A spatial approach to jointly estimate Wright's neighborhood size and long-term effective population size

Zachary B. Hancock[1*], Rachel H. Toczydlowski[2], Gideon S. Bradburd[1]

1 Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, MI 481103, USA

2 Northern Research Station, United States Forest Service, Rhinelander, WI 54501, USA

Corresponding author: hancockz@umich.edu

Running title: A spatial approach to jointly estimate Wright's neighborhood size and long-term effective population size

Keywords: isolation by distance, continuous space, effective population size, spatial population genetics, Bayesian, *Amphiprion bicinctus*

Corresponding author:

Zachary B. Hancock

Department of Ecology & Evolutionary Biology

University of Michigan

Ann Arbor, MI 48103

Email: hancockz@umich.edu

Abstract

Spatially continuous patterns of genetic differentiation, which are common in nature, are often poorly described by existing population genetic theory or methods that assume panmixia or discrete, clearly definable populations. There is therefore a need for statistical approaches in population genetics that can accommodate continuous geographic structure, and that ideally use georeferenced individuals as the unit of analysis, rather than populations or subpopulations. In addition, researchers are often interested describing the diversity of a population distributed continuously in space, and this diversity is intimately linked to the dispersal potential of the organism. A statistical model that leverages information from patterns of isolation-by-distance to jointly infer parameters that control local demography (such as Wright's neighborhood size), and the long-term effective size (N_e) of a population would be useful. Here, we introduce such a model that uses individual-level pairwise genetic and geographic distances to infer Wright's neighborhood size and long-term N_e. We demonstrate the utility of our model by applying it to complex, forward-time demographic simulations as well as an empirical dataset of the Red Sea clownfish (*Amphiprion bicinctus*). The model performed well on simulated data relative to alternative approaches and produced reasonable empirical results given the natural history of clownfish. The resulting inferences provide important insights into the population genetic dynamics of spatially structure populations.

Introduction

In many species, individual (or gamete) dispersal is geographically limited, leading to spatial structure. This spatial structure can in turn give rise to a pattern of isolation by distance (Wright, 1943, 1946; Meirmans 2012), in which a focal individual is, on average, more closely related to an individual sampled nearby than it is to another individual sampled farther away. Much of early population genetic theory was derived under the assumption of random mating, which may have adequately described the model organism populations common in empirical population genetic studies of the time (e.g., *Drosophila* in vials; Prout 1954; Merrell 1953; Dobzhanksy & Spassky 1962) but is less well-suited to describing spatially continuous genetic structure. While several theoretical approaches have relaxed these assumptions by modeling the population as partitioned into “demes" with some constant rate of migration between them (e.g., Wright’s [1943] “island model” and Kimura & Weiss’ [1964] “stepping-stone model”), each approach still maintained panmixia within demes, and neither effectively captures population genetic dynamics in continuous space. The island and stepping-stone models inspired a series of statistical approaches that rely on partitioning samples into discrete populations with some level of genetic differentiation (sometimes estimated) between them (e.g., Wright 1951; Pritchard et al. 2000; Pickrell & Pritchard 2012; Peter 2016). These approaches have been expanded to estimate the degree of admixture between these discrete populations (e.g., ADMIXTURE – Alexander et al. 2009), where individuals inferred to have ancestry proportions in multiple inferred clusters are described as “admixed.” However, when the true pattern of genetic variation is

continuous across the landscape, the inferred *K* clusters in a method like STRUCTURE or individual admixture proportions in ADMIXTURE are mere artifacts of the sampling scheme (Frantz et al 2009, Bradburd et al 2018). An example of this phenomenon in action can be found in studies that group human genetic variation by continent, which often generates an apparent pattern of discrete clusters (Rosenberg et al. 2002; Li et al. 2008); however, when *individuals* are the unit of investigation, human genetic ancestry is continuous and defies simple continental or population groupings (Ramachandran et al. 2005; Lewis et al. 2022; Carlson et al. 2022).

Wright (1943) and Malécot (1946) were the first to consider population models in which individuals were continuously distributed across one- and two-dimensions in geographic space. Wright (1943, 1946) introduced the concept of “neighborhood size” ($\mathcal{N}$) as a statistic to describe natural populations; $\mathcal{N}$ was meant to capture the number of potential parents within a given radius of a focal individual, where that radius was defined by the dispersal distance in two-dimensions. Their early theoretical work has since been expanded by Malécot (1948), Maruyama (1971), Nagylaki (1975; 1978), Felsenstein (1976), Barton et al. (2002), Barton et al. (2013), among others. Despite these theoretical advances, relatively few statistical methods exist for examining populations in continuous space. Wright’s neighborhood size ($\mathcal{N}$) provides important information about the dispersal potential and the rate of genetic drift in continuously distributed populations at a localized level, and is therefore a useful quantity to know, both for conservation purposes and a general understanding of the evolutionary context of a particular population or species.

Rousset (1997; 2000) introduced a method for the estimation of Wright's neighborhood size as the inverse of the slope of a regression between pairwise F_{ST} / (1 – F_{ST}) and the logarithm of pairwise geographic distance. While this method is limited in that it assumes a constant population density, it has the useful benefit of providing a single estimate of $\mathcal{N}$ for the entire population (Shirk & Cushman 2014). Many popular programs enable researchers to estimate $\mathcal{N}$ by implementing this expected relationship (e.g., SPAGeDI – Hardy & Vekemans 2002; Rousset & Leblois 2011). Importantly, the decision of which measure of F_{ST} to use is non-trivial and can produce dramatically different results (Pearse & Crandall 2004; Bhatia et al. 2013). Furthermore, researchers must decide to estimate F_{ST} between individuals or artificially designated subpopulations. The former is very sensitive to individual measures of genetic diversity, which can be particularly noisy and impacted by bioinformatic decisions that affect the presence of missing data and rare variants (Bhatia et al. 2013); this noisiness can be smoothed by lumping individuals into subpopulations, but this "lumping" approach is not ideal when sampling covers a large geographic area and there is a continuous pattern of isolation-by-distance (Pearse & Crandall 2004). An alternative method, introduced by Shirk & Cushman (2014) and implemented in sGD (Shirk & Cushman 2011), estimates $\mathcal{N}$ following a user-specified neighborhood radius, and then used the Burrow's method of linkage disequilibrium (Cockerham & Weir 1977) to estimate the number of breeding individuals of the sample. However, this method does not utilize the information from the shape of the isolation by distance curve like Rousset's. Furthermore, it relies on the investigator having strong prior knowledge on the dispersal potential of the organism

such that they can adequately predict how to discretize space and, thus, how many samples should be included within the neighborhood estimation.

Another quantity that, like Wright's neighborhood size, is useful for understanding and conserving species is the effective population size (N_e). In a conservation setting, N_e may serve as a rough proxy for census size and the adaptive potential of threatened populations (Exposito-Alonso et al. 2022; Theodoridis et al. 2021). Others have used N_e to investigate the relationship between range size or dispersal ability (Leigh et al. 2021; De Kort et al. 2021). Estimation of the inbreeding effective size N_e (Wright 1931) often relies on its relationship with genetic diversity, which, when the mutation rate is μ and the population is at mutation-drift equilibrium, is given by $4N_e\mu$ (termed the "population mutation rate"; Kimura & Crow 1964). This definition is distinct from the variance N_e, which describes the rate of genetic drift between successive generations (Crow & Kimura 1970; Wang & Caballero 1999; Charlesworth 2009). A common estimator of the population mutation rate is Watterson's θ_w, which is K / a_n where K is the number of segregating sites in the sample and $a_n = \sum_{i=1}^{n-1} \frac{1}{i}$ (Watterson 1975). However, Watterson's θ_w is naive to the spatial structure of the sample and is known to be upwardly biased relative to random mating expectations when neighborhood sizes are very low (Battey et al. 2020). Another estimator of the population mutation rate is $\hat{\pi}$, which is $\hat{\pi} = \frac{n}{n-1}\sum_{ij} x_i x_j \pi_{ij}$, where n is the number of samples, x_i and x_j are the frequencies of the ith and jth sequence, and π_{ij} is the number of nucleotide differences between the ith and jth site (Nei & Tajima 1981). At mutation-drift equilibrium and assuming no selection, $\theta_w = \hat{\pi} = 4N_e\mu$.

In continuously distributed populations, inbreeding N_e is not independent of $\mathcal{N}$ as each is impacted by dispersal (Barton et al. 2002; Wilkins 2004). When dispersal is high, N_e converges to the population census size (N_c) but tends to be much larger when dispersal is low. Similarly, because dispersal dictates the radius of $\mathcal{N}$, higher rates lead to more potential parents being included within $\mathcal{N}$. An ideal model of spatial population structure would thus be individual-based such that researchers would not need to arbitrarily group individuals into subpopulations and would co-estimate Wright's neighborhood size and inbreeding N_e while explicitly accounting for the shared influence of dispersal across timescales.

In this paper, we take steps towards this goal by introducing a model that jointly estimates $\mathcal{N}$ and long-term N_e from data on pairwise π and geographic distance between individuals. We validate the model's performance using individual-based forward-time simulations and evaluate our model's performance against Rousset's (1997) method when F_{ST} is estimated between individuals. Finally, we apply our model to an empirical dataset of Red Sea clownfish (*Amphiprion bicinctus*; Saenz-Agudelo et al. 2015).

Methods

Model Intuition

The population genetic pedigree is ultimately shaped by an organism's life-history, including its dispersal potential, generational structure, and mating strategies. Because mutations occur along the branches of the genetic genealogy contained within the

population pedigree, the shape of the pedigree fundamentally determines the diversity of a sample as well as a whole host of additional summary statistics (Fig. 1). The field of statistical population genetics is predicated on the idea that information about processes shaping the population pedigree (e.g., selection, demography) leave their imprint in patterns of genetic diversity and divergence observable in a modern-day sample.

For populations that are spatially structured such that there is autocorrelation between geographic location and genetic ancestry, the pedigree becomes distorted relative to random-mating expectations. This pedigree distortion occurs in two phases: the *scattering phase* and the *collecting phase* (Wakeley 1999, Wilkins 2004; Wilkins & Wakeley 2002). Going backward in time from the present, the scattering phase happens first, and is characterized by lineages that are geographically near one another coalescing more rapidly than expected under random mating (i.e., with a probability greater than $1 / 2N_e$; Wilkins 2004; Wilkins & Wakeley 2002). This signature is especially strong when dispersal is low, as nearby individuals are more likely to be more closely related than a pair of individuals selected from the population at random (with respect to geography). The parameter that governs the rate of coalescence in this phase of the population pedigree is Wright's neighborhood size, which is defined as $\mathcal{N} = 4\pi\rho\sigma^2$, where π is the mathematical constant, ρ is population density, and σ is the standard deviation of the effective dispersal distance defined by a normal distribution with mean zero (Wright 1940, 1943, 1946, 1949). Forwards-in-time, $\mathcal{N}$ describes the number of potential mates within a circle of radius $2\sigma^2$, within which breeding occurs approximately at random with respect to geographic position. Backwards-in-time, Wright's neighborhood size can be thought of as the "pool of possible parents" of a focal

individual (i.e., the number of reproductively mature individuals within two effective dispersal distances of the focal individual in any direction).

Further in the past — exactly how far depends on the rate of dispersal and the habitat geometry (Wilkins 2004) — the coalescent process shifts to the second phase, known as the *collecting phase*, in which the rate of coalescence is independent of the geographic distribution of the modern-day sample of individuals. This independence arises because, following a focal individual's pedigree backward through time and across space, the geographic distribution of its genetic ancestors expands until it ceases to be correlated with that focal individual's location (Fig. 1b; Bradburd & Ralph 2019). In the collecting phase, relatedness between lineages is well-represented by a neutral coalescent process (Kingman 1982), in which the rate of coalescence per generation is $1 / 2N_e$ and the average time to the most recent common ancestor is $4N_e$ generations.

This two-phase distortion of the shape of the pedigree relative to random-mating expectations affects estimates of the genetic diversity of the population. In a spatially structured population at migration-drift equilibrium, dispersal limitations shrink the depth of the coalescent tree locally (i.e., individuals nearby are more related on average than expected under panmixia) while expanding it globally (farther away, individuals are more distantly related than expected). As a result, estimates of $\hat{\pi}$ or $\hat{\theta}$ are highly dependent on the spatial scale of sampling, and are generally downwardly biased relative to the equilibrium π in the collecting-phase; this effect is particularly strong when dispersal is very low relative to the length of the range and the geographic area encompassed by the genotyped samples is small (Exposito-Alonso et al 2022).

Our model (explained below) relies on the relationship between the spatial population genetic pedigree and the isolation-by-distance curve (Fig. 1b, 2). At short geographic distances, there is strong spatial autocorrelation of relatedness - this captures the scattering-phase of the spatial pedigree, and it decays rapidly (Fig. 1b). As the curve flattens, geographic distance ceases to be explanatory of relatedness and the population approaches expectations under panmixia - this is capturing the transition to the collecting-phase. The shape of the decay of relatedness over short spatial scales carries information about $\mathcal{N}$, whereas the inferred asymptote of relatedness over large geographic distances represents a diversity equilibrium (what we term "collecting-phase π", π_c), and is most informative about long-term inbreeding N_e.

Model

We first introduce the model of isolation-by-distance (IBD) that we use - both its form and its assumptions, then describe how we fit this model to observed genomic data. Briefly, our model is closely related to previous theoretical models of genetic differentiation in continuous space (e.g., Wright 1943; Malécot 1946; Barton et al. 2002, 2010, 2013; Ringbauer et al. 2017), and describes the decay in pairwise homozygosity with the geographic distance between samples assuming a homogeneous landscape with isotropic dispersal. We implement this model in a Bayesian framework to estimate the posterior distribution of model parameters conditioned on observed pairwise sample homozygosity and pairwise geographic distance between samples.

Our model seeks to capture two important components of spatially structured populations: 1) that samples covary in their allele frequencies, with a covariance that

decays with geographic distance during the scattering phase, and 2) that there exists an equilibrium level of minimum divergence between individuals that is established during the collecting phase (Fig. 1). Furthermore, our model assumes that populations exist in continuous space in two-dimensions and that dispersal is random and diffusive. In a single dimension, tracking diffusive dispersal backwards-in-time, two lineages will eventually exist in the same location at the same time at some point in the past. However, in two-dimensions, lineages diffusing via Brownian motion will never arrive in the same place at the same time (e.g., Nagylaki 1978, Barton et al. 2002). Modeling a spatial coalescent process in two dimensions is therefore tricky. This issue has been circumvented in the past (Wright 1943; Malécot 1946) by assuming that individuals need not be in the exact same location at the same time, but merely within a given radius of one another. Within this radius, individuals are assumed to interact in a way that is independent of the geographic distance between them, so that the probability of coalescence can be described as 1 / 2ρ (Barton et al. 2002), where ρ is population density. For habitats that are relatively homogenous, such that geographic distance is the primary explanation of covariance, with constant ρ and σ through time and across space, the probability of samples *i* and *j* being identical-by-descent (F_{ij}) can be estimated by

$$F_{ij} = \frac{1}{4\pi\rho\sigma^2} K_0\left(\sqrt{2\mu}\frac{d_{ij}}{\sigma}\right) \quad (1)$$

where K_0 is the modified Bessel function of the second kind of order 0, d_{ij} is the pairwise geographic distance between samples *i* and *j*, and π is the mathematical constant (Wright 1943, Malécot 1946, Barton et al. 2002). Barton et al. (2002) notes that this

approximation diverges as $d_{ij} \rightarrow 0$; to account for this, following Ringbauer et al. (2017) we designate a short distance, κ, within which the rate of coalescence becomes a constant γ that no longer depends on the geographic distance between samples. We refer to γ as the "in-deme" rate of coalescence. Theoretically, γ should converge on $1 / 2\rho$ and an optimal value for κ would be approximately 2σ (Barton et al. 2002).

The classic Wright-Malécot formula (Eqn. 1) describes the probability of identity-by-descent, which, in an infinite population, theoretically decays to zero. However, our model breaks the assumptions of the Wright-Malécot model in two important ways. First, we assume that populations are finite, meaning that there will be a maximum time at which all individuals in the population are identical-by-descent. Second, we choose to model identity-by-state, rather than identity-by-descent, as we assume more empiricists will have access to identity-by-state information than identity-by-descent (particularly in non-model organisms). Therefore, we must incorporate into our model a background rate of genetic similarity at which all individuals in the population are identical-by-state. The expected homozygosity of a pair of samples, *i* and *j*, is thus

$$\widehat{H}_{ij} = F_{ij} + \left(1 - F_{ij}\right)s \quad (2)$$

where $s = \frac{1}{N}\sum_{n=1}^{N} p_n^2 + (1 - p_n^2)$ and p_n is the population frequency of an allele at the nth of N loci (Ringbauer et al. 2018). The quantity s represents the "background" rate of sequence similarity and can be thought of as the complement of the amount of genetic diversity in a population at equilibrium: $1 - s = 2p_n(1 - p_n) = \pi_c$, which we define as "collecting-phase π" and is related to long-term N_e. Unlike $\hat{\pi}$ (mean pairwise genetic distance in a population, Nei & Tajima 1981), estimates of π_c (defined by the asymptote

of the IBD curve, which showcases the transition to the collecting phase; Fig. 1) should be insensitive to the size of a sampled area.

Inference and Implementation

We assume users' data consist of allele frequencies taken across *L* unlinked, biallelic single nucleotide polymorphisms (SNPs) genotyped across a set of *N* samples. Each sample may consist of a single individual or a group of individuals collected at a single location. Allele frequencies may be estimated from genotype data (e.g., the frequency of the *l*th allele in the *n*th sample is simply the number of times that allele is observed divided by the total number of genotyped haplotypes in that sample at that locus) or from pooled sequencing data. From these data, we compute the sample pairwise homozygosity, which is the complement of the pairwise diversity between the samples (i.e., $1 - D_{xy}$). We note that users working with low-coverage sequence data may wish to generate estimates of D_{xy} without conditioning on allele frequencies (e.g., Buerkle & Gompert 2012; Ellegren 2014). Pairwise homozygosity between samples *i* and *j* gives the probability that, at a locus chosen at random, a pair of alleles sampled at random from *i* and *j* respectively, are the same. We calculate it as:

$$\widehat{H}_{ij} = 1 - \frac{1}{L}\sum_{\ell=1}^{L} \hat{f}_{i,\ell}\left(1 - \hat{f}_{j,\ell}\right) + \hat{f}_{j,\ell}\left(1 - \hat{f}_{i,\ell}\right) \quad (3)$$

where, $\widehat{H}_{ij}$ gives the sample homozygosity between samples *i* and *j* calculated across all *L* loci, and $\hat{f}_{i,\ell}$ gives the sample allele frequency in the *i*th sample at the *l*th locus. Pairwise homozygosity is a measure of absolute genetic similarity, so it is not sensitive

to the sampling configuration. Additionally, $\widehat{H}$ is proportional to the allelic diversity defined in Bradburd et al. (2018), so we proceed by assuming it can be reasonably modeled as Wishart-distributed, and the framework we use for statistical inference is similar to that of Bradburd et al (2018) (see also Ringbauer et al [2018]).

Specifically, we construct a parametric expected homozygosity matrix using a modified version of the Wright-Malécot model of isolation by distance introduced in Equation 2 and calculate the likelihood of the sample homozygosity as a draw from a Wishart distribution parameterized by the parametric homozygosity. Concretely, we write that the probability of identity by descent between samples *i* and *j*, F_{ij}, is

$$F_{ij} = \begin{cases} \gamma, & \text{if } D_{ij} \leq \kappa \\ \dfrac{K_0\left(\sqrt{m}D_{ij}\right)}{\mathcal{N}}, & \text{if } D_{ij} \geq \kappa \end{cases} \quad (4)$$

where γ is the "in-deme" rate of identity by descent (i.e., the probability of being identical by descent at distances short enough that mating can reasonably be considered panmictic, $D_{ij} \leq \kappa$), and F_{ij} for $D_{ij} \geq \kappa$ is given by the Wright-Malécot function introduced in Equation 2. We then construct our parametric homozygosity between samples *i* and *j*, Ω_{ij}, as

$$\Omega_{\mathrm{ij}} = \mathrm{F_{ij}} + \left(1 - \mathrm{F_{ij}}\right)\mathrm{s} + \delta_{\mathrm{ij}}\eta_{\mathrm{i}} \quad (5)$$

where *s* is the rate of identity by state not due to identity by descent, δ_{ij} is the Kronecker δ, and η_{i} is a parameter (often called a "nugget" in the geostatistical literature, Diggle et al 1998) that captures inbreeding specific to the *i*th sample. We then calculate our likelihood as

$$\mathrm{P}\left(\widehat{\mathrm{H}} \mid \Omega\right) = \mathcal{W}\left(\mathrm{L}\widehat{\mathrm{H}} \mid \Omega, \mathrm{L}\right) \tag{6}$$

where *L* is the number of independent genomic loci used in the calculation of $\widehat{\mathrm{H}}$.

We take a Bayesian approach to infer the parameters of this model. The posterior probability density of our parameters is given by

$$P\left(\gamma, m, \mathcal{N}, \vec{\eta} \mid \widehat{H}\right) \propto P\left(\widehat{H} \,\middle|\, \Omega(\gamma, m, \mathcal{N}, \vec{\eta})\right) P(\gamma)P(m)P(\mathcal{N})P(\vec{\eta}) \tag{7}$$

where $\Omega(\gamma, \mathrm{m}, \mathcal{N}, \vec{\eta})$ denotes the dependence of Ω on its constituent parameters $(\gamma, \mathrm{m}, \mathcal{N}, \vec{\eta})$, and $P(\theta)$ denotes the prior probability of a given parameter θ. Table S1 describes the prior probability distributions we implement for each parameter in our model. We implement this model in Rstan (Stan Development Team 2023) and use STAN's Hamiltonian Monte Carlo algorithm to characterize the posterior distribution of the parameters.

Because pairwise homozygosity often varies over a very small absolute range (e.g., 0.99-0.999), we take several steps to facilitate inference on the parameters of the model. First, we estimate the parameters *m*, γ, and $\vec{\eta}$ in log space, which should help chains mix over the posterior density. Second, we scale both the sample homozygosity $\widehat{\Omega}$ and the parametric homozygosity Ω to vary between 0 and 1:

$$\begin{aligned} \mathrm{a} &= \left(\widehat{\Omega} - \min\left(\widehat{\Omega}\right)\right) \\ \mathrm{b} &= \max\left(\widehat{\Omega} - \min\left(\widehat{\Omega}\right)\right) \\ \widehat{\Omega'} &= \frac{\widehat{\Omega} - \mathrm{a}}{\mathrm{b}} \\ \Omega' &= \frac{\Omega - \mathrm{a}}{\mathrm{b}} \end{aligned} \tag{8}$$

This scaling is not without drawbacks, as the variance of a Wishart distribution parameterized by Ω is not the same as that parameterized by Ω'. However, we feel that the benefits it offers outweigh the costs, particularly because, to our knowledge, there is no theoretically motivated "correct" variance on homozygosity.

Simulations

We evaluated model performance using individual-based forward-time simulations performed in SLiM v3.6 (Haller & Messer 2019). Simulated individuals were diploid ($2n$) and hermaphroditic, with haploid genomes of 100 Mb, non-overlapping generations, and a uniform recombination rate of 10^{-9} per base-pair per generation. Mate choice, spatial competition, and dispersal are controlled by a constant value, σ. Individuals were simulated on a continuous, two-dimensional 25x25 landscape with reflecting boundaries. Total population density was regulated by an enforced carrying-capacity, K, to avoid spatial clumping (Felsenstein 1975). To reduce the impact of edge effects, we designate this density-dependent competition using the function *localPopulationDensity()* in SLiM, which computes the total interaction strength around a focal individual first as $2\pi\sigma^2$, then divides this strength by the integral of the interaction after clipping by the bounds of the specified landscape (Haller & Messer 2022). This has the desired effect of rescaling competition relative to the occupiable area. Edge-effects reduced the census population size by ~35% relative to Kw^2, where w is the width of the simulated landscape. The maximum distance at which individuals experience spatial competition was capped at 3σ. Similar to spatial competition, mates were chosen within a maximum distance of 3σ, with each potential mate assigned a weight drawn from a

Gaussian distribution with max $1 / 2\pi\sigma^2$. The number of offspring produced per mating event was drawn from a Poisson distribution with shape parameter $\lambda = 2$, which on average replaces the parents. Finally, offspring dispersal distance was drawn from a normal distribution with mean 0 and standard deviation σ and maximum of 3σ.

We performed simulations across a range of values of K (2, 5, 10, 25) and σ (0.5, 0.75, 1.0, 1.25, 1.5, 2.0), which varied theoretical neighborhood size from a minimum of 6.25 to a maximum of 861.81, and total census size from ~800–10,000. We performed 10 replicates for each combination of parameter values, for a total of 240 simulations. Each simulation was run for 100,000 generations to ensure time for dispersal to shape patterns of genetic diversity. The output from SLiM were tree-sequences and these were parsed in Python using the package *pyslim* v.1.0.1 (Kelleher et al. 2018). For trees in which multiple roots existed (i.e., coalescence had not yet occurred during the SLiM run), we performed "recapitation," which simulates a neutral coalescent process among remaining, uncoalesced lineages (Haller et al. 2018). Mutations were then simulated onto the tree-sequences using *msprime* v.1.2.0 (Kelleher et al. 2016) at a rate of 10^{-7} per base-pair per generation. We then randomly sampled 100 individuals alive in the final generation and we calculated pairwise π between each pair of individuals using *tskit* v.0.5.3 (Kelleher et al. 2018). We output the pairwise individual π matrix and geographic coordinate matrix, the latter of which was converted into a distance matrix in R (R Core Team 2022) using the package *fields* (Nychka et al. 2021). These two matrices were then used as input to our model, which we used to infer a range of parameters ($\mathcal{N}$, π_c, γ, and μ). Note that of these parameters, we only explicitly designated μ; the remaining values have theoretical expectations given the parameters

we chose for each individual simulation but are not explicitly defined. Theoretical N_e in a square habitat is

$$N_e = N\left(1 + \frac{2\log\left(k\frac{L}{\sigma}\right)}{\mathcal{N}}\right) \quad (9)$$

where k = 0.24 when dispersal is Gaussian, and L is the length of the major axis (Wilkins 2004; Barton et al. 2002; Charlesworth et al. 2003). Importantly, as the rate of dispersal approaches the length of the range and $\mathcal{N}$ approaches the census size, N_e approximately equals the population census size, N. However, theoretical N_e is likely an overestimate of the true N_e in our simulations due to reduced population density near the edges as opposed to the expected homogenous distribution. Thus, we estimate the "true" N_e in each simulation as $N_e = \pi_{d_{ij}>20}/4\mu$, where $\pi_{d_{ij}>20}$ is the mean pairwise diversity for samples at distances greater than 20, ensuring they all occur in the collecting phase. For $\mathcal{N}$, we do not use a similar proxy and instead rely on the theoretical $\mathcal{N}$ given that the impacts of deviations from theory in our simulations for areas of relevance to $\mathcal{N}$ are expected to be much smaller than for N_e.

For each dataset, we performed four MCMC chains of 4000 steps each. We pruned the first 1e3 steps in each chain as burn-in and thinned the remaining steps by sampling every 4th iteration, for a total of 250 sampled post burn-in iterations per chain. Because some chains displayed poor mixing, we selected the chain with the highest mean posterior probability as the one from which to report results. We visually verified that the chains had achieved convergence by inspecting the trace plots of the posterior probability and parameter values of interest.

Simulations and model estimation were performed on the Michigan State Institute for Cyber-Enabled Research High Performance Computing Cluster (ICER HPCC). Code, including the ICER HPCC slurm workflow, SLiM recipes, and Python scripts, can be found at https://github.com/zachbhancock/WM_model.

Empirical dataset

We used a reduced-representation genomic dataset of Red Sea clownfish (*Amphiprion bicinctus*) from Saenz-Agudelo et al. (2015) to explore how this model performed on empirical data. This dataset is composed of 103 wild individuals sampled from 10 unique locations spread across the species' range (Red Sea). Sequence reads were generated using double-digest restriction-associated DNA and sequenced 1 x 101 base pairs. We downloaded the raw (demultiplexed) sequence reads for each of the 103 individuals from the International Nucleotide Sequence Database Collaboration (BioProject PRJNA294760, https://www.ncbi.nlm.nih.gov/bioproject/PRJNA294760) using the fasterq-dump function in the SRA Toolkit (V2.10.7, Kodama, Shumway, & Leinonen 2012). We dropped reads with uncalled bases and where the mean Phred score was < 15 (sliding window 15% of read length; Stacks2 process_radtags module V2.54 – Rochette, Rivera-Colón, & Catchen 2019). We confirmed that no common adapter sequences were present (3' end of reads) and that all reads were a uniform length (81 bp post demultiplexing and trimming by Saenz-Agudelo et al.). Next, we assembled these reads *de novo* using Stacks2 (V2.54, Rochette, Rivera-Colón, & Catchen 2019) and optimized assembly parameters (see Saenz-Agudelo et al. 2015). Post Stacks, we dropped loci that were scored in <50% of individuals. Finally, we

calculated pairwise pi and pairwise geographic distance between each pair of individuals in the dataset. Geographic distance here is measured as the distance traveling via the sea and avoiding crossing land (marmap – Pante & Simon-Bouhet 2013).

Results

Our model performed well on both the simulated and empirical datasets. In the former, the model converged on the theoretical expectations for both $\mathcal{N}$ and N_e. Furthermore, it performed favorably relative to Rousset's method, which was unbiased but had much higher variance than our model, especially at lower $\mathcal{N}$. Finally, as applied to the empirical dataset, our model produced reasonable results given the known natural history of clownfish.

Simulation results

The individual-based simulations were performed by varying both dispersal distance (σ) and local carrying-capacity (K), which generated a range of $\mathcal{N}$ from 6.28 to 861.81. Given the simulated geographic range area, this range of values of encompasses $\mathcal{N}$ extreme/strong spatial structure that is likely never well-described by the Kingman's coalescent at the lower end (Wilkins 2004) and virtually panmictic populations at the upper end.

At all simulated values of σ and K, the 95% equal-tailed credible interval of $\mathcal{N}$ in the consistently included the theoretical $\mathcal{N}$; however, at larger σ and K it did so with

high variance, with the median falling below the theoretical value (Fig. 3a). This is likely due to the reliance of the model on detectible signatures of covariance between geography and ancestry, which decays at large $\mathcal{N}$ (Fig. 2). At small dispersal distances, the variance in our estimates of $\mathcal{N}$ was relatively low even at higher *K*. However, at larger dispersal distance the variance subsequently increased regardless of *K*, indicating that σ likely has the largest impact on model precision. Notably, our model estimated an $\mathcal{N}$ greater than the theoretical expectation only at the lowest K and σ combination, which may indicate that the Wright-Malécot model is a poor approximation of relatedness when $\mathcal{N} < 10$ for a population of this geographic range with finite boundaries. Furthermore, we found that while Rousset's method is an unbiased estimator, its variance becomes larger at lower *K* and σ pairs than our model and can even take on negative values (Fig. 4).

The model also performed well at estimating π_c. As noted above, the theoretical N_e should scale negatively with increasing $\mathcal{N}$ and N_e, but positively with global census size and local density (regulated by *K*). While our model does not estimate N_e directly, we estimate a proxy in π_c; under neutrality, $\pi_c \approx 4N_e\mu$. Rearranging, $N_e = \pi_c / 4\mu$, and we compare this value with the true N_e estimated from $\pi_{d_{ij}>20}$ and the theoretical N_e in Equation 9. When N_e is estimated from $\pi_{d_{ij}>20}$, our model performs exceptionally well (Fig. 3B), generally falling along equality for all values of σ and *K*. When compared to the purely theoretical N_e, our model underestimates N_e, especially at higher values of σ and *K* (Fig. S2). We also compared estimated π_c across values of σ and *K* and found a general pattern of decreasing π_c at increasing σ, irrespective of population density (Fig. S1). This matches our expectations, based on theory (Wilkins 2004), that the total

amount of genetic diversity in a finite, spatially structured population should increase with the degree of spatial structure, and therefore decrease with the scale of dispersal.

Empirical results

For the Red Sea clownfish (*Amphiprion bicinctus*) dataset, our model estimated a $\mathcal{N}$ of 52.16 (95% CI: 50.54–54.04) (Fig. 5, S2). This neighborhood size is consistent with a recent localized population census study in the Gulf of Eilat, which found that the number of clownfish in the census year 2015 within a 200 x 50 m area was 52 fish (Howell et al. 2016). It is important to note that we have no independent information on the relevant spatial scale of dispersal or mating in Red Sea clownfish, and thus the correspondence between our estimates and the population census survey could be coincidental. We estimated a π_c of 0.003 and our estimated *m* (which is a compound parameter that includes long-distance dispersal and mutation) as 1.4e-9. Substituting *m* for the mutation rate, this produces an estimate of the long-term N_e for clownfish as 535,714. It should be noted that while *A. bicinctus* is currently listed by the IUCN as "least concern" (Myers et al. 2017), several studies have suggested that clownfish are undergoing population declines, which could theoretically lead to a decoupling between $\mathcal{N}$ and N_e (i.e., $\mathcal{N}$ may reflect recent demographic shifts that N_e has yet to be impacted by; Nanninga et al. 2015; Howell et al. 2016; Yosef et al. 2022).

Discussion

An extensive historical literature exists on the biases that spatial structure introduces to commonly employed population genetic summary statistics (e.g., Bradburd & Ralph 2019; Battey et al. 2020). However, many studies still use measures of effective population size (such as Watterson's θ_w and π) that are naive to population structure (and the geographic sampling of individuals) in empirical systems that are spatially structured. Other attempts to develop statistical population genetic approaches that account for population structure by discretizing the habitat into demes that may be defined by sampling region or violation of Hardy-Weinberg (e.g. STRUCTURE – Pritchard 2000; Montana & Hoggart 2007), but in reality, many populations display a continuous pattern of isolation-by-distance that cannot be adequately reduced to a discrete stepping-stone model (Kimura & Weiss 1964). Indeed, even in a fine-scaled lattice population, Battey et al. (2020) showed that biases in estimates of Watterson's θ and Tajima's *D* emerge as the sample size per deme approaches the local effective size.

To investigate populations with continuous patterns of isolation-by-distance, we suggest that a single summary diversity statistic often does not capture the dynamics of interest and is skewed by the shape of the population genetic pedigree. For example, $\hat{\pi}$ is dragged down by rapid coalescence in the scattering-phase but inflated by the influence of low dispersal in the collecting phase. Hence, $\hat{\pi}$ ceases to adequately reflect either process – it cannot tell us about local demography because it is upwardly inflated by deep-time coalescence, and it cannot tell us about ancient events because it is downwardly biased by local demography. Indeed, dispersal acts in opposing ways in these two phases: low dispersal causes local individuals to be more related on average

and increases the length of the scattering phase, but over large distances and deeper timescales it inflates $\hat{\pi}$ (Equation 9; Wilkins 2004). Our model generates estimates of a diversity statistic (π_c) that is insensitive to the geography of sampling, and simultaneously provides estimates of Wright's neighborhood size, thereby better capturing the spatial dynamics of the population.

Furthermore, our model is robust to a wide range of $\mathcal{N}$, though the variance of our estimates of $\mathcal{N}$ increases as the true $\mathcal{N}$ grows large. Theory and previous simulation-based research predicts that populations with very large $\mathcal{N}$ (≥10,000) behave in a way that is approximately panmictic (Wright 1943), and populations need a considerably smaller $\mathcal{N}$ (≤100) to differ substantially from random mating expectations (Battey et al. 2020). However, when subsampling the population randomly with respect to space, our simulated IBD curves indicate that $\mathcal{N}$ as low as ~200 generate results that superficially resemble panmixia. This finding is in line with theory from Wilkins (2002), who found that most of the population genetic pedigree occurs in the collecting phase. For IBD to be detectable at $\mathcal{N}$ > 200, researchers would need to exhaustively sample local areas to ensure the collection of recent pedigree relatives due to how shallow the scattering phase is relative to the collecting.

One important caveat in interpreting the performance of our model on simulations is that there is not always a perfect correspondence between the quantities we are trying to estimate in our model (e.g., $\mathcal{N}$ or π_c) and parameter values we set in our simulations. For example, the Wright-Malécot IBD model assumes an infinite landscape, whereas our simulation model, in an effort to incorporate greater biological realism, has absorbing boundaries. We contend this is a more realistic depiction of species' ranges, but it causes our simulation parameters to diverge from theory. Edge effects decrease both $\mathcal{N}$ and N_e relative to theoretical expectations because they reduce population density along the periphery of the range. Combined with lack of gene flow from even farther-flung regions, this reduction in diversity leads to neighboring individuals near edges being more related to one another on average than those sampled from the range center (Wilkins 2004; Rogan et al. 2023). Therefore, our model may be capturing this depression of $\mathcal{N}$ and N_e relative to theoretical predictions, which may be especially apparent at high values of K and σ.

The other major way in which our simulation parameters diverge from their theoretical counterparts is in the difference between forward- and backward-time dynamics. The dispersal parameter we set in our SLiM simulations (σ) determines the midparent-offspring dispersal kernel in two dimensions forward-in-time, while the dispersal parameter in Equation 1, which affects both our estimates of $\mathcal{N}$ and π_c, is the *effective* dispersal rate, and describes the dispersal kernel connecting offspring to their parents backwards-in-time (Cayuela et al 2018). The model of spatial competition and density-dependence (the implications of which are discussed more below) that we implement in our simulations causes the forward- and backward-time dispersal

dynamics to differ. This discrepancy arises because, in our simulations, individuals are most likely to be born in regions of higher relative population density (due to the proximity required for parents to mate), where, if they remain, they will experience, on average, lower fitness. Therefore, individuals who disperse *farther* from their birth location are more likely to have higher fitness be represented in the population pedigree, which increases the effective dispersal rate relative to the forward-time rate we specify in our SLiM model. The impact of this discrepancy on the "true" value of $\mathcal{N}$ appears to be offset by the fact that effective population density is smaller than the value of K (for similar reasons) that we specify in our SLiM model. Nonetheless, it is helpful to be explicit about the differences between the meanings of the parameters we specify in our simulation model and that we estimate in our inference model.

Model assumptions and shortcomings

While our model performs well on simulated data and the empirical dataset presented here, it makes several important assumptions that may often be violated in natural systems. Firstly, we assume that dispersal is random (within a specified dispersal kernel) and non-directional; thus, our model may poorly approximate patterns of isolation-by-distance in organisms that have directional movement, such as some marine planktonic species that may be carried by ocean currents or wind-dispersed pollen grains. Second, our model assumes that the habitat is relatively homogenous such that there are no major barriers to gene flow – i.e., patterns of isolation-by-distance are generated solely by the traversable Euclidean distance between two individuals. Strong physical or environmentally mediated barriers to dispersal may affect model

performance (Ringbauer et al 2018, Wang & Bradburd 2014, Bradburd et al 2013). In a similar vein, our model ignores confounding factors such as local adaptation that may drive clinal patterns of relatedness (e.g., Pruisscher et al. 2018; Jofre & Rosenthal 2021).

As discussed in the Methods, the Wright-Malécot model relies on assumptions that are mutually incompatible (independent dispersal and homogeneous population density). In SLiM, we modeled populations with density-dependent selection to overcome Felsenstein's "pain in the torus" (Felsenstein 1975) and maintain a roughly homogenous distribution of individuals across space. In this way, comparing the theoretical expectations from Wright-Malécot with a population model in which dispersal and density are not strictly independent is inexact. However, despite this discrepancy between theory and the simulated population, our model reasonably captured theoretical expectations, indicating that the Wright-Malécot formulation which underpins our model is robust to this violation.

Finally, our model relies on a user-defined κ, interpreted as the minimum-distance between two individuals in which the Wright-Malécot formulation breaks down and relatedness converges on $1 / 2\rho$ (Barton et al. 2002; Ringbauer et al. 2018). Preliminary exploration of the model indicates that model performance is poor at arbitrarily high κ, but not at low κ. This is due to the fact that, at higher κ, less of the scattering-phase is being captured by the model, leading to poor mixing. Ideally, κ would be estimated like the other parameters of the model; however, doing so has led to dramatic increases in computation time, and thus for the current work we opted to set a constant κ.

Conclusions

For many organisms, geographical distance influences mate-choice, leading to patterns of continuous spatial structure. An important parameter governing the strength of isolation-by-distance is Wright's neighborhood size ($\mathcal{N}$), a theoretical quantity that describes the number of potential breeding individuals within a given dispersal radius. Previously, empirical researchers interested in estimating $\mathcal{N}$ relied upon Rousset's (1997; 2000) method, but this method requires either subsetting individuals into pseudo-populations (arbitrarily discretizing a potentially continuous reality) and calculating F_{ST} between them or estimating pairwise F_{ST} between individuals. The latter approach introduces a significant amount of noise into patterns of isolation-by-distance, and, in our simulation study, demonstrates poor performance due to high variance in results. Unlike Rousset's estimator, our method shows good performance across values of $\mathcal{N}$, and offers the additional benefit of generating an estimate of the long-term effective population size (N_e), which is linked to $\mathcal{N}$ by dispersal. Here, we have presented a model that jointly estimates $\mathcal{N}$ and long-term N_e using an individual-based approach (i.e., it does not require arbitrary discretization via the lumping of samples). The introduced model produced reasonable estimates of the theoretical expectations from simulated data and performed well on an empirical dataset of clownfish. Future work will aim to generate an R package for ease of use for researchers (though code for performing the presented model is available in the Github link above). In addition, the model could be expanded to incorporate the potential for directional migration or patterns of isolation-by-environment (Wang & Bradburd 2014).

References

Alexander DH, Novembre J, Lange K. 2009. Fast model-based estimation of ancestry in unrelated individuals. Genome Research, 19: 1655–1664.

Barton NH, Depaulis F, Etheridge AM. 2002. Neutral evolution in spatially continuous populations. Theoretical Population Biology, 61(1): 31–48.

Barton NH, Kelleher J, Etheridge AM. 2010. A new model for extinction and recolonization in two dimensions: Quantifying phylogeography. Evolution, 64(9): 2701–2715.

Barton NH, Etheridge AM, Véber A. 2013. Modelling evolution in a spatial continuum. Journal of Statistical Mechanics: Theory and Experiment, 2013: DOI:10.1088/1742-5468/2013/01/P01002.

Battey CJ, Ralph PL, Kern AD. 2020. Space is the place: Effects of continuous spatial structure on analysis of population genetic data. GENETICS, 215(1): 193–214.

Bhatia G, Patterson N, Sankararaman S, Price AL. 2013. Estimating and interpreting F_{ST}: The impact of rare variants. Genome Research, 23: 1514–1521.

Bradburd GS, Coop GM, Ralph PL. 2018. Inferring continuous and discrete population genetic structure across space. GENETICS, 210(1): 33–52.

Buerkle CA, Gompert Z. 2012. Population genomics based on low coverage sequencing: how low should we go? Molecular Ecology, 22(11): 3028–3035.

Carlson J, Henn BM, Al-Hindi DR, Ramachandran S. 2022. Counter the weaponization of genetics research by extremists. Nature, 610: 444–447.

Charlesworth B. 2009. Effective population size and patterns of molecular evolution and variation. Nature Review Genetics, 10: 195–205.

Charlesworth B, Charlesworth D, Barton NH. 2003. The effects of genetic and geographic structure on neutral variation. Annual Review of Ecology, Evolution, and Systematics, 34: 99–125.

Cockerham CC, Weir BS. 1977. Digenic descent measures for finite populations. Genetics Research, 30(2): 121–147.

Crow JF, Kimura M. 1970. *An Introduction to Population Genetics Theory*. Burgess, Minneapolis, MN.

De Kort H, Prunier JG, Ducatez S, Honnay O, Baguette M, Stevens VM, Blanchet S. 2021. Life history, climate and biogeography interactively affect worldwide genetic diversity of plant and animal populations. Nature Communications, 12(516: https://doi.org/10.1038/s41467-021-20958-2.

Dobzhanksy T, Snassky NP. 1962. Genetic drift and natural selection in experimental populations of *Drosophila pseudoobscura*. Proceedings of the National Academy of Sciences, 48(2): 148–156.

Ellegren H. 2014. Genome sequencing and population genomics in non-model organisms. Trends in Ecology & Evolution, 29(1): 51–63.

Exposito-Alonso M, Booker TR, Czech L, Gillespie L, Hateley S, Kyriazis CC, Lang PLM, Leventhal L, Nogues-Bravo D, Pagowski V, Ruffley M, Spence JP, Toro Arana SE, Weiss CL, Zess E. 2022. Genetic diversity loss in the Anthropocene. Science, 377(6613): 1431–1435.

Felsenstein J. 1975. A pain in the torus: Some difficulties with models of isolation by distance. American Naturalist, 109(967): 359–368.

Felsenstein, J. 1976. The theoretical population genetics of variable selection and migration. Annual Review of Genetics, 10: 253–280.

Frantz AC, Cellina S, Krier A, Schley L, Burke T. 2009. Using spatial Bayesian methods to determine the genetic structure of a continuously distribution population: clusters or isolation by distance? Journal of Applied Ecology, 46(2): 493–505.

Haller BC, Galloway J, Kelleher J, Messer PW, Ralph PL. 2018. Tree-sequence recording in SLiM opens new horizons for forward-time simulation of whole genomes. Molecular Ecology Resources, 19(2): 552–566.

Haller BC, Messer PW. 2019. SLiM 3: Forward genetic simulations beyond the Wright-Fisher model. Molecular Biology and Evolution, 36(3): 632–637.

Haller BC, Messer PW. 2022. SLiM 4: Multispecies co-evolutionary modeling. American Naturalist, in press.

Hardy OJ, Vekemans X. 2002. SPAGeDi: a versatile computer program to analyse spatial genetic structure at the individual or population levels. Molecular Ecology Resources, 2(4): 618–620.

Howell J, Goulet TL, Goulet D. 2016. Anemonefish musical chairs and the plight of th two-band anemonefish, *Amphiprion bicinctus*. Environmental Biology of Fishes, 99: 873–886.

Jofre, GI, Rosenthal GG. 2021. A narrow window for geographic cline analysis using genomic data: Effects of age, drift, and migration on error rates. Molecular Ecology Resources, 21(7): 2278–2287.

Kelleher J, Etheridge AM, McVean G. 2016. Efficient coalescent simulation and genealogical analysis for large sample sizes. PLoS Computational Biology, 12(5): e1004842.

Kelleher J, Thornton KR, Ashander J, Ralph PL. 2018. Efficient pedigree recording for fast population genetics simulation. PLoS Computational Biology, 14(11): e1006581.

Kimura M, Crow JF. 1964. The number of alleles that can be maintained in a finite population. GENETICS, 49(4): 725–738.

Kimura M, Weiss GH. 1964. The stepping stone model of population structure and the decrease of genetic correlation with distance. GENETICS, 49(4): 561–576.

Kingman JFC. 1982. The coalescent. Stochastic Processes and their Applications, 13(3): 235–248.

Kodama Y, Shumway M, Leinonen R. 2012. The Sequence Read Archive: explosive growth of sequencing data. Nucleic Acids Research, 40(D1), D54-D56.

Leigh DM, van Rees CB, Millette KL, Breed MF, Schmidt C, Bertola LD, Hand BK, Hunter ME, Jensen EL, Kershaw F, Liggins L, Luikart G, Manel S, Mergeay J, Miller JM, Segelbacher G, Hoban S, Paz-Vinas I. 2021. Opportunities and challenges of macrogenetic studies. Nature Review Genetics, 22: 791–807.

Lewis AC, Molina SJ, Appelbaum PS, Dauda B, Rienzo AD, Fuentes A, Fullerton SM, Garrison NA, Ghosh N, Hammonds EM, Jones DS, Kenny EE, Kraft P, Lee SSJ, Mauro M, Novembre J, Panofsky A, Sohail M, Neale BM, Allen D. 2022. Getting genetic ancestry right for science and society. Science, 376(6590): 250–252.

Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Rmachandran S, Cann HM, Barsh GS, Feldman M, Cavalli-Sforza LL, Myers RM. 2008. Worldwide human relationships inferred from genome-wide patterns of variation. Science, 319(5866): 1100–1104.

Malécot G. 1948. Les mathematiques de l'heredite. Masson and Cie, Paris.

Maruyama T. 1971. The rate of decrease of heterozygosity in a population occupying a circular or a linear habitat. GENETICS, 67(3): 437–454.

Merrell DJ. 1953. Gene frequency changes in small laboratory populations of *Drosophila melanogaster*. Evolution, 7(2): 95–101.

Merimans PG. 2012. The trouble with isolation by distance. Molecular Ecology, 21(12): 2839–2846.

Myers R, Rocha LA, Allen G. 2017. *Amphiprion bicinctus*. *The IUCN Red List of Threatened Species* 2017, e.T18832A1856510. https://dx.doi.org/10.2305/IUCN.UK.2017-2.RLTS.T188320A1857510.en

Nagylaki T. 1975. Conditions for the existence of clines. GENETICS, 80(3): 595–615.

Nagylaki T. 1978. A diffusion model for geographically structured populations. Journal of Mathematical Biology, 6: 375–382.

Nanninga GB, Saenz-Agudelo P, Zhan P, Hoteit I, Berumen ML. 2015. Not finding Nemo: limited reef-scale retention in a coral reef fish. Coral Reefs, 34: 383–392.

Nei M, Tajima F. 1981. DNA polymorphism detectable by restriction endonucleases. GENETICS, 97(1): 145–163.

Nychka D, Furrer R, Paige J, Sain S. 2021. fields: Tools for spatial data. R package version 14.1. https://github.com/dnychka/fieldsRPackage.

Pante E, Simon-Bouhet B. 2013. marmap: A Package for Importing, Plotting and Analyzing Bathymetric and Topographic Data in R. PLoS ONE 8(9): e73051.

Pearse DE, Crandall KA. 2004. Beyond F_{ST}: Analysis of population genetic data for conservation. Conservation Genetics, 5: 585–602.

Peter BM. 2016. Admixture, population structure, and *F*-statistics. GENETICS, 202(4): 1485–1501.

Pickrell J, Pritchard J. 2012. Inference of population splits and mixtures from genome-wide allele frequency data. Nature Precedings, https://doi.org/10.1038/npre.2012.6956.1.

Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. GENETICS, 155(2): 945–959.

Prout T. 1954. Genetic drift in irradiated experimental populations of *Drosophila melanogaster*. GENETICS, 39(4): 529–545.

Pruisscher P, Nylin S, Gotthard K, Wheat CW. 2018. Genetic variation underlying local adaptation of diapause induction along a cline in a butterfly. Molecular Ecology, 27(18): 3613–3626.

R Core Team. 2022. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Ramachandran S, Deshpande O, Roseman CC, Cavalli-Sforza LL. 2005. Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. Proceedings of the National Academy of Sciences, 102(44): 15942–15947.

Ringbauer H, Coop G, Barton NH. 2017. Inferring recent demography from isolation by distance of long shared sequence blocks. GENETICS, 205(3): 1335–1351.

Ringbauer H, Kolesnikov A, Field DL, Barton NH. 2018. Estimating barriers to gene flow from distorted isolation-by-distance patterns. GENETICS, 208(3): 1231–1245.

Rochette NC, Rivera-Colón AG, Catchen JM. 2019. Stacks 2: Analytical methods for paired-end sequencing improve RADseq-based population genomics. Molecular Ecology, 28(21), 4737-4754.

Rogan JE, Parker MR, Hancock ZB, Earl AD, Buchholtz EK, Chyn K, Martina J, Fitzgerald LA. 2023. Genetic and demographic consequences of range contraction patterns during biological annihilation. Scientific Reports, 13(1691): https://doi.org/10.1038/s41598-023-28927-z.

Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, Feldman MW. 2002. Genetic structure of human populations. Science, 298(5602): 2381–2385.

Rousset F. 1997. Genetic differentiation and estimation of gene flow from *F*-statistics under isolation by distance. GENETICS, 145(4): 1219–1228.

Rousset F. 2000. Genetic differentiation between individuals. Journal of Evolutionary Biology, 13: 58–62.

Rousset F, Leblois R. 2012. Likelihood-based inferences under isolation by distance: Two-dimensional habitats and confidence intervals. Molecular Biology and Evolution, 29(3): 957–973.

Saenz-Agudelo P, Dibattista JD, Piatek MJ, Gaither MR, Harrison HB, Nanninga GB, Berumen ML. 2015. Seascape genetics along environmental gradients in the Arabian Peninsula: insights from ddRAD sequencing of anemonefishes. Molecular Ecology, 24(24), 6241-6255.

Shirk AJ, Cushman SA. 2014. Spatially-explicit estimation of Wright's neighborhood size in continuous populations. Frontiers in Ecology and Evolution, 2: https://doi.org/10.3389/fevo.2014.00062.

Stan Development Team. 2023. RStan: the R interface to Stan. R package version 2.21.8, https://mc-stan.org/.

Theodoridis S, Rahbek C, Nogues-Bravo D. 2021. Exposure of mammal genetic diversity to mid-21st century global change. Ecography, 44(6): 817–831.

Wakeley J. 1999. Nonequilibrium migration in human history. GENETICS, 153(4): 1863–1871.

Wang IJ, Bradburd GS. 2014. Isolation by environment. Molecular Ecology, 23(23): 5649–5662.

Wang J, Caballero A. 1999. Developments in predicting the effective size of subdivided populations. Heredity, 82(2): 212–226.

Watterson GA. 1975. On the number of segregating sites in genetical models without recombination. Theoretical Population Biology, 7(2): 256–275.

Wilkins JF. 2004. A separation-of-timescales approach to the coalescent in a continuous population. GENETICS, 168(4): 2227–2244.

Wilkins JF, Wakeley J. 2002. The coalescent in a continuous, finite, linear population. GENETICS, 161(2): 873–888.

Wright S. 1931. Evolution in Mendelian populations. GENETICS, 16(2): 97–159.

Wright S. 1940. Breeding structure in relation to speciation. American Naturalist, 74(752): 232–248.

Wright S. 1943. Isolation by distance. GENETICS, 28(2): 114–138.

Wright S. 1946. Isolation by distance under diverse systems of mating. GENETICS, 31(1): 39–59.

Wright S. 1949. Population structure evolution. Proceedings of the American Philosophical Society, 93(6): 471–478.

Wright S. 1951. The genetical structure of populations. Annals of Eugenics, 15: 323–354.

Yosef R, Nachshonov T, Zduniak P. 2022. Anthropopause positively influenced Red Sea Clownfish (*Amphiprion bicinctus*) populations but not the host sea anemone (*Aciniaria* spp.) in Eilat, Israel. Marine Policy, 145: 105280.

Acknowledgments:

We would like to thank members of the Bradburd lab – Leonard Jones, Meaghan Clark, Alex Lewanski, and Mike Grundler – as well as Teresa Pegan, John Wares, and Cynthia Riginos for helpful feedback on this manuscript. We are also grateful to feedback on developing this method from Luis Zaman and his lab, as well as Peter Ralph. This work was supported in part through computational resources and services provided by the Institute for Cyber-Enabled Research at Michigan State University. Research reported in this publication was supported by the National Institute of General Medical Sciences of the National Institutes of Health under Award Number R35GM137919 (awarded to G.S.B.). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

Data accessibility statement: Scripts for running the Wright-Malécot model and SLiM recipes are available at https://github.com/zachbhancock/WM_model.

Figures:

Figure 1. Relationship between the separation-of-timescales of the coalescent and the isolation-by-distance (IBD) curve in continuous space. A) A population genetic tree showing the relationships between sampled individuals (colored circles) across a continuous landscape with dimensionality (*x*, *y*). Relative to a focal sample (dotted circle), the transition to the collecting phase occurs as the rate of coalescence converges to a neutral Kingman's coalescent. B) An IBD plot relative to the focal individual. The transition to the collecting phase occurs when geographic distance is no longer predictive of genetic distance. The red dotted line denotes *s* (or $1 - \pi_c$), which is the estimated mean minimum relatedness between individuals in the population.

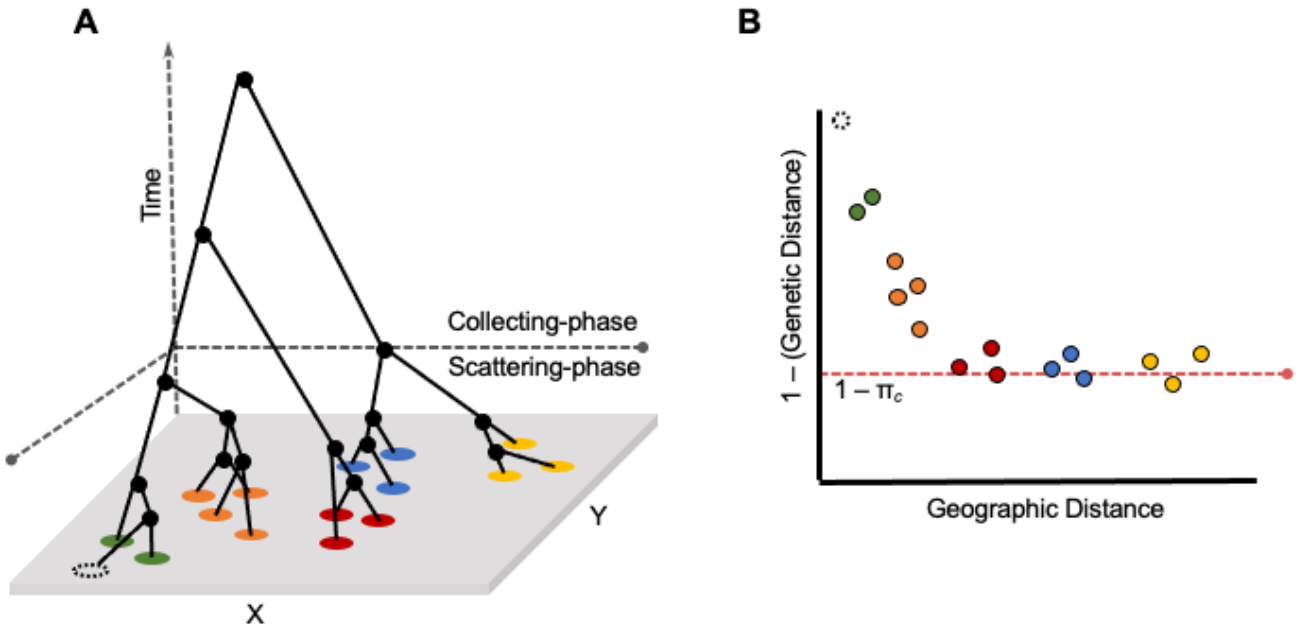


Figure 2. Expected relatedness decay curves of $\Omega_{i,j}$ with distance given s = 0.95 for various values of neighborhood size ("Nbhd"). See Equation 5 in the text.

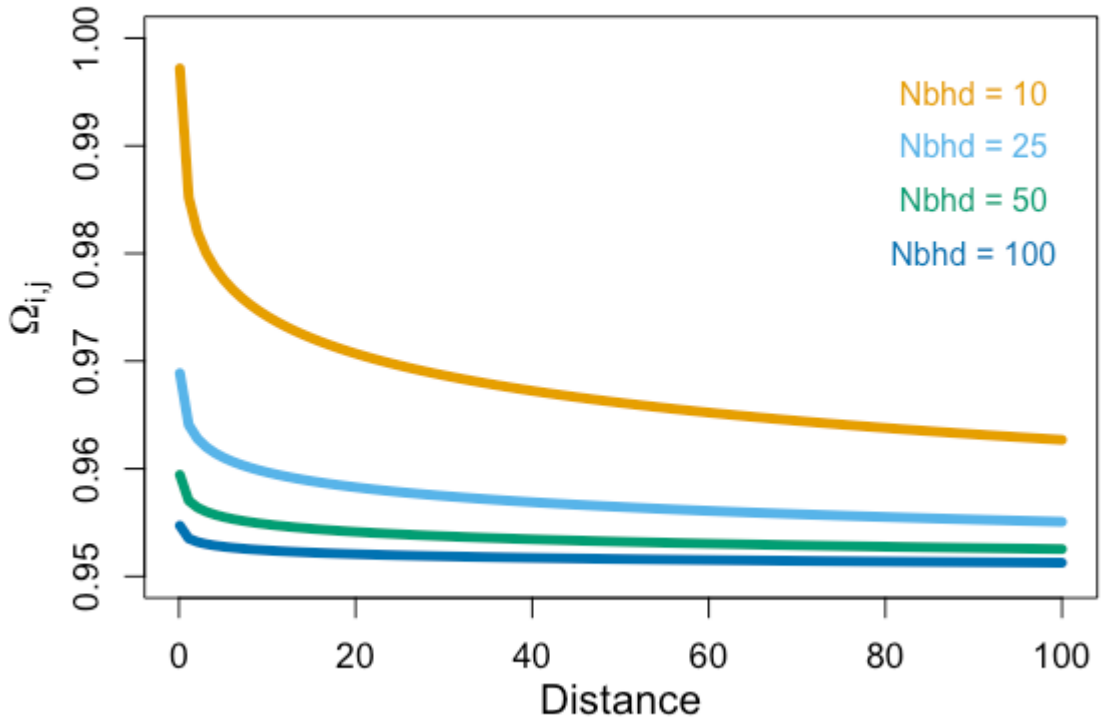


Figure 3. Model accuracy and precision. A) Estimated Wright's neighborhood size ("Nbhd") against the theoretical expectations along the 1:1 dashed line. B) Inferred N_e from π_c against true N_e estimated from the simulations (see text for details). Each panel represents simulated values of dispersal rate (σ), colors are different values of population density (K). Each point represents the posterior density from the best MCMC chain across all 10 simulation replicates.

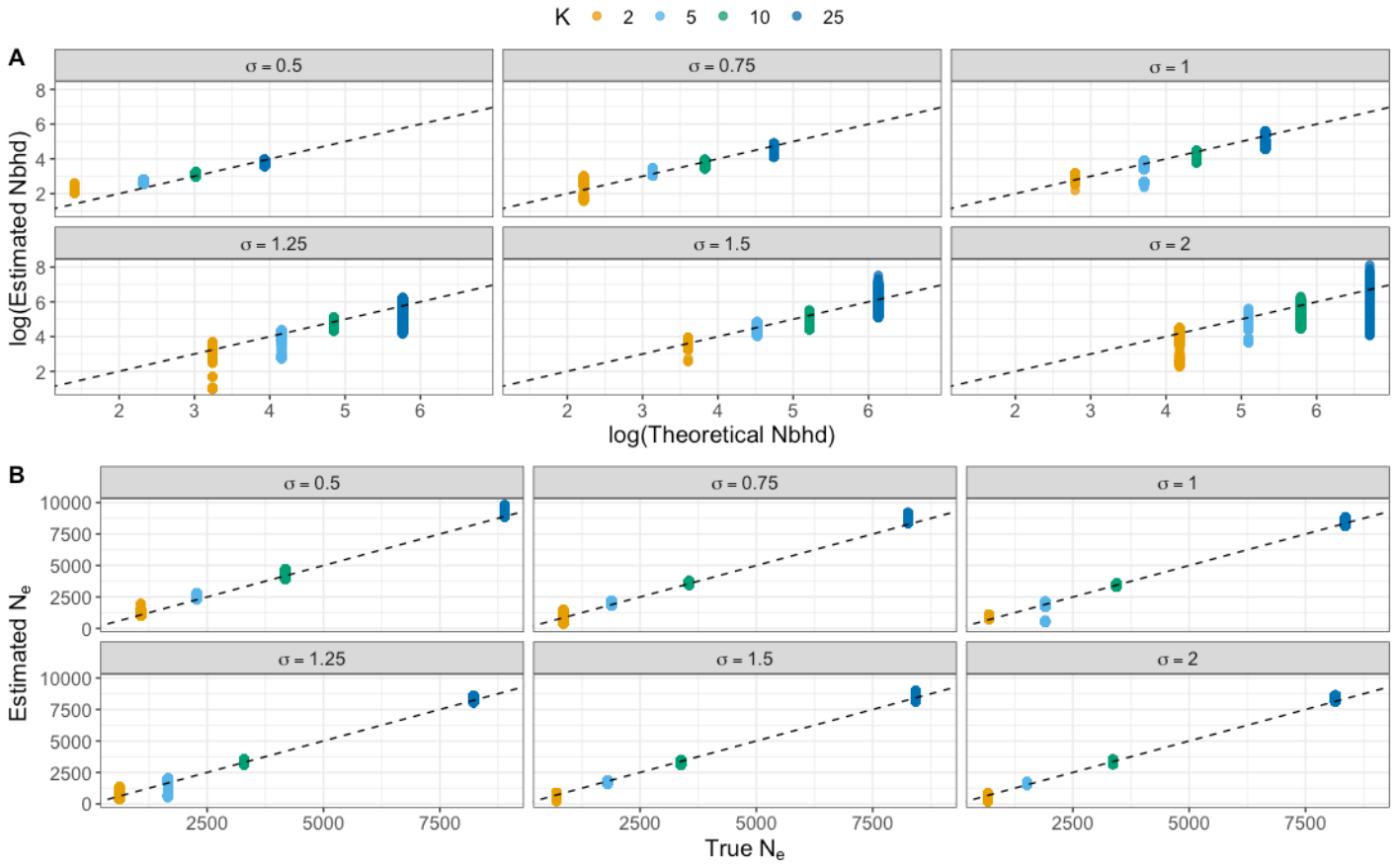


Figure 4. Comparison of the model presented in the text and Rousset's method for estimating $\mathcal{N}$. Black, open circles are the mean estimate using Rousset's method and circles colored by *K* are the estimates under the model presented here. Vertical bars represent the 95% quantile, black for Rousset's method and colored for the model in the text. The right panel shows the full range of Rousset's 95% quantile for a pair of extreme values; the left panel is an inset capturing the narrower range of the 95% quantile of our estimates (colored bars within the longer black bars). The black dotted line represents the 1:1 relationship between the *x* and *y*.

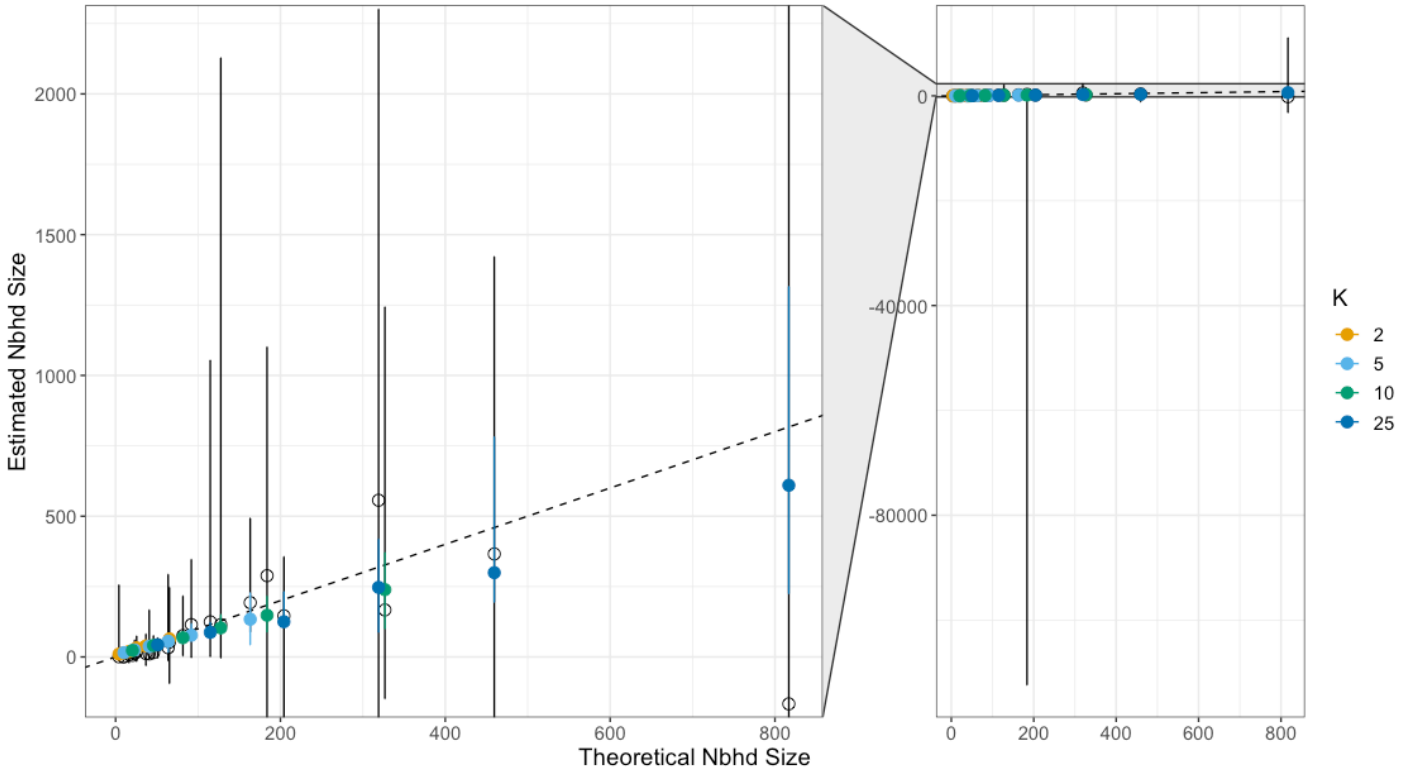
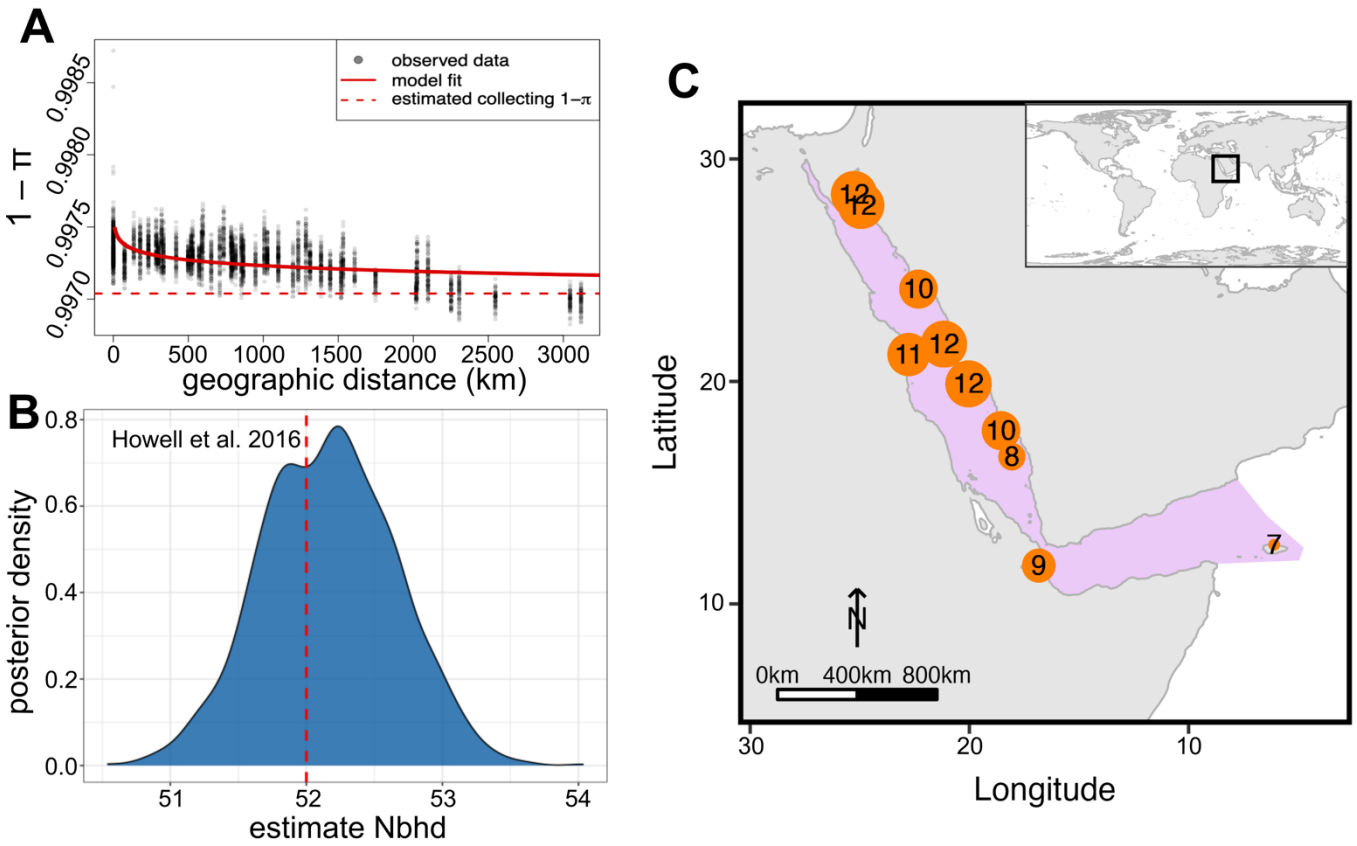


Figure 5. *Amphiprion bicinctus* dataset. A) Model fit, dark circles are the empirical observations and red line is the fit; dashed red line is estimated $1 - \pi_c$. B) Posterior density estimate for $\mathcal{N}$ ("Nbhd"), with the red dashed line representing the number of individuals observed in the year 2015 in the Gulf of Eilat (Howell et al. 2016). C) Sample locations of *A. bicinctus*, with numbers representing sample size and circles scaled by sample size relative to other locations (Saenz-Agudelo et al. 2015).



SUPPLEMENTARY MATERIALS

Table S1. Prior distributions for the model.

Parameter	**Distribution**
s	normal(0, 1)
log($\boldsymbol{\mu}$)	normal(–5, 1)
$\mathcal{N}$	normal(100, 1000)
log(γ)	normal(0, 1)
log(nugget)	normal(0, 1)
lik(pHom)	Wishart(L, pHom)

Figure S1. Estimated diversity during the collecting-phase (π_c).

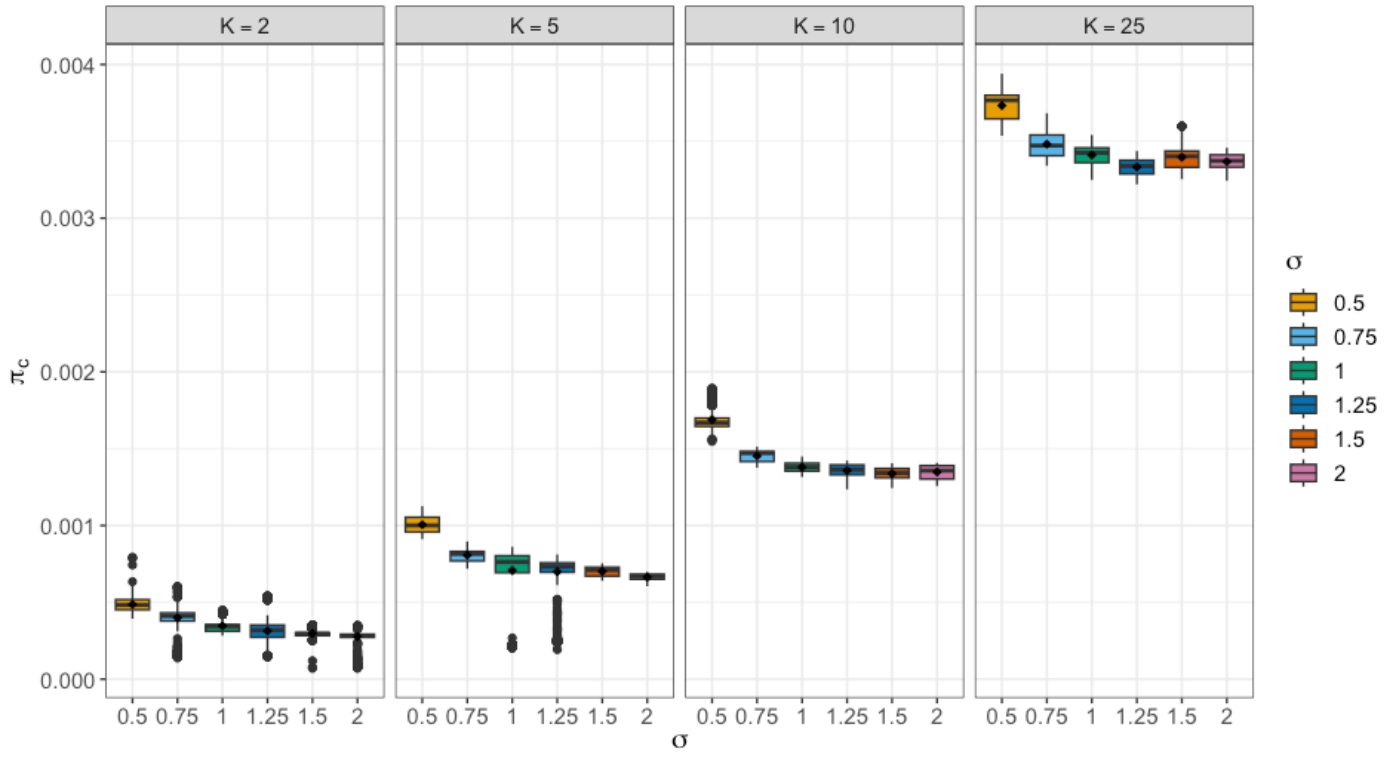


Figure S2. Estimated N_e against theoretical N_e from Eqn. 9. Unlike the results from Fig. 3B, we see the estimated N_e is on average less than the theoretical; this is likely due to a mismatch between the simulated landscape that includes finite edges that varies population density across the range and the theoretical expectation of a homogenous distribution (see main text for further discussion).

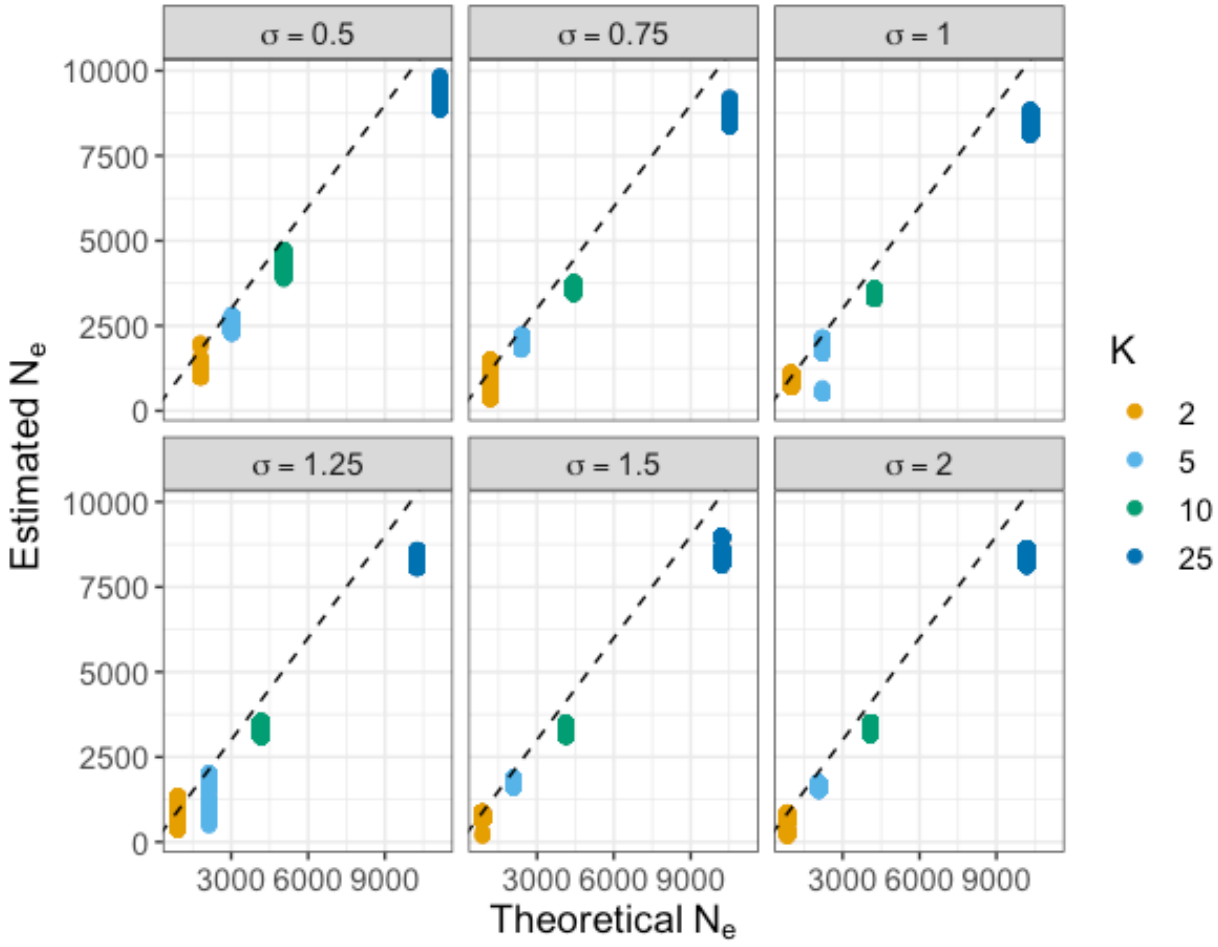


Figure S2. Trace plots of estimated parameter values for the *Amphiprion bicinctus* dataset; *m* is the compound mutation and long-distance migration rate; nbhd is Wright's neighborhood size; *s* is the minimum relatedness between samples (see main text); "inDeme" is the coalescent rate for individuals within κ of one another; and the nugget is individual-level inbreeding.

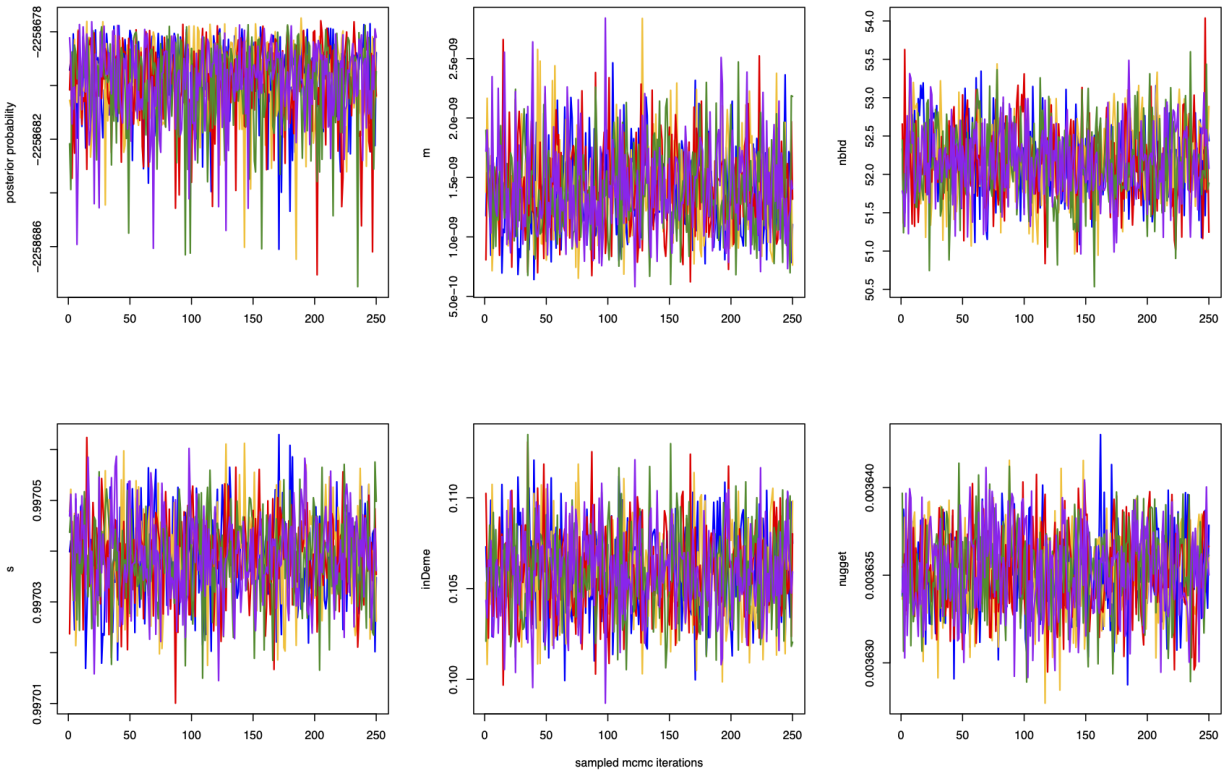


Figure S3. Autocorrelation plots between estimated parameter values for the *Amphiprion bicinctus* dataset; *m* is the compound mutation and long-distance migration rate; nbhd is Wright's neighborhood size; *s* is the minimum relatedness between samples (see main text); "inDeme" is the coalescent rate for individuals within κ of one another; and the nugget is individual-level inbreeding.