# OXFORD GENETICS

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

Zachary B. Hancock,[1,*] Rachel H. Toczydlowski [ID],[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: Zachary B. Hancock, 1105 North University Avenue, Biological Sciences Building, University of Michigan, Ann Arbor, MI 48103, USA. Email: hancockz@umich.edu

Spatially continuous patterns of genetic differentiation, which are common in nature, are often poorly described by existing population genetic theory or methods that assume either 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 in describing the diversity of a population distributed continuously in space; this diversity is intimately linked to both the dispersal potential and the population density 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 two-form bumblebee (*Bombus bifarius*). The model performed well on simulated data relative to alternative approaches and produced reasonable empirical results given the natural history of bumblebees. The resulting inferences provide important insights into the population genetic dynamics of spatially structured populations.

Keywords: isolation by distance; continuous space; effective population size; spatial population genetics; Bayesian

## Introduction

In many species, 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 is significantly more tractable mathematically, and may also have adequately described the model organism populations common in empirical population genetic studies of the time (e.g. *Drosophila* in vials; Merrell 1953; Prout 1954; Dobzhansky and Spassky 1962) but is less well-suited to describing spatially continuous genetic structure.

Several theoretical approaches have relaxed these assumptions by modeling a population as partitioned into demes with some constant rate of migration between them [e.g. Wright's (1931) "island model" and Kimura and Weiss' (1964) "stepping-stone model"]. 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 between them (e.g. Wright 1951; Pritchard *et al.*

2000; Pickrell and Pritchard 2012; Peter 2016). This genetic differentiation is estimated via $F_{ST}$, defined as the variance in allele frequencies within subpopulations relative to the total population. 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, these models (and the statistical approaches they inspired) still maintain panmixia within demes, and hence do not effectively capture population genetic dynamics in continuous space. For example, when patterns of ancestry are truly continuous, 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, a procedure that 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; Novembre *et al.* 2008; Carlson *et al.* 2022; Lewis *et al.* 2022).

Wright (1943) and Malécot (1948) were among 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, and was defined as half the inverse of the probability that two alleles sampled in nearby individuals were derived from the same ancestral allele in the previous generation. Wright (1946) showed that, for populations occupying an infinitely large range, in one-dimension $\mathcal{N} = 2\sqrt{\pi}\sigma\rho$ and in two-dimensions $\mathcal{N} = 4\pi\sigma^2\rho$, where $\pi$ is the mathematical constant, $\sigma$ is the dispersal distance along one axis defined as the standard deviation of a normal distribution with mean zero, and $\rho$ is the local population density. The early theoretical work of Wright and Malécot has been further expanded by Kimura and Weiss (1964); Nagylaki (1975, 1978); Felsenstein (1976); Barton *et al.* (2002); Barton *et al.* (2013), among others.

Despite these theoretical advances, there are far fewer statistical methods for examining populations in continuous space than as discrete entities. 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 for 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 across space and time, it has the useful benefit of providing a single estimate of $\mathcal{N}$ for the entire population (Shirk and Cushman 2014). Several popular programs implement this expected relationship to enable researchers to estimate $\mathcal{N}$ (e.g. SPAGeDI—Hardy and Vekemans 2002; Rousset and Leblois 2012 and GENEPOP—Rousset 2007). Importantly, the decision of which estimator of $F_{ST}$ to use is nontrivial and can produce dramatically different results (Pearse and Crandall 2004; Bhatia *et al.* 2013). Furthermore, researchers must decide whether to estimate $F_{ST}$ between individuals or between 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); the latter, lumping individuals into subpopulations, can smooth this noisiness, 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 and Crandall 2004).

Another quantity that, like Wright's neighborhood size, is useful for understanding and conserving species is the effective population size ($N_e$). While $\mathcal{N}$ is a measure of the local rate of drift and dispersal (analogous to the number of migrants in Wright's island model in two dimensions; Wright 1931), $N_e$ is a global measure, capturing the long-term rate of coalescence. In a conservation setting, $N_e$ may serve as a rough proxy for census size and the adaptive potential of threatened populations (Theodoridis *et al.* 2021; Exposito-Alonso *et al.* 2022). Others have used $N_e$ to investigate its relationship with range size or dispersal ability (De Kort *et al.* 2021; Leigh *et al.* 2021). Estimation of the coalescent effective size, $N_e$, which describes the expected rate of coalescence in a population (Kingman 1982; Wang and Caballero 1999; Sjödin *et al.* 2005; Charlesworth 2009), often relies on its relationship with genetic diversity, which, when the mutation rate

is $\mu$ and the population is at mutation-drift equilibrium, is given by $4N_e\mu$ (Kimura and Crow 1964). While often presented as a single $N_e$ value, the coalescent effective size is more accurately represented as the inverse of the rate of coalescence at time t, $N_e(t)$, and hence technically has no single value. One estimator of the population mutation rate is Wu and Watterson's $\theta_w$, which is S/$a_n$ where S is the number of segregating sites in a sample of $n$ sequences and $a_n = \sum_{i=1}^{n-1} \frac{1}{i}$ (Watterson 1975). However, Wu and Watterson's $\theta_w$ 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 (Wilkins 2004; Battey *et al.* 2020). Another estimator of the population mutation rate, also biased in the presence of spatial structure, 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 $i$th and $j$th sequence, and $\pi_{ij}$ is the proportion of nucleotide differences between the $i$th and $j$th sequence (Nei and Tajima 1981). At mutation-drift equilibrium, and assuming no selection, $\theta_w = \hat{\pi} = 4N_e\mu$.

An important issue emerges in these estimators of $N_e$ in the presence of population structure with respect to sampling scheme. Often, sampling is under-dispersed due to the vagaries of field collection techniques that prioritize easy-to-sample locations and are therefore often spatially under-dispersed relative to the population's range. When sampling is spatially under-dispersed and populations are spatially structured, common estimators of $N_e$ are downwardly biased by the inclusion of nearby individuals who are more closely related than expected by chance. Hence, studies that are interested in comparing estimates of $N_e$ across species may be significantly underestimating long-term $N_e$ in the presence of spatial structure when these estimates rely on summary statistics like $\hat{\pi}$. However, in general, $F_{ST}$ is low, and in many systems, this issue may not play a major role in shaping genetic diversity.

In continuously distributed populations at migration-drift equilibrium, inbreeding $N_e$ is not independent of $\mathcal{N}$, as each is impacted by dispersal (Supplementary Fig. 1; Barton *et al.* 2002; Wilkins 2004). When dispersal is high, $N_e$ converges to the population census size of reproductive adults ($N_c$), but it tends to be much larger when dispersal is low (assuming constant density). Similarly, because higher dispersal leads to more potential parents, it increases $\mathcal{N}$. An ideal model of spatial population structure would thus be individual-based so 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 addition, estimators of $N_e$ within the model would be robust to sampling scheme and, ideally, not force researchers to make arbitrary decisions about which individuals are "too related" to be included within the dataset.

In this paper, we take steps toward this goal by introducing a model that jointly estimates $\mathcal{N}$ and long-term diversity (which is related to $N_e$) from data on pairwise sequence divergence and geographic distance between individuals. We validate the model's behavior using individual-based forward-time simulations and compare its performance against Rousset's (1997). Finally, we apply our model to an empirical dataset of *Bombus bifarius*, the two-form bumblebee, to evaluate its utility in practical applications.

## Methods
### Model intuition

A pedigree is shaped by an organism's life-history, including its dispersal potential, generational structure, and mating strategies.
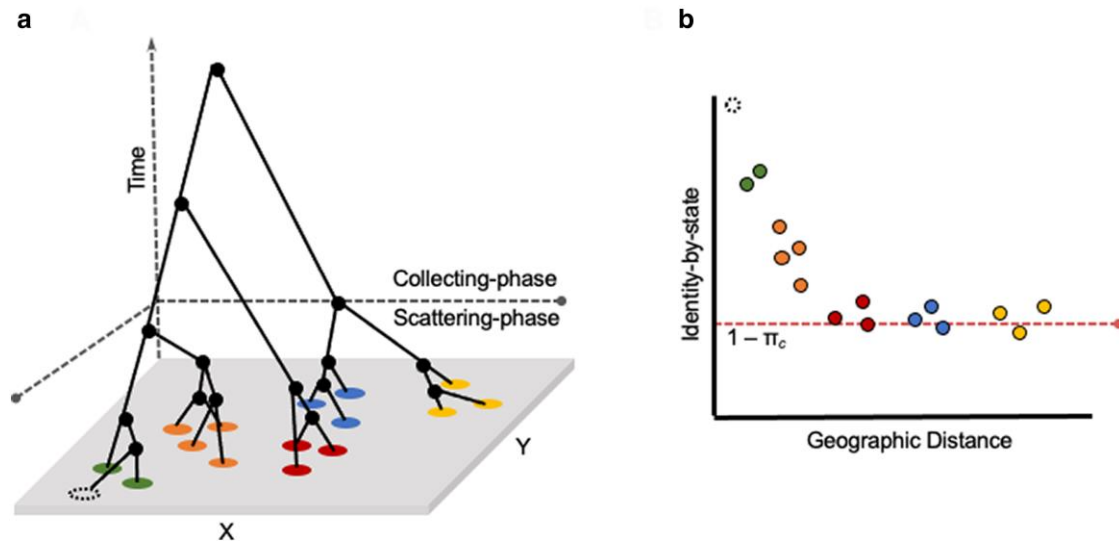
**Fig. 1.** Relationship between the separation-of-timescales of the coalescent and the isolation-by-distance curve in continuous space. a) A single representative gene tree showing the relationships between sampled individuals (colored circles) across a continuous landscape with dimensionality $(x, y)$. 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 coalescence. b) An isolation-by-distance plot relative to the focal individual. The transition to the collecting phase occurs when geographic distance is no longer predictive of genetic distance (i.e. pairwise heterozygosity). The red dotted line denotes s (or $1 - \pi_c$), which is the estimated mean minimum relatedness between individuals in the population.

Because mutations occur along the branches of the genetic genealogies embedded within the 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 pedigree (e.g. selection and 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 correlation between geographic location and genetic ancestry, the set of genetic genealogies contained within the pedigree (the Ancestral Recombination Graph, or "ARG;" Lewanski *et al.* 2024) becomes distorted relative to random-mating expectations (e.g. Anderson-Trocmé *et al.* 2023). In two-dimensions, the coalescent process can be considered as occurring in two distinct phases: the *scattering phase* and the *collecting phase* (Wright 1943; Wakeley 1999; Wilkins and Wakeley 2002; Wilkins 2004). Going backward in time from the present, the scattering phase occurs first, and is characterized by lineages that are geographically near one another coalescing, on average, more rapidly than expected under random mating (i.e. with a probability greater than $1/2N_e$, Wilkins and Wakeley 2002; Wilkins 2004). 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 ARG is Wright's neighborhood size, which, in two-dimensions, is defined as $\mathcal{N} = 4\pi\rho\sigma^2$, where $\pi$ is the mathematical constant, $\rho$ is population density, and $\sigma$ is the standard deviation of the effective dispersal distance along an axis 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 $2\sigma$, 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 a circle of radius $2\sigma$ centered on the focal individual). In Fig. 1a, we show a single representative gene tree that depicts coalescence happening more quickly between nearby individuals than distant individuals (although we note that the pattern of relatedness depicted in Fig. 1b is generated by the aggregation of many gene trees that adhere to this geographical pattern).

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, Wilkins 2004; Bradburd and Ralph 2019). In the collecting phase, the rate of coalescence in a simple model (e.g. the island model, the stepping-stone model, or in continuous, two-dimensional space) is well-represented by a neutral coalescent process (Kingman 1982), in which the rate of coalescence $t$ generations in the past is $1/2N_et$ and the average time to the most recent common ancestor is $4N_e$ generations. Importantly, $N_e$ is not independent of $\sigma$ and will, in general, be larger at lower $\sigma$.

This two-phase distortion of the shape of the ARG 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 dependent on the spatial scale of sampling; the genetic diversity estimated from a sample of nearby (and therefore likely more related than a sample of individuals taken at random from across the species' range) individuals is likely lower than that of the entire population or species. This effect is particularly strong when dispersal is very low relative to the length of the range and the
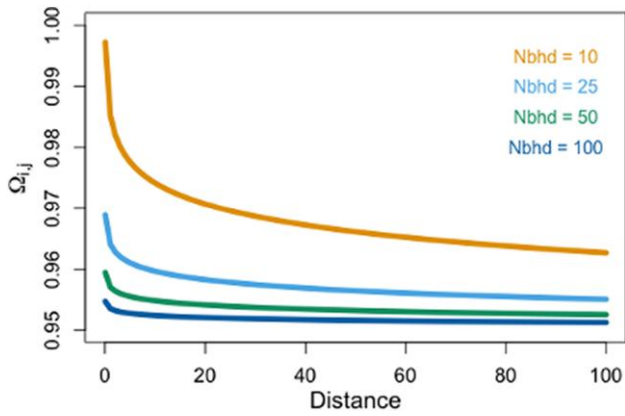
**Fig. 2.** Expected relatedness decay curves of $\Omega_{i,j}$ with distance given $s = 0.95$ for various values of $\mathcal{N}$ (Equation 5 in the text).

geographic area encompassed by the genotyped samples is small (e.g. Exposito-Alonso *et al.* 2022).

Our model (explained below) relies on the relationship between gene genealogies and the pattern of isolation by distance (Fig. 1b, 2). At short geographic distances, there is strong spatial autocorrelation of relatedness—this captures the scattering phase of the ARG, and it decays rapidly (Fig. 1b). As the curve flattens, geographic distance ceases to explain 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 $\pi$", $\pi_c$), and is most informative about long-term inbreeding $N_e$. We define the quantity $\pi_c$ as an estimate of the expected maximum heterozygosity between a pair of individuals in the population. Because a random (or even an exhaustive) sample of a population will contain many pairs of closely related individuals, expected heterozygosity itself should be lower than $\pi_c$; the latter effectively ignores relationships in the scattering phase and hence behaves as if samples were collected in a spatially over-dispersed way (Supplementary Fig. 2).

## Model

We first introduce the model of isolation by distance (IBD) that we use—both its form and its assumptions—and 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 1948; 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 maximum 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 1948) 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. The rate of coalescence between individuals within this radius is determined by the effective population density, $\rho_e$ (Barton *et al.* 2002), which defines the probability of identity by descent $1/(2\rho_e)$ for nearby genes in some previous time slice, integrated over all separations within that "nearby" distance. For habitats that are relatively homogenous, such that geographic distance is the primary explanation of covariance, with constant $\rho$ and $\sigma$ through time and across space, the probability of samples $i$ and $j$ being identical by descent ($F_{ij}$) can be estimated by

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

where $K_0$ is a modified Bessel function of the second kind of order 0, $d_{ij}$ is the geographic distance between samples $i$ and $j$, and $\pi$ is the mathematical constant (Wright 1943, Malécot 1948; Barton *et al.* 2002). Barton *et al.* (2002) note that this approximation diverges as $d_{ij} \to 0$; to account for this, following Ringbauer *et al.* (2017), we designate a short distance, $\kappa$, within which the rate of coalescence becomes a constant $\gamma$ that describes the mean probability of being identical by descent between all pairs of individuals closer than $\kappa$ to each other. There is no optimal value for $\kappa$ as it depends on the local population structure, but is generally of order $\sigma$ (Barton *et al.* 2002).

The classic Wright-Malécot formula (Eqn. 1) describes the theoretical probability of identity-by-descent, which, in an infinite population, decays to zero as the distance between a pair of sampled individuals goes to infinity (see also Maruyama 1972). Our model breaks the assumptions of the Wright–Malécot model in two important ways. First, we assume that populations are finite, meaning that all individuals are identical by descent at some point in the past. 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

$$\hat{H}_{ij} = F_{ij} + (1 - F_{ij})s \tag{2}$$

where $s = \frac{1}{L}\sum_{\ell=1}^{L} \hat{p}_\ell^2 + (1 - \hat{p}_\ell)^2$ and $\hat{p}_\ell$ is the frequency of a particular allele in a spatially over-dispersed sample of individuals at the $\ell$th of $L$ biallelic loci (Ringbauer *et al.* 2018). The quantity $s$ represents the "background" rate of sequence similarity (the mean minimum sequence similarity between any pair of individuals) and can be thought of as the complement of the amount of genetic diversity in a population at equilibrium during the collecting phase. We can therefore define a quantity "collecting phase

$\pi$": $1 - s = \frac{1}{L}\sum_{\ell=1}^{L} 2\hat{p}_\ell(1 - \hat{p}_\ell) = \pi_c$. Collecting phase $\pi$ can be used as an estimator of long-term inbreeding $N_e$, independent of the scattering phase. Relative to $\hat{\pi}$ (mean pairwise genetic distance in a population, Nei and Tajima 1981), estimates of $\pi_c$ (defined by the asymptote of the IBD curve, which showcases the transition to the collecting phase; Fig. 1) should be less sensitive 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 an allele at the $\ell$th locus 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 and Gompert 2012; Ellegren 2014). We assume that SNPs are ascertained without bias; as with other approaches for statistical inference from population genetic data, ascertainment bias may strongly impact results, although in ways that are difficult to predict, depending on the nature of the bias. 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:

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

where, $\hat{H}_{ij}$ 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 $\ell$th locus. Pairwise homozygosity is a measure of absolute genetic similarity, so it is not sensitive to the sampling configuration. Additionally, $\hat{H}$ is proportional to the allelic diversity defined by 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). Note that, in doing so, we are assuming that allele frequencies are multivariate normally distributed across samples and are independent between loci (i.e. not in linkage disequilibrium). Linkage disequilibrium (LD) will have the effect of decreasing the number of independent observations we have, thereby decreasing the actual number of degrees of freedom of the Wishart distribution.

To infer parameter values, 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. Because the individual parameters of the Wright–Malécot model (in particular, $\sigma$ and $\rho$ as well as $\sigma$ and $\mu$) are partially non-identifiable, we instead implement a model

that employs the compound parameters [following Ringbauer et al. (2018)] $\mathcal{N}$, which is neighborhood size, and is defined as $4\pi\rho\sigma^2$, and $m$, which is defined as $\frac{2\mu}{\sigma^2}$ (Barton et al. 2002; Ringbauer et al. 2018). Concretely, we write the probability of identity by descent between samples $i$ and $j$, $F_{ij}$, as

$$F_{ij} = \begin{cases} \gamma & \text{if } d_{ij} \leq \kappa \\ \dfrac{K_0\left(\sqrt{m}d_{ij}\right)}{\mathcal{N}} & \text{if } d_{ij} > \kappa \end{cases} \qquad (4)$$

where $\gamma$ is the probability of being identical by descent at distances $d_{ij} \leq \kappa$ (i.e. short enough that mating might reasonably be considered panmictic), and $F_{ij}$ for $d_{ij} > \kappa$ is given by the Wright-Malécot function introduced in Equation 2. We then construct our parametric homozygosity between individuals $i$ and $j$, $\Omega_{ij}$, as

$$\Omega_{ij} = F_{ij} + (1 - F_{ij})s + \delta_{ij}\eta_i \qquad (5)$$

where $(1 - F_{ij})s$ is the rate of identity by state not due to identity by descent, $\delta_{ij}$ is the Kronecker $\delta$, and $\eta_i$ is a statistical parameter we include in our model to describe inbreeding specific to the $i$th individual. We then calculate our likelihood as

$$P(\Omega) = W(\Omega, \ L) \qquad (6)$$

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

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

$$P(\hat{H}) \propto P(\Omega(\gamma, m, \mathcal{N}, s, \vec{\eta}))P(\gamma)P(m)P(\mathcal{N})P(s)P(\vec{\eta}) \qquad (7)$$

where $\Omega(\gamma, m, \mathcal{N}, s, \vec{\eta})$ denotes the dependence of $\Omega$ on its constituent parameters $(\gamma, m, \mathcal{N}, s, \vec{\eta})$, and $P(\theta)$ denotes the prior probability of a given parameter $\theta$. The parameters of our model are therefore $\mathcal{N}$, $m$, and $\gamma$ (which determine $F$ as defined in Equation 4), as well as $s$ and $\eta$ (which, along with $F$, determine $\Omega$, as defined in Equation 5). Note that $\mathcal{N}$ and $m$ are the two mechanistic (ontological) parameters of our model and entirely determine the expected probability of identity-by-descent between all pairs of individuals found greater than distance $\kappa$ from each other. The parameters $\gamma$ and $\eta$ do not appear in the Wright–Malécot model of isolation by distance (Equation 1) and are included as phenomenological parameters here to facilitate statistical inference. However, we note that $\gamma$, which is intended to capture variation in relatedness over very small spatial scales, and $\eta$, which is intended to capture variation in inbreeding specific to the individual (i.e. above and beyond that found between individuals that occur very close to one another) are closely related to the quantities $F_{ST}$ and $F_{IS}$, respectively. The value of $\kappa$ is specified by the user, a modeling choice we discuss further below. Supplementary Table 1 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 (a type of Markov chain Monte Carlo, or MCMC, approach; Betancourt and Girolami 2015) 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 $m$, $\gamma$, and $\vec{\eta}$ in log space, which should help chains mix over the posterior density. Second, we scale the sample homozygosity $\hat{\Omega}$ so that it varies between 0 and 1 (we apply the same scaling to the parametric homozygosity):

$$
\begin{aligned}
a &= \min(\hat{H}) \\
b &= \max(\hat{H} - a) \\
\hat{H}' &= \frac{\hat{H} - a}{b} \\
\Omega' &= \frac{\Omega - a}{b}
\end{aligned}
\tag{8}
$$

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

## Simulations

We evaluated model performance using individual-based forward-time simulations implemented in SLiM v3.6 (Haller and Messer 2019). Our simulations had non-overlapping generations, diploid and hermaphroditic individuals with haploid genomes of 100 Mbp, and a uniform recombination rate of $10^{-9}$ per base-pair per generation. Mate choice, spatial competition, and dispersal are controlled by a single parameter, $\sigma$. Distance units in SLiM are arbitrary; for convenience, we will refer to them as kilometers. Individuals were simulated on a continuous, two-dimensional $25 \times 25$ km landscape with reflecting boundaries. Total population density was regulated by an enforced local carrying-capacity, $K$, to avoid spatial clumping (Felsenstein 1975). This was achieved by first computing the spatial distance, $d$, between a focal individual and all individuals within distance $3\sigma$. The strength of each interaction is determined by a Gaussian density, $g(d) = \exp(-d^2/2\sigma^2)/(2\sigma^2)$, and the total competitive interactions felt by individual $i$ is $c = \sum_j g(d_{ij})$, where $d_{ij}$ is the distance between individual $i$ and $j$ (Battey et al. 2020). Local population density was then regulated by scaling the fitness of individual $i$ as $fitness_i = K/c$. Fitness was implemented via mortality, which was evaluated before mating occurred within the simulation. To reduce the impact of edge effects, we then divided this strength by the integral of the interaction after clipping by the bounds of the specified landscape (Haller and Messer 2022), which has the desired effect of rescaling competition relative to the occupiable area. Edge-effects and spatial competition collectively reduced the census population size by ~35% relative to $Kw^2$, where $w$ is the width of the square simulated landscape ($w = 25$ km). We also simulated scenarios in which the edge width was set to 50 km, as well as scenarios in which there were no edges (i.e. on a torus) for comparison (Supplementary Fig. 3–5). We implemented spatial mating dynamics in a way similar to spatial competition: mates were chosen within a maximum distance of $3\sigma$, with each potential mate assigned a weight that was an inverse function of its geographic distance from a focal individual. These distances were converted to weights using a Gaussian function with a max $1/2\pi\sigma^2$; the lower the total probability of all mates, the higher the probability that a focal individual might potentially not choose any mate. The number of offspring produced per mating event was drawn from a Poisson distribution with shape parameter $\lambda = 2$. Finally, offspring dispersal was drawn from a normal distribution with mean 0, a standard deviation $\sigma$, and a maximum of $3\sigma$.

We performed simulations across a range of values of $K$ (2, 5, 10, 25) and $\sigma$ (0.5, 0.75, 1.0, 1.25, 1.5, 2.0), where $\sigma$ is the distance moved along a single axis, 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 (note these values are scaled by the observed density instead of the input $K$). We performed 10 simulation 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, which 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 using the parameter-combination specific census population size (Haller et al. 2018). Only 25% of simulations (n = 60) had not coalesced and required recapitation; of these, 59 had two roots, and a single simulation had 3 roots. As such, our decision about which population size we used for recapitation had minimal impact on overall results (Supplementary Fig. 6). 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 randomly sampled 100 individuals alive in the final generation and calculated pairwise $\pi$ (the average genetic divergence between a pair of samples) between each pair of individuals using *tskit* v.0.5.3 (Kelleher et al. 2018). We output the pairwise individual $\pi$ matrix and geographic coordinate matrix, the latter of which was converted into a distance matrix in R (R Core Team 2022) using the function *rdist* in the package *fields* v.2.9–1 (Nychka et al. 2021). These two matrices were then used as input to our statistical model, which we used to perform inference on all model parameters ($\mathcal{N}$, $s$, $\lambda$, and $m$, and $\vec{\eta}$). These values have theoretical expectations given the parameters we chose for each individual simulation, but are not explicitly defined because we were not simulating directly under the inference model.

To examine how recent population fluctuations might impact our model inference, we also investigated simple contraction and expansion scenarios. In each scenario, the value of $\sigma$ was set to 1. The expansion scenario was initialized with $K = 5$ (census population size of ~2200), the contraction scenario with $K = 10$ (census population size of ~4400). For each, after 100,000 generations of the initial population size, we instituted an instantaneous size change, either a doubling (expansion) or halving (contraction) in size, and sampled individuals after a number of generations equal to 0.1, 1, and 10% of the post-change census population size. We simulated 10 replicates of each scenario and followed the same procedure as the constant-size simulations, sampling 100 individuals, re-capitating (using the initial population size) if necessary, and, for each simulation, outputting the pairwise individual $\pi$ matrix and geographic coordinate matrix. Because $\mathcal{N}$ is expected to be a reflection of the dynamics of the scattering-phase, we expected that it should largely reflect the new population census size, whereas $\pi_c$ should approximately reflect diversity prior to the size change, as it captures deeper-time dynamics.

We ran the implementation of our model on each simulated dataset, running four independent MCMC chains of 4,000 steps each. We pruned the first 2,000 steps in each chain as burn-in and thinned the remaining steps by sampling every eighth iteration, for a total of 250 sampled post burn-in iterations per

chain. Because some chains displayed poor mixing, we visually verified that the chains had achieved convergence by inspecting the trace plots of the posterior probability and parameter values of interest, then selected the (well-behaved) chain with the highest mean posterior probability as the one from which to report results. For purposes of testing the model against simulations, we set $\kappa = 0.25$ km.

Simulations and model estimation were performed on the Advanced Research Computing (ARC) Great Lakes High Performance Computing Cluster at the University of Michigan. Code, including the ARC HPCC SLURM workflow, SLiM recipes, and Python scripts, can be found at https://github.com/zachbhancock/WM_model.

Finally, we compared our estimated values to the true $\mathcal{N}$, which we estimate as $4\pi\rho_e\sigma_e^2$. Here, $\rho_e$ is calculated as $2N_p/A$, where $N_p$ is the number of parents in the generation preceding sampling, $A$ is the area of the simulation, and $\sigma_e$ is the maximum likelihood estimate of the variance of parent-offspring dispersal (the average squared parent-offspring displacement in each dimension); both $\rho_e$ and $\sigma_e$ are calculated for each simulation. We include a comparison of $\mathcal{N}$ as calculated using the parameter values we input into SLiM to this effective $\mathcal{N}$ in Supplementary Fig. 7. To evaluate our performance in estimating $\pi_c$, we compared our model estimate with mean pairwise $\pi$ calculated between all sampled pairs of individuals that were at least 20 km away; the number of pairs compared varied by simulation because samples were taken randomly with respect to space. This distance was chosen because it is the minimal distance that ensured in all simulations, irrespective of the value of $\sigma$, there was no longer spatial signal in the isolation-by-distance curves, indicating that the populations had transitioned to the collecting phase. The distance at which the isolation-by-distance curve levels off varies between simulations is a function of the value of $\sigma$ used to simulate the data; we pick a single cutoff to apply across all simulations for ease of comparison. We explored how the choice of this cutoff distance value affected estimates of $\pi_c$ by also using cutoffs of 15 and 25 km. We found that, when we used a distance cutoff of 15 km, we included significantly more related individuals, leading to an overestimation of $\pi_c$ relative to the sample, while at distances >25 km, there was no discernable difference in estimated diversity compared to samples >20 km (Supplementary Fig. 8).

## Empirical dataset

We also tested the performance of our method on a published empirical spatial population genetic dataset of 383 individuals of the two-form bumblebee *Bombus bifarius* (Jackson *et al.* 2018a). Bumblebees have been widely studied due to their ecological and commercial importance as pollinators (Garibaldi *et al.* 2013). Like many hymenopterans, bumblebees are eusocial, and hence each nest represents a single reproductive female. Furthermore, there is evidence that some bumblebee species (including *B. bifarius*) tend to avoid mating with nestlings (Foster 1992); the number of nests within a given dispersal distance can thus be a reasonable proxy for $\mathcal{N}$ in these species.

We analyzed a published RADseq dataset for *B. bifarius* generated by Jackson *et al.* (2018a), in which the full molecular methods used to generate the data can be found. Briefly, sequence reads were generated from thoracic muscle tissue using RADseq and restriction enzyme *PstI* and sequenced single-end in multiplexed libraries (1 × 100 bp, Illumina HiSeq 2,000 or 4000). We downloaded the raw (demultiplexed) sequence reads for each of the 383 *B. bifarius* individuals archived in the National Center for Biotechnology Information (BioProject PRJNA473221, SRP149031;

Jackson *et al.* 2018b) using the fasterq-dump function in the SRA Toolkit (V2.10.7, Kodama, Shumway, and Leinonen 2012). We dropped reads with uncalled bases and for which the mean Phred score was < 15 (sliding window 15% of read length; Stacks2 process_radtags module V2.54—Rochette, Rivera-Colón, and Catchen 2019). We confirmed that there were no over-represented sequences (sequences present in > 5% of reads per individual, averaged across all individuals) and trimmed all reads to a uniform length (80 bp). Next, we assembled these reads *de novo* using Stacks2 and nine combinations of assembly parameters (v.2.54, Rochette, Rivera-Colón, and Catchen 2019). We used the R80 rule (Paris, Stevens, and Catchen 2017) to select an optimal assembly (in this case, when parameters were: $m = 3$, $M = 3$, and $n = 3$) and dropped loci that were scored in <50% of individuals. This resulted in 58,970 loci that were used for the subsequent analyses. We then calculated pairwise $\pi$ and pairwise geographic distance between each pair of individuals in the dataset. Geographic distance here is measured as great circle distance (*rdist.earth* function of the *fields* R package; v.14.1—Nychka *et al.* 2021). Lastly, we chose $\kappa = 1$ km for our empirical dataset; as discussed above, this theoretically represents the distance within which mating is panmictic (less than the per-generation dispersal distance of reproducing individuals). The scale of distance in this dataset is kilometers, and thus individuals are modeled as equally likely to be related if they are within 1 km of one another, which covers the scale of sample sites in Jackson *et al.* (2018a) but is much less than the distance between localities.

For comparison, we also estimated neighborhood size using Rousset's method. To do so, we compared two measures of $F_{ST}$: a SNP-based estimation from Jackson *et al.* (2018a), in which individuals are subsetted into "populations" (sample locations); and an individual-based estimation using the pairwise $\pi$ matrix used as input for our model. In this second, individual-based approach, we estimated $F_{ST}$ as $1 - 2(\pi_i + \pi_j)/(\pi_i + 2\pi_{ij} + \pi_j)$, where $\pi_i$ is individual diversity, and $\pi_{ij}$ is the divergence between $i$ and $j$. We regressed $F_{ST}$ against both geographic distance and log(geographic distance) and estimated $\mathcal{N}$ as $1/\beta$, where $\beta$ was the slope of that regression.

## Results

Our model performed well on both the simulated and empirical datasets (Supplementary Fig. 9–10). In the former, the model converged on the theoretical expectations for both $\mathcal{N}$ and $\pi_c$, which can be used to estimate $N_e$. 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 bumblebees.

## Simulation results

The individual-based simulations were performed by varying both dispersal distance ($\sigma$) and local carrying-capacity ($K$), which generated a range of $\mathcal{N}$ from 6.25 to 861.81. Given the simulated geographic range area, this range of values encompasses extreme/strong spatial structure at the lower end and almost panmictic populations at the upper end.

At almost all simulated values of $\sigma$ and $K$, the 95% equal-tailed credible interval of $\mathcal{N}$ consistently included the theoretical $\mathcal{N}$; however, at larger $\sigma$ 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 detectable signatures of
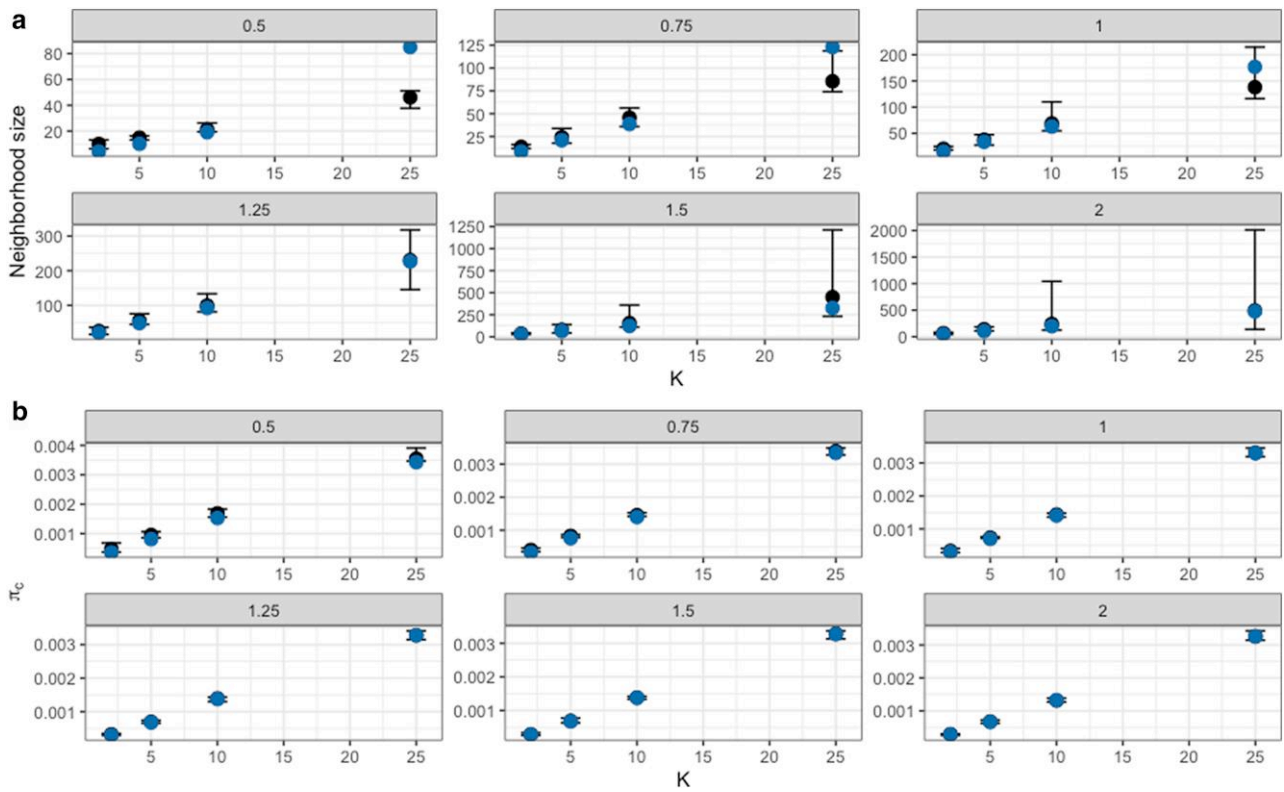
**Fig. 3.** Model accuracy and precision. a) Estimates of Wright's neighborhood size for each value of $K$ and $\sigma$ (gray header); black circles are medians and error bars are the 95% quantile. Blue circles are the theoretical $\mathcal{N}$. b) As in (a), but for estimates of $\pi_c$. Blue circles represent $\pi$ when $d_{ij} > 20$.

covariance between geography and ancestry, which decays at large $\mathcal{N}$ (Wilkins 2004) as illustrated in Fig. 2 (see also Supplementary Fig. 11). At small dispersal distances, the variance in our estimates of $\mathcal{N}$ was relatively low even at higher $K$. However, at larger dispersal distances, the variance subsequently increased regardless of $K$, indicating that $\sigma$ likely has the largest impact on model precision. Notably, our model estimated a value of $\mathcal{N}$ greater than the true value in the simulation only at the lowest K and $\sigma$ combination. In this region of parameter space, the Wright–Malécot model we have implemented is a poor approximation of relatedness. For datasets with small $\mathcal{N}$, it would be more appropriate to use the small $\mathcal{N}$ approximation to the Wright–Malécot model (Eqn 15, Barton *et al.* 2002). We discuss this further below. Although we found that Rousset's method is an unbiased estimator, the variance in its estimates across replicates is much larger at lower K and $\sigma$ pairs than that of our model; Rousset's method can even generate negative values, the interpretation of which is murky (Fig. 4).

On simulated scenarios with recent population size changes, the model performed reasonably well (Supplementary Fig. 12–13). We slightly underestimated $\mathcal{N}$ in the expansion scenario (relative to the theoretical value of the post-expansion conditions); this underestimation was most pronounced at the most recent sampled time point after the population size had changed (0.1% $N_c$ generations; Supplementary Fig. 13). Similarly, in the contraction scenario, we accurately estimated the new $\mathcal{N}$ at all times except at 0.1% $N_c$ generations, when it was slightly overestimated. Our model accurately estimated the expected deep-time levels of diversity ($\pi_c$), which is the diversity prior to the population size change (Supplementary Fig. 13). Collectively, these results indicate that our estimation of $\mathcal{N}$ reasonably tracks very recent population size changes, whereas $\pi_c$ consistently is influenced

by deeper-time dynamics and is insensitive to recent demography.

The model also performed well at estimating $\pi_c$ (with the true value set as $\pi_{d_{ij}>20}$ or $\pi_{d_{ij}>25}$, see Supplementary Fig. 8), and did so with little variance across simulations. In the simulations with the lowest $\sigma$, there was a slight overestimation of $\pi_c$, which may indicate that the empirical IBD curve still maintains a subtle signal at $d_{ij} > 20$; this would cause the model to extrapolate that the IBD curve hits its asymptote at greater distances, possibly beyond the range edge. Interestingly, we also find that the levels of $\pi_c$ are largely affected by values of $K$ rather than values of $\sigma$ (with the exception of the lowest $\sigma$). Theory predicts 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 (Wilkins 2004). The deviation from theory that we observed likely results from the fact that, in our models, K changes by a larger magnitude than $\sigma$ (with respect to the total simulated area) across simulation parameter values.

## Empirical results

For the empirical dataset of *Bombus bifarius* (two-form bumblebee), we performed five independent MCMC analyses, each with 5,000 iterations. The model mixed well (Supplementary Fig. 14), but the joint marginal plots showed that several parameters were interacting: $m$ and $s$ were positively correlated, and $\gamma$ was negatively correlated with $m$ and $s$ (Supplementary Fig. 15). Furthermore, there was a weak but significant ($R^2 = 0.02512$; $P = 1.056\text{e-}8$) negative relationship between $\mathcal{N}$ and $m$. Similarly, there was a very weak but significant ($R^2 = 0.0028$; $P = 0.0318$) relationship between $\pi_c$ and $\mathcal{N}$. These relationships reflect relationships between parameters in our model, despite which we were nonetheless able to infer parameter values well.
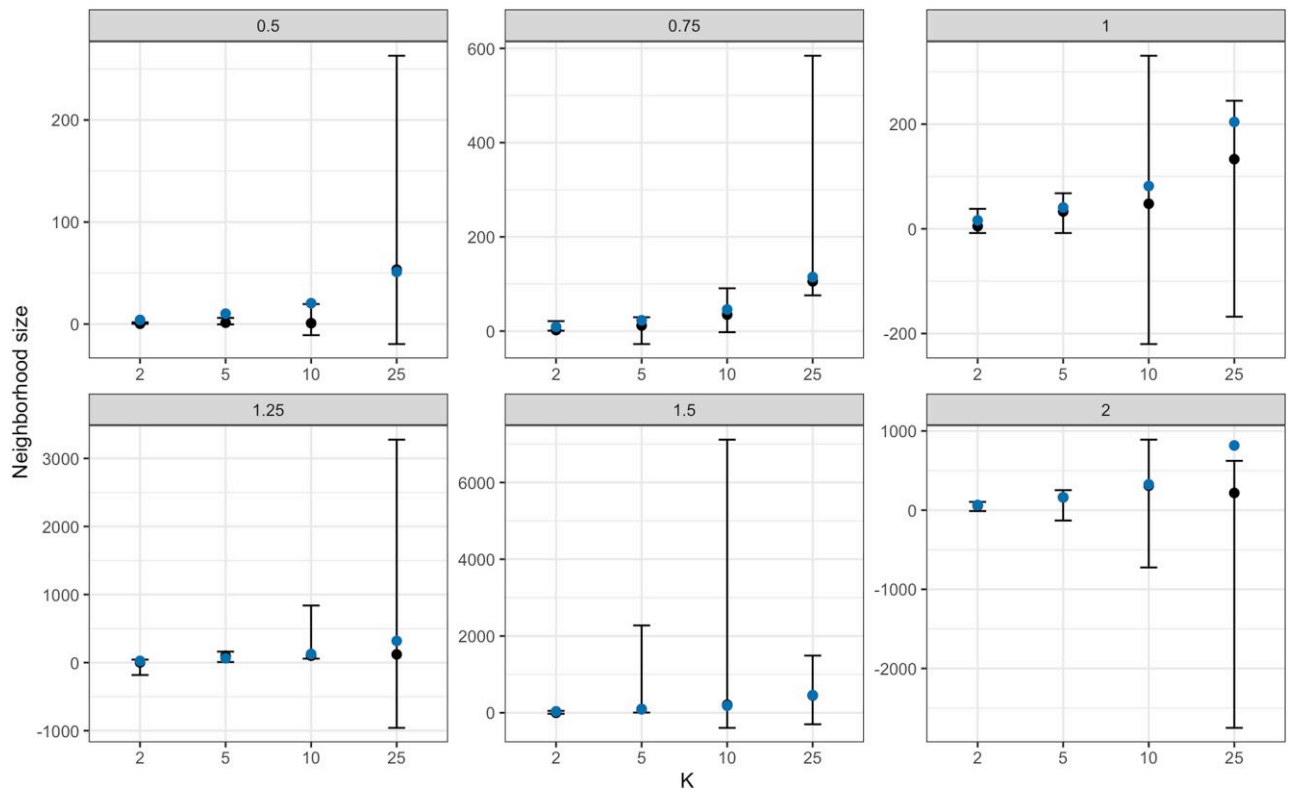
**Fig. 4.** Results of Rousset's method for estimating $\mathcal{N}$ across datasets simulated with different values of K (population density in SLiM; x-axis) and $\sigma$ (dispersal; gray header). As in Fig. 3a, black dots and bars are the estimated values with the median and 95% quantile; blue dots are the true $\mathcal{N}$ for each simulation. Notice the axes differ greatly between Fig. 3a and this figure, demonstrating the wide variance in Rousset's estimator. Supplementary Figs. 15–17 show Rousset's estimator conditioned by various maximum distance thresholds.

We estimated that $\hat{\mathcal{N}} = 137.18$ ($\pm 0.9$) and $\hat{\pi}_c = 0.002$ ($\pm 2\text{e}{-}6$) for this two-form bumblebee dataset (Fig. 5). While the true $\mathcal{N}$ is unknown for most species, given the performance of the model on simulated data and the natural history of bumblebees, we think this is a reasonable estimate. We explore the biological implications (and caveats) of these results further in the Discussion.

## Discussion

An extensive literature exists on the biases that spatial structure introduces to commonly employed population genetic summary statistics (e.g. Miermans 2012; Bradburd and Ralph 2019; Battey et al. 2020). However, many studies still use measures of effective population size (such as Wu and Watterson's $\theta_w$ and $\pi$) that are naïve to population structure (and the geographic sampling of individuals) in empirical systems that are spatially structured. Other statistical population genetic approaches attempt to account for population structure by discretizing the habitat into demes defined by sampling region or violation of Hardy–Weinberg (e.g. STRUCTURE—Pritchard et al. 2000; Montana and 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 and Weiss 1964). Indeed, even in a fine-scaled lattice population, Battey et al. (2020) showed that biases in estimates of Wu and Watterson's $\theta$ and Tajima's $D$ emerge as the sample size per deme approaches the deme 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. 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 $\hat{\pi}$ (Wilkins 2004). Our model generates estimates of a diversity statistic ($\pi_c$) 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.

## Simulations

Our model performs well over a wide range of true $\mathcal{N}$, though the variance of our estimates of $\mathcal{N}$ increases as the true $\mathcal{N}$ grows large. In contrast, Rousset's (1997) estimator consistently had extremely large variances and would often take negative values (Fig. 4), particularly when dispersal was high (resulting in low differentiation across the sampled range). At small $\mathcal{N}$, our model overestimates $\mathcal{N}$ slightly; this is likely due to the fact that we have implemented the "large $\mathcal{N}$" approximation of the Wright–Malécot model, which must be modified for small $\mathcal{N}$ as shown in Equation 15 of Barton et al. (2002). However, the departure between the truth and our model estimates is relatively small, so we feel like this shortcoming is outweighed by the convenience to the empirical user of being able to implement a single inference
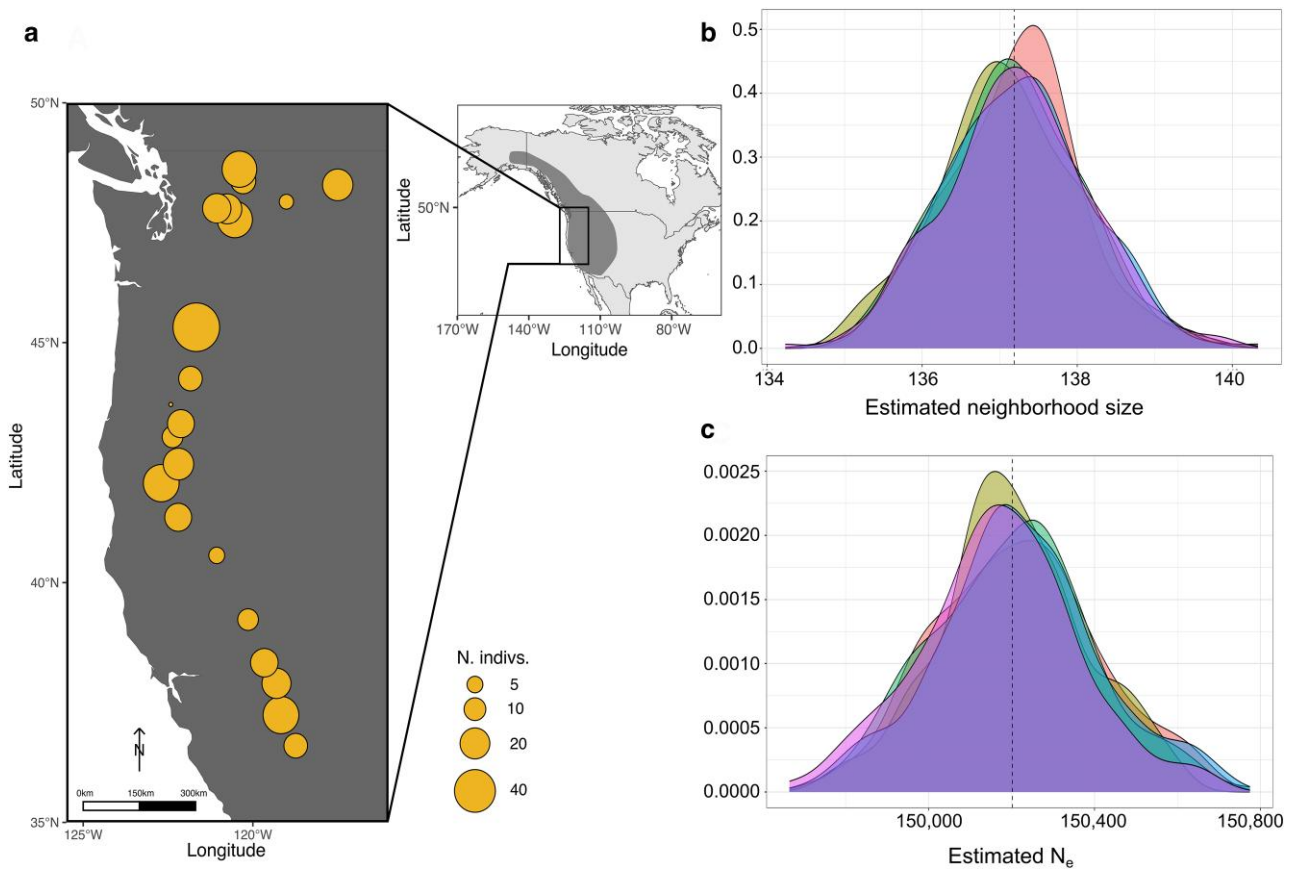
**Fig. 5.** Empirical application of the method. a) Sample sites of *B. bifarius* from Jackson *et al.* (2018a), size of circles represent the number of individuals sampled, and dark gray shading is the known range extent. b) Density of $\mathcal{N}$ estimates across all five chains. c) Density of $N_e$ estimates, where $N_e = \hat{\pi}_c/4\mu$, and $\mu$ is the empirically estimated mutation rate for bumblebees (Liu *et al.* 2017).

model, rather than one that might vary based on the specifics of their system (and their choice of $\kappa$).

In addition, in our simulations, the spatial scale of dispersal, mate choice, and competition were all set to be the same (controlled by a single parameter, $\sigma$). This will obviously not be the case for most empirical systems; however, we believe that even in such systems, our method will still be able to accurately infer the standard deviation of effective dispersal.

## Empirical application

Few empirical systems have reliable estimates of effective population density and dispersal distance such that estimates of Wright's neighborhood size made using genetic data can be directly compared against a known biological quantity. However, because bumblebees are an important commercial pollinator, their populations are better characterized than many natural systems, and we can use this natural history knowledge to inform the interpretation of our model-based estimates. Goulson *et al.* (2010) examined bumblebee nest density across a $10 \times 20$ km area for *B. lapidarius* and *B. pascuorum*, two species that are closely related to our model species (*Bombus bifarius*). On the scale of their sampling, they found significant $F_{ST}$ between, but not within, sites, where each "site" is a $200 \times 10$-meter strip within the $10 \times 20$ km censused area, indicating that the distance between sampling site locations might reasonably be expected to exceed the average dispersal distance in a generation. Within sites, they estimated a mean of 114 nests per site for *B. lapidarius* and 87 nests per site for *B. pascuorum*. Because bumblebees are eusocial, each nest

represents a single mating female, and queens tend to avoid mating with nestlings (Foster 1992). If we therefore assume that the mean number of observed nests per site is roughly the same as the pool of possible parents within two dispersal distances of a focal individual in *B. lapidarius* and *B. pascuorum*, our estimate of $\mathcal{N}$ in *B. bifarius* ($\hat{\mathcal{N}} = 137.18$ ($\pm 0.9$)) is quite close to the observed values of $\mathcal{N}$ in two closely related species. One thing to note is that the uncertainty associated with this estimate (as well as other estimates output by the model) is a function of the number of degrees of freedom specified for the Wishart distribution, which, in this model, is the number of specified loci. Because that number can be very large, it can lead to unrealistically low uncertainty in parameter estimates, which should be interpreted with caution.

We applied Rousset's method to the SNP dataset from Jackson *et al.* 2018a both assuming a two-dimensional population (i.e. with geographic distance logged) and one-dimensional (raw geographic distance). The former produced an estimate of $\hat{\mathcal{N}} = 34141 \pm 1821494$ (Supplementary Fig. 16), with the latter estimating $\mathcal{N}$ as $66979$ ($\pm 6887$) (Supplementary Fig. 17; see also Supplementary Fig. 18–20). In light of the extremely high variance in Rousset's estimator's results from both our simulations and in this empirical example, it is difficult to evaluate the accuracy of these results. We also do not know the extent to which edge effects are impacting this system; however, given that our model performs well in the presence of edges, we do not think these will systematically bias our estimates.

We are making a number of assumptions regarding the comparability of these census values taken from *B. lapidarius* and

*B. pascuorum* and our estimate of $\mathcal{N}$ in *B. bifarius*. First, we are assuming that belonging to the same genus is predictive of similarity in $\mathcal{N}$ between species, which seems reasonable, but for which, to our knowledge, there is little empirical support. Second, we are assuming that each "site" in Goulson *et al.*'s (2010) study represents an area approximately equal to that of a circle of radius $2\sigma$, such that the number of nests within a site can be thought of as the pool of possible parents within two dispersal distances of a focal individual. Third, and perhaps most crucially, we are assuming that there is a good concordance between these censused quantities and the effective $\mathcal{N}$ that can be estimated from genetic data. There are many reasons why this assumption may be violated, most notably, if there is either a high variance in reproductive output between queens, such that the effective density differs from the census density, or if there is a relationship between dispersal distance and fitness, such that effective $\sigma$ differs systematically from the standard deviation of the forward-time parent-offspring dispersal distance distribution. The validity of these assumptions is difficult to evaluate without more data taken across generations, but it is nonetheless reassuring that our estimate of $\mathcal{N}$ in *B. bifarius* is of the correct order of magnitude compared to the census neighborhood sizes in two closely related species.

### 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. First, 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 (Wang and Bradburd 2014; Bradburd *et al.* 2013; Ringbauer *et al.* 2018). 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 and Rosenthal 2021).

Another important consideration is that the Wright–Malécot model relies on assumptions that are mutually incompatible (independent dispersal and homogeneous population density). In SLiM, we modeled populations with local density dependence to overcome Felsenstein's "pain in the torus" (Felsenstein 1975) and maintain a stochastically 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 the true simulation neighborhood size, indicating that the Wright–Malécot formulation that underpins our model is robust to this violation.

Finally, our model relies on a user-defined $\kappa$, 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 $\kappa$, but not at low $\kappa$ (Supplementary Fig. 21). This is due to the fact that, at higher $\kappa$, less of the scattering phase is being captured by the model, leading to poor mixing. Ideally, $\kappa$ would be estimated like the other parameters of the model; however, an attempt to implement such a model led to dramatic increases in computation time, and thus for the current work we opted to set a constant $\kappa$.

## Conclusions

For many organisms, geographical distance influences dispersal, competition, and 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.

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). Unlike Rousset's estimator, our method shows good performance across values of $\mathcal{N}$, and offers the additional benefit of generating an estimator of the long-term effective population size ($N_e$) that is insensitive to recent demographic shifts. The introduced model produced reasonable estimates of the theoretical expectations from simulated data and performed well on an empirical dataset of two-form bumblebees. Future work will aim to develop an *R* package for ease of use for researchers (though code for performing the presented model is available in the Github link).

## Data Availability

## Acknowledgments

## Funding

## Conflicts of interest

The author(s) declare no conflict of interest.

## Literature cited

Alexander DH, Novembre J, Lange K. 2009. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 19(9): 1655–1664. doi:10.1101/gr.094052.109.

Anderson-Trocmé L, Nelson D, Zabad S, Diaz-Papkovich A, Kryukov I, Baya N, Touvier M, Jeffery B, Dina C, Vézina H, *et al.* 2023. On the genes, genealogies, and geographies of Quebec. Science. 380(6647):849–855. doi:10.1126/science.add5300.

Barton NH, Depaulis F, Etheridge AM. 2002. Neutral evolution in spatially continuous populations. Theoretical Population Biol. 61(1): 31–48. doi:10.1006/tpbi.2001.1557.

Barton NH, Etheridge AM, Véber A. 2013. Modelling evolution in a spatial continuum. J Stat Mech: Theory Exp. 2013(01|1):1–38. doi:10.1088/1742-5468/2013/01/P01002.

Barton NH, Kelleher J, Etheridge AM. 2010. A new model for extinction and recolonization in two dimensions: quantifying phylogeography. Evolution. 64(9):2701–2715. doi:10.1111/j.1558-5646.2010.01019.x.

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. doi:10.1534/genetics.120.303143.

Betancourt MJ, Girolami M. 2015. Hamiltonian Monte Carlo for hierarchical models. Curr Trends Bayesian Methodol Appl. 79(30): 2–4.

Bhatia G, Patterson N, Sankararaman S, Price AL. 2013. Estimating and interpreting $F_{ST}$: the impact of rare variants. Genome Res. 23(9):1514–1521. doi:10.1101/gr.154831.113.

Bradburd GS, Coop GM, Ralph PL. 2018. Inferring continuous and discrete population genetic structure across space. Genetics. 210(1): 33–52. doi:10.1534/genetics.118.301333.

Bradburd GS, Ralph PL. 2019. Spatial population genetics: it's about time. Annu Rev Ecol Evol Syst. 50:427–449. doi:10.1146/annurev-ecolsys-110316-022659.

Bradburd GS, Ralph PL, Coop GM. 2013. Disentangling the effects of geographic and ecological isolation on genetic differentiation. Evolution. 67(11): 3258–3273. doi:10.1111/evo.12193.

Buerkle CA, Gompert Z. 2012. Population genomics based on low coverage sequencing: how low should we go? Mol Ecol. 22(11): 3028–3035. doi:10.1111/mec.12105.

Carlson J, Henn BM, Al-Hindi DR, Ramachandran S. 2022. Counter the weaponization of genetics research by extremists. Nature. 610(7932):444–447. doi:10.1038/d41586-022-03252-z.

Charlesworth B. 2009. Effective population size and patterns of molecular evolution and variation. Nat Rev Genet. 10(3):195–205. doi:10.1038/nrg2526.

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. Nat Commun. 12(1):516. doi:10.1038/s41467-021-20958-2.

Dobzhanksy T, Snassky NP. 1962. Genetic drift and natural selection in experimental populations of *Drosophila pseudoobscura*. Proc Nat Acad Sci U S A. 48(2):148–156. doi:10.1073/pnas.48.2.148.

Ellegren H. 2014. Genome sequencing and population genomics in non-model organisms. Trend Ecol Evol. 29(1):51–63. doi:10.1016/j.tree.2013.09.008.

Exposito-Alonso M, Booker TR, Czech L, Gillespie L, Hateley S, Kyriazis CC, Lang PLM, Leventhal L, Nogues-Bravo D, Pagowski

V, *et al.* 2022. Genetic diversity loss in the Anthropocene. Science. 377(6613):1431–1435. doi:10.1126/science.abn5642.

Felsenstein J. 1975. A pain in the torus: some difficulties with models of isolation by distance. Am Nat. 109(967):359–368. doi:10.1086/283003.

Felsenstein J. 1976. The theoretical population genetics of variable selection and migration. Ann Rev Genet. 10(1):253–280. doi:10.1146/annurev.ge.10.120176.001345.

Foster RL. 1992. Nestmate recognition as an inbreeding avoidance mechanism in bumble bees (Hymenoptera: Apidae). J Kansas Entomol Soc. 65(3):238–243. https://www.jstor.org/stable/25085362.

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? J App Ecol. 46(2):493–505. doi:10.1111/j.1365-2664.2008.01606.x.

Garibaldi LC, Steffan-Dewenter I, Winfree R, Aizen MA, Bommarco R, Cunningham SA, Kremen C, Carvalheiro LG, Harder LD, *et al.* 2013. Wild pollinators enhance fruit set of crops regardless of honey bee abundance. Science. 339(6127):1608–1611. doi:10.1126/science.1230200.

Goulson D, Lepais O, O'Connor S, Osborne JL, Sanderson RA, Cussans J, Goffe L, Darvill B. 2010. Effects of land use at a landscape scale on bumblebee nest density and survival. J App Ecol. 47(6): 1207–1215. doi:10.1111/j.1365-2664.2010.01872.x.

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. Mol Ecol Res. 19(2):552–566. doi:10.1111/1755-0998.12968.

Haller BC, Messer PW. 2019. Slim 3: forward genetic simulations beyond the Wright-Fisher model. Mol Biol Evol. 36(3):632–637. doi:10.1093/molbev/msy228.

Haller BC, Messer PW. 2022. Slim 4: multispecies co-evolutionary modeling. Am Nat. 201(5):E127–E139. doi:10.1086/723601.

Hardy OJ, Vekemans X. 2002. SPAGedi: a versatile computer program to analyse spatial genetic structure at the individual or population levels. Mol Ecol Res. 2(4):618–620. doi:10.1046/j.1471-8286.2002.00305.x.

Jackson JM, Pimsler ML, Oyen KJ, Koch-Uhuad JB, Herndon JD, Strange JP, Dillon ME, Lozier JD. 2018a. Distance, elevation and environment as drivers of diversity and divergence in bumble bees across latitude and altitude. Mol Ecol. 27(14):2926–2942. doi:10.1111/mec.14735.

Jackson JM, Pimsler ML, Oyen KJ, Koch-Uhuad JB, Herndon JD, Strange JP, Dillon ME, Lozier JD. 2018b. Biogeography and population genetics of diversity and divergence in two North American bumble bees species. National Center for Biotechnology Information. BioProject Accession: PRJNA473221.

Jofre GI, Rosenthal GG. 2021. A narrow window for geographic cline analysis using genomic data: effects of age, drift, and migration on error rates. Mol Ecol Res. 21(7):2278–2287. doi:10.1111/1755-0998.13428.

Kelleher J, Etheridge AM, McVean G. 2016. Efficient coalescent simulation and genealogical analysis for large sample sizes. PLoS Comput Biol. 12(5):e1004842. doi:10.1371/journal.pcbi.1004842.

Kelleher J, Thornton KR, Ashander J, Ralph PL. 2018. Efficient pedigree recording for fast population genetics simulation. PLoS Comput Biol. 14(11):e1006581. doi:10.1371/journal.pcbi.1006581.

Kimura M, Crow JF. 1964. The number of alleles that can be maintained in a finite population. Genetics. 49(4):725–738. doi:10.1093/genetics/49.4.725.

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. doi:10.1093/genetics/49.4.561.

Kingman JFC. 1982. The coalescent. Stoch Proc App. 13(3):235–248. doi:10.1016/0304-4149(82)90011-4.

Kodama Y, Shumway M, Leinonen R. 2012. The sequence read archive: explosive growth of sequencing data. Nucl Acid Res. 40(D1): D54–D56. doi:10.1093/nar/gkr854.

Leigh DM, van Rees CB, Millette KL, Breed MF, Schmidt C, Bertola LD, Hand BK, Hunter ME, Jensen EL, Kershaw F, *et al.* 2021. Opportunities and challenges of macrogenetic studies. Nat Rev Genet. 22(12):791–807. doi:10.1038/s41576-021-00394-0.

Lewanski AL, Grundler MC, Bradburd GS. 2024. The era of the ARG: an introduction to ancestral recombination graphs and their significance in empirical evolutionary genomics. Plos Genetics. 20(1):e1011110.

Lewis AC, Molina SJ, Appelbaum PS, Dauda B, Rienzo AD, Fuentes A, Fullerton SM, Garrison NA, Ghosh N, Hammonds EM, *et al.* 2022. Getting genetic ancestry right for science and society. Science. 376(6590):250–252. doi:10.1126/science.abm7530.

Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Rmachandran S, Cann HM, Barsh GS, Feldman M, Cavalli-Sforza LL, *et al.* 2008. Worldwide human relationships inferred from genome-wide patterns of variation. Science. 319(5866):1100–1104. doi:10.1126/science.1153717.

Liu H, Jia Y, Sun X, Tian D, Hurst LD, Yang S. 2017. Direct determination of the mutation rate in the bumblebee reveals evidence for weak recombination-associated mutation and approximate rate constancy in insects. Mol Biol Evol. 34(1):119–130. doi:10.1093/molbev/msw226.

Malécot G. 1948. Les Mathematiques de L'heredite. Paris: Masson and Cie.

Maruyama T. 1972. Rate of decrease of genetic variability in a two-dimensional continuous population of finite size. Genetics. 70(4):639–651. doi:10.1093/genetics/70.4.639.

Merimans PG. 2012. The trouble with isolation by distance. Mol Ecol. 21(12):2839–2846. doi:10.1111/j.1365-294X.2012.05578.x.

Merrell DJ. 1953. Gene frequency changes in small laboratory populations of *Drosophila melanogaster*. Evolution. 7(2):95–101. doi:10.2307/2405744.

Montana G, Hoggart C. 2007. Statistical software for gene mapping by admixture linkage disequilibrium. Brief Bioinform. 8(6): 393–395.

Nagylaki T. 1975. Conditions for the existence of clines. Genetics. 80(3):595–615. doi:10.1093/genetics/80.3.595.

Nagylaki T. 1978. A diffusion model for geographically structured populations. J Mathemat Biol. 6(4):375–382. doi:10.1007/BF02463002.

Nei M, Tajima F. 1981. DNA polymorphism detectable by restriction endonucleases. Genetics. 97(1):145–163. doi:10.1093/genetics/97.1.145.

Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko A, Auton A, Indap A, King KS, Bergmann S, Nelson MR, *et al.* 2008. Genes mirror geography within Europe. Nature. 456(7218):98–101. doi:10.1038/nature07331.

Nychka D, Furrer R, Paige J, Sain S. Fields: tools for spatial data. R package version 14.1. https://github.com/dnychka/fieldsRPackage2023 2021.

Paris JR, Stevens JR, Catchen JM. 2017. Lost in parameter space: a road map for STACKS. Method Ecol Evol. 8(10):1360–1373. doi:10.1111/2041-210X.12775.

Pearse DE, Crandall KA. 2004. Beyond $F_{ST}$: analysis of population genetic data for conservation. Conservat Genet. 5(5):585–602. doi:10.1007/s10592-003-1863-4.

Peter BM. 2016. Admixture, population structure, and *F*-statistics. Genetics. 202(4):1485–1501. doi:10.1534/genetics.115.183913.

Pickrell J, Pritchard J. 2012. Inference of population splits and mixtures from genome-wide allele frequency data. PLoS Genet. 8(11):e1002967. doi:10.1038/npre.2012.6956.1.

Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. Genetics. 155(2): 945–959. doi:10.1093/genetics/155.2.945.

Prout T. 1954. Genetic drift in irradiated experimental populations of *Drosophila melanogaster*. Genetics. 39(4):529–545. doi:10.1093/genetics/39.4.529.

Pruisscher P, Nylin S, Gotthard K, Wheat CW. 2018. Genetic variation underlying local adaptation of diapause induction along a cline in a butterfly. Mol Ecol. 27(18):3613–3626. doi:10.1111/mec.14829.

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. Proc Nat Acad Sci U S A. 102(44):15942–15947. doi:10.1073/pnas.0507611102.

R Core Team. 2022. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Ringbauer H, Coop G, Barton NH. 2017. Inferring recent demography from isolation by distance of long shared sequence blocks. Genetics. 205(3):1335–1351. doi:10.1534/genetics.116.196220.

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. doi:10.1534/genetics.117.300638.

Rochette NC, Rivera-Colón AG, Catchen JM. 2019. Stacks 2: analytical methods for paired-end sequencing improve RADseq-based population genomics. Mol Ecol. 28(21):4737–4754. doi:10.1111/mec.15253.

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. doi:10.1126/science.1078311.

Rousset F. 1997. Genetic differentiation and estimation of gene flow from *F*-statistics under isolation by distance. Genetics. 145(4): 1219–1228. doi:10.1093/genetics/145.4.1219.

Rousset F. 2000. Genetic differentiation between individuals. J Evol Biol. 13(1):58–62. doi:10.1046/j.1420-9101.2000.00137.x.

Rousset F. 2007. GENEPOP'007: a complete re-implementation of the GENEPOP software for windows and Linux. Mol Ecol Res. 8(1): 103–106. doi:10.1111/j.1471-8286.2007.01931.x.

Rousset F, Leblois R. 2012. Likelihood-based inferences under isolation by distance: two-dimensional habitats and confidence intervals. Mol Biol Evol. 29(3):957–973. doi:10.1093/molbev/msr262.

Shirk AJ, Cushman SA. 2014. Spatially-explicit estimation of Wright's neighborhood size in continuous populations. Fron Ecol Evol. 2: 1–12. doi:10.3389/fevo.2014.00062.

Sjodin P, Kaj I, Krone S, Lascoux M, Nordborg M. 2005. On the meaning and existence of an effective population size. Genetics. 169(2): 1061–1070.

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. doi:10.1111/ecog.05588.

Wakeley J. 1999. Nonequilibrium migration in human history. Genetics. 153(4):1863–1871. doi:10.1093/genetics/153.4.1863.

Wang IJ, Bradburd GS. 2014. Isolation by environment. Mol Ecol. 23(23):5649–5662. doi:10.1111/mec.12938.

Wang J, Caballero A. 1999. Developments in predicting the effective size of subdivided populations. Heredity (Edinb). 82(2):212–226. doi:10.1038/sj.hdy.6884670.

Watterson GA. 1975. On the number of segregating sites in genetical models without recombination. Theor Pop Biol. 7(2):256–275. doi:10.1016/0040-5809(75)90020-9.

Wilkins JF. 2004. A separation-of-timescales approach to the coalescent in a continuous population. Genetics. 168(4):2227–2244. doi:10.1534/genetics.103.022830.

Wilkins JF, Wakeley J. 2002. The coalescent in a continuous, finite, linear population. Genetics. 161(2):873–888. doi:10.1093/genetics/161.2.873.

Wright S. 1931. Evolution in Mendelian populations. Genetics. 16(2):97–159. doi:10.1093/genetics/16.2.97.

Wright S. 1940. Breeding structure in relation to speciation. Am Nat. 74(752):232–248. doi:10.1086/280891.

Wright S. 1943. Isolation by distance. Genetics. 28(2):114–138. doi:10.1093/genetics/28.2.114.

Wright S. 1946. Isolation by distance under diverse systems of mating. Genetics. 31(1):39–59. doi:10.1093/genetics/31.1.39.

Wright S. 1949. Population structure in evolution. Proc Am Phil Soc. 93(6):471–478. https://www.jstor.org/stable/3143336.

Wright S. 1951. The genetical structure of populations. Ann Eugen. 15(1):323–354. doi:10.1111/j.1469-1809.1949.tb02451.x.

*Editor: N. Barton*