Genetic Differentiation and Estimation of Gene Flow from *F*-Statistics Under Isolation by Distance

François Rousset

Laboratoire Génétique et Environnement, Institut des Sciences de l'Évolution, Université de Montpellier II, 34095 Montpellier, France

Manuscript received September 17, 1996
Accepted for publication December 20, 1996

ABSTRACT

I reexamine the use of isolation by distance models as a basis for the estimation of demographic parameters from measures of population subdivision. To that aim, I first provide results for values of *F*-statistics in one-dimensional models and coalescence times in two-dimensional models, and make more precise earlier results for *F*-statistics in two-dimensional models and coalescence times in one-dimensional models. Based on these results, I propose a method of data analysis involving the regression of $F_{ST}/(1 - F_{ST})$ estimates for pairs of subpopulations on geographic distance for populations along linear habitats or logarithm of distance for populations in two-dimensional habitats. This regression provides in principle an estimate of the product of population density and second moment of parental axial distance. In two cases where comparison to direct estimates is possible, the method proposed here is more satisfactory than previous indirect methods.

ANALYSES of the structure of natural populations are often based on the island or stepping stone models. Functions of probabilities of identity of genes within and between units, such as F_{ST}, are estimated and compared to expectations under the island model. The relationship between F_{ST} and the number of migrants according to this model is often used to quantify gene flow. This relationship has been shown to approximate the relationship between F_{ST} and the number of migrants in stepping stone models on a two-dimensional space (KIMURA and MARUYAMA 1971; CROW and AOKI 1984; SLATKIN and BARTON 1989), and is used to obtain indirect estimates of number of migrants or "neighborhood size" from genetic data. These models can also be used to study the distribution of coalescence times (SLATKIN 1991, 1993), and therefore to obtain the value of measures of population subdivision that have been proposed for analyzing differences in allele size distributions at microsatellite loci, number of nucleotide differences, and maybe quantitative traits (HUDSON 1990; LANDE 1992; SLATKIN 1995). Analytical approximations have been obtained for coalescence times in the one-dimensional stepping stone model (SLATKIN 1991).

Here, I present a new method of analysis that may be deduced from isolation by distance models. The estimation method uses estimates of F_{ST} for pairs of subpopulations rather than a single *F*-statistic for the entire set of subpopulations. To that aim, I first provide results concerning expected values of measures of population subdivision under isolation by distance. Then, I propose an indirect estimator of the product of population density and second moment of parental axial distance. Theory suggests that this method is more reliable than currently used methods of analysis, particularly for types of dispersal distributions that may be common in natural populations.

There may be some correlation between direct and indirect estimates of demographic parameters, (*e.g.*, HASTINGS and HARRISON 1994; SLATKIN 1994; WARD *et al.* 1994), but detailed case studies generally argue for discrepancies between the two approaches (*e.g.*, CAMPBELL and DOOLEY 1992; SCHILTHUIZEN and LOMBAERTS 1994; JOHNSON and BLACK 1995). Cases where it is possible to compare "indirect" estimates to "direct" estimates obtained from observation of population densities and individual dispersal remain scarce. For the two most detailed data sets I have found, where such a comparison is possible, I find a better agreement between direct estimates and indirect estimates obtained by the present method than with other indirect methods.

ANALYTIC THEORY

Identity by descent in one and two dimensions: I will consider discrete generation models for populations on finite and infinite lattices in one or two dimensions, *i.e.*, for subpopulations on a circle or two-dimensional torus of finite or infinite size. Because exact results are available for these models, it is possible to discuss their interpretation without concern for problems of mathematical formulation. Some models of potentially continuously distributed populations may yield similar results

Address for correspondence: François Rousset, Laboratoire Génétique et Environnement, Institut des Sciences de l'Évolution, CC065, USTL, Place E. Bataillon, 34095 Montpellier Cedex 05, France.
E-mail: rousset@isem.univ-montp2.fr

but their formulation remains difficult (SAWYER and FELSENSTEIN 1981).

MARUYAMA (1970) and SAWYER (1976) provide detailed and complementary mathematical expositions of the lattice models. They are lengthy and will not be repeated here. The APPENDIX summarizes some results for the finite and infinite population models (*i.e.*, with a finite or infinite number of subpopulations). Below are the main assumptions, meaning of parameters, and some approximations for the infinite lattice models based on results of SAWYER (1977).

In their basic formulation the models assume that two gametes produced in the same subpopulation have probability $1/(2N)$ of being copies of the same gene. Under the multinomial sampling scheme of the Wright-Fisher model this amounts to assume that there are either $2N$ "breeding" haploid individuals per subpopulation, or N diploid individuals and that dispersal occurs through gametes only. The results are accurate for zygotic dispersal (NAGYLAKI 1983) and some other mating systems are accounted for by use of "effective population size" arguments (SAWYER 1976; TACHIDA and YOSHIMARU 1996). Within these models, the "effective population size" is a measure of the rate at which genes coalesce per generation.

These models also assume that there is a finite second moment of the distance between a gene and its parent in the previous generation. For symmetric dispersal in one dimension, this is also the variance σ^2 of parental position X relative to offspring position. σ^2 is not the variance Var $(|X|)$ of unsigned distance $|X|$ as $\sigma^2 = \text{Var}(X) = \text{Var}(|X|) + E(|X|)^2$. For isotropic dispersal in two dimensions σ^2 is defined as the variance of parental "axial distance" X, that would be measured along one dimension. The noncentral second moment of parent-offspring euclidian distance is $2\sigma^2$ and should not be confused with the variance Var (R) of the Euclidian distance as $2\sigma^2 = \text{Var}(R) + E(R)^2$ (*e.g.*, CRAWFORD 1984).

If u is the mutation rate per generation, and θ_j the probability of identity by descent of a pair of genes at j steps from each other, then using the results of SAWYER (1977) as explained in the APPENDIX, one obtains in one dimension (all starred symbols will refer to the infinite lattice models):

$$\frac{\theta_j^*}{1-\theta_0^*} \approx \frac{e^{-\sqrt{2u}\,j/\sigma}}{4N\sigma\sqrt{2u}}. \qquad (1)$$

This is an approximation for large geographic distances. On the other hand,

$$\frac{\theta_0^*}{1-\theta_0^*} \approx \frac{1}{4N\sigma\sqrt{2u}} + \frac{A_1}{4N\sigma}, \qquad (2)$$

where A_1 is a constant dependent on the dispersal distribution, but not on N or u. Its definition is given by SAWYER (1977), Equation (2.4) and its biological significance is similar to that of A_2 discussed below.

In two dimensions, for the probability $\theta_{j,k}$ of identity of genes at j steps from each other in one dimension and k in the other, one has

$$\frac{\theta_r^*}{1-\theta_0^*} \approx \frac{K_0(\sqrt{2u}r/\sigma)}{4N\pi\sigma^2}, \qquad (3)$$

where θ_r stands for $\theta_{j,k}$, $r = \sqrt{j^2 + k^2}$ is the distance between genes and K_0 is the modified Bessel function of second kind and zero order.

A different formula must be considered when $r = 0$:

$$\frac{\theta_0^*}{1-\theta_0^*} \approx \frac{-\ln(\sqrt{2u}) + 2\pi A_2}{4N\pi\sigma^2}. \qquad (4)$$

A_2 is of the same nature as A_1 above and an explicit definition is given in the APPENDIX (Equation A11). Its biological significance is discussed later.

***F_{ST}* and related quantities:** The previous theory provides values of the probabilities of identity by descent. Hence it can be used to provide values of the correlation $\beta_j \equiv (\theta_0 - \theta_j)/(1 - \theta_j)$ of genes within populations with respect to genes at some distance j. This quantity is different from F_{ST_j}, which is better defined as $(Q_0 - Q_j)/(1 - Q_j)$, where the Q's are probabilities of identity in state rather than identity by descent, and from the ratio of average coalescence times, $C_{ST_j} \equiv (T_j - T_0)/T_j$ where the T's are average coalescence times of pairs of genes at distance j. These distinctions are necessary to understand how F-statistics are affected by the mutation rate and mutation process (ROUSSET 1996). However, properties of F_{ST} can be deduced from those of C_{ST} and β, and in finite populations in the limit of low mutation rate, the values of all three parameters are identical (SLATKIN 1991; ROUSSET 1996).

When $\theta_j = 0$, β reaches its maximum possible value, which is θ_0. This is the limit value of β at long distances. Under the assumption that $F_{ST} \approx \beta \approx \theta_0$, Equation 4 implies $1/F_{ST} - 1 \approx 4N\pi\sigma^2/(-\ln(\sqrt{2u}) + 2\pi A_2)$ where the denominator is a function of the mutation rate and distribution of dispersal, but not of distance. For the stepping stone model $1/F_{ST} - 1 \approx 2Nm\pi/(-\ln(\sqrt{2u}) + 2\pi A_2)$ where m is the fraction of migrants. Similar formulas have been obtained by KIMURA and WEISS (1964) and SLATKIN and BARTON (1989), and proposed as a basis for estimating either $4N\pi\sigma^2$ or Nm by the latter authors.

It appears useful to reconsider the underlying models. For a one-dimensional infinite population one obtains

$$\frac{\beta_j^*}{1-\beta_j^*} \approx \frac{A_1}{4N\sigma} + \frac{1 - e^{-\sqrt{2u}\,j/\sigma}}{4N\sigma\sqrt{2u}}, \qquad (5)$$

which is the difference between (1) and (2) (see APPENDIX). For a two-dimensional infinite population,

$$\frac{\beta^*_{jk}}{1-\beta^*_{jk}} \approx \frac{-\ln(\sqrt{2u}) - K_0(\sqrt{2u}r/\sigma) + 2\pi A_2}{4N\pi\sigma^2}, \quad (6)$$

which is the difference between (3) and (4). The expression in SLATKIN and BARTON (1989) for F_{ST} at short distances is equivalent except that they do not give a definition for the equivalent of A_2 in their formulas.

The low mutation limits of the above expressions yield

$$\frac{C^*_{ST}}{1-C^*_{ST}} \approx \frac{A_1}{4N\sigma} + \frac{j}{4N\sigma^2}. \quad (7)$$

If A_1 is neglected, this is in agreement with a result of SLATKIN (1993) for the stepping stone model. The two-dimensional result is

$$\frac{C^*_{ST}}{1-C^*_{ST}} \approx \frac{\ln(r/\sigma) - 0.116 + 2\pi A_2}{4N\pi\sigma^2}, \quad (8)$$

because $K_0(x) \approx -(\ln(x/2) + \gamma_e)$ for a small x (ABRAMOVITZ and STEGUN 1972, eq. 9.6.13), where $\gamma_e = 0.5772\ldots$ is Euler's constant. Note that $C_{ST}/(1-C_{ST})$ is the coalescent approximation for $1/(2\hat{M})$ where $\hat{M}$ is the quantity discussed by SLATKIN (1993). For the log-log plots of $\hat{M}$ *vs.* distance discussed there,

$$\log \hat{M} \approx \log(2N\sigma^2) - \log(A_1\sigma + j) \quad (9)$$

in one dimension, and

$$\log \hat{M} \approx \log(2N\pi\sigma^2) - \log(\ln(r/\sigma) - 0.116 + 2\pi A_2) \quad (10)$$

in two dimensions.

Numerical evaluation of β_j and β_{jk} for finite population structures with different distributions of dispersal (detailed in the APPENDIX) are compared to the infinite population coalescent approximation with and without the A_1 or A_2 term in Figures 1 and 2. The theory is remarkably accurate if the A's are taken into account.

Schematically, the variables considered have a linear relationship the slope of which is determined by $N\sigma^2$ only and the intercept determined by both N and more complex features of the dispersal distribution embodied in the definition of the A's. It is not easy to relate the A values to particular features of the distribution of dispersal. However, a given value of σ^2 may be due to a relatively large number of migrants at short distances or to a few long distance migrants. Differentiation between neighboring subpopulations should be more efficiently prevented in the former case than in the latter, resulting in negative A_2/σ^2 values if all migrants are from neighboring populations (*e.g.*, the stepping stone model with $\sigma^2 = 1/200$), and in positive A_2/σ^2 values

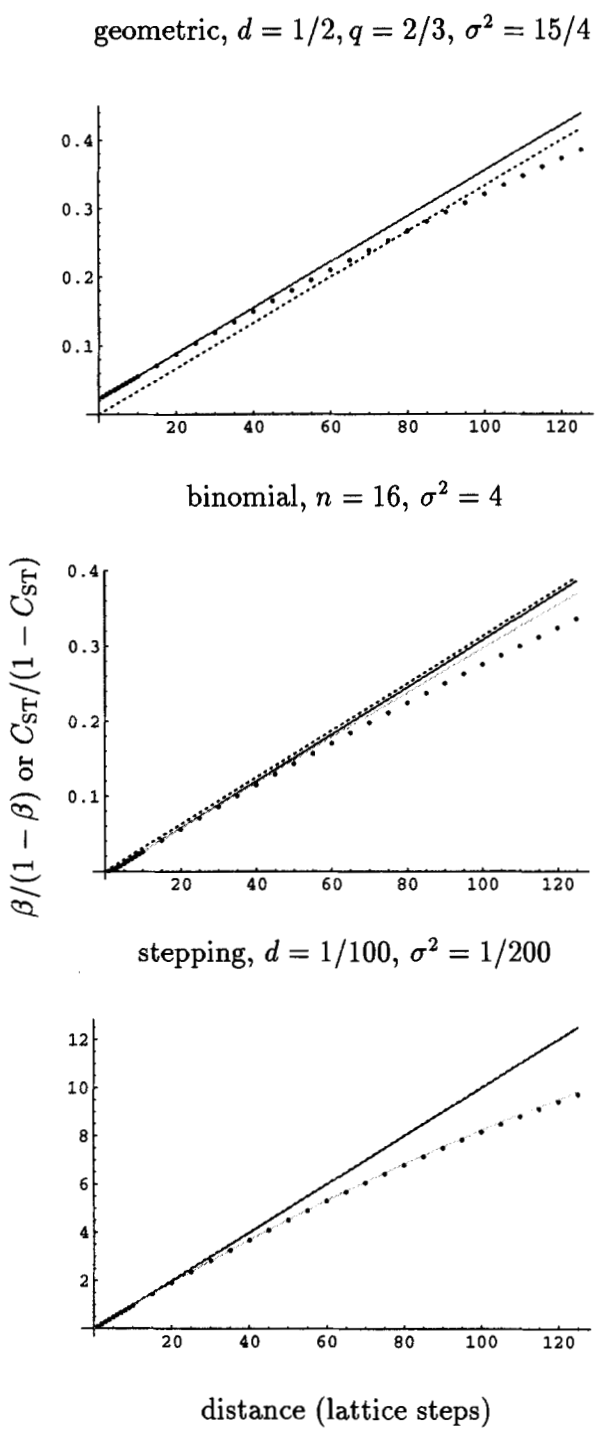

FIGURE 1.—Coalescence and identity measures in one-dimensional models. Exact values of $\beta/(1-\beta)$ (· · ·), computed from (A3) and (A7), are compared to the asymptotic approximation for $C_{ST}/(1-C_{ST})$ in finite one-dimensional lattices for different distributions of migration (Equation 7, dark gray line) and the same approximation without the A_1 term (-----). $u = 10^{-6}$ and $N = 20$ in all cases (a different value of N would only change the y-axis scale). Lattice size is 1000 steps and differentiation shown for distances up to 125 steps. Note the differences between identity by descent measures in the finite population (· · ·) and infinite population (light gray line) models.

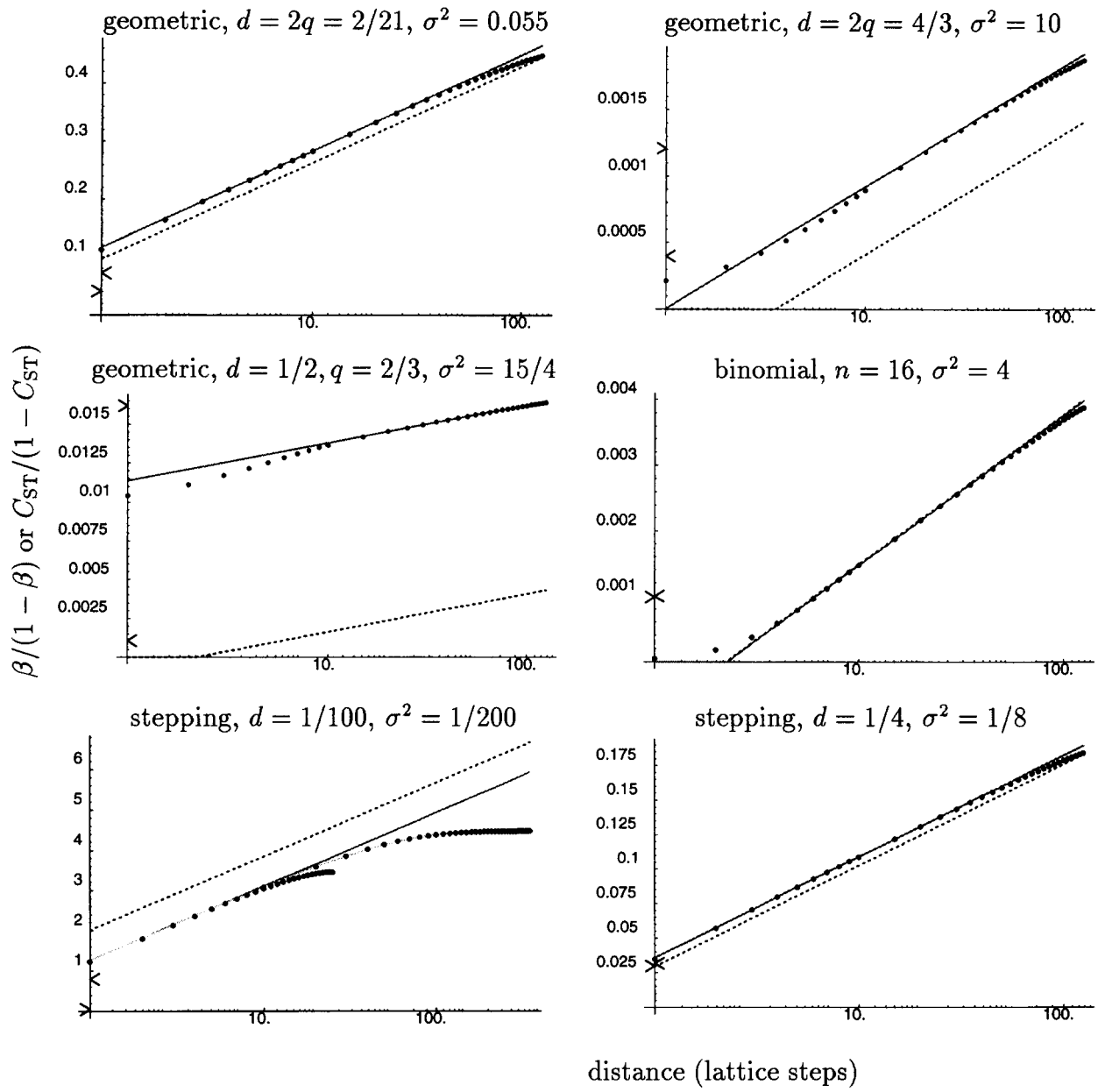

FIGURE 2.—Coalescence and identity measures in two-dimensional models. Note the logarithmic scale for distance. Exact values of $\beta/(1-\beta)$ ($\cdots$), computed from (A5) and (A7), are compared to the asymptotic approximation for $C_{ST}/(1-C_{ST})$ in finite two-dimensional lattices for different distributions of migration (Equation 8, dark gray line) and the same approximation without the A_2 term (-----). The value of $1/(4N\pi\sigma^2)$ is indicated by $<$, and the inverse of WRIGHT's neighborhood size (Equation A12) by $>$. $u = 10^{-6}$ and $N = 20$ in all cases (a different value of N would only change the y-axis scale). Lattice size is 500×500 and differentiation shown for distances up to 125 steps in all cases except for the $m = 1/100$ stepping stone example where 50×50 and 700×700 lattices are considered and differentiation is shown up to half their length. The infinite population result for $\beta/(1-\beta)$ is also shown (light gray line) in the latter case. This example shows the relationship between finite and infinite torus models.

(*e.g.*, the geometric model) if most migrants are from distant populations. Larger values are obtained when the fraction of migrants decreases for a given distribution of dispersal distance among migrants (the geometric cases with $q = {}^2/_3$ and $d = {}^4/_3$ *vs.* $d = {}^1/_2$). The infinite island model may be considered an extreme illustration of this case, where $F_{ST}/(1-F_{ST}) \approx 1/(4Nm)$ and the slope is $1/(4N\pi\sigma^2) = 0$. It shows that differentiation can be arbitrarily much larger than $1/(4N\pi\sigma^2)$. The binomial model is an intermediate case where $|A_2/\sigma^2|$ is small, in agreement with the fact that A_2 is null for Gaussian distributions (SAWYER 1977).

Leptokurtic dispersal distributions are commonly observed in natural populations (ENDLER 1977), for example, in the two data sets discussed below (note that, as for σ^2, kurtosis is not defined here from the central

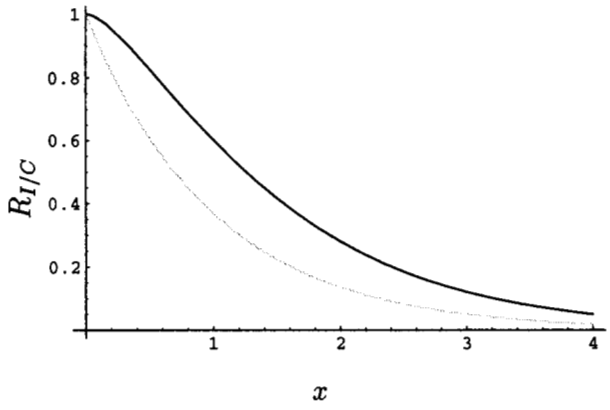

FIGURE 3.—The difference between coalescence and identity measures. These differences are measured by $R_{I/C} = -xK_0'(x)$ (plain line) or $=e^{-x}$ (gray line), in two or one dimension, respectively. See text for usage.

moments of unsigned distance, but from the noncentral moments). They would tend to have high A_2/σ^2 values, so differentiation among populations at low distances ($r < \sigma$) could be common. However, kurtosis is not a perfect descriptor of the extent of migration between neighboring populations in the present models nor of A_2/σ^2. For some extreme lattice models strong kurtosis is compatible with migration only between neighboring subpopulations (the stepping stone example with $\sigma^2 = 1/200$ has the highest kurtosis of all examples in Figure 2).

β and C_{ST} are practically identical at short distances but progressively depart from each other. In two dimensions, the slope of the relationship between $\beta/(1-\beta)$ and logarithm of distance, which is independent of A_2, will be $R_{I/C}$ times that for $C_{ST}/(1-C_{ST})$ at distance r where $dK_0(\sqrt{2u}r/\sigma)/d \ln(r) = R_{I/C}$. This ratio of slopes can be deduced from the value of the function $-xK_0'(x)$ (Figure 3): at distance $r = x\sigma/\sqrt{2u}$, $R_{I/C} = -xK_0'(x)$. For example, $R_{I/C} = 0.8$ for $r \approx 0.56\sigma/\sqrt{2u}$, and $R_{I/C} = 0.2$ for $r \approx 2.4\sigma/\sqrt{2u}$. It is necessary to know both u and σ to determine the distance at which some discrepancy between coalescence and identity measures is reached. For example, if $u = 10^{-4}$ and $\sigma^2 = 0.1$ km^2, the distance where $R_{I/C} = 0.2$ is only 53.7 km, and 1697 km, if $u = 10^{-6}$ and $\sigma^2 = 1$ km^2. In the one-dimensional model, the slope of the relationship between $\beta/(1-\beta)$ and distance (not its logarithm) is given by values of e^{-x} rather than $-xK_0'(x)$ (see Equation 5). The deviation from the coalescent approximation occurs at a shorter distance in one than in two dimensions.

These values are valid only for identity by descent measures. The distance will be shorter, $\sqrt{(k-1)/k}$ times those given above, for a symmetric k-allele model, and longer for stepwise mutation models (ROUSSET 1996). In two dimensions the differences on F_{ST} values

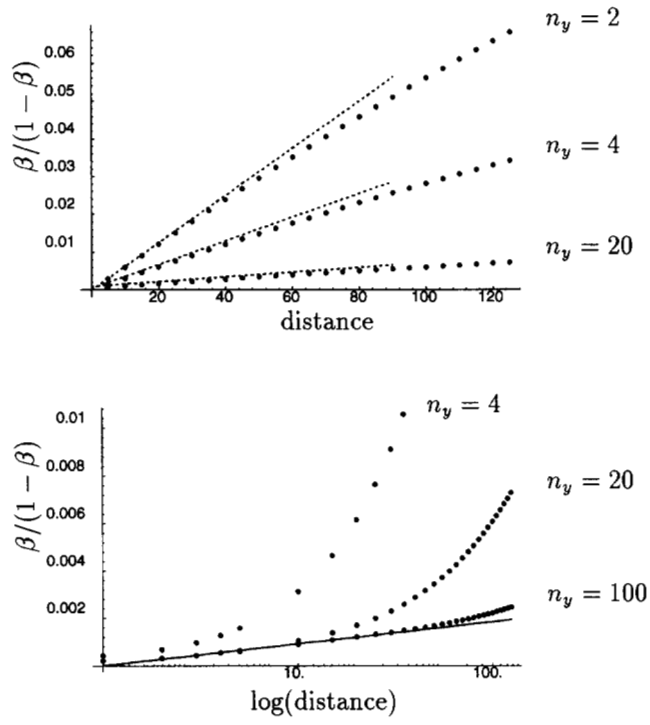

FIGURE 4.—Differentiation in an elongated habitat. This figure shows differentiation as a function of distance and logarithm of distance along the long axis of an habitat of 1000 × n_y subpopulations, for values of n_y from 2 to 100. Dispersal follows the geometric model with $d = 2q = 4/3$. Exact values of $\beta/(1-\beta)$ ($\cdots$) were computed from (A5) and (A7). --- show expected slopes for one-dimensional habitats with the same linear density Nn_y. The dark gray line is the asymptotic approximation for $C_{ST}/(1-C_{ST})$ (Equation 8). $u = 10^{-6}$ and $N = 20$.

due to differences in mutation rates may not be considerable. For example, for $N = 6.23$, $\sigma^2 = 2.72$, and $A_2 = 0.745$ (values chosen to fit the indirect estimates in the example from human populations below), the maximum β^* value is 0.05 for $u = 10^{-6}$ and 0.04 for $u = 10^{-4}$. Differences due to different mutational processes are smaller and may be difficult to detect (details not shown).

When should a narrow elongated habitat be considered one- or two-dimensional is not obvious. Numerical examples (Figure 4) suggest that in such habitats, differentiation between populations at distance smaller than half the width of the habitat follows the two-dimensional model in that there is a linear relationship between $F_{ST}/(1-F_{ST})$ and logarithm of distance, and differentiation between populations at distances larger than the width of the habitat follows the one-dimensional model, with a linear relationship between $F_{ST}/(1-F_{ST})$ and distance and slope determined by the density of individuals per unit length. Thus, differentiation in elongated habitats can be analyzed using the two-dimensional model when habitat is "locally" two-dimensional at the scale of study defined by the distance

between samples, and using the one-dimensional model at larger distances.

IMPLICATIONS FOR DATA ANALYSIS

The previous results show that an appropriate representation of data is a plot of estimates of $F_{ST}/(1 - F_{ST})$ against the distance in one dimension or logarithm of distance in two dimensions. In the latter case, the expected relationship is approximately linear, $y = a + bx$ with slope $b = 1/(4N\pi\sigma^2)$ and intercept $a = -\ln(\sigma) + \gamma_e - \ln(2) + 2\pi A_2$. The slope of the regression may be used to estimate $1/(4N\pi\sigma^2)$. The quantity

$$e^{-a/b} = 2\sigma e^{-2\pi A_2 - \gamma_e} \approx 1.123\sigma e^{-2\pi A_2} \qquad (11)$$

would be independent of subpopulation size. If we could somehow discard A_2, it would be possible to estimate both σ and N, for example, by $\hat{\sigma} \equiv e^{-\hat{a}/\hat{b}}/1.123$ and $\hat{N} \equiv 1/(4\pi\hat{b}\hat{\sigma}^2)$, if $A_2 = 0$. This would be a very poor method for most examples in Figure 2, and there is no reason to consider A_2 negligible in real situations, hence σ and N cannot be estimated separately.

The linear relationship never perfectly holds, and if possible it is preferable to take into account only the differentiation observed at distances $r > \sigma$ and r lower than the value determined by some acceptable $R_{I/C}$ value, for example, $r < 0.56\sigma/\sqrt{2u}$ for $R_{I/C} = 0.8$ in two dimensions. If only a minimal estimate of σ is available, it provides an upper bound to the bias measured by $R_{I/C}$.

Interpretation of parameters: In many recent studies attempting to estimate a demographic parameter from genetic data, it is considered appropriate to estimate a number of migrants between subpopulations and this creates a need to define subpopulations. This raises two difficulties, that of estimating number of migrants when models of isolation by distance do not suggest any simple way to do so and that of defining subpopulations. The surface occupied by such subpopulations is often estimated by the surface within which no differentiation is detected, or equated to the neighborhood area as given by WRIGHT's formulas. These procedures have little theoretical support, and it is not the purpose of the present paper to provide such support. Nor does it provide any ground to define subpopulations that can be considered panmictic in some sense, another problematic interpretation of WRIGHT's neighborhood.

In fact, the present method of analysis does not require the definition of subpopulations on a lattice, but only the knowledge of the distances between samples. In their basic formulation the lattice models assume that the distance between neighboring subpopulations ϵ is 1. In data analyses the distance between neighboring subpopulations is often unknown or even difficult to define, so it is necessary to identify quantities that interpretation does not depend on the assumption that $\epsilon = 1$, or equivalently is not affected by a change of scale.

The slope is such a quantity: whatever the scale it is inversely proportional to the product of population density D_ϵ by the second moment of dispersal distance σ_ϵ^2. In two dimensions, the slope does not depend on the spatial scale because of the logarithmic effect of distance, and $D_\epsilon\sigma_\epsilon^2$ is always $N\sigma^2$: when the distance between steps ϵ is 1, this is density, $D_\epsilon = N$, times second moment, $\sigma_\epsilon^2 = \sigma_1^2$, and if scale is changed this is still density, $D_\epsilon = N/\epsilon^2$, times second moment, $\sigma_\epsilon^2 = \sigma_1^2\epsilon^2$. In one dimension, the slope is $N\sigma_1^2\epsilon$: it depends on the spatial scale but is always density, $D_\epsilon = N/\epsilon$, times second moment, $\sigma_\epsilon^2 = \sigma_1^2\epsilon^2$.

Then, in the special case of the two-dimensional stepping stone model, $\sigma_\epsilon^2 = \sigma_1^2\epsilon^2 \approx m\epsilon^2/2$ and $2D_\epsilon\sigma_\epsilon^2 = (N/\epsilon^2)m\epsilon^2 = Nm$ is the number of migrants per subpopulation. Thus Nm can be estimated even if ϵ is unknown, but obviously this quantity provides no information about movements of individuals unless ϵ is known. In the one-dimensional stepping stone model, $D_\epsilon\sigma_\epsilon^2 = Nm\epsilon$ is not the number of migrants per subpopulation when $\epsilon \neq 1$, so the number of migrants cannot be estimated if ϵ is unknown.

Examples: In this section, I give two applications of the approach described above. First, I have applied it to Gainj- and Kalam-speaking people of New Guinea for which both genetic differentiation and demographic properties have been extensively studied (*e.g.*, WOOD 1987; LONG *et al.* 1987), which permits a comparison of different estimates of parameters.

The natal dispersal data of WOOD *et al.* (1985) provide the position of parents of individuals that reproduced in some place. They can be used to estimate σ^2. Women have $\hat{\sigma}_f^2 = 2.21$ km^2 and men have $\hat{\sigma}_m^2 = 3.23$ km^2. Hence $\hat{\sigma}^2 = (\hat{\sigma}_f^2 + \hat{\sigma}_m^2)/2 = 2.72$ km^2. The population density is ~24 individuals · km^{-2}, and age structure and distribution of number of offspring may reduce the "effective population size" by a factor of 2 (WOOD 1987), hence a "direct" estimate of $4D\pi\sigma^2$ is 410.

I reanalyzed the five loci studied by LONG *et al.* (1987) (Figure 5). The slope of the regression is 0.0047, hence the "slope" estimate of $4D\pi\sigma^2$ is 213 individuals, about half the direct estimate (the slope estimate is 265 individuals if differentiation at distances larger than $\hat{\sigma}$ only is taken into account). The estimate of F_{ST} computed from all subpopulations is 0.025, hence by the "$1/F_{ST} - 1$" method one would obtain 40, a poorer estimate of $4D\pi\sigma^2$. The high differentiation at short distances may be at least in part explained by the nature of the migration distribution, which is strongly leptokurtic (the kurtosis for the axial migration distribution, inferred under the assumption of isotropic migration, is 14.6), but may also be due to other factors not included in the model.

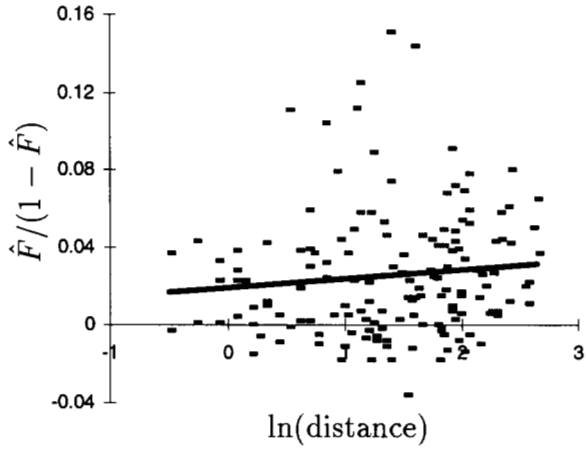

FIGURE 5.—Differentiation among Gainj- and Kalam-speaking peoples. Multilocus estimates of pairwise differentiation are plotted against logarithm of map distances (in Km). The regression is $y = 0.0047x + 0.0191$ and the maximum distance between two subpopulations is 14 km. Genotypic data appear in LONG *et al.* (1986). F_{ST} was estimated according to WEIR and COCKERHAM (1984).

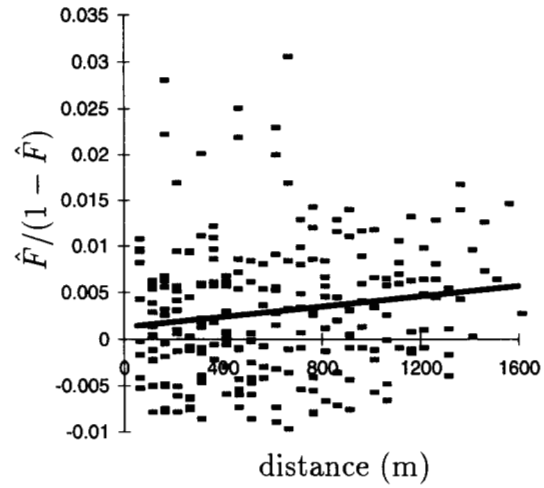

FIGURE 6.—Differentiation among Bembicium snails. Multilocus estimates of pairwise differentiation are plotted against distance. The regression is $y = 2.76\ 10^{-6}x + 0.0013$. F_{ST} was estimated according to WEIR and COCKERHAM (1984).

The other example allowing comparison of direct and indirect estimates that I have found is that of the intertidal snail *Bembicium vittatum* (JOHNSON and BLACK 1995). The habitat is linear (~3 m wide) and the results will be analyzed according to the one dimensional model.

I used the estimate of density at Noddy Shore, $D = 111$ adults $\cdot\ m^{-1}$, which is typical of populations along the "1600 m" transect (JOHNSON and BLACK 1995). Dispersal was studied by mark-recapture experiments within this transect, and I computed an estimate of the second moment σ^2 of dispersal distance over one generation (about 12 months) as 2.4 times the estimate of the second moment over 5 months (see JOHNSON and BLACK 1995). From their Table 1, $\hat{\sigma}^2 = 2.4(6.4^2 + 181.5) = 533.9$ m^2. Then a direct estimate of $4D\sigma^2$ is $2.4\ 10^5$ individuals $\cdot$ m (the latter unit may surprise, but in one dimension the product of linear density times second moment of dispersal distance necessarily scales as a number of individuals times a distance). I reanalyzed the genotypic data for the 1600 m transect (13 loci, provided by M. S. JOHNSON) and found that the slope of the regression of $\hat{F}/(1-\hat{F})$ to distance is $2.76\ 10^{-6}$ (individuals $\cdot$ m)$^{-1}$ (Figure 6). According to the one-dimensional model the inverse value $3.6\ 10^5$ individuals $\cdot$ m is an estimate of $4D\sigma^2$, 1.5 times the direct estimate.

JOHNSON and BLACK's (1995) estimate of Nm, based on the approximation $\log \hat{M} \approx \log (Nm)$ at one interdeme distance (SLATKIN 1993), is 22 at 150 m. Considering more general migration distributions, and neglecting the A_1 term in Equation 9, this could be interpreted as an estimate of $(D\sigma^2)$/distance so an estimate of $4D\sigma^2$ is $88 \times 150 = 13200$ individuals $\cdot$ m, far from the "direct" estimate. This discrepancy may be due to neglecting A_1 as well as estimation problems since log − log representations prevent the use of unbiased estimators of F_{ST}.

The analytical theory can be used to assess to which extent the difference between the slope expected from coalescence measures and the slope expected from F-statistics may bias the analysis. In the human example the ratio $R_{I/C}$ as given by Figure 3 is only 0.9996 at 14 km, and in the one-dimensional example this is 0.907 at 1600 m (assuming $u = 10^{-6}$ and using estimated values of σ^2 in both cases). Thus in the one-dimensional case a slight bias in estimation is expected. It should result in an overestimation of $N\sigma^2$ by some factor $<1/0.907$, the exact value depending on the location of all samples.

DISCUSSION

The method defined here is based on F_{ST} values for pairs of populations. F_{ST}-based analyses appear appropriate in a number of ways (CROW and AOKI 1984; SLATKIN 1991, 1993, 1994). F_{ST} values are relatively independent of mutation rate and mutation process, and of total population size, in contrast to probabilities of identity. The formulas for F_{ST}, obtained from earlier asymptotic results for large distances and infinite populations, turn out to be remarkably accurate for the finite population models and at short distances, particularly in two dimensions. They can be related to average coalescence times, though the maximum value at long distance is not given by coalescence theory and depends on both mutation rate and mutation process. The representation introduced here should clarify the relationship between identity and coalescence measures in isolation by distance models. It also emphasizes the role of the

kurtosis of the distribution of dispersal on expected levels of differentiation.

In his formulation of isolation by distance models WRIGHT (1946) took into account the nature of the dispersal distribution by way of the neighborhood size, a number of individuals, and the neighborhood area, the surface occupied by this number of individuals. His most precise definition of neighborhood size is the reverse of "the chance that two uniting gametes came from the same individual" (see Equation A12). It is apparent from Figure 2 that this quantity does not describe any simple feature of the model, neither does the neighborhood area that is a proportional to the neighborhood size for all examples in Figure 2. The quantity $4N\pi\sigma^2$ that determines the value of the slope in the two-dimensional model should not be confused with the neighborhood size. WRIGHT found that the value of the neighborhood size is $4N\pi\sigma^2$ for Gaussian dispersal, but is different for other distributions. Here the result that the slope is $1/(4N\pi\sigma^2)$ arises without reference to a Gaussian distribution of parental distances, and holds more generally.

The theoretical results agree with earlier numerical studies of F_{ST} and related quantities that showed that differentiation in the two-dimensional stepping stone (*i.e.*, nearest neighbor dispersal) model is roughly as expected under the island model, and increases more rapidly with distance in one-dimensional models (KIMURA and MARUYAMA 1971; CROW and AOKI 1984; SLATKIN and BARTON 1989; SLATKIN 1991). An important but frequently neglected message from the two-dimensional stepping stone model is that subpopulations that never exchange migrants may not exhibit much higher F_{ST} values than those that do.

Another important implication of these models is that an apparent absence of a pattern of isolation by distance may be due not only to range expansions (SLATKIN 1993), but also to sampling at large distances (so that that $R_{I/C} \ll 1$), or to large values of $D\sigma^2$. The capacity to detect isolation by distance depends also on the range of (logarithm of) distance values investigated, and on the variance of estimators that is probably lower at short distances.

However, the models show that the variation of pairwise F_{ST} values with distance may be more easily interpretable that the F_{ST} values themselves. Using F_{ST} values themselves to estimate demographic parameters is not straightforward, and the examples confirm these expectations. In both of them, the "slope" and direct estimates differ by less than twofold. This agreement may be due in part to the fact that the estimates are based on differentiation at a relatively small geographical scale, where stochastic equilibrium is approached more rapidly (SLATKIN 1993), and where differentiation should be independent of the details of the mutation process. Some other complicating factors such as spatial variation of demographic parameters or selection variable in space are also more easily avoided at short distances.

The relatively small discrepancies between direct and indirect estimates may be due to minor inadequacies of the models as well as imprecision of the estimators. Good mark-recapture estimates of σ^2 may also be difficult to obtain because of long distance migration outside the study area. More studies of this kind would be necessary before systematic differences between different kind of estimates can be detected and interpreted. Nevertheless, the variation of $F_{ST}/(1 - F_{ST})$ with distance contains the most easily interpretable information, and the available examples show that there is a better match between direct and indirect estimates of $D\sigma^2$ obtained in this way than with indirect estimates of this quantity obtained by other methods.

I thank M. S. JOHNSON for providing the *Bembicium* data set and useful comments, and J. BRITTON-DAVIDIAN, C. CHEVILLON, P. JARNE, M. RAYMOND, M. SLATKIN and particularly Y. MICHALAKIS for providing references, helpful suggestions or comments on the manuscript. Some of the programs used in the data analyses were written in collaboration with M. RAYMOND and will be included in future versions of the Genepop package (RAYMOND and ROUSSET 1995) available by ftp at ftp.cefe.cnrs-mop.fr. This work was supported by the Programme Environnement du Centre National de la Recherche Scientifique (GDR 11.05) and the Centre de Biosystématique de Montpellier. This is paper 97-019 of the Institut des Sciences de l'Évolution.

LITERATURE CITED

ABRAMOVITZ, M., and I. A. STEGUN (Editors), 1972 *Handbook of Mathematical Functions.* Dover, New York.

CAMPBELL, D. R., and J. L. DOOLEY, 1992 The spatial scale of genetic differentiation in a hummingbird-pollinated plant: comparison with models of isolation by distance. Am. Nat. **139:** 735–748.

CRAWFORD, T., 1984 The estimation of neighborhood parameters for plant populations. Heredity **52:** 273–283.

CROW, J. F., and K. AOKI, 1984 Group selection for a polygenic behavioural trait: estimating the degree of population subdivision. Proc. Natl. Acad. Sci. USA **81:** 6073–6077.

ENDLER, J. A., 1977 *Geographical Variation, Speciation, and Clines.* Princeton University Press, Princeton, NJ.

HASTINGS, A., and S. HARRISON, 1994 Metapopulation dynamics and genetics. Ann. Rev. Ecol. Syst. **25:** 167–188.

HUDSON, R. R., 1990 Gene genealogies and the coalescent process. Oxford Surv. Evol. Biol. **7:** 1–44.

JOHNSON, M. S., and R. BLACK, 1995 Neighborhood size and the importance of barriers to gene flow in an intertidal snail. Heredity **75:** 142–154.

KIMURA, M., and T. MARUYAMA, 1971 Pattern of neutral polymorphism in a geographically structured population. Genet. Res. **18:** 125–131.

KIMURA, M., and G. H. WEISS, 1964 The stepping stone model of population structure and the decrease of genetic correlation with distance. Genetics **49:** 561–576.

LANDE, R., 1992 Neutral model of quantitative genetic variance in an island model with local extinction and recolonization. Evolution **46:** 381–389.

LONG, J. C., J. M. NAIDU, H. W. MOHRENWEISER, H. GERSHOWITZ, P. L. JOHNSON *et al.*, 1986 Genetic characterization of Gainj- and Kalam-speaking peoples of Papua New Guinea. Am. J. Phys. Anthropol. **70:** 75–96.

LONG, J. C., P. E. SMOUSE and J. J. WOOD, 1987 The allelic correlation structure of Gainj- and Kalam-speaking people. II. The genetic distance between population subdivisions. Genetics **117:** 273–283.

MALÉCOT, G., 1950 Quelques schémas probabilistes sur la variabilité des populations naturelles. Annales de l'Université de Lyon A **13:** 37–60.

MALÉCOT, G., 1951 Un traitement stochastique des problèmes linéaires (mutation, linkage, migration) en génétique des populations. Annales de l'Université de Lyon A **14:** 79–117.

MARUYAMA, T., 1970 Effective number of alleles in a subdivided population. Theor. Pop. Biol. **1:** 273–306.

NAGYLAKI, T., 1976 The decay of genetic variability in geographically structured populations. II. Theor. Pop. Biol. **10:** 70–82.

NAGYLAKI, T., 1983 The robustness of neutral models of geographical variation. Theor. Pop. Biol. **24:** 268–294.

RAYMOND, M., and F. ROUSSET, 1995 GENEPOP Version 1.2: population genetics software for exact tests and ecumenicism. J. Hered. **86:** 248–249.

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

SAWYER, S., 1976 Results for the stepping stone model for migration in population genetics. Ann. Prob. **4:** 699–728.

SAWYER, S., 1977 Asymptotic properties of the equilibrium probability of identity in a geographically structured population. Adv. Appl. Prob. **9:** 268–282.

SAWYER, S., and J. FELSENSTEIN, 1981 A continuous migration model with stable demography. J. Math. Biol. **11:** 193–205.

SCHILTHUIZEN, M., and M. LOMBAERTS, 1994 Population structure and levels of gene flow in the mediterranean land snail *Albinaria corrugata* (Pulmonata: Clausiliidae). Evolution **48:** 577–586.

SLATKIN, M., 1991 Inbreeding coefficients and coalescence times. Genet. Res. **58:** 167–175.

SLATKIN, M., 1993 Isolation by distance in equilibrium and nonequilibrium populations. Evolution **47:** 264–279.

SLATKIN, M., 1994 Gene flow and population structure, pp. 3–17 in *Ecological Genetics*, edited by L. A. REAL. Princeton University Press, Princeton, NJ.

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 indirect methods for estimating average levels of gene flow. Evolution **43:** 1349–1368.

SPITZER, F., 1976 *Principles of Random Walk*, Ed. 2. Springer-Verlag, Berlin.

STROBECK, C., 1987 Average number of nucleotide differences in a sample from a single subpopulation: a test for population subdivision. Genetics **117:** 149–153.

TACHIDA, H., and H. YOSHIMARU, 1996 Genetic diversity in partially selfing populations with the stepping-stone structure. Heredity **77:** 469–475.

WARD, R. D., M. WOODWARD and D. O. F. SKIBINSKI, 1994 A comparison of genetic diversity levels in marine, freshwater, and anadromous fishes. J. Fish Biol. **44:** 213–232.

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

WOLFRAM, S., 1991 *Mathematica*, Ed. 2. Addison Wesley, Redwood City, CA.

WOOD, J. W., 1987 The genetic demography of the Gainj of Papua New Guinea. 2. Determinants of effective population size. Am. Nat. **129:** 165–187.

WOOD, J. W., P. E. SMOUSE and J. C. LONG, 1985 Sex-specific dispersal patterns in two human populations of highland New Guinea. Am. Nat. **125:** 747–768.

WRIGHT, S., 1946 Isolation by distance under diverse systems of mating. Genetics **31:** 39–59.

WRIGHT, S., 1969 *Evolution and the Genetics of Populations. II. The Theory of Gene Frequencies.* University of Chicago Press, Chicago.

Communicating editor: W. J. EWENS

APPENDIX

Identity by descent: Consider migration in one dimension. Let $\gamma \equiv (1 - u)^2$ and m_l be the probability that the parent of some gene in population j was in population $j + l$, then at stochastic equilibrium the probabilities θ_j of identity by descent of pairs of genes in subpopulations i and $i + j$ obey the relationship (MALÉCOT 1951)

$$\theta_j = \gamma\left(\sum_k \sum_l m_k m_{k+l}\theta_{j+l} + \sum_k m_k m_{k-j}\frac{1-\theta_0}{2N}\right). \quad \text{(A1)}$$

An explicit solution is expressed as follows. For n subpopulations on a circle, let $\tilde{n}$ be the integer part of $n/2$, $\psi(z) \equiv \sum_{-\infty}^{+\infty} m_l z^l$ be the generating function of the m's, and define

$$\xi(x) \equiv \psi^2(e^{ix});$$

$$\xi_k \equiv \xi\left(\frac{2\pi k}{n}\right) = \psi^2(e^{i2\pi k/n}). \quad \text{(A2)}$$

Then

$$\theta_j = \frac{(1-\theta_0)}{2Nn}\sum_{k=0}^{\tilde{n}}\frac{\Delta_k\gamma\xi_k}{(1-\gamma\xi_k)}\cos\left(\frac{2\pi kj}{n}\right), \quad \text{(A3)}$$

where $\Delta_0 = \Delta_{n/2} = 1$ and $\Delta_k = 2$ otherwise [see MARUYAMA (1970) with notations changed and slightly different value of $\tilde{n}$]. The limit when n goes to infinity is (MALÉCOT 1950; NAGYLAKI 1976; SAWYER 1977)

$$\theta_j^* = \frac{(1-\theta_0^*)}{2N\pi}\int_0^\pi \frac{\gamma\psi^2(e^{ix})}{1-\gamma\psi^2(e^{ix})}\cos(jx)\,dx. \quad \text{(A4)}$$

In the same way, for a two-dimensional torus of $n_x \times n_y$ subpopulations,

$$\theta_{jk} = \frac{(1-\theta_{00})}{2Nn_xn_y}\sum_{l=0}^{\tilde{n}_x}\sum_{m=0}^{\tilde{n}_y}\frac{\gamma\xi_l\xi_m}{1-\gamma\xi_l\xi_m}\Delta_l\Delta_m \times \cos\left(\frac{2\pi jl}{n_x}\right)\cos\left(\frac{2\pi km}{n_y}\right), \quad \text{(A5)}$$

where the $\tilde{n}$'s, ξ's and ψ's are defined as above, one for each dimension. When both n_x and n_y goes to infinity, the above result converges to

$$\theta_{jk}^* = \frac{(1-\theta_{00}^*)}{2N\pi^2}\int_0^\pi\int_0^\pi \frac{\gamma\psi_x^2(e^{ix})\psi_y^2(e^{iy})}{1-\gamma\psi_x^2(e^{ix})\psi_y^2(e^{iy})} \times \cos(jx)\cos(ky)\,dxdy. \quad \text{(A6)}$$

Evaluation of the integrals in (A4) and (A6) is detailed by SAWYER (1977) and yields Equations 1–4. These formulas are asymptotic results for low mutation rates, *i.e.*, for γ in the neighborhood of 1. Taken as function of γ these are approximations for the generating functions of coalescence times $\theta(\gamma)$ in the neighborhood of 1.

Values of β and C_{ST}: It is useful to consider the quantity

$$\frac{\beta_j}{1-\beta_j} = \frac{\theta_0 - \theta_j}{1-\theta_0}. \quad \text{(A7)}$$

At the low mutation limit, it can be interpreted as a ratio of average coalescence times,

$$\lim_{\gamma\to 1} \frac{\beta_j}{1-\beta_j} = \frac{C_{ST}}{1-C_{ST}} = \frac{T_j}{T_0} - 1. \quad \text{(A8)}$$

This ratio has a finite limit when the number of subpopulations increases though the average coalescence times themselves become infinite. In finite populations $T_0 = 2Nn_p$, where n_p is the number of subpopulations (STROBECK 1987), hence T_j is readily obtained when C_{ST} is known.

From Equation A3,

$$\frac{\beta_j}{1-\beta_j} = \frac{1}{2Nn}\sum_{k=1}^{\tilde{n}} \frac{\Delta_k \gamma \xi_k}{(1-\gamma\xi_k)} \times \left(1 - \cos\left(\frac{2\pi k j}{n}\right)\right), \quad \text{(A9)}$$

and when n goes to infinity,

$$\frac{\beta_j^*}{1-\beta_j^*} = \frac{1}{2N\pi}\int_0^\pi \frac{\gamma\psi^2(e^{ix})}{1-\gamma\psi^2(e^{ix})} \times (1-\cos(jx))\,dx. \quad \text{(A10)}$$

The integral is no more than a difference between two integrals discussed by SAWYER (1977). In this way one obtains Equation 5 as the difference between Equations 1 and 2. Likewise Equation 6 is the difference between Equations 3 and 4.

Discrete migration distributions, A_2 values, and Wright's neighborhood size: The examples make use of the following distributions of parent-offspring distance. In some cases the probability of migration by l steps in one dimension is $m_l(d, n) = dC_n^{l+n/2}2^{-n} + (1-d)\delta_{l0}(n \text{ even})$, $\psi(e^{ix}) = 1 - d(1-\cos^n(x/2))$ and $\sigma^2 = dn/4$. When $n = 2$ it corresponds to the stepping stone model with migration rate in each dimention $d/2$. (The migration rate $m \approx d$ in two dimensions if d is small.) When $d = 1$ and $n \geq 2$, this is a "shifted" binomial that may be considered a discrete equivalent of Gaussian migration. In other cases $m_0 = 1 - d/2$ and

$$m_l = (1-q)\,q^{l-1}\,d/4 \text{ for } l \neq 0,$$

$$\psi(e^{ix}) = 1 - [1 - (1-q)(\cos(x) - q)/(1 - 2\cos(x)q + q^2)]\,d/2$$

and

$$\sigma^2 = (d/2)(1+q)/(1-q)^2.$$

Thus, the fraction of migrants is determined by d, and among migrants distance follows a geometric distribution described by q.

A_2 is defined by SAWYER (1977), Equation (3.4). Since $C_{ST}/(1-C_{ST})$ is almost identical to the "potential kernel" of the random walk defined by the ψ^2's (SPITZER 1976), some results of SPITZER (1976), p. 124, can be used to obtain a somewhat more explicit formula in the case of isotropic migration:

$$A_2 = \frac{\sigma^2}{\pi^2}\int_0^\pi\int_0^\pi \frac{\psi_x^2\psi_y^2}{1-\psi_x^2\psi_y^2} - \frac{1}{\sigma^2(x^2+y^2)}\,dxdy + \frac{\ln(2\pi\sigma)}{2\pi} - \frac{\lambda}{\pi^2}, \quad \text{(A11)}$$

where $\lambda = 0.9159\ldots$ is Catalan's constant. A_1 and A_2 were computed using *Mathematica* (WOLFRAM 1991).

In the present notations, WRIGHT's neighborhood size (WRIGHT 1969, Equations 12.40–12.41) can be written

$$\frac{N\pi^2}{\int_0^\pi\int_0^\pi \psi_x^2\psi_y^2\,dxdy} \quad \text{(A12)}$$

for the lattice models. It has no simple relationship to N and to A_2/σ^2 or the integral in the definition of A_2.