Convergence to the Island-Model Coalescent Process in Populations with Restricted Migration

Frederick A. Matsen* and John Wakeley†,1

*Program for Evolutionary Dynamics and the Department of Mathematics, Harvard University, Cambridge, Massachusetts 02138

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

September 28, 2005

Running head: Convergence to the Island Model

Keywords: Coalescent, random walks on graphs, island migration model, stepping stone models, isolation by distance

[1]Corresponding author:

John Wakeley
2102 Biological Laboratories
16 Divinity Avenue
Cambridge, MA 02138
e-mail: wakeley@fas.harvard.edu
phone: 1-617-495-1564

ABSTRACT

In this paper we apply some graph-theoretic results to the study of coalescence in a structured population with migration. The graph is the pattern of migration among subpopulations, or demes, and we use the theory of random walks on graphs to characterize the ease with which ancestral lineages can traverse the habitat in a series of migration events. We identify conditions under which the coalescent process in populations with restricted migration, such that individuals cannot traverse the habitat freely in a single migration event, nonetheless becomes identical to the coalescent process in the island migration model in the limit as the number of demes tends to infinity. Specifically, we first note that a sequence of symmetric graphs with Diaconis-Stroock constant bounded above have an unstructured Kingman-type coalescent in the limit for a sample of size two from two different demes. We then show that circular and toroidal models with long-range but restricted migration have a upper bound on this constant, and so have an unstructured-migration coalescent in the limit. We investigate the rate of convergence to this limit using simulations.

1 Introduction

A classical dichotomy exists in population genetics between measurements of gene flow based on the observed movement of individuals and those based on patterns of genetic variation (Slatkin 1985). These direct and indirect measurements often disagree, and the possible explanations for this have been debated extensively. Here we are concerned with cases in which direct measurements of individual movement, or intuition based on the mobility of organisms, seems to predict that a correlation between genetic distance and geographic distance should be observed but indirect measurements of gene flow show no such correlation. We offer a new explanation for this, based on a surprising mathematical result. It contrasts with the prevailing notion, which was articulated recently by Ouborg *et al.* (1999) in a review of genetic studies of dispersal in plants: "If no isolation by distance is detected, then either dispersal is not distance dependent (which is, apart from very small spatial scales, unlikely in plants) or no equilibrium exists." The statement that no equilibrium exists means that changes in population size or structure must have occurred in the recent past.

Observations of the kind we consider do not abound in the literature, but there are some. Bohonak (1999) found significant structure among demes of water mites in the genus *Arrenurus* in North America, but found no correlation between genetic distance and geographic distance. While

genetic diversity in these species is likely to be strongly influenced by post-glacial population expansion, one of them (*A. birgei*) was suggested to be at or near migration-drift equilibrium. Durand *et al.* (2000) found low but significant population structure with no detectable isolation by distance in the widely distributed and long-term demographically stable African savanna grass species *Hyparrhenia diplandra*, whose dispersal ability is apparently low. Vandewoestijne *et al.* (1999) observed a high level of genetic diversity but a low level of structure, and no correlation between geographic distance and genetic distance, in the butterfly *Aglais urticae* for which very long distance dispersal is considered unlikely. A fourth example is Strand *et al.* (1996) who found high diversity and a high degree of population structure among local populations of plants in the genus *Aquilegia* in the southwestern United States and adjacent parts of Mexico, and used the lack of a correlation between genetic distance and geographic distance to argue for a history of isolation from a common ancestral population without subsequent gene flow.

We do not claim that historical events such as population expansion or the splitting of populations are not important factors. On the contrary, for many species these may be the most significant determinants of current patterns of diversity. We only point out that not all patterns of restricted migration are expected to produce a pattern of isolation by distance.

The more typical situation, in which there is some correlation between genetic distance and geographic distance, is easily explained. The movement of individuals or the dispersal of gametes is usually local, and the birth and death of individuals in a population of this sort establishes a pattern where relatedness tends to decrease with the distance between organisms. When DNA sequences or other genetic data are sampled from many individuals separated by different distances, a pattern of isolation by distance will be observed (Wright 1943). The models most commonly invoked to explain the correlation between genetic distance and geographic distance are the one-dimensional and two-dimensional stepping stone models (Kimura 1953; Kimura and Weiss 1964; Weiss and Kimura 1965; Maruyama 1970; Maruyama 1971; Sawyer 1976). In these models, the demes that make up the population are arrayed either on a line or on a two-dimensional lattice, and migration occurs only between neighboring demes.

Wright's island model and its variants (Wright 1931; Latter 1973; Maruyama 1974) are the only equilibrium migration models that have been invoked to explain a lack of correlation between genetic distance and geographic distance. In the island model, a single migration event is equally

likely to move the individual or gamete to any other deme in the population. At equilibrium, a sample taken from a single deme will show less genetic variation than a sample taken from more than one deme, but there will be no isolation by distance. Although the assumption of island-model migration is unrealistic for most species, estimates of gene flow are typically made under this assumption. Recent work has shown that the coalescent process, which is the retrospective process by which contemporary samples of genetic data trace back through ancestral lineages to reach common ancestors, is relatively simple in the island model when the number of demes is very large (Wakeley 1998). It is characterized by a slow process for between-deme samples, which is a Kingman-type coalescent process (Kingman 1982), and a fast process for within-deme samples, during which they either coalesce or end up in different demes (then enter the slow process).

Charlesworth *et al.* (2003) suggested that a similar two-phase coalescent process, in which the recent ancestry of a sample depends on the sample locations but the memory of these locations fades quickly as lineages trace their genealogy farther into the past, should be observed in the two-dimensional stepping stone model. Some recent theoretical work supports this. Wilkins (2004) identified this sort of behavior in an analysis of coalescence in a continuous two-dimensional habitat with symmetric Gaussian dispersal, although in that case the ancestral process is not necessarily an unstructured coalescent. The assumptions of the model we describe below are similar in spirit to those of Wilkins (2004) in that both allow individuals to migrate to a potentially very large number of locations. However, they differ in that our model is a discrete-deme model and does not assume a particular form for the distribution of dispersal distances.

Following an earlier paper of Cox and Durrett (2002) and making use of some special properties of random walks in two dimensions, Zähle *et al.* (2005) found a two-phase coalescent process in a two-dimensional stepping stone model with direct migration possible between demes separated by K steps or fewer along the lattice. While on the surface their model and results appear similar to ours, there are important differences. We seek an ancestral limit process which approximates the behavior of a population comprised of very many demes. In comparison to Zähle *et al.* (2005), who consider K to be a fixed, relatively small number compared to the total number of demes in the population, we assume that K is very large, on the order of the number of demes. The consequence of this is that the behavior of our model converges to that of the island model while theirs continues to predict a pattern of isolation by distance even in the limit as the number of demes tends to infinity.

2 Model and Results

Our model consists of D demes, each containing N haploid individuals. We will make use of the framework of graph theory, so we represent each deme as a node on a graph, where the edges are potential single-step migration paths. Any discrete-deme model with migration can be represented in this way. We will restrict ourselves to the case of vertex-transitive graphs, which we will simply call symmetric graphs. These graphs are homogeneous in the sense that they look the same from every node. Migration patterns of this sort have been called 'isotropic' in the population genetics literature (Strobeck 1987). An example is given in Figure 1. Clearly, the circular and toroidal stepping stone models are symmetric graphs. Trivially, the island model, which is represented by the complete graph, is also symmetric. Each node of a symmetric graph contacts the same number of edges. The number of edges that a single node contacts is called the degree of the symmetric graph, and is denoted d. Because each edge connects two nodes, the total number of edges in the graph is equal to $dD/2$.

We assume a continuous-time Moran model of reproduction. Each of the ND individuals in the population dies at some rate λ per unit time. When an individual is chosen to die, it is replaced by the offspring of an individual chosen uniformly at random either from the same deme, with probability $1 - m$, or from one of the d demes it is connected to in the graph, with probability m. In both cases, the same individual can be chosen to reproduce and to die. Consider the ancestry of a sample of size two in this model. When viewed backwards in time, the forward-time process of migration and reproduction becomes a coalescing random walk of ancestral lineages on the graph, with the caveat that the two ancestral lineages can occupy the same node without coalescing. The assumption of Moran-type reproduction simplifies the analysis somewhat because only one lineage can move at a time. Later, we consider the case of Wright-Fisher reproduction and the possibility that a sample of size greater than two is taken from the population.

Let the random variable X_{ij}, $i, j \in \{1, 2, \ldots, D\}$, be the time back to the most recent common ancestor for a pair of samples taken from deme i and deme j. We first consider the expected value $E[X_{ij}]$. Looking back in time, each lineage encounters its birth with rate λ, and is either a migrant or a non-migrant, with probabilities m and $1 - m$, respectively. By conditioning on the first event

in the continuous time Markov process that describes the ancestry of a sample, we have

$$E[X_{ii}] = \frac{1}{2\lambda} + (1-m)\left(1-\frac{1}{N}\right)E[X_{ii}] + m\sum_{j\in\Omega_i}\frac{1}{d}E[X_{ij}], \tag{1}$$

in which Ω_i is the set of labels of the d demes accessible by a single migration event from deme i. The terms on the right in Equation 1 are, from left to right: the waiting time for one or the other lineage to be the offspring in a reproduction event; the probability that the event is neither a migration event nor a coalescent event times the expected time given this; and the probability that the event is a migration event to a particular deme j times the expected time given this, summed over all j.

Because the graph is symmetric and every deme has the same size, $E[X_{ii}] = E[X_{jj}]$ for all i and j. The main result we present below is that when D is large, the time for a pair of lineages sampled in distinct demes to enter the same deme does not depend on the initial choice of demes. It follows that the distribution of X_{ij}, $i \neq j$, does not depend on i and j in the limit as D tends to infinity. Now, let us define τ_0 to be the average time for the pair of lineages to enter the same deme, thus giving them a chance to coalesce. For the expected value of X_{ij}, we can write

$$E[X_{ij}] \approx \tau_0 + \left(1-\frac{1}{N}\right)E[X_{ii}]. \tag{2}$$

This says that the expected coalescence time for a sample of size two from two different demes is equal to the time τ_0 plus the probability they do not immediately coalesce when they enter the same deme times the expected coalescence time given they are now in the same deme. Slatkin (1987) and Strobeck (1987) showed that the expected within-deme pairwise coalescence time under symmetric migration is the same as the expected pairwise coalescence time in a single, panmictic population of the same total size. In our model, this is $E[X_{ii}] = ND/(2\lambda)$. Substituting Equation 2 into Equation 1 and using $E[X_{ii}] = ND/(2\lambda)$ to then solve for τ_0 gives $\tau_0 \approx D/(2\lambda m)$, where the approximation is valid if D is large. Using Equation 2 again, we obtain

$$E[X_{ij}] \approx \frac{ND}{2\lambda}\left(1+\frac{1-m}{Nm}\right), \tag{3}$$

which is the same as the result for the island model with a large number of demes and Moran-type reproduction; *e.g.* see Slade and Wakeley (2005).

We use the Diaconis-Stroock bound (Diaconis and Stroock 1991) to prove that the limit ($D \to \infty$) distribution of X_{ij}, $i \neq j$, does not depend on i and j. The crucial fact of the analysis

is that it is possible for a random walk on a graph to have a finite "mixing time" even when the number of nodes of the graph goes to infinity. The mixing time, denoted τ_2, quantifies the amount of time it takes for a Markov chain to near the stationary distribution. If the mixing time is short compared to the amount of time before a coalescent event, then the Markov chain acts close to as if it had been started with the stationary distribution. In particular, the initial location of the samples is irrelevant.

A bound for the mixing time is provided by the Diaconis-Stroock bound. To apply this theory, one chooses a "distinguished set of paths" Γ connecting every ordered pair of nodes in the state space. For this Γ one finds L, which is the length of the longest path, and B, the maximum number of paths going through a given edge, to gain the bound for our symmetric graph

$$\tau_2 \leq \frac{LBd}{D}.$$

In the cases of interest to us, we will be able to find a set of paths which have maximal path length L bounded above and the maximal number of paths going through an edge B will be bounded above. The degree d is always less than or equal to the number of nodes D. In biological terms we might think of this as saying that spatial structure is unimportant in the limit if first, it is possible to get across the space in a small number of migration events, and second, there are no narrow channels through which migration must occur. We also note that the bound (2) is the simplest example of the paths technique; further development can be found in Saloff-Coste (1997).

Using this mixing time bound we can show using a theorem of Aldous (1989) that the distribution of the time for any pair of lineages to enter the same deme converges to an exponential distribution, with mean equal to one when time is rescaled appropriately. The scaling is by the average time τ_0, introduced above, and part of the proof is to show that τ_0 converges to $D/(2\lambda m)$ as D increases. The details of the proof are given in the Appendix. Briefly, we note that the random walk of two lineages on a symmetric graph can be treated by fixing the position of one of them, and letting the other move with twice the rate, in this case $2\lambda m$. We then label the nodes of the graph so that deme one is the fixed position of the first lineage, and define Y_i to be the time for the moving lineage to reach deme one given that it starts in deme i. We use H_i to denote the expected value $E[Y_i]$. The overall average is $\tau_0 = \frac{1}{D}\sum_{i=2}^{D} H_i$. The interesting result is that the distribution of Y_i/τ_0 converges to an exponential distribution with mean equal to one in the limit as D tends to infinity.

Figure 2 illustrates the behavior of the scaled expected times H_i/τ_0 for a series of graphs like the one in Figure 1. To produce Figure 2, we calculated H_i analytically using the spectral formula for hitting times (Equation 9) and the properties of circulant matrices (Davis 1979). All H_i/τ_0 should be equal to one in the limit we consider. The figure shows H_i/τ_0 for a series of circular graphs with increasing D, where each node is connected to all nodes that are within 1/10 of the graph away from it. This corresponds to a species in which an individual can move to any deme that is within 10% of the total distance around its circular habitat. When the number of demes is small there is a substantial difference between samples that are close together and those that are far apart. However, when the number of demes becomes large this difference decreases. As might be expected, there is also a visible jump at $i/D = 1/10$, so that samples within migration-range of each other have shorter times than more distant samples, but the magnitude of this jump becomes negligible in the limit as D tends to infinity.

The graphs in Figures 1 and 2 can be thought of as extensions of the familiar circular stepping stone model (Kimura and Weiss 1964), but where we have added some extra branches representing longer-range migration. We also consider a toroidal model which is the corresponding extension of the two-dimensional stepping stone model. Conditions under which the toroidal model will have mixing time bounded above, and will thus fall under our convergence result, are identified in the Appendix. In both the circular case and the toroidal case, it is sufficient that a migrant can migrate freely within a neighborhood of non-vanishing size (measured as a fraction of the total number of demes) as D tends to infinity. The neighborhood need not be large, so migration is restricted even in the limit model. For example, in Figures 1 and 2 it will take at least ten migration steps to move once around the habitat.

We can now return to the coalescence time, X_{ij}, for two samples starting in demes i and j. In the limit as D tends to infinity, we have shown that the waiting time for the two lineages to enter the same deme does not depend on i and j, and is exponentially distributed with mean equal to one when time is rescaled by the factor $D/(2\lambda m)$. When they first enter the same deme, there is a chance $1/N$ that the two lineages coalesce. If they do not coalesce immediately (probability $1-1/N$) then after some number of reproduction events they will either coalesce, with probability

$$\frac{2\lambda(1-m)/N}{2\lambda m + 2\lambda(1-m)/N} = \frac{1-m}{Nm+1-m}, \tag{4}$$

or one of the lineages will migrate out of the deme. The time it takes for one or the other of these

events to occur will have mean equal to $N/(2\lambda(Nm+1-m))$, and when D is large this will be much less than the average time $D/(2\lambda m)$ for the pair to meet in the same deme.

Thus, we can appeal to the "separation of time scales" between this within-deme process and the process of migration-movement of lineages across the population, treated above and in the Appendix. We can show that the distribution of X_{ij} in the limit is also exponential with mean equal to one when it is rescaled appropriately. Overall, the probability of coalescence given the two lineages enter the same deme is equal to

$$\frac{1}{N} + \left(1 - \frac{1}{N}\right) \frac{1-m}{Nm+1-m} = \frac{1}{Nm+1-m}. \tag{5}$$

Again, if they do not coalesce, then one or the other lineage will migrate to a different deme. The number of times the two lineages will have to repeat this process of entering the same deme and having a chance to coalesce before they finally do coalesce will be geometrically distributed with parameter equal to Equation 5. MÖHLE (1998) has developed a formal method for treating Markov processes with two time scales which applies here, but also admits more generality.

Our result that the time for a pair of lineages to enter the same deme is exponentially distributed requires that time is rescaled by τ_0, which recall converges to $D/(2\lambda m)$ as D tends to infinity. On this timescale and as $D \to \infty$, the durations of the periods when the lineages are together in the same deme become negligible. Therefore the distribution of $X_{ij}2\lambda m/D$ is given by the sum of a geometric number of exponential random variables which can be shown, *e.g.* as in WAKELEY (1999), to be exponential with mean equal to $Nm + 1 - m$. If we rescale time again, by this new factor, so that our new unit of time is equal to

$$N_e := \frac{D}{2\lambda m}(Nm + 1 - m) = \frac{ND}{2\lambda}\left(1 + \frac{1-m}{Nm}\right) \tag{6}$$

of the original units, then the limit distribution of the scaled coalescence time $T_{ij} = X_{ij}/N_e$ will exponential with mean equal to one. Note that our "effective population size" N_e is identical to the expression for $E[X_{ij}]$ given in Equation 3.

We have followed WAKELEY (1999) in defining N_e so that the time-rescaled coalescent process for a sample of lineages from *different* demes is given by Kingman's coalescent. Samples from the same deme will undergo an instantaneous process of migration and coalescence, called the "scattering phase" in WAKELEY (1999). A pair of lineages sampled (without replacement) from a single deme will coalesce with probability given by Equation 4. If they do not coalesce, then one or the other

will migrate and they will enter the Kingman coalescent, or "collecting phase." The difference in the distributions of coalescence times for within-deme versus between deme samples can be seen in their respective cumulative distribution functions (CDFs):

$$P\{T_{ii} < t\} = 1 - \frac{Nm}{Nm + 1 - m} e^{-t} \tag{7}$$

$$P\{T_{ij} < t\} = 1 - e^{-t}. \tag{8}$$

In the limit model, single-deme samples of size two have a probability mass of $(1-m)/(Nm+1-m)$ at $t = 0$, followed by the usual exponential decay for $t > 0$.

We used simulations to assess the convergence of the rescaled coalescence time X_{ij}/N_e to the exponential distribution. The source code of a program which simulates the exact model is available from the authors upon request. Some results are shown in Figure 3, which compares the CDF of X_{ij}/N_e in simulations to the limit results given in Equations 7 and 8, for a series of increasing D. Figure 3 presents results for samples of size two: (a) from the same deme, (b) from adjacent demes, and (c) from maximally-distant demes. The demes were arrayed on an $l \times l$ torus and we allowed direct migration to any deme within $l/16$ steps from the current deme in either dimension. When $D = l^2$ is large, this corresponds to a species in which individuals can migrate to any deme in an area equal to 1/64, or about 1.56%, of the total, two-dimensional, species range. We set $\lambda = 1$, $N = 50$, and $m = 0.02$, and considered $l = 16, 32, 62, 128$. As the number of demes increases, the CDFs for adjacent and maximally-distant samples converge to that of the same exponential distribution, with mean equal to one and given by Equation 8. The CDF for a single-deme sample converges to Equation 7. The theory presented above and in the Appendix assures convergence in the limit, but Figure 3 illustrates that the limit result can be approximately true even when D is only moderately large.

3 Discussion

We have shown that the distribution of coalescence times for a sample of size two from a subdivided population with restricted migration may not depend on the distance between sampling locations. This result holds in the limit as the number of demes tends to infinity, and with some restrictions on the pattern of migration. Samples from the same deme differ from samples from different demes

— on average having shorter coalescence times — but in the limit this will be the only evidence of subdivision. Thus, although migration may be fairly restricted, the usual pattern of isolation by distance may not be observed. We suspect that this will be important for only a limited number of species, but we hope that the result adds something to the ongoing debate about the role of gene flow in structuring genetic variation.

Our result is similar to other recent results (Wakeley and Aliacar 2001) and can be understood with reference to the strong-migration limit of Nagylaki (1980), which was taken up in a genealogical setting by Notohara (1993). In particular, when the number of demes is large in our model, the memory of the original sampling locations of lineages is lost quickly in comparison to the rate at which their common ancestor is reached. For this reason, it seems reasonable to speculate that the same kind of result will hold for populations with asymmetric migration patterns (*e.g.*, populations with edges), populations with different migration rates between different pairs of demes, populations with other kinds of reproduction (*e.g.*, Wright-Fisher reproduction), as well as for samples of size greater than two.

In fact, we have verified the last two of these predictions using simulations, although we do not present the results. However, we note that the rate of convergence to Kingman's result (as D increases) for the CDF of time to the first coalescent in the sample decreases as the sample size increases. This is to be expected since the validity of Kingman's coalescent depends roughly on the square of the sample size being much smaller than the effective size of the population (Kingman 1982). Of course, studies of isolation by distance are nearly always based on pairwise comparisons between samples, and our results for samples of size two hold marginally for pairs in samples of any size.

We thank David Aldous and Rick Durrett for helpful comments. F.A.M. was supported by a Graduate Research Fellowship from the National Science Foundation. J.W. was supported by a Presidential Early Career Award for Scientists and Engineers (DEB-0133760) from the National Science Foundation and by a Fellowship from the Radcliffe Institute for Advanced Study.

LITERATURE CITED

Aldous, D., 1989 Hitting times for random walks on vertex-transitive graphs. Math. Proc. Camb. Phil. Soc. **106:** 179–191.

Bohonak, A. J., 1999 Effect of insect-mediated dispersal on the genetic structure of postglacial water mite populations. Heredity **82:** 451–461.

Charlesworth, B., D. Charlesworth and N. Barton, 2003 The effects of genetic and geographic structure on neutral variation. Annu. Rev. Ecol. Evol. Syst. **34:** 99–125.

Cox, J. T., and R. Durrett, 2002 The stepping stone model: New formulas expose old myths. Ann. Appl. Probab. **12:** 1348–1377.

Davis, P. J., 1979 *Circulant Matrices.* Wiley Interscience, New York.

Diaconis, P., and D. Stroock, 1991 Geometric bounds for eigenvalues of Markov chains. Ann. Appl. Prob. **1:** 36–61.

Durand, J., L. Garnier, I. Dajoz, S. Mousset and M. Veuille, 2000 Gene flow in a facultative apomictic Poacea, the savanna grass *Hyparrhenia diplandra.* Genetics **156:** 823–831.

Kimura, M., 1953 "Stepping stone" model of population. Ann. Rept. Nat. Inst. Genetics, Japan **3:** 62–63.

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

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

Latter, B. D. H., 1973 The island model of population differentiation: a general solution. Genetics **73:** 147–157.

Lovasz, L., 1993 Random walks on graphs: a survey, pp. 1–46 in *Paul Erdös is Eighty* volume 2. Jànos Bolyai Mathematical Society, Budapest, Hungary.

Maruyama, T., 1970 Analysis of population I. One-dimensional stepping-stone models of finite length. Ann. Hum. Genet., Lond. **34:** 201–219.

MARUYAMA, T., 1971 Analysis of population II. Two-dimensional stepping-stone models of finite length and other geographically structured populations. Ann. Hum. Genet., Lond. **35:** 179–196.

MARUYAMA, T., 1974 A simple proof that certain quantities are independent of the geographical structure of population. Theoret. Pop. Biol. **5:** 148–154.

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

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

NOTOHARA, M., 1993 The strong migration limit for the genealogical process in geographically structured populations. J. Math. Biol. **31:** 115–122.

OUBORG, N. J., Y. PIQUOT and J. M. VAN GROENENDAEL, 1999 Population genetics, molecular markers and the study of dispersal in plants. J. Ecology **87:** 551–568.

SALOFF-COSTE, L., 1997 Lectures on finite Markov chains, pp. 301–413 in *Lectures on Probability Theory and Statistics*. Springer, Berlin Lecture notes in mathematics 1665.

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

SLADE, P. F., and J. WAKELEY, 2005 The structured ancestral selection graph and the many-demes limit. Genetics **169:** 1117–1131.

SLATKIN, M., 1985 Gene flow in natural populations. Ann. Rev. Ecol. Syst. **16:** 393–430.

SLATKIN, M., 1987 The average number of sites separating DNA sequences drawn from a subdivided population. Theoret. Pop. Biol. **32:** 42–49.

STRAND, A. E., B. G. MILLIGAN and C. M. PRUITT, 1996 Are populations islands? Analysis of chloroplast DNA variation in *Aquilegia*. Evolution **50:** 1822–1829.

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

VANDEWOESTIJNE, S., G. NÈVE and M. BAGUETTE, 1999 Spatial and temporal population genetic structure of the butterfly *Aglais urticae* l. (Lepidoptera, Nymphalidae). Mol. Ecol. **8:** 1539–1543.

WAKELEY, J., 1998 Segregating sites in Wright's island model. Theoretical Population Biology **53:** 166–174.

WAKELEY, J., 1999 Non-equilibrium migration in human history. Genetics **153:** 1863–1871.

WAKELEY, J., and N. ALIACAR, 2001 Gene genealogies in a metapopulation. Genetics **159:** 893–905 [Corrigendum (Figure 2): *Genetics* 160:1263-1264].

WEISS, G. H., and M. KIMURA, 1965 A mathematical analysis of the stepping stone model of genetic correlation. J. Appl. Probab. **2:** 129–149.

WILKINS, J., 2004 A separation-of-timescales approach to the coalescent in a continuous population. Genetics **168:** 2227–2244.

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

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

ZÄHLE, I., J. T. COX and R. DURRETT, 2005 The stepping stone model II: genealogies and the infinite sites model. Ann. Appl. Probab. **15:** 671–699.

Appendix

In this section we present and prove the technical results for the paper. For simplicity we will consider random walks with rate one rather than walks with rate $2\lambda m$ as in the main text. The expectations of the time to enter the same deme (the hitting time) scale appropriately; for example if a given hitting time is H for the rate one case, the corresponding hitting time would be $H/(2\lambda m)$ in the previously considered case. Also, in order to conform with the large and well established literature of graph theory and random processes on graphs, we will use n to denote the number of nodes of a graph rather than D, which is used in the main text.

First we recall some basic facts about random walks on graphs. Let M denote the Markov transition matrix for the random walk on our symmetric graph. M is symmetric and thus can be diagonalized with an orthogonal matrix V with entries v_{ij}. Let us order the eigenvalues $1 = \lambda_1 > \lambda_2 \geq \cdots \geq \lambda_n \geq -1$. The eigenvalues are bounded above by one in absolute value because large powers of the Markov matrix are bounded.

There is a nice formula for expected hitting times in terms of these eigenvalues and vectors:

$$H_i = n \sum_{k \geq 2} \frac{1}{1 - \lambda_k} \left(v_{1k}^2 - v_{ik} v_{1k} \right). \tag{9}$$

Using the orthogonality of the matrix V gives

$$\tau_0 = \sum_{k \geq 2} \frac{1}{1 - \lambda_k}. \tag{10}$$

where τ_0 is again the average hitting time. For proofs see, Lovasz (1993) section 3. The aforementioned mixing time is defined in terms of the second largest eigenvalue: $\tau_2 = (1 - \lambda_2)^{-1}$. As mentioned above, the mixing time quantifies the amount of time the chain will take to near the stationary distribution. Under our hypotheses the mixing time is bounded above, which is the crucial fact that allows our conclusions.

The Diaconis-Stroock bound is found in the literature as Proposition 1 of (Diaconis and Stroock 1991):

Theorem 1

$$\lambda_2 \leq 1 - 1/\kappa \tag{11}$$

where κ is a constant which contains information about a chosen set of "distinguished allowable paths" Γ connecting every pair of states in the Markov chain. In the case of a random walk on a

symmetric graph, Γ is simply a chosen set of paths on the graph connecting every pair of nodes. Furthermore, as a special case of Corollary 1 of (DIACONIS and STROOCK 1991), κ can be bounded above by LBd/n, where L is the longest path in Γ and B is the maximal number of paths which traverse a given edge e. As before, d is the degree and n is the number of nodes in the graph.

If we can find an upper bound for λ_2 which is independent of n, we can apply the following proposition, proven below.

Proposition 1 *Let $G_1, G_2, \ldots$ be a sequence of symmetric graphs where G_n has at least n nodes. Assume*

$$\lambda_2(G_n) \leq \Delta. \tag{12}$$

for some $0 < \Delta < 1$. Assume that $i(n)$ is a sequence of nodes such that $i(1)$ is a node of G_1, $i(2)$ is a node of G_2, and so on. Then the distribution of $Y_{i(n)}/\tau_0$ converges to the exponential distribution with mean 1 in the limit of n going to infinity.

Putting these two together, we would like to have a upper bound for κ which is independent of n. The following proposition gives such a bound for the torus with uniform migration across a given fraction of the habitat:

Proposition 2 *Define $G_{\theta,l}$ to be $l \times l$ toroidal lattice with an additional edge between any two nodes whose horizontal and vertical distance is less than or equal to $\lceil \theta l \rceil + 1$. There exists a set of distinguished paths Γ on $G_{\theta,l}$ such that κ is bounded above by a number which depends only on θ.*

Note that the hypotheses of the proposition are certainly satisfied for large $l \times l$ tori which have connections between any two nodes which are within ρl steps of each other along the lattice, where ρ some fixed fraction and is greater than zero. We also note that an upper bound can be proven for the one-dimensional cyclic population model with edges connecting any two nodes within a fixed fraction of the circle; this version is easier to prove and we omit its proof.

Now we prove the propositions, starting with Proposition 1. By the ordering of the eigenvalues the hypothesis implies that for any $k > 1$

$$\frac{1}{1-\lambda_k} \leq \frac{1}{1-\Delta}$$

Then we recall that for vertex-transitive chains (see Proposition 5 of ALDOUS (1989))

$$H_i \geq \frac{n}{2}$$

for $i \neq 1$. Thus

$$\tau_0 \geq \frac{n-1}{2}.$$

We apply this fact as follows:

$$\begin{aligned}
|H_i/\tau_0 - 1| &= \frac{1}{\tau_0}|H_i - \tau_0| \\
&\leq \frac{2}{n-1}|H_i - \tau_0| \\
&= \frac{2}{n-1}\left|\sum_{k\geq 2}\frac{1}{1-\lambda_k}(n(v_{1k}^2 - v_{ik}v_{1k}) - 1)\right| \\
&\leq \frac{2n}{(n-1)(1-\Delta)}\left|\sum_{k\geq 2}\left(v_{1k}^2 - v_{ik}v_{1k} - \frac{1}{n}\right)\right| \\
&= \frac{2n}{(n-1)(1-\Delta)}\left|\frac{1}{n} - v_{11}^2 + v_{i1}v_{11}\right|
\end{aligned}$$

where the last step is via the definition of an orthogonal matrix. The entry v_{11} is the first coordinate of the first eigenvector, which is scaled to have norm one. As the first eigenvector corresponds to the stationary distribution, which is in this case the uniform distribution, $v_{11} = n^{-1/2}$. Therefore this upper bound goes to zero.

The following fact is a case of Proposition 8 of Aldous (1989):

Theorem 2 *Consider a sequence of vertex-transitive graphs such that*

1. *The number of nodes goes to infinity.*
2. $\tau_2/\tau_0 \to 0$
3. $H_i/\tau_0 \to 1$ *for any sequence of i's.*

Then the distribution of Y_i/τ_0 converges to the exponential distribution with mean one for any i.

Under the hypotheses of Proposition 1, λ_2 is bounded away from 1 and therefore τ_2 is bounded above. Clearly τ_0 goes to infinity and by the above calculation H_i/τ_0 converges to one, therefore we can apply Aldous' theorem to prove Proposition 1.

Now we prove Proposition 2. Recall that we need to find a set of distinguished paths Γ such that L and B are bounded by fixed constants, where again L is the maximum length of any path in Γ, and B is the maximal number of paths in Γ going through a given edge.

We set $S = \lceil \theta^{-1} \rceil$. It is easy to see that we can travel from any node to any other node along existing edges in at most S steps, as

$$\lceil l/2 \rceil \leq \lceil \theta^{-1} \rceil \cdot \lceil \theta l \rceil \leq S \cdot (\lceil \theta l \rceil + 1).$$

Therefore we can choose Γ to contain paths of maximal length S, which will make L bounded.

In fact, we will choose paths in Γ to be paths of exactly length S as often as possible. To specify the class of paths, we pick a node p and then choose a set of paths Γ_p from p to all of the points of the torus. We then translate this set of paths around the torus to get a complete set of paths Γ.

To describe Γ_p, set up coordinates on the torus such that p is at the point $(\lceil l/2 \rceil, \lceil l/2 \rceil)$. Now, for any point q on the torus, there exist integers a and b such that

$$\begin{aligned} p_x + aS &\leq q_x \leq p_x + (a+1)S \\ p_y + bS &\leq q_y \leq p_y + (b+1)S. \end{aligned}$$

We note that by hypothesis an edge will exist between any two nodes which are $|a|+1$ or less apart in the horizontal direction and $|b|+1$ apart or less in the vertical direction. For example, when l is sufficiently large,

$$|a| + 1 = \left\lceil \frac{|p_x - q_x|}{S} \right\rceil \leq \lceil \theta \cdot (l/2 + 1) \rceil \leq \lceil \theta l + 1 \rceil$$

Therefore we choose a path from p to q using edges going only a or $a+1$ in the horizontal direction, and only b or $b+1$ in the vertical direction. This path will be at most S steps long. We repeat this process for all q to construct Γ_p, then translate to construct the whole of Γ.

Now we need to show that B is bounded above by a constant independent of l. Pick an edge e traversed by a path γ, and assume that it goes a steps in the horizontal direction, and b steps in the vertical direction. By the above construction the start p and the terminus of the path q will satisfy the inequalities

$$\begin{aligned} p_x + (a-1)S &< q_x < p_x + (a+1)S \\ p_y + (b-1)S &< q_y < p_y + (b+1)S. \end{aligned}$$

Note that in a given Γ_p there can be at most $4S^2$ paths possible terminal points in this region. Therefore there are at most $4S^2$ paths in Γ_p which contain the edge e.

Now, how many translates of paths in Γ_p contain e? Let us denote by $E(\Gamma_{p'})$ the union of edges traversed by paths in $\Gamma_{p'}$. Note that if $e \in E(\Gamma_{p'})$, then we have made a choice of $e' \in E(\Gamma_p)$ to

map onto e. This choice determines the base of translation p' uniquely. Because each path is at most S edges long, and there are at most $4S^2$ paths, there are at most $4S^3$ choices of such an e', and therefore at most $4S^3$ possible translations. Each choice of translation has at most $4S^2$ paths traversing e, therefore B is at most $16S^5$. This proves Proposition 2.

FIGURE 1 — A simple symmetric graph.

FIGURE 2 — H_i/τ_0 for a random walk on a graph which has complete connections out to distance $D/10$. The x axis is the number of nodes D. The y axis is the scaled distance between node i and node one.

FIGURE 3 — Convergence to the limit distributions on the torus for a sample of size two: (a) from the same deme, (b) from adjacent demes, and (c) from maximally-distant demes. The dotted line in (a) represents Equation 7, and the dotted lines and (b) and (c) represent Equation 8. Other lines represent the CDFs of pairwise coalescence times on the $l \times l$ torus described in the text, for $l = 16, 32, 64, 128$. The lines farthest away from the dotted line are for $l = 16$; as l increases the lines move toward the dotted one. The simulations counted the number of times X_{ij}/N_e fell in each of 100 bins of increasing width between 0 and 20, so that an exponential random variable with mean one would have an equal probability of falling in each bin and its CDF would fall along the diagonal. Additional simulations were done to obtain an accurate picture of the shape of the curves in (a) for small values of X_{ij}/N_e.

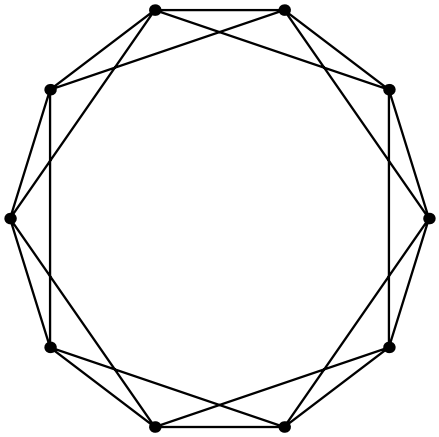

FIGURE 1: Matsen and Wakeley

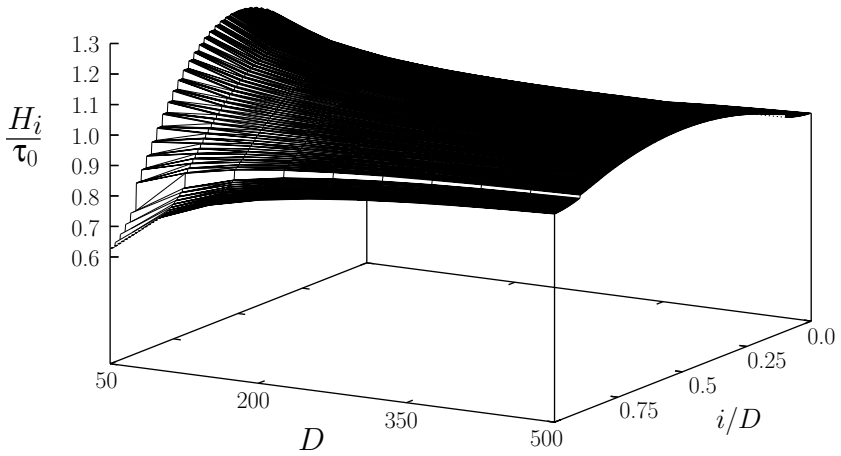

FIGURE 2: Matsen and Wakeley

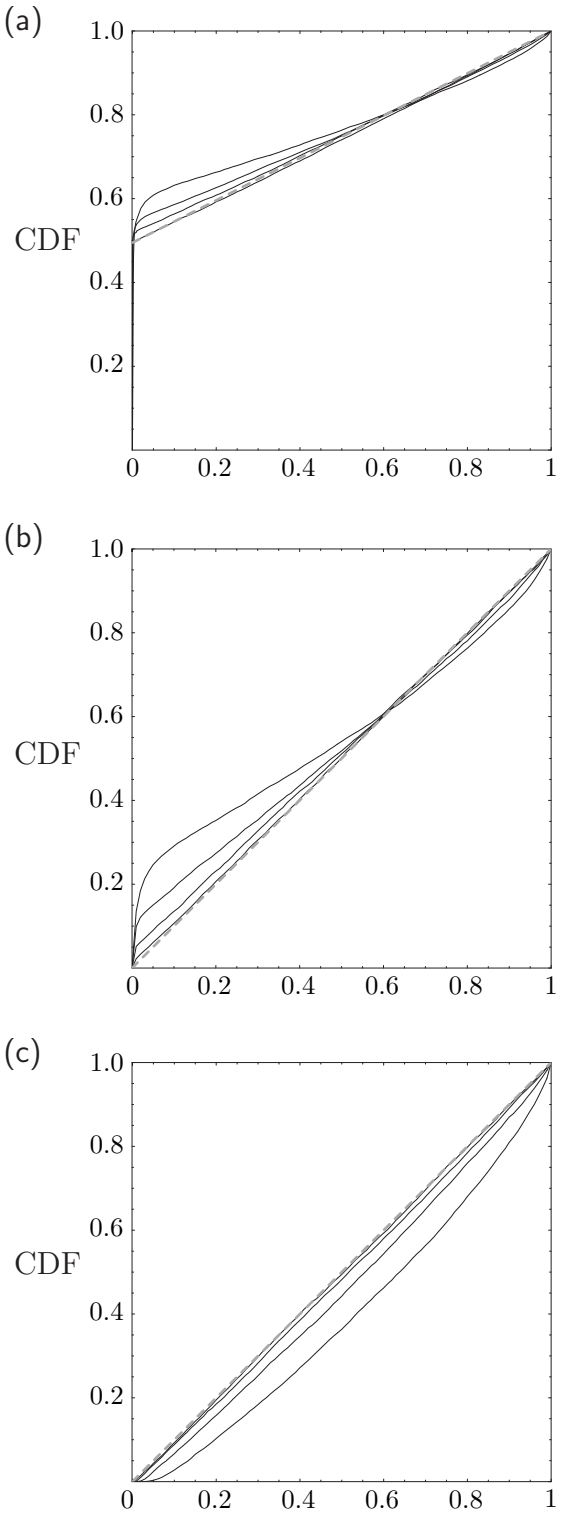

FIGURE 3: Matsen and Wakeley