Proc. Natl. Acad. Sci. USA
Vol. 84, pp. 8512–8514, December 1987
Evolution

Correlations, descent measures: Drift with migration and mutation

(*F* statistics/equilibrium/island model/genic variance/population structure)

C. Clark Cockerham and B. S. Weir

Department of Statistics, North Carolina State University, Box 8203, Raleigh, NC 27695-8203

Contributed by C. Clark Cockerham, August 14, 1987

ABSTRACT **The analysis of gene frequencies for a nested structure of genes within individuals, individuals within subpopulations, and subpopulations within populations is considered. Alternative parameterizations are provided by measures of correlation and of identity by descent, but the latter parameters provide more flexibility. The effects of population size, mating system, mutation, and migration can be incorporated into transition equations for identity measures and the structure of equilibrium populations can be determined; the procedures are illustrated for a finite island model. With parameters defined before estimation procedures are developed, problems of estimates depending on the numbers of sampled subpopulations are avoided, while the descent measures also avoid the approximations found in other treatments.**

In the analysis of gene frequencies in natural populations, it is important to have a parametric model elucidating the kinds of variation to be encountered. In this way, various assumptions about the model are clarified and estimation can be guided by the model. Without any information, just correlations and variances often must suffice. Even so, these should also be accurately parameterized as a guide to the appropriate analysis (1, 2).

In some cases, the correlations bear the interpretation of identity by descent parameters. These parameters are very useful in studies of the consequences of mating system, finite population size, migration, and even mutation in certain circumstances.

The purpose of this note is to relate correlation and identity by descent parameters and to provide an illustration of the versatility of identity by descent parameters for a finite island model at equilibrium with respect to migration and mutation.

Genic Structure

Since individual genes are identified, the genic hierarchical structure to be considered is genes within individuals, individuals within subpopulations, and subpopulations within populations. This means that distinct pairs of genes fall into the following categories: genes within individuals, genes in different individuals in the same subpopulation, genes in different subpopulations in the same population, and genes in different independent replicate populations.

Variance and Correlation Parameters

This development is the same as that of Cockerham (1, 2). Essential details will be reviewed for completeness with some extensions. We utilize a measure $x_{il} = 1$ if the gene is A_l, the lth allele, where i identifies the location of the gene in the hierarchy, and $x_{il} = 0$ if the gene is another allele, A_k, $k \neq l$. Then, for a random gene $\mathscr{E}x_{il} = p_l$, where $\mathscr{E}$ denotes expectation and p_l is the parametric gene frequency. The variance among random independent genes is $\mathscr{E}x_{il}^2 - (\mathscr{E}x_{il})^2 = p_l(1 - p_l)$.

We next consider the expectations for random pairs of genes in the different categories, $\mathscr{E}x_{il}x_{i'l} = (\mathscr{E}x_{il})(\mathscr{E}x_{i'l}) + \mathscr{C}x_{il}x_{i'l} = p_l^2 + \theta_{ii'}\mathrm{p}_l(1 - p_l)$, where $\mathscr{C}$ denotes covariance and θ denotes correlation since x_{il} and $x_{i'l}$ have the same expectation and variance. We let $\theta_k = \theta_{ii'}$, where i and i' are k steps apart in the hierarchy. For $k = 0$, $\theta_0 = 1$ is the correlation of genes with themselves. The other correlations are θ_1 for genes within individuals, θ_2 for genes in different individuals in the same subpopulation, θ_3 for genes in different subpopulations in the same population, and $\theta_4 = 0$ for genes in different independent populations.

If one assumes that the correlations are the same for all alleles, then the expectations can be summed

$$Q_k = \sum_l \mathscr{E}x_{il}x_{i'l} = q + \theta_k(1 - q) = \theta_k + (1 - \theta_k)q,$$

where $q = \Sigma_l p_l^2$. If the correlations differ, one can replace θ_k with $\bar{\theta}_k = \Sigma_l\, \theta_{kl}p_l(1 - p_l)/(1 - q)$ for a more complex situation, or else treat each allele separately. In practice, and in the context of identity by descent parameters to be considered later, Q_k can be interpreted as the frequency with which genes are alike.

The case was developed by Cockerham (1, 2) that linear functions of the Qs could be interpreted as components of variance appropriate for the analysis of variance model. These components are

$\sigma_0^2 = 1 - Q_1 = (1 - \theta_1)(1 - q)$ within individuals,

$\sigma_1^2 = Q_1 - Q_2 = (\theta_1 - \theta_2)(1 - q)$ individuals within subpopulations,

$\sigma_2^2 = Q_2 - Q_3 = (\theta_2 - \theta_3)(1 - q)$ subpopulations within populations, and

$\sigma_3^2 = Q_3 - Q_4 = \theta_3(1 - q)$ among populations,

which sum to the total variance, $\sigma^2 = 1 - Q_4 = 1 - q$. This of course extends the procedures for estimation to include small sample analysis of variance techniques.

Frequency and Identity by Descent Parameters

When certain forces are acting on populations, identity by descent parameters can be utilized to provide a more informative genetic interpretation than do correlations. Mating system, finite population size, and migration are forces that operate on all genes equally, and transitional and equilibrium results can be formulated in terms of descent measures. One can further include some mutation models (3).

The interpretation of Q_k is straightforward in this case. With probability θ_k, genes are identical by descent and are alike. They are not identical by descent with probability 1 −

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

θ_k but are identical in state with frequency $\Sigma_l p_l^2 = q$. Thus,

$$Q_k = \theta_k + (1 - \theta_k)q = q + \theta_k(1 - q)$$

has the same form as for the correlation model.

As a descriptor for analysis at any time, the two parameterizations are the same, but it is the descent measures that allow us to elaborate the effects of mating system, population size, migration, and mutation.

Variations in Population Structure

Considerable explanation was given (2) about adjusting the parametric model to the situation at hand, limitations in estimable functions, ignoring an existing hierarchy, and so on. A few examples will suffice. If we consider only independent subpopulations with no hierarchy of populations, then $\theta_3 = 0$ and $\theta_1 = F_{IT}$ and $\theta_2 = F_{ST}$, where the Fs are Wright's (4) F statistics. With the addition of the hierarchy of populations, the list of F statistics has to be extended to include θ_3.

With only a single subpopulation, the only estimable components of variance are $1 - Q_1 = (1 - \theta_1)(1 - q)$ and $Q_1 - Q_2 = (\theta_1 - \theta_2)(1 - q)$ with a total of $1 - Q_2 = (1 - \theta_2)(1 - q)$ and the only estimable correlation is $F_{IS} = (\theta_1 - \theta_2)/(1 - \theta_2)$ for genes within individuals within subpopulations, as is well known [Wright's F_{IS} (4)].

With monoecious populations and random union of gametes, genes within individuals and between individuals in the same subpopulation have the same correlation—i.e., $\theta_1 = \theta_2$ and $Q_1 = Q_2$. In addition, suppose we restrict our consideration to a single population of subpopulations. We still have to consider a parametric framework of independent populations. Otherwise, we have no basis for taking expectations. Since $\theta_1 = \theta_2$, we have only two components of variance: $1 - Q_2 = (1 - \theta_2)(1 - q)$ and $Q_2 - Q_3 = (\theta_2 - \theta_3)(1 - q)$ with a total of $1 - Q_3 = (1 - \theta_3)(1 - q)$. The only estimable correlation is $\beta = (\theta_2 - \theta_3)/(1 - \theta_3) = (Q_2 - Q_3)/(1 - Q_3)$, which is the correlation of genes in subpopulations within populations and is similar in concept to F_{IS}. It is this situation that will be considered in the following example.

Finite Island Model with Migration and Mutation

Crow and Aoki (5) considered a finite island model with mutation and migration, and they worked with equilibrium values for the Qs. The procedure is complex and requires approximations during the development.

Recently, in a study of quantitative genetic variation within and between finite populations for an additive genetic model with mutation and migration, it was found that the equilibrium values could be written as simple functions of an equilibrium descent measure and the variance in an infinite equilibrium population (3). The transitional value of the descent measure was also useful in expressing the transitional values of the variance within populations. The situation is even simpler for the genic model x_l, which of course is entirely additive.

We consider here only equilibrium conditions. For the mutation model, a random gene mutates at rate v_l to the allele A_l, including no change in state of the gene. Let the total mutation rate be $u = \Sigma_l v_l$, $l = 1, 2, \ldots, k$ for k alleles. At equilibrium, regardless of the founder population, $p_l = v_l/u$ and $q = \Sigma_l p_l^2 = (1 + c^2)/k$, where c is the coefficient of variation of the equilibrium frequencies (6).

We now proceed to formulate the descent measures θ_2 and θ_3 with mutation and migration. We consider n subpopulations, each with N individuals in each generation. Migration is gametic at rate m at the time of reproduction and the migrant gamete has an equal chance of coming from each of the other $n - 1$ subpopulations.

A notation by Nagylaki (7) and Crow and Aoki (5) is useful. After migration, the frequency of pairs of genes from the same subpopulation in the previous generation is $a = (1 - m)^2 + m^2/(n - 1)$ for genes in one subpopulation and $b = (1 - a)/(n - 1)$ for genes between subpopulations. When genes are from the same subpopulation, they are identical by descent with probability $1/2N + (1 - 1/2N)\theta_2$, and when from different subpopulations they are identical by descent with probability θ_3 provided neither gene has mutated with probability $\rho = (1 - u)^2$. Letting $\gamma = 1 - 1/2N$ and $(a - b) = (1 - m\alpha)^2 = d$, where $\alpha = n/(n - 1)$

$$\begin{aligned}
\theta_{2,t+1} &= \rho[a(1 - \gamma + \gamma\theta_{2,t}) + (1 - a)\theta_{3,t}] \\
\theta_{3,t+1} &= \rho[b(1 - \gamma + \gamma\theta_{2,t}) + (1 - b)\theta_{3,t}] \\
\theta_{2,t+1} - \theta_{3,t+1} &= \rho d[\gamma(\theta_{2,t} - \theta_{3,t}) + (1 - \gamma)(1 - \theta_{3,t})].
\end{aligned}$$

Note that identity by descent requires that neither gene has mutated. At equilibrium, the θs do not change. To solve for β directly, we find $\theta_2 - \theta_3 = (1 - \theta_3)\rho d(1 - \gamma)/(1 - \rho d\gamma)$. Consequently,

$$\begin{aligned}
\beta &= (\theta_2 - \theta_3)/(1 - \theta_3) = \rho d/[2N(1 - \rho d) + \rho d] \\
&\simeq 1/(1 + 4Nu + 4Nm\alpha) = \tilde{\beta}.
\end{aligned}$$

The latter approximation is for small u and m. If n is infinite, $\theta_3 = 0$ and $\beta = \theta_2 \simeq 1/(1 + 4Nu + 4Nm)$, as found by Cockerham and Tachida (3). Equilibrium values for θ_2 and θ_3 are $\theta_2 = \rho(a - \rho d)/2NW$ and $\theta_3 = \rho b/2NW$, where $W = 1 - \rho + \rho b - \rho a\gamma + \rho^2\gamma d$, which are solutions obtained by Nagylaki (7).

Crow and Aoki (5) utilize Nei's (8) $G_{ST} = (Q_2 - \bar{Q})/(1 - \bar{Q})$, where $\bar{Q} = [Q_2 + (n - 1)Q_3]/n$. There is, of course, a direct relationship between G_{ST} and β, based on the parametric relationship $Q_k = \theta_k + (1 - \theta_k)q$,

$$G_{ST} = \frac{(n - 1)\beta}{n - \beta}, \qquad \beta = \frac{nG_{ST}}{G_{ST} + n - 1}.$$

If we substitute $\tilde{\beta}$ into G_{ST}, we obtain $\tilde{G}_{ST} = 1/(1 + 4Nu\alpha + 4Nm\alpha^2)$, which agrees with their result except for the term $4Nu\alpha$. They assumed $u << m$, so that $4Nu\alpha$ is negligible compared to $4Nm\alpha^2$. Takahata and Nei (9) review several approximate formulas for G_{ST}.

Discussion

Since there is not a great deal of difference between β and G_{ST}, then why bother to distinguish between them? Actually, n is a parameter in the model and is generally unknown. With a sample of r subpopulations, $r \geq 2$, an unbiased estimate of β can be obtained in the sense that the numerator and denominator of the estimator are unbiased. What does one do for G_{ST}? Weight the estimates of Qs with r in the same manner as n is used parametrically? If so, the estimate of G_{ST} is directly affected by the number of subpopulations sampled, a very undesirable property, particularly for small r.

There is a more compelling reason for preferring β. Each step in the hierarchy represents a potential degree of differentiation. Each degree of differentiation should be utilized in arriving at the total differentiation instead of averaging over some steps. Also, β makes sense from the standpoint of partitioning variation and the role that intraclass correlations play in this partitioning. Considerable statistical literature is available on the theory and methodology for intraclass correlations. It is the correlations that measure the degrees of differentiation in the population.

A parametric model is essential for understanding the situation and as a guide to estimation. Crow and Aoki's (5) development of the parametric Qs is a valid approach and the only one available in certain circumstances. It just so happens in this situation that identity by descent coefficients, θ_2 and θ_3, simplify the development of the results considerably and they are the relevant correlations. However, in terms of a single population of subpopulations, they are not estimable and the parameters are reduced to $\beta = (\theta_2 - \theta_3)/(1 - \theta_2)$, the only correlation that is estimable. The more complete model clarifies the estimable function for the restricted model.

The identity by descent coefficents provide an additional advantage of accommodating a fairly general mutation model for any number of alleles with unequal mutation rates.

All estimates must be of the Qs or linear functions of them such as $1 - Q_2$ and $Q_2 - Q_3$. Small sample estimation procedures have been considered in some detail (1, 2, 10, 11). All procedures give the same results when the number of individuals in each subpopulation sampled is equal but different results when the numbers vary. It should be noted that the method of symmetrical products in ref. 12 when applied to the x_is defined earlier provides direct estimates of the Qs.

We have treated only a single locus. As mentioned previously, drift and migration affect all loci in the same manner. Different loci may well have different numbers of alleles, equilibrium frequencies, and overall mutation rates. Also, in practice other forces such as selection will lead to differences among loci. Crow and Aoki (5) point out that there are now many molecular variants that are believed to be neutral or nearly neutral and that these are appropriate candidates for the study of differentiation among subpopulations. Also, if mutation rates are much less than migration rates, the differences in β due to mutation will be minor. Unfortunately, there are no good tests of significance for heterogeneity of the βs. An average $\bar{\beta}$ can be obtained by summing the numerators and denominators of the individual βs,

$$\bar{\beta} = \frac{\sum_i (Q_{2i} - Q_{3i})}{\sum_i (1 - Q_{3i})} = \frac{\bar{\theta}_2 - \bar{\theta}_3}{1 - \bar{\theta}_3},$$

where $\bar{\theta}_k = \Sigma_i\, \theta_{ki}(1 - q_i)/\Sigma_i\,(1 - q_i)$. ($i$ now indexes loci.)

Crow and Aoki (5) suggested that the finite island model is not the most realistic and that migrants are more likely to come from nearby groups. They did numerical calculations for a stepping-stone model in the form of an abstract torus in a study of the effects of mutation rates, total population size, and shape of the habitat. G_{ST} was affected little by mutation rate, increased with n, and increased as the habitat became long and narrow. Their results for G_{ST} also apply to β because of the functional relationship between the two. Nagylaki (7) considered several migration models and concluded that gametic migration gave a good approximation to diploid migration, particularly with small mutation and migration rates and large colonies.

The primary emphasis in this note, however, has been on adopting a model appropriate for the situation at hand. Considerable theory in population genetics has been based on the infinite allele model. As in our example, the correlations are insensitive to the number of alleles, but with the infinite allele model $Q_k = \theta_k$, and the use of this assumption in the model for estimation can lead to considerable error. The variation available is dependent to a considerable extent on the measuring device whether it be gel or other electrophoresis, cutter locations for endonucleases, sequence data, or other. As pointed out in ref. 6, silent variation—i.e., variation not recognized by the measuring device—is appropriately ignored. Of course, the greater the number of alleles, the more information there is on the correlations and population differentiation, but in practice one must deal with a finite, often small, number of alleles.

We thank Drs. J. F. Crow and C. H. Langley for helpful comments. This is paper no. 11126 of the Journal Series of the North Carolina Agricultural Research Service (Raleigh, NC 27695-7601). This investigation was supported in part by Research Grant GM 11546 from the National Institute of General Medical Sciences.

1. Cockerham, C. C. (1969) *Evolution* **23,** 72–84.
2. Cockerham, C. C. (1973) *Genetics* **74,** 679–700.
3. Cockerham, C. C. & Tachida, H. (1987) *Proc. Natl. Acad. Sci. USA* **84,** 6205–6209.
4. Wright, S. (1951) *Ann. Eugen.* **15,** 323–354.
5. Crow, J. F. & Aoki, K. (1984) *Proc. Natl. Acad. Sci. USA* **81,** 6073–6077.
6. Cockerham, C. C. (1984) *Proc. Natl. Acad. Sci. USA* **81,** 530–534.
7. Nagylaki, T. (1983) *Theor. Pop. Biol.* **24,** 268–294.
8. Nei, M. (1973) *Proc. Natl. Acad. Sci. USA* **70,** 3321–3323.
9. Takahata, N. & Nei, M. (1984) *Genetics* **107,** 501–504.
10. Weir, B. S. & Cockerham, C. C. (1984) *Evolution* **38,** 1358–1370.
11. Cockerham, C. C. & Weir, B. S. (1986) *Ann. Hum. Genet.* **50,** 271–281.
12. Koch, G. G. (1967) *Technometrics* **9,** 93–118.