2$), estimated from $n_{ij}/2$ individuals sampled from the $i$th population for the $j$th locus, BALAKRISHNAN and SANGHVI (1968) suggested an estimator of the genetic distance between the two populations, given by

$$G_c^2 = \sum_{j=1}^{r} \mathbf{d}_j' \mathbf{C}_j^{-1} \mathbf{d}_j, \tag{1}$$

where $\mathbf{d}_j$ is a column vector of dimension $s_j$ (one less than the number of segregating alleles at the $j$th locus, $s_j + 1$) whose $k$th element is $d_{jk} = p_{1jk} - p_{2jk}$, and $\mathbf{C}_j$ is a square matrix of size $s_j \times s_j$ whose elements are

$$C_{jkl} = \begin{cases} p_{jk}(1 - p_{jk}), & \text{for } k = l, \\ -p_{jk}p_{jl}, & \text{for } k \neq l \end{cases} \tag{2}$$

for $k, l = 1, 2, \ldots, s_j$; in which $p_{jk}$ is the average of the $k$th allele frequency at the $j$th locus across populations; *i.e.*,

$$p_{jk} = \sum_i n_{ij}p_{ijk} / \sum_i n_{ij}. \tag{3}$$

Obviously, the quadratic form of equation (1) is the analog of Mahalanobis-$D^2$ (MAHALANOBIS 1936) since $\mathbf{C}_j$, given by (2), is the common dispersion matrix of the "shortened" vector of allele frequencies, estimated from the average allele frequencies across populations. Equation 1 may be written in the algebraic form

$$G_c^2 = \sum_{j=1}^{r} \sum_{k=1}^{s_j} \sum_{l=1}^{s_j} (p_{1jk} - p_{2jk}) C_j^{kl} (p_{1jl} - p_{2jl}), \quad (4)$$

where $C_j^{kl}$ is the $(k,l)$th element of the $\mathbf{C}_j^{-1}$ matrix.

In order to show that $G_c^2$ is dependent on all allele frequencies, KURCZYNSKI (1970) noted that the inverse of the matrix $\mathbf{C}_j$ (of Equation 2) has the form

$$C_j^{kl} = \begin{cases} p_{jk}^{-1} + p_{j,s_j+1}^{-1}, & \text{for } k = l, \\ p_{j,s_j+1}^{-1}, & \text{for } k \neq l, \end{cases} \quad (5)$$

for $k, l = 1, 2, \ldots, s_j$.

Inserting (5) in (4) and noting that

$$\sum_{k=1}^{s_j} (p_{1jk} - p_{2jk})^2 + \sum_{k \neq j=1}^{s_j} \sum (p_{1jk} - p_{2jk})(p_{1jl} - p_{2jl}) = \left[ \sum_{k=1}^{s_j} (p_{1jk} - p_{2jk}) \right]^2 = (p_{1j,s_j+1} - p_{2j,s_j+1})^2,$$

we obtain

$$G_c^2 = \sum_{j=1}^{r} \sum_{k=1}^{s_{j+1}} (p_{1jk} - p_{2jk})^2 / p_{jk}, \quad (6)$$

which depends on frequencies of every segregating allele, irrespective of which alleles are being dropped in the definition of the $\mathbf{d}_j$-vectors or $\mathbf{C}_j$ matrices. Equation 6 not only shows the involvement of all allele frequencies in the estimation, but also it is numerically simpler to compute than Equation 4. Note that the above proof also applies to SMOUSE and WILLIAM's (1982) measure of disease-gene association, where such equivalence is stated without a formal derivation. Furthermore, it demonstrates that BALAKRISHNAN and SANGHVI's (1968) measure is equivalent to the original estimator of SANGHVI (1953), except a multiplication factor. In addition, the above derivation shows that the alternative two estimators ($G_c^2$ and $G_s^2$) proposed by BALAKRISHNAN and SANGHVI (1968) are mathematically identical.

Another advantage of the representation of Equation 6 is that it clearly shows how SANGHVI's estimator of genetic distance is related to others. For example, considering the allele frequencies at a single locus (say, the $j$th locus), BHATTACHARYYA (1946) defined a distance statistic, $\theta^2$, between populations, which satisfies the relationship

$$\text{Cos } \theta = \sum_{k=1}^{s_{j+1}} \{p_{1jk} p_{2jk}\}^{1/2}, \quad (7)$$

which can be written as

$$\begin{aligned} \text{Cos } \theta &= \frac{1}{2} \cdot \sum_{k=1}^{s_{j+1}} [(p_{1jk} + p_{2jk})^2 - (p_{1jk} - p_{2jk})^2]^{1/2} \\ &= \frac{1}{2} \cdot \sum_{k=1}^{s_{j+1}} (p_{1jk} + p_{2jk}) \left[ 1 - \frac{(p_{1jk} - p_{2jk})^2}{(p_{1jk} + p_{2jk})^2} \right]^{1/2} \quad (8) \\ &= 1 - \frac{1}{4} \cdot \sum_{k=1}^{s_{j+1}} \frac{(p_{1jk} - p_{2jk})^2}{p_{1jk} + p_{2jk}}. \end{aligned}$$

However, since $\text{Cos } \theta \approx 1 - \theta^2/2$, for small $\theta$, Equation 8 approximates to

$$\theta^2 \approx \frac{1}{2} \cdot \sum_{k=1}^{s_{j+1}} (p_{1jk} - p_{2jk})^2 / (p_{1jk} + p_{2jk}), \quad (9)$$

showing that for genetically close populations (*i.e.*, for small $\theta$), SANGHVI's (1953) and BHATTACHARYYA's (1946) distance estimators are equivalent, barring a multiplication factor. Equivalence of Equations 9 and 6 with 4 further shows that they are analogs of Mahalanobis-$D^2$ for categorial data. Several other such equivalence relationships between various distance functions are discussed in CHAKRABORTY and RAO (1991) who utilize representations such as Equation 6.

The same logic provides a formal proof of the assertion that in the absence of disequilibria (WEIR 1979), SMOUSE and SPIELMAN's (1977) multivariate distance function based on multiple-allele genotype score vectors reduces to the form of Equation 6.

*Weighted least square estimate of admixture proportions*: For a dihybrid population whose gene pool consists of a fraction $M$ of genes from a parental population 1 and a fraction $(1 - M)$ from parental population 2, LONG (1991) recently suggested a weighted least square estimator of $M$, which in matrix notation takes the form

$$m_j = (\mathbf{x}_j' \mathbf{V}_j^{-1} \mathbf{x}_j)^{-1} \mathbf{x}_j' \mathbf{V}_j^{-1} \mathbf{y}_j, \quad (10)$$

where $\mathbf{x}_j$ and $\mathbf{y}_j$ are column vectors of dimension $s_j$ (one less than the number of segregating alleles, $s_{j+1}$, at the $j$th locus), with their $k$th elements defined by $x_{jk} = p_{1jk} - p_{2jk}$ and $y_{jk} = p_{hjk} - p_{2jk}$, for $k = 1, 2, \ldots, s_j$, and $\mathbf{V}_j$ is a $s_j \times s_j$ matrix with elements

$$V_{jkl} = \begin{cases} E(p_{hjk}) \cdot [1 - E(p_{hjk})], & \text{for } k = l \\ -E(p_{hjk}) \cdot E(p_{hjl}), & \text{for } k \neq l \end{cases} \quad (11)$$

in which $p_{ijk}$ is the frequency of the $k$th allele ($k = 1, 2, \ldots, s_{j+1}$) at the $j$th locus in the $i$th population ($i =$ 1 or 2 for the parental populations) and $E(p_{hjk})$ is the expected allele frequency in the admixed population under the admixture model.

The estimator $m_j$ (Equation 10) is based on the "shortened" vectors of allele frequency differences (dropping the $(s_{j+1})$th allele). However, noting that the elements of the $\mathbf{V}_j^{-1}$ matrix are given by

$$V_j^{kl} = \begin{cases} 1/E(p_{hjk}) + 1/E(p_{hj,s_j+1}), & \text{for } k = l, \\ 1/E(p_{hj,s_j+1}), & \text{for } k \neq l, \end{cases} \quad (12)$$

for $k, l = 1, 2, \ldots, s_j$, LONG (1991) verified that the estimator $m_j$ of Equation 10 is invariant of the allele dropped from the analysis. To show explicitly that Equation 10 does not depict that it depends on all allele frequencies, first note that

$$\mathbf{x}_j'\mathbf{V}_j^{-1}\mathbf{x}_j = \sum_{k=1}^{s_j}\sum_{l=1}^{s_j}(p_{1jk} - p_{2jk})V_j^{kl}(p_{1jk} - p_{2jk}), \quad (13a)$$

and

$$\mathbf{x}_j'\mathbf{V}_j^{-1}\mathbf{y}_j = \sum_{k=1}^{s_j}\sum_{l=1}^{s_j}(p_{1jk} - p_{2jk})V_j^{kl}(p_{hjk} - p_{2jk}). \quad (13b)$$

Invoking (12) in (13a) and (13b), and noting that

$$p_{1j,s_j+1} - p_{2j,s_j+1} = -\sum_{i=1}^{s_j}(p_{ijk} - p_{2jk}), \quad (14a)$$

and

$$p_{hj,s_j+1} - p_{2j,s_j+1} = -\sum_{i=1}^{s_j}(p_{hjk} - p_{2jk}), \quad (14b)$$

we can rewrite (13a) and (13b) as

$$\mathbf{x}_j'\mathbf{V}_j^{-1}\mathbf{x}_j = \sum_{k=1}^{s_j+1}(p_{1jk} - p_{2jk})^2/E(p_{hjk}), \quad (15a)$$

and

$$\mathbf{x}_j'\mathbf{V}_j^{-1}\mathbf{Y}_j = \sum_{k=1}^{s_j+1}(p_{1jk} - p_{2jk})(p_{hjk} - p_{2jk})/E(p_{hjk}), \quad (15b)$$

so that Equation 10 becomes

$$m_j = \frac{\sum_{k=1}^{s_j+1}(p_{1jk} - p_{2jk})(p_{hjk} - p_{2jk})/E(p_{hjk})}{\sum_{k=1}^{s_j+1}(p_{1jk} - p_{2jk})^2/E(p_{hjk})}, \quad (16)$$

which is an equation of scalers. Expressed in this fashion, $m_j$ involves each of the $s_{j+j}$ segregating allele frequencies of both parental populations and the admixed one.

This representation (Equation 16) of the weighted least squares (WLS) estimator of LONG (1991) further shows that $m_j$ (the WLS estimator) is identical to the classical BERNSTEIN (1931) estimator of admixture proportion for a bi-allelic locus, noted in LONG and SMOUSE (1983). With the notation $p_{ij}$ and $q_{ij} = (1 - p_{ij})$ of the two allele frequencies at a locus the numerator and denominator of Equation 16 become

$$\frac{(p_{1j} - p_{2j})(p_{hj} - p_{2j})}{E(p_{hj})} + \frac{(q_{1j} - q_{2j})(q_{hj} - q_{2j})}{E(q_{hj})} = \frac{(p_{1j} - p_{2j})(p_{hj} - p_{2j})}{E(p_{hj})\cdot E(q_{hj})},$$

and

$$\frac{(p_{1j} - p_{2j})^2}{E(p_{hj})} + \frac{(q_{1j} - q_{2j})^2}{E(q_{hj})} = \frac{(p_{1j} - p_{2j})^2}{E(p_{hj})\cdot E(q_{hj})},$$

so that the cancellation of their common denominators results in

$$m_j = (p_{hj} - p_{2j})/(p_{1j} - p_{2j}) = (q_{hj} - q_{2j})/(q_{1j} - q_{2j}),$$

establishing the identity of the WLS and Bernstein's estimators for bi-allelic loci.

The combined estimator for allele frequency data on $r$ loci, based on LONG's (1991) method, becomes

$$m = \frac{\sum_{j=1}^{r}\sum_{k=1}^{s_j+1}(p_{1jk} - p_{2jk})(p_{hjk} - p_{2jk})/E(p_{hjk})}{\sum_{j=1}^{r}\sum_{k=1}^{s_j+1}(p_{1jk} - p_{2jk})^2/E(p_{hjk})} \quad (17)$$

which avoids matrix manipulations of even bigger dimension.

The sampling error of $m$ also has a corresponding scalar form. In LONG's notation, the sampling variance is

$$V(m) = \text{MSE}\cdot(\mathbf{x}'\mathbf{V}^{-1}\mathbf{x})^{-1}, \quad (18)$$

where the mean square error (MSE) of the admixture model is

$$\text{MSE} = (\mathbf{y} - m\mathbf{x})'\mathbf{V}^{-1}(\mathbf{y} - m\mathbf{x})/\sum_{j=1}^{r} s_j. \quad (19)$$

Invoking (12) in these quadratic forms, and using the identities (14a) and (14b), we can rewrite (19) as

$$\text{MSE} = \frac{\sum_{j=1}^{r}\sum_{k=1}^{s_{j+1}}[(p_{hjk} - p_{2jk}) - m(p_{1jk} - p_{2jk})]^2/E(p_{hjk})}{\sum_{j=1}^{r} s_j} \quad (20)$$

which yields the sampling variance of $m$,

$$V(m) = \frac{\sum_{j=1}^{r}\sum_{k=1}^{s_j+1}[(p_{hjk} - p_{2jk}) - m(p_{1jk} - p_{2jk})]^2/E(p_{hjk})}{\left[\sum_{j=1}^{r} s_j\right]\cdot\left[\sum_{j=1}^{r}\sum_{k=1}^{s_j+1}(p_{1jk} - p_{2jk})^2/E(p_{hjk})\right]}. \quad (21)$$

The variance of the admixture estimate based on the $j$th locus data is

$$V(m_j) = \text{MSE}\cdot\left[\sum_{k=1}^{s_j+1}(\mathrm{p}_{1jk} - p_{2jk})^2/E(p_{hjk})\right]^{-1}, \quad (22)$$

in which the expression (20) is used for evaluating the MSE.

In addition to the demonstration that Equations 16 and 22, or 17 and 21 involve all allele frequencies

from each population, their computational simplicity remain unaltered even if the sample sizes for different loci are different. Since the $V$ matrix refers to the expected allele frequencies in the admixed population, all terms of the summation over $j$ will have to be weighted by $n_{hj}$, the number of genes sampled for the $j$th locus from the admixed population. For example, the combined estimator becomes

$$m = \frac{\sum_{j=1}^{r} n_{hj} \sum_{k=1}^{s_j+1} (p_{1jk} - p_{2jk})(p_{hjk} - p_{2jk})/E(p_{hjk})}{\sum_{j=1}^{r} n_{hj} \sum_{k=1}^{s_j+1} (p_{1jk} - p_{2jk})^2/E(p_{hjk})}. \quad (23)$$

The corresponding changes in its sampling variance are also similar.

Other population genetic applications of algebraic representations of quadratic forms involving inverses of multinomial variance-covariance matrices include the estimation of Wright's fixation indices in the context of analysis of population structure. Using approaches similar to the above, LONG's (1986) multiallelic generalizations of COCKERHAM's (1969, 1973) variance-covariance estimators of the fixation indices can also be reduced to algebraic forms, which indicate their relationship with some existing estimators suggested earlier (see *e.g.*, LI and HORVITZ 1953; CURIE-COHEN 1982; ROBERTSON and HILL 1984; WEIR and COCKERHAM 1984).

To close this commentary, I must mention that the algebraic reductions of the matrix estimators such as the ones mentioned above are not meant to denigrate the utility of matrix notations in population genetics. Matrix representations of functions of allele frequencies at multiallelic loci have their importance and place that cannot be denied. They serve the purpose of establishing the basis of the method of estimation that is not always obvious in the closed form algebraic expression. In some instances matrix estimators are unavoidable. For example, the estimator of admixture contributions from multiple (more than two) ancestral populations is straightforward in matrix notation (ELSTON 1971; CHAKRABORTY 1986) and the incorporation of all orders of disequilibria (WEIR 1979) in estimating parameters of population structure and genetic distance analyses requires matrix notations, although nearly equivalent algebraic forms are also available (see *e.g.*, WEIR and COCKERHAM 1984). Nevertheless, the primary intent of this note has been to demonstrate that the principle that these are independent of which allele is dropped from the analysis.

This work was supported by U.S. Public Health Service research grants GM 41399 from the National Institutes of Health and 90-IJ-CX-0038 from the National Institute of Justice. I thank P. E. SMOUSE, B. S. WEIR and an anonymous reviewer for their comments and suggestions on the work.

RANAJIT CHAKRABORTY
Center for Demographic and Population Genetics
University of Texas Graduate School of Biomedical Sciences
P. O. Box 20334
Houston, Texas 77225

## LITERATURE CITED

BALAKRISHNAN, V., 1973 Use of distance in hybrid analysis, pp. 268–273 in *Genetic Structure of Populations*, edited by N. E. MORTON, University of Hawaii Press, Honolulu.

BALAKRISHNAN, V., and L. D. SANGHVI, 1968 Distances between populations on the basis of attribute data. Biometrics, **24:** 859–865.

BERNSTEIN, F., 1931 Die geographische Verteilung der Blutgruppen und ihre anthropologische Bedeutung. Comitato Italiano per lo Studio dei Problemi della Popülazione. Instituto Poligraphico dello Stato, Rome.

BHATTACHARYYA, A., 1946 On a measure of divergence between two multinomial populations. Sankhya **7:** 401–406.

CHAKRABORTY, R., 1986 Gene admixture in human populations: models and predictions. Yearb. Phys. Anthropol. **29:** 1–43.

CHAKRABORTY, R., and C. R. RAO, 1991 Measurement of genetic variation for evolutionary studies, in *Handbook of Statistics*, Vol. 8, edited by C. R. RAO and R. CHAKRABORTY. Elsevier, Amsterdam (in press).

COCKERHAM, C. C., 1969 Variance of gene frequencies. Evolution **23:** 72–84.

COCKERHAM, C. C., 1973 Analyses of gene frequencies. Genetics **74:** 679–700.

CURIE-COHEN, M., 1982 Estimates of inbreeding in a natural population: a comparison of sampling properties. Genetics **100:** 339–358.

ELSTON, R. C., 1971 The estimation of admixture in racial hybrids. Ann. Hum. Genet. **35:** 9–17.

KURCZYNSKI, T. W., 1970 Generalized distance and discrete variables. Biometrics **26:** 525–534.

LI, C. C., and D. G. HORVITZ, 1953 Some methods of estimating the inbreeding coefficient. Am. J. Hum. Genet. **5:** 107–117.

LONG, J. C., 1986 The allelic correlation structure of Gainj- and Kalam-speaking people. I. the estimation and interpretation of Wright's *F*-statistics. Genetics **112:** 629–647.

LONG, J. C., 1991 The genetic structure of admixed populations. Genetics **127:** 417–428.

LONG, J. C., and P. E. SMOUSE, 1983 Intertribal geneflow between the Ye'cuana and Yanamamö: genetic analysis of an admixed village. Am. J. Phys. Anthropol. **61:** 411–422.

MAHALANOBIS, P. C., 1936 On the generalized distance in statistics. Proc. Natl. Inst. Sci. India **12:** 49–55.

ROBERTS, D. F., and R. W. HIORNS, 1962 The dynamics of racial admixture. Am. J. Hum. Genet. **14:** 261–277.

ROBERTSON, A., and W. G. HILL, 1984 Deviations from Hardy-Weinberg proportions: sampling variances and use in estimation of inbreeding coefficients. Genetics **107:** 703–718,

SANGHVI, L. D., 1953 Comparison of genetical and morphological methods for a study of biological differences. Am. J. Phys. Anthropol. **11:** 385–404.

SMOUSE, P. E., and R. S. SPIELMAN, 1977 How allocation of individuals depends on genetic differences among populations. Excerpta Med. Congr. Ser. No. **411:** 255–260.

SMOUSE, P. E., and R. C. WILLIAMS, 1982 Multivariate analysis of HLA-disease associations. Biometrics **38:** 757–768.

WEIR, B. S., 1979 Inferences about linkage disequilibrium. Biometrics **35:** 235–254.

WEIR, B. S., and C. C. COCKERHAM, 1984 Estimating F-statistics for the analysis of population structure. Evolution **38:** 1358–1370.