DOI: 10.1534/genetics.108.094342

Coalescence Times and F_{ST} Under a Skewed Offspring Distribution Among Individuals in a Population

Bjarki Eldon[1] and John Wakeley

Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, Massachusetts 02138

Manuscript received July 23, 2008
Accepted for publication November 25, 2008

ABSTRACT

Estimates of gene flow between subpopulations based on F_{ST} (or N_{ST}) are shown to be confounded by the reproduction parameters of a model of skewed offspring distribution. Genetic evidence of population subdivision can be observed even when gene flow is very high, if the offspring distribution is skewed. A skewed offspring distribution arises when individuals can have very many offspring with some probability. This leads to high probability of identity by descent within subpopulations and results in genetic heterogeneity between subpopulations even when Nm is very large. Thus, we consider a limiting model in which the rates of coalescence and migration can be much higher than for a Wright–Fisher population. We derive the densities of pairwise coalescence times and expressions for F_{ST} and other statistics under both the finite island model and a many-demes limit model. The results can explain the observed genetic heterogeneity among subpopulations of certain marine organisms despite substantial gene flow.

NATURAL populations of organisms are often subdivided by geography. Individuals may or may not migrate between these subpopulations. Modeling gene flow between subpopulations can be traced back to Wright (1931), whose island model describes a population subdivided into discrete, local subpopulations by geography, with limited migration between individual subpopulations. With the advent of techniques to characterize genetic variation, several measures of population subdivision have been proposed on the basis of probabilities of identity (see Rousset 2002 for a review). These include Wright's (1951) F_{ST}, Nei's (1982) γ_{ST}, and Lynch and Crease's (1990) N_{ST}.

The quantity F_{ST} can, under the assumption of equilibrium, be used to estimate levels of gene flow from allozyme data (Wright 1951). The quantity γ_{ST} can be calculated from DNA sequence data and is equivalent to F_{ST} if the mutation rate is very low (Slatkin 1991). Levels of gene flow between subpopulations can thus also be estimated from DNA sequence data. However, Slatkin (1991) argues that F_{ST} is appropriate for allozyme data, whereas the gene genealogy-based method of Slatkin and Maddison (1989) is appropriate for DNA sequence data. As F_{ST} continues to be used in investigations of population structure and, recently, as a tool for identifying loci under selection (*e.g.*, Murray and Hare 2006), we are concerned with F_{ST} and related measures below.

We derive expressions for F_{ST} and N_{ST} under the island model of population subdivision with symmetric migration (Nagylaki 1980; Strobeck 1987) and skewed offspring distribution among individuals in a population. When the offspring distribution is skewed, individuals have some nonnegligible probability of having very many offspring. The population model of skewed offspring distribution we adopt in this work can result in an ancestral process with asynchronous multiple mergers (Eldon and Wakeley 2006). An ancestral process with asynchronous multiple mergers, or Λ-coalescent, was introduced by Pitman (1999) and also derived by Sagitov (1999) from a Cannings (1974) model. In a Λ-coalescent, any number of ancestral lines can coalesce at once to a single ancestor. In contrast, the Kingman coalescent (Kingman 1982a,b) allows only two lines to coalesce each time. For a single population, the ancestral process obtained from the population model of Eldon and Wakeley (2006), and employed in this work, is a special case of the Λ-coalescent of Pitman (1999) and Sagitov (1999).

Type III survivorship curve, and high fecundity, characterize a diverse group of organisms (*e.g.*, many plants and marine animals). A prime example are marine species with broadcast spawning, including Atlantic cod (*Gadus morhua*; Árnason 2004), Pacific oysters (*Crassostrea gigas*; Beckenbach 1994; Hedgecock 1994a), and red drum (*Sciaenops ocellatus*; Turner *et al.* 2002). A model of skewed offspring distribution, in which individuals can have very many offspring with a nonnegligible probability, may therefore better apply in such cases than the Wright–Fisher (Fisher 1930; Wright 1931) or the Moran (Moran 1958, 1962) models.

[1]*Corresponding author:* 4100 Biological Laboratories, 16 Divinity Ave., Cambridge, MA 02138. E-mail: eldon@fas.harvard.edu

Genetic observations from these species also argue against the standard population models. Genetic diversity is observed to be much lower than expected on the basis of population size for some marine populations (Hedgecock *et al.* 1982; Nei and Graur 1984; Avise *et al.* 1988; Avise 1994). In particular, low effective to actual population size ratios have been reported for Atlantic cod (Árnason 2004), red drum (Turner *et al.* 2002), and the Pacific oyster (Hedgecock 1994a), and this has been explained by high variance in offspring distribution (Crow and Kimura 1970; Hedrick 2005). Second, models of skewed offspring distribution predict a large number of singleton variants (Eldon and Wakeley 2006; Sargsyan and Wakeley 2008), a feature observed, for example, in Pacific oysters (Boom *et al.* 1994), Atlantic cod (Árnason 2004), and some hydrothermal vent taxa (Won *et al.* 2003; Hurtado *et al.* 2004; Johnson *et al.* 2006; Young *et al.* 2008).

Genetic heterogeneity on a small spatial scale has been observed for many marine populations, including the purple sea urchin (*Strongylocentrotus purpuratus*; Edmands *et al.* 1996), even though planktonic larvae disperse over wide-ranging habitats (Johnson and Black 1984; Watts *et al.* 1990; Hedgecock 1994b; David *et al.* 1997). A range of explanations has been proposed for the observed heterogeneity (see Burton 1983; Palumbi 1994). Our aim is to address, by analytic methods, the problem concerning the genetic population structure of a highly fecund species with potentially highly skewed offspring distribution, like the Atlantic cod (Árnason *et al.* 2000).

We obtain the probability distributions of pairwise coalescence times, and expressions for F_{ST}, for both the finite island and a many-demes limit model. Our main result is that evidence of population subdivision can be observed in genetic data even if the usual migration rate Nm is very large. In essence, a skewed offspring distribution leads to high probabilities of identity by descent within subpopulations and thus high F_{ST}. Therefore, patterns in genetic data indicating population subdivision cannot be taken to indicate low levels of gene flow in a population with a skewed offspring distribution. In fact, estimates of migration rate based on F_{ST} (or N_{ST}) are confounded by the reproduction parameters of our model of skewed offspring distribution. These results may explain the genetic heterogeneity among subpopulations of some marine species like the purple sea urchin (*S. purpuratus*; Edmands *et al.* 1996), despite the potential for wide dispersal of long-lived planktotrophic larvae (Burton 1983; Palumbi 1994).

METHODS AND RESULTS

Throughout we are concerned with neutral genetic diversity at a single nonrecombining locus in a haploid population. As usual, N is the population size. The results should hold for a diploid population with gametic migration if we replace N with $2N$. The population model we consider is a modification of the well-known Moran model of reproduction (Moran 1958, 1962). In the Moran model, a single randomly chosen individual reproduces each time step. To keep the population size constant a randomly chosen individual, but not the offspring, dies to make room for the offspring.

In our model, which was first presented in Eldon and Wakeley (2006), a single randomly chosen individual (the parent) reproduces each time step. With probability $1 - \varepsilon$ the parent has one offspring. Alternatively, with probability ε the parent has $\psi N - 1$ offspring (a large reproduction event) with $0 < \psi < 1$. To keep population size constant when a large reproduction event occurs, a total of $\psi N - 1$ individuals die to make room for the new offspring. In our model the parent always persists. The parameter ψ represents the fraction of the population that is replaced by the offspring of the parent. Eldon and Wakeley (2006) show that this modified Moran model of overlapping generations gives rise to a coalescent process that allows for asynchronous multiple mergers of ancestral lines, *i.e.*, is of the same type as the ancestral process considered by Pitman (1999) and Sagitov (1999).

For ease of presentation, we define the following quantities: N_γ, c_N, λ_γ, and I_A. The quantity N_γ is the coalescence timescale in our model. The coalescence timescale is proportional to the number of time steps, on average, it takes for two individuals to coalesce (in a single population). It depends on the value of ε that we assume has the form $\varepsilon \equiv 2\phi/N^\gamma$ for some constants ϕ and γ with $0 < \phi, \gamma < \infty$. In our model, the coalescence timescale is $N^\gamma/2$ time steps when $0 < \gamma < 2$. In the usual Moran model, the timescale is $N^2/2$ time steps, which is also the value of N_γ when $\gamma \geq 2$.

For a single population, Eldon and Wakeley (2006) show that different coalescent processes result depending on γ. Multiple mergers of ancestral lines are allowed in the coalescent process when $0 < \gamma \leq 2$, while Kingman's coalescent (Kingman 1982a,b) results when $\gamma > 2$. The probability that two individuals do coalesce in a single time step is denoted by c_N and depends on ε. The rate λ_γ of coalescence of two individuals is obtained from c_N by "speeding up" time by a factor of N_γ. When $0 < \gamma \leq 2$, λ_γ depends on the reproduction parameters ϕ and ψ. In mathematical notation, N_γ is expressed as $N_\gamma \equiv (1/2)\min(N^\gamma, N^2)$, and the coalescence probability c_N is

$$c_N = \frac{2(1-\varepsilon) + \psi N(\psi N - 1)\varepsilon}{N(N-1)}. \tag{1}$$

For notational convenience, we also define the indicator function I_A as

$$I_A \equiv \begin{cases} 1 & \text{if } A \text{ is true,} \\ 0 & \text{otherwise.} \end{cases}$$

For example, $I_{\gamma<2} = 1$ if $\gamma < 2$, and zero otherwise. In our model a large reproduction event occurs when the number of offspring of the parent equals $\psi N - 1$. These events occur with probability ε. Our choice of $\varepsilon = 2\phi/N^{\gamma}$ results in the coalescence timescale being N_γ. The rate λ_γ of coalescence is then

$$\lambda_\gamma \equiv \lim_{N\to\infty} N_\gamma c_N = I_{\gamma\geq 2} + \phi\psi^2 I_{\gamma\leq 2}. \tag{2}$$

The coalescence rate λ_γ is a key quantity in nearly all of our results below.

Model of subdivision: We now consider the finite island model of population subdivision with the simplifying assumption that migration does not change the sizes of the subpopulations (Nagylaki 1980; Strobeck 1987; Herbots 1997). Reproduction in all the subpopulations follows the modified Moran model described above. The discrete-time ancestral process for a sample of size 2 is a Markov chain with transition probabilities given in Equation A1 in the appendix.

We are concerned with small migration rates, specifically those on the order of $1/N_\gamma$ time steps. This means that a single individual resides in the same subpopulation for $2N_\gamma$ time steps, on average, before migrating to a different subpopulation. When $0 < \gamma < 2$, each individual resides in the same subpopulation for only N^γ time steps, on average. This time can be much shorter (when $0 < \gamma < 1$) than the usual average of N time steps assumed in Wright–Fisher populations. In other words, a large number of individuals migrate during N time steps when $0 < \gamma < 1$. We let m denote the probability that a single individual resided in a different subpopulation in the previous time step and model m as $m = m_\gamma \equiv \kappa/(2N_\gamma)$ in which κ is a finite constant $(0 < \kappa < \infty)$.

To illustrate the difference between our migration rate κ and the usual migration rate Nm let $M^* \equiv N^2 m_\gamma$ denote a migration rate scaled in units of N^2 time steps (or N generations). This corresponds to the usual "Nm" in the Wright–Fisher model. Substituting for m_γ gives $M^* = (I_{\gamma\geq 2} + N^{2-\gamma} I_{\gamma<2})\kappa$. If, for example, $\gamma = \frac{3}{2}$, then $M^* = \sqrt{N}\kappa$. When $\gamma < 2$ the migration rate M^* is very high; *i.e.*, $M^* \to \infty$ as $N \to \infty$ since κ is finite. However, in our modified model of reproduction coalescence also occurs on the timescale of $N^{3/2}$ time steps (or $\sqrt{N}$ generations when $\gamma = \frac{3}{2}$) and thus "counteracts" the effects of high migration rate.

The main results of this work concern expected coalescence times (Equations 3 and 5) and F_{ST}-like measures (Equations 10–12). We also derive the densities of the coalescence times (see appendix). The densities are used to derive distribution functions for the number of segregating sites between two sequences (see the appendix), which in turn yield expressions for F_{ST}-like measures including mutation (Equations 13 and 14).

The distributions of the coalescence times are functions of λ_γ: DNA sequences differ because they have accumulated mutations from the time of their most recent common ancestor until they are sampled. By assuming a very low mutation rate, Slatkin (1991) derived an expression for F_{ST} in terms of expected values of coalescence times. The time until two genes coalesce is therefore a fundamental quantity in theoretical work on structured populations. Given two genes sampled from a structured population, two different coalescence times arise that are of interest: the time T_0 until two genes sampled from the same subpopulation coalesce and time T_1 until two genes sampled from different subpopulations reach a common ancestor. The densities of T_0 and T_1 were previously derived under the structured coalescent by Takahata (1988) and Nath and Griffiths (1993) in the case of two subpopulations and by Herbots (1997) for any finite number of subpopulations.

Given the transition rates in Equation A2, we can obtain the distributions of the coalescence times T_0 and T_1 (see the appendix). Figure 1 shows the distributions of T_0 and T_1, respectively, as functions of time for different values of ψ (the fraction of the population replaced by the offspring of a single individual). As ψ increases (*i.e.*, tends to 1), the coalescence times T_0 and T_1 become very short.

The expected value and variance of T_0 are both less than the corresponding quantities for T_1. Specifically,

$$\begin{aligned} E(T_0) &= \frac{D}{\lambda_\gamma}, \\ E(T_1) &= \frac{D}{\lambda_\gamma} + \frac{D-1}{\kappa} \end{aligned} \tag{3}$$

and

$$\begin{aligned} \mathrm{Var}(T_0) &= \frac{D^2}{\lambda_\gamma^2} + \frac{2(D-1)^2}{\kappa\lambda_\gamma}, \\ \mathrm{Var}(T_1) &= \frac{D^2}{\lambda_\gamma^2} + \frac{2(D-1)^2}{\kappa\lambda_\gamma} + \frac{(D-1)^2}{\kappa^2}. \end{aligned} \tag{4}$$

Equation 3 holds a key result, namely that $E(T_0)$ is always less than $E(T_1)$.

The significance of the result in Equation 3 is best understood by an example. When $\gamma < 2$, say $\frac{3}{2}$, then the timescale is $N_\gamma = N^{3/2}$, and $\lambda_\gamma = \psi^2$ (assuming $\phi = 1$). Our migration parameter is then $\kappa = m_\gamma N_\gamma = m_\gamma N^{3/2}$. Migration is scaled in units of N^2 time steps in a standard Moran population. If we let $M^* \equiv N^2 m_\gamma$ be a scaled migration rate in units of N^2 time steps, then if m_γ is of order $1/N^{3/2}$ as above, M^* becomes very high in a large population. Specifically, since $\gamma = \frac{3}{2}$, we have $M^* = \kappa\sqrt{N} \to \infty$ (as $N \to \infty$), since κ is a constant. The result in Equation 3 says that even when $M^* \to \infty$ one will still see evidence of population structure in DNA sequence data, since coalescence occurs on a timescale of $N^{3/2} \ll N^2$

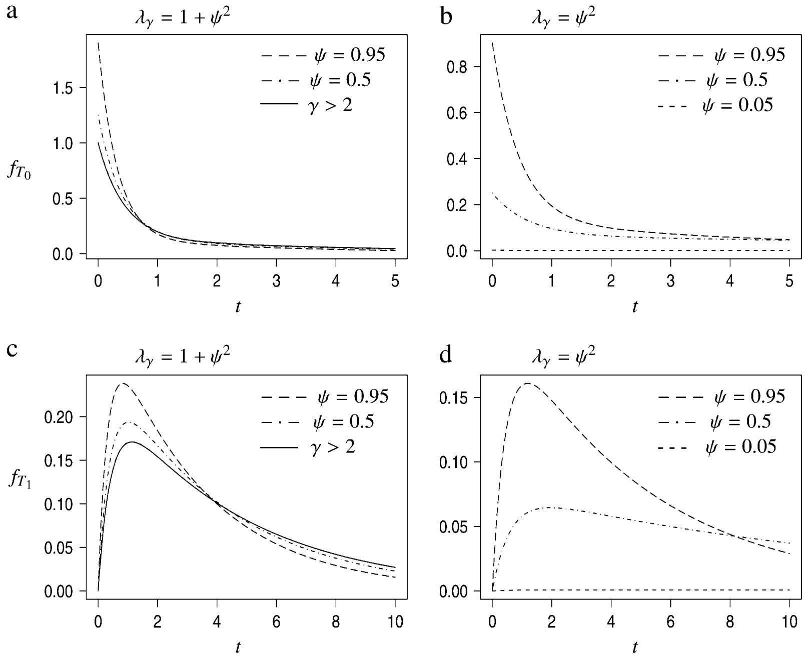

Figure 1.—The densities f_{T_0} and f_{T_1} of times to coalescence for two genes sampled from the same (T_0), or different (T_1), subpopulations as functions of time for different values of ψ when the number of subpopulations $D = 3$ and $\phi = \kappa = 1$. The coalescence timescale is $N^2/2$ in a and c and $N^\gamma/2$ with $0 < \gamma < 2$ in b and d. The solid lines in a and c are the densities obtained under the standard coalescent (*i.e.*, $\gamma > 2$).

time steps (in a large population) when $\gamma = \frac{3}{2}$. In fact, $M^* \to \infty$ as $N \to \infty$ whenever $0 < \gamma < 2$.

Similarly, $\mathrm{Var}(T_0)$ is always less than $\mathrm{Var}(T_1)$. In addition, the expected value and variance of T_0 are inversely proportional to λ_γ and thus will be small when the probability of large reproduction events is close to one. The expressions for $E(T_0)$ and $E(T_1)$ (Equation 3) obtained under the usual reproduction models (Nei and Feldman 1972; Li 1976; Griffiths 1981) can be recovered by assuming that large reproduction events occur on a longer timescale ($\gamma > 2$) than usual (*e.g.*, Wright–Fisher) sampling, in which case $\lambda_\gamma = 1$. The variances of T_0 and T_1 were first derived by Hey (1991) under the structured coalescent and can be recovered in the same way from Equation 4.

A many-demes limit: The structured coalescent simplifies under certain migration mechanisms when the number of subpopulations is taken to be much greater than the sample size of DNA sequences (Wakeley 1998). The convergence of the ancestral process under a many-demes limit (*i.e.*, when $D \to \infty$) follows from the work of Möhle (1998), which shows how events in a stochastic process that occur on different timescales can be separated (see the appendix for a more detailed description). We consider the ancestral process in the limit $D \to \infty$ and $N \to \infty$. Switching the order of the limits leads to the same coalescent process (see the appendix).

The limit process of two genes sampled from a population subdivided into very many subpopulations ($D \to \infty$), each of which is very large ($N \to \infty$), is of the form $\mathbf{P}^* e^{t\mathbf{G}^*}$ in which $\mathbf{P}^*$ and $\mathbf{G}^*$ are given by Equations A16 and A19, respectively. The form of $\mathbf{P}^*$ tells us that the ancestral process immediately enters the continuous-time process if the two genes are sampled from two different demes. If the two genes are sampled from the same subpopulation, they coalesce with probability $\lambda_\gamma/(\lambda_\gamma + \kappa)$ or enter the continuous-time process by moving to different subpopulations with probability $\kappa/(\lambda_\gamma + \kappa)$. In the continuous-time process the two lines wait with exponential time with rate $\kappa\lambda_\gamma/(\kappa + \lambda_\gamma)$ on a timescale of DN_γ time steps until they coalesce. The ancestral process under the many-demes limit model (Equation A19) differs from the limit process obtained when the number of subpopulations is finite (Equation A2), in that $\mathbf{G}^*$ has a zero entry for the transition where the two alleles enter the same subpopulation, after having been separated. When $D < \infty$, the corresponding rate is $\kappa/(D - 1)$ (Equation A2). Ancestral lines can coalesce, however, only if they reside in the same subpopulation. The matrix $\mathbf{B}^*$ (Equation A18) ensures that the two lines do arrive in the same subpopulation.

Again we are interested in the coalescence times T_0 and T_1 of two genes sampled from the same, or different, subpopulations, respectively. The distribution of T_0 is a mixture distribution (see appendix), and we obtain

$$\begin{aligned} E(T_0) &= \frac{1}{\lambda_\gamma}, \\ E(T_1) &= \frac{1}{\lambda_\gamma} + \frac{1}{\kappa} \end{aligned} \qquad (5)$$

and

$$\begin{aligned} \mathrm{Var}(T_0) &= \frac{1}{\lambda_\gamma^2} + \frac{2}{\kappa\lambda_\gamma}, \\ \mathrm{Var}(T_1) &= \frac{1}{\lambda_\gamma^2} + \frac{2}{\kappa\lambda_\gamma} + \frac{1}{\kappa^2}. \end{aligned} \qquad (6)$$

The expressions for the expected value and variance of T_0 and T_1 obtained under the many-demes limit model (Equations 5 and 6) are functions of λ_γ and κ in the same way as the corresponding expected values and variances (Equations 3 and 4) obtained for a finite number of subpopulations. In particular, we always expect a shorter coalescence time for two ancestral lines sampled from the same subpopulation than if they were sampled from different subpopulations.

Deriving F_{ST} and N_{ST}: The quantity F_{ST} is commonly used to assess levels of population subdivision. The inbreeding coefficient of an individual relative to a collection of subpopulations, F_{IT}, can be attributed to nonrandom mating within a subpopulation (F_{IS}) and to differences among subpopulations (F_{ST}; Wright 1951). Two sequences are identical by descent if they have not experienced mutation from the time of their most recent common ancestral sequence until they are sampled. If we let f_0 and f denote the probability of identity by descent of two genes sampled from the same subpopulation (f_0) and at random from the collection of subpopulations (f), we can express F_{ST} as

$$F_{\mathrm{ST}} = \frac{f_0 - f}{1 - f} \tag{7}$$

(Nei 1973). By the definition of F_{ST} in terms of inbreeding coefficients (as in Equation 7), F_{ST} depends on the mutation rate (μ). By forcing μ to be very low Slatkin (1991) derived an approximation of F_{ST} that is a function of expectation of coalescence times and is given by

$$F_{\mathrm{ST}}^{(0)} \equiv \lim_{\mu \to 0^+} F_{\mathrm{ST}} = \frac{E(T) - E(T_0)}{E(T)}, \tag{8}$$

in which T is the coalescence time of two lines randomly sampled from the collection of subpopulations, T_0 is the time to coalescence of two lines from the same subpopulation, and μ is the mutation rate.

To obtain an expression of $F_{\mathrm{ST}}^{(0)}$ in terms of coalescence times under skewed offspring distribution, we can proceed by first obtaining the expected coalescence time $E(T)$ of two genes randomly sampled from the collection of subpopulations, which is readily obtained from Equations 3 and A10 and is given by

$$E(T) = \frac{D}{\lambda_\gamma} + \frac{(D-1)^2}{D\kappa}. \tag{9}$$

When the number of subpopulations D is finite, the general form of $F_{\mathrm{ST}}^{(0)}$ is

$$F_{\mathrm{ST}}^{(0)} = \frac{1}{(\kappa/\lambda_\gamma)(D/(D-1))^2 + 1}. \tag{10}$$

For example, when $0 < \gamma < 2$, the rate of coalescence is $\lambda_\gamma = \psi^2$ (with $\phi = 1$) and Equation 10 gives $F_{\mathrm{ST}}^{(0)} = (\kappa\psi^{-2}(D/(D-1))^2 + 1)^{-1}$. The expression for $F_{\mathrm{ST}}^{(0)}$ in Equation 10 has the same form as the one derived by Slatkin (1991) under the standard coalescent. The key difference is that, under skewed offspring distribution, F_{ST} is a function of the rate λ_γ (Equation 2) of coalescence and thus a function of the reproduction parameters ϕ and ψ. The result that Slatkin (1991) obtained can be recovered from Equation 10 by taking $\gamma > 2$, in which case $\lambda_\gamma = 1$ (recall that the probability of large reproduction events $\propto 1/N_\gamma$).

When the number of subpopulations $D \gg 1$, we obtain from Equation 10

$$F_{\mathrm{ST}}^{(0)} \approx \frac{1}{1 + \kappa/\lambda_\gamma}. \tag{11}$$

In Equation 11 we have taken two limits: $N \to \infty$ and $D \to \infty$. Switching the order of the limits gives the same limit result for F_{ST} in Equation 11.

Following Wright (1951), the value of F_{ST} has often been used to estimate levels of gene flow. Figure 2 shows $\hat{\kappa} \equiv \lambda_\gamma((1/F_{\mathrm{ST}}) - 1)$, obtained from Equation 11, as a function of ψ for different values of F_{ST} (Figure 2a) and ϕ (Figure 2b) and for two different values of λ_γ. Since F_{ST} is a function of ψ and ϕ, so is any estimate of gene flow based on F_{ST}, as Figure 2 clearly shows.

Lynch and Crease (1990) used the number of pairwise sequence differences of DNA sequences to estimate levels of genetic heterogeneity. In that context, Lynch and Crease (1990) introduced the quantity N_{ST} that has the form $\hat{v}_1/(\hat{v}_1 + \hat{v}_0)$ in which $\hat{v}_1$ and $\hat{v}_0$ are the average number of pairwise differences between sequences sampled from different, or the same, subpopulations, respectively. If mutation rate is constant and mutations occur according to the infinite-sites model (Watterson 1975), then N_{ST} estimates $(E(T_1) - E(T_0))/E(T_1)$ (Slatkin 1993). Using the results obtained for expected coalescence times (Equation 3), we obtain $N_{\mathrm{ST}} = F_{\mathrm{ST}}^{(0)}$ as in Equation 11 for the many-demes limit model of population subdivision and

$$N_{\mathrm{ST}} = \frac{1}{1 + (\kappa/\lambda_\gamma)D/(D-1)} \tag{12}$$

when $D < \infty$. The effect of skewed offspring distribution is the same on N_{ST} as it is on F_{ST}. Under the infinite-sites mutation model we do not need an assumption of small mutation rate to obtain an expression of N_{ST} in terms of coalescence times, unlike the case for F_{ST}. As N_{ST} is defined, the mutation parameter cancels out (Slatkin 1993).

Number of segregating sites between pairs of sequences: By the definition of F_{ST} in terms of probabilities of identity by descent (Equation 7), F_{ST} depends on mutation. Eldon and Wakeley (2006) show that the limit process (as $N \to \infty$) of our model of skewed offspring distribution predicts nonzero levels of genetic variation only when $\gamma > 1$. If we (as in Eldon and Wakeley 2006) let μ denote the probability of mutation

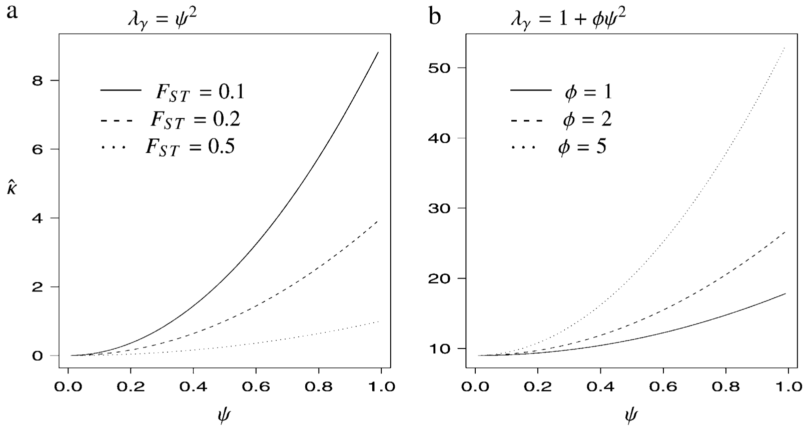

Figure 2.—The estimate $\hat{\kappa} \equiv \lambda_\gamma\ (1/(F_{\mathrm{ST}}) - 1)$ of migration rate from Equation 11 as a function of ψ. (a) $\psi^2(1/(F_{\mathrm{ST}}) - 1)$ when $F_{\mathrm{ST}} = 0.1$ (solid line), $F_{\mathrm{ST}} = 0.2$ (dashed line), and $F_{\mathrm{ST}} = 0.5$ (dotted line). (b) $(1 + \phi\psi^2)(1/(F_{\mathrm{ST}}) - 1)$ with $F_{\mathrm{ST}} = 0.1$ and $\phi = 1$ (solid line), $\phi = 2$ (dashed line), and $\phi = 5$ (dotted line).

for each offspring in a single time step, we define the mutation rate θ as $\theta \equiv \lim_{N\to\infty}(N_\gamma/N)\mu$ (and $\gamma > 1$). We can include mutation in an expression for F_{ST} by first obtaining the probability distributions of the number of segregating sites, under the infinite-sites model (Watterson 1975), between two genes given a model of population subdivision with migration. Let K_0 denote the number of segregating sites between two genes sampled from the same subpopulation and K denote the number of segregating sites between two genes sampled randomly from the collection of subpopulations. The distributions of K_0 and K are derived in the appendix, along with the distribution of the number of segregating sites K_1 between two genes sampled from different subpopulations. Then by the definition of F_{ST} given in Equation 7 we obtain

$$\begin{aligned} F_{\mathrm{ST}} &= \frac{P(K_0 = 0) - P(K = 0)}{1 - P(K = 0)} \\ &= \frac{1}{1 + (D/(D-1))^2(\kappa/\lambda_\gamma) + (D/(D-1))((\theta/2)/\lambda_\gamma)}. \end{aligned} \tag{13}$$

When $D \gg 1$,

$$F_{\mathrm{ST}} \approx \frac{1}{1 + \kappa/\lambda_\gamma + (\theta/2)/\lambda_\gamma}. \tag{14}$$

From Equation 14 we conclude that mutation can affect F_{ST} only if θ is large relative to λ_γ. The expression for F_{ST} in Equation 14 has the same form as the one derived by Wilkinson-Herbots (1998) and by Nei (1975) and Takahata (1983) by other methods, under the Wright–Fisher model, including mutation. In Figure 3, F_{ST} from Equation 14 is graphed as a function of ψ for different values of θ and κ. The interpretation of Figure 3 is that F_{ST}, as a function of ψ, can vary considerably when the timescale of coalescence (and migration) is in units of $N^\gamma/2$ generations with $1 < \gamma < 2$ (Figure 3, b and d).

Nei's genetic distance *d*: Not all indicators of separation between populations depend on λ_γ. Nei's (1972) genetic distance is more appropriate for estimating divergence time between species, and F_{ST}-like quantities are more suitable for inferring population structure within species (Slatkin 1991). Nei's (1972) genetic distance measure is given by $d_N = -\ln(f_1/f_0)$ in which f_0 and f_1 are the probabilities of identity by descent of two genes sampled from the same or different subpopulations, respectively, and we add the subscript N to remind us that time is discrete. If we now assume that $0 < \mu E(t_i) < 1$ for $i = 0, 1$, then using the Maclaurin series expansion of the logarithmic function $\ln(1 - \mu E(t_i))$ we obtain $d_N \approx \mu(E(t_1) - E(t_0))$ (previously obtained by Slatkin 1991) in which t_0 and t_1 are the coalescence times for two genes sampled from the same, or different, subpopulations, respectively. To obtain an expression of d for continuous time, we assume that the product $(N_\gamma/N)\mu$ converges to a constant θ as $N \to \infty$ (and $\gamma > 1$). Rewriting the approximation for d_N gives

$$d_N \approx \frac{N_\gamma}{N}\mu\left(E\left(\frac{t_1}{N_\gamma/N}\right) - E\left(\frac{t_0}{N_\gamma/N}\right)\right), \tag{15}$$

which has the continuous-time limit

$$d = \lim_{N\to\infty} d_N \approx \theta(E(T_1) - E(T_0)). \tag{16}$$

However, using the expressions for $E(T_1)$ and $E(T_0)$ (Equation 3), we obtain $d \approx \theta(D-1)/\kappa$ and so Nei's (1972) genetic distance is independent of λ_γ. Another way of deriving an expression for d is to note that the probability of identity by descent of two genes is the same as the probability that no mutations occur from the time they are sampled until they reach a common ancestor. Thus $f_i = P(K_i = 0)$ for $i = 0, 1$. We can therefore write

$$d = -\ln\left(\frac{P(K_1 = 0)}{P(K_0 = 0)}\right) \tag{17}$$

for any model of population subdivision. For the many-demes limit model under consideration,

$$d = -\ln\left(\frac{1}{1 + (\theta/2)/\kappa}\right).$$

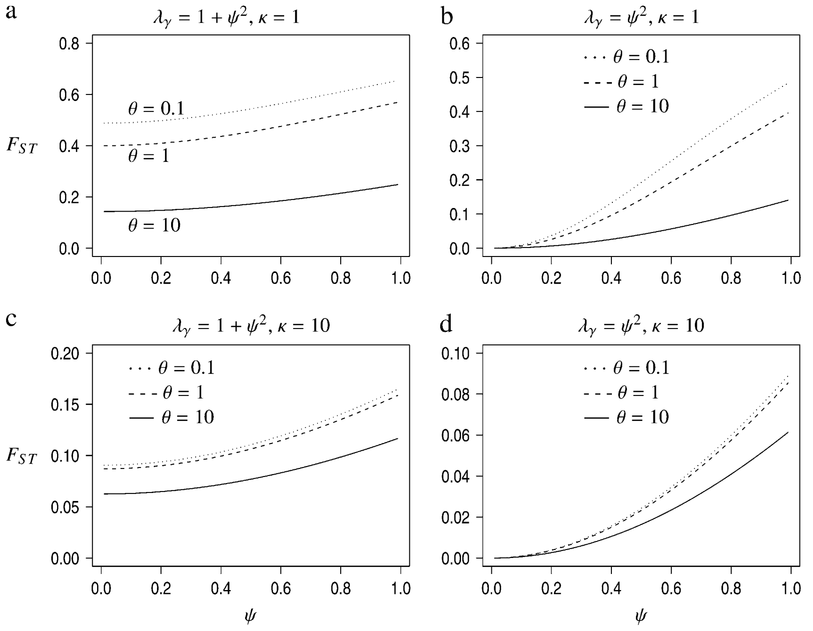

FIGURE 3.—The quantity F_{ST} from Equation 14 as a function of ψ (with $\phi = 1$) for different values of θ, κ, and rate of coalescence (λ_γ). Solid lines, $\theta = 10$; dashed lines, $\theta = 1$; dotted lines, $\theta = 0.1$.

Using either the limit approach (Equation 16) or the substitution approach (Equation 17) in the many-demes limit model, and assuming small θ/κ (*i.e.*, $0 < \theta/\kappa < 1$), d is of the form θ/κ. The same result is obtained for a finite number of subpopulations. Indeed, when D is finite, we obtain from Equations A28 and A29

$$\frac{P(K_1 = 0)}{P(K_0 = 0)} = \frac{1}{1 + (\theta/2)(D - 1)/\kappa}. \qquad (18)$$

Thus, if $0 < (\theta/2)(D - 1)/\kappa < 1$, we have from Equations 17 and 18

$$d \approx \frac{\theta}{2}\frac{D - 1}{\kappa}. \qquad (19)$$

Even if $(\theta/2)(D - 1)/\kappa > 1$, we have from Equations 17 and 18 that d is not a function of λ_γ. Thus NEI's (1972) genetic distance can be used to estimate divergence times of species even if one or both species have skewed offspring distribution, since d is proportional to the time of separation of two populations (NEI 1972; SLATKIN 1991).

DISCUSSION

Some organisms, for example Pacific oysters (BECKENBACH 1994; HEDGECOCK 1994a) and Atlantic cod (BEKKEVOLD *et al.* 2002; ÁRNASON 2004), may exhibit skewed offspring distribution among individuals in a population. Both BECKENBACH (1994) and HEDGECOCK (1994a) describe the reproductive mode of oysters, for example, as a lottery, in which only the offspring of a few lucky females survive. Oyster and cod females have very high reproductive potential, as they may produce millions of eggs in a single spawning (MAY 1967; STRATHMANN 1987; CHAMBERS and WAIWOOD 1996; KJESBU *et al.* 1996). The Wright–Fisher model does not capture the skewed offspring distribution possibly exhibited by organisms with high fecundities and high early mortality. The models of PITMAN (1999) and SAGITOV (1999), and later of ELDON and WAKELEY (2006) and SARGSYAN and WAKELEY (2008) for overlapping generations in a single population, incorporate the skewness and may thus better apply to organisms with highly fecund individuals and sweepstakes-style recruitment. By deriving distributions of coalescence times for two genes sampled from a subdivided population, we show how skewed offspring distribution confounds estimates of migration rate between subpopulations when based on F_{ST}-like measures of population subdivision.

An important result of this work is that F_{ST} depends not only on the migration rate κ but also on the parameters (ψ and ϕ) of our model of large offspring numbers. Demographic processes such as population size fluctuations, founder effects, or skewed offspring distribution have been thought to increase genetic differentiation among subpopulations. As defined and calculated from genetic data, common indicators of population subdivision then take on high values, thus suggesting low levels of migration (BOILEAU *et al.* 1992; WHITLOCK 1992; SLATKIN 1993; HEDGECOCK 1994a). Our main conclusions are twofold. First, F_{ST} is shown to depend on the parameters controlling the size (ψ) and frequency (ϕ) of large reproduction events (the probability that the offspring of a single individual replace a fraction ψ of the population is $\varepsilon = 2\phi/N^\gamma$) and can thus indicate high or low levels of genetic heterogeneity depending on ψ and ϕ. To illustrate, consider the expression for F_{ST} derived under the many-demes limit model without mutation (Equation 11), and let the timescale of coalescence occur on $N^{\gamma}/2$ time steps (*i.e.*, $0 < \gamma < 2$). In that case the rate of coalescence $\lambda_{\gamma} = \psi^2$ (by taking $\phi = 1$), and the rate of large reproduction events is high. By fixing κ, we see that F_{ST} ranges from very low (when ψ is low), to $\approx 1/(1 + \kappa)$, when $\psi \approx 1$. Second, to the extent that F_{ST} (or N_{ST}) is used in estimating levels of gene flow, these estimates are confounded by λ_{γ} and thus by ϕ and ψ. Also, migration in our model is not the usual *Nm* quantity, but is given by $\kappa = m_{\gamma}N_{\gamma}$. This means that even when *Nm* is very large, we may still observe genetic heterogeneity, since the rate of large reproduction events is also large. In a population where individuals can have very many offspring, gene flow is not the only demographic force that influences genetic heterogeneity.

The coalescence times T_0 and T_1 (for genes sampled from the same or different subpopulations, respectively) are fundamental quantities of the ancestral process of genes in subdivided populations. The time during which DNA sequences accumulate mutations is determined by T_0 and T_1. As we have shown, the coalescence times depend on the skewness of the offspring distribution through the rate λ_{γ} (Equation 2) of coalescence. By deriving the distributions of T_0 and T_1 for two genes in a structured population, we have obtained insight into how skewed offspring distribution shapes the genetics of structured populations. Since T_0 and T_1 are functions of ϕ and ψ through the rate of coalescence λ_{γ}, all the quantities of interest in regard to investigation of the genetics of structured populations, including expected values, number of segregating sites, and indicators of population subdivision, are functions of λ_{γ}.

One such insight is that genetic heterogeneity can be observed in genetic data even if gene flow is very high by the usual standard ($Nm \gg 1$). Edmands *et al.* (1996) found significant genetic heterogeneity among subpopulations of the purple sea urchin *S. purpuratus* sampled along the coast of California and Baja California. The ecology and physiology of *S. purpuratus* indicate the capacity for highly skewed offspring distribution: external fertilization and very high fecundity. Despite a planktonic larval period of several weeks (Strathmann 1978), and thus a potential for high dispersal, both allozyme and mtDNA sequence data revealed genetic differentiation, even over short distances (Edmands *et al.* 1996). We have shown that, regardless of the timescale of migration, $E(T_0) < E(T_1)$. Genetic heterogeneity can, therefore, be observed in DNA sequence data even if gene flow is very high, in a population with skewed offspring distribution. Population turnover in a metapopulation model when demes that become extinct are recolonized by one or a few individuals can also lead to increased F_{ST} (Wade and McCauley 1988; Whitlock and McCauley 1990; Pannell 2003). Indeed, a model of metapopulation structure that allows only one founder for every deme that is recolonized necessarily results in a coalescent process with multiple mergers, if the founder can have many offspring.

In summary, we consider the coalescence times of a subdivided population following a sweepstakes-style recruitment. The expected coalescence time for two genes sampled from the same subpopulation is always less than the expected coalescence time for two genes sampled from different subpopulations, even when migration occurs on a very short timescale. Estimates of migration rate based on F_{ST} are confounded by the rate λ_{γ} of coalescence, since F_{ST}-like measures of genetic heterogeneity are a function of the reproduction parameters of our model of skewed offspring distribution. These results underscore the importance of choosing an appropriate limit process for the population under consideration.

We thank two anonymous reviewers for helpful comments and suggestions.

LITERATURE CITED

Árnason, E., 2004 Mitochondrial cytochrome *b* variation in the high-fecundity Atlantic cod: trans-Atlantic clines and shallow gene genealogy. Genetics **166:** 1871–1885.

Árnason, E., P. H. Petersen, K. Kristinsson, H. Sigurgíslason and S. Pálsson, 2000 Mitochondrial cytochrome *b* DNA sequence variation of Atlantic cod from Iceland and Greenland. J. Fish Biol. **56:** 409–430.

Avise, J. C., 1994 *Molecular Markers, Natural History and Evolution.* Chapman & Hall, New York.

Avise, J. C., R. M. Ball and J. Arnold, 1988 Current versus historical population sizes in vertebrate species with high gene flow: a comparison based on mitochondrial DNA lineages and inbreeding theory for neutral mutations. Mol. Biol. Evol. **5:** 331–344.

Beckenbach, A. T., 1994 Mitochondrial haplotype frequencies in oysters: neutral alternatives to selection models, pp. 188–198 in *Non-Neutral Evolution*, edited by B. Golding. Chapman & Hall, New York.

Bekkevold, D. M., M. Hansen and V. Loeschcke, 2002 Male reproductive competition in spawning aggregations of cod (*Gadus morhua* l.). Mol. Ecol. **11:** 91–102.

Boileau, M. G., P. D. N. Hebert and S. S. Schwartz, 1992 Non-equilibrium gene frequency divergence: persistent founder effects in natural populations. J. Evol. Biol. **5:** 25–39.

Boom, J. D. G., E. G. Boulding and A. T. Beckenbach, 1994 Mitochondrial DNA variation in introduced populations of Pacific oyster, *Crassostra gigas*, in British Columbia. Can. J. Fish. Aquat. Sci. **51:** 1608–1614.

Burton, R. S., 1983 Protein polymorphisms and genetic differentiation of marine invertebrate populations. Mar. Biol. Lett. **4:** 193–206.

Cannings, C., 1974 The latent roots of certain Markov chains arising in genetics: a new approach. Adv. Appl. Probab. **6:** 260–290.

Chambers, R. C., and K. G. Waiwood, 1996 Maternal and seasonal differences in egg sizes and spawning characteristics of captive Atlantic cod, *Gadus morhua*. Can. J. Fish. Aquat. Sci. **53:** 1986–2003.

Crow, J. F., and M. Kimura, 1970 *Introduction to Population Genetics Theory.* Harper & Row, New York.

David, P., M. Perdieu, A. Pernot and P. J. Jarne, 1997 Fine-grained spatial and temporal population genetic structure in the marine bivalve *Spisula ovalis* l. Evolution **51:** 1318–1322.

Edmands, S., P. E. Moberg and R. S. Burton, 1996 Allozyme and mitochondrial DNA evidence of population subdivision in the purple sea urchin *Strongylocentrotus purpuratus*. Mar. Biol. **126:** 443–450.

Eldon, B., and J. Wakeley, 2006 Coalescent processes when the distribution of offspring number among individuals is highly skewed. Genetics **172:** 2621–2633.

Fisher, R. A., 1930 *The Genetical Theory of Natural Selection*. Clarendon Press, Oxford.

Griffiths, R. C., 1981 The number of heterozygous loci between two randomly chosen completely linked sequences of loci in two subdivided population models. J. Math. Biol. **12:** 251–261.

Hedgecock, D., 1994a Does variance in reproductive success limit effective population sizes of marine organisms?, pp. 1222–1344 in *Genetics and Evolution of Aquatic Organisms*, edited by A. Beaumont. Chapman & Hall, London.

Hedgecock, D., 1994b Temporal and spatial genetic structure of marine animal populations in the California Current. Calif. Coop. Ocean Fish. Invest. Rep. **35:** 73–81.

Hedgecock, D., M. Tracey and K. Nelson, 1982 Genetics, pp. 297–403 in *The Biology of Crustacea*, Vol. 2, edited by L. G. Abele. Academic Press, New York.

Hedrick, P. W., 2005 Large variance in reproductive success and the N_e/N ratio. Evolution **59:** 1596–1599.

Herbots, H. M., 1997 The structured coalescent, pp. 231–255 in *Progress of Population Genetics and Human Evolution*, edited by P. Donnelly and S. Tavaré. Springer, New York.

Hey, J., 1991 A multi-dimensional coalescent process applied to multi-allelic selection models and migration. Theor. Popul. Biol. **39:** 30–48.

Hurtado, L. A., R. A. Lutz and R. C. Vrijenhoek, 2004 Distinct patterns of genetic differentation among annelids of eastern Pacific hydrothermal vents. Mol. Ecol. **13:** 2603–2615.

Johnson, M. S., and R. Black, 1984 Pattern beneath the chaos: the effect of recruitment on genetic patchiness in an intertidal limpet. Evolution **38:** 1371–1383.

Johnson, S. B., C. R. Young, W. J. Jones, A. Warén and R. C. Vrijenhoek, 2006 Migration, isolation, and speciation of hydrothermal vent limpets (gastropoda; lepetodrilidae) across the Blanco Transform Fault. Biol. Bull. **210:** 140–157.

Kingman, J. F. C., 1982a The coalescent. Stoch. Proc. Appl. **13:** 235–248.

Kingman, J. F. C., 1982b On the genealogy of large populations. J. Appl. Probab. **19A:** 27–43.

Kjesbu, O. S., P. Solemdal, P. Bratland and M. Fonn, 1996 Variation in annual egg production in individual captive Atlantic cod (*Gadus morhua*). Can. J. Fish. Aquat. Sci. **53:** 610–620.

Li, W., 1976 Distribution of nucleotide difference between two randomly chosen cistrons in a subdivided population: the finite island model. Theor. Popul. Biol. **10:** 303–308.

Lynch, M., and T. J. Crease, 1990 The analysis of population survey data on DNA sequence variation. Mol. Biol. Evol. **7:** 377–394.

May, A. W., 1967 Fecundity of Atlantic cod. J. Fish. Res. Brd. Can. **24:** 1531–1551.

Möhle, M., 1998 A convergence theorem for Markov chains arising in population genetics and the coalescent with selfing. Adv. Appl. Probab. **30:** 493–512.

Moran, P. A. P., 1958 Random processes in genetics. Proc. Camb. Philos. Soc. **54:** 60–71.

Moran, P. A. P., 1962 *Statistical Processes of Evolutionary Theory*. Clarendon Press, Oxford.

Murray, M. C., and M. P. Hare, 2006 A genomic scan for divergent selection in a secondary contact zone between Atlantic and Gulf of Mexico oysters, *Crassostrea virginica*. Mol. Ecol. **15:** 4229–4242.

Nagylaki, T., 1980 The strong migration limit in geographically structured populations. J. Math. Biol. **9:** 101–114.

Nath, H. B., and R. C. Griffiths, 1993 The coalescent in two colonies with symmetric migration. J. Math. Biol. **31:** 841–852.

Nei, M., 1972 Genetic distance between populations. Am. Nat. **106:** 283–292.

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

Nei, M., 1975 *Molecular Population Genetics and Evolution*. Elsevier, New York.

Nei, M., 1982 Evolution of human races at the gene level, pp. 167–181 in *Human Genetics, Part A: The Unfolding Genome*, edited by B. Bohhe-Tamir, P. Cohen and R. N. Goodman. Alan R. Liss, New York.

Nei, M., and M. Feldman, 1972 Identity of genes by descent within and between populations under mutation and migration pressure. Theor. Popul. Biol. **3:** 460–465.

Nei, M., and D. Graur, 1984 Extent of protein polymorphism and the neutral mutation theory. Evol. Biol. **17:** 73–118.

Palumbi, S. R., 1994 Genetic divergence, reproductive isolation, and marine speciation. Annu. Rev. Ecol. Syst. **25:** 547–572.

Pannell, J. R., 2003 Coalescence in a metapopulation with recurrent local extinction and recolonization. Evolution **57:** 949–961.

Pitman, J., 1999 Coalescents with multiple collisions. Ann. Probab. **27:** 1870–1902.

Rousset, F., 2002 Inbreeding and relatedness coefficients: What do they measure? Heredity **88:** 371–380.

Sagitov, S., 1999 The general coalescent with asynchronous mergers of ancestral lines. J. Appl. Probab. **36:** 1116–1125.

Sargsyan, O., and J. Wakeley, 2008 A coalescent process with simultaneous multiple mergers for approximating the gene genealogies of many marine organisms. Theor. Popul. Biol. **74:** 104–114.

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., and W. P. Maddison, 1989 A cladistic measure of gene flow inferred from the phylogenies of alleles. Genetics **123:** 603–613.

Strathmann, M. F., 1987 *Reproduction and Development of Marine Invertebrates of the Northern Pacific Coast*. University of Washington Press, Seattle.

Strathmann, R. R., 1978 The length of pelagic period in echinoderms with feeding larvae from the northeastern pacific. J. Exp. Mar. Biol. Ecol. **34:** 23–27.

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

Takahata, N., 1983 Gene identity and genetic differentiation of populations in the finite island model. Genetics **104:** 497–512.

Takahata, N., 1988 The coalescent in two partially isolated diffusion populations. Genet. Res. Camb. **52:** 213–222.

Turner, T. F., J. P. Wares and J. R. Gold, 2002 Genetic effective size is three orders of magnitude smaller than adult size census size in an abundant, estuarine-dependent marine fish (*Sciaenops ocellatus*). Genetics **162:** 1329–1339.

Wade, M. J., and D. E. McCauley, 1988 Extinction and recolonization: their effects on the genetic differentiation of local populations. Evolution **42:** 995–1005.

Wakeley, J., 1998 Segregating sites in Wright's island model. Theor. Popul. Biol. **53:** 166–174.

Wakeley, J., 2008 *Coalescent Theory: An Introduction*. Roberts & Company, Greenwood Village, CO.

Watterson, G. A., 1975 On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. **7:** 256–276.

Watts, R. J., M. S. Johnson and R. Black, 1990 Effects of recruitment on genetic patchiness in the urchin *Echinometra mathaei* in Western Australia. Mar. Biol. **105:** 145–151.

Whitlock, M. C., 1992 Temporal fluctuations in demographic parameters and the genetic variance among populations. Evolution **46:** 608–615.

Whitlock, M. C., and D. E. McCauley, 1990 Some population genetic consequences of colony formation and extinction: genetic correlations within founding groups. Evolution **44:** 1717–1724.

Wilkinson-Herbots, H. M., 1998 Genealogy and subpopulation differentiation under various models of population structure. J. Math. Biol. **37:** 535–585.

Won, Y., C. R. Young, R. A. Lutz and R. C. Vrijenhoek, 2003 Dispersal barriers and isolation among deep-sea mussel populations (mytilidae: *Bathymodiolus*) from eastern Pacific hydrothermal vents. Mol. Ecol. **12:** 169–184.

Wright, S., 1931 Evolution in Mendelian populations. Genetics **16:** 97–159.

Wright, S., 1951 The genetical structure of populations. Ann. Eugen. **15:** 323–354.

Young, C. R., S. Fujio and R. C. Vrijenhoek, 2008 Directional dispersal between mid-ocean ridges: deep-ocean circulation and gene flow in *Ridgeia piscesae*. Mol. Ecol. **17:** 1718–1731.

APPENDIX

The discrete-time transition matrix: The probability transition matrix $\mathbf{\Pi}_N$ (Equation A1) for the ancestral process over one time step for the case of arbitrary fixed number $D \geq 2$ of subpopulations has three states: (1) two lines in the same deme but not coalesced, (2) two lines in different demes, and (3) the two lines have coalesced. We do not distinguish between subpopulations. The transition probabilities in $\mathbf{\Pi}_N$ are derived under the assumption that migration does not alter the subpopulation sizes (Nagylaki 1980; Strobeck 1987; Herbots 1997). We let m denote the single backward migration fraction. The matrix $\mathbf{\Pi}_N$ is

$$\mathbf{\Pi}_N = \begin{pmatrix} \left(M_{10} + M_{12}^{(s)}\right)(1 - c_N) & M_{11} + M_{12}^{(d)} & \left(M_{10} + M_{12}^{(s)}\right)c_N \\ \left(M_{21}^{(s)} + M_{22}^{(s)}\right)(1 - c_N) & M_{20} + M_{21}^{(d)} + M_{22}^{(d)} & \left(M_{21}^{(s)} + M_{22}^{(s)}\right)c_N \\ 0 & 0 & 1 \end{pmatrix} \tag{A1}$$

in which c_N is the coalescence probability and

$$M_{10} = \frac{N(1-m)(N(1-m)-1)}{N(N-1)}$$

$$M_{11} = \frac{2Nm(1-m)N}{N(N-1)}$$

$$M_{12}^{(s)} = \frac{Nm(Nm-1)/(D-1)}{N(N-1)}$$

$$M_{12}^{(d)} = \frac{Nm(Nm-1)(D-2)/(D-1)}{N(N-1)}$$

$$M_{20} = (1-m)^2$$

$$M_{21}^{(s)} = \frac{2m(1-m)}{D-1}$$

$$M_{21}^{(d)} = \frac{2m(1-m)(D-2)}{D-1}$$

$$M_{22}^{(s)} = \frac{m^2(D-2)}{(D-1)^2}$$

$$M_{22}^{(d)} = m^2\left(\frac{1}{D-1} + \frac{(D-2)^2}{(D-1)^2}\right).$$

The matrix in Equation A1 is a generalization of the matrix for the same migration mechanism (*cf.* Wakeley 2008) obtained under the usual Wright–Fisher model of reproduction. In Equation A1, the probability of coalescence is c_N, instead of the usual $1/N$, as is the case for the haploid Wright–Fisher model. The corresponding continuous-time process has rate matrix $\mathbf{G}$ given by

$$\mathbf{G} = \lim_{N\to\infty} N_\gamma(\mathbf{\Pi}_N - \mathbf{I}) = \begin{pmatrix} -\kappa - \lambda_\gamma & \kappa & \lambda_\gamma \\ \frac{\kappa}{D-1} & \frac{\kappa}{1-D} & 0 \\ 0 & 0 & 0 \end{pmatrix}. \tag{A2}$$

Distribution functions of the coalescence times when D is finite: In this section we derive the distribution functions for the coalescence times T_0 and T_1 when the number of subpopulations is finite. Given the distributions of T_0 and T_1 we can determine the distribution of the coalescence time T of two genes sampled at random from the collection of subpopulations. The distributions of these coalescence times allow us to derive expressions for F_{ST} with or without mutation.

We can use the rate matrix (Equation A2) to obtain the density functions for T_0 and T_1. Using Laplace transforms (see Herbots 1997) we obtain

$$f_{T_0}(t) = A_1 e^{r_1 t} + A_2 e^{r_2 t} \tag{A3}$$

in which

$$A_i = \frac{\kappa_\gamma\lambda_\gamma + (D-1)\lambda_\gamma r_i}{D-1}\,\frac{(-1)^i}{r_2 - r_1}, \quad i = 1, 2 \tag{A4}$$

and

$$r_i = \frac{-D\kappa_\gamma - (D-1)\lambda_\gamma + (-1)^i\sqrt{(D\kappa_\gamma + (D-1)\lambda_\gamma)^2 - 4(D-1)\kappa_\gamma\lambda_\gamma}}{2(D-1)} \tag{A5}$$

for $i = 1, 2$. To obtain the density function of T_1, we note that T_1 can be represented as a sum of two independent random variables, $T_1 = Y_1 + T_0$, where Y_1 is an exponential random variable with rate $\kappa_\gamma/(D-1)$. By direct calculation,

$$f_{T_1}(t) = \frac{A_1\kappa_\gamma}{\kappa_\gamma + (D-1)r_1}\left(e^{r_1 t} - e^{-\kappa_\gamma t/(D-1)}\right) + \frac{A_2\kappa_\gamma}{\kappa_\gamma + (D-1)r_2}\left(e^{r_2 t} - e^{-\kappa_\gamma t/(D-1)}\right). \tag{A6}$$

The form of the continuous functions f_{T_0} and f_{T_1} (Equations A3 and A6) immediately yields the cumulative densities F_{T_0} and F_{T_1} for T_0 and T_1, respectively. Namely, writing $\lambda_i = -r_i$ for $i = 1, 2$,

$$F_{T_0}(t) = \frac{A_1}{\lambda_1}(1 - e^{-\lambda_1 t}) + \frac{A_2}{\lambda_2}(1 - e^{-\lambda_2 t}) \tag{A7}$$

and

$$F_{T_1}(t) = \frac{B_1}{\lambda_1}(1 - e^{-\lambda_1 t}) + \frac{B_2}{\lambda_2}(1 - e^{-\lambda_2 t}) - \frac{(B_1 + B_2)(D-1)}{\kappa_\gamma}(1 - e^{-\kappa_\gamma t/(D-1)}) \tag{A8}$$

in which

$$B_i = \frac{A_i\kappa_\gamma}{\kappa_\gamma + (D-1)r_i}, \quad i = 1, 2.$$

Let T denote the time to coalescence for two genes sampled at random from the collection of subpopulations. Then, with probability $1/D$ the two genes are sampled from the same subpopulation, and with probability $1 - 1/D$ they are sampled from different subpopulations. The cumulative density function (c.d.f.) F_T of T is then given by

$$F_T(t) = \frac{1}{D}F_{T_0}(t) + \frac{D-1}{D}F_{T_1}(t) \tag{A9}$$

in which F_{T_i} denotes the c.d.f. of T_i for $i = 0, 1$. The expected value of T is

$$E(T) = E(T_0)/D + E(T_1)(D-1)/D \tag{A10}$$

in which $E(T_0)$ and $E(T_1)$ are given by Equation 3 and the variance of T is

$$\mathrm{Var}(T) = \frac{1}{D}\mathrm{Var}(T_0) + \frac{D-1}{D}\mathrm{Var}(T_1) + \frac{D-1}{D^2}(E(T_0) - E(T_1))^2 \tag{A11}$$

in which $\mathrm{Var}(T_0)$ and $\mathrm{Var}(T_1)$ are given by Equation 4. Note that although the expected value of T lies between $E(T_0)$ and $E(T_1)$, Equation A11 tells us that the variance of T may not lie between the variance of T_0 and T_1. If $D \gg 1$, then $\mathrm{Var}(T) > \mathrm{Var}(T_1)$.

A many-demes limit model: In this section we derive the ancestral process for two genes in the many-demes limit ($D \to \infty$) and as $N \to \infty$. Since now $D \to \infty$, the single-generation backward transition matrix, following Möhle (1998), can be written

$$\mathbf{\Pi}_D = \mathbf{A} + \frac{\mathbf{B}}{D} + O(1/D) \tag{A12}$$

in which the matrices **A** and **B** are given below. The matrix **A** describes the probabilities of the transitions that occur on a timescale of time steps. The matrix $\mathbf{B}/D$ describes transitions that occur on a timescale of D time steps, thus forming the continuous-time part of the ancestral process. The limit process (as $D \to \infty$) is then given by $\mathbf{\Pi}(t) = \mathbf{P}e^{t\mathbf{PBP}}$ in which $\mathbf{P} = \lim_{r\to\infty} \mathbf{A}^r$ describes the equilibrium process of the events that occur on the timescale of time steps (Möhle 1998). The ancestral history of a sample is first adjusted by an instantaneous process described by **P** and then enters a continuous-time process described by the rate matrix **PBP**. Given the ancestral process, we can derive the distributions of T_0 and T_1.

The instantaneous matrix **A** is

$$\mathbf{A} = \begin{pmatrix} \frac{(1-m)(N(1-m)-1)}{N-1}(1-c_N) & \frac{2Nm(1-m)+m(Nm-1)}{N-1} & \frac{(1-m)(N(1-m)-1)}{N-1}c_N \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{A13}$$

with eigenvalues $\lambda_1 = \lambda_2 = 1$ and

$$\lambda_3 = \frac{(1-c_N)(1-m)(N(1-m)-1)}{N-1}$$

and we observe that $\lim_{r\to\infty}\lambda_3^r = 0$. By calculating the corresponding eigenvectors (not shown) we obtain the limit matrix **P** by diagonalizing **A**,

$$\mathbf{P} = \lim_{r\to\infty} \mathbf{A} = \begin{pmatrix} 0 & P_{12} & 1-P_{12} \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \tag{A14}$$

in which

$$P_{12} = \frac{2mN - m - Nm^2}{Nc_N + 2Nm - c_N - m + c_N m - 2Nc_N m - Nm^2 + Nc_N m^2}, \tag{A15}$$

and in the limit $N \to \infty$ we obtain

$$\mathbf{P}^* = \lim_{N\to\infty} \mathbf{P} = \begin{pmatrix} 0 & \frac{\kappa}{\kappa+\lambda_\gamma} & \frac{\lambda_\gamma}{\kappa+\lambda_\gamma} \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{A16}$$

When $D \to \infty$ but holding N finite, the ancestral process is given by $\mathbf{P}e^{t\mathbf{PBP}}$, in which the matrix **B** is given by

$$\mathbf{B} = \begin{pmatrix} \frac{(1-c_N)m(mN-1)}{N-1} & \frac{m(mN-1)}{N-1} & \frac{c_N m(mN-1)}{N-1} \\ (1-c_N)(2-m)m & -(2-m)m & c_N(2-m)m \\ 0 & 0 & 0 \end{pmatrix} \tag{A17}$$

and we obtain

$$\mathbf{B}^* = \lim_{N\to\infty} N_\gamma \mathbf{B} = \begin{pmatrix} 0 & 0 & 0 \\ \kappa & -\kappa & 0 \\ 0 & 0 & 0 \end{pmatrix}. \tag{A18}$$

The rate matrix $\mathbf{G}^* = \mathbf{P}^*\mathbf{B}^*\mathbf{P}^*$ then takes the general format

$$\mathbf{G}^* = \begin{pmatrix} 0 & -\frac{\kappa^2\lambda_\gamma}{(\kappa+\lambda_\gamma)^2} & \frac{\kappa^2\lambda_\gamma}{(\kappa+\lambda_\gamma)^2} \\ 0 & -\frac{\kappa\lambda_\gamma}{\kappa+\lambda_\gamma} & \frac{\kappa\lambda_\gamma}{\kappa+\lambda_\gamma} \\ 0 & 0 & 0 \end{pmatrix}. \quad \text{(A19)}$$

The ancestral process in the limit $D \rightarrow \infty$ and $N \rightarrow \infty$ is $\mathbf{P}^* e^{t\mathbf{G}^*}$ and immediately yields the density functions for T_0 and T_1 as follows. The time T_1 to coalescence for two lines sampled from different demes follows in each case ($\gamma > 2$, $\gamma = 2$, and $0 < \gamma < 2$) an exponential distribution with rate $\kappa\lambda_\gamma/(\kappa + \lambda_\gamma)$. Now consider the time T_0 to coalescence for two lines sampled from the same subpopulation. Going back in time, two lines in the same subpopulation can either coalesce with probability $\lambda_\gamma/(\kappa + \lambda_\gamma)$ or they enter the continuous-time process with probability $\kappa/(\kappa + \lambda_\gamma)$, in which case one of the two lines migrates to a different subpopulation. Thus T_0 follows a mixture distribution with cumulative density function

$$F_{T_0}(t) = \frac{\lambda_\gamma}{\kappa+\lambda_\gamma} + \frac{\kappa}{\kappa+\lambda_\gamma}(1 - e^{-t\kappa\lambda_\gamma/(\kappa+\lambda_\gamma)}). \quad \text{(A20)}$$

Order of limits irrelevant in the many-demes limit model: In this section we show that the same ancestral process is obtained irrespective of the order of the limits $N \rightarrow \infty$ and $D \rightarrow \infty$. We have already derived the process when first $D \rightarrow \infty$ and then $N \rightarrow \infty$. Now we show that the same ancestral process is obtained when first $N \rightarrow \infty$ and then $D \rightarrow \infty$.

As $N \rightarrow \infty$, we obtain the ancestral process described by the rate matrix

$$\mathbf{G} = \mathbf{A} + \mathbf{B}/D = \begin{pmatrix} -\lambda_\gamma - \kappa & \kappa & \lambda_\gamma \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} + \frac{1}{D}\begin{pmatrix} 0 & 0 & 0 \\ \kappa & -\kappa & 0 \\ 0 & 0 & 0 \end{pmatrix}. \quad \text{(A21)}$$

We remark that $\mathbf{A}^n = (-\kappa - \lambda_\gamma)^{n-1}\mathbf{A}$ for $n \geq 1$. Since $\mathbf{A}$ is a rate matrix, we have, with $a = \kappa + \lambda_\gamma$,

$$e^{t\mathbf{A}} = \mathbf{I} + \sum_{n=1}^{\infty} \frac{t^n}{n!}\mathbf{A}^n = \mathbf{I} - \frac{1}{a}\sum_{n=1}^{\infty}\frac{(-at)^n}{n!}\mathbf{A} = \mathbf{I} + (e^{-at} - 1)\mathbf{A}. \quad \text{(A22)}$$

Equation A22 immediately gives us the instantaneous matrix

$$\mathbf{P} = \lim_{t\rightarrow\infty} e^{t\mathbf{A}} = \mathbf{I} - \frac{1}{a}\mathbf{A} = \begin{pmatrix} 0 & \frac{\kappa}{\kappa+\lambda_\gamma} & \frac{\lambda_\gamma}{\kappa+\lambda_\gamma} \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}. \quad \text{(A23)}$$

Equations A21 and A23 then give us the rate matrix $\boldsymbol{G} = \boldsymbol{PBP}$ after first taking the limit $N \rightarrow \infty$ and then $D \rightarrow \infty$.

By similar arguments we can show that the ancestral process does indeed result in coalescence regardless of initial state. Indeed,

$$\mathbf{G}^n = \left(-\frac{\kappa_\gamma\lambda_\gamma}{\kappa_\gamma + \lambda_\gamma}\right)^{n-1}\mathbf{G}, \quad n \geq 1, \quad \text{(A24)}$$

which gives, writing $b = \kappa\lambda_\gamma/(\kappa + \lambda_\gamma)$,

$$e^{t\mathbf{G}} = \mathbf{I} - \frac{1}{b}\sum_{n=1}^{\infty}\frac{(-bt)^n}{n!}\mathbf{G} = -\frac{1}{b}(e^{-bt} - 1)\mathbf{G}. \quad \text{(A25)}$$

The equilibrium distribution (as $t \rightarrow \infty$) is then

$$\mathbf{L} = \lim_{t\rightarrow\infty} e^{t\mathbf{G}} = \mathbf{I} + \frac{1}{b}\mathbf{G} \quad \text{(A26)}$$

and we obtain that two genes do reach a common ancestor regardless of initial state—*i.e.*,

$$\lim_{t\rightarrow\infty} \mathbf{P}e^{t\mathbf{G}} = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}. \quad \text{(A27)}$$

Number of segregating sites: We now consider the number of segregating sites, under the infinite-sites model (Watterson 1975), between two genes in the two models of population subdivision discussed previously. First, consider the model of finite number of subpopulations. Let K_0 and K_1 denote the number of segregating sites for two genes sampled from the same or different subpopulations, respectively. We let θ denote the scaled mutation rate and note that, given a specific length of time t, the number of segregating sites is Poisson distributed with rate $\theta t/2$. The probability distribution for the number of segregating sites for two genes sampled from the same subpopulation is

$$P(K_0 = k) = \frac{A_1}{\theta/2 + \lambda_1}\left(\frac{\theta/2}{\theta/2 + \lambda_1}\right)^k + \frac{A_2}{\theta/2 + \lambda_2}\left(\frac{\theta/2}{\theta/2 + \lambda_2}\right)^k \tag{A28}$$

in which A_1 and A_2 are given in Equation A4, and $\lambda_i = -r_i$ from Equation A5.

The probability distribution for the number of segregating sites between two genes sampled from different subpopulations is

$$\begin{aligned} P(K_1 = k) = {} & \frac{B_1}{\theta/2 + \lambda_1}\left(\frac{\theta/2}{\theta/2 + \lambda_1}\right)^k + \frac{B_2}{\theta/2 + \lambda_2}\left(\frac{\theta/2}{\theta/2 + \lambda_2}\right)^k \\ & - \frac{B_1 + B_2}{\theta/2 + \kappa_\gamma/(D-1)}\left(\frac{\theta/2}{\theta/2 + \kappa_\gamma/(D-1)}\right)^k, \quad k = 0, 1, \ldots, \end{aligned} \tag{A29}$$

in which $B_i = A_i\kappa/(\kappa + (D-1)r_i)$ for $i = 1, 2$. The expected value and variance of K_0 are both less than the corresponding quantities for K_1. Indeed,

$$E(K_0) = \frac{\theta}{2}\frac{D}{\lambda_\gamma} < \frac{\theta}{2}\left(\frac{D}{\lambda_\gamma} + \frac{D-1}{\kappa}\right) = E(K_1) \tag{A30}$$

and

$$\begin{aligned} \mathrm{Var}(K_0) & = \frac{\theta}{2}\frac{D}{\lambda_\gamma} + \frac{\theta^2}{4}\left(\frac{D^2}{\lambda_\gamma^2} + \frac{2(D-1)^2}{\kappa\lambda_\gamma}\right) < \mathrm{Var}(K_0) + \frac{\theta}{2}\left(\frac{D-1}{\kappa} + \frac{\theta}{2}\frac{(D-1)^2}{\kappa^2}\right) \\ & = \mathrm{Var}(K_1). \end{aligned} \tag{A31}$$

The probability mass distribution of the number of segregating sites K for two genes sampled at random from the collection of subpopulations is

$$\begin{aligned} P(K = k) = {} & \frac{C_1}{\theta/2 + \lambda_1}\left(\frac{\theta/2}{\theta/2 + \lambda_1}\right)^k + \frac{C_2}{\theta/2 + \lambda_2}\left(\frac{\theta/2}{\theta/2 + \lambda_2}\right)^k \\ & + \frac{C_3}{\theta/2 + \kappa/(D-1)}\left(\frac{\theta/2}{\theta/2 + \kappa/(D-1)}\right)^k, \end{aligned} \tag{A32}$$

in which $C_i = A_i/D + (D-1)B_i/D$ for $i = 1, 2$ and $C_3 = (1-D)(B_1 + B_2)/D$. One can write $K \sim (1/D)K_0 + (1-1/D)K_1$; *i.e.*, K is a mixture distribution. In fact, K is distributed as K_0 with probability $1/D$ and as K_1 with probability $1 - 1/D$. Hence,

$$E(K) = \frac{\theta}{2}\left(\frac{D}{\lambda_\gamma} + \frac{(D-1)^2}{D\kappa}\right) \tag{A33}$$

and so $E(K_0) < E(K) < E(K_1)$. However, the variance of K (Equation A34), which can be obtained in the same way as the variance of the time to coalescence of two genes sampled at random from the collection of subpopulations (Equation A11), may not lie between the variance of K_0 and K_1. The variance of K is

$$\begin{aligned} \mathrm{Var}(K) = {} & \frac{\theta}{4D^2\kappa^2\lambda_\gamma^2}(-\theta\lambda_\gamma^2 + D^4(\kappa + \lambda_\gamma)^2\theta + 2D(\kappa + \theta)\lambda_\gamma^2 \\ & + 2D^2\kappa\lambda_\gamma(\theta - 2\lambda_\gamma) + 2D^3\lambda_\gamma(\kappa^2 + \kappa(\lambda_\gamma - 2\theta) - \theta\lambda_\gamma)). \end{aligned} \tag{A34}$$

Number of segregating sites under the many-demes limit model: Under the many-demes limit model of population subdivision, the probability mass distribution for the number of segregating sites K_1 between two genes sampled from different subpopulations is

$$P(K_1 = k) = \frac{2\kappa\lambda_\gamma}{\theta(\kappa + \lambda_\gamma) + 2\kappa\lambda_\gamma}\left(\frac{\theta(\kappa + \lambda_\gamma)}{\theta(\kappa + \lambda_\gamma) + 2\kappa\lambda_\gamma}\right)^k. \tag{A35}$$

The expected number and the variance of the number of segregating sites between two genes sampled from different subpopulations is then

$$E(K_1) = \frac{\theta}{2}\left(\frac{\kappa + \lambda_\gamma}{\kappa\lambda_\gamma}\right) \tag{A36}$$

$$\mathrm{Var}(K_1) = \frac{\theta}{2}\left(\frac{\kappa + \lambda_\gamma}{\kappa_\gamma\lambda_\gamma}\right)\left(1 + \frac{\theta}{2}\left(\frac{\kappa + \lambda_\gamma}{\kappa\lambda_\gamma}\right)\right). \tag{A37}$$

The probability mass function for K_0, the number of segregating sites between two genes sampled from the same subpopulation, is ($k = 0, 1, \ldots$)

$$P(K_0 = k) = \frac{\lambda_\gamma}{\kappa + \lambda_\gamma} I_{k=0} + \frac{\kappa}{\kappa + \lambda_\gamma} P(K_1 = k), \tag{A38}$$

which gives expected value

$$E(K_0) = \frac{\theta}{2}\frac{1}{\lambda_\gamma}, \tag{A39}$$

and the variance of K_0 is

$$\mathrm{Var}(K_0) = \frac{\theta}{2}\frac{1}{\lambda_\gamma}\left(1 + \frac{\theta}{2}\left(\frac{\kappa + \lambda_\gamma}{\kappa\lambda_\gamma}\right)\right) + \frac{\theta^2}{4}\frac{1}{\kappa\lambda_\gamma}. \tag{A40}$$