Estimating Barriers to Gene Flow from Distorted Isolation-by-Distance Patterns

Harald Ringbauer,^{*,1} Alexander Kolesnikov,^{*} David L. Field,[†] and Nicholas H. Barton^{*}

^{*}Institute of Science and Technology Austria, Klosterneuburg A-3400, Austria, and [†]Department of Botany and Biodiversity Research, University of Vienna, A-1030, Austria

ORCID IDs: 0000-0002-4884-9682 (H.R.); 0000-0002-4014-8478 (D.L.F.); 0000-0002-8548-5240 (N.H.B.)

ABSTRACT In continuous populations with local migration, nearby pairs of individuals have on average more similar genotypes than geographically well-separated pairs. A barrier to gene flow distorts this classical pattern of isolation by distance. Genetic similarity is decreased for sample pairs on different sides of the barrier and increased for pairs on the same side near the barrier. Here, we introduce an inference scheme that uses this signal to detect and estimate the strength of a linear barrier to gene flow in two dimensions. We use a diffusion approximation to model the effects of a barrier on the geographic spread of ancestry backward in time. This approach allows us to calculate the chance of recent coalescence and probability of identity by descent. We introduce an inference scheme that fits these theoretical results to the geographic covariance structure of biallelic genetic markers. It can estimate the strength of the barrier as well as several demographic parameters. We investigate the power of our inference scheme to detect barriers by applying it to a wide range of simulated data. We also showcase an example application to an *Antirrhinum majus* (snapdragon) flower-color hybrid zone, where we do not detect any signal of a strong genome-wide barrier to gene flow.

KEYWORDS barriers to gene flow; isolation by distance; identity by descent; demographic inference

MANY populations are distributed across geographically extended habitats that are sometimes interrupted by barriers to gene flow. They can arise due to physical obstacles that reduce migration, but can also be caused by genetic incompatibilities, which reduce gene flow across a hybrid zone (Barton 1979). Barriers can prevent locally adapted populations from being swamped by dispersal and they can facilitate divergence, ultimately leading to speciation. Therefore, they play a central role not only in conservation, but also in evolutionary biology and ecology. As direct observations of individual movement and reproduction are time consuming, expensive, and can only give a snapshot in time, there is much interest in indirect methods that infer such barriers from observed geographic genetic structure.

Such methods to detect barriers from genetic data can be grouped into two distinct approaches (Guillot *et al.* 2009): clustering methods, which detect geographic genetic discontinuities

between populations by grouping individuals into population units based on genetic similarity (Dupanloup *et al.* 2002; Falush *et al.* 2003; Guillot *et al.* 2005); and edge-detection methods, which identify areas of sharp genetic change (Womble 1951; Manni *et al.* 2004; Cercueil *et al.* 2007). Neither of these approaches is directly linked to any spatial population genetic model. They can therefore infer the existence of a barrier, but cannot give meaningful and biologically interpretable estimates of its strength. In addition, these approaches are often confounded by isolation-by-distance patterns (Safner *et al.* 2011; Meirmans 2012), whereby individuals nearby are more similar than distant individuals (Wright 1943) due to recent coancestry. While the description of this effect has a long history in theoretical population genetics of homogeneous populations (Malécot 1948; Slatkin 1993; Rousset 1997; Hardy and Vekemans 1999; Barton *et al.* 2002), it has not been included into a practically applicable method to estimate the strength of a barrier to gene flow.

Here, we fill this gap and introduce a method that infers the strength of a barrier in a two-dimensional population by fitting a population genetic model. Our method uses the fact that a barrier to gene flow distorts classical isolation-by-distance patterns (Figure 1). Based on the theoretical work of Nagylaki (1988),

Copyright © 2018 by the Genetics Society of America

doi: <https://doi.org/10.1534/genetics.117.300638>

Manuscript received October 19, 2017; accepted for publication December 23, 2017; published Early Online January 8, 2018.

Supplemental material is available online at www.genetics.org/lookup/suppl/doi:10.1534/genetics.117.300638/-/DC1.

¹Corresponding author: Institute of Science and Technology Austria, Am Campus 1, 3400 Klosterneuburg, Austria. E-mail: hringbauer@ist.ac.at

Figure 1 Geographic fluctuations of allele frequencies. We simulated individual biallelic genotype data for one locus in a two-dimensional habitat in absence of a barrier to gene flow (left) and in presence of a strong barrier (right). We used the backward simulation scheme on a grid outlined below with an axial variance of dispersal $\sigma^2 = 1$, one genotype per deme, a long-distance migration rate $\mu = 0.001$, and a complete barrier $\kappa = 0$. Top: Discrete genotype data. The two colors in each grid position code for the two possible allelic states. Bottom: Allele frequencies are smoothed with a Gaussian kernel for better visualization. This figure demonstrates the underlying idea of our method: a barrier distorts random geographic fluctuations of allele frequencies. The average strength of these local fluctuations increases next to the barrier, and there is less correlation across the barrier. This signal can be used to infer the strength of the barrier.

Barton (2008) constructed a theoretical framework. He showed that in two spatial dimensions, where fluctuations of allele frequencies are more localized than in one dimension, these effects of a barrier on allele-frequency fluctuations can already be significant for intermediate barrier strengths. This signal therefore has great potential for demographic inference. The derivation of Barton (2008) also shows that the effect of a barrier depends primarily on short-lived, localized fluctuations. In general, isolation-by-distance patterns equilibrate relatively quickly and depend mostly on recent demography (Barton *et al.* 2013; Aguillon *et al.* 2017). Therefore, an inference scheme based on distorted isolation-by-distance patterns infers contemporary barriers to gene flow, and should be robust to confounding effects of ancestral structure.

Here, we first expand previous theoretical results that describe the effect of a barrier on classical isolation-by-distance patterns (Nagylaki 1988; Barton 2008). We introduce a model where ancestry diffuses backward in time and is partially reflected by a barrier. This allows us to numerically calculate the probability of recent coancestry, which can then be fitted to genetic data. As single nucleotide polymorphism (SNP) data sets are currently widely used, we develop and implement ways to fit such biallelic genetic markers.

We test our inference scheme on synthetic data simulated under an explicit population genetics model and investigate how it is affected by confounding factors, for instance in a scenario of secondary contact. We also show a practical application in which we apply our inference scheme to a hybrid zone population of

Antirrhinum majus, in which a sharp transition in flower color and a causal flower-color gene occurs (Whibley *et al.* 2006). We apply our method to test whether there is also a genome-wide barrier to contemporary gene flow. To this end, we analyze a data set of 12,389 individuals and 60 suitable SNP markers.

Materials and Methods

We first outline the model underlying our inference scheme and discuss its assumptions, and then describe how our method fits this model to observed genotype data. In brief, we use a diffusion approximation for the spread of ancestry to calculate the probability of recent identity by descent between pairs of samples. We then fit our model to data by finding the demographic parameters that maximize the fit of observed homozygosity between all sample pairs.

Model

No model can capture all complexities of the real demographic history of a population. Therefore, the aim is not to have a mathematically exact model, but one that robustly captures general patterns of spatial fluctuations of allele frequencies. We use a model of a two-dimensional continuous habitat that is interrupted by a barrier and assume that the demographic parameters are the same on both sides. In short, we calculate the chance of pairwise coalescence before a long-distance migration or mutation event. We use a diffusion model to trace lineages backward in time and assume that rare long-distance

migration events, which counteract the buildup of local allele-frequency fluctuations, occur at a constant rate. To calculate the equilibrium identity-by-descent pattern, *i.e.*, the probability that two lineages coalesce before the occurrence of a rare long-distance migration or mutation event, we first derive the coalescence probability at specific times t in the past and then integrate over t .

Our model is closely related to previous theoretical treatments of allelic identity by state in presence of a barrier to gene flow (Nagylaki 1988; Barton 2008). For a population occupying a linear habitat, Nagylaki (1988) derived continuous equations for identity by state by taking the limit of a model of linear demes that exchange migrants (the so-called stepping-stone model). Barton (2008) expanded this by solving an analogous equation for two-dimensional populations. His formulas are given as a numerical Fourier transform that diverges for nearby individuals. These equations for two-dimensional populations are formally problematic as they were not obtained by rescaling (which is impossible in two spatial dimensions), but Barton (2008) demonstrated that the solution is in close agreement with the solutions from a discrete stepping-stone model for all but very close distances.

Here, we base our inference scheme on a different approach. We use a diffusion approximation to describe the spread of ancestry backward in time. While the results are formally equivalent, we found our model to be computationally more robust and, most importantly, more efficient. This approach allows us to apply our method to sample sizes of hundreds to thousands of individuals, and dozens to hundreds of loci.

Diffusion approximation: We model the spread of ancestry using a geographic diffusion approximation, which has a long history in population genetics (Fisher 1937; Wright 1943; Malécot 1948; Nagylaki 1978). Tracing a lineage of one locus backward in time, the total spatial movement is the sum of many independent migration events. If these events are sufficiently uncorrelated, the central limit theorem establishes that the total displacement tends toward a normal distribution. Therefore, the overall spread of ancestry can be approximated as a random walk process. This approximation is accurate as long as rare large-scale events do not significantly influence the movement of ancestral lineages. The diffusion approximation is expected to be most accurate on recent to intermediate timescales, on which large-scale events such as colonizations often play only a minor role.

In the absence of a barrier, the process is a free diffusion and the probability density function of finding an ancestor at position x at time t back along a given axis is given by a Gaussian probability density function around the current position x_0 and variance $\sigma^2 t$ that increases linearly backward in time:

$$G_0(x, x_0, t) := \frac{1}{\sqrt{2\pi\sigma^2 t}} \exp\left(-\frac{(x-x_0)^2}{2\sigma^2 t}\right). \quad (1)$$

The dispersal rate σ describes the speed of the spread of ancestry. In the case of a homogeneous density of individuals across the landscape, the backward dispersal probability den-

sity equals the probability density of lineages moving forward in time. In this particular case, the dispersal rate σ^2 can be interpreted as the axial variance of the one generational dispersal kernel (Rousset 1997). The diffusion approximation is fully determined by the equation:

$$\frac{\partial}{\partial t} G(x, x_0, t) = \frac{\sigma^2}{2} \frac{\partial^2}{\partial x^2} G(x, x_0, t), \quad (2)$$

with initial condition:

$$G(x, x_0; t = 0) = \delta(x - x_0). \quad (3)$$

Partial barrier to gene flow: We model a barrier as a partially permeable barrier to diffusion of ancestry. For a barrier at $x = 0$, the following interface boundary conditions have to be supplied (Grebakov *et al.* 2014):

$$\begin{aligned} \lim_{\epsilon \rightarrow 0} \left(\frac{\partial}{\partial x} G(x, x_0, t) \right)_{x=+\epsilon} &= \lim_{\epsilon \rightarrow 0} \left(\frac{\partial}{\partial x} G(x, x_0, t) \right)_{x=-\epsilon} \\ &= \frac{2\kappa}{\sigma^2} \left(\lim_{\epsilon \rightarrow 0} G(\epsilon, x_0; t) - \lim_{\epsilon \rightarrow 0} G(-\epsilon, x_0; t) \right). \end{aligned} \quad (4)$$

The first line describes the constancy of the flux across the barrier. For $\kappa = 0$, there is no flux across the barrier and the barrier is infinitely strong. On the other hand, the case $\kappa = \infty$ implies the continuity of the probability density across the barrier and the solution reduces to free diffusion. Comparing to the differential equation 54 in Nagylaki (1988), which is derived by rescaling a stepping-stone model, gives an intuitive interpretation of κ : This parameter corresponds to the fraction of successful migrants across a barrier if demes are spaced one dispersal unit apart. The quotient $\gamma := \frac{2\kappa}{\sigma^2}$ corresponds to an equivalent factor in equation 54 of Nagylaki (1978). It is also the inverse of the barrier strength parameter B defined by Barton and Bengtsson (1986), which has the dimension of distance.

Equations 2–4 allow for an analytic solution for the probability density with a barrier (Grebakov *et al.* 2014):

$$\begin{aligned} G(x, x_0, t) = & \left\{ \begin{aligned} & \frac{\exp\left(-\frac{(x-x_0)^2}{2\sigma^2 t}\right) + \exp\left(-\frac{(x+x_0)^2}{2\sigma^2 t}\right)}{\sqrt{2\pi\sigma^2 t}} - \\ & - \frac{2\kappa}{\sigma^2} \exp\left(\frac{4\kappa}{\sigma^2}(x_0 - x + 2\kappa t)\right) \operatorname{erfc}\left(\frac{x_0 - x + 4\kappa t}{\sqrt{2\sigma^2 t}}\right), x > 0; \\ & \frac{2\kappa}{\sigma^2} \exp\left(\frac{4\kappa}{\sigma^2}(x_0 - x + 2\kappa t)\right) \operatorname{erfc}\left(\frac{x_0 - x + 4\kappa t}{\sqrt{2\sigma^2 t}}\right), x < 0, \end{aligned} \right. \end{aligned} \quad (5)$$

where $\operatorname{erfc}(z)$ denotes the complementary error function. These expressions are valid for starting positions $x_0 > 0$, and their extension to $x_0 < 0$ is straightforward by the symmetry

Figure 2 Comparison of analytical diffusion formulas (Equation 4) with discrete random walk simulations for $t = 500$ in the past. We simulated a one-dimensional random walk on an array of linear discrete nodes at all integers and a barrier at $x = 0.5$. Every generation, a random step to one side is made (which implies that $\sigma = 1$). If a movement would be across the barrier (red line), it is realized with probability κ (otherwise no step is made). For each of four barrier strengths, we simulated 10^6 replicates starting at $x = 20$. The blue line depicts the corresponding Gaussian probability density of free diffusion in absence of a barrier.

$x, x_0 \rightarrow -x, -x_0$. The transition density converges to the Gaussian of free Brownian motion for $\kappa \rightarrow \infty$. For a two-dimensional diffusion process with a linear barrier, the full transition density is obtained by multiplying the one-dimensional density functions from Equation 2 for movement parallel to the barrier with Equation 5 for movement orthogonal to the barrier.

Forien (2017) showed that the one-dimensional transition density can be also obtained as a scaling limit of a one-dimensional discrete random walk with a partially reflective barrier. This result indicates that the continuous model is also a good approximation for stepping-stone barrier models. We compared the continuous formula to discrete random walk simulations, and there is indeed close agreement (Figure 2).

Pairwise coalescent probabilities: We use the diffusion approximation to model the distribution of coalescence times for pairs of individuals. We approximate the chance of coancestry originating in a small time interval dt around time t ago as the product of the probability of coming close and a rate of local coalescence $1/2D_e$ (Ringbauer *et al.* 2017). The local density D_e describes a rate with which nearby lineages coalesce (Wright 1943). In a stepping-stone model, D_e corresponds to the number of diploid individuals per deme (Barton *et al.*

2002). This approximation ignores that lineages do not move apart again once they have coalesced. An equivalent simplification is made by Barton (2008), and it is accurate as long as coalescence is sufficiently rare (Wilkins 2004). Since the dispersal process is symmetrical in time in our model of constant population density, the chance that two lineages at current position p_1 and p_2 are close at time t back equals the probability density that a single lineage moves from p_1 to p_2 in time $2t$. Formally, the probability density $\psi(p_1, p_2, t)$ of coalescence T_c at time t in the past is approximated by

$$\psi(p_1, p_2, t) = \frac{\Pr(T_c \in [t, t + dt])}{dt} \approx G(p_1, p_2, 2t) \frac{1}{2D_e}. \quad (6)$$

Empirical simulations for a stepping-stone model confirm that this model approximates the pairwise coalescence time distribution on recent to intermediate timescales (Supplemental Material, File S1).

Identity by descent: We define identity by descent F of two samples at p_1 and p_2 as the chance that two lineages coalesce before a long-distance migration or, equivalently (but unlikely for SNPs), a mutation event occurs along one of the lineages. This definition is closely related to the widely used

Figure 3 Decay of identity by descent F in presence of a strong barrier to gene flow ($\kappa = 0.01$) and moderate neighborhood size ($\sigma = 1$, $D_e = 5$, and $\mu = 0.003$). We calculate F for an individual at $x = 1$ (blue illustration), an individual at $x = -1$ (red illustration), and an individual at $x = -5$ (yellow illustration) with, for all of them, other individuals to the left of the barrier at $x = 0$ that only differ in their x -coordinate (right). The analytical solution in the absence of a barrier is given by the Wright–Malécot formula (Equation 8), the others are integrated numerically (Equation 7).

fixation index F_{ST} , and both definitions agree in the limiting case of an infinite population (see Rousset 2002 for a review). If one assumes that mutation or long-distance migration occurs at a constant rate μ , it is straightforward to calculate the probability that two lineages coalesce before a long-distance event:

$$F(p_1, p_2) = \int_0^\infty \psi(p_1, p_2, t) \exp(-2\mu t) dt. \quad (7)$$

In the absence of a barrier, the probability of identity by descent depends only on the Euclidean distance between two individuals. In this special case, the integral in Equation 7 has an analytical solution, the classical Wright–Malécot formula (Barton *et al.* 2002, 2013):

$$F(r) = \frac{1}{4\pi D_e \sigma^2} K_0\left(\sqrt{2\mu} \frac{r}{\sigma}\right), \quad (8)$$

where K_0 is the modified Bessel function of the second kind of degree zero. A caveat of this analytical solution is that K_0 diverges logarithmically as $r \rightarrow 0$ (Barton *et al.* 2002). Similarly, the integral in Equation 7 diverges for nearby individuals. As for the Wright–Malécot formula, this is caused by a behavior of the diffusion approximation for short timescales, as the chance of two lineages being close diverges as $1/t$ for $t \rightarrow 0$. A solution to circumvent this problem is to start integration at time $t_0 > 0$. Here, we choose one generation time, a biologically plausible value. We could not find an analytical solution, but Equation 7 can be numerically integrated.

Our calculations show that if a barrier to gene flow is present, identity is decreased across the barrier, and increased for points on the same side of the barrier (Figure 3). Interestingly, the increase of F for a pair of points p_1 and p_2 on the

same side of the barrier equals the decrease of F between points p_1 and p'_2 , where p'_2 is the reflection of p_2 across the barrier. This symmetry originates from a reflection principle of the underlying random-walk model, as lineages that do not cross the barrier behave as if they were reflected. This symmetry already occurs in the barrier point density function (Equation 5). It implies that, for a complete barrier, identity by descent can increase to at most twice of the value in absence of a barrier, as observed in the equivalent case of a range boundary (Wilkins 2004).

As already noted by Barton (2008), a barrier to gene flow mostly affects recent coancestry, whereas deeper coalescence probabilities are not distorted much (File S1). In particular, weighted mean coalescence times for individuals from the same deme remain completely invariant in deme models with conservative migration (*i.e.*, constant population density), independent of the specific migration scheme (Nagylaki 1998; Laporte and Charlesworth 2002). We used models of this type in our simulations below to approximate our continuous spatial theory. However, this invariance of mean coalescence times is not problematic for our inference method. When calculating allele frequency correlations, pairwise coalescence probabilities are weighted by a factor $\exp(-2\mu t)$, which accounts for rare, but nonnegligible long-distance migration events. It decays exponentially with time. Therefore, our method is based on a signal of mostly recent coalescence events. Reassuringly, this is also the timescale on which barriers affect pairwise coalescence probabilities the most, and on which the spatial diffusion approximation is likely most accurate (File S1).

Rescaling: Not all parameters in Equation 7 are independent. Consequently, they cannot be estimated separately, as in

Figure 4 Model used to generate synthetic data sets. Ancestry is traced back on a two-dimensional grid of demes (pink circles) with a linear barrier (red line) in discrete time steps. An ancestral deme is randomly picked according to a dispersal kernel and then an ancestor is chosen at random from within this deme. If lineages fall on the same ancestor, they coalesce into a single lineage. For each unique lineage and every step, there is constant chance that a long-distance migration event occurs. In this case, the lineage and all corresponding individuals pick the same allele randomly drawn from a mean allele frequency.

absence of a barrier (Rousset 1997; Barton *et al.* 2013). Therefore, we replace the four demographic parameters $\vec{\theta} : D_e, \kappa, \mu, \sigma$ in Equation 6 with three compound parameters $\vec{\theta}$: neighborhood size $N_{bh} := 4D_e\sigma^2\pi$ [a classical parameter that goes back to Wright (1943)], a scaled barrier parameter $\gamma := 2\kappa/\sigma^2$ (corresponding to the inverse of Barton's B), and a scaled long-distance migration rate $m := 2\mu/\sigma^2$ (Appendix).

Fitting the model to data

A typical data set consists of diploid genotypes $g_1^i, \dots, g_n^i \in \{0, 0.5, 1\}$ for a marker i and individuals at geographic positions $\vec{p}_1, \dots, \vec{p}_n$. To infer the underlying demographic parameters $\vec{\theta}$ from observed data, we have to develop a way to fit our model to such data.

In principle, it is straightforward to transform the probability of identity by descent as calculated by our model into expected allele-frequency covariances. Denoting the probability of identity by descent between a pair of samples k and l by F_{kl} and the population mean allele frequency of a marker i by $\bar{p}_i$, the expected covariance between their genotypes is given by

$$\text{Cov}(g_k^i, g_l^i) = F_{kl}\bar{p}_i(1 - \bar{p}_i).$$

However, often mean allele frequencies have to be estimated from the data, and estimating the means of allele frequencies for many markers would lead to overfitting. To circumvent this caveat, we developed and tested different methods to fit identity by descent to genotype data without directly estimating all mean allele frequencies. We included one approach that models individual genotypes as binomial draws from latent allele frequencies modeled by a Gaussian random field (File S2).

Fitting pairwise homozygosity: Of all tested fitting methods, a relatively simple approach that fits the fraction of pairwise homozygosity (defined as the fraction of identical genotypes)

has the least bias and sampling variation (File S2). Throughout this work, we use this method for data analysis. In the following we give a brief outline of how we calculate expected homozygosity and how we fit it to data.

The observed average homozygosity h_{kl} for a pair of individuals k and l with genotypes g_k^i and $g_l^i \in \{0, 0.5, 1\}$ at markers $i = 1, 2, \dots, n$ can be straightforwardly calculated from the data:

$$h_{kl} = \frac{1}{n} \sum_{i=1}^n (g_k^i g_l^i + (1 - g_k^i)(1 - g_l^i)). \quad (9)$$

Our model predicts the pairwise chance of identity by descent F , and these probabilities can be used to calculate the expected values of average homozygosities $\mathbb{E}(h_{kl})$. The expected pairwise homozygosity of a pair of samples k and l at a marker i is given by:

$$\mathbb{E}(h_{kl}^i) = F_{kl} + (1 - F_{kl})(\bar{p}_i^2 + (1 - \bar{p}_i)^2).$$

The first term gives the probability of having the same genotype due to identity of descent and the second describes the probability of having the same genotype by chance. To avoid estimating all mean allele frequencies, we average over all n markers to get the expected fraction of pairwise identical genotypes:

$$\mathbb{E}(h_{kl}) = F_{kl} + (1 - F_{kl}) \underbrace{\frac{1}{n} \sum_i (\bar{p}_i^2 + (1 - \bar{p}_i)^2)}_{:=s}. \quad (10)$$

Instead of fitting all unknown allele frequencies, now only one additional compound parameter s has to be fit in addition to the demographic parameters θ . We tried fitting this formula to observed data with a composite likelihood approach. However, we found that minimizing the sum of all squared deviations between the expected and observed pairwise homozygosities

Figure 5 Parameter estimates on simulated data. For each replicate, we simulated a population of 60×40 individuals spaced one node apart, and then applied the inference scheme to fit neighborhood size Nb , barrier strength γ , scaled long-distance migration rate m , as well as the allele frequency variance parameter s . To keep run times manageable, we binned individuals into 2×2 demes. We simulated 10 different barrier strengths ($\gamma = 0, 0.01, 0.02, 0.05, 0.1, 0.15, 0.2, 0.25, 0.5, 1.0$), with 20 replicate runs each. Each dot represents an estimate for one such replicate. Horizontal lines depict the true parameters used in the simulations [$\sigma = 1, D_e = 5, \mu = 0.003, \sigma(\rho) = 0.1$]. We split up the barrier plot into two parts with different axes (blue and red) for better visibility.

$$\bar{\theta} = \min_{\theta, s} \sum_{k < l} \left(\mathbb{E}(h_{kl}(\theta, s)) - h_{kl} \right)^2 \quad (11)$$

gives almost identical results, while also being much faster (File S2).

Fitting pairwise homozygosities can be extended to deme data easily, where nearby individuals have been binned, by plugging deme allele frequencies into Equation 9.

Estimation uncertainties: To learn about estimation uncertainty, we bootstrap over genetic markers. Unlinked markers contain almost independent information because their spatial movements are typically correlated only on very short time-scales. Therefore, resampling loci at random is expected to yield accurate empirical confidence intervals.

Implementation

In brief, our inference scheme runs the following three computational steps for a given set of demographic parameters $\vec{\theta}$:

1. Calculate pairwise F for all pairs of samples (integral Equation 7).
2. Use the pairwise F to calculate the expected pairwise homozygosity for all pairs of samples (Equation 10).

3. Calculate the sum of squared differences between the expected and observed pairwise homozygosities (Equation 11).

Our program then finds the parameters $\bar{\theta}$ that minimize this function by using the Levenberg–Marquardt algorithm, as implemented in the Python package Scipy.

We implemented the described simulation and inference methods mostly in Python. To speed up calculations, we parallelized the calculations for pairwise F so that they can be run simultaneously on different central processing units (CPUs). The evaluation of the integrand in Equation 7 is a computational bottleneck. We implemented this calculation in C, to make use of the superior speed of a compiled language.

The inference scheme has to compute the expected identity by state for every pair of samples. It therefore scales quadratically with the number of individuals, as there are $n(n-1)/2$ such pairs. It has a run time of several hours for an individual/deme number of 1000 when run on a single standard desktop CPU. To produce a sufficient number of replicates and bootstraps, we used a scientific computer cluster at the Institute of Science and Technology Austria. To speed up run time, individuals can be grouped into demes. If clustering is done on

Figure 6 Barrier strength estimates for different demographic parameters. We simulated populations of 60×40 individuals spaced one node apart. We fixed default demographic parameters [$D_e = 5$, $m = 0.006$, $\sigma(\bar{p}) = 0.1$, $\sigma = 1$] and barrier strength ($\gamma = 0.02$) and varied demographic parameters, one by one, to five different values each. For every set of parameters, we simulated 100 replicates and applied our inference scheme to estimate γ . Top: Mean estimates of γ for each parameter set, the error bars depict the estimated SEM. Bottom: Empirical SD of estimates of γ across replicates for each parameter set.

small scales for bins at most a few σ in diameter, it does not significantly affect the estimation scheme (File S2). In our results, we capped γ to a maximum value of 1, as for $\gamma > 1$ the effect of a barrier becomes negligible (Figure 2).

Simulations

We extensively tested our inference scheme on simulated data sets. We used a stepping-stone model with D_e individuals per deme, and we traced ancestry backward in time (Figure 4). Every generation, each individual picks an ancestor at random with probabilities given by a dispersal kernel. Here, we use a discretized Laplace distribution as an axial dispersal kernel. Due to the rapid convergence to the continuous diffusion approximation (Figure 2), the specific choice of dispersal kernel has no significant impact as long as its axial variance σ^2 remains finite. If two lineages happen to pick the same ancestor, they coalesce into a single ancestral lineage. We simulate long-distance migration events to occur at a constant rate. If they occur, the corresponding lineage picks an allele at random from the population mean allele frequency $\bar{p}$. To model the effects of a barrier, we follow Nagylaki (1988) and realize migration events across the barrier only with relative probability κ , the barrier strength parameter. For constant deme sizes, this backward model is equivalent to a forward model in which a large number of gametes disperse with the same dispersal kernel (Nagylaki 1988).

After a preset maximum number of generations, every lineage picks an allele at random according to the mean allele frequency $\bar{p}$. Different, unlinked loci were simulated as in-

dependent runs. We picked mean allele frequencies at random according to a predetermined distribution, usually Gaussian with SD $\sigma(\bar{p})$ around an overall mean of 0.5. We also investigated how robust our inference scheme is to scenarios of secondary contact. We simulated them by assigning each ancestral lineage an allele with probability $\bar{p}_l$ or $\bar{p}_r$, according to its location at time of secondary contact.

We stress that the data are simulated under a process very similar to our model. For real data, there could be other deviations which might further reduce the power of the inference scheme. Therefore, our results should be seen as limits for the inference scheme in the case of ideal data.

A. majus data

To show the practical utility of our inference scheme, we applied it to data from an *A. majus* hybrid zone. This hybrid zone is located in a valley in the Eastern Pyrenees. It shows a geographically narrow transition between two flower-color morphs, in which a range of hybrid flower-color phenotypes occur. This transition is mainly determined by three major flower-color genotypes that regulate the intensity and patterning of flower color (Whibley *et al.* 2006; Bradley *et al.* 2017). We applied our method to a data set of 12,389 plants collected between 2009 and 2014, which were genotyped for 112 SNP markers.

To satisfy the assumptions of our model as good as possible, we filtered markers based on four quality criteria: minor allele frequency, large-scale geographic correlation, linkage disequilibrium, and deviations from local Hardy-Weinberg equilibrium. SNP design and filtering are explained in detail in File

Figure 7 Bootstrap estimates on simulated data: We simulated 60×40 individuals on a grid spaced one distance unit apart. We simulated five replicates of four different barrier strengths ($\gamma = 0, 0.1, 0.5, 1.0$). For each of the 20 data sets, we inferred the parameters and additionally did 20 estimates when bootstrapping over loci. Horizontal lines depict the true parameters used in the simulations [$\sigma = 1$, $D_e = 5$, $\mu = 0.003$, $\sigma(\bar{p}) = 0.1$].

S3. After this data cleaning step, we were left with 60 unlinked polymorphic SNPs that were spaced throughout most of the genome (File S3).

Data availability

The source code for the implementation of our inference scheme is freely accessible at the Github repository <https://github.com/hringbauer/BarrierInferPublic.git>.

The *A. majus* data set is a subset of samples collected from 2009 to 2014 with the long-term goal to build a pedigree. The details of this data set and data filtering are described in File S3.

Results

Inference on simulated data

We investigated the overall capability of this method to estimate barrier strengths and the accuracy of empirical bootstrap uncertainty estimates. Our tests show that the inference scheme can reliably recover barrier strengths as well as demographic parameters (Figure 5 and Figure 7). Estimates of the neighborhood size are robust, but show a slight upward bias. These slight biases are likely due to the fact that a continuous model is used to fit to discrete simulations. Estimates of the long-distance migration rate m are more variable, but are not significantly biased. In all cases, the range of bootstrap estimates mostly overlaps with the true value to the expected degree. This result indicates that bootstrapping gives accurate uncertainty estimates (Figure 7).

The power to infer a barrier depends on several factors, including the particular sampling scheme, the number of sampled loci, various demographic parameters, as well as the strength of the barrier. We simulated several specific scenarios to explore overall patterns. Our results indicate that

estimates of the barrier strength γ get more variable with increasing γ (Figure 5, notice the two different scales). Weak barriers ($\gamma \approx 0.5$) were sometimes inferred as no barrier at all, and vice versa. In contrast, strong barriers ($\gamma < 0.1$) were mostly also estimated as such. When varying demographic parameters, estimation uncertainty of γ grows with increasing long-distance migration rate m and neighborhood size (Figure 6). Changing the variance of the mean allele frequencies has comparatively little effect. In addition, we performed power simulations for varying numbers of sampled loci and sampled individuals. They indicate that at least a few dozen biallelic markers and several hundred samples are needed to robustly infer a strong barrier to gene flow (File S2).

Secondary contact: Barriers to gene flow sometimes coincide with areas of secondary contact, for instance in secondary hybrid zones (Barton and Hewitt 1985). Allele frequencies might have diverged before this contact, and present day allele-frequency differences are not caused by the presence of a barrier. However, these clines resemble the effect of a barrier and might be mistakenly inferred as such. One salient way to deal with this problem is to remove markers that show a large-scale geographic structure of allele frequency. One can base inference on a subset of markers that have a similar mean allele frequency across the whole population range and only display fluctuations on small geographic scales, which equilibrate quickly (Barton *et al.* 2013). We tested this approach on simulated data. When applying the inference scheme to a simulated scenario of secondary contact with divergent allele frequencies, it wrongly infers a barrier in case of no filtering (File S2). However, when using the subset of loci that show no large-scale correlation with geography, the false positive signal decreases. Moreover, filtering out loci with a large-scale structure does not remove the signal in

Figure 8 Testing various putative barrier positions on synthetic data. We simulated a single data set, as in Figure 5, with a barrier strength of $\gamma = 0.05$. We applied our inference scheme to fit 15 putative barrier locations, and bootstrapped 20 times over loci for each putative barrier to visualize the uncertainty of the inferred parameters.

case of a true barrier (File S2). However, if sampling is only done on small spatial scales, such filtering could become problematic as one might remove signal from local fluctuations as well. Therefore, we advise to always check that the sampling area is bigger than the spatial scale of isolation-by-distance patterns before any markers are removed.

Unknown barrier locations: Our inference scheme assumes that the location of the putative barrier is known *a priori*. In practice, one might not always have this information or one perhaps wants to test the hypothesis of barriers in different locations. In this case, one can repeatedly apply the inference scheme and fit different potential barrier positions. When testing this approach on simulated data, the inference scheme only inferred a strong barrier near the true position (Figure 7 and Figure 8). The estimated uncertainties on the habitat edges are inflated. This effect is caused by limited power to infer barriers near sampling edges: one needs a sufficient number of samples on both sides of a barrier to fit the strength of a barrier.

Hybrid-zone analysis

We observed a clear isolation-by-distance pattern across the *Antirrhinum* hybrid zone (Figure 9). On average, mean identity by state for nearby plants is elevated by $\sim 2\%$ above the background level, and falls away with increasing pairwise

distance, most rapidly over the first 2000 m. Our inference scheme fits this pattern well, with an estimated neighborhood size of 188 (95% bootstrap confidence interval: 120 – 240; Figure 10).

Obviously, there are demographic complications that are not captured by our model, such as heterogeneities in plant distributions and density. The population is not distributed uniformly in two dimensions, as plants are often found in patches of suitable habitat (Figure 10). However, our analysis indicates that isolation-by-distance patterns are neither strongly influenced by the cardinal direction or relative plant positions, nor the geographic location within the hybrid zone (File S3). These observations imply that our model assumptions are not grossly violated.

For putative barriers in the center of the hybrid zone, our inference scheme estimates no barrier ($\gamma > 1$). Bootstrap estimates rarely fall below $\gamma = 0.5$, and all of them are above $\gamma = 0.1$. When testing for barriers toward the flank of the hybrid zone, estimates get more variable. Some estimates in both flanks indicate an intermediate barrier to gene flow, but in each case some of the corresponding bootstrap estimates also take the value of no barrier ($\gamma > 1$). This signal likely reflects the lower power to infer barriers in these regions, as there is a higher sampling density in the center of the hybrid zone (58.7% of the samples originate from within 500 m of the flower-color transition). Only in one case, for the

Figure 9 Decay of pairwise homozygosity with geographic distance for hybrid zone data. All pairwise homozygosities for the filtered data set were binned according to pairwise distance. We also plot the best fit ($N_{bh} = 188, m = 4.4 \times 10^{-4}, s = 0.5247$). Equation 10 can be used to translate pairwise homozygosity into pairwise F .

leftmost tested barrier, the bootstrap estimates do not overlap with the value of the null hypothesis ($\gamma = 1.0$). This area also shows the strongest small-scale isolation-by-distance pattern (File S3), and the density of plants is very low in sampled patches (M. Melo, personal communication). Therefore, a potential explanation for this significant deviation from the null hypothesis of constant demography and no barrier to gene flow is an exceptionally low plant density in this area.

Given the overall good fit of isolation by distance with our inference scheme (Figure 9), our results indicate that there is no strong genome-wide barrier to contemporary gene flow that coincides with the flower-color transition. As such a strong barrier would require many barrier loci spaced densely throughout the genome (Barton and Bengtsson 1986), this result comes as no surprise. At the moment, no other traits apart from flower color are known to be divergent across the hybrid zone, despite much work to detect them (M. Melo, personal communication).

Previous results suggest the presence of a barrier to exchanging flower-color alleles (Whibley *et al.* 2006; Bradley *et al.* 2017) and indicate that selection maintains differences in flower color (Ellis 2016). Therefore, we applied our method to a subset of polymorphic markers in the genetic neighborhood of two genes, *Rosea* and *Eluta*, known to affect flower-color variation in the hybrid zone. However, bootstrap estimates varied widely for all tested barrier locations (results not shown), which indicates that there is not sufficient signal in the data. This lack of power is likely due to the low number of suitable SNP markers without steep allele-frequency clines near this region in our data set (< 10). Simulations confirmed that for such a low number of markers there is not enough power to detect even strong barriers (File S2).

Discussion

To our knowledge, our scheme is the first method that infers the strength of the barrier based on an explicit spatial population genetic model. There are several similarities with the inference method BEDASSLE (Bradburd *et al.* 2013), which aims to disentangle the effects of geographic isolation by distance and differences in ecological variables. This scheme is not based on an explicit population genetic model, but rather fits the decay of genetic similarity with distance to a heuristic formula. In light of possibly very complex demographic structures, this approach is not necessarily worse than fitting our spatial model. However, this approach does not take the increased covariances on the same side of a barrier into account, and does not make use of some valuable signal because of that. Moreover, it uses an MCMC approach that is based on a model of Gaussian random fields, which is computationally too expensive to apply to hundreds or even thousands of individuals or demes. In contrast, our method is well suited to data sets of this magnitude.

Another widely used method to infer barriers to gene flow, Geneland, clusters individuals using their explicit geographic coordinates (Guillot *et al.* 2005). Safner *et al.* (2011) identified it as one of the most potent methods to infer barriers to gene flow. Therefore, we tested its performance on some of the data sets which we have generated to test our scheme (File S4). As described previously (Guillot and Santos 2009; Safner *et al.* 2011), Geneland's ability to accurately infer barriers to gene flow decreases when isolation-by-distance patterns are present, as its underlying model assumptions of discrete populations with well-defined allele frequencies are violated. Indeed, it fails to detect a barrier to gene flow in our test data sets, which all exhibit such isolation-by-distance patterns (File S4). In contrast, the method introduced here can give accurate estimates of the barrier strength in these scenarios. It is not confounded by isolation-by-distance patterns; in fact, it relies on the presence of this signal. Our method can therefore be seen as complementary to Geneland.

In contrast to BEDASSLE, Geneland, and most other existing methods, which all heuristically describe the strength of the parameter; the inference scheme introduced here fits an explicit spatial population genetics model. It corresponds to Nagylaki's γ (Nagylaki 1978), whose inverse $1/\gamma$ is equal to Barton's barrier strength B (Barton and Bengtsson 1986). This correspondence makes the inferred barrier strength parameter γ interpretable directly in terms of population genetic theory. For instance, the parameter B/σ^2 , which has a dimension of time, must be large to retard the spread of even a neutral allele (Barton and Bengtsson 1986).

Our method can reliably estimate the presence of strong barrier ($\gamma < 0.1$), but there is little power to distinguish between a weak barrier and no barrier (Figure 5 and Figure 7). The reason for this is not a shortcoming of our inference scheme, but due to the fact that relatively weak barriers do not significantly affect the spread of ancestry (Figure 2).

Figure 10 Inference of various putative barrier positions on data from an *A. majus* population. We used our inference scheme to fit 15 barrier locations and bootstrapped 20 times over loci for each putative barrier (bootstrap estimates are shown in orange and cyan). Bottom: Geographic distribution of 12,389 samples. The color of each dot represents flower-color phenotype.

Therefore, it is infeasible to estimate the strength of weak barriers to gene flow, simply because they do not have a significant effect on allele frequency covariances. This effect was already observed by Barton *et al.* (2013), who found that in two spatial dimensions the effect of a barrier starts to have appreciable effects on the spatial pattern of genetic marker alleles when barrier strength B (which corresponds to $1/\gamma$) is $\approx \sigma$.

The exact power of the inference method depends on a range of factors. As a barrier mostly affects nearby allele-frequency fluctuations, a high sampling density spread evenly on both sides of a putative barrier is ideal (Figure 8). Our results demonstrate that estimates of the barrier strength get more variable for increasing long-distance migration/mutation rate, or increasing neighborhood size (Figure 6). These patterns are not surprising, since the strength of isolation by distance, which constitutes the signal for our inference method, decreases when m or N_{bh} grow. Our method is data hungry as it needs at least a few dozen SNP markers and at least hundreds of sampled individuals (File S2). This wealth of data is required since recent pairwise coalescence events, which constitute the signal for our inference scheme, are rare events in most realistic scenarios. As a rough rule of thumb,

our inference scheme can be applied in cases in which there is sufficient power to detect an overall isolation-by-distance pattern.

Any realistic scenario has its own set of parameters and specific sampling scheme. Our power simulations can only test a tiny fraction of possible combinations of these. They helped to elucidate general underlying patterns, but cannot cover all specific cases. Therefore, we recommend doing customized power simulations. Simulating the specific sampling scheme, the marker number together with likely demographic parameters will help to determine whether the inference scheme has sufficient power to detect putative barriers for the specific scenario of interest.

Outlook

The inference scheme introduced here fits a linear barrier, the most straightforward model for a barrier in two dimensions. We used an analytical formula (Equation 5) to model the spread of ancestry, which in turn allows one to reduce calculations for pairwise F to a single numerical integral (Equation 7). However, barriers might be geographically more complex in practice. There could also be multiple barriers in different locations which would be partly indicated by

our method, but also invalidate its underlying model of a single barrier. Such more-complicated scenarios will most likely not allow for simple formulas, and calculations for the chances of recent coancestry become much more challenging. One could trace the geographic ancestry distribution back with discretized simulations, and use them to first calculate the expected distributions of recent pairwise coalescence (Equation 6) and consequently identity-by-descent patterns (Equation 7). This salient extension to our model poses a numerical challenge, but it seems to be within reach of present day computational power.

Our method fits allele frequencies that can be confounded by deeper ancestral patterns. By filtering loci that show large-scale geographic variation, one can in principle remove some of this ancestral genetic structure, but one might accidentally remove true signal as well by doing so. This problem can be a severely confounding factor when applying the inference method to scenarios where ancestral structure is present, for instance in zones of putative secondary contact.

One promising way to overcome this problem would be to base inference on identity-by-descent blocks, the direct genetic traces of recent coancestry (Browning and Browning 2012). As blocks of ancestral genetic material are split up at a constant rate by recombination, the probability of sharing a block of length l decays exponentially back in time (Ralph and Coop 2013). For example, blocks >5 cM are therefore very unlikely to originate from coancestry older than 100 generations, even under relatively extreme demographic scenarios. Moreover, the length of the blocks contains information about the time of coalescence. Identifying such blocks is a nontrivial task, particularly when only unphased genotype data are available (Browning and Browning 2012), as it requires dense genotype data and linkage information. But in cases where identity-by-descent blocks can be robustly called—as is already possible for humans and some model organisms—an inference scheme based on this signal holds great potential. Our method to model the spread of ancestry can be combined with formulas for block sharing (Ralph and Coop 2013; Ringbauer *et al.* 2017) to calculate the expected number of shared identity-by-descent blocks in presence of a barrier. These results could be used to fit observed block-sharing data.

In summary, our method is only a first step to robustly infer barriers to gene flow from genotype data. The techniques outlined here can be expanded in various directions to better deal with the complexities of real data and to make full use of the opportunities within the era of population genomics. We hope that this will ultimately lead to a better understanding of barriers to gene flow within many natural populations.

Acknowledgments

We thank Himani Sachdeva and Raphaël Forien for several useful discussions and for comments on previous drafts of this article. We thank Maria Melo for many discussions about the *Antirrhinum* system, as well as members of the

Barton and Coen laboratory and many volunteers for organizing and carrying out the field work that gave rise to the data set analyzed here. We thank Jon Wilkins and an anonymous reviewer for their very helpful suggestions on how to improve the manuscript.

Literature Cited

- Aguillon, S. M., J. W. Fitzpatrick, R. Bowman, S. J. Schoech, A. G. Clark *et al.*, 2017 Deconstructing isolation-by-distance: the genomic consequences of limited dispersal. *PLoS Genet.* 13: e1006911.
- Barton, N. H., 1979 Gene flow past a cline. *Heredity* 43: 333–339.
- Barton, N. H., 2008 The effect of a barrier to gene flow on patterns of geographic variation. *Genet. Res.* 90: 139–149.
- Barton, N. H., and G. M. Hewitt, 1985 Analysis of hybrid zones. *Annu. Rev. Ecol. Syst.* 16: 113–148.
- Barton, N. H., and B. O. Bengtsson, 1986 The barrier to genetic exchange between hybridising populations. *Heredity* 57: 357–376.
- Barton, N. H., F. Depaulis, and A. M. Etheridge, 2002 Neutral evolution in spatially continuous populations. *Theor. Popul. Biol.* 61: 31–48.
- Barton, N. H., A. M. Etheridge, J. Kelleher, and A. Véber, 2013 Inference in two dimensions: allele frequencies vs. lengths of shared sequence blocks. *Theor. Popul. Biol.* 87: 105–119.
- Bradburd, G. S., P. L. Ralph, and G. M. Coop, 2013 Disentangling the effects of geographic and ecological isolation on genetic differentiation. *Evolution* 67: 3258–3273.
- Bradley, D., P. Xu, I.-I. Mohorianu, A. Whibley, D. Field *et al.*, 2017 Evolution of flower color pattern through selection on regulatory small RNAs. *Science* 358: 925–928.
- Browning, S. R., and B. L. Browning, 2012 Identity by descent between distant relatives: detection and applications. *Annu. Rev. Genet.* 46: 617–633.
- Cercueil, A., O. François, and S. Manel, 2007 The genetical bandwidth mapping: a spatial and graphical representation of population genetic structure based on the wombling method. *Theor. Popul. Biol.* 71: 332–341.
- Dupanloup, I., S. Schneider, and L. Excoffier, 2002 A simulated annealing approach to define the genetic structure of populations. *Mol. Ecol.* 11: 2571–2581.
- Ellis, T. J., 2016 The role of pollinator-mediated selection in the maintenance of a flower color polymorphism in an *Antirrhinum majus* hybrid zone. Ph.D. Thesis, IST Austria, Klosterneuburg, Austria.
- Falush, D., M. Stephens, and J. K. Pritchard, 2003 Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* 164: 1567–1587.
- Fisher, R. A., 1937 The wave of advance of advantageous genes. *Ann. Eugen.* 7: 355–369.
- Forien, R., 2017 Gene flow across geographical barriers - scaling limits of random walks with obstacles. *arXiv preprint arXiv: 1711.02589*.
- Grebenkov, D. S., D. Van Nguyen, and J.-R. Li, 2014 Exploring diffusion across permeable barriers at high gradients. I. Narrow pulse approximation. *J. Magn. Reson.* 248: 153–163.
- Guillot, G., and F. Santos, 2009 A computer program to simulate multilocus genotype data with spatially autocorrelated allele frequencies. *Mol. Ecol. Resour.* 9: 1112–1120.
- Guillot, G., A. Estoup, F. Mortier, and J. F. Cosson, 2005 A spatial statistical model for landscape genetics. *Genetics* 170: 1261–1280.
- Guillot, G., R. Leblois, A. Coulon, and A. C. Frantz, 2009 Statistical methods in spatial genetics. *Mol. Ecol.* 18: 4734–4756.

- Hardy, O. J., and X. Vekemans, 1999 Isolation by distance in a continuous population: reconciliation between spatial autocorrelation analysis and population genetics models. *Heredity* 83: 145–154.
- Laporte, V., and B. Charlesworth, 2002 Effective population size and population subdivision in demographically structured populations. *Genetics* 162: 501–519.
- Malécot, G., 1948 *Les mathématiques de l'hérédité*. Masson et Cie, Paris.
- Manni, F., E. Guérard, and E. Heyer, 2004 Geographic patterns of (genetic, morphologic, linguistic) variation: how barriers can be detected by using monmonier's algorithm. *Hum. Biol.* 76: 173–190.
- Meirmans, P. G., 2012 The trouble with isolation by distance. *Mol. Ecol.* 21: 2839–2846.
- Nagylaki, T., 1978 A diffusion model for geographically structured populations. *J. Math. Biol.* 6: 375–382.
- Nagylaki, T., 1988 The influence of spatial inhomogeneities on neutral models of geographical variation: I. Formulation. *Theor. Popul. Biol.* 33: 291–310.
- Nagylaki, T., 1998 The expected number of heterozygous sites in a subdivided population. *Genetics* 149: 1599–1604.
- Ralph, P., and G. Coop, 2013 The geography of recent genetic ancestry across Europe. *PLoS Biol.* 11: e1001555.
- Ringbauer, H., G. Coop, and N. H. Barton, 2017 Inferring recent demography from isolation by distance of long shared sequence blocks. *Genetics* 205: 1335–1351.
- Rousset, F., 1997 Genetic differentiation and estimation of gene flow from f-statistics under isolation by distance. *Genetics* 145: 1219–1228.
- Rousset, F., 2002 Inbreeding and relatedness coefficients: what do they measure? *Heredity* 88: 371–380.
- Safner, T., M. P. Miller, B. H. McRae, M.-J. Fortin, and S. Manel, 2011 Comparison of Bayesian clustering and edge detection methods for inferring boundaries in landscape genetics. *Int. J. Mol. Sci.* 12: 865–889.
- Slatkin, M., 1993 Isolation by distance in equilibrium and non-equilibrium populations. *Evolution* 47: 264–279.
- Whibley, A. C., N. B. Langlade, C. Andalo, A. I. Hanna, A. Bangham *et al.*, 2006 Evolutionary paths underlying flower color variation in *antirrhinum*. *Science* 313: 963–966.
- Wilkins, J. F., 2004 A separation-of-timescales approach to the coalescent in a continuous population. *Genetics* 168: 2227–2244.
- Womble, W. H., 1951 Differential systematics. *Science* 114: 315–322.
- Wright, S., 1943 Isolation by distance. *Genetics* 28: 114.

Communicating editor: J. Wakeley

Appendix

Here we give the full formula we fit and describe the rescaling of a set of independent effective parameters. Let x_1, x_2 denote the x -coordinate of the samples and $\Delta x (= x_1 - x_2)$ and Δy their separation along each axis. For the identity by state on different sides of the barrier, plugging into Equation 7 gives for pairwise F :

$$\int_0^\infty \frac{1}{2D_e} \underbrace{\exp\left(\frac{4\kappa}{\sigma^2}(\Delta x + 4kt)\right) \frac{2\kappa}{\sigma^2} \operatorname{erfc}\left(\frac{\Delta x + 8\kappa t}{\sqrt{4\sigma^2 t}}\right)}_{\text{Movement x - axis}} \underbrace{\frac{1}{\sqrt{4\pi\sigma^2 t}} \exp\left(-\frac{(\Delta y)^2}{4\sigma^2 t}\right)}_{\text{Movement y - axis}} \underbrace{\exp(-2\mu t)}_{\text{mutation}} dt.$$

We now rescale time such that $t' = \sigma^2 t$. The integral ($dt' = \sigma^2 dt$) transforms to:

$$\int_0^\infty \frac{\sqrt{\pi}}{\sqrt{t'}} \frac{2\kappa}{4\pi D_e \sigma^4} \exp\left(\frac{4\kappa}{\sigma^2} \left(\Delta x + \frac{4k}{\sigma^2} t'\right)\right) \operatorname{erfc}\left(\frac{\Delta x + 4\frac{2\kappa}{\sigma^2} t'}{2\sqrt{t'}}\right) \exp\left(-\frac{(\Delta y)^2}{4t'}\right) \exp\left(-\frac{2\mu}{\sigma^2} t'\right) dt'.$$

Defining $\text{Nbh} := 4\pi D_e \sigma^2$, $\gamma := 2\kappa/\sigma^2$, and $m = 2\mu/\sigma^2$ gives the full formula used for inference:

$$\int_0^\infty \frac{1}{\text{Nbh}} \sqrt{\frac{\pi}{t'}} \exp(2\gamma(\Delta x + 2\gamma t')) \gamma \operatorname{erfc}\left(\frac{\Delta x + 4\gamma t'}{2\sqrt{t'}}\right) \exp\left(-\frac{(\Delta y)^2}{4t'}\right) \exp(-mt') dt'.$$

The formula for the same sides of the barrier is rescaled analogously. The additional term in the integrand becomes:

$$\frac{1}{\text{Nbh}} \frac{1}{2t} \left\{ \exp\left(-\frac{(x_1 - x_2)^2}{4t'}\right) + \exp\left(\frac{(x_1 + x_2)^2}{4t'}\right) \right\} \exp\left(-\frac{(\Delta y)^2}{4t'}\right) \exp(-mt').$$