Evolution, 52(4), 1998, pp. 950–956

ESTIMATING HIERARCHICAL *F*-STATISTICS

RONG-CAI YANG
Department of Renewable Resources, University of Alberta, Edmonton, Alberta T6G 2H1, Canada
E-mail: rcyang@rr.ualberta.ca

Abstract.—This paper presents an analysis of variance (ANOVA) approach by which estimation of *F*-statistics can be made from data with an arbitrary *s*-level hierarchical population structure. Assuming a complete random-effect model, a general ANOVA procedure is developed to estimate *F*-statistics as ratios of different variance components for all levels of population subdivision in the hierarchy. A generalized relationship among *F*-statistics is also derived to extend the well-known relationship originally found by Sewall Wright. Although not entirely free from the bias particular to small number of subdivisions at each hierarchy and extreme gene frequencies, the ANOVA estimators of *F*-statistics consider sampling effects at each level of hierarchy, thus removing the bias incurred in the other estimators that are commonly based on direct substitution of unknown gene frequencies by their sample estimates. Therefore, the ANOVA estimation procedure presented here may become increasingly useful in analyzing complex population structure because of increasing use of the estimated hierarchical *F*-statistics to infer genetic and demographic structures of natural populations within and among species.

Key words.—Analysis of variance, estimation, *F*-statistics, genic correlations, hierarchical population structure.

Received October 15, 1997. Accepted March 24, 1998.

Wright's (1951, 1965) *F*-statistics have proven to be an extremely useful tool in elucidating the pattern and extent of genetic variation residing within and among natural populations of animal and plant species. For a total population that is subdivided into many subpopulations, Wright (1951, 1965) defined three *F*-statistics (correlations between uniting gametes), to relate the departure from panmixia in the total population (F_{IT}) to the genetic divergence among subdivisions (F_{ST}) and to averaged departures from panmixia within subdivisions (F_{IS}),

$$(1 - F_{IT}) = (1 - F_{IS})(1 - F_{ST}) \qquad (1)$$

In the absence of disturbing forces such as selection, these correlations can be interpreted in terms of hierarchical genetic sampling from a single ancestral population. An important property of population subdivision (first noted by Wahlund 1928) is that even if each subpopulation is under random mating ($F_{IS} = 0$), the genotypic frequencies in the total population may deviate from Hardy-Weinberg proportions ($F_{IT} = F_{ST} \neq 0$) because of variation in gene frequencies among subpopulations. Extensions and refinements of Wright's original definitions have since been made to account for further population subdivision and multiple alleles at a locus (e.g., Nei 1977; Wright 1978; Chakraborty 1980). However, when attempting to estimate *F*-statistics, these studies have rarely made a distinction between parameters and statistics. The commonly used procedure is the direct substitution of unknown parametric values (e.g., population gene frequencies) by their respective sample values to obtain estimates of *F*-statistics. Thus, the resulting estimates may be biased, particularly when the numbers of individuals and subdivisions sampled are small.

Cockerham (1969, 1973) and Weir (1996) have developed the analysis of variance (ANOVA) approach to define and estimate the *F*-statistics. As noted by Cockerham and Weir (1987, 1993), this alternative formulation of *F*-statistics has a number of desirable properties in estimating population structure. For example, in estimating gene flow from an *n*-island model, the ANOVA estimators of *F*-statistics do not depend on the unknown quantity *n*, whereas the estimators by Nei (1973) and Crow and Aoki (1984) do.

In the ANOVA approach, the *F*-statistics are various intraclass correlations that are defined as ratios of variance components (Cockerham 1969, 1973). The ANOVA estimators of *F*-statistics are simply those of intraclass correlations (Weir and Cockerham 1984; Weir 1996). By relating *F*-statistics to variance components from an appropriate nested ANOVA, it is straightforward to add additional levels,of population subdivisions in the sample. Weir (1996) presented the detailed estimation procedures for three- and four-level hierarchies. However, no general estimation procedure has been given for data with an arbitrary *s*-level hierarchy. In the present study, a general hierarchical genetic structure will be presented to provide an appropriate basis for a subsequent description of a general procedure of estimating *F*-statistics in *s*-level hierarchy. The generality emphasizes the nature of the parameters being estimated and facilitates attempts to implement the estimation procedure into a computer program.

HIERARCHICAL GENIC STRUCTURE

When individual genes are identifiable, the basic unit for population genetic analysis is the gene itself. Thus, genic structure in an arbitrary *s*-level hierarchy of population subdivisions is genes within individuals (diploids), individuals within the ($s - 1$)th level of subpopulations (or in short, subpopulations^{s-1}), subpopulations^{s-1} within subpopulations^{s-2}, . . . , subpopulations within populations. A pair of genes can fall into any of the following categories: genes in the same individuals, genes in different individuals but in the same subpopulations^{s-1}, genes in different subpopulations^{s-1} but in the same subpopulations^{s-2}, . . . , genes in different subpopulations but in the same populations, and genes in different independent populations. This genic structure is simply an extension of the well-known development of Cockerham (1969, 1973) and Weir (1996, ch. 5).

For a pair of alleles (A and $\bar{A}$) at a locus with global population frequencies p and $(1 - p)$, an indicator variable X is defined as $X = 1$ if the gene is A and $X = 0$ if the gene is

$\bar{A}$. The mean and variance of X are p and $\sigma^2 = p(1 - p)$, respectively. The total variance σ^2 can be partitioned into components of variation due to differences between genes at different levels of the hierarchy, that is, genes in the same individuals ($\sigma^2_{s+1(s)}$), genes in different individuals but in the same subpopulations^{s-1} ($\sigma^2_{s(s-1)}$), genes in different subpopulations^{s-1} but in the same subpopulations^{s-2} ($\sigma_{s-1(s-2)}$), genes in different subpopulations but in the same populations ($\sigma^2_{2(1)}$), and genes in different independent populations ($\sigma^2_{1(0)}$). These variance components sum to the total variance σ^2,

$$\sigma^2 = p(1 - p) = \sum_{j=1}^{s+1} \sigma^2_{j(j-1)} \tag{2}$$

and are related to Wright's genic correlations (F-statistics) as follows,

$$\sigma^2_{j(j-1)} = (\theta_j - \theta_{j-1})\sigma^2, \tag{3}$$

where $j = 1, 2 \ldots s + 1$ and θ_j is the correlation of two genes in the same subpopulation at the jth level in the hierarchy relative to the zero correlation of unrelated genes in the noninbred and nonsubdivided founder population, as clearly noted by Cockerham (1969, 1973). In particular, $\theta_{s+1} = 1$ is the correlation of genes with themselves and $\theta_0 = 0$ is the correlation of genes in different independent populations. In Cockerham (1969, 1973) or Weir (1996), $\theta_s = F$ is the correlation between genes within the same individual. It is easily seen that the genetic correlations (θ_j) are simply different variance ratios, all being relative to the total variance,

$$\theta_j = \sum_{i=1}^{j} \sigma^2_{i(i-1)}/\sigma^2 \tag{4}$$

for $j = 1, 2 \ldots s + 1$. These correlations can be used to determine other, complementary correlations (f_j) with varying references,

$$f_j = \frac{\sigma^2_{j(j-1)}}{\sum_{i=j}^{s+1} \sigma^2_{i(i-1)}} = \frac{\theta_j - \theta_{j-1}}{1 - \theta_{j-1}}, \tag{5}$$

where $j = 1, 2, \ldots, s + 1$ and two trivial values are $f_1 = \theta_1$ and $f_{s+1} = \theta_{s+1} = 1$. Rearranging equation (5), we have,

$$\begin{aligned}(1 - \theta_j) &= (1 - f_j)(1 - \theta_{j-1})\\ &= \prod_{i=1}^{j} (1 - f_i),\end{aligned} \tag{6}$$

a generalization of well-known expressions by Wright (1978, p. 86–89).

A few special cases should be noted. Equation (1) can be obtained by settings $s = 2$, that is, a two-level hierarchy, alleles within diploid individuals and individuals within populations. In this case, equation (6) reduces to: $(1 - \theta_2) = (1 - f_1)(1 - \theta_1)$. It is easy to identify Wright's (1951) notation, $F_{IT} = \theta_2$, $F_{IS} = f_1$ and $F_{ST} = \theta_1$ and Weir and Cockerham's (1984) notation $F = \theta_2, f = f_1$, and $\theta = \theta_1$. With monoecious populations and individuals resulting from random union of gametes, there is no need to distinguish between the correlations of genes within and among individuals, that is, $\theta_s = \theta_{s-1}$. Likewise, the distinction between θ_s and θ_{s-1} is also dropped for haploid populations. Thus, essentially the same population structure is revealed from the analysis of random mating and haploid populations.

Estimation

Data Structure

Suppose that a total of n_0 diploid individuals are sampled from a species or any taxon with an arbitrary s-level hierarchical population structure, that is, independent populations, subpopulations within populations, subpopulations^2 within subpopulations . . . individuals within subpopulations^{s-1}. Each diploid individual has two alleles at a locus. The linear model of the genic indicator variable X as defined earlier for this hierarchical population structure can be written as

$$\begin{aligned}X_{i_1i_2i_3\ldots i_{s-1}i_si_{s+1}} &= X_{123\ldots s-1,s,s+1}\\ &= p_0 + a_{1(0)} + b_{2(1)} + c_{3(2)} + \ldots + u_{s-1(s-2)}\\ &\quad + v_{s(s-1)} + w_{s+1(s)},\end{aligned} \tag{7}$$

where p_0 is the frequency of gene A in the total population; $a_{1(0)}$ is the effect of i_1th population ($i_1 = 1, 2 \ldots I_1$); $b_{2(1)}$ is the effect of i_2th subpopulation within i_1th population ($i_2 = 1, 2 \ldots I_2$); $c_{3(2)}$ is the effect of i_3th subpopulation^2 within i_2th subpopulation ($i_3 = 1, 2 \ldots I_3$); $u_{s-1(s-2)}$ is the effect of i_{s-1}th subpopulation^{s-2} within i_{s-2}th subpopulation^{s-3} ($i_{s-1} = 1, 2 \ldots I_{s-1}$); $v_{s(s-1)}$ is the effect of i_sth individual within i_{s-1}th subpopulation^{s-2} ($i_s = 1, 2 \ldots n_{s-1}$); and $w_{s+1(s)}$ is the effect of i_{s+1}th allele within i_sth individual ($i_{s-1} = 1, 2$). All effects except for p_0 are assumed to be *random* and uncorrelated, with variances being $\sigma^2_{1(0)}, \sigma^2_{2(1)}, \sigma^2_{3(2)} \ldots \sigma^2_{s-1(s-2)}$, $\sigma^2_{s(s-1)}$, and $\sigma^2_{s-1(s)}$, respectively.

To clarify the intricate data structure described above, it is necessary to compact counts of individuals and alleles within individuals at different level hierarchies in abbreviated form. Thus, n_j is used to represent $n_{i_1i_2\ldots i_j}$, the number of individuals in the jth subpopulation^{s-2} down the hierarchy. For example, let $n_{s-2} = n_{i_1i_2\ldots i_{s-2}}$ be the number of diploid individuals in the $(s - 2)$th subpopulation. Then total number of alleles in the i_{s-2}th subpopulation^{s-3} is

$$2n_{i_1i_2\ldots i_{s-2}} = 2 \sum_{i_{s-1}=1}^{I_{s-1}} n_{i_1i_2\ldots i_{s-2}i_{s-1}} \tag{8}$$

and its abbreviated form is written as

$$2n_{s-2} = 2 \sum_{i_{s-1}=1}^{I_{s-1}} n_{s-1}. \tag{9}$$

In general, the compact sums of alleles at different levels of hierarchy may be expressed as

$$\begin{aligned}2n_0 &= 2 \sum_{i_1=1}^{I_1} n_1 = 2 \sum_{i_1=1}^{I_1} \sum_{i_2=1}^{I_2} n_2 = \ldots\\ &= 2 \sum_{i_1=1}^{I_1} \sum_{i_2=1}^{I_2} \ldots \sum_{i_{s-1}=1}^{I_{s-1}} n_{s-1},\end{aligned} \tag{10}$$

where n_0 is the total number of individuals in the sample.

Sums of the numbers of subdivisions at different levels of

TABLE 1. Analysis of variance foı an arbitrary *s*-level hierarchical population structure.

Source of variation	df	Mean squares
Among populations	d_1	$m_{1(0)} = \frac{1}{d_1}\left[\sum \frac{X_1^2}{2n_1} - \frac{X_0^2}{2n_0}\right] = \frac{2}{d_1}\sum n_1(\hat{p}_1 - \hat{p}_0)^2$
Among subpopulations within populations	d_2	$m_{2(1)} = \frac{1}{d_2}\left[\sum \frac{X_2^2}{2n_2} - \sum \frac{X_1^2}{2n_1}\right] = \frac{2}{d_2}\sum n_2(\hat{p}_2 - \hat{p}_1)^2$
Among subpopulations^2 within subpopulations	d_3	$m_{3(2)} = \frac{1}{d_3}\left[\sum \frac{X_3^2}{2n_3} - \sum \frac{X_2^2}{2n_2}\right] = \frac{2}{d_3}\sum n_3(\hat{p}_3 - \hat{p}_2)^2$
⋮	⋮	⋮
Among subpopulations^{s-2} within subpopulations^{s-3}	d_{s-1}	$m_{s-1(s-2)} = \frac{1}{d_{s-1}}\left[\sum \frac{X_{s-1}^2}{2n_{s-1}} - \sum \frac{X_{s-2}^2}{2n_{s-2}}\right] = \frac{2}{d_{s-1}}\sum n_{s-1}(\hat{p}_{s-1} - \hat{p}_{s-2})^2$
Among individuals within subpopulations^{s-2}	d_s	$m_{s(s-1)} = \frac{1}{d_s}\left[\sum \frac{X_s^2}{2} - \sum \frac{X_{s-1}^2}{2n_{s-1}}\right] = \frac{2}{d_s}\sum n_{s-1}\left[\hat{p}_{s-1} + \hat{P}_{s-1} - 2\hat{p}_{s-1}^2\right]$
Among alleles within individuals	d_{s+1}	$m_{s+1(s)} = \frac{1}{d_{s+1}}\left[\sum X_{s+1}^2 - \sum \frac{X_s^2}{2}\right] = \frac{1}{2d_{s+1}}\sum n_{s-1}(\hat{p}_{s-1} - \hat{P}_{s-1})$

Note: X_i, n_i, and $\hat{p}_i$ are, respectively, the total of the indicator variable, the number of individuals, and the estimated gene frequency of a subdivision at the *i*th level of hierarchy. $\hat{P}_{s-1}$ is the estimated genotypic frequency of subdivision at the $(s - 1)$th level of hierarchy.

the hierarchy are also needed to derive degrees of freedom. Let r_i be the number of subdivisions in level *i* of the hierarchy. Then, recognizing that $I_j = I_{i_1 i_2 \ldots i_{j-1}}$, we have

$$\begin{aligned} r_1 &= I_1 \\ r_2 &= \sum_{i_1=1}^{I_1} I_2 \\ r_3 &= \sum_{i_1=1}^{I_1}\sum_{i_2=1}^{I_2} I_3 \\ &\vdots\ \ \vdots\ \vdots \\ r_{s-1} &= \sum_{i_1=1}^{I_1}\sum_{i_2=1}^{I_2}\cdots\sum_{i_{s-2}=1}^{I_{s-2}} I_{s-1} \\ r_s &= \sum_{i_1=1}^{I_1}\sum_{i_2=1}^{I_2}\cdots\sum_{i_{s-1}=1}^{I_{s-1}} n_{s-1} = n_0 \\ r_{s+1} &= 2n_0 \end{aligned} \tag{11}$$

and degrees of freedom for the *i*th source of variation is given by

$$d_i = r_i - r_{i-1},$$

with $r_0 = 1$. The total degrees of freedom is $\Sigma_{i=1}^{s+1} d_i = 2n_0 - 1$.

The same abbreviation is used to compact the sums of the indicator variable *X*. Let $X_{s+1} = X_{123\ldots s-1,s,s+1}$ be the observed values of individual alleles (i.e., zeros or ones). Further, let X_s, $X_{s-1} \ldots X_1$, and X_0 be the totals of the i_s stage, the i_{s-1} stage . . . the i_1 stage, and the grand total, respectively. Then the following relationship holds:

$$\begin{aligned} X_0 &= \sum_{i_1=1}^{I_1} X_1 = \sum_{i_1=1}^{I_1}\sum_{i_2=1}^{I_2} X_2 = \ldots \\ &= \sum_{i_1=1}^{I_1}\sum_{i_2=1}^{I_2}\cdots\sum_{i_s=1}^{n_{s-1}}\sum_{i_{s+1}=1}^{2} X_{s+1}. \end{aligned} \tag{12}$$

Clearly, X_s, $X_{s-1} \ldots X_1$, and X_0 are counts of *A* genes in the *s*th individual, in the subpopulation^{s-1} . . . in the population, and in the total population, respectively. Thus, the frequency of gene *A* at the *i*th level of hierarchy is estimated by

$$\hat{p}_i = X_i/2n_i$$

for $i = 0, 1, 2 \ldots s - 1$. Counts of genotype *AA* may be obtained by defining a new indicator variable *G* as the product of the two indicator variables representing the two alleles within the same individuals, i.e.,

$$G_{123\ldots s-1,s} = (X_{123\ldots s-1,s,1})(X_{123\ldots s-1,s,2}).$$

Thus, counts of genotype *AA* at different levels of hierarchy (G_s, $G_{s-1} \ldots G_1$, and G_0) are related to each other by

$$G_0 = \sum_{i_1=1}^{I_1} G_1 = \sum_{i_1=1}^{I_1}\sum_{i_2=1}^{I_2} G_2 = \ldots = \sum_{i_1=1}^{I_1}\sum_{i_2=1}^{I_2}\cdots\sum_{i_s=1}^{n_{s-1}} G_s \tag{13}$$

and the frequency of genotype *AA* at the *i*th level of hierarchy is estimated by

$$\hat{P}_i = G_i/n_i$$

for $i = 0, 1, 2 \ldots s - 1$. Using these notations, the ANOVA table including source of variation, degrees of freedom, sums of squares, and mean squares is readily constructed for the *s*-level hierarchical population structure (Table 1).

ANOVA Estimators

The ANOVA method of estimating variance component is to equate mean squares from ANOVA to the expected values. Let $\mathbf{m}$ be the vector of the mean squares, $\mathbf{m} = [m_{1(0)}\ m_{2(1)}\ m_{3(2)} \cdots m_{s-1(s-2)}\ m_{s(s-1)}\ m_{s+1(s)}]'$ as given in Table 1, and $\boldsymbol{\sigma}^2$ be the vector of variance components, $\boldsymbol{\sigma}^2 = [\sigma^2_{1(0)}, \sigma^2_{2(1)}, \sigma^2_{3(2)} \cdots \sigma^2_{s-1(s-2)}, \sigma^2_{s(s-1)}, \sigma^2_{s+1(s)}]'$. The expected mean squares are $E(\mathbf{m}) = \mathbf{K}\boldsymbol{\sigma}^2$, where

$$\mathbf{K} = \begin{bmatrix} k_{1,1} & k_{1,2} & k_{1,3} & \cdots & k_{1,s-1} & k_{1,s} & k_{1,s+1} \\ 0 & k_{2,2} & k_{2,3} & \cdots & k_{2,s-1} & k_{2,s} & k_{2,s+1} \\ 0 & 0 & k_{3,3} & \cdots & k_{3,s-1} & k_{3,s} & k_{3,s+1} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & k_{s-1,s-1} & k_{s-1,s} & k_{s-1,s+1} \\ 0 & 0 & 0 & \cdots & 0 & k_{s,s} & k_{s,s+1} \\ 0 & 0 & 0 & \cdots & 0 & 0 & k_{s+1,s+1} \end{bmatrix}. \tag{14a}$$

It is easy to see that $k_{i,s+1} = 1$ for $i = 1, 2, \ldots, s+1$ and $k_{i,s} = 2$ for $i = 1, 2, \ldots, s$. For $1 \le j \le s-1$, we have

$$k_{i,j} = \frac{2}{d_i}\left[\sum \frac{n_j^2}{n_i} - \sum \frac{n_j^2}{n_{i-1}}\right] \qquad (j \ge i). \tag{14b}$$

An alogorithm to calculate $k_{i,j}$ is given in the Appendix. Thus, the ANOVA estimator of $\boldsymbol{\sigma}^2$ is $\hat{\boldsymbol{\sigma}}^2$, obtained from $\mathbf{m} = \mathbf{K}\hat{\boldsymbol{\sigma}}^2$ as

$$\hat{\boldsymbol{\sigma}}^2 = \mathbf{K}^{-1}\mathbf{m}, \tag{15}$$

provided that $\mathbf{K}$ is nonsingular. Clearly, the estimators of variance components in equation (15) are unbiased because $E(\boldsymbol{\sigma}^2) = \mathbf{K}^{-1}E(\mathbf{m}) = \mathbf{K}^{-1}\mathbf{K}\boldsymbol{\sigma}^2 = \boldsymbol{\sigma}^2$. Because $\mathbf{K}$ is an upper triangle matrix, its inverse $\mathbf{K}^{-1} = \mathbf{L} = \{l_{i,j}\}$ is also an upper triangle matrix with

$$l_{i,j} = \begin{cases} 1/k_{i,i}, & j = i \\ -\sum\limits_{t=i+1}^{j} k_{i,t}l_{t,j}/k_{i,i}, & j > i \\ 0, & j < i. \end{cases} \tag{16}$$

Thus, variance components ($\sigma^2_{i(i-1)}$) are estimated by

$$\hat{\sigma}^2_{i(i-1)} = \sum_{j=i}^{s+1} l_{i,j}m_{j(j-1)} \tag{17}$$

for $i = 1, 2, \ldots, s, s+1$. These estimates of variance components lead to estimates of genic correlations,

$$\hat{\theta}_j = \sum_{i=1}^{j} \hat{\sigma}^2_{i(i-1)}/\hat{\sigma}^2, \tag{18}$$

and their complements,

$$\hat{f}_j = \frac{\hat{\sigma}^2_{j(j-1)}}{\sum\limits_{i=j}^{s+1} \hat{\sigma}^2_{i(i-1)}} = \frac{\theta_j - \theta_{j-1}}{1 - \theta_{j-1}}, \tag{19}$$

for $j = 1, 2 \ldots s, s+1$.

A Numerical Example

To illustrate the estimation procedure described above, let us consider the hypothetical example given in the Appendix. A total of $n_0 = 232$ diploid individuals or $2n_0 = 464$ gametes are sampled from a four-level sampling hierarchy ($s = 4$). Thus, there are five sources of variation: among populations, among subpopulations within populations, among sub-sub-populations within subpopulations, among individuals within sub-subpopulations, and among alleles within individuals. As shown in Table A1, the number of subdivisions at each level are censused, $r_1 = 4$, $r_2 = 11$, $r_3 = 31$, and $r_4 = 232$, along with $r_0 = 1$ and $r_5 = 464$. Thus, the degrees of freedom (d_i, $i = 1, 2, 3, 4, 5$) are $d_1 = 3$, $d_2 = 7$, $d_3 = 20$, $d_4 = 201$, and $d_5 = 232$.

The sums of squares and mean squares are calculated using formulas given in Table 1 for $s = 4$. The uncorrected sums of squares for the indicator variables at different levels of hierarchy are first calculated: $\Sigma X_0^2/2n_0 = 114.01$; $\Sigma X_1^2/2n_1 = 117.26$; $\Sigma X_2^2/2n_2 = 117.74$; $\Sigma X_3^2/2n_3 = 123.53$; $\Sigma X_4^2/2 = 173$; and $\Sigma X_5^2 = 230$. The corrected sums of squares are: $ss_{1(0)} = 117.26 - 114.01 = 3.25$; $ss_{2(1)} = 117.74 - 117.26 = 0.48$; $ss_{3(2)} = 123.53 - 117.74 = 5.79$; $ss_{4(3)} = 173 - 123.53 = 49.47$; and $ss_{5(4)} = 230 - 173 = 57$. The vector of mean squares, which is calculated as $m_{i(i-1)} = ss_{i(i-1)}/d_i$ for $i = 1, 2 \ldots 5$, is $\mathbf{m} = [1.0832\ 0.0693\ 0.2893\ 0.2461\ 0.2457]'$. The ANOVA estimator of variance components is obtained by substituting the inverse of the $\mathbf{K}$-matrix in equation (A2) and $\mathbf{m}$ into equation (15):

$$\hat{\sigma}^2 = \begin{bmatrix} 0.0092 & -0.0111 & -0.0001 & 0.002 & 0 \\ 0 & 0.0259 & -0.0358 & 0.0098 & 0 \\ 0 & 0 & 0.0781 & -0.0781 & 0 \\ 0 & 0 & 0 & 0.5 & -0.5 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \times \begin{bmatrix} 1.0832 \\ 0.0693 \\ 0.2893 \\ 0.2461 \\ 0.2457 \end{bmatrix} = \begin{bmatrix} 0.0097 \\ -0.0061 \\ 0.0034 \\ 0.0002 \\ 0.2457 \end{bmatrix}.$$

Thus, estimates of genic correlations (*F*-statistics) for this four-level hierarchy are as follows:

$$\hat{\theta}_1 = \frac{\hat{\sigma}^2_{1(0)}}{\hat{\sigma}^2} = 0.0375$$

$$\hat{\theta}_2 = \frac{\hat{\sigma}^2_{1(0)} + \hat{\sigma}^2_{2(1)}}{\hat{\sigma}^2} = 0.0375$$

$$\hat{\theta}_3 = \frac{\hat{\sigma}^2_{1(0)} + \hat{\sigma}^2_{2(1)} + \hat{\sigma}^2_{3(2)}}{\hat{\sigma}^2} = 0.0505$$

$$\hat{\theta}_4 = \frac{\hat{\sigma}^2_{1(0)} + \hat{\sigma}^2_{2(1)} + \hat{\sigma}^2_{3(2)} + \hat{\sigma}^2_{4(3)}}{\hat{\sigma}^2} = 0.0513$$

$$\hat{\theta}_5 = 1,$$

where $\hat{\sigma}^2 = \Sigma_{i=1}^5\ \hat{\sigma}^2_{i(i-1)} = 0.2590$, which is slightly higher than the expected value of 0.25 for gene frequency of 0.5.

The negative estimate of $\hat{\sigma}^2_{2(1)}$ is set to zero for estimating F-statistics, leading to $\hat{\theta}_1 = \hat{\theta}_2$. The correspondence between this model and that of Weir (1996, table 5.6) for the four-level hierarchy is evident: $\theta_1 = \theta_P$, $\theta_2 = \theta_S$, $\theta_3 = \theta_{SS}$, and $\theta_4 = F$. The complementary F-statistics are estimated using equation (19):

$$\hat{f}_1 = \hat{\theta}_1 = 0.0375$$

$$\hat{f}_2 = \frac{\hat{\theta}_2 - \hat{\theta}_1}{1 - \hat{\theta}_1} = 0$$

$$\hat{f}_3 = \frac{\hat{\theta}_3 - \hat{\theta}_2}{1 - \hat{\theta}_2} = 0.0135$$

$$\hat{f}_4 = \frac{\hat{\theta}_4 - \hat{\theta}_3}{1 - \hat{\theta}_3} = 0.0009$$

$$\hat{f}_5 = 1.$$

These estimates can be used to verify the parametric relationships between F-statistics (cf. eq. 6)

$$\begin{aligned}
(1 - \hat{\theta}_1) &= (1 - \hat{f}_1) \\
(1 - \hat{\theta}_2) &= (1 - \hat{\theta}_1)(1 - \hat{f}_2) \\
(1 - \hat{\theta}_3) &= (1 - \hat{\theta}_1)(1 - \hat{f}_2)(1 - \hat{f}_3) = (1 - \hat{\theta}_2)(1 - \hat{f}_3) \\
(1 - \hat{\theta}_4) &= (1 - \hat{\theta}_1)(1 - \hat{f}_2)(1 - \hat{f}_3)(1 - \hat{f}_4) \\
&= (1 - \hat{\theta}_2)(1 - \hat{f}_3)(1 - \hat{f}_4) = (1 - \hat{\theta}_3)(1 - \hat{f}_4)
\end{aligned}$$

Discussion

This paper has developed a general nested ANOVA framework by which estimation of F-statistics can be made from data with an arbitrary s-level hierarchical population structure. A generalized relationship among F-statistics (eq. 6) is derived to extend Wright's (1951) original relationship (eq. 1). Parametric relationship of intraclass correlations (F-statistics) with variance components (eqs. 4 and 5) and estimators of F-statistics (eqs. 18 and 19) are simply a generalization of the well-known development of Cockerham (1969, 1973), Weir and Cockerham (1984), and Weir (1996). For example, when population subdivisions are three- or four-level hierarchies (i.e., $s = 3$ or 4), my results are reduced to those of Weir (1996, tables 5.5 and 5.6), as shown above for $s = 4$.

My inquiry into a general procedure of estimating F-statistics is motivated by the following considerations. First, this general estimation procedure removes the need to have a different set of formulas for the sample with a different population structure, as long as levels of hierarchy are appropriately identified. This feature is particularly useful if an attempt is made to implement this estimation procedure into a computer program. Second, although most studies on population structure have estimated F-statistics from samples with four or fewer hierarchies (for which detailed estimation procedures are given in Weir 1996), this procedure will stimulate an interest in estimating F-statistics from samples with higher levels of hierarchy. This is particularly fitting with the increasing use of F-statistics to characterize interspecific differentiation (Porter 1990; Wolf and Soltis 1992; Mayer et al. 1994) or even higher hierarchical taxonomic structure.

To clearly describe this estimation procedure for a general hierarchical population structure, the ANOVA estimators of F-statistics are developed for one of the alleles at one locus. If a locus has only two alleles, either allele will give the same estimates of variance components so that only one allele is needed for estimating F-statistics as shown in this paper. There are two methods that can extend our treatment to cases of multiple alleles and loci. One way to combine information from different alleles and loci is to average variance components across alleles and loci and then to compute F-statistics based on the averaged variance components (Weir and Cockerham 1984; Weir 1996). The alternative way is to extend the ANOVA procedure directly, using multivariate analysis of variance (MANOVA) techniques (see Long 1986). In the MANOVA approach, indicator variable X is replaced with an **X**-vector of length $k - 1$ for k segregating alleles at a locus. The MANOVA table with the same structure as Table 1 can be constructed using the vector analog of linear model (7). The sum squares and mean squares are replaced by the matrices of sum squares and cross-products and mean squares and cross-products, respectively. The coefficient matrix as given in equation (14) remains the same, with variance components ($\sigma^2_{j(j-1)}$) replaced everywhere by matrices of covariance components ($\Sigma_{j(j-1)}$). Thus, genic correlations (F-statistics) can be estimated using the matrix analog of equation (18):

$$\hat{\theta}_j = \frac{1}{k-1}\mathrm{tr}\left[\hat{\Sigma}^{-1/2}(\hat{\Sigma}_{1(0)} + \hat{\Sigma}_{2(1)} + \ldots + \hat{\Sigma}_{j(j-1)})\hat{\Sigma}^{-1/2}\right],$$

where the total covariance matrix $\Sigma = \Sigma_{1(0)} + \Sigma_{2(1)} + \ldots + \Sigma_{s+1(s)}$ with the dimensions of $(k - 1) \times (k - 1)$ and tr denote the trace of a matrix. Evidently, for a multiallelic locus, the F-statistics estimated in this way are essentially the averages of estimates for all alleles at the locus. With multiple, independent loci, the **X**-vector is simply extended for successive loci, resulting in a larger covariance matrix (Long 1986). However, the sensitivity of Long's (1986) extension to linkage disequilibrium and sampling error remain to be evaluated.

The ANOVA estimators of F-statistics given in this paper can be appropriately used to estimate demographic and genetic parameters (e.g., gene flow) if the mutation process at given loci fits to the infinite-alleles model with low mutation rate. This model is a reasonable approximation for most of allozyme loci, but probably not for many microsatellite loci expressing higher mutation rate (> 10^{-3} per generation) and a stepwise pattern of mutation (e.g., Weber and Wong 1993). For this reason, new measures of population subdivision that are equivalent to F-statistics but are based on the stepwise mutation model have been recently proposed to take into account allele sizes for microsatellite loci (Slatkin 1995; Michalakis and Excoffier 1996; Rousset 1996). These new statistics are estimated using the average sum of squares of the differences in allele size within and among subdivisions. As shown by Michalakis and Excoffier (1996), the ANOVA estimators of the new statistics can be obtained because the sum of squared differences *(SSD)* between allele sizes is the same as a conventional sum of squares (Li 1976, p. 64). Analogous to the ANOVA in Table 1, the total

SSD can be decomposed into *s* components corresponding to *s*-level hierarchy:

$$\text{Total } SSD = \sum_{i=1}^{s} SSD_{i(i-1)} = \sum_{i=1}^{s} \left(\sum \frac{\delta_i^2}{2n_i} - \sum \frac{\delta_{i-1}^2}{2n_{i-1}} \right),$$

where δ_i^2 is the sum of Euclidean distances between pairs of gametes or haplotypes at the *i*th level of hierarchy. Thus, the substitution of the mean squares by the mean squared differences ($MSD_{i(i-1)} = SSD_{i(i-1)}/d_i$) in equation (15) would lead to the required ANOVA estimators of new statistics for microsatellite data. It should be noted that because, for diploid data, the *SSD* analysis assumes a random union of gametes at the *s*th level of hierarchy (i.e., no distinction between the correlations of genes within and among individuals), it does not allow for an estimation of departure from random mating, as is done in the above ANOVA analysis.

Although the ANOVA estimators of variance components are unbiased (cf. eq. 15), the ratios of these estimators (i.e., estimated *F*-statistics) are inherently biased downward. This inherent downward bias is well known in the genetic analysis of quantitative traits (e.g., Ginsburg 1973; Ponzoni and James 1978), but is not well investigated for discrete characters. For a given gene frequency, the degree of the downward bias is partly dependent on the sampling of different subdivisions and individuals within subdivisions. For balanced data from *r* random mating populations of size *n* (i.e., one-level hierarchy), the expected bias of estimated intraclass correlation ($\hat{\theta}_1$) is approximated by (cf. Ponzoni and James 1978):

$$E(\hat{\theta}_1 - \theta_1) \cong \frac{-2(1-\theta_1)[\theta_1 + (1-\theta_1)/2n][\theta_1 + (1-\theta_1)/2rn]}{r-1}.$$

The inverse relationship between the bias and the number of sampled populations (*r*) indicates *r* is the primary determinant of the bias. To examine the effect of gene frequencies on estimated *F*-statistics, I also carried out computer simulations based on the beta distribution of gene frequencies among populations arising from Wright's (1943) island model. These simulation results (unpubl.) show that extreme gene frequencies (<0.2) contribute further to the downward bias of estimated *F*-statistics particularly with 10 or fewer populations sampled. More research is needed to clarify the interrelationships between gene frequencies, sample dimension, and bias from estimated *F*-statistics. These discussions serve to emphasize that although the ANOVA estimators of *F*-statistics as given in this and earlier studies (e.g., Weir and Cockerham 1984; Weir 1996) have considered sample sizes and number of subdivisions sampled at each level of hierarchy and have removed some of the bias incurred in the other estimation procedures (e.g., Nei 1973, 1977; Crow and Aoki 1984), they are *not* unbiased as Slatkin and Barton (1989, p. 1357) and Michalakis and Excoffier (1996, p. 1062) have implied. As Weir and Cockerham (1984, p. 1359) clearly noted, the ANOVA estimators of *F*-statistics are unbiased only if one is willing to "take the expectation of a ratio to be the ratio of expectations." This certainly is not the case for small samples and extreme gene frequencies as shown in the above simulation results.

Because subdivisions at all levels of the hierarchy are assumed to be completely random, the resulting estimators of *F*-statistics can be used to make evolutionary inferences about the existing population structure under varying population genetic models (e.g., Cockerham and Weir 1987; Weir 1996, pp. 179–183). One important use of estimated *F*-statistics is the inference about gene flow among subdivisions (Crow and Aoki 1984; Slatkin and Barton 1989; Cockerham and Weir 1993). Although these studies are limited to the inference with the one-level hierarchy, the extension of Slatkin and Voelm's (1991) hierarchical island model or other similar models may enable inference about gene flow at multiple levels of the hierarchy using the estimated hierarchical *F*-statistics derived in this paper. Recently, there has been a growing interest in carrying this inference beyond intraspecific differentiation (Porter 1990; Wolf and Soltis 1992; Mayer et al. 1994). These studies have also suggested that the ANOVA estimation procedure (e.g., Weir and Cokerham 1984) is preferable to the method of Nei (1973, 1977) because the ANOVA estimators of *F*-statistics provide a more realistic estimate of gene flow based on known sibling-species relationships. In a comparative assessment of the ANOVA estimators vis-á-vis Nei's original estimators and Nei and Chesser's (1983) estimators with a correction for sampling effect, Chakraborty and Danker-Hopfe (1991) concluded that empirical differences in the two sets of estimators were generally insignificant and that the differences were more philosophical than numerical. However, the philosophical differences (random model with ANOVA estimators vs. fixed model with Nei's estimators) may be most relevant and important in inferring genetic and demographic structures of natural populations (e.g., gene flow) using estimated *F*-statistics.

Acknowledgments

I thank Dr. E. Zouros and two anonymous reviewers for helpful comments. This research has been supported in part by the Natural Sciences and Engineering Research Council of Canada grant OGP3983.

Literature Cited

Chakraborty, R. 1980. Gene-diversity analysis in nested subdivided populations. Genetics 96:721–723.

Chakraborty, R., and H. Danker-Hopfe. 1991. Analysis of population structure: a comparative study of different estimators of Wright's fixation indices. Pp. 203–254 *in* C. R. Rao and R. Chakraborty, eds. Handbook of statistics. Vol. 8. Elsevier, North-Holland.

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

———. 1973. Analysis of gene frequencies. Genetics 74:679–700.

Cockerham, C. C., and B. S. Weir. 1987. Correlations, descent measures: drift with migration and mutation. Proc. Natl. Acad. Sci. USA 84:8512–8514.

———. 1993. Estimation of gene flow from *F*-statistics. Evolution 47:855–863.

Crow, J. F., and K. Aoki. 1984. Group selection for a polygenic behavioral trait: estimating the degree of subdivision. Proc. Natl. Acad. Sci. USA 81:6073–6077.

Ginsburg, E. H. 1973. On the planning of the experiment on estimation of intraclass correlation. Biom. Z. 15:47–52.

Li, C. C. 1976. First course in population genetics. Boxwood, Pacific Grove, CA.

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

MAYER, M. S., P. S. SOLTIS, AND D. E. SOLTIS. 1994. The evolution of the *Streptanthus glandulosus* complex (Cruciferae): genetic divergenece and gene flow in serpentine endemics. Am. J. Bot. 81: 1288–1299.

MICHALAKIS, Y., AND L. EXCOFFIER. 1996. A generic estimation of population subdivisions using distances between alleles with special reference to microsatellite loci. Genetics 142:1061–1064.

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

———. 1977. *F*-statistics and analysis of gene diversity in subdivided populations. Ann. Hum. Genet. 41:225–233.

NEI, M., AND R. K. CHESSER. 1983. Estimation of fixation indices and gene diversities. Ann. Hum. Genet. 47:253–259.

PONZONI, R. W., AND J. W. JAMES. 1978. Possible biases in heritability estimates from intraclass correlation. Theor. Appl. Genet. 53:25–27.

PORTER, A. H. 1990. Testing nominal species boundaries using gene flow statistics: the taxonomy of the hybridizing admiral butterflies (*Limenitis:* Nymphalida). Syst. Zool. 39:131–147.

ROUSSET, F. 1996. Equilibrium values of measures of population subdivision for stepwise mutation processes. Genetics 142:1357–1362.

SLATKIN, M. 1995. A measure of population subdivision based on microsatellite allele frequencies. Genetics 139:457–462.

SLATKIN, M., AND N. H. BARTON. 1989. A comparison of three direct methods for estimating average levels of gene flow. Evolution 43: 1349–1368.

SLATKIN, M., AND L. VOELM. 1991. F_{ST} in a hierarchical island model. Genetics 127:627–629.

WAHLUND, S. 1928. Zusammensetzung von population und korrelationserscheinung vom standpunkt der vererbungslehre aus betrachtet. Hereditas 11:65–106.

WEBER, J. L., AND C. WONG. 1993. Mutation of human short tandem repeats. Hum. Mol. Genet. 2:1123–1128.

WEIR, B. S. 1996. Genetic data analysis. II. Sinauer Associates, Sunderland, MA.

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

WOLF, P. G., AND P. S. SOLTIS. 1992. Estimates of gene flow among populations, geographic races, and species in the *Ipomopsis aggregata* complex. Genetics 130:639–647.

WRIGHT, S. 1943. Isolation by distance. Genetics 28:114–138.

———. 1951. The genetic structure of populations. Ann. Eugen. 15: 323–354.

———. 1965. The interpretation of population structure by *F*-statistics with special regard to systems of mating. Evolution 19:395–420.

———. 1978. Evolution and the genetics of populations. Vol. 4. Variability within and among natural populations. Univ. of Chicago Press, Chicago.

Corresponding Editor: E. Zouros

APPENDIX

In the text, I obtained a general expression (14b) for deriving all coefficients of expected mean squares from the nested ANOVA of indicator variable X representing one of the two alleles at a locus. In this and other ANOVA estimation procedures, calculation of these coefficients is a key step to estimation of variance components and F-statistics. Here I will describe, through a hypothetical example, an algorithm that may facilitate an implementation of this general estimation procedure into a computer program.

Consider the hypothetical example given in Table A1 for a four-level sampling hierarchy with a total of $n_0 = 232$ diploid individuals or $2n_0 = 464$ gametes sampled. A sampled gamete can take the value of zero or one, depending on whether an computer-generated uniform variate is less than or equal to or greater than the global gene frequency of 0.5, respectively. The number of individuals in each subdivision at each level of the hierarchy (n_0, n_1, n_2, and n_3) are counted. The number of subdivisions at each level (r_1, r_2, r_3, and r_4) are also censussed with $r_0 = 1$; the degrees of freedom (d_i) are calculated as $d_i = r_i - r_{i-1}$. To describe the algorithm, equation (14b) is rewritten as:

$$k_{i,j} = \frac{2}{d_i}(v_{i,j} - v_{i-1,j}) \qquad \text{(A1)}$$

TABLE A1.

Numbers of populations (POP), subpopulations (SPOP), sub-sub-populations (SSPOP), and individuals (n) sampled in a hypothetical example; $r_1 = 4$, $r_2 = 11$, $r_3 = 31$, $r_4 = 232$.

POP	SPOP	SSPOP	n_3	n_2	n_1	n_0
1	1	1	10			
		2	12			
		3	13	35		
	2	1	7			
		2	11			
		3	5			
		4	3	26		
	3	1	10	10	71	
2	1	1	1			
		2	3			
		3	5			
		4	3	12		
	2	1	14			
		2	11			
		3	10			
		4	12	47		
	3	1	4			
		2	2			
		3	13	19	78	
3	1	1	13	13	13	
4	1	1	2			
		2	5			
		3	9			
		4	7	23		
	2	1	13	13		
	3	1	7	7		
	4	1	2			
		2	12			
		3	2			
		4	9			
		5	2	27	70	232

The quantities $v_{i,j}$ can be calculated using the following steps: (1) square the numbers in column j; (2) divide each square by the number in the same or next occupied cell of column $i \leq j$; and (3) sum the quantities so obtained to produce $v_{i,j}$.

To illustrate the application of this algorithm, let us, for example, calculate $k_{1,3}$:

$$v_{0,3} = (10^2 + 12^2 + \ldots + 9^2 + 2^2)/232 = 9.89$$

$$v_{1,3} = \frac{10^2 + 12^2 + \ldots + 10^2}{71} + \frac{1^2 + 3^2 + \ldots + 13^2}{78} + \frac{13^2}{13} + \frac{2^2 + 5^2 + \ldots + 2^2}{70} = 42.05$$

The coefficient $k_{1,3}$ is calculated by substituting these two quantities into equation (A1), $k_{1,3} = 2(42.05 - 9.89)/3 = 21.44$. The K-matrix containing all coefficients of expected mean squares is obtained by repeatedly applying the above algorithm for all pairs of columns with different sums of individuals in Table A1:

$$\mathbf{K} = \begin{bmatrix} 108.13 & 46.43 & 21.44 & 2 & 1 \\ 0 & 38.55 & 17.67 & 2 & 1 \\ 0 & 0 & 12.81 & 2 & 1 \\ 0 & 0 & 0 & 2 & 1 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}. \qquad \text{(A2)}$$