**F_{ST} in a Hierarchical Island Model**

Montgomery Slatkin and Lianne Voelm

*Department of Integrative Biology, University of California, Berkeley, California 94720*

Manuscript received August 6, 1990
Accepted for publication November 8, 1990

ABSTRACT

It is shown that in a hierarchical island model, in which demes within a neighborhood exchange migrants at a much higher rate than do demes in different neighborhoods, hierarchical \( F \) statistics introduced by S. Wright can indicate the extent of gene flow within and between neighborhoods. At equilibrium, the within-neighborhood inbreeding coefficient, \( F_{SN} \), is approximately \( 1/(1 + 4Nm) \) where \( N \) is the deme size and \( m \) is the migration rate among demes in the same neighborhood. The between-neighborhood inbreeding coefficient, \( F_{NT} \), is approximately \( 1/(1 + 4Nd m) \) where \( d \) is the number of demes in a neighborhood and \( m_2 \) is the migration rate among demes in different neighborhoods.

In this note we will describe the population genetic consequences of a species having a hierarchical population structure. The particular structure we will be concerned with is what we will call the hierarchical island model in which local populations are grouped into neighborhoods within which there is considerable gene flow and between which there is less gene flow. Carmelli and Cavalli-Sforza (1976) analyzed this model and called it the “multiuniform model,” and Sawyer and Felsenstein (1983) also examined this model. We will show that under biologically reasonable assumptions, hierarchical \( F \) statistics introduced by Wright (1951) can be used to estimate average levels of gene flow within and between neighborhoods.

We will consider a population of monocious, diploid individuals and assume that there are discrete nonoverlapping generations. We assume that the population is made up of \( n \) neighborhoods each of which contains \( d \) demes of effective size \( N \). We will consider only neutral alleles and, for simplicity, assume that dispersal occurs at the gamete stage. We assume that a randomly chosen gamete after migration has a probability \( 1 - m_1 - m_2 \) of being a nonimmigrant, a probability \( m_1 \) of being an immigrant from a different randomly chosen deme in the same neighborhood and a probability \( m_2 \) of being an immigrant from a randomly chosen deme in a different neighborhood. We will assume that both \( m_1 \) and \( m_2 \) are small.

Our concern is with the extent of inbreeding in different hierarchical levels in the population. Wright (1951) showed how inbreeding coefficients could be partitioned within and between levels. For a population in which individuals are grouped into distinct subpopulations, Wright defined \( F_{ST} \) to be the inbreeding coefficient of a subpopulation relative to the whole population. For populations that are themselves grouped, as in the hierarchical island model, Wright defined inbreeding coefficients for each level in the hierarchy. Here we will use \( F_{SN} \) for the inbreeding coefficient of a deme relative to its neighborhood and \( F_{NT} \) for the inbreeding coefficient of a neighborhood relative to the whole population.

To relate these hierarchical inbreeding coefficients to the parameters of the hierarchical island model, we will use NEI’s (1973) formulas for expressing inbreeding coefficients in terms of probabilities of identity by descent. With a single level of subdivision, NEI defined \( G_{ST} \) which is equivalent to \( F_{ST} \) to be

\[
F_{ST} = f_0 - \tilde{f} \over 1 - \tilde{f},
\]

where \( f_0 \) is the probability of identity by descent of two genes drawn at random from the same subpopulation and \( \tilde{f} \) is the average probability of identity by descent of two genes drawn at random from the whole population. NEI (1973) defined hierarchical coefficients similarly. In our notation, NEI’s definitions can be written

\[
F_{SN} = f_0 - \tilde{f}_N \over 1 - \tilde{f}_N,
\]

and

\[
F_{NT} = \tilde{f}_N - \tilde{f}_T \over 1 - \tilde{f}_T,
\]

where \( \tilde{f}_N \) is the average probability of identity by descent of two genes drawn at random from a single neighborhood and \( \tilde{f}_T \) is the average probability of identity by descent of two genes drawn at random from the entire population.

The problem is to find the values of the various \( \tilde{f} \)'s for the hierarchical island model. One way would be to derive the equations for \( f_0 \), \( \tilde{f}_N \) and \( \tilde{f}_T \) that are analogous to those of Maruyama (1970) for the island...
model. That approach leads to three linear equations whose equilibrium solutions are, in general, complicated functions of the parameters. Here, we will use a different method that depends on the relationship between inbreeding coefficients and average coalescence times of pairs of genes. This approach provides relatively simple results under the assumption that the mutation rate at the locus of interest is small.

To show the relationship between identity by descent and coalescence times, consider any pair of genes and let \( P(t) \) be the probability that those genes have their first common ancestor at generation \( t \) in the past. The probability that these two genes are identical by descent is the probability that neither gene has mutated between the present and the time of first common ancestry. If the mutation rate per generation is \( \mu \), then \( f \), the probability of identity by descent, is

\[
f = \sum_{i=1}^{\infty} (1 - \mu)^{2i} P(i)
\]

which is the net probability that two genes coalesced before either mutated (Hudson 1990; M. Slatkin, manuscript in preparation). If \( \mu \) is small, then (4) can be approximated by

\[
f \approx 1 - 2\mu \sum_{i=1}^{\infty} iP(i) = 1 - 2\mu \bar{t},
\]

where \( \bar{t} \) is the average coalescence time. The approximation in (5) is valid if \( \mu \bar{t} \ll 1 \).

We can use this result to find the approximate values of \( f_0, \bar{t}_n \) and \( \bar{t}_f \) in terms of the corresponding average coalescence times. We begin by noting that \( \bar{t}_f \), the average coalescence time of two genes sampled at random from the same deme, is \( 2Ndn \). That follows from the more general result that the average coalescence time of two genes drawn from the same local population in any connected subdivided population is equal to the total number of copies of the gene in the population (Strobeck 1987; Hey 1990).

We define \( \bar{t}_0 \) to be the average coalescence time of two genes drawn at random from within a neighborhood and note that

\[
\bar{t}_0 = \frac{1}{d} \bar{t}_f + \frac{d - 1}{d} \bar{t}_1
\]

where \( \bar{t}_1 \) is the average coalescence time of genes drawn from different demes in the same neighborhood. Similarly, we define \( \bar{t}_f \) to be the average coalescence time of two genes drawn at random from the entire population, so

\[
\bar{t}_f = \frac{1}{n} \bar{t}_N + \frac{n - 1}{n} \bar{t}_2
\]

where \( \bar{t}_2 \) is the mean coalescence time of two genes drawn from different neighborhoods.

It is relatively easy to find \( \bar{t}_1 \) and \( \bar{t}_2 \) by recognizing that the coalescence of two genes occurs in two phases.

If we think of the process as going backwards in time, there will be the time before the two genes are first present in the same deme and then the time after they are first present in the same deme but before they finally coalesce. We already know the average time until coalescence once they are in the same deme: \( \bar{t}_0 = 2Ndn \). Although it is true that after genes are first present in the same deme they may be present in different demes before they ultimately coalesce, that possibility is included in the value of \( \bar{t}_0 \). We need to find then only the average times until two genes are first present in the same deme given that they are sampled from either different demes in the same neighborhood or from different neighborhoods. We will denote these times by \( \bar{t}_1 \) and \( \bar{t}_2 \). Thus, \( \bar{t}_i = 2Ndn + \bar{t}_i \) (\( i = 1,2 \)).

To find the \( \bar{t}_i \), we formulate the process going backwards in time as a three state Markov chain: state 0, which for our purposes is the absorbing state, indicates that the ancestors of the two genes sampled are in the same deme; state 1 indicates that the ancestors of the two genes sampled are in different demes in the same neighborhood; and state 2 indicates that the ancestors of the two genes sampled are in different neighborhoods. The elements of the transition matrix \( P \) for this Markov chain are \( p_{ij} \), the probability of state \( j \) in generation \( t + 1 \) in the past given state \( i \) in generation \( t \) in the past, and are determined by working out which migration events result in a state change. The values of the \( p_{ij} \) that are relevant here are as follows:

\[
\begin{align*}
p_{11} &= 1 - \frac{2m_1}{d - 1} - \frac{2m_2}{d - 1} & p_{12} &= \frac{2m_2}{d - 1} \\
p_{21} &= \frac{2(d - 1)m_2}{d(n - 1)} & p_{22} &= 1 - \frac{2m_2}{d - 1}
\end{align*}
\]

The values in (8) assume that \( m_1 \) and \( m_2 \) are sufficiently small that at most one ancestral gene migrates per generation.

The general theory of Markov chains, conveniently summarized by Ewens (1979), tells us that the \( \bar{t}_i \) are the solutions to the matrix equation

\[
P \begin{pmatrix} \bar{t}_1 \\ \bar{t}_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}
\]

which implies

\[
\bar{t}_i = \frac{2m_1}{d - 1} + \frac{2m_2}{d - 1} \approx \frac{n(d - 1)}{2m_1}
\]

and

\[
\bar{t}_2 = \bar{t}_1 + \frac{2m_2}{(d - 1)(n - 1)} \approx \frac{n}{2m_2}
\]
The second approximate equalities in (10) and (11) are obtained by assuming $m_1 \gg m_2$, the case of interest here.

We can now return to equations (2) and (3) and use (5) to write

$$ F_{SN} \approx \frac{\hat{F}_{SN} - \hat{F}_0}{\hat{F}_{SN}} \quad (12) $$

and

$$ F_{NT} \approx \frac{\hat{F}_{NT} - \hat{F}_N}{\hat{F}_{NT}}. \quad (13) $$

These equations allow us to find the equilibrium values of $F_{SN}$ and $F_{NT}$ for small $\mu$. If $m_1 \gg m_2$,

$$ F_{SN} \approx \frac{1}{1 + 4N \frac{d^2}{(d - 1)^2} m_1} \quad (14) $$

and

$$ F_{NT} \approx \frac{1}{1 + 4Nd \frac{n^2}{(n - 1)^2} m_2}. \quad (15) $$

These results have a simple interpretation if we recall that the equilibrium value of $F_{ST}$ in the island model is

$$ F_{ST} \approx \frac{1}{1 + 4N \frac{d^2}{(d - 1)^2} m} \quad (16) $$

where $m$ is the migration rate and $d$ is the number of demes (Takahata 1983; Crow and Aoki 1984). In light of (16), equation (14) tells us that $F_{SN}$, the inbreeding coefficient of a deme relative to its neighborhood, is the value that would be expected if the neighborhood were completely isolated from the other neighborhoods. Equation (15) tells us that each neighborhood can, for our purposes, be regarded as a single deme containing $Nd$ individuals with immigration rate $m_2$.

Our results suggest that estimating hierarchical $F$ statistics from gene frequency data will permit inference of average levels of gene flow within and between levels in the hierarchy if the rate of gene flow within neighborhoods is much larger than rates among neighborhoods. Although we have done no simulations of the hierarchical island model, the simulation results of Slatkin and Barton (1989) suggest that the methods of Nei (1973) and Weir and Cockerman (1984) would perform equally well for moderate and low levels of gene flow. Hierarchical $F$ statistics could also be used to identify neighborhoods within which there are relatively high levels of gene flow.

As an example of the application of our results, we consider the data of Patton (1990) for pocket gophers, Thomomys bottae. Patton (1990, Figure 3b) summarized extensive electrophoretic data and estimated hierarchical $F$ values of demes relative to local populations ($F_{PP} \approx 0.08$), local populations relative to clusters of local populations ($F_{PR} \approx 0.3$), clusters relative to geographic segments ($F_{RS} \approx 0.3$) and finally segments relative to the total ($F_{ST} \approx 0.28$). Our results indicate that if this species is in genetic equilibrium, then there are relatively high levels of gene flow among demes within what Patton recognizes as local populations but extremely low levels of gene flow among other levels in the hierarchy.

This research has been supported in part by National Institutes of Health research grant GM40822. We thank J. Felsenstein for helpful discussions of this topic and two anonymous reviewers for useful comments on an earlier version of this paper.

**LITERATURE CITED**


Communicating editor: A. G. Clark