Evolution, 38(6), 1984, pp. 1358–1370

ESTIMATING *F*-STATISTICS FOR THE ANALYSIS OF POPULATION STRUCTURE[1]

B. S. WEIR AND C. CLARK COCKERHAM
Department of Statistics, North Carolina State University, Raleigh, NC 27695-8203

Received June 28, 1983. Accepted May 7, 1984

This journal frequently contains papers that report values of *F*-statistics estimated from genetic data collected from several populations. These parameters, F_{ST}, F_{IT}, and F_{IS}, were introduced by Wright (1951), and offer a convenient means of summarizing population structure. While there is some disagreement about the interpretation of the quantities, there is considerably more disagreement on the method of evaluating them. Different authors make different assumptions about sample sizes or numbers of populations and handle the difficulties of multiple alleles and unequal sample sizes in different ways. Wright himself, for example, did not consider the effects of finite sample size.

The purpose of this discussion is to offer some unity to various estimation formulae and to point out that correlations of genes in structured populations, with which *F*-statistics are concerned, are expressed very conveniently with a set of parameters treated by Cockerham (1969, 1973). We start with the parameters and construct appropriate estimators for them, rather than beginning the discussion with various data functions. The extension of Cockerham's work to multiple alleles and loci will be made explicit, and the use of jackknife procedures for estimating variances will be advocated. All of this may be regarded as an extension of a recent treatment of estimating the coancestry coefficient to serve as a measure of genetic distance (Reynolds et al., 1983). Other methods of estimating coancestry were evaluated in that paper and found to be generally unsatisfactory as measures of distance for divergence due only to drift.

We begin by giving some detail of the estimation procedures using just one of the alleles at one locus, with emphasis on incorporating the effects of sample sizes and of the number of populations sampled. A weighting scheme over alleles which is expected to be close to minimum variance and unbiased in some situations, but which appears to be generally of minimum mean square error, is then presented. It is hoped that the generality of the estimators given will be useful in comparing the results from different studies with different sample structures.

One Allele

The procedures are most easily developed for one of the alleles at one locus and then the resulting estimators combined appropriately for several alleles and loci. We are concerned with the following parameters: F, the correlation of genes within individuals ("inbreeding"); θ, the correlation of genes of different individuals in the same population ("coancestry"); and f, the correlation of genes within individuals within populations. The three parameters are related by

$$f = (F - \theta)/(1 - \theta).$$

Cockerham (1969, 1973) showed that, for all intents and purposes, these parameters are related to Wright's *F*-statistics as

$$F = F_{IT}, \qquad \theta = F_{ST}, \qquad f = F_{IS}.$$

[1] Paper No. 8906 of the Journal Series of the North Carolina Agricultural Research Service, Raleigh, North Carolina 27695. This investigation was supported in part by NIH Research Grant No. GM 11546 from the National Institute of General Medical Sciences.

We retain the F, θ, f notation, however, because of the considerable confusion in the literature about F-statistics. For example, Nei (1976 and many other places) says that F_{ST} varies according to the number of populations observed, indicating that he regards F_{ST} as a statistic and not as a parameter. Our parameters have precise and unambiguous definitions—F for pairs of alleles as constituted within individuals, and θ for pairs of alleles between individuals within populations. They depend on population size and history, but are unaffected by aspects of the sampling scheme—numbers of alleles observed per locus, numbers of individuals sampled per population or numbers of populations sampled. Only f can be estimated from frequency data for a single population.

The sampling model is as follows. A number, r, of populations of the same size are considered to have descended separately from a single ancestral population that was in both Hardy-Weinberg equilibrium and linkage equilibrium. The populations are considered to have been maintained under the same conditions, so that samples from them differ because of the genetic sampling between generations and the statistical sampling of the individuals for observation. We construct statistics that have expected values differing from the parameters of interest (F and θ) only by multiples of the expected allelic frequencies. These unknown (ancestral) frequencies can be removed by taking appropriate ratios.

Expectations of statistics are taken over all possible samples and all possible replicate populations. The requirement of equal population (but not sample) sizes is needed to ensure that only one F or θ results from the expectation process. Otherwise, our procedures provide estimates of complex functions of unequal F's and θ's, as do other procedures. Although we will discuss their paper in more detail elsewhere, we point out here that Nei and Chesser (1983), in estimating F-statistics, do not consider replication over replicate populations.

Following the approach of Cockerham (1969, 1973), we perform analyses of variance of the frequencies for the allele A under consideration. Observed components of variance: a for between populations; b for between individuals within populations; and c for between gametes within individuals have expectations

$$\begin{aligned} Ea &= p(1 - p)\theta, \\ Eb &= p(1 - p)(F - \theta), \\ Ec &= p(1 - p)(1 - F), \end{aligned}$$

where p is the expected frequency of the allele, and is equal to its frequency in the ancestral population. The following estimators for the three parameters are suggested immediately

$$\begin{aligned} 1 - \hat{F} &= \frac{c}{a + b + c}, \\ \hat{\theta} &= \frac{a}{a + b + c}, \\ 1 - \hat{f} &= \frac{c}{b + c}. \end{aligned} \tag{1}$$

To the extent that we may take the expectation of a ratio to be the ratio of expectations, these estimators are all unbiased. The simulation results we present later confirm the satisfactory nature of these estimators, and some expressions for bias are given in the Appendix.

While there are other statistics with the same expectations as a, b, c, these three quantities were obtained from a weighted analysis of variance (Cockerham, 1973). If $\tilde{p}_i$ is the frequency of allele A in the sample of size n_i from population i ($i = 1, 2, \ldots, r$) and $\tilde{h}_i$ is the observed proportion of individuals heterozygous for allele A, then

$$a = \frac{\bar{n}}{n_c}\left\{s^2 - \frac{1}{\bar{n} - 1}\left[\bar{p}(1 - \bar{p}) - \frac{r - 1}{r}s^2 - \frac{1}{4}\bar{h}\right]\right\} \tag{2}$$

$$b = \frac{\bar{n}}{\bar{n} - 1}\left[\bar{p}(1 - \bar{p}) - \frac{r - 1}{r}s^2 - \frac{2\bar{n} - 1}{4\bar{n}}\bar{h}\right] \quad (3)$$

$$c = \frac{1}{2}\bar{h} \quad (4)$$

where:

$\bar{n} = \sum_i n_i/r$, the average sample size

$n_c = \left(r\bar{n} - \sum_i n_i^2/r\bar{n}\right)/(r - 1) = \bar{n}(1 - C^2/r)$, with C^2 the squared coefficient of variation of sample sizes

$\bar{p} = \sum_i n_i\tilde{p}_i/r\bar{n}$, the average sample frequency of allele A

$s^2 = \sum_i n_i(\tilde{p}_i - \bar{p})^2/(r - 1)\bar{n}$, the sample variance of allele A frequencies over populations

$\bar{h} = \sum_i n_i\tilde{h}_i/r\bar{n}$, the average heterozygote frequency for allele A.

Notice that if the same allele is fixed in all samples, then (1) provides $\hat{\theta} = 0/0$, so that the estimate is undefined. We feel that this is an appropriate outcome since there is no way of knowing from presently observed homozygosity whether the different populations have just become homozygous or have been homozygous for some time that may have extended back to the founding population. When several loci are used, we see below that loci fixed for the same allele properly make no contribution to the estimators. We also note that, in the finite monoecious random mating situation, fixation corresponds to complete identity by descent of all genes present, or to $\theta = 1$, so that we do not subscribe to the view of Nei and Chakravarti (1977) that fixation should give an estimate of 0 for θ.

To complete this section, we manipulate equations (1) into forms that illustrate the effects of $\bar{n}$, r and variation in sample size (see the equation at the bottom of this page). Notice that these formulae contain explicit corrections for a small number of populations and for small or unequal sample sizes over and above the weighting present in means ($\bar{p}$) and variances (s^2).

Special Cases

We advocate the use of general expressions (1)–(4) since they involve no ap-

$$1 - \hat{F} = \frac{\left(1 - \frac{C^2}{r}\right)\bar{h}}{2\left[1 - \frac{\bar{n}C^2}{r(\bar{n} - 1)}\right]\bar{p}(1 - \bar{p}) + 2\left[1 + \frac{(r - 1)\bar{n}C^2}{r(\bar{n} - 1)}\right]\frac{s^2}{r} + \left[\frac{C^2}{r(\bar{n} - 1)}\right]\frac{\bar{h}}{2}}, \quad (5)$$

$$\hat{\theta} = \frac{s^2 - \frac{1}{\bar{n} - 1}\left[\bar{p}(1 - \bar{p}) - \frac{r - 1}{r}s^2 - \frac{\bar{h}}{4}\right]}{\left[1 - \frac{\bar{n}C^2}{r(\bar{n} - 1)}\right]\bar{p}(1 - \bar{p}) + \left[1 + \frac{(r - 1)\bar{n}C^2}{r(\bar{n} - 1)}\right]\frac{s^2}{r} + \left[\frac{C^2}{r(\bar{n} - 1)}\right]\frac{\bar{h}}{4}}, \quad (6)$$

$$1 - \hat{f} = \frac{\bar{h}}{\left[\frac{2\bar{n}}{\bar{n} - 1}\right]\bar{p}(1 - \bar{p}) - \left[\frac{2\bar{n}(r - 1)}{r(\bar{n} - 1)}\right]s^2 - \left[\frac{1}{\bar{n} - 1}\right]\frac{\bar{h}}{2}}. \quad (7)$$

$$1 - \hat{F} = \frac{\bar{h}}{2\bar{p}(1 - \bar{p}) + 2s^2/r}$$

$$\hat{\theta} = \frac{s^2 - \frac{1}{n-1}\left[\bar{p}(1 - \bar{p}) - \frac{r-1}{r}s^2 - \frac{1}{4}\bar{h}\right]}{\bar{p}(1 - \bar{p}) + s^2/r}$$

proximations and would lead to computational uniformity among published values of such estimators by different investigators using different structured populations. By making a series of assumptions in (5)–(7), we can show how the general formulae reduce to some of those in common use.

Equal Sample Sizes.—When all samples are of the same size $n = \bar{n}$, $C^2 = 0$ and the equations at the top of this page pertain with no change, other than $n = \bar{n}$, in $1 - \hat{f}$ (equation 7).

Large Number of Populations.—Although it is not often the case, there is considerable simplification when a sufficient number of populations are sampled that terms in $1/r$ may be ignored. In such cases, see the equation at the bottom of this page.

Large Sample Sizes.—If samples are large enough that terms in $1/\bar{n}$ can be ignored, see the equation at the top of the next page, with this last expression also being given by Nei (1977).

Common Expressions.—We have looked at the most recent 20 or so papers in this journal that have presented F-statistics, and have found that they fall into a few categories.

Many papers do not give computational formulae, but generally refer to work by Wright (1943, 1951, 1965, 1973) or Nei (1973, 1977), and any assumptions made about sample sizes are not stated. Such papers are by Chesser (1983), Ferrari and Taylor (1981), Fleischer (1983), Hiebert and Hamrick (1983), Patton and Feder (1981), Patton and Yang (1977), Schaal and Smith (1980), Schoen (1982), Schwaegerle and Schaal (1979), Sites and Greenbaum (1983), and Ward (1980).

The most common explicit computational formula is

$$F_{ST} = \frac{\sigma_p^2}{\bar{p}(1 - \bar{p})}, \tag{8}$$

which we will rewrite as

$$\hat{\theta} = \frac{s^2}{\bar{p}(1 - \bar{p})}. \tag{9}$$

This formula has appeared where sample mean and variance of gene frequencies are either unweighted or weighted for un-

$$1 - \hat{F} = \frac{\bar{h}}{2\bar{p}(1 - \bar{p})},$$

$$\hat{\theta} = \frac{s^2 - \frac{1}{\bar{n}-1}[\bar{p}(1 - \bar{p}) - s^2 - \bar{h}/4]}{\bar{p}(1 - \bar{p})},$$

$$1 - \hat{f} = \frac{(\bar{n} - 1)\bar{h}}{2\bar{n}\bar{p}(1 - \bar{p}) - 2\bar{n}s^2 - \bar{h}/2}.$$

$$1 - \hat{F} = \frac{\left[1 - \frac{C^2}{r}\right]\bar{h}}{2\left[1 - \frac{C^2}{r}\right]\bar{p}(1 - \bar{p}) + 2\left[1 + \frac{r-1}{r}C^2\right]s^2/r},$$

$$\hat{\theta} = \frac{s^2}{\left[1 - \frac{C^2}{r}\right]\bar{p}(1 - \bar{p}) + \left[1 + \frac{r-1}{r}C^2\right]s^2/r},$$

$$1 - \hat{f} = \frac{\bar{h}}{2\bar{p}(1 - \bar{p}) - 2(r-1)s^2/r},$$

equal sample size. From the above approximations, we see that this result supposes that there are a large number of large samples, so that both $1/r$ and $1/\bar{n}$ can be ignored. Papers using this formula are by Avise and Felley (1979), Baker (1981), Baker et al. (1982), Brittnacher et al. (1978), Eanes and Koehn (1978), Foltz and Hoogland (1983), Guries and Ledig (1982), and Winans (1980). It should be pointed out that (8) is also given in the most commonly cited theoretical papers, including those by Nei and Wright mentioned above, as well as Kirby (1975) and Workman and Niswander (1970).

Several authors are concerned with removing sampling bias from their estimates. Sample size is explicitly accounted for in our formulations, but several attempts have been made to remove sampling effects by subtracting terms involving the reciprocal of sample size. Such an approach is suggested by Nei and Imaizumi (1966), made explicit by Workman and Niswander (1970), discussed by Wright (1978), and used by Chesser (1983), Ellstrand and Levin (1980), Levin (1977, 1978), Levin et al. (1979), Pamilo (1983), and Ryman et al. (1980). The approach assumes $\bar{p}(1 - \bar{p})$ to be a constant, and removes within population sample variance of gene frequencies from s^2. We cannot derive any of the formulae used from our expressions, but if we substitute the Hardy-Weinberg heterozygote frequencies $\tilde{h}_i = 2\tilde{p}_i(1 - \tilde{p}_i)$ and ignore terms in $1/r$, we find that

$$\hat{\theta} = \frac{2\bar{n} - 1}{2(\bar{n} - 1)} \frac{s^2}{\bar{p}(1 - \bar{p})} - \frac{1}{2(\bar{n} - 1)}.$$

We do find more agreement with formulae often used for f or F_{IS}. If a fixation index $\tilde{F}_i$ is calculated in sample i as

$$\tilde{F}_i = 1 - \frac{\tilde{h}_i}{2\tilde{p}_i(1 - \tilde{p}_i)},$$

then the weighted sum (Kirby, 1975; Nei, 1977)

$$\sum_i n_i\tilde{p}_i(1 - \tilde{p}_i)\tilde{F}_i / \sum_i n_i\tilde{p}_i(1 - \tilde{p}_i)$$

is the same as our $\hat{f}$, provided terms in $1/\bar{n}$ are ignored. The same would not hold if the $\tilde{F}_i$ were weighted just by sample size (Workman and Niswander, 1970). Papers using the weighted average of fixation indices are by Avise and Felley (1979), Baker et al. (1982), Guries and Ledig (1982), and Ryman et al. (1980).

The differences likely between using (5)–(7) and approximations such as (9) are illustrated in the simulation results presented later.

Random Union of Gametes

When gametes unite at random within populations, the correlations between pairs of genes are the same whether those genes are located in one or two individuals. There is no need then to distinguish between F and θ, $Eb = 0$, and we pool the

$$\hat{\theta} = \frac{s^2 - \frac{1}{2\bar{n} - 1}\left[\bar{p}(1 - \bar{p}) - \frac{r - 1}{r}s^2\right]}{\left[1 - \frac{2\bar{n}C^2}{(2\bar{n} - 1)r}\right]\bar{p}(1 - \bar{p}) + \left[1 + \frac{2\bar{n}(r - 1)C^2}{(2\bar{n} - 1)r}\right]\frac{s^2}{r}}.$$

mean squares for between and within individuals. The pooled value is

$$d = \frac{2\bar{n}}{2\bar{n} - 1}\left[\bar{p}(1 - \bar{p}) - \frac{r - 1}{r}s^2\right]$$

with

$$\varepsilon d = \bar{p}(1 - \bar{p})(1 - \theta)$$

(Reynolds et al., 1983). Consequently, θ is estimated as

$$\hat{\theta} = \frac{a}{a + d}$$

where the quantity a is changed from that given above. The general expression is given in the equation at the top of this page. As expected, there is no need to employ heterozygote frequencies. Various approximations can again be used, and if $1/\bar{n}$ and $1/r$ can be ignored,

$$\hat{\theta} = \frac{s^2}{\bar{p}(1 - \bar{p})}.$$

Several Alleles at a Locus

Under our neutral model, any allele at any locus may be used to construct variance components *a, b, c* and so provide estimates of the parameters *F*, θ, *f*. If a locus has only two alleles, either allele will give the same values of *a, b, c* so that only one allele needs to be used in the analysis. In the more general case of several alleles at a locus, we index the alleles by u and could calculate variance components a_u, b_u, c_u for each allele, and form separate estimates such as

$$\hat{\theta}_u = \frac{a_u}{a_u + b_u + c_u}.$$

The problem of combining estimates over alleles is considered in the appendix, but we note here that the simple average over alleles $u = 1, 2, \ldots, v$

$$\hat{\theta}_U = \frac{1}{v}\sum_u \hat{\theta}_u,$$

or the average of estimates weighted by $(1 - \bar{p}_u)$

$$\hat{\theta}_{RH} = \frac{1}{v - 1}\sum_u (1 - \bar{p}_u)\hat{\theta}_u$$

suggested by Robertson and Hill (1984) does not appear to be as satisfactory as the weighting we advocated previously for distances (Reynolds et al., 1983)

$$\hat{\theta}_W = \frac{\sum_u a_u}{\sum_u (a_u + b_u + c_u)}.$$

Notice that

$$E\sum_u a_u = \theta\sum_u p_u(1 - p_u) = \theta\left(1 - \sum_u p_u^2\right) = \theta\alpha$$

$$E\sum_u (a_u + b_u + c_u) = \left(1 - \sum_u p_u^2\right) = \alpha,$$

so that $\hat{\theta}_W$ (as well as $\hat{\theta}_U$ and $\hat{\theta}_{RH}$) are unbiased, to a first approximation. This procedure is essentially a weighted average over alleles.

Several Loci

Under the neutral model, each locus could be used separately to estimate the same parameters *F*, θ, *f*. If the variance

components a, b, c are also subscripted with l to denote the lth locus, then

$$E \sum_l \sum_u a_{lu} = \theta \sum_l \alpha_l$$

$$E \sum_l \sum_u (a_{lu} + b_{lu} + c_{lu}) = \sum_l \alpha_l$$

leading to

$$\hat{\theta}_W = \sum_l \sum_U a_{lu} \Big/ \sum_l \sum_u \cdot(a_{lu} + b_{lu} + c_{lu}) \qquad (10)$$

as a single estimate based on the whole data set. Previous work (Reynolds et al., 1983) has shown that such weighted averages over loci perform better than unweighted averages. The weighted average over alleles and loci avoids any problems of zero denominators in the ratio estimates that can arise for some loci when there is only one allele in the entire sample.

Simulation Study

We have performed some simulations to illustrate both the effects of including sample size and population number corrections, and of our method of combining estimates over alleles and loci. The basic simulation followed finite, randomly mating monoecious populations of size 50 over 50 generations, giving a range of $F = \theta$ values from 0 to .39. Ten loci were followed, each with three alleles. Initial allelic frequencies were either made equal or widely disparate ($p_1 = .6$, $p_2 = .3$, $p_3 = .1$), and linkage λ between adjacent loci was either zero or .9 (recombination coefficients of .5 or .05). In generations 5, 25, and 50, samples of size 15, 20, and 25 were drawn from three replicate populations, and the whole procedure replicated 50 times. In Table 1 we present the means and mean square errors over replicates of a series of estimates (where v_l is the number of alleles found over the three populations at locus l and loci for which there is allelic variation).

$$\hat{\theta}_W = \sum_l \sum_u a_{lu} \Big/ \sum_l \sum_u (a_{lu} + b_{lu} + c_{lu})$$

$$\hat{\theta}_U = \frac{1}{m} \sum_l \frac{1}{v_l} \cdot \sum_u [a_{lu}/(a_{lu} + b_{lu} + c_{lu})]$$

$$\hat{\theta}_{RH} = \sum_l \sum_u (1 - \bar{p}_{lu}) \cdot [a_{lu}/(a_{lu} + b_{lu} + c_{lu})] \Big/ \cdot \sum_l (v_l - 1)$$

$$\hat{\theta}_T = \frac{1}{m} \sum_l \frac{1}{v_l} \sum_u [s_{lu}^2/\bar{p}_{lu}(1 - \bar{p}_{lu})]$$

$\hat{\theta}_M$, a matrix estimate defined in the Appendix.

It is clear from the table that $\hat{\theta}_W$ is a very satisfactory estimator over a wide range of conditions. While it is not necessarily the best estimator in every situation, it is the least biased in seven out of nine situations in Table 1 and has the smallest mean square error in four of the nine situations. It appears to do better for larger θ values and for linked loci. No single alternative estimator is better for more than one θ value. The "traditional" estimator $\hat{\theta}_T$ is the worst in terms of bias and mean square error for low θ values, with bias sometimes exceeding 50% of the true value, and better than only $\hat{\theta}_M$ for large θ values. It should be pointed out that $\hat{\theta}_M$ has the smallest variance in every case studied.

Further Population Subdivision

It is often the case that we recognize a demic structure within populations, so that we can replace the parameter θ by θ_1 for pairs of alleles between individuals within demes, and θ_2 for pairs of alleles between demes within populations (Cockerham, 1973). There are now variance components for populations (a), demes within populations (b_2), individuals within demes (b_1), and gametes within individuals (c). The expectations are now

$$Ea = p(1 - p)\theta_2$$

$$Eb_2 = p(1 - p)(\theta_1 - \theta_2)$$

$$Eb_1 = p(1 - p)(F - \theta_1)$$
$$Ec = p(1 - p)(1 - F)$$

leading to

$$1 - \hat{F} = \frac{c}{a + b_1 + b_2 + c},$$
$$\hat{\theta}_1 = \frac{a + b_2}{a + b_1 + b_2 + c},$$
$$\hat{\theta}_2 = \frac{a}{a + b_1 + b_2 + c}.$$

If observations are made on n_{ij} individuals for the jth deme ($j = 1, 2, \ldots, m_i$) sampled in the ith population ($i = 1, 2, \ldots, r$), then the computing formulae for the variance components are

$$c = MSG,$$
$$b_1 = \frac{1}{2}(MSI - MSG),$$
$$b_2 = \frac{1}{2n_3}(MSD - MSI),$$
$$a = \frac{1}{2n_2n_3}[n_3MSP - n_1MSD - (n_3 - n_1)MSI]$$

where

$$n_1 = \frac{1}{r - 1}\sum_i\sum_j \frac{(n.. - n_{i.})n_{ij}^2}{n_{i.}n..} \qquad n_{i.} = \sum_j n_{ij}$$

$$n_2 = \frac{1}{r - 1}\left[n.. - \frac{1}{n..}\sum_i n_{i.}^2\right] \qquad n.. = \sum_i n_{i.}$$

$$n_3 = \frac{1}{m. - r}\left[n.. - \sum_i\sum_j \frac{n_{ij}^2}{n_{i.}}\right] \qquad m. = \sum_i m_i$$

and

$$MSP = 2\sum_i n_{i.}(\tilde{p}_{i.} - \tilde{p}..)^2/(r - 1)$$
$$MSD = 2\sum_i\sum_j n_{ij}(\tilde{p}_{ij} - \tilde{p}_{i.})^2/(m. - r)$$

TABLE 1. Simulation results for estimating θ from $r = 3$ monoecious populations of size $N = 50$ mating at random, using $m = 10$ loci with $v = 3$ alleles each.

	$p_1 = p_2 = p_3 = 1/3$				$p_1 = 2p_2 = 6p_3 = 3/5$	
	$\lambda = .0$		$\lambda = .90$		$\lambda = .0$	
	Mean[1]	MS error[1,2]	Mean[1]	MS error[1,2]	Mean[1]	MS error[1,2]
$\theta = .049$						
W	.055	.032	.046	.023	.055	.033
U	.053	.026	.044	.022	.050	.020
RH	.052	.025	.044	.022	.049	.017
T	.079	.120	.071	.068	.077	.099
M	.051	.023	.043	.022	.048	.016
$\theta = .222$						
W	.226	.227	.220	.171	.234	.229
U	.206	.198	.201	.187	.201	.172
RH	.201	.210	.196	.210	.190	.194
T	.254	.358	.248	.275	.249	.274
M	.194	.230	.191	.231	.168	.375
$\theta = .395$						
W	.392	.349	.392	.409	.403	.476
U	.344	.527	.344	.569	.339	.719
RH	.331	.636	.328	.720	.318	.876
T	.434	.650	.433	.686	.428	.877
M	.313	.891	.311	.971	.240	2.663

[1] Over 50 replicates.
[2] MS error × 10^2.

$$MSI = \left[2\sum_i\sum_j n_{ij}\tilde{p}_{ij}(1 - \tilde{p}_{ij}) - \frac{1}{2}\sum_i\sum_j n_{ij}\tilde{h}_{ij}\right]/(n.. - m.)$$
$$MSG = \sum_i\sum_j n_{ij}\tilde{h}_{ij}/2n..$$

with

$$\tilde{p}_{i.} = \sum_j n_{ij}\tilde{p}_{ij}/n_{i.},$$
$$\tilde{p}.. = \sum_i\sum_j n_{ij}\tilde{p}_{ij}/n..$$

Notice that $\tilde{p}_{ij}$ and $\tilde{h}_{ij}$ are the observed allelic and heterozygote frequencies for one particular allele in the jth deme of the ith population. The formulae can be added over alleles and loci as before, providing the sum of the $\tilde{h}_{ij}$'s over alleles is recognized to give twice the total frequency of heterozygotes in that deme.

Variances of Estimates

The usefulness of estimates such as those just described is greatly increased if estimates of their variances are also available. Nei and Chakravarti (1977) use Taylor series expansions to provide approximate analytical expressions for the variances of their estimators, but we advocate the calculation of numerical estimates of variance using the jackknife procedure (Miller, 1974; Efron, 1982). This procedure has the advantage, for computer based analyses, of using the same equations as are used for calculating the estimators.

If we were estimating some parameter ϕ from a sample of n observations, the jackknife procedure for estimating the variance of the estimator $\hat{\phi}$ consists of omitting each of the n observations in turn and using the variation among the resulting n estimates. If $\hat{\phi}_{(i)}$ is the estimate resulting when observation i is omitted, then the jackknife variance of $\hat{\phi}$ is

$$\text{var}(\hat{\phi}) \triangleq \frac{n-1}{n} \cdot \sum_{i=1}^{n} \left(\hat{\phi}_{(i)} - \frac{1}{n} \sum_{i=1}^{n} \hat{\phi}_{(i)} \right)^2$$

and a less biased estimator of ϕ is

$$\hat{\phi}^* = n\hat{\phi} - \frac{n-1}{n} \sum_{i=1}^{n} \hat{\phi}_{(i)}.$$

This jackknife procedure has been found to be suitable for ratio estimators.

There may be several ways of proceeding in the present case, but we suggest jackknifing over loci by omitting one locus at a time. The estimate of θ, for example, obtained by omitting locus L is

$$\hat{\theta}_{(L)} = \sum_{l \neq L} \sum_{u} a_{lu} / \sum_{l \neq L} \sum_{u} (a_{lu} + b_{lu} + c_{lu}),$$

so that the jackknife variance of $\hat{\theta}$ in the m locus case is

$$\text{var}(\hat{\theta}) \triangleq \frac{m-1}{m} \sum_{L=1}^{m} \left(\hat{\theta}_{(L)} - \frac{1}{m} \sum_{L=1}^{m} \hat{\theta}_{(L)} \right)^2.$$

This procedure was found to work satisfactorily in the two-population situation (Reynolds et al., 1983) and also for simulations involving many loci. We are making use of the fact that loci are expected to be acting as (nearly) independent replicates with very little dependence being introduced by linkage, especially for a large number of loci. For the case of a single locus, it may be satisfactory to jackknife over populations. We find little difference between the original estimate $\hat{\theta}$ and the jackknife estimate $\hat{\theta}^*$.

Discussion

The comparison of measures of population structure from different sources is more valid when the same method of estimation is used in each. To accommodate the possibility of quite different sample sizes in terms of populations observed, individuals sampled, loci scored and alleles observed, we advocate the use of the general formulae given in (1) and (10). While they do not appear as simple as those generally used, their evaluation is not a difficult matter on a computer.

The formulae make explicit use of sample sizes and observed allelic and heterozygote frequencies. We have refrained from expressing our development in terms of "expected heterozygosities." It seems unwarranted to use such terminology before indications of random mating departures, such as $F \neq \theta$, are obtained.

We feel that the use of a standard set of formulae, clearly derived as estimators for well defined parameters, will prevent any confusion of statistical concepts and terminology. Wright's original development was clearly intended to apply to parameters, so that symbols such as σ^2 were appropriate for variances, and there was no need to introduce sample consid-

erations. It leads only to confusion when, in an estimation procedure, sample mean notation ($\bar{p}$) is used but sample variance notation (s^2) is not. It is also confusing when F_{ST} is introduced verbally as a parameter but defined symbolically as a data function, as in equation (8). This equation, now shown to rest upon a series of assumptions about the magnitude of the data set, has had the effect of diverting attention from the meaning of F_{ST}. The parameter θ is defined as a correlation of gene frequencies, but we see F_{ST} referred to as a "standardized variance of gene frequencies" or even as an "effective inbreeding coefficient."

Finally, we wish to comment on the phenomenon of sampling. The random processes involved in gamete formation give rise to variation that is largely beyond our control to correct for with increased sample sizes. Even if we were to census the entire population, our results are still affected by "sampling" in that the particular population sampled is but one of the many possible replicates that could have arisen under the same conditions.

Summary

Formulae are given for estimators for the parameters F, θ, f (F_{IT}, F_{ST}, F_{IS}) of population structure. As with all such estimators, ratios are used so that their properties are not known exactly, but they have been found to perform satisfactorily in simulations. Unlike the estimators in general use, the formulae do not make assumptions concerning numbers of populations, sample sizes, or heterozygote frequencies. As such, they are suited to small data sets and will aid the comparisons of results of different investigators. A simple weighting procedure is suggested for combining information over alleles and loci, and sample variances may be estimated by a jackknife procedure.

Acknowledgments

The revision of this paper was aided by comments from J. Felsenstein and P. Smouse, and from discussions with A. Robertson and W. G. Hill. Appreciation is extended to the Department of Genetics, University of Edinburgh for hospitality and computing facilities while B. S. Weir held a Guggenheim Fellowship in that department.

Literature Cited

Avise, J. C., and J. Felley. 1979. Population structure of freshwater fishes. I. Genetic variation of bluegill (*Lepomis macrochirus*) populations in man-made reservoirs. Evolution 33: 15–26.

Baker, A. E. M. 1981. Gene flow in house mice: introduction of a new allele into free-living populations. Evolution 35:243–258.

Baker, M. C., D. B. Thompson, G. L. Sherman, M. A. Cunningham, and D. F. Tomback. 1982. Allozyme frequencies in a linear series of song dialect populations. Evolution 36:1020–1029.

Balakrishnan, V., and L. D. Sanghvi. 1968. Distance between populations on the basis of attribute data. Biometrics 24:859–865.

Brittnacher, J. G., S. R. Sims, and F. J. Ayala. 1978. Genetic differentiation between species of the genus *Speyeria* (Lepidoptera: Nymphalidae). Evolution 32:199–210.

Chesser, R. K. 1983. Genetic variability within and among populations of the black-tailed prairie dog. Evolution 37:320–331.

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

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

Cockerham, C. C., and B. S. Weir. 1983. Variance of actual inbreeding. Theoret. Pop. Biol. 23:85–109.

Eanes, W. F., and R. K. Koehn. 1978. An analysis of genetic structure in the monarch butterfly, *Danaus plexippus* L. Evolution 32:784–797.

Efron, B. 1982. The Jackknife, the Bootstrap and Other Resampling Plans. Society for Industrial and Applied Mathematics, Philadelphia.

Ellstrand, N. C., and D. A. Levin. 1980. Recombination system and population structure in *Oenothera*. Evolution 34:923–933.

Ferrari, J. A., and C. E. Taylor. 1981. Hierarchical patterns of chromosome variation in *Drosophila subobscura*. Evolution 35:391–394.

Fleischer, R. 1983. A comparison of theoretical and electrophoretic assessment of genetic structure in populations of the house sparrow (*Passer domesticus*). Evolution 37:1001–1089.

Foltz, D. W., and J. L. Hoogland. 1983. Genetic evidence of outbreeding in the black-tailed prairie dog (*Cynomys ludovicianus*). Evolution 37:273–281.

Guries, R. P., and F. T. Ledig. 1982. Genetic diversity and population structure in pitch pine (*Pinus rigida* Mill.). Evolution 36:387–402.

Hiebert, R. D., and J. L. Hamrick. 1983. Pat-

terns and levels of genetic variation in great basin bristlecone pine, *Pinus longaeva.* Evolution 37:302–310.

KENDALL, M. G., AND A. STUART. 1969. The Advanced Theory of Statistics, Vol. 1 (3rd ed.). Griffin, London.

KIRBY, G. C. 1975. Heterozygote frequencies in small populations. Theoret. Pop. Biol. 8:31–48.

LEVIN, D. A. 1977. The organization of genetic variability in *Phlox Drummondii.* Evolution 31: 477–494.

———. 1978. Genetic variation in annual phlox: self-compatible versus self-incompatible species. Evolution 32:245–263.

LEVIN, D. A., K. RITTER, AND N. C. ELLSTRAND. 1979. Protein polymorphism in the narrow endemic *Oenothera organensis.* Evolution 33:534–542.

MILLER, R. G. 1974. The jackknife—a review. Biometrika 61:1–15.

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

———. 1976. Mathematical models of speciation and genetic distance, p. 723–765. *In* S. Karlin and E. Nevo (eds.), Population Genetics and Ecology. Academic Press, N.Y.

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

NEI, M., AND A. CHAKRAVARTI. 1977. Drift variances of F_{ST} and G_{ST} statistics obtained from a finite number of isolated populations. Theoret. Pop. Biol. 11:307–325.

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

NEI, M., AND Y. IMAIZUMI. 1966. Genetic structure of human populations. II. Differentiation of blood group gene frequencies among isolated populations. Heredity 21:183–190.

PAMILO, P. 1983. Genetic differentiation within subdivided populations of *Formica* ants. Evolution 37:1010–1022.

PATTON, J. L., AND J. H. FEDER. 1981. Microspatial genetic variation in pocket gophers: nonrandom breeding and drift. Evolution 35:912–920.

PATTON, J. L., AND S. Y. YANG. 1977. Genetic variation in *Thomomys bottae* pocket gophers: macrogeographic patterns. Evolution 31:697–720.

REYNOLDS, J., B. S. WEIR, AND C. C. COCKERHAM. 1983. Estimation of the coancestry coefficient: basis for a short-term genetic distance. Genetics 105:767–779.

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.

RYMAN, N., C. REUTERWALL, K. NYGREN, AND T. NYGREN. 1980. Genetic variation and differentiation in Scandinavian moose (*Alces alces*): are large mammals monomorphic? Evolution 34:1037–1049.

SCHAAL, B. A., AND W. G. SMITH. 1980. The apportionment of genetic variation within and among populations of *Desmodium nudiflorum.* Evolution 34:214–221.

SCHOEN, D. J. 1982. Genetic variation and the breeding system of *Gilia archilleifolia.* Evolution 36:361–370.

SCHWAEGERLE, K. E., AND B. A. SCHAAL. 1979. Genetic variability and founder effect in the pitcher plant *Sarracenia purpurea* L. Evolution 33:1210–1218.

SITES, J. W., AND I. F. GREENBAUM. 1983. Chromosome evolution in the iguanid lizard *Sceloporus grammicus.* II. Allozyme variation. Evolution 37:54–65.

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

WARD, P. S. 1980. Genetic variation and population differentiation in the *Rhytidoponera impressa* group, a species complex of Ponerine ants (Hymenoptera: Formicidae). Evolution 34: 1060–1076.

WINANS, G. A. 1980. Geographic variation in the milkfish *Chanos chanos.* I. Biochemical evidence. Evolution 34:558–574.

WORKMAN, P. L., AND J. D. NISWANDER. 1970. Population studies on southwestern Indian tribes. II. Local genetic differentiation in the Papago. Amer. J. Hum. Genet. 22:24–29.

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

———. 1951. The genetical 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.

———. 1973. The origin of the F-statistics for describing the genetic aspects of population structure, p. 3–26. *In* N. E. Morton (ed.), Genetic Structure of Populations. Univ. Press Hawaii, Honolulu.

———. 1978. Evolution and the Genetics of Populations, Vol. 4. Variability Within and Among Natural Populations. Univ. Chicago Press, Chicago.

Corresponding Editor: J. Felsenstein

APPENDIX

Combining Estimates over Alleles

The optimal weighting of single allele estimates over the alleles at one locus requires knowledge of the variances and covariances of these estimates. Contrary to the thrust of this paper, however, determination of expressions for such moments seems to be possible only via Taylor series expressions of ratios, with the implied assumption of large sample sizes. The following treatment is for illustration only then, while the simulations reported in the paper

$$Es_u^2 = p_u(1 - p_u)\theta$$
$$E\bar{p}_u(1 - \bar{p}_u) = p_u(1 - p_u)(1 - \theta/r)$$
$$\text{var}(s_u^2) = [p_u(1 - p_u)\delta + p_u^2(1 - p_u)^2(3\Delta - 6\delta - \theta^2)]/r$$
$$\text{var}[\bar{p}_u(1 - \bar{p}_u)] = p_u(1 - p_u)(1 - 2p_u)^2\theta/r$$
$$\text{cov}(s_u^2, s_{u'}^2) = [p_u p_{u'}\delta + p_u p_{u'}(1 - p_u - p_{u'})(\Delta - 2\delta - \theta^2) + p_u^2 p_{u'}^2(3\Delta - 6\delta - \theta^2)]/r$$
$$\text{cov}[\bar{p}_u(1 - \bar{p}_u), \bar{p}_{u'}(1 - \bar{p}_{u'})] = -p_u p_{u'}(1 - 2p_u)(1 - 2p_{u'})\theta/r$$
$$\text{cov}[s_u^2, \bar{p}_u(1 - \bar{p}_u)] = p_u(1 - p_u)(1 - 2p_u)^2\gamma/r$$
$$\text{cov}[s_u^2, \bar{p}_{u'}(1 - \bar{p}_{u'})] = -p_u p_{u'}(1 - 2p_u)(1 - 2p_{u'})\gamma/r$$

take proper account of small and unequal sized samples from small numbers of populations.

Suppose that the n_i and r are large enough that we may take, for allele u at some locus,

$$\hat{\theta}_u = \frac{s_u^2}{\bar{p}_u(1 - \bar{p}_u)}.$$

The variances and covariances of the $\hat{\theta}_u$'s, taking account of between and within population variation, can be expressed in terms of descent measures (e.g., Cockerham and Weir, 1983). For a finite monoecious population mating at random, the probabilities that any two, three, four or two pairs of genes are identical by descent are written as θ, γ, δ or Δ, respectively. Ignoring terms of order r^{-2} or n^{-1}, we can derive the expressions at the top of this page where different alleles are indicated by $u \neq u'$.

From Taylor series expressions (e.g., Kendall and Stuart, 1969 p. 232) then,

$$\text{var}(\hat{\theta}_u) = \frac{1}{r}\left[\frac{X}{p_u(1 - p_u)} + Y\right]$$
$$\text{cov}(\hat{\theta}_u, \hat{\theta}_{u'}) = \frac{1}{r}\left[\frac{X}{(1 - p_u)(1 - p_{u'})} + Y + \frac{Z(1 - p_u - p_{u'})}{(1 - p_u)(1 - p_{u'})}\right]$$

where

$$X = \delta - 2\theta\gamma + \theta^3,$$
$$Y = 3\Delta - 6\delta + 8\theta\gamma - \theta^2 - 4\theta^3,$$
$$Z = -2(\Delta - 2\delta + 2\theta\gamma - \theta^3).$$

We wish to construct a combined estimator of the form

$$\hat{\theta} = \sum_u w_u\hat{\theta}_u$$

with $\hat{\theta}$ unbiased and of minimum variance. This requires us to seek weights w_u and Lagrangian multiplier Λ to minimize the function

$$\sum_u w_u^2\text{var}(\hat{\theta}_u) + \sum\sum_{u \neq u'} w_u w_{u'}\text{cov}(\hat{\theta}_u, \hat{\theta}_{u'}) + \Lambda\left(\sum_u w_u - 1\right),$$

and we find that

$$w_u \propto \frac{p_u(1 - p_u)}{X - Zp_u}.$$

In general, there does not appear to be any simple function of the observations that has expectation w_u and so can be used as a weight for the optimal combination of the $\hat{\theta}_u$'s. There are some special cases that are tractable however, apart from the trivial case of two alleles per locus where $\hat{\theta}_1 \equiv \hat{\theta}_2$ and all weightings are equivalent.

For equally frequent alleles, each $\hat{\theta}_u$ should be given equal weight, and we have the simple average

$$\hat{\theta}_U = \frac{1}{v}\sum_{u=1}^{v}\frac{s_u^2}{\bar{p}_u(1 - \bar{p}_u)}.$$

for a locus with v equally frequent expected allelic frequencies.

For populations that have recently descended from a non-inbred ancestral population, θ is small and good approximations are given by

$$\gamma \approx \theta^2, \quad \delta \approx \theta^3, \quad \Delta \approx \theta^2,$$

so that $X \approx 0$ and

$$w_u \propto (1 - p_u) \quad \text{or} \quad w_u = \frac{1 - p_u}{v - 1}.$$

This higher weighting for less frequent alleles was noticed by Robertson and Hill (1984) and the above derivation of w_u follows their suggestion. Using sample values for the weights then,

$$\hat{\theta}_{RH} = \frac{1}{v - 1}\sum_{u=1}^{v}\frac{s_u^2}{\bar{p}_u}.$$

If, on the other hand, inbreeding levels are high (time since the ancestral population of the order of population size), good approximations were found by Cockerham and Weir (1983) to be

$$\gamma \approx (3\theta - 1)/2, \quad \delta \approx (8\theta - 3)/5, \Delta \approx (9\theta - 4)/5,$$

so that $Z \approx 0$ and

$$w_u \propto p_u(1 - p_u).$$

Now the denominator of $\hat{\theta}_u$ has this expectation, so we are led to

$$\hat{\theta}_W = \frac{\sum_{u=1}^{v} s_u^2}{\sum_{u=1}^{v} \bar{p}_u(1 - \bar{p}_u)}.$$

This procedure was also given by Nei and Chakravarti (1977). Note that, for small r, the denominator of $\hat{\theta}_u$ is increased by s_u^2/r, but this does not affect the variances given above.

Our suggested weighting scheme is optimal then for large θ, and the simulations show that it performs well for other θ values. That $\hat{\theta}_W$ is less biased than the other estimators can be confirmed by the following expressions for bias, also derived from Taylor series expansions (Kendall and Stuart, 1969 p. 44) that ignore terms of order r^{-2} and n^{-1}.

$$B_U = E\hat{\theta}_U - \theta = \frac{\theta^2}{r}\left[1 + \frac{1}{v}\sum_u \frac{(1 - 2p_u)^2}{p_u(1 - p_u)}\right]$$

$$B_{RH} = E\hat{\theta}_{RH} - \theta = \frac{\theta^2}{r}\frac{1}{v - 1}\sum_u \left[\frac{(1 - p_u)^2}{p_u}\right]$$

$$B_W = E\hat{\theta}_W - \theta = \frac{\theta^2}{r}\left[1 + \frac{\sum_u p_u(1 - p_u)(1 - 2p_u)^2}{\left[\sum_u p_u(1 - p_u)\right]^2}\right].$$

It can be shown that

$$B_{RH} \geq B_U \geq B_W$$

and the simulations suggest that the lower bias of $\hat{\theta}_W$ is sufficient to give it a lower mean square error.

Another approach has been suggested by Smouse (e.g., Smouse and Williams, 1982). Instead of performing an analysis of variance for the frequencies of each allele separately, he suggests a multivariate analysis of variance for all alleles together. In the random union of gametes situation, the components a_u, d_u are replaced by matrices **A, D** with diagonal elements $a_{uu} = a_u$, $d_{uu} = d_u$ and off-diagonal elements $a_{uu'}$, $d_{uu'}$ obtained from these by changing $p_u(1 - p_u)$ to $-p_u p_{u'}$, s_u^2 to $s_{uu'}$ and $\bar{h}_u$ to $\bar{h}_{uu'}$. Here $s_{uu'}$ is the covariance over populations and $\bar{h}_{uu'}$ the average frequency of heterozygotes over populations for alleles u and u'. If **P** is the matrix with diagonal elements $P_{uu} = p_u(1 - p_u)$ and off-diagonal elements $P_{uu'} = -p_u p_{u'}$, then

$$E\mathbf{A} = \theta\mathbf{P}, \quad E\mathbf{D} = \mathbf{P}.$$

Since the determinants $|\mathbf{A}|$, $|\mathbf{D}|$ are zero, we omit the row and column for any one allele, k say (indicated by the subscript $[k]$), and estimate θ as the average of the diagonal elements of a matrix product

$$\hat{\theta}_M = \frac{1}{v - 1}\,\text{trace}\,[\mathbf{D}_{(k)}^{-1}\mathbf{A}_{(k)}].$$

For large n and r

$$a_{uu} = s_u^2, \qquad a_{uu'} = s_{uu'},$$
$$d_{uu} = \bar{p}_u(1 - \bar{p}_u), \quad d_{uu'} = -\bar{p}_u\bar{p}_{u'},$$

and we find that $\hat{\theta}_M \equiv \hat{\theta}_{RH}$.

The formulae given for optimal weights assume knowledge of the parameters p_u, and optimality is not guaranteed when observed values $\bar{p}_u$ are used. This is illustrated by $\hat{\theta}_M$ having a smaller variance than $\hat{\theta}_{RH}$ in our simulation study, even when the true θ is small.

Combining Estimates over Loci

A similar treatment for combining estimates over loci does not appear to be possible. For alleles u, u' at loci l, l', we can show that

$$\text{cov}(\hat{\theta}_{lu}, \hat{\theta}_{l'u'}) = \frac{1}{r}\frac{\Delta_{ll'}^* - \theta^2}{(1 - p_{lu})(1 - p_{l'u'})}$$

where the two-locus descent measure $\Delta_{ll'}^*$ is the probability that a pair of genes at locus l and a pair at locus l', the four genes being on separate gametes, are identical by descent. This is for zero initial linkage disequilibrium and so precludes the recovery of the variance of $\hat{\theta}_{lu}$ by setting recombination to zero.

The simulations support our procedure of summing numerators and denominators of the $\hat{\theta}_{lu}$ separately, while a natural weighting scheme for the procedure of Robertson and Hill is to use $(v_l - 1)$, leading to

$$\hat{\theta} = \sum_l \sum_u \frac{s_{lu}^2}{\bar{p}_{lu}} \Big/ \sum_l (v_l - 1).$$

This is analogous to the distance formula of Balakrishnan and Sanghvi (1968).