Spatial Genetic Structure Under Limited Dispersal:

Theory, Methods and Consequences of Isolation-by-Distance

by

Tara N. Furstenau

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved April 2016 by the
Graduate Supervisory Committee:

Reed A. Cartwright, Chair
Michael S. Rosenberg
Jesse Taylor
Melissa Wilson-Sayres

ARIZONA STATE UNIVERSITY

May 2016

ABSTRACT

Isolation-by-distance is a specific type of spatial genetic structure that arises when parent-offspring dispersal is limited. Many natural populations exhibit localized dispersal, and as a result, individuals that are geographically near each other will tend to have greater genetic similarity than individuals that are further apart. It is important to identify isolation-by-distance because it can impact the statistical analysis of population samples and it can help us better understand evolutionary dynamics. For this dissertation I investigated several aspects of isolation-by-distance. First, I looked at how the shape of the dispersal distribution affects the observed pattern of isolation-by-distance. If, as theory predicts, the shape of the distribution has little effect, then it would be more practical to model isolation-by-distance using a simple dispersal distribution rather than replicating the complexities of more realistic distributions. Therefore, I developed an efficient algorithm to simulate dispersal based on a simple triangular distribution, and using a simulation, I confirmed that the pattern of isolation-by-distance was similar to other more realistic distributions. Second, I developed a Bayesian method to quantify isolation-by-distance using genetic data by estimating Wright's neighborhood size parameter. I analyzed the performance of this method using simulated data and a microsatellite data set from two populations of Maritime pine, and I found that the neighborhood size estimates had good coverage and low error. Finally, one of the major consequences of isolation-by-distance is an increase in inbreeding. Plants are often particularly susceptible to inbreeding, and as a result, they have evolved many inbreeding avoidance mechanisms. Using a simulation, I determined which mechanisms are more successful at preventing inbreeding associated with isolation-by-distance.

DEDICATION

To my parents, Devon and Rhonda, my husband, Chris, and the rest of my family who have been a constant source of support that kept me going. Also to the memory of my grandpa Wally, who always encouraged me to work hard at “the brain factory”.

ACKNOWLEDGMENTS

I would like to thank Reed Cartwright for being a supportive and encouraging dissertation mentor and for guiding my development as a computational biologist. I am extremely grateful to have been a member of the Cartwright lab because it has been an invaluable learning experience, and I was afforded many exciting opportunities to share my research. I also want to thank my dissertation committee members — Michael Rosenberg, Jay Taylor, and Melissa Wilson-Sayres — for their valuable advice and input which greatly improved my dissertation and for their instructive discussions at journal club meetings. Finally, I am thankful for the feedback and support I received from the members of the Cartwright lab: Rachel Schwartz, David Winter, Steven Wu, Christian Sievert, Kael Dai, and Adam Orr.

TABLE OF CONTENTS

APPENDIX

LIST OF TABLES

LIST OF FIGURES

Figure Page

Figure Page

Chapter 1

INTRODUCTION

Spatial genetic structure (SGS) is observed when there is a non-uniform distribution of allele types across space. Typically, these alleles are positively autocorrelated such that individuals that are close together are more likely to share a similar allele than individuals that are further apart. This autocorrelation manifests in large patches of genetic similarity in the population.

Spatial genetic structure can be a consequence of many physical, geographical, ecological, or behavioral factors but the most common and inherent cause of SGS is isolation-by-distance (Epperson, 2003). Isolation-by-distance describes the SGS caused by limited parent-offspring dispersal and was first introduced by Sewall Wright in 1943. In nature, the offspring of many species tend to disperse only a short distance from their birth site before mating (Caine et al., 2000; Howe and Smallwood, 1982; Kot et al., 1996), and consequently, the mating pool will contain an excess of related partners. After a number of generations of local mating and local dispersal, a stationary pattern of isolation-by-distance will become established. The degree of isolation is a function of the genotype-independent dispersal ability of the individuals that make up the population. Patterns of isolation-by-distance can be observed at different spatial scales such as within a continuous population or between different subpopulations (Epperson, 2003; Rousset, 2004); the former will be the main focus of this dissertation.

Tests for isolation-by-distance are routine in molecular ecology studies and a positive result should be used to inform downstream data analysis (Meirmans, 2012). Statistical tests often rely on the assumption that individuals have been sampled randomly and independently from a panmictic population. The presence of isolation-by-distance violates this assumption causing a reduction in the effective sample size and biased estimates of population parameters when appropriate steps are not taken to handle the correlation. When population analyses do not account for isolation-by-distance, such patterns can easily be conflated with other environmental or evolutionary processes (Meirmans, 2012). When used properly, however, the pattern of isolation-by-distance for neutral genetic loci can serve as a null hypothesis from which deviations can be more accurately detected and interpreted.

In addition to the practical statistical reasons, the study of isolation-by-distance is important because it influences many population and evolutionary processes. Novembre et al. (2008) published a study which showed that the pattern of SGS in sampled Europeans corresponded closely with the geography of Europe and follows a pattern of isolation-by-distance. The study of isolation-by-distance in human populations allowed Novembre et al. (2008) to correctly identify the origin of 90% of European individuals. Such information can be applied to tests of genetic ancestry and can serve as a null geographical distribution of neutral variation from which important deviations (e.g. selection or region specific disease alleles) can be detected. Ecologists and conservation biologists are interested in isolation-by-distance because it can influence local adaptation, population differentiation and responses to habitat fragmentation (Zhao et al., 2013; Leonardi et al., 2012; Andrew et al., 2012). Isolation-by-distance also facilitates inbreeding which will be covered in more detail in Chapter 4.

Isolation-by-distance is often analyzed in two ways. It can be described qualitatively using a correlogram which shows the genetic relationship between individuals over geographical distance. If isolation-by-distance is present, pairs of individuals that are close together in space will show significantly higher genetic similarity than individuals that are further apart. Isolation-by-distance can also be quantified using Wright's (1943; 1946) neighborhood size, N_b. Neighborhood size describes the effective number of individuals in the local vicinity from which parents can be drawn at random (Rousset, 1997; Wright, 1943). Neighborhood size generally increases as dispersal distances increase.

For this dissertation I will explore several different aspects of isolation-by-distance. In the first chapter I will look at how the shape of the dispersal kernel affects the pattern of isolation-by-distance using a spatially-explicit lattice-based simulation. My goal was to verify well established analytical results that suggest that the pattern of isolation-by-distance depends mostly on the neighborhood size. To do this, I simulated dispersal using distributions that were parameterized to have to same neighborhood size and analyzed the resulting pattern of isolation-by-distance. I also present an algorithm for efficiently simulating dispersal using the triangular distribution which produces a uniform distribution over the neighborhood area. I argue that triangular dispersal is more in line with the idea of neighborhood size representing a local panmictic unit. These results were published in Furstenau and Cartwright (2016).

In the second chapter I present a method I developed for the estimation of neighborhood size. This is the first method developed to estimate neighborhood size using a Bayesian approach. In this chapter I analyze the performance of this method on data generated from the model, data from a lattice based simulation and data from two populations of *Pinus pinaster* Aiton. I look at how well the method performs when different assumptions are violated and I determine which sampling schemes provide better estimates.

In the third chapter I examine bi-parental inbreeding which is one of the consequences of isolation-by-distance. Many plant species have evolved self-incompatibility (SI) systems to avoid self-fertilization and genetic forms of SI have been particularly successful across the angiosperms. In addition to preventing self-fertilization, it is often assumed that the success of genetic SI species is a result of their ability to reduce bi-parental inbreeding. To test this assumption, I developed a spatially-explicit individual based simulation to model populations of self-incompatible plants under isolation-by-distance and analyze the amount of inbreeding in the populations compared to a non-genetic form of SI.

Chapter 2

THE EFFECT OF THE DISPERSAL KERNEL ON ISOLATION-BY-DISTANCE IN CONTINUOUS POPULATIONS

Abstract

Under models of isolation-by-distance, population structure is determined by the probability of identity-by-descent between pairs of genes according to the geographic distance between them. Well established analytical results indicate that the relationship between geographical and genetic distance depends mostly on the neighborhood size of the population which represents a standardized measure of gene flow. To test this prediction, I model local dispersal of haploid individuals on a two-dimensional landscape using seven dispersal kernels: Rayleigh, exponential, half-normal, triangular, gamma, Lomax and Pareto. When neighborhood size is held constant, the distributions produce similar patterns of isolation-by-distance, confirming predictions. Considering this, I propose that the triangular distribution is the appropriate null distribution for isolation-by-distance studies. Under the triangular distribution, dispersal is uniform over the neighborhood area which suggests that the common description of neighborhood size as a measure of a local panmictic population is valid for popular families of dispersal distributions. I further show how to draw random variables from the triangular distribution efficiently and argue that it should be utilized in other studies in which computational efficiency is important.

Introduction

For many populations, individuals do not exist in discrete patches or demes; instead they are spread across a continuous landscape. Although there are no barriers separating individuals, dispersal distances are often limited, and individuals that are near one another tend to be more similar genetically than individuals further apart. This phenomenon is known as isolation-by-distance and introduces a spatial component that should be considered when studying population genetic processes. Unfortunately, incorporating multiple dimensions of space at fine scales into analytical models is often analytically intractable (Epperson et al., 2010). Therefore, many researchers have turned to spatially-explicit, individual-based computer simulations which offer a more flexible way to incorporate spatial complexity into biological models (e.g. Barton et al., 2013; Cartwright, 2009; Epperson, 2003; Novembre et al., 2008; Rousset, 2004; Slatkin, 1993).

A dispersal kernel describes the distribution of Euclidean distances between birth site and reproduction site. Ideally, when modeling dispersal, the dispersal distribution would be selected based on how well it fits the dispersal kernel estimated from natural populations. Classically, dispersal has been modeled as a diffusion process with Gaussian displacement; however, the observed dispersal kernels in many species tend to be more leptokurtic with a higher probability of short and long distance dispersal (Bateman, 1950). In plants, the shape of the dispersal kernel near the origin depends on the mechanism of dispersal; for example, there may be a high peak near the origin for gravity or animal dispersal whereas there may be a minimum near the origin for wind dispersal (Barluenga et al., 2011; Clark et al., 2005).

The shape of the dispersal kernel impacts many population processes including the rate of population expansion (Clark et al., 2001; Kot et al., 1996), responses to environmental changes (Nathan et al., 2011), local adaptation (Berdahl et al., 2015), speciation (Hoelzer et al., 2008), and the spatial distribution of genetic diversity (Bialozyt et al., 2006; Ibrahim et al., 1996). Fat-tailed dispersal kernels, with a higher probability of long-distance dispersal, are a good fit to many empirical data sets (Bullock and Clarke, 2000; Clark et al., 2005; Gonzàlez-Martìnez et al., 2006; Martìnez and Gonzàlez-Taboada, 2009; Klein et al., 2006). Many studies have shown that population models behave differently when fat-tailed dispersal distributions are used instead of Gaussian dispersal. Kot et al. (1996) demonstrated that population spread is sensitive to the shape of the dispersal kernel and models using a normal distribution underestimated the rate of invasion compared to fat-tailed distributions. Nathan et al. (2011) found that long distance dispersal plays a large role during range shifts of wind-dispersed trees in response to projected climate changes. Houtan et al. (2007) showed that heavy tailed dispersal kernels were a better fit for dispersal of Amazonian birds but the shape of the dispersal kernel can change in response to forest fragmentation.

While the shape of the dispersal kernel impacts many population processes at different scales, it remains unclear how it affects patterns of isolation-by-distance within a continuous population. It has been argued that the number of long-distance dispersal events will not have a noticeable effect because new long-distance alleles are more likely to be lost due to drift than become established at the new location (Epperson, 2007; Ibrahim et al., 1996). On the other hand, the shape of the dispersal kernel near the origin may have a significant impact on the overall rate of migration. In plants, this could result in a higher probability of self-fertilization and/or a reduction in the number of successful offspring when there is density dependent regulation (Barluenga et al., 2011; Howe et al., 1985; Moyle, 2006).

Isolation-by-distance theory predicts that the probability of identity-by-descent between two neutral genes will decrease as the geographic distance between them increases and this pattern can help quantify spatial genetic structure. The analytical model developed by Malécot (1969) depends on the effective population density, the mutation rate, the spatial dimensions of the population, and the dispersal distribution. Much of the isolation-by-distance work has focused on the lattice model which forces a constant population density (Malécot, 1969; Maruyama, 1970; Sawyer, 1977) but these results hold when considering continuously distributed populations with spatial clustering (Barton et al., 2013).

In two dimensions, the relationship between the probability of identity-by-descent and the log of distance is linear over a certain range of distances and the relationship is proportional to $1/(D_e\sigma^2)$ where D_e is the effective population density and $2\sigma^2$ is the mean squared distance of dispersal (i.e. non-central second moment of Euclidean distance; Barton et al., 2013; Malécot, 1969; Rousset, 1997, 2004; Wright, 1946). Over this range, the slope of the probability of identity-by-descent function is independent of most aspects of the dispersal distribution except for $2\sigma^2$; however, when the distance between individuals falls below the range, the shape of the dispersal distribution becomes important (Rousset, 1997). This suggests that as long as $2\sigma^2$ stays constant, any dispersal distribution will produce similar patterns of isolation-by-distance. However, Rousset (1997, 2008a) argues that the magnitude of genetic differentiation will always depend on the shape of the distribution. Rousset (1997) numerically evaluated the correlation of the probability of identity-by-descent between pairs of genes a certain distance apart relative to the probability of identity-by-descent between two genes within an individual using different discrete dispersal models. While the slope of the relationship over the log of distance for two-dimensional space depends only on $2\sigma^2$, the y-intercept is determined by more complicated features of the dispersal distribution. The numerical analysis was considered for a lattice model of discrete demes, however the theory extends to lattice models of continuous populations with one individual per lattice node (Rousset, 2000).

Despite the increase in the use of spatially explicit simulations in studies of spatial genetic structure, it remains unclear whether the shape of the dispersal kernel should be considered. There has not been a clear comparison of how the shape of different dispersal kernels affect observable patterns of isolation-by-distance in these simulations. Here I attempt to offer such a comparison using a spatially-explicit, individual-based model to simulate local dispersal in a continuous population to determine if patterns of isolation-by-distance vary based on the shape of several different dispersal distributions: Rayleigh, half-normal, exponential, triangular, gamma, Lomax, and Pareto. Each dispersal distribution has a different shape, but they can be parameterized such that their non-central second moment is $2\sigma^2$. If the simulations reveal a similar pattern of isolation-by-distance across all dispersal distributions, I can conclude that, for a wide range of dispersal distributions, $2\sigma^2$ is the main determining factor of how genetic similarity declines with increasing distance in a continuous population. Consequently, when designing isolation-by-distance simulations, researchers may choose a dispersal distribution based on computational needs instead of biological fit.

Wright (1946) uses the term “neighborhood” to describe a local population from which parents are randomly drawn. He measures the magnitude of the effective size of the neighborhood, N_b, as the inverse of the probability that two gametes at the same location came from the same parent. Assuming dispersal is normally distributed along each axis, he calculated that $N_b = 4\pi\sigma^2 D_e$, where D_e is the effective density of individuals, and $2\sigma^2$ is the mean squared distance of dispersal. — In his model this captures 86.5% of parents of central individuals. — Although Wright assumed Gaussian dispersal, his formula can be used to calculate N_b for many different dispersal models at equilibrium due to the central-limit theorem. N_b is important because it helps define the rate of decay of genetic similarity over spatial distance, i.e. the amount of isolation-by-distance in a population (Barton et al., 2013; Rousset, 1997, 2000).

If a neighborhood is supposed to represent a local panmictic unit, then in the ideal model parents should be chosen uniformly from within a circle of radius 2σ centered on an offspring, and the Euclidean distance between parents and offspring should follow a triangular distribution: $f(r;\sigma) = r/(2\sigma^2)$, where $2\sigma^2$ is again the non-central second moment. This type of neighborhood is similar to the neighborhood defined in the spatially continuous Fleming-Voit disc model in which a number of parents, v, are chosen uniformly at random from a disc with radius r to replace a fraction u of the population (Barton et al., 2013). In this model, neighborhood size is defined by the ratio v/u and the individuals occupying the disc constitute a panmictic population. If 100% of the population is replaced ($u = 1$), the definition of neighborhood size reduces to the number of individuals competing for the central location.

Below, I demonstrate that patterns of isolation-by-distance in continuous populations at equilibrium are similar for different dispersal kernels with the same second moment, and discuss the use of the triangular distribution to model dispersal in a continuous population.

Methods

Simulation

In my individual-based simulation, a population exists on a 100×100 rectangular lattice (cf. Epperson, 1995; Epperson and Li, 1997; Epperson, 2007; Hardy and Vekemans, 1999). Individuals are uniformly spaced with a single individual per cell. Each individual contains one haploid locus. The initial population of 10,000 individuals each carry a unique allele. Generations are discrete, and individuals reproduce by generating a fixed number of clonal offspring that experience mutations according to the infinite alleles model at rate μ. All starting and mutant alleles are selectively neutral.

The offspring disperse from the parent cell following a given dispersal distribution and when offspring disperse off of the lattice they are lost. Offspring that land in the same cell will compete to become a parent in the next generation. Because all alleles are selectively neutral, a single successful offspring is uniformly selected for each cell. To avoid storing all the offspring in memory until dispersal is completed, I use a reservoir sampling method to immediately accept or reject offspring when they land on a cell (Vitter, 1985). This method allows us to keep track of two randomly chosen offspring per cell. The first offspring becomes a parent in the next generation and the second individual is recorded to measure the probability of identity-by-descent for offspring competing for the same cell. In the simulation, it is possible for some cells to remain empty after dispersal; however, I determined that when each parent generates 15 offspring, the number of empty cells per generation is negligible. Therefore, the results presented here are from simulations where individuals produce 15 offspring, and I assume a constant homogeneous population density. Simulations with larger numbers of offspring produce similar results, but I did not test any simulations where the number of offspring varied for different individuals.

Modeling Dispersal

The simulation is spatially-explicit with space represented on a rectangular lattice. Due to the discrete nature of the lattice, the dispersal kernels will be discretized approximations of continuous distributions (Chesson and Lee, 2005; Chipperfield et al., 2011). The dispersal kernel function, $f(r, \theta; \sigma)$, takes a parameter σ and returns continuous polar coordinates. The σ parameter is the square root of one-half the second moment of dispersal distance. The polar coordinates include the angle, $\theta \in [0, 2\pi]$, which is uniformly distributed to ensure isotropic dispersal, and distance, $r > 0$, which is drawn from a continuous distribution.

Once the angle and distance are drawn, the final position is determined by converting the polar coordinates into rectangular coordinates and adding them to the parent's position. The new coordinates are then rounded to determine the integer coordinates of the destination cell. This dispersal scheme is similar to the centroid-to-area approximation of continuous dispersal kernels described by Chipperfield et al. (2011), which showed minimal deviation from expectations especially when cell length is less than the expected value of the dispersal distance distribution.

I looked at seven different dispersal distance kernels (Table 1): Rayleigh, exponential, half-normal, triangular, gamma, Pareto, and Lomax. I chose these distributions because they provide a range of shapes for short, intermediate, and long distance dispersal.

The Rayleigh is a distribution of Euclidean distances that result from bivariate normal displacement along the x and y axis. The Rayleigh distribution follows the assumptions of Wright (1946)'s two-dimensional isolation-by-distance model.

The exponential distribution is more leptokurtic with a higher probability of dispersal at short and long distances, and a lower probability at intermediate distances. The exponential tail is the boundary that separates truly heavy-tailed distributions with potentially infinite higher moments from distributions with all moments finite. The distinction is important because leptokurtic, heavy-tailed dispersal kernels are typically a better fit to observed dispersal in nature (Clark, 1998).

The half-normal distribution is equivalent to a normal distribution that has been folded over the y-axis. In this case, Euclidean distance is simply the absolute value of normally distributed random variables. The half-normal is a monotonically decreasing distribution with a convex shoulder near zero. This distribution has a higher probability of dispersal at intermediate distances compared to the exponential.

The triangular distribution is typically defined using three points: a lower limit, a, an upper limit, b, and a mode, c. Here I use a special case of the triangular distribution where $a = 0$ and $b = c = 2\sigma$. I chose this special case of the triangular distribution because in my dispersal function it will return polar coordinates that are uniformly sampled from within a circle with area $4\pi\sigma^2$, which is the same as the neighborhood area (See proof in Appendix A). The triangular distribution is also the only one of my distributions that has a finite range, $r \in [0, 2\sigma]$.

Unlike the previous single parameter distributions, the final three distributions have an additional α shape parameter. The gamma distribution is equivalent to the exponential distribution when $\alpha = 1$, and as α increases the distribution becomes more symmetrical with a higher probability for intermediate distances and a lower probability for short distances.

The Lomax and Pareto distributions are both heavy-tailed power-law distributions. The n-th moments are finite only when $\alpha > n$. The support for the Pareto distribution, $r \in [x_{min}, +\infty)$, begins at a parameter $x_{min} > 0$. The Lomax distribution is a Pareto distribution that has been shifted so that the support begins at zero. I chose values of α between 2 and 3 so that the second moment of the distribution would be finite but higher moments are infinite.

The dispersal function is executed over 100-billion times per simulation, and thus it is important to make the implementation as efficient as possible. With this aim, I used an xorshift algorithm for uniform pseudo-random number generation and the ziggurat rejection sampling algorithm when applicable (Marsaglia and Tsang, 2000b; Marsaglia, 2003). I used two different versions of the ziggurat algorithm to draw distances from the exponential and half-normal distribution. For the gamma distribution I used a rejection sampling method that uses the ziggurat algorithm to draw normal variates (Marsaglia and Tsang, 2000a). Random variables from the Pareto distribution are generated by $x_{min}e^{U}$ where U is an exponentially distributed random variable that is drawn using the ziggurat algorithm. The Lomax distribution is sampled the same way as the Pareto distribution but it is shifted by $-x_{min}$.

In addition to generating random distances, the dispersal function requires costly conversions from polar to Cartesian coordinates. I was able to avoid this conversion for the Rayleigh and triangular distributions. I simulated the Rayleigh distribution by drawing vertical offsets from independent normal distributions using the ziggurat algorithm. For the triangular distribution, I developed a discrete sampling algorithm using the Alias method that allows the vertical and horizontal offsets to be drawn simultaneously in constant time (Vose, 1991). See Appendix A for a description of the algorithm.

To compare the run time for the different dispersal functions, I simulated one dispersal event from each cell on a 100×100 landscape 100,000 times for a total of 10^9 dispersal events. For each simulation $\sigma = 1$, and $\alpha = 3$ for the two parameter distributions. The CPU time was averaged over 5 different runs.

Analysis

A simulation was run for each of the seven dispersal distributions under 4 levels of dispersal (σ = 1, 1.5, 2, and 4) with a mutation rate of $\mu = 10^{-4}$. Each simulation was run for a burn-in period of 10,000 generations to allow the population to reach a mutation-drift equilibrium. After the burn-in, data was collected from populations that were 1,000 generations apart to decrease the correlation between them for a total of 2,000 replicate populations per simulation. In each population, a straight transect of 50 individuals was sampled from the center of the landscape to avoid measuring edge effects.

From the transect, all possible pairs of individuals were placed into distance classes based on the geographical distance between the pair. The number of pairs that shared an identical allele was determined and recorded as a proportion of the total number of pairs in the distance class. The probabilities for each distance class were then averaged over all sampled populations. Under this sampling scheme, the number of pairs per distance class decreases as distance increases so in distance class 50 there is only one pair sampled per population.

The parameters for each dispersal distribution were calculated so that $E[X^2] = 2\sigma^2$; the calculations are reflected in the probability distribution functions in Table 1. Due to the discrete nature of the lattice, some parameters values were adjusted slightly until the simulations produced an average, observed, squared distance between parent and offspring, s^2, that was within 5% of the expected value, σ^2. Three of the distributions require a second α parameter. For the gamma distribution I used $\alpha = 1, 2, 4,$ and 8. For the Lomax and Pareto distributions I used $\alpha = 2.4, 2.6, 2.8,$ and 3.0, all of which result in distributions that are infinite in the 3rd and higher moments.

Under isolation-by-distance, individuals geographically near one another will tend to be genetically similar, and this similarity will decrease as the distance between pairs of individuals increases. Therefore, isolation-by-distance is described by constructing correlograms of genetic similarity between individuals versus the geographical distance between them. Genetic similarity can be measured using identity-by-descent, identity-by-state, relatedness, conditional kinship, or F-coefficients and can be based on coalescent times, an ancestral population, or the current population (Hardy and Vekemans, 1999; Hardy, 2003; Malécot, 1969; Rousset, 1997, 2002; Wang, 2014). For two-dimensional populations, genetic similarity is often plotted against the log-distance separating pairs because theory predicts that this relationship is approximately linear over a certain range (Barton et al., 2013; Hardy and Vekemans, 1999; Rousset, 2000).

I recorded the probability of identity-by-descent for pairs of individuals in each distance class. Under the infinite alleles model, pairs of individuals were considered identical-by-descent if they shared the same allele. The probability of identity-by-descent in each distance class depends on the mutation rate; the probability will be greater when there are fewer alleles. For more consistent results that are nearly independent of mutation rate, the probability of identity is often calculated as a ratio that measures genetic similarity (or differentiation) relative to a particular reference group. I calculated the kinship coefficient which measures the correlation of genetic similarity between pairs of individuals a certain distance apart relative to the genetic similarity in the whole sample.

$$F_r = \frac{p_{ij} - \bar{p}}{1 - \bar{p}} \approx \frac{E[T] - E_{ij}[T]}{E[T]} \tag{2.1}$$

Here p_{ij} is the probability of identity-by-descent between haploid individuals i and j at distance r and $\bar{p}$ is the probability of identity-by-descent between random haploid individuals in the current sample (Hardy and Vekemans, 1999). The kinship coefficient is related to differences in the expected coalescent times, T, between a specific pair of individuals and a random pair in the population (Barton et al., 2013). Kinship coefficients were calculated for each transect and then averaged across transects for each distance class. Since this statistic is highly dependent on the sampling scheme, I sampled the same transect in all simulations.

I also calculated the a_r parameter of Rousset (2000):

$$a_r = \frac{p_0 - p_{ij}}{1 - p_0} \tag{2.2}$$

which measures genetic differentiation over distance relative to the probability of identity-by-descent within a location. The a_r parameter is independent of sampling scheme, but it does depend on the level of local identity-by-descent, p_0, in the population such that a_r approaches infinity as p_0 approaches one (Vekemans and Hardy, 2004). Typically, p_0 is estimated from the amount of autozygosity in the population; however, I estimated p_0 as the probability that an individual shared an allele with one of the offspring that it competed with for the cell, which is suitable for haploid organisms and better fits its definition (Vekemans and Hardy, 2004).

For each simulation, I calculated the average number of unique alleles in a 50-individual transect ($\bar{k}$) and the average squared distance between parents and offspring ($2s^2$). Using $\bar{k}$, I estimated the population-level diversity, $\hat{\theta}_k$ (Ewens, 2004, eq. 9.32) and estimated effective haploid population size as $\hat{N}_e = \hat{\theta}_k/2\mu$ and effective density as $\hat{D}_e = \hat{N}_e/A$, where $A = 10,000$.

Finally, I estimated neighborhood size using two different methods. First I used the estimated demographic parameters to calculate neighborhood size as the product $\hat{N}_b = 4\pi s^2 \hat{D}_e$. I calculated an estimate from samples from each population and calculated an average over all populations. I then estimated neighborhood size using the regressions of both F_r and a_r on the log of distance. The slope of the a_r regression is an estimate of $1/2\pi\sigma^2 D_e$ and the slope of F_r regression is an estimate of $-(1 - F_0)/2\pi\sigma^2 D_e$ (Barton et al., 2013; Hardy and Vekemans, 1999; Rousset, 2000). I performed the regression for distance classes between 5 and 35. I estimated the slope from each population sample and then pooled the data from all the samples to get a combined slope estimate.

Results

Behavior of Dispersal Distributions

Figure 1 shows the empirical cumulative distributions generated from 10,000 simulated dispersal events from each distribution. The probability of not dispersing from the original cell is indicated by the height of the left-most horizontal line for each distribution. The more leptokurtic distributions (exponential, gamma-1 and Lomax) with a high probability peak near zero have a much higher probability of not dispersing from the original cell, especially when σ is low. The Pareto distribution, which has a fat tail but has been shifted so it does not have a peak at zero, has a very low probability of not dispersing. Under the gamma distribution as the α parameter increases, the probability of remaining at the origin decreases; when $\alpha = 8$ the probability is nearly zero for all values of σ.

The average squared parent-offspring dispersal distance, s^2, observed for each distribution was very similar with a relative error of less than 5% from the expected σ^2 value (Table 2); however, the distribution of these values over sampled generations varied (Fig. 2A). Expectedly, the thin tailed or no-tail (triangular) dispersal distributions have the smallest variance because their properties are easier to represent with a small number of samples. The Lomax distribution has the highest variance with the median falling slightly below the expected value.

Figure 2B shows the distribution of the average cubed parent-offspring dispersal distances, s^3, for each transect. The theoretical third moment of the Lomax and Pareto distributions is infinite and while it is not possible to simulate this on a finite landscape, I do observe values of s^3 that are several orders of magnitude larger than distributions with finite third moments. The distribution of s^2 and s^3 for the Lomax and Pareto distributions both have a large positive skew.

Allelic Diversity

The distribution of the number of unique alleles is similar for most of the dispersal kernels with the median falling near the expected value under the infinite alleles model (Fig. 3). The expected number of alleles under the infinite alleles model (gray horizontal line) is equal to $\sum_{i=0}^{n-1} \theta/(\theta + i) = 7.03$ where $n = 50$ is the number of individuals in the sampled transect. The Lomax distributions have a higher median number of alleles at lower values of σ but this gets closer to the expected value when $\sigma > 2$. The average diversity is also slightly elevated for the exponential and gamma-1 simulations.

Differences in effective population size between simulations can be measured by comparing the number of unique alleles observed in the transects. Different dispersal kernels produce similar levels of diversity, except for the Lomax distributions which have a higher θ_k and consequently a larger effective population size (Table 2).

Spatial Autocorrelation and Isolation-by-Distance

To describe the patterns of isolation-by-distance, I first measured the average probability of identity-by-descent for each sampled population as a function of distance. When σ is small, the probability of identity-by-descent between pairs of individuals is greater at small distance and decreases as the distance between individuals increases; the strength of this relationship decreases as the dispersal parameter increases and it flattens out when $\sigma = 4$ (Fig. 4). The thin tailed dispersal kernels produced very similar patterns of isolation-by-distance. This is also true for the Pareto dispersal kernels which are fat tailed but are shifted so that the probability of not dispersing from the original cell is lower compared to the Lomax dispersal kernels. For simulations run with the Lomax dispersal kernels, the probability of identity-by-descent has a steeper decrease at short distances. Additionally, the overall pattern seen under the Lomax kernel is shifted lower than the other distributions and this is most pronounced when dispersal is limited. This overall decrease in the probability of identity-by-descent is associated with the increase in the number of unique alleles, $\bar{k}$. Differences between the different dispersal distributions are more apparent when the distance between individuals is small. The more leptokurtic dispersal distributions have a steeper incline as distance decreases and they have a higher probability of autozygosity at distance class zero. The plots for the triangular distribution nearly perfectly overlap the plots for the Rayleigh distribution in all cases.

Because the probability of identity-by-descent is sensitive to differences in the number of alleles present in the sample, I also calculated the pairwise-kinship coefficient over the log of distance (Fig. 5). When the value of the kinship coefficient is greater than zero (horizontal gray dashed line), pairs of individuals at that distance are more genetically similar than random pairs of individuals in the sample as a whole, and when the kinship coefficient is less than zero, pairs of individuals are less genetically similar than random pairs of individuals. The kinship coefficient is nearly independent of differences in allele number and there is much better overlap of the plots for the different dispersal distributions. When the kinship coefficient is plotted against the log of distance there is a negative linear relationship over a certain range of distances (Hardy and Vekemans, 1999). For short distances, the relationship does depend on the shape of the dispersal distribution; however, over the linear range, the different distributions more closely overlap, especially when $\sigma > 1$.

Finally, I plotted Rousset (2000)'s a_r parameter, a measure of genetic differentiation, against the log of distance. Again, the relationship at small distances depends on the dispersal distribution but at larger distances, there is a positive linear relationship between a_r and the log of distance (Fig. 6). The overall magnitude of the relationship is different for the different distributions and the more leptokurtic distributions typically show greater differentiation. The overall increase in genetic differentiation with the log of distance (the slope) is fairly similar among the dispersal kernels for different values of σ.

Estimated Neighborhood Size

The $\hat{N}_{b(\theta_k)}$ estimates are shown in Table 2 and Figure 7A. Table 2 shows the average estimate over all population samples. The colored dots in Figure 7A show this same average relative to the expected values and the bars represent the middle 50% of the individual sample estimates. As mentioned previously, the populations with Lomax dispersal tend to have a greater number of unique alleles and this translates to higher $\hat{\theta}_k$, higher effective population size, and ultimately higher effective density. The estimates for s^2 were highly variable but skewed towards lower values. As a result, the estimates of $\hat{N}_{b(\theta_k)}$ for the Lomax distribution appear to be higher on average but the estimates are skewed. Otherwise, the estimates for the other dispersal distributions are similar and close to the expected values.

Table 2 shows the $\hat{N}_{b(a_r)}$ estimates calculated as the twice the inverse of the regression of a_r and the log of distance for the pooled sample data. Estimates using the slope of the F_r statistics were identical so they are not shown. The colored dots in Figure 7B show the slope estimate of the combined data relative to the expected slope, and the bars represent the middle 50% of the slopes from individual populations. All of the dispersal distributions have similar slopes. When $\sigma = 4$, the actual spread of the slope values is smaller than the the spread of the slopes for the other values of σ (not shown), but in Figure 7 the values are relative so the middle 50% is wider.

Relative Execution Time of Dispersal Functions

My implementation of the triangular distribution was the most efficient followed by the Rayleigh which took about 26.8% longer on average (Table 3). The half-normal and the exponential functions had similar execution times but took nearly 5 times longer than the triangular function. The gamma, Pareto, and Lomax were the least efficient functions running over 5 times longer than the triangular function.

Discussion

Approximating continuous dispersal on a discrete lattice will introduce obvious biases when the dispersal distance is small compared to the scale of the lattice nodes (Chipperfield et al., 2011). This bias can be seen in Fig. 1 by the jagged nature of the empirical cumulative distribution (ECDF) (especially when σ is small) compared to the CDF of the continuous distribution. In the simulation, the distance between nodes is one lattice unit so dispersal has to exceed at least a distance of 0.5 lattice units to leave the original cell. For Lomax simulations with small σ, the high probability density near zero falls rapidly before a distance of 0.5 lattice units has been reached. This means that the majority of dispersal events do not leave the parent cell. The Pareto and Lomax distributions share a similar shape and a wide tail, but unlike the Lomax distribution, the mode of the Pareto is greater than zero and almost all dispersal events leave the original cell. I refer back to the differences between the Lomax and the Pareto when I discuss whether I can differentiate results that are specific to dispersal with a high peak at zero or are more general to wide-tailed dispersal.

Allelic diversity is near the expected value predicted by the infinite alleles model for most distributions. The Lomax distributions tend to have a higher number of alleles up until $\sigma = 4$. This appears to be in agreement with Maruyama (1972) which showed that the effective population size is larger than the census size when $\sigma < 1$ which is the case in many of the Lomax simulations (Fig. 2). Because the median allele number for the Pareto simulations falls near the expected value, it seems likely that the higher allelic diversity in the Lomax simulations is due to the high probability of not dispersing. This is supported by the fact that the average diversity is slightly higher for the exponential and gamma-1 as well. When dispersal is unlikely to occur outside of the original cell, the number of migrants is low and the pool of offspring before competition will consist mostly of offspring from the same parent. It is unlikely that migrants will become established at their new location after competition and thus more alleles will be maintained.

Much of the theory of isolation-by-distance in continuous populations is based on infinite or periodic lattice models. Here I simulated dispersal in a continuous population occupying a finite lattice with absorbing boundaries to better understand the effect of the dispersal kernel on isolation-by-distance models on a more natural landscape. As expected under isolation-by-distance, the probability of identity-by-descent between neutral alleles in pairs of individuals decreases as the distance between them increases. When neighborhood size is small, the relationship is very pronounced with a high initial probability that quickly declines. As neighborhood size increases ($\sigma = 4$), this relationship nearly disappears. This is likely an affect of the size of the lattice, because on any finite landscape, dispersal greater than a certain threshold will effectively lead to panmixia.

Simulations with the different dispersal kernels show a strikingly similar pattern of isolation-by-distance. However, theory predicts that when distance is small, deviation in the shape of the dispersal kernel relative to the Rayleigh distribution will become important (Rousset, 1997, 2000). This is evident in my results when I compare the probabilities of identity-by-descent at small distances between the different dispersal kernels. When the dispersal kernel is leptokurtic, the probability is higher between individuals occupying the same location and there is a steeper decrease in identity at short distances compared to the Rayleigh results. The pattern of identity-by-descent in the thin tailed distributions, including the triangular are nearly identical to the Rayleigh. The situation is similar for the pairwise kinship except there is even greater similarity between the different dispersal kernels.

Rousset (2008a) makes it clear that the increase of genetic differentiation with distance is robust to the shape of the dispersal kernel but the overall magnitude of differentiation will depend on the shape of the kernel. Looking at the relationship between a_r and the log of distance for the simulations, I can see that the slope for each distribution is similar over larger distance values but the plots are shifted up or down depending on kurtosis. The two fat tailed distributions, the Lomax and the Pareto, have very different magnitudes with the Pareto being closer in magnitude to the thin tailed distributions. This suggests that the magnitude of this relationship greatly depends on the amount of dispersal at the origin. The a_r statistic is a ratio that compares the amount genetic differentiation between individuals at certain distance to the differentiation within a single individual. When the probability of identity-by-descent within an individual is high, the differentiation between neighbors will appear much higher due to a steep initial drop in identity. As a result, the a_r statistic will be greater for leptokurtic distributions even if the actual probability of identity is similar to other distributions.

As expected, the neighborhood-size estimates are similar to the expected value for all simulations.

Neighborhood size was slightly higher for the Lomax simulations when using allele diversity to estimate effective density. Otherwise, the slopes of the regression methods were similar and thus predicted similar neighborhood sizes. This reconfirms that neighborhood size is a robust descriptor of the decrease of genetic identity with distance. However, the actual pattern of isolation-by-distance at close distances does depend on the shape of the dispersal distribution. Additionally, the pattern of isolation-by-distance may depend on the size and scale of the landscape. The results shown here are from simulations on the same sized landscape and the distributions are truncated at the landscape edge. Therefore, it is possible that the pattern of isolation-by-distance, especially for the fat-tailed distributions, will depend on the size and scale of the landscape.

The triangular distribution has not been considered as a reasonable distribution to use for modeling biological dispersal. However, as discussed previously, it arises from the simple assumption that dispersal is locally panmictic, making it potentially useful. When I compared the triangular distribution against some of the more popular dispersal models, there was not a large difference between the resulting patterns of isolation-by-distance.

The triangular dispersal model can serve as a null model for the probability that two lineages will meet and coalesce in a previous generation. Identity-by-descent may be defined as the total probability of coalescence between the current generation, t_0, and a generation at some time t in the past (Rousset, 2002). When a population is not panmictic due to limited dispersal, the time to coalescence depends on the probability that the two lineages will move close enough together so that there is some probability that they shared a parent in the previous generation. When the dispersal kernel has an infinite tail, there is always some small probability that two individuals coalesce even if they are very far apart. Because the triangular distribution is finite with a maximum distance of 2σ, the probability that two individuals coalesce in the previous generation is $1/(4\pi\sigma^2 D)$ if they are separated by a distance less than 2σ and zero otherwise.

The triangular distribution allows us to simulate dispersal more efficiently than other dispersal kernels because it is uniform over a finite area. It allows us to easily pre-compute probabilities of dispersal to neighboring cells and use an efficient discrete sampling algorithm to sample dispersal positions. A similar approach is possible for other dispersal distributions. For distributions with infinite tails this would require defining a truncated distribution which captures the bulk of the dispersal probabilities. Then, for two dimensions, double integrals would need to be calculated to determine the probabilities of dispersal to locations on the lattice. These pre-computations are laborious because in addition to the double integrals, many cells will have non-zero probabilities. For the triangular distribution, only cells in a radius of 2σ will have non-zero probability and since the distribution is uniform, the probabilities are easy to calculate. The triangular distribution algorithm introduced here is more efficient than the other dispersal distributions that were implemented. The differences in absolute time shown in Table 3 are not dramatically different because the simulations were run for comparatively few generations; however, when running the full simulation, there was a time savings of several hours.

The results suggest that the relationship between probability of identity-by-descent and distance is similar for a wide range of dispersal kernels in a continuous population and both theoretical and computational concerns suggest that triangular distributions should be included in the molecular ecologists toolkit. However, these results should not be taken to mean that it is always safe to ignore the shape of the dispersal kernel. As I demonstrate here, the high number of extremely limited dispersal events under the Lomax distribution increases the probability of identity-by-descent within a cell. In a hermaphroditic plant this could translate into a higher rate of self-fertilization. The shape of the tail can impact the number of long distance dispersal events which may affect the rate of population expansion, colonization, responses to climate change, population fragmentation and the movement of genes between locally adapted populations. Each of these processes will be affected by the dispersal distribution chosen for the simulation. However, when simulating a finite, isolated population at equilibrium, in many cases the shape of the dispersal kernel does not appear to have a strong effect on the resulting pattern of isolation-by-distance. Because speed is an important factor in deploying isolation-by-distance simulations in analytical contexts, e.g. approximate Bayesian computation, I recommend using the triangular distribution when long distance dispersal and other features of the dispersal kernel can safely be ignored.

Table 1. The dispersal function, range, probability density function, and two-dimensional density plots when $\sigma = 1$. The counts for the two-dimensional density plots were square-root transformed and where applicable the α parameter used for the two-dimensional density plot is indicated in bold.

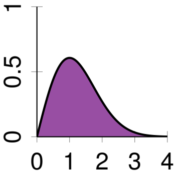
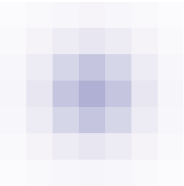
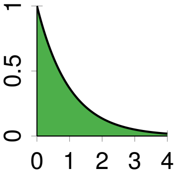
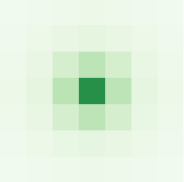
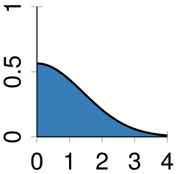
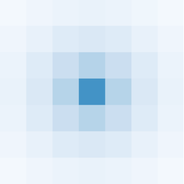
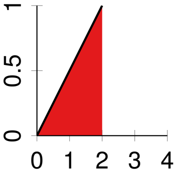
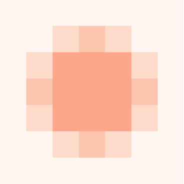
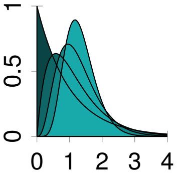
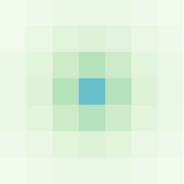
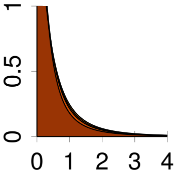
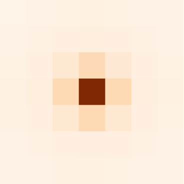
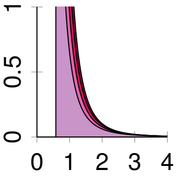
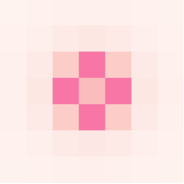
	Probability Density Function (PDF)	PDF	2D Density Plot
Rayleigh	$f(r,\theta;\sigma) = \frac{1}{2\pi}\frac{r}{\sigma^2}e^{\frac{-r^2}{2\sigma^2}}$ $r \geq 0$		
Exponential	$f(r,\theta;\sigma) = \frac{1}{2\pi}\frac{1}{\sigma}e^{\frac{-r}{\sigma}}$ $r \geq 0$		
Half-Normal	$f(r,\theta;\sigma) = \frac{1}{2\pi}\frac{1}{\sigma\sqrt{\pi}}e^{\frac{-r^2}{4\sigma^2}}$ $r \geq 0$		
Triangular	$f(r,\theta;\sigma) = \frac{1}{2\pi}\frac{r}{2\sigma^2}$ $0 \leq r \leq 2\sigma$		
Gamma	$f(r,\theta;\sigma,\alpha) = \frac{1}{2\pi}\frac{\left(\frac{\sqrt{a(a+1)}}{\sqrt{2}\sigma}\right)^{\alpha}}{\Gamma(a)}r^{\alpha-1}e^{-r\frac{\sqrt{a(a+1)}}{\sqrt{2}\sigma}}$ $\alpha = 1, \mathbf{2}, 4, 8$ $r \geq 0$		
Lomax	$f(r,\theta;\sigma,\alpha) = \frac{1}{2\pi}\frac{\alpha\left(1+\frac{r}{\sqrt{(\sigma^2(\alpha-2)(\alpha-1)}}\right)^{-(\alpha+1)}}{\sqrt{\sigma^2(\alpha-2)(\alpha-1)}}$ $\alpha = \mathbf{2.4}, 2.6, 2.8, 3.0$ $r \geq 0$		
Pareto	$f(r,\theta;\sigma,\alpha) = \frac{1}{2\pi}\frac{\alpha\sqrt{\frac{2\sigma^2(\alpha-2)}{\alpha}}}{r^{a+1}}$ $\alpha = \mathbf{2.4}, 2.6, 2.8, 3.0$ $r \geq \sqrt{\frac{2\sigma^2(\alpha-2)}{\alpha}}$		

Table 2. Estimates of allele diversity, $\hat{\theta}_k$, effective population density, $\hat{D}_e$, dispersal, s^2, and neighborhood size. Neighborhood size is estimated two different ways. $\hat{N}_{b(\theta)}$ is $4\pi s^2 \hat{D}_e$ where $\hat{D}_e$ is estimated from $\hat{\theta}_k$. $\hat{N}_{b(a_r)}$ is twice the inverse of the slope of a_r and the log of distance. The expected neighborhood size ($4\pi\sigma^2 \cdot 1$) is 12.56, 28.28, 50.26, and 201.06 for σ =1, 1.5, 2, and 4, respectively.

	σ									
	1					1.5				
	$\hat{\theta}_k$	$\hat{D}_e$	s^2	$\hat{N}_{b(\theta_k)}$	$\hat{N}_{b(a_r)}$	$\hat{\theta}_k$	$\hat{D}_e$	s^2	$\hat{N}_{b(\theta_k)}$	$\hat{N}_{b(a_r)}$
Ray	1.82	0.91	0.99	11.31	13.07	1.83	0.91	2.33	26.79	31.16
Exp	2.09	1.04	1.04	13.70	14.32	2.04	1.02	2.26	29.04	29.00
Nor	1.94	0.97	0.98	11.94	13.49	1.91	0.95	2.31	27.69	30.61
Tri	1.82	0.91	1.00	11.37	13.58	1.83	0.92	2.36	27.18	31.02
Gam 1	2.07	1.04	1.05	13.63	14.41	2.01	1.00	2.32	29.22	30.07
Gam 2	1.89	0.94	0.98	11.62	12.80	1.85	0.92	2.32	26.98	30.13
Gam 4	1.83	0.92	1.00	11.49	12.88	1.87	0.94	2.32	27.31	28.27
Gam 8	1.80	0.90	1.01	11.45	13.31	1.79	0.90	2.32	26.16	29.91
Lom 2.4	2.97	1.49	1.06	19.70	13.41	2.62	1.31	2.16	35.53	26.65
Lom 2.6	2.88	1.44	0.97	17.61	13.23	2.47	1.24	2.34	36.25	26.11
Lom 2.8	2.73	1.36	1.04	17.78	12.82	2.41	1.21	2.22	33.66	25.30
Lom 3	2.72	1.36	1.00	17.07	14.28	2.36	1.18	2.34	34.71	28.50
Par 2.4	1.98	0.99	0.98	12.18	11.71	1.93	0.97	2.19	26.56	27.12
Par 2.6	1.95	0.98	1.04	12.74	13.82	1.81	0.91	2.28	25.98	27.95
Par 2.8	1.90	0.95	0.97	11.57	12.25	1.85	0.93	2.25	26.16	30.85
Par 3	1.89	0.95	0.99	11.80	13.56	1.89	0.94	2.24	26.54	29.79
	2					4				
	$\hat{\theta}_k$	$\hat{D}_e$	s^2	$\hat{N}_{b(\theta_k)}$	$\hat{N}_{b(a_r)}$	$\hat{\theta}_k$	$\hat{D}_e$	s^2	$\hat{N}_{b(\theta_k)}$	$\hat{N}_{b(a_r)}$
Ray	1.97	0.99	4.07	50.39	58.81	2.02	1.01	16.11	204.93	236.23
Exp	2.02	1.01	4.08	51.88	49.60	2.09	1.05	16.16	212.48	154.94
Nor	1.95	0.97	4.08	49.87	55.00	2.04	1.02	16.04	205.76	189.69
Tri	1.94	0.97	4.11	50.13	54.57	2.09	1.04	16.09	210.87	245.02
Gam 1	2.03	1.01	4.06	51.74	52.25	2.16	1.08	16.15	218.67	257.28
Gam 2	1.89	0.95	4.12	48.88	54.39	2.02	1.01	16.08	204.41	214.04
Gam 4	1.94	0.97	4.08	49.80	55.60	1.98	0.99	15.94	197.97	191.02
Gam 8	1.89	0.94	4.06	48.21	52.47	2.02	1.01	16.11	203.96	231.04
Lom 2.4	2.48	1.24	3.98	62.01	47.94	2.19	1.09	16.06	220.82	180.03
Lom 2.6	2.36	1.18	3.94	58.49	48.10	2.15	1.07	15.45	208.62	219.14
Lom 2.8	2.27	1.13	4.16	59.23	51.08	2.14	1.07	15.81	212.44	241.05
Lom 3	2.24	1.12	3.97	56.05	47.23	2.07	1.04	16.55	215.21	211.19
Par 2.4	1.93	0.97	4.13	50.12	48.20	2.03	1.02	16.04	204.74	192.65
Par 2.6	1.95	0.97	4.11	50.29	51.74	2.03	1.02	15.91	203.23	189.19
Par 2.8	1.98	0.99	4.02	49.95	47.73	1.95	0.97	15.53	189.90	219.90
Par 3.	1.98	0.99	4.10	50.92	49.58	2.01	1.00	16.30	205.48	169.53

Table 3. Triangular dispersal algorithm is the most efficient. Execution time and relative time for 10^9 dispersal events from different dispersal functions ordered from most to least efficient.

Dispersal Function	**CPU Seconds**	**Relative Time**
Triangular	21.853	1.000
Rayleigh	27.713	1.268
Exponential	106.434	4.870
Half Normal	106.771	4.886
Gamma	119.357	5.462
Pareto	127.218	5.822
Lomax	127.376	5.829

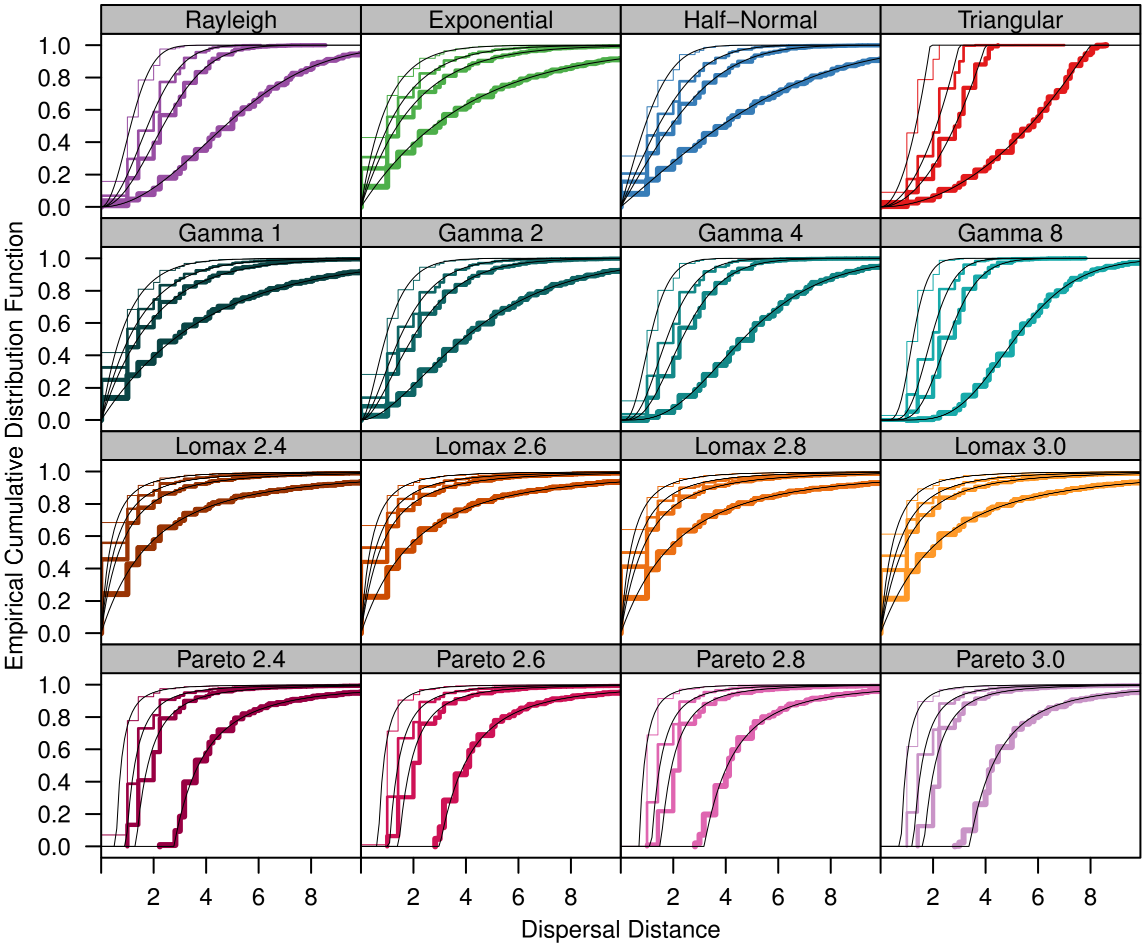


Figure 1. Effect of discretization of continuous dispersal distributions. The plot shows the empirical cumulative distribution function for each dispersal distribution on a discrete lattice compared to the CDF of its continuous counterpart (black line). The different plots in each panel represent simulations run using different σ parameters: 1, 1.5, 2, and 4. An increase in the thickness of the line corresponds to increasing σ parameter.

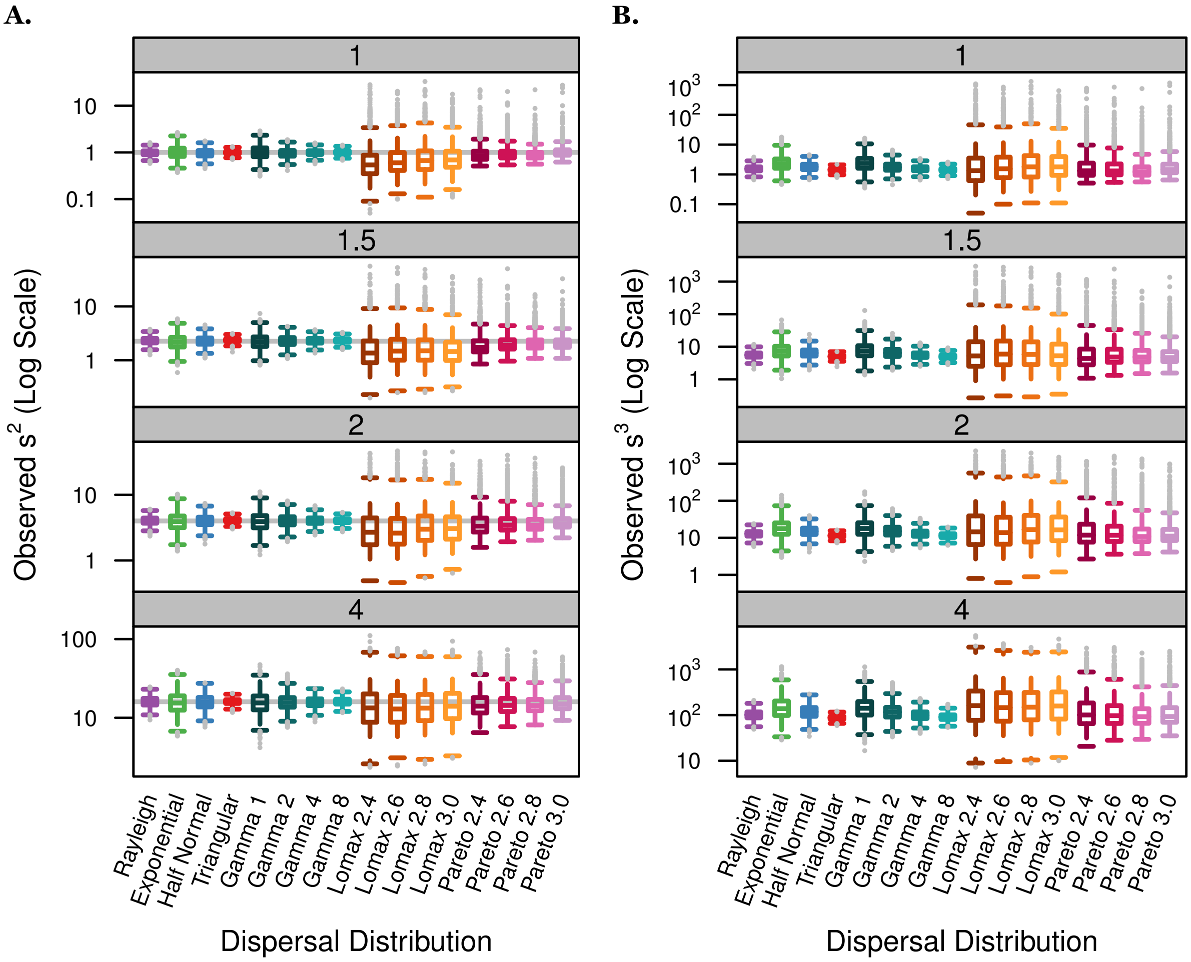


Figure 2. Different dispersal kernels have equivalent second moments but different third moments. Each panel represents groups of simulations run with different σ parameters and contains box-whisker plots summarizing the distribution of the average (A) squared or (B) cubed parent-offspring distance of 2,000 sampled transects. The top and bottom of the boxes represent the 75% and 25% quartiles, while the central bar represents the median. The gray dots outside the whiskers represent outliers. The gray horizontal line in A represents the expected σ^2 value. The observed values are shown on a log scale which is different in some panels.

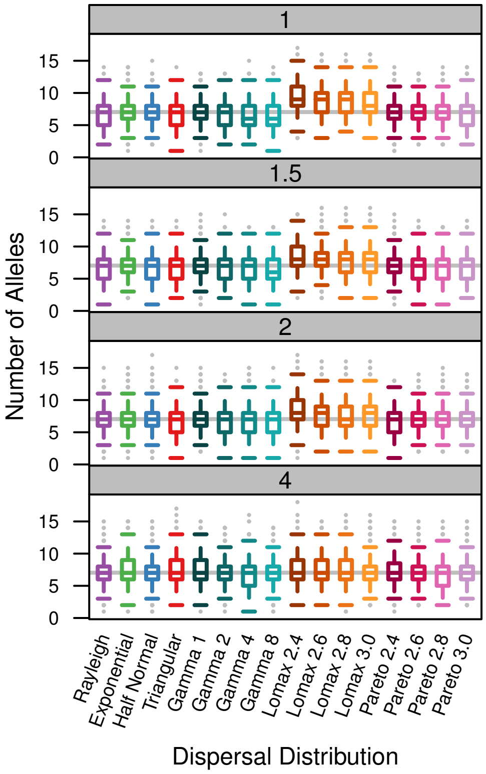


Figure 3. The distribution of unique alleles is similar for most dispersal kernels. Each panel represents simulations run with a the σ parameter provided in the gray box. For each dispersal distribution, the box-whisker plot summarizes the number of unique alleles (k) found in 2,000 50-individual transects. The gray horizontal line represents the expectation under the infinite alleles model. The features of the box-whisker summary are the same as Fig. 2.

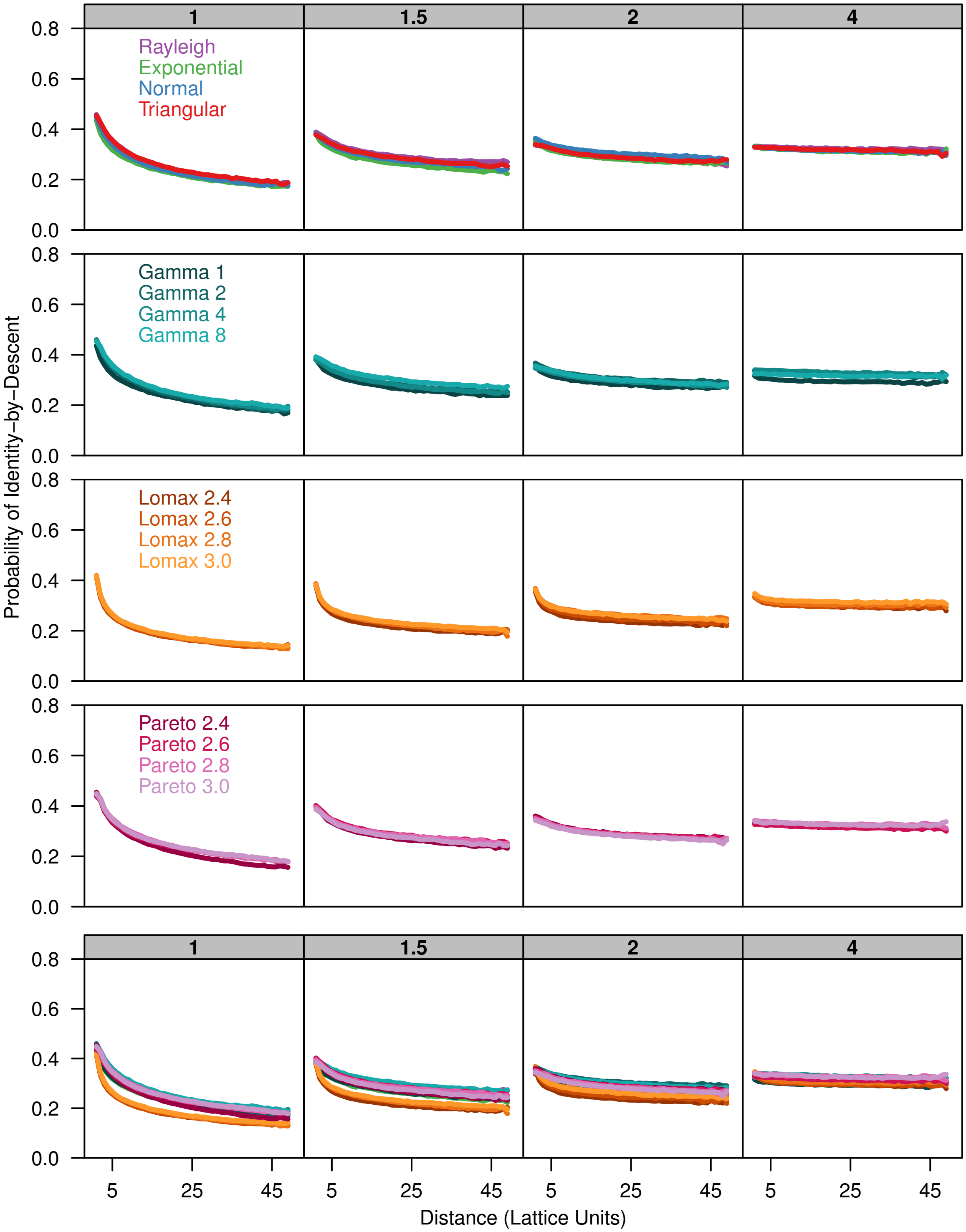


Figure 4. Identity-by-descent is similar between different dispersal models. Each plot shows the average probability of identity-by-descent for pairs of individuals in each distance class. Each panel represents simulations run with different σ parameters (gray box) for different groups of dispersal distributions; all of the dispersal distributions are plotted together in the last row.

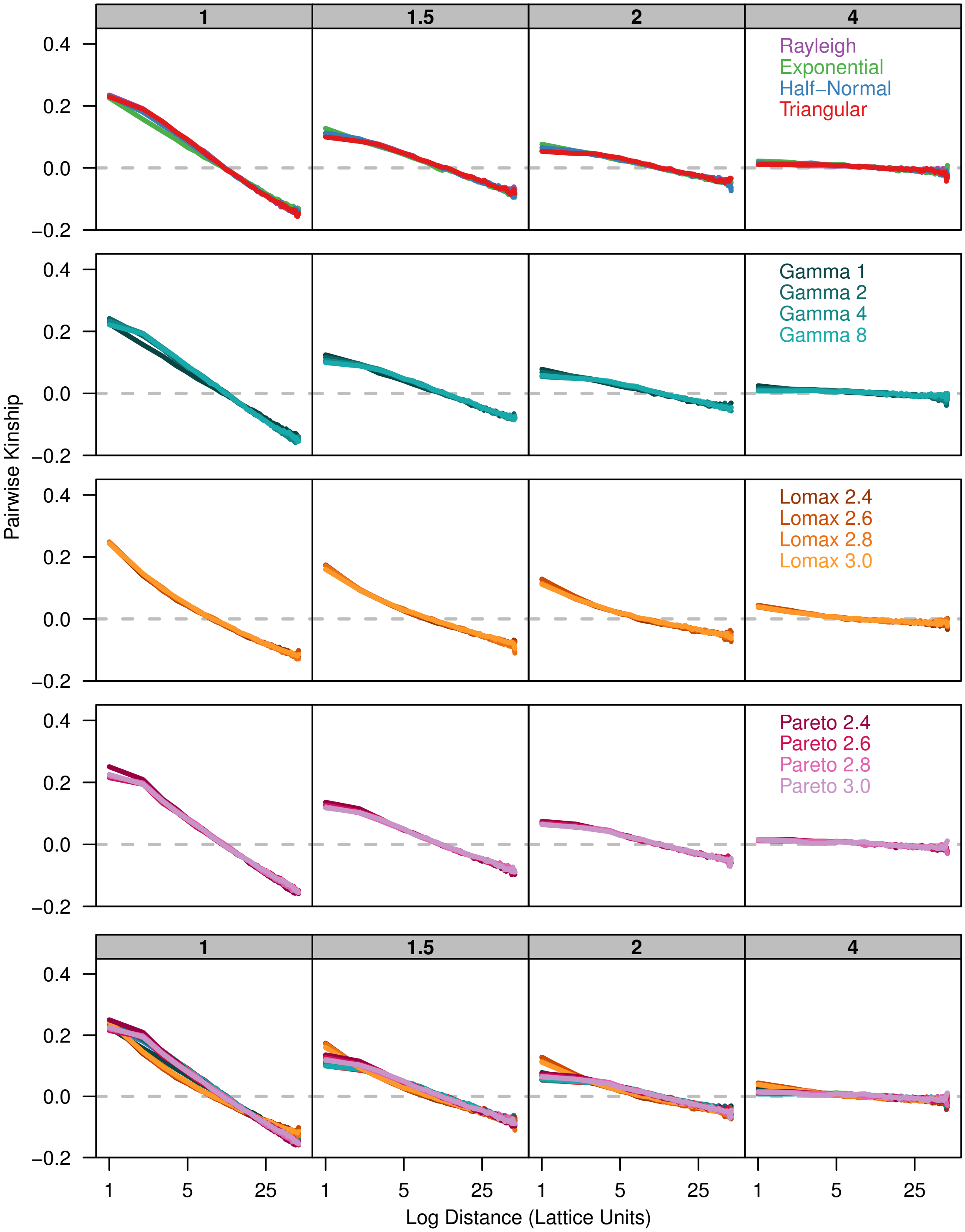


Figure 5. Kinship coefficients are similar between different dispersal models. Each plot shows the average kinship coefficient for pairs of individuals over the log of the distance between them. Each panel represents simulations run with different σ parameters (gray box) for different groups of dispersal distributions; all of the dispersal distributions are plotted together in the last row.

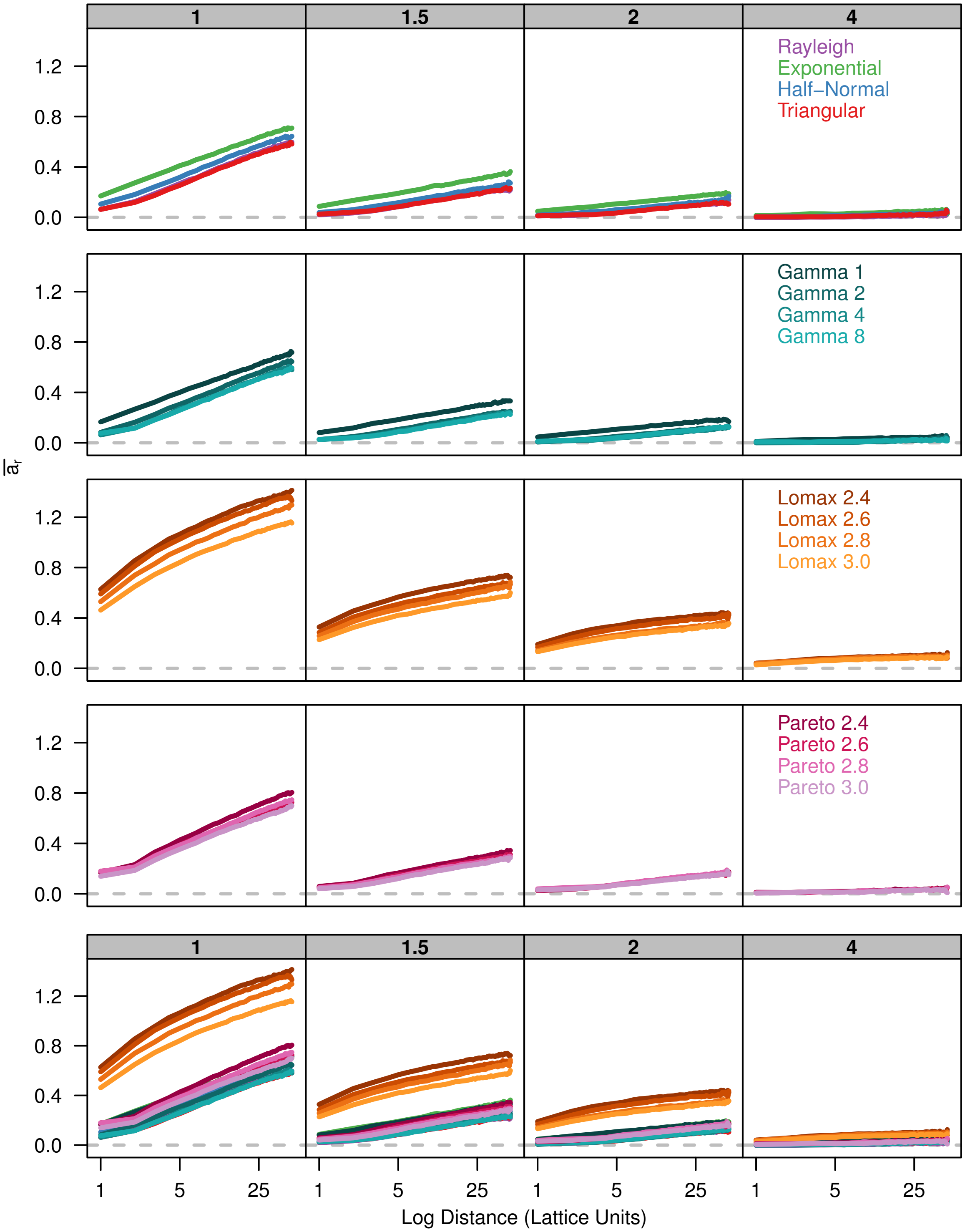


Figure 6. Slopes of genetic differentiation are similar between different dispersal models. Each plot shows the average differentiation, a_r, for pairs of individuals over the log of the distance between them. Each panel represents simulations run with different σ parameters (gray box) for different groups of dispersal distributions; all of the dispersal distributions are plotted together in the last row.

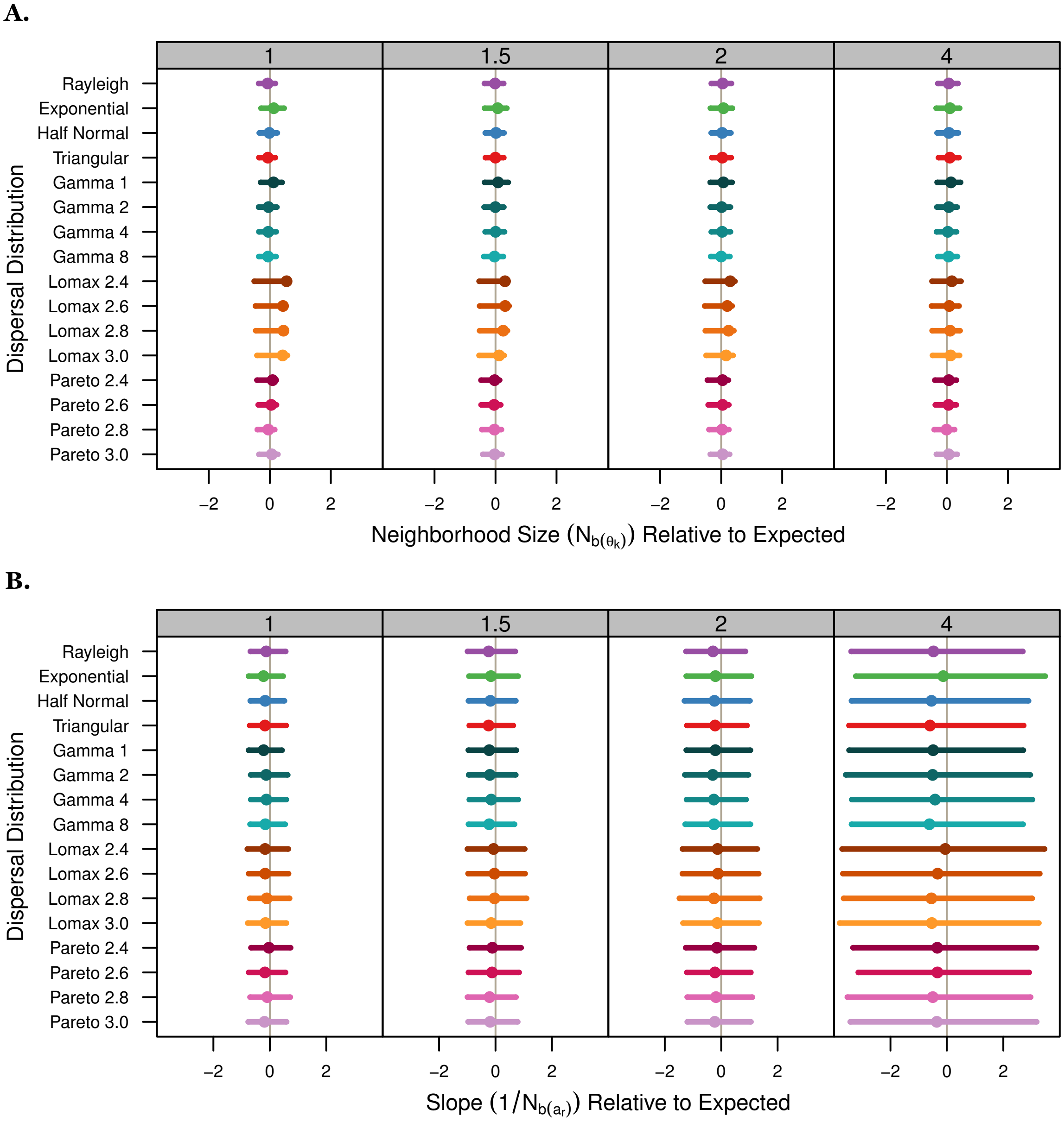


Figure 7. Estimated neighborhood sizes are similar across all dispersal distributions. Neighborhood size is estimated in two different ways. (A) $N_{b(\theta_k)}$ is $4\pi s^2 \hat{D}_e$ where $\hat{D}_e$ is estimated from $\hat{\theta}_k$. The dot is the average from all populations samples and the bars represent the middle 50% of estimates from individual samples. (B) The slope estimates, $\frac{2}{N_{b(ar)}}$, of a_r and the log of distance. The dots represent the slope estimate from the combined data from all samples and the bars represent the middle 50% of slopes from individual samples.

Chapter 3

BAYESIAN ESTIMATION OF NEIGHBORHOOD SIZE USING A COMPOSITE MARGINAL LIKELIHOOD

Abstract

Wright's neighborhood size is a density-dependent measure of gene flow that quantifies the degree of spatial genetic structure that is due to isolation-by-distance in a population. The neighborhood size formula, $N_b = 4\pi\sigma^2 D_e$, contains two important demographic parameters: the mean squared parent-offspring dispersal distance, $2\sigma^2$, and the effective density of individuals in the population, D_e. Several methods have been devised that make reasonable point estimates of neighborhood size but so far none have attempted estimates in a Bayesian framework. Here I describe a Bayesian method to estimate neighborhood size using a composite marginal likelihood (CML) in place of a full likelihood. The model uses an approximation of the Wright-Malècot (WM) formula to link observed patterns of isolation-by-distance to the neighborhood size parameter. Data on the proportion of pairs that are identical-in-state (IIS) at different distances and at different neutral loci can then be modeled using individual binomial likelihoods; the product of each of these likelihoods is the CML. The neighborhood size parameter is modeled using a log-normal prior and an MCMC algorithm approximates the neighborhood size marginal posterior distribution. I tested this method using data generated directly from the model and from a spatially-explicit lattice simulation and show that the estimates have high coverage and low error but can be biased when fewer markers are analyzed. I examine how well the model performs under different sampling schemes and when certain assumptions are violated. I also applied this method to analyze microsatellite data from two populations of maritime pine (*Pinus pinaster* Aiton) and compared my estimates to those obtained using a different method. Finally, one of the advantages of using this Bayesian approach is the ability to incorporate prior information about density to get a better estimate of the dispersal parameter; I demonstrate that the distribution of the dispersal parameter estimates reflects the level of certainty in the density prior.

Introduction

The dispersal ability of many species is spatially limited and a large proportion of offspring tend to reproduce close to their birth site (Caine et al., 2000; Howe and Smallwood, 1982; Kot et al., 1996). Wright (1943, 1946) recognized that limited dispersal can lead to localized breeding and genetic differentiation within continuous populations and he developed the isolation-by-distance model to better understand the genetic consequences under these conditions.

In a panmictic population, parents are equally likely to come from any part of the population; however, if dispersal is limited, potential parents are restricted to only the individuals within a local region. The size of this local region depends on dispersal capability and the number of individuals in this region depends on population density. Wright introduced the term "neighborhood" size (N_b) to describe the effective number of individuals within this local population.

Wright measured the magnitude of the neighborhood as the inverse of the probability that two gametes at the same location came from the same parent ($1/N_b$). For a two-dimensional population, Wright assumed that individuals disperse according to a normal distribution with a standard deviation σ, along each axis, and he calculated that $N_b = 4\pi\sigma^2 D_e$, where $2\sigma^2$ is the mean squared dispersal distance, and D_e is the the effective population density. Under this model, the neighborhood size captures 86.5% of potential parents of central individuals.

Neighborhood size is important because it quantifies the spatial structure that arises as a result of isolation-by-distance, and it contains information about two important population parameters: dispersal and density. Unfortunately, these parameters, particularly the dispersal parameter (Slatkin, 1987), are difficult to measure directly so neighborhood size is often inferred indirectly from observed patterns of isolation-by-distance.

Under isolation-by-distance, the genetic similarity shared between individuals decreases as the geographical distance between them increases. Malécot (1969) described the relationship between individuals in terms of the probability that their alleles are identical-by-descent. Two homologous genes are identical by descent if they are both descended from the same common ancestor and no mutations have occurred. For a population at mutation-drift-migration equilibrium, the Wright-Malècot (WM) formula calculates the probability of identity-by-descent (IBD) between pairs of individuals given the neighborhood size and the distance between the pair. When the pattern of isolation-by-distance is strong, neighborhood size is small relative to the size of the population, and when pattern of isolation-by-distance is weak or undetectable, neighborhood size is large and the population is approaching panmixia. The WM formula provides the necessary link between the observed pattern of isolation-by-distance and neighborhood size, and it can be used to make inferences about neighborhood size if information about the probability of IBD between individuals is known.

Existing Methods for Estimating Neighborhood Size

There are several existing methods that use the WM theory to estimate neighborhood size within continuous populations. For two-dimensional landscapes, Rousset (2000) uses the regression of pairwise measures of genetic differentiation on the logarithm of geographical distance. The inverse of the slope of the regression provides an estimate of $4\pi\sigma^2 D_e$. The relationship between genetic differentiation and the log of distance is only linear over a certain range of distances. When the distance between pairs is smaller than σ, the relationship depends on the shape of the dispersal kernel; when the distance is greater than $0.56\sigma/\sqrt{2\mu}$ the relationship depends on the mutation rate (Rousset, 1997). An implementation of the method is provided in the GENEPOP software package (Raymond and Rousset, 1995; Rousset, 2008b).

Hardy and Vekemans (1999) describe a similar approach using pairwise kinship coefficients. The regression slope of the pairwise kinship coefficients and the log of distance is $-(1-F)/4\pi\sigma^2 D_e$ where F is the inbreeding coefficient (Hardy and Vekemans, 1999; Vekemans and Hardy, 2004). This method is implemented in the SPAGeDi (Spatial Pattern Analysis of Genetic Diversity) software package (Hardy and Vekemans, 2002). Neighborhood size estimates from these methods are robust to different mutation models, different mutation rates (under $\mu = 10^{-3}$), the shape of the dispersal kernel, and spatial heterogeneity (Leblois et al., 2003, 2004).

The pairwise statistics used in the previous two methods are defined in terms of the probability of identity-in-state (ISS) rather than the probability of IBD because the latter is usually difficult to measure directly. Two genes are IIS if they share the same allele but are not necessarily descendants from the same common ancestor. The statistics are also defined in the form of a ratio that compares the probability of IIS for pairs of individuals at a certain distance to the probability of IIS in some reference group of genes (Rousset, 2002; Vekemans and Hardy, 2004). The Rousset (2000) statistic uses pairs of genes sampled from within individuals as a reference, and the Hardy and Vekemans (1999) statistic uses random genes sampled from the population. Statistics based on the probability of IIS depend on the mutation process and are not generally equivalent to the probability of IBD, however when the mutation rate is low, the ratio of IIS probabilities is approximately equal to similar ratios based on probability of IBD (Rousset, 2002; Vekemans and Hardy, 2004).

Barton et al. (2013) take a different approach to obtain a maximum likelihood estimate of neighborhood size. They assume that fluctuations in allele frequencies are small and can be modeled using a bivariate Gaussian likelihood. To model spatial correlations they calculate covariances of allele frequencies for observed data and for expected values generated from an approximation of the WM formula. For large distances, their WM approximation includes a modified Bessel function of the second kind and degree zero. At smaller distances, this function diverges from the WM formula so they substitute a constant value. Currently, this method has not been implemented in any publicly available software package.

Current Method

The method described here is the first to use a Bayesian approach to estimate neighborhood size. A Bayesian approach offers several advantages. First, prior knowledge about the model parameters can be incorporated, and I will demonstrate that priors can be used to get better estimates for certain model parameters. Second, MCMC methods make it is easier to create more complex models (e.g. hierarchical models) and could allow greater flexibility in handling different types of data.

This method is also unique because it applies a composite marginal likelihood (CML). As models become more complex, CMLs are becoming increasingly popular because they serve as a good substitute when the full likelihood is impractical to compute. The CML combines marginal likelihoods for different subsets of data as if they were independent and ignores any higher order dependencies that may exist. Despite this misspecification, inferences based on CML methods perform well in many cases (Pauli et al., 2011; Xu and Reid, 2011). One problem with CML methods is that they often lead to deceptively precise inference which, in a Bayesian framework, means underestimation of the width of the credible interval (Menéndez et al., 2014; Pauli et al., 2011). This issue becomes more pronounced when additional strain is placed on the already tenuous assumption of independence between the marginal likelihoods.

This method attempts to fit an observed pattern of isolation-by-distance using an approximation of the WM formula. To approximate the WM formula, I used a Taylor series for short distances and a Bessel function, similar to Eq. A.6 in Barton et al. (2013), for long distances. For a given distance class and marker locus, the proportion of individuals that are IIS is modeled using a binomial likelihood and each likelihood is combined to form a composite marginal likelihood. An MCMC algorithm is used to make inferences from the model using the composite marginal likelihood and a prior distribution for the neighborhood size parameter.

Here I analyze the performance of this method under different dispersal parameters, mutation rates, mutation models, dispersal distributions, and sampling schemes using data generated from the model and from a spatially-explicit lattice based simulation. I also apply the method to microsatellite data from two populations of *Pinus pinaster* Aiton (De-Lucas et al., 2009a) and compare my estimates to those obtained using SPAGeDi (Hardy and Vekemans, 2002).

Methods

Model

The Wright-Malècot (WM) formula defines the probability that genes from two individuals are identical-in-state (IIS) as a function of the distance between them (see derivation from Barton et al., 2013).

$$\phi(x) = \frac{1-\phi_{(0)}}{2N_b} \sum_{t=1}^{\infty} \frac{e^{\left(-2\mu t + \frac{-x^2}{4\sigma^2 t}\right)}}{t} \tag{3.1}$$

In the formula, x is the distance between the pair, μ is the mutation rate, and $N_b = 4\pi\sigma^2 D_e$ is neighborhood size where $2\sigma^2$ is the second moment of the dispersal distribution, and D_e is the effective population density or the effective number of individuals per unit of area. Equation 3.1 assumes the infinite alleles mutation model (IAM) where $e^{-2\mu}$ represents the probability that neither gene mutates in a single generation. Assuming Gaussian dispersal, the summation represents the probability that neither gene has mutated since they shared a common ancestor, and under the IAM this can be defined as the probability of IBD (Rousset, 1996). Other mutation models could be substituted, but due to homoplasy they would model the probability of IIS. (Rousset, 1996).

Because of the infinite series, calculating probabilities directly from the WM formula presents a computational challenge. To simplify the computation, I approximated the WM formula in two different ways. First, I use a modified Bessel function of the second kind and degree zero ($\mathcal{K}_0$) which is a good approximation for long distances but it begins to diverge a shorter distances (Barton et al., 2013; Rousset, 1997; Sawyer, 1977). To approximate the WM formula at short distances, I derived a Taylor polynomial. I determined empirically that the Taylor polynomial should be calculated to 34 terms because this allows it to be accurate up to a distance of 5σ which is where the relative error of the Bessel approximation falls below 10^{-6}.

$$\phi(x) \approx \begin{cases} \frac{\sum_{t=0}^{34} \frac{\mathrm{Li}_{(t+1)}(e^{-2\mu}) \cdot x^{2t} \cdot -1^{t}}{(2t)!! \cdot 2^{t+1} \cdot \sigma^{2t}}}{N_b - \log\left(\sqrt{1-e^{-2\mu}}\right)} & x \leq 5\sigma \\ \\ \frac{\mathcal{K}_0\left(\frac{|x|}{\sigma}\sqrt{1-e^{-2\mu}}\right)}{N_b - \log\left(\sqrt{1-e^{-2\mu}}\right)} & x > 5\sigma \end{cases} \tag{3.2}$$

The Taylor polynomial in 3.2 is still a fairly complex calculation due to the double factorial and the polylog function ($\mathrm{Li}_s(z)$). To speed up the calculation, I treated the mutation rate as a constant so that I could pre-calculate and cache the polylog terms.

The WM formula assumes an infinite population size, so the next challenge was to adjust the WM approximation to fit data from a finite population. According to the WM formula, the probability of IBD between two individuals will approach zero as the distance between them increases; however, in a finite population, the probability approaches the average probability of IBD in the population. To correct for this, I exploited the fact that the correlation of the probabilities over distance is approximately equal for finite and infinite populations:

$$\underbrace{\frac{\phi_x - \bar{\phi}}{1 - \bar{\phi}}}_{\text{infinite}} \approx \underbrace{\frac{p_x - \bar{f}}{1 - \bar{f}}}_{\text{finite}} \tag{3.3}$$

where ϕ_x and p_x are probabilities of IBD between individuals separated by distance x and $\bar{\phi}$ and $\bar{f}$ are average probabilities of IBD in an infinite and a finite population, respectively. I then solve for the probability of IBD in a finite population.

$$p_x = \bar{f} + (1 - \bar{f}) \cdot \frac{\phi_x - \bar{\phi}}{1 - \bar{\phi}} \tag{3.4}$$

This equation simultaneously solves another issue with the model. As previously mentioned, the data that will be provided to the model will likely be probabilities of IIS rather than probabilities of IBD. These probabilities are not directly equivalent but they are approximately equal in the ratio form presented in 3.3. Consequently, Equation 3.4 will allow the model to fit IIS data using the IBD values from the WM formula, assuming the mutation rate is not too high (Rousset, 2002; Vekemans and Hardy, 2004).

Composite Marginal Likelihood

Together with the probability of IIS (p_x) determined by the WM approximation, data on the proportion of IIS pairs can be modeled using a binomial likelihood. Unfortunately, such a likelihood only describes the relationship for a single distance class and a single locus. Modeling the full relationship across many distance classes and loci would be extremely complex so instead I opted to use a composite marginal likelihood.

The composite marginal likelihood treats each of the marginal binomial likelihoods as if they were independent and is therefore the product over each marginal likelihood or equivalently the sum over each log likelihood.

$$l_c = \sum_{i=1}^{d} \sum_{j=1}^{m} y_{ij} \log(p_{ij}) + (n_{ij} - y_{ij}) \log(1 - p_{ij}) \tag{3.5}$$

Equation 3.5 is the log CML for the model where y_{ij} is the number of identical alleles out of the total n_{ij} alleles examined and p_{ij} is the probability of IIS determined by the finite WM approximation for distance class, d, and locus, m.

The misspecification of the model that results from using a CML is magnified when dependence between the marginal likelihoods is increased. To minimize the dependence, one can avoid reusing the same genes to estimate the IIS proportion for multiple distance classes.

Bayesian Inference

The log CML depends on p_{ij} which is determined by the finite WM approximation which in turn depends on several parameters: μ, $\bar{f}$, and N_b. I treat the mutation rate, μ, as a constant that is provided to the model. The average probability of IIS, $\bar{f}$ is estimated from the sample as the total number of IIS pairs among all sampled pairs of individuals. The parameter of interest, N_b, can be broken down into the product $4\pi\sigma^2 D_e$, where D_e and σ are non-identifiable parameters. Neighborhood size, N_b, and effective density, D_e, are assigned independent log-normal priors, and the dispersal parameter, σ, is calculated deterministically as $\sigma = \sqrt{\frac{N_b}{4\pi D_e}}$. The posterior is then proportional to:

$$p(N_b, D_e|y) \propto p(y|N_b, D_e)p(N_b)p(D_e) \tag{3.6}$$

An MCMC algorithm was used to sample from the posterior. The model was programmed in Python using the PyMC library (Patil et al., 2010) and posterior samples were generated using the Metropolis-Hastings algorithm (Hastings, 1970; Metropolis et al., 1953). MCMC chains were run for 30,000 iterations after a 10,000 iteration burn-in and thinned every 6th iteration for a total of 5,000 samples.

Model Comparison

The neighborhood size estimate can only be obtained when a pattern of isolation-by-distance is discernible in the population sample. To establish whether a sample exhibits a detectable pattern of isolation-by-distance, I compared the Deviance Information Criterion (DIC) (Spiegelhalter et al.,

2002) for the model to the DIC for a null model of no isolation-by-distance where the probability of IIS was set to $\bar{f}$ for every distance class. The DIC is calculated as

$$\text{DIC } = p_D + \bar{D}$$

where p_D is a measure of the effective number of parameters and $\bar{D}$ is the expectation of the model deviance. Methods for calculating the DIC value are implemented in PyMC. For simple Bayesian model comparisons, models with a smaller DIC are preferred. Therefore, large $\Delta\text{DIC} = \text{DIC}_{H_o} - \text{DIC}_{H_a}$ values indicate a strong pattern of isolation-by-distance while small ΔDIC values indicate weak or no isolation-by-distance. The magnitude of the DIC values depend on the likelihood function which can vary for different data sets, so I used relative $\delta\text{DIC} = \Delta\text{DIC}/\text{DIC}_{H_o}$ to get more standardized differences.

Generating Data from the Model

To evaluate the performance of the MCMC algorithm and the influence of the prior, I tested the model using data generated from independent binomial counts based on the likelihood in 3.5. Unless otherwise noted, the parameters were set to $D_e = 1$, $\sigma = 1$, $\mu = 0.0001$, and $\bar{f} = 0.39$. The value of $\bar{f}$ was chosen because it was the average value from lattice simulations with $\mu = 0.0001$. Given the parameters, I was able to generate data by drawing random counts from a binomial distribution for each distance class and for a certain number of loci. When running the model, the mean for the neighborhood size prior was set to twice the true value, and the starting value was set to four times the true value. The mean for the density prior was set to the true value and the starting value was twice the true value. Unless otherwise noted, the precision for both priors was set to $\tau = 0.0001$.

Data from Spatially-Explicit Lattice Simulation

I also tested the model using data generated from a spatially-explicit lattice-based simulation. In the simulation, diploid individuals occupied a 100×100 toroidal lattice with a constant density of one individual per node ($D_e = 1$). In each discrete generation, parents generated 100 gametes that were displaced from the parent cell a normally distributed distance along each axis where $\sigma_x = \sigma_y$. Gametes carried a certain number of independent marker loci that were inherited through independent assortment, and each locus mutated according to the infinite alleles model with rate $\mu = 0.0001$ to a new selectively neutral allele. After dispersal, two gametes from each lattice cell were randomly chosen to become a parent in the next generation. Some simulations deviate from the above description of mutation rate, mutation model, dispersal model, or sampling scheme and these changes are noted in the appropriate sections.

Each simulation was run for a 20,000 generation burn-in to reach a drift-mutation equilibrium and then population samples were collected every 10,000 generations for nearly independent samples. The populations were sampled in two different ways. For the first sampling method, 20 pairs of individuals were randomly chosen without replacement for each of 20 distance classes (1–20). This resulted in a sample of 400 pairs or 800 individuals. Only orthogonal pairs along the same row or column were chosen to avoid non-integer distances. The number of identical alleles between the pairs was counted and totaled for each distance class and for each marker.

In the second sampling method, a $10\sigma \times 10\sigma$ grid of individuals was sampled from the population. When $\sigma = 1$, this resulted in a sample size of 100 individuals. All possible pairwise comparisons were made between the individuals in the sample, and the total number of identical alleles between pairs was counted and totaled for each distance class and for each marker. This sampling scheme differs from the previous method because fewer individuals are sampled and individuals are reused in multiple pairings. For this sampling scheme there were ten distance classes (1–10) at one lattice unit intervals, followed by an 11th class for individuals between 10 and 13 units apart; the distance classes are inclusive of the upper bound so that individuals that are one unit apart are placed in the first distance class. The total number of pairs that were analyzed per distance class was: 180, 322, 556, 596, 774, 632, 564, 554, 424, 258, and 90.

In addition to the infinite alleles model (IAM), simulations were run using the K-alleles model (KAM), and the step-wise mutation model (SMM). Under the IAM, each mutation results in a completely unique mutation, and any alleles that are identical-in-state are also identical-by-descent (Kimura and Crow, 1964). Under the KAM, each mutation results in a new allele selected from the remaining $K - 1$ possible alleles (Kimura, 1968), in the simulations $K = 20$. The SMM mutation model was developed to model the increase or decrease in repeat number in microsatellite loci. Under the SMM, each mutation results in either an increment or decrement of the repeat number by one (Ohta and Kumura, 1973). The number of possible repeats was contained to 20 with reflecting boundaries at the maximum and minimum values. All alleles were selectively neutral regardless of the mutation model.

In some simulations, different distributions are used to model gamete dispersal distance. These additional distributions include exponential, half-normal, triangular, gamma, Pareto, and Lomax. A description of each distribution and implementation details can be found in Section 2. The α parameters for the gamma distribution were 1.5, 2, 4, and 8 and the α parameters for the Pareto and the Lomax were 2.4, 2.6, 2.8, and 3.0. I used the Rayleigh distribution to model the isotropic bivariate normal that is assumed in the WM formula.

Application to Real Datasets

I applied my model to microsatellite data sets from two populations of maritime pine (*Pinus pinaster* Aiton) from De-Lucas et al. (2009a). Microsatellite markers are ideal for these types of studies because a large number of alleles are often maintained in the population allowing greater statistical power to be achieved (Epperson, 2005; Lynch and Ritland, 1999). In this study, De-Lucas et al. showed that spatial genetic structure is stronger within fragmented populations of maritime pine than in larger un-fragmented populations in the Iberian Peninsula. They analyzed data from six polymorphic nuclear microsatellite markers, and used SPAGeDi version 1.2 (Hardy and Vekemans, 2002) to estimate neighborhood size. I retrieved their data for the two fragmented populations (Fuentelapeña and Quatretonda) from the Dryad Digital Repository (De-Lucas et al., 2009b), and compared neighborhood size estimates from my model to the estimates they reported.

The data set contained genotypes for 78 individuals from the Fuentelapeña population, and 85 individuals from the Quatretonda population at 6 microsatellite loci. I used the same distance classes as the original study which had 6 distance classes (0–60 m) at 10 meter intervals, and a final distance class for distances greater than 60 meters. Following a similar approach that was used for the simulated data, I analyzed the data in two different ways. First, I analyzed the data using all possible pairwise comparisons, and calculated the number of shared alleles between pairs of individuals for each distance class and for each microsatellite marker. For this analysis scheme, the number of pairs per distance class varied for different markers due to missing data but on average the Fuentelapeña population had 171, 407, 508, 520, 502, 422, and 360 pairs per distance distance class, and the Quatretonda population had 227, 528, 622, 663, 541, 402, and 479 pairs per distance class.

For the second analysis approach, I attempted to limit the number of times that the data from each individual was reused. To accomplish this, I generated data sets where pairs were drawn at random without replacement for each distance class and for each marker. However, because the sample sizes were small for each population, I combined the counts for 20 data sets generated in this way. I generated 10 different data sets using this sampling scheme, and ran the MCMC model for each set. Again, the number of pairs per distance class varied due to missing data and random sampling, but on average the Fuentelapeña population had 115, 118, 115, 106, 102, 100, and 95 pairs per distance class, and the Quatretonda population had 122, 123, 120, 118, 116, 112, and 105 pairs per distance class. The average distance in each distance class was 6.76, 15.39, 24.81, 34.94, 44.89, 54.39, and 67.15 for the Fuentelapeña population, and 6.76, 15.49, 25.04, 35.06, 44.86, 54.62, and 72.98 for the Quatretonda population. The prior for neighborhood size had a mean of 2, a precision of 0.0001, and the density prior had a mean of 1, and precision of 0.0001.

Results

Performance on Data Generated from the Model

I generated 100 data sets with 20 pairs of individuals at 20 distance classes (1–20) with either 10, 20, or 30 independent loci. Figure 8 shows the estimated mean of the neighborhood size posterior (gray dots) for each data set arranged in increasing order. The expected neighborhood size is 12.56, indicated by the black dashed line. The vertical lines are the 95% credible interval for each estimate, and the credible intervals that do not cover the true value are indicated in red. The coverage percentage, the relative mean-squared error (MSE), and the bias is indicated in each panel. The percentage of credible intervals that cover the true value is close to 95% in each group. The relative MSE is low and decreases when more markers are used. The negative bias indicates that the model is more likely to underestimate neighborhood size, but the bias is reduced as more markers are used.

Neighborhood Size Estimates for Different Dispersal Parameters

To analyze the performance of the model for a range of neighborhood sizes, I generated 30 data sets from the model for each of six different dispersal parameters ($\sigma\ = 0.25, 0.5, 1, 1.5, 2$, and 4), and $D_e = 1$. The data sets represent 20 pairs of individuals with 10 independent loci for each of 20 distance classes (1–20). Figure 9 shows box-whisker plots that summarize the distribution of the relative squared error of the neighborhood size estimates (blue dots) for each dispersal parameter. The error is low for small neighborhood sizes but gets very large and more variable when σ is larger than 1.5.

Table 4 summarizes the performance of the estimates for different neighborhood sizes. When σ is less than 1.5, the average of the 30 neighborhood size estimates ($\bar{N}_b$) is close to the expected value (N_b) and the MSE, the bias, and the average width of the credible intervals is low. When σ is greater than 1.5, the average of the estimates is far from the expected value, and the relative mean-squared error, the bias, and the average credible interval widths are very high. The difference between the DIC values for the null and alternative model (ΔDIC) decreases as σ and N_b increase. This indicates that the pattern of isolation-by-distance is not strong enough to be detected so the model cannot provide a good estimate of neighborhood size; although, if the populations were sampled over greater distances, there may have been detectable isolation-by-distance.

Increasing the number of markers and the number of samples should lead to better estimates, but it is useful to determine which will have a bigger impact on the performance of the estimate. To better understand this relationship, I generated two different data sets. The first data set had a small neighborhood size ($4\pi \cdot 0.5^2 \cdot 1 = 3.14$), which according to Table 4, produced accurate and precise estimates when the sample contained 20 pairs of individuals per distance class with 10 markers. To determine the minimum amount of data that is sufficient to achieve an accurate estimate, I generated 30 data sets each with either 5, 10, or 15 markers and 1, 2, 5, or 10 pairs of individuals for each of 20 distance classes (1–20). The results in Table 5 show that increasing from 5 to 10 markers reduces the relative MSE more than increasing from 10 to 15 markers or increasing from 5 to 10 pairs per distance class. Increasing from 1 or 2 pairs per distance class to 5 or 10 pairs nearly halves the relative MSE in each case, and it continues to decrease when 20 pairs are sampled (Table 4). The average width of the credible intervals is an indicator of the precision of the estimates. When the number of pairs per distance class is low, there is a large reduction in the width of the credible interval when increasing from 5 to 10 markers, but the reduction is not as large when increasing from 10 to 15 markers. Comparing the decrease in width from 5 to 10 pairs per distance class and from 5 to 10 markers, the increase in the number of pairs has the greater impact.

For the second data set, I wanted to see how large of a sample and how many markers are necessary to get a good estimate for a larger neighborhood size ($4\pi \cdot 2^2 \cdot 1 = 50.27$). According to the model, there is a weak pattern of isolation-by-distance for this parameter set, but Table 4 shows that a sample with 10 markers and 20 pairs of individuals per distance class is not enough to detect it. Therefore, I generated 30 data sets each with either 10, 20, or 30 markers, and 20, 30, 40, or 50 pairs of individuals for each of 20 distance classes (1–20). Table 6 shows that a good estimate is possible with at least 30 markers and 50 pairs per distance class with a relative MSE of 0.059 and an average credible interval width of 4.876. On average, increasing the number of markers reduces the relative MSE and the width of the credible interval more than increasing the number of pairs by 10.

Neighborhood Size Estimates for Different Levels of Genetic Diversity

When generating data from the model, the $\bar{f}$ parameter determines the average probability of IIS in the sample. Genetic diversity and $\bar{f}$ have an inverse relationship such that alleles in a highly diverse population will have a lower probability of being IIS. To look at the effect of genetic diversity on neighborhood size estimates, I generated 30 data sets each with $\bar{f}$ = 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, or 0.9, $D_e = 1$, $\sigma = 1$, and expected neighborhood size of 12.56. Each data set contained 20 pairs of individuals with 10 independent markers for each of 20 distance classes (1–20). Figure 10 shows the distribution of the 30 neighborhood size estimates ($\hat{N_b}$) for $\bar{f}$ = 0.1-0.8; for clarity, estimates for $\bar{f} = 0.9$ are not shown because some estimates were extremely large. When $\bar{f}$ is small the neighborhood size estimates are clustered close together near the true neighborhood size (gray horizontal line), but as $\bar{f}$ increases the estimates become more variable. Table 7 shows that the relative MSE and the average width of the credible intervals increase as $\bar{f}$ increases.

Performance on Data from Lattice Simulation

To examine how well the model performs on data that is not a direct result of the model, I analyzed data sets from a spatially-explicit lattice simulation. I collected samples from populations simulated with $\sigma = 1$, $D_e = 1$, and $\mu = 0.0001$. Independent samples of 20 pairs of individuals at 20 distance classes (1–20) were collected from 100 different populations with either 10, 20, or 30 independent marker loci. Figure 11 follows the same format as Figure 8. The performance on the simulated data is similar to the performance on the data generated from the model, with high coverage percentage and low relative MSE that decreases when more markers are used; however, the bias is negative when 10 or 20 markers are used, but becomes positive for 30 markers.

Using the same parameters, I ran a second set of simulations with a different sampling scheme. Here I collected 100 individuals from each population and analyzed the samples using all possible pairwise combinations the individuals. The performance of the estimates for these data sets had low error, but the percentage of credible intervals that covered the true value was greatly reduced (Fig. 12). Therefore, all subsequent simulation results are based on data from independent pairs.

Figures 13 and 14 show an example of the posterior predictive fit for a single data set from the non-pairwise and the pairwise sampling schemes, respectively for each of the 10 marker loci. The gray dots represent the proportion of IIS pairs for each distance class estimated from the data. The blue lines represent the mean (horizontal curve) and the 95% credible interval (vertical lines) of the distribution of hypothetical values for the data that would be likely given the estimated posterior distribution. Model fit can be assessed by comparing how likely the observed data would be under the posterior predictive distribution. In both cases, the model produces replicate data sets that fit the observed data fairly well. However, the credible intervals for the replicated data are extremely narrow for the pairwise data so many of real data points fall far outside of the predicted credible interval. For the independent pairs, 89.5% of the data points fall within the credible intervals compared to only 62.7% for the pairwise data.

Effect of the Dispersal Kernel

The WM formula that is implemented in the model assumes that dispersal is isotropic and follows a normal distribution along each axis; however, at equilibrium, the pattern of isolation-by-distance is similar for a wide range of dispersal distributions (Fig. 4). Using the spatially-explicit lattice simulation, I generated thirty data sets using either the Rayleigh, exponential, half-normal, triangular, gamma, Pareto, or Lomax dispersal kernels. The parameters for the dispersal kernel were set so that the average squared parent-offspring distance (s^2) would be approximately 1 for an expected neighborhood size of 12.56. Each data set contained 20 pairs of individuals with 20 independent markers for each of 20 distance classes (1–20). Figure 15 summarizes the distribution of the 30 neighborhood size estimates (blue dots) for each dispersal kernel. Table 8 provides the dispersal parameters that were set for each distribution (σ and α), the observed mean-squared, parent-offspring dispersal distance (s^2), the average of the neighborhood size estimates, the relative MSE, the bias, and the average width of the credible intervals. Most of the distributions result in estimates that are close to the expected neighborhood size with low relative MSE and similar CI widths. The neighborhood size estimates for the Lomax distributions are lower on average with a higher relative MSE.

Effect of Different Mutation Rates

In the model, the mutation rate is treated as a constant parameter that must be provided. Assuming that an accurate estimate of the mutation rate is not always available, I wanted to determine how much an inaccurate mutation parameter will affect the neighborhood size estimate. To test this, I simulated 30 data sets from the lattice simulation with different mutation rates ($\mu = 10^{-2}$, 10^{-3}, 10^{-4}, 10^{-5}, and 10^{-6}). The dispersal parameter was set to 1 for an expected neighborhood size of 12.56. Each data set contained 20 pairs of individuals with 20 independent markers for each of 20 distance classes (1–20). For each data set, the MCMC algorithm was provided a different mutation rate estimate ($\hat{\mu} = 10^{-2}$, 10^{-3}, 10^{-4}, 10^{-5}, and 10^{-6}). Figure 16 shows the distribution of neighborhood size estimates for each combination of simulated and provided mutation rates. Table 9 shows the average observed heterozygosity ($\bar{H}_o$), the relative MSE, the bias, and the average width of the credible intervals. In most cases, the median of the neighborhood size estimates is closer to the true value when the mutation parameters agree. However, even when the mutation rate provided to the model is orders of magnitude higher or lower, the estimates are still close to the true value.

Better estimates are obtained when the true mutation rate is high (10^{-2}, 10^{-3}, or 10^{-4} vs. 10^{-5}, or 10^{-6}). When the mutation rate is low, the observed heterozygosity is also low ($\bar{H}_o$ = 0.131 when $\mu = 10^{-5}$) and there is less genetic diversity. These results complement the findings for generated data with high $\bar{f}$.

Effect of Different Mutation Models

The WM formula implemented here assumes an infinite alleles mutation model, but I wanted to determine how well the model performs using data simulated under different mutation models. Figure 17 shows the distribution of neighborhood size estimates for 30 simulations using either the infinite alleles model (IAM), the K-alleles model (KAM), or the step-wise mutation model (SMM) and four different mutation rates, $\mu = 10^{-2}$, 10^{-3}, 10^{-4}, and 10^{-5}. The expected neighborhood size is 12.56 and each data set contained 20 pairs of individuals with 30 independent markers for each of 20 distance classes (1–20). The neighborhood size estimates are close to the expected value for each mutation model, and there does not seem to any clear bias in the estimates for the different mutation models. Table 10 shows the average number of alleles and the average frequency of heterozygotes observed for each group of simulations. For every mutation rate except for 10^{-5}, the populations under the IAM maintained more alleles and had a slightly higher frequency of heterozygotes. Here, similar to results shown in Section 3 for $\mu = 10^{-5}$, neighborhood size estimates become more variable and had higher error when heterozygosity is low. When $\mu = 10^{-5}$, the average frequency of heterozygotes is higher under the SMM model, and the estimates are clustered more tightly around the true value.

Effect of the Sampling Scale

Next, I tested whether the sampling scale affects neighborhood size estimates. I generated 30 data sets from the model with σ = 0.5, 1.0, 2.0, or 10.0, and the density was set so that the expected neighborhood size was 12.56 in each case. The data sets contained 20 pairs of individuals with 10 independent markers for each of 20 distance classes from either 1–20 (constant), or 1σ–20σ (scaled). Figure 18 shows the distribution of neighborhood size estimates for each sampling scheme and for each dispersal parameter. The estimates for the constant and the scaled sampling schemes are similar when $\sigma =$ 0.5, 1.0, and 2.0, but the scaled sampling scheme improves the estimates when $\sigma = 10$.

Effect of Different Sampling Schemes

To test different sampling schemes, I simulated 30 data sets with $\sigma = 1$ for an expected neighborhood size of 12.56. Each data set contained 20 pairs of individuals with 20 independent markers for each of 40 distance classes (1–40). Neighborhood size estimates were made using different subsets of the 40 distance classes. The following sampling schemes were used: (1) all 40 distance classes, (2) the first 20 distance classes, (3) the first 10 distance classes, (4) each distance class from 11–20, (5) each distance class from 6–15, (6) every other distance class from 1–19, (7) every third distance class from 1–28, and (8) every fourth distance class from 1–37. Figure 19 shows the distribution of neighborhood size estimates for each sampling scheme. Using all 40 of the distance classes does not seem to improve neighborhood size estimates more than using the first 20 distance classes, but reducing the sampling scheme to only the first 10 distance classes results in more variable estimates. Sampling schemes that do not sample short distance classes (schemes 4 and 5) are more variable and tend to underestimate neighborhood size. Sampling 10 distance classes over a larger range (schemes 6, 7, and 8) provides better estimates than sampling the first 10 distance classes. Table 11 shows the average of the neighborhood size estimates, the relative MSE, the bias, and the average width of the credible intervals for each sampling scheme. The average of the neighborhood size estimates is close to the expected value for all schemes except for 4 and 5. Sampling the first 20 distance classes has the lowest relative MSE. When only 10 distance classes are sampled, the average width of the credible intervals is larger but it decreases when distance classes are sampled over a larger range.

Estimating Dispersal Parameter with Different Density Priors

In the model, a prior is assigned to the neighborhood size parameter and the density parameter, and the dispersal parameter is calculated deterministically from these values. If density can be estimated separately, information from the estimate, including the level of certainty, can be be included in the prior to get a better estimate for the dispersal parameter. To demonstrate this, I generated 30 data sets from the model with $\sigma = 0.5, 1.0, 2.0$, or 10.0, and the density parameter was set so that the expected neighborhood size is 12.56 in each case ($D_e = 12.56/4\pi\sigma^2$). The data sets contained 20 pairs of individuals with 10 independent markers for each of 20 distance classes (1σ–20σ). For each data set, the prior distribution for the density parameter had a mean equal to the true value and a precision of either $\tau = 0.0001$, 1, or 100. Figure 20 shows the distribution of relative squared error for the dispersal parameter estimates. The error decreases when the precision, τ, of the density prior is larger. Table 12 shows the dispersal, σ, and the density, D_e, parameters that were used to generate the data, the density prior precision, τ, the average for the neighborhood size estimates, $\bar{N}_b$, the average for the dispersal parameter estimates, $\bar{\sigma}$, the relative MSE, the bias, and the average width of the credible interval for the dispersal estimates. The average of the neighborhood size estimates is close to the expected value for each data set. The average of the dispersal estimates are closer to the expected value, the relative MSE is lower, and the average width of the credible intervals is smaller when τ is larger.

Performance on Pinus pinaster Aiton

The neighborhood size estimates reported in De-Lucas et al. (2009a) were 37.86 for the Fuentelapeña population and 51.03 for the Quatretonda population. Using the data sets that contained all possible pairwise comparisons, my method estimated the Fuentelapeña neighborhood size as 48.36 with 95% CI [27.97,77.07], and the Quatretonda neighborhood size as 21.27 with 95% CI [5.58,48.44]. The estimated credible interval for the Fuentelapeña population included the original point estimate from De-Lucas et al. (2009a). The neighborhood size estimate for the Quatretonda population was much lower than the original estimate, and the original estimate falls outside of the estimated credible interval.

Figure 21 shows the mean of the neighborhood size posterior (gray dots), and the 95% credible intervals for the 10 data sets where pairs were chosen without replacement for each population. The estimates for the Fuentelapeña population were similar to the previously published value (black dashed line). The estimates for the Quatretonda population were not as close to the reported value and were more variable. Table 13 shows the average width of the credible intervals, the average DIC values for the model (DIC_{H_a}) and the null hypothesis (DIC_{H_o}), the difference between the two DIC values ($\Delta\text{DIC} = \text{DIC}_{H_o} - \text{DIC}_{H_a}$), and the relative difference $\delta\text{DIC} = \Delta\text{DIC}/\text{DIC}_{H_o}$ for both populations. The Fuentelapeña estimates have tighter credible intervals and a higher relative δDIC compared to the Quatretonda population.

Figure 22 shows the posterior predictive fit for the pairwise data and Fig. 23 shows the posterior predictive fit for the second data set shown in Fig. 21 for each population for each of the microsatellite loci. While the total number of genotyped samples are the same, the credible intervals for the replicated data in Fig. 22 are much narrower than the credible intervals in Fig. 23. In Fig. 22, 69.0% of the data points fall within the 95% credible intervals compared to 76.2% in Fig. 23.

Discussion

When a clear pattern of isolation-by-distance exists, this method is able to make accurate estimates of neighborhood size using data generated directly from the model. The estimates have high coverage around 95% and low error. The slightly negative bias indicates that the model is more likely to underestimate the neighborhood size, but the bias is reduced when more markers are analyzed. The method performed similarly well when independent pairs of individuals were collected from the lattice-based simulation. The bias of the estimates from the simulated data showed a slightly different pattern than the generated data, but this is likely due to the placement of the expected value. The expected value was based on the parameters that were provided to the simulation, but due to the discrete nature of the lattice, it is unlikely that the simulation will be a perfect reflection of the parameters.

The performance of the method suffers when non-independent pairs are sampled from the lattice simulation. In these simulations, fewer samples were collected, but the data was analyzed using all possible pairwise comparisons of the sampled individuals. Because this sampling scheme reuses the data from individuals many times, the total number of pairs that are counted per distance class is much larger. The model incorrectly interprets the counts as if they were from a large sample of independent pairs when they are actually based on highly dependent pairs from a small sample. As a result of this discrepancy, the model tends to estimate artificially narrow credible intervals which have as low as 47% coverage rather than the expected 95%. The error and the bias indicate that the estimates are close to the true value, and they are comparable to the results from the independent data sets; however, due to the lower coverage, the true value is less likely to be contained within the credible interval. The pairwise sampling scheme used in these simulations is typically applied when estimating neighborhood size using the regression methods (Rousset, 2000; Vekemans and Hardy, 2004), but it should be avoided in favor of a more independent sampling scheme when using this method.

The method performed well on data that was generated with different dispersal parameters when σ was less than 1.5. The relative MSE for $\sigma = 0.25$ was slightly elevated but this may be because the sampling scale was not fine enough for such low dispersal. When σ was 1.5 or greater, the estimates became unreliable with higher relative error and larger credible intervals. At the higher dispersal values, the observed pattern of isolation-by-distance in the data was very weak, and it was not detected by the model. When this occurs, the model makes extremely large and variable estimates of neighborhood size. A relative δDIC value less than 0.01 appears to indicate the point where isolation-by-distance is no longer detected.

When the pattern of isolation-by-distance is strong, estimates are accurate with fewer samples and fewer markers. On the other hand, when the pattern of isolation-by-distance is weak, accurate estimates can be made if more data is available. In some cases, increasing the number of markers had diminishing returns, but larger sample sizes generally improved the estimates.

The method performs better when genetic diversity is high in the sample for both the generated and simulated data. For the generated data, I adjusted the $\bar{f}$ parameter which determines the average probability of identity in the sample. The $\bar{f}$ parameter has an inverse relationship with genetic diversity, so when $\bar{f}$ was high, the neighborhood size estimates were less accurate and more variable. For the simulated data, a higher mutation rate increased the genetic diversity, and the method provided better estimates when the mutation rate was high. When genetic diversity is low, it is difficult to detect isolation-by-distance but perhaps better estimates can be made if more data is available. Generally this should not be a concern because only highly polymorphic markers would be considered in such a study. Leblois et al. (2003) report that Rousset's regression method (Rousset, 2000) is also sensitive to genetic diversity, but they point out that the heterozygosity at microsatellite loci is usually within the range of 0.5-0.8; which is where both methods perform well. Markers that are less polymorphic (e.g. single nucleotide polymorphisms or allozymes) could be used but likely many more markers or samples would need to be analyzed to produce good estimates.

The mutation rate for each marker is treated as a constant parameter that must be provided to the model. I found that when the true mutation rate is high ($\mu = 10^{-2}$, 10^{-3}, and 10^{-4}), the provided mutation rate can differ by several orders and the model will still provide decent estimates of neighborhood size but values that are closer to the true value perform better. This suggests that the provided mutation rate can be a very rough approximation of the assumed mutation rate.

The model is robust to data generated using different mutation models. I compared estimates for data simulated under the infinite alleles model, the K-alleles model, and the step-wise mutation model and found that the neighborhood size estimates were close to the true value in each case. The number of possible alleles was restricted to 20 for both the KAM and SMM models, and when the mutation rate was high, it resulted in a large difference in the average number of alleles in the population compared to the IAM, but the average number of heterozygotes was approximately the same for each model. Leblois et al. (2003), using a more stringent $K = 10$, showed similar results using the regression method from Rousset (2000). They suggest that the pattern of local differentiation corresponds to events that have occurred in the recent past, and therefore it is less dependent on the mutation process.

The WM approximation assumes that dispersal follows a normal distribution along each axis, but as shown in Figure 4, the pattern of isolation-by-distance is similar for a wide range of dispersal distributions. I found that the neighborhood size estimates were close to the expected value for many of the dispersal distributions. Similar to the likelihood-based method from Novembre and Slatkin (2009), the neighborhood size estimates using my method also show a negative bias when the dispersal distribution is highly leptokurtic (i.e. Lomax). This negative bias may be due to the steeper decrease in genetic identity at short distances that is observed for these distributions 4. Because the model assumes normal a normal distribution, this pattern may be interpreted as a stronger pattern of isolation-by-distance which would result in an smaller neighborhood size estimate.

Many studies have looked at how the sampling scale can influence the observed pattern of isolation-by-distance and ultimately neighborhood size estimates (Epperson and Li, 1997; Leblois et al., 2003; Vekemans and Hardy, 2004). Rousset (2000) recommends sampling most individuals within an area of $10\sigma \times 10\sigma$ but Leblois et al. (2003) suggest that such a sampling scale could become prohibitive in practice. They tested deviations from the recommended scheme and found that the MSE remained low, but the estimates did become negatively biased if the scale was too small and positively biased if the scale was too large. I tested the impact of the sampling scale in my model and I found that scaling the distance classes with σ did not have a large impact on the neighborhood size estimate unless σ is very large in which case it is better to sample distance classes $1\sigma - 20\sigma$. I also tested different sampling patterns, and I found estimates are more accurate when the nearest neighbors are included in the sample. This is expected because isolation-by-distance is often only detected at the shortest spatial scales (Epperson and Li, 1997; Vekemans and Hardy, 2004). Leblois et al. (2003) recommend that distances do not exceed $10\sigma - 50\sigma$ because at larger scales the mutation rate and the mutation model can no longer be neglected especially when the mutation rate is high. Most of the data sets tested here included data for 20 distance class but estimates using 10 distance classes showed low MSE. The method presented here does not technically require that the data be grouped into distance classes, but the computational efficiency of the model would suffer if each pair is modeled separately.

Neighborhood size contains information about effective population density and the dispersal ability of a species. Many biologist are particularly interested in information about dispersal because it can be difficult to measure directly. Unfortunately, none of the indirect methods described here allow these parameters to be estimated independently of neighborhood size. However, if density can be measured separately, this information can be used to calculate a point estimates of the dispersal parameter as $\sigma^2 = N_b/4\pi D_e$. One of the advantages of my method is that the level of certainty in the density estimate can be included in the model and reflected in the credible intervals for the dispersal estimate. I demonstrated that different levels of certainty in the density prior led tighter credible intervals and accurate estimates for the density parameter.

Finally, I tested the method on microsatellite data from samples from two populations of *Pinus pinaster* Aiton. De-Lucas et al. (2009a) reported neighborhood sizes estimates for these samples using SPAGeDi which uses data sets based on all possible pairwise comparisons between samples. I estimated neighborhood size using a data set that contained all pairwise data as well as data sets where I attempted to limit the number of times that each sample was reused. My estimates for the Fuentelapeña samples were fairly consistent and close to the value estimated in the original paper but for the Quatratonda population, my estimates were more variable and much lower than the value estimated in the paper. The variability in the estimates may be due to the lower relative δ DIC value for the Quatratonda which indicates that isolation-by-distance is weaker. The posterior predictive fit is good for both the pairwise and non-pairwise data sets, but the credible intervals are more narrow for the pairwise data. As discussed previously, this is likely due to the reuse of data in the pairwise sampling scheme.

The estimate provided by the SPAGeDi package is only a point estimate of neighborhood size. Leblois et al. (2003) introduced a method to compute 95% confidence intervals for estimates based on the slope of Rousset (2000)'s a_r statistic. However, they reported that their ABC bootstrap procedure produced inaccurate intervals with lower than expected coverage. An accurate estimate of uncertainty is important when making inferences. My Bayesian approach naturally provides a probabilistic statement about the certainty of the estimated parameter in the form of the credible interval, and I have shown that when samples are independent, the estimated credible intervals have high coverage. Additionally, my method is able to provide a posterior predictive check which allows one to access model fit for each of the markers independently. This will help determine if the behavior of certain markers are driving the overall estimate.

MCMC methods can be cumbersome in certain situations but I did not experience any major issues with convergence and I was able to generate many independent samples from the posterior in a short amount of time. The average time to run the *Pinus pinaster* Aiton data was 209.9 ± 59 seconds for 30,000 iterations plus a 10,000 iteration burn-in.

Overall, this method performs well and is robust to certain violations of the model assumptions. This method is recommended for estimating neighborhood size from genetic data sets collected from continuous populations at drift-mutation-dispersal equilibrium. The credible intervals for the estimates behave best when pairs of individuals are independent. The model assumes that spatial genetic structure observed in the data is a consequence of isolation-by-distance and estimates will be more accurate when the pattern of isolation-by-distance is strong. When the pattern of isolation-by-distance is weak, including more samples and more markers should produce more accurate estimates of neighborhood size.

Table 4. The neighborhood size estimates are accurate when dispersal is local but they become less accurate and more variable as dispersal distance increases. Thirty data sets were generated from the model with $D_e = 1$, $\bar{f} = 0.39$, and $\sigma =$ 0.25, 0.5, 1, 1.5, 2, or 4. Twenty pairs of individuals with 10 independent markers were generated for each of 20 distance classes (1–20). The table shows the average of the neighborhood size estimates ($\bar{N}_b$), the relative mean squared error (MSE), the bias, the average width of the credible intervals, the DIC values for both the model and the null hypothesis of no isolation-by-distance, their difference (H_o DIC $-$ H_a DIC), and their relative difference (δDIC $=$ ΔDIC$/H_o$ DIC).

σ	0.25	0.5	1	1.5	2	4
D_e	1	1	1	1	1	1
N_b	0.785	3.142	12.566	28.274	50.265	201.062
$\bar{N}_b$	0.763	2.769	11.123	543.587	37334.619	43974.325
MSE	0.263	0.053	0.047	4753.022	8323973.557	327725.582
Bias	-0.022	-0.373	-1.444	515.313	37284.354	43773.263
CI Width	1.651	2.708	11.068	4816.192	331098.758	387265.153
H_o DIC	1876.341	1546.122	1212.816	1160.790	1144.325	1151.911
H_a DIC	1137.717	1138.801	1145.810	1147.562	1140.767	1151.296
ΔDIC	738.624	407.321	67.006	13.229	3.558	0.615
δDIC	0.394	0.263	0.055	0.011	0.003	0.001

Table 5. Accurate estimates can be made with fewer samples and fewer markers when isolation-by-distance is strong. Thirty data sets were generated from the model with $D_e = 1$, $\sigma = 0.5$, and $\bar{f} = 0.39$ for an expected neighborhood size of 3.14. Each data set contained either 1, 2, 5, or 10 pairs of individuals with either 5, 10, or 15 independent loci for each of 20 distance classes (1–20). The table shows the relative mean square error, the bias, and the average width of the credible intervals.

			Number of Pairs Per Distance Class			
			1	2	5	10
Number of Loci	5	MSE	0.526	0.530	0.284	0.289
		Bias	-1.252	-0.769	-1.300	-1.300
		CI Width	14.653	11.852	4.897	3.916
	10	MSE	0.311	0.321	0.144	0.142
		Bias	-1.509	-1.404	-0.854	-0.766
		CI Width	6.779	5.476	4.319	3.208
	15	MSE	0.314	0.292	0.124	0.083
		Bias	-0.767	-1.357	-0.786	-0.573
		CI Width	6.790	4.724	3.715	2.863

Table 6. When isolation-by-distance is weak, more samples and more markers can improve estimates. Thirty data sets were generated from the model with $D_e = 1$, $\sigma = 2$, and $\bar{f} = 0.39$ for an expected neighborhood size of 50.27. Each data set contained either 20, 30, 40, or 50 pairs of individuals with either 10, 20, or 30 independent markers for each of 20 distance classes (1–20). The table shows the relative mean square error, the bias, and the average width of the credible intervals.

			Number of Pairs Per Distance Class			
			20	30	40	50
Number of Loci	10	MSE	7055.300	287.408	15131.937	9.696
		Bias	1066.537	336.470	1490.264	42.134
		CI Width	8077.658	2874.703	11450.843	726.678
	20	MSE	29.559	6.487	0.370	0.195
		Bias	122.509	15.579	-0.994	-1.165
		CI Width	1440.785	376.768	157.077	124.136
	30	MSE	4.053	0.935	0.668	0.059
		Bias	37.527	10.392	8.882	-0.768
		CI Width	165.385	67.050	59.472	4.876

Table 7. Neighborhood size estimates are more accurate and precise when genetic diversity is high (low $\bar{f}$). Thirty data sets were generated from the model with $D_e = 1$, $\sigma = 1$, and $\bar{f} =$ 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, or 0.9 for an expected neighborhood size of 12.56. Each data set contained 20 pairs of individuals with 10 independent markers for each of 20 distance classes (1–20). The table shows the relative mean square error, the bias, and the average width of the credible intervals.

	Average Probability of Identity ($\bar{f}$)							
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8
MSE	0.006	0.021	0.030	0.036	0.102	0.127	0.599	274.151
Bias	-0.546	-1.312	-1.081	-1.085	-2.574	-2.615	-1.493	63.851
CI Width	4.105	6.716	9.208	11.826	13.289	19.041	42.104	739.914

Table 8. Neighborhood size estimates are close to the expected value for most dispersal distributions. Thirty data sets were simulated with different dispersal distributions. The parameters of the dispersal distributions were set so that the average squared parent-offspring distance (s^2) would be approximately 1 for an expected neighborhood size of 12.56. Each data set contained 20 pairs of individuals with 20 independent markers for each of 20 distance classes (1–20). The table shows the dispersal distribution, the dispersal parameters (s^2), the average of 30 neighborhood size estimates, the relative mean squared error, the bias, and the average width of the credible intervals.

Distribution	σ	α	s^2	$\bar{N}_b$	MSE	Bias	CI Width
Rayleigh	1	–	1.088	13.518	0.036	0.952	9.248
Exponential	1	–	1.065	11.766	0.060	-0.800	8.647
Half-Normal	1	–	1.066	12.404	0.048	-0.162	8.476
Triangular	1	–	1.120	13.264	0.026	0.698	9.686
Gamma	1	1.5	1.064	11.859	0.078	-0.708	8.960
Gamma	1	2	1.065	11.797	0.036	-0.769	8.072
Gamma	1	4	1.098	12.206	0.035	-0.360	8.475
Gamma	1	8	1.082	13.775	0.035	1.209	9.753
Pareto	0.98	2.4	1.076	10.219	0.071	-2.348	6.768
Pareto	0.962	2.6	1.095	11.836	0.030	-0.730	6.799
Pareto	0.917	2.8	1.077	10.775	0.037	-1.791	6.395
Pareto	0.944	3	1.112	11.515	0.034	-1.052	7.270
Lomax	1.25	2.4	0.971	7.990	0.188	-4.577	7.764
Lomax	1.108	2.6	1.138	7.707	0.229	-4.859	7.275
Lomax	1.058	2.8	1.066	8.479	0.187	-4.087	8.068
Lomax	1.05	3	1.144	8.344	0.207	-4.222	8.694

Table 9. The mutation rate provided to the model has a small impact on the neighborhood size estimate. Thirty data sets were simulated with different mutation rates $\mu = 10^{-2}$, 10^{-3}, 10^{-4}, and 10^{-5}, $\sigma = 1$, and an expected neighborhood size of 12.56. Each data set contained 20 pairs of individuals with 20 independent markers for each of 20 distance classes (1–20). For each data set, the MCMC algorithm was run with a different mutation rate parameter $\mu = 10^{-2}$, 10^{-3}, 10^{-4}, 10^{-5}, or 10^{-6}. The table shows the average observed heterozygosity ($\bar{H}_o$), the relative mean square error, the bias, and the average width of the credible intervals.

				MCMC Mutation Rate				
		$\bar{H}_o$		10^{-2}	10^{-3}	10^{-4}	10^{-5}	10^{-6}
Simulation Mutation Rate	10^{-2}	0.851	MSE	0.012	0.045	0.022	0.016	0.025
			Bias	0.699	2.117	0.974	-0.182	-1.226
			CI Width	2.485	3.550	3.612	3.624	3.772
	10^{-3}	0.779	MSE	0.032	0.015	0.031	0.034	0.031
			Bias	-1.993	0.915	1.775	1.846	1.567
			CI Width	2.511	3.502	4.626	5.403	5.783
	10^{-4}	0.537	MSE	0.066	0.021	0.028	0.054	0.062
			Bias	-2.837	-0.274	0.488	0.216	-0.001
			CI Width	5.064	6.675	8.521	9.064	9.569
	10^{-5}	0.131	MSE	0.528	0.675	0.189	15.654	29.264
			Bias	-3.448	-2.090	-3.033	4.359	8.949
			CI Width	38.095	44.762	30.501	107.715	143.118

Table 10. When the mutation rate is high, more alleles are maintained under the infinite alleles model and the heterozygote frequency is high. Samples were collected from populations simulated under the IAM, KAM, or SMM and $\mu = 10^{-2}$, 10^{-3}, 10^{-4}, or 10^{-5}. The table shows the average number of alleles ($\bar{k}$), and the average observed frequency of heterozygotes ($\bar{H}_o$).

		Model	$\bar{k}$	$\bar{H}_o$
Mutation Rate	10^{-2}	IAM	525	0.851
		KAM	20	0.807
		SMM	19	0.785
	10^{-3}	IAM	78	0.779
		KAM	20	0.741
		SMM	10	0.674
	10^{-4}	IAM	12	0.528
		KAM	9	0.500
		SMM	5	0.441
	10^{-5}	IAM	2	0.131
		KAM	2	0.150
		SMM	2	0.159

Table 11. Neighborhood size estimates are less precise when fewer distance classes are sampled and estimates are not accurate when small distances are not sampled (4 and 5). Thirty populations were simulated with $\sigma = 1$ for an expected neighborhood size of 12.56. Each data set contained 20 pairs of individuals with 30 independent markers for each of 40 distance classes (1–40). The table shows the average neighborhood size estimates, the relative mean squared error, the bias, and the average credible interval width for different subsets of the 40 distance classes. The following sampling schemes were used: (1) all 40 distance classes, (2) the first 20 distance classes, (3) the first 10 distance classes, (4) each distance class from 11–20, (5) each distance class from 6–15, (6) every other distance class from 1–19, (7) every third distance class from 1–28, and (8) every fourth distance class from 1–37.

	Sampling Scheme							
	1	2	3	4	5	6	7	8
$\bar{N}_b$	13.6	13.010	12.472	5106.555	342.229	13.501	13.351	13.669
MSE	0.024	0.019	0.066	2078145.007	16856.458	0.053	0.039	0.046
Bias	1.099	0.444	-0.095	5093.989	329.663	0.935	0.785	1.102
CI Width	4.640	5.915	14.938	49373.045	2306.524	9.840	8.475	7.239

Table 12. Estimates of the dispersal parameter are closer to the true value when the density prior has high precision (τ). Thirty data sets were generated from the model with different combinations of parameters for dispersal (σ) and density (D_e) that resulted in an expected neighborhood size of 12.56. Each data set contained 20 pairs of individuals with 10 independent markers for each of 20 distance classes (1σ–20σ). For each data set, the prior distribution for the density parameter had a mean equal to the true value and a precision of $\tau = 0.001$, 1, or 100. The table shows the average of the neighborhood size estimates, the average of the dispersal estimates ($\bar{s}$), the mean squared error, the bias, and the average width of the credible intervals.

σ	D_e	τ	$\bar{N}_b$	$\bar{\sigma}$	MSE	Bias	CI Width
0.5	4	0.001	11.108	0.534	0.412	0.034	1.478
1	1	0.001	11.710	1.055	0.313	0.055	2.988
2	0.25	0.001	11.012	1.663	0.348	-0.337	4.715
10	0.01	0.001	11.167	11.441	0.698	1.441	28.661
0.5	4	1	11.927	0.508	0.074	0.008	0.830
1	1	1	12.973	1.065	0.041	0.065	1.778
2	0.25	1	12.835	1.959	0.032	-0.041	3.275
10	0.01	1	12.258	10.484	0.092	0.484	17.244
0.5	4	100	12.722	0.501	0.006	0.001	0.174
1	1	100	13.321	1.026	0.004	0.026	0.366
2	0.25	100	12.639	1.963	0.036	-0.037	0.694
10	0.01	100	12.620	9.813	0.038	-0.187	3.456

Table 13. Estimates for the Fuentelapeña population have tighter credible intervals and the ΔDIC values are larger compared to the Quatretonda population. The table shows the average credible interval width, the DIC value for the model (DIC$_{H_a}$) and the DIC value for the null model of no isolation-by-distance (DIC$_{H_o}$), the difference between the two DIC values ($\Delta\text{DIC} = \text{DIC}_{H_o} - \text{DIC}_{H_a}$), and the relative difference ($\delta\text{DIC} = \Delta\text{DIC}/\text{DIC}_{H_o}$) for the independent samples from both populations.

Population	CI Width	DIC$_{H_a}$	DIC$_{H_o}$	ΔDIC	δDIC
Fuentelapeña	55.515	358.455	385.562	27.106	0.070
Quatretonda	78.915	342.822	361.520	18.698	0.052

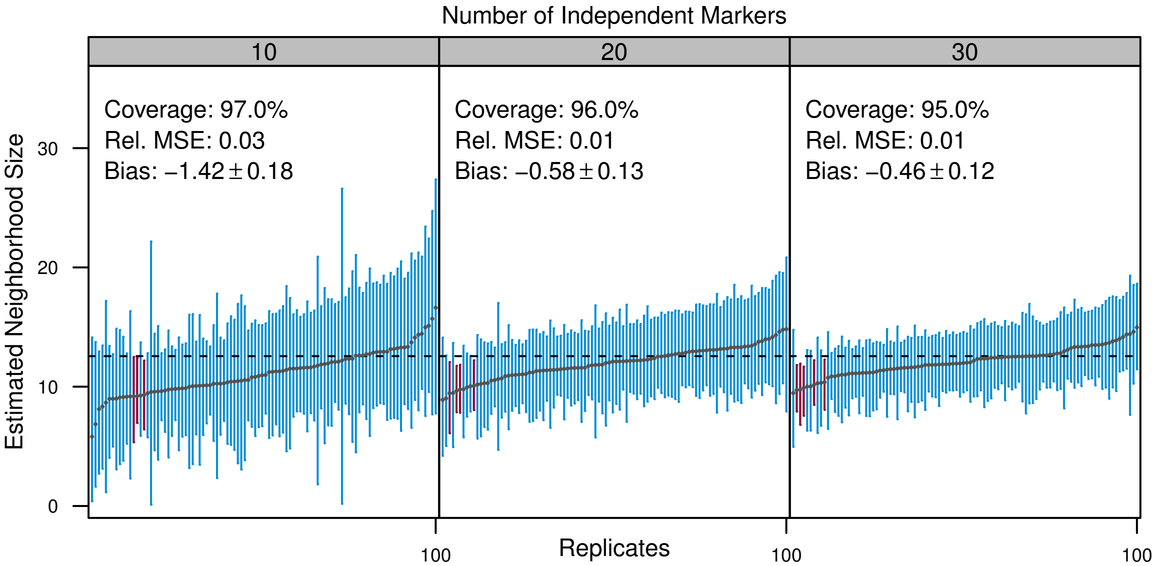


Figure 8. The method performs well on data generated directly from the model but there is a bias toward lower estimates. Each panel shows the neighborhood size estimates for 100 data sets randomly generated with $\sigma = 1$, $D_e = 1$, and $\bar{f} = 0.39$ for an expected neighborhood size of 12.56 indicated by the black dashed line. Twenty pairs of individuals were generated for each of 20 distance classes (1–20). Different panels represent data generated for either 10, 20, or 30 independent markers. The gray dots are the mean of the neighborhood size posterior distribution shown in increasing order. The vertical lines are the 95% credible intervals; the intervals that do not cover the true value are shown in red. Each panel indicates the percent of estimates that cover the true value, the relative mean squared error, and the bias.

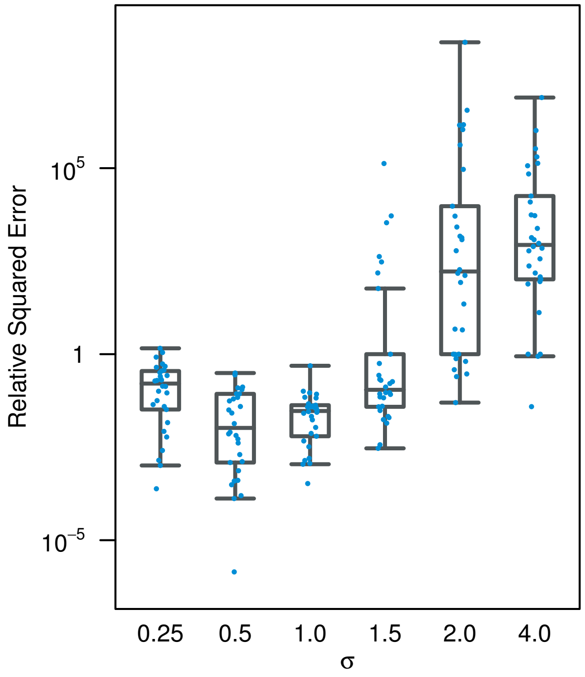


Figure 9. The relative error is low for small neighborhood sizes but estimates are more variable and less accurate after neighborhood size reaches 50.27 ($\sigma = 2$). Thirty data sets were generated from the model with $D_e = 1$, $\bar{f} = 0.39$, and $\sigma =$ 0.25, 0.5, 1, 1.5, 2, or 4. Twenty pairs of individuals with 10 independent loci were generated for each of 20 distance classes (1–20). The box-whisker plot summarizes the distribution of the relative squared error (log scale) of the neighborhood size estimates for each data set (blue dots), where the top and bottom of the boxes represent the 25% and 75% quartiles and the center bar represents the median.

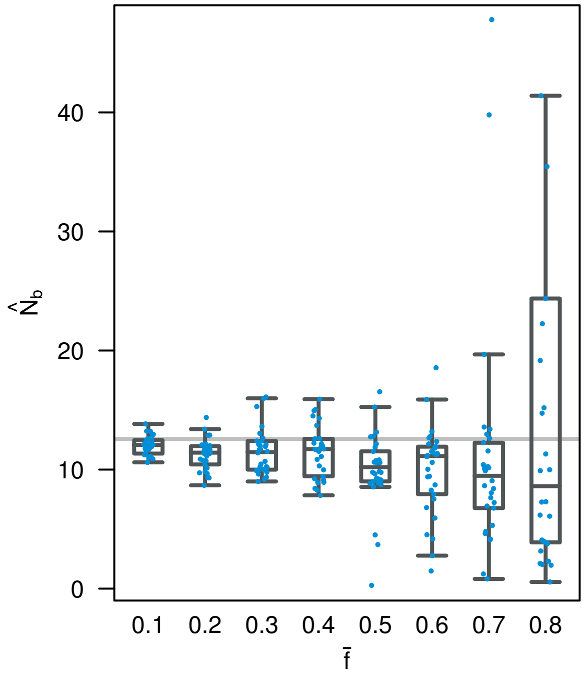


Figure 10. Neighborhood size estimates are more accurate and precise when genetic diversity is high (low $\bar{f}$). Thirty data sets were generated from the model with $D_e = 1$, $\sigma = 1$, and $\bar{f}$= 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, or 0.9 for an expected neighborhood size of 12.56 (gray horizontal line). Each data set contained 20 pairs of individuals with 10 independent markers for each of 20 distance classes (1–20). The box-whisker plots summarize the distribution of neighborhood size estimates (blue dots) for each data set. The features of the box-whisker plot are the same as in Figure 9.

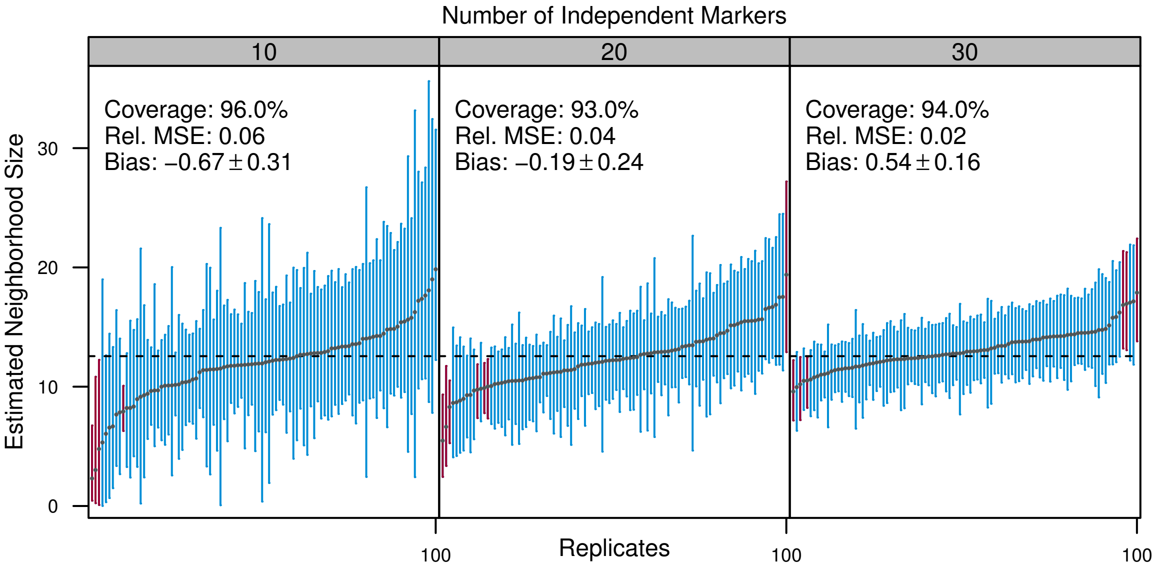


Figure 11. The method performs well on simulated data. Each panel shows the neighborhood size estimates for 100 data sets from the lattice simulation with $\sigma = 1$ and $D_e = 1$ for an expected neighborhood size of 12.56 indicated by the black dashed line. Each data set contained 20 pairs of individuals with either 10, 20, or 30 independent markers for each of 20 distance classes (1-20). The features of the plot are the same as described in Fig. 8.

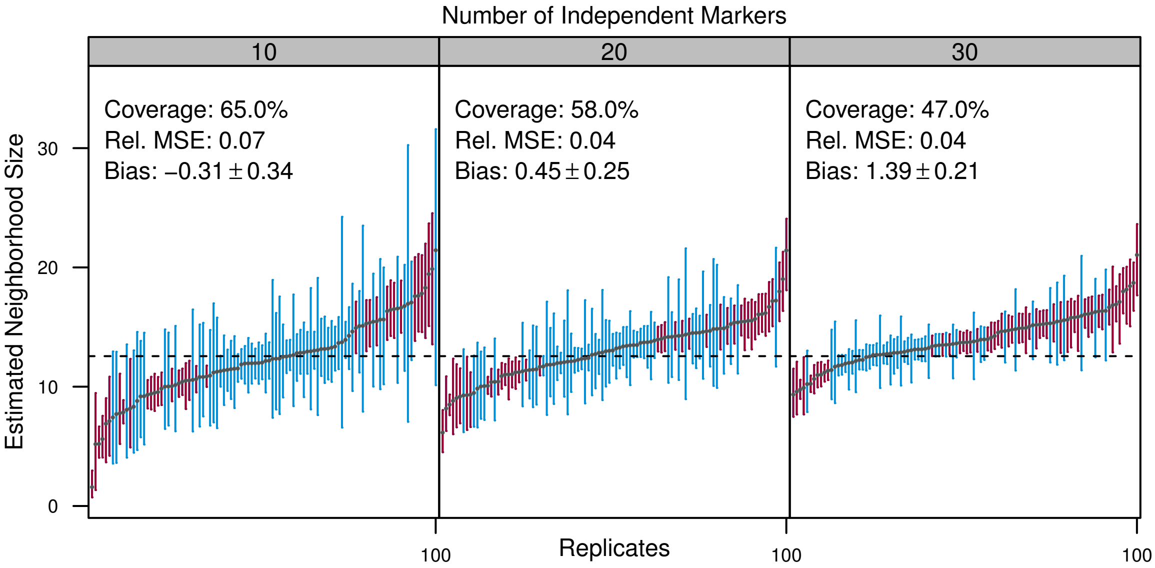


Figure 12. The coverage percentage suffers when pairwise data are used. Each panel shows the neighborhood size estimates for 100 pairwise data sets from the lattice simulation with $\sigma = 1$ and $D_e = 1$, for an expected neighborhood size of 12.56 indicated by the black dashed line. Each data set was a sample of 100 individuals from a $10\sigma \times 10\sigma$ grid and each individual had either 10, 20, or 30 independent markers. The features of the plot are the same as described in Fig. 8.

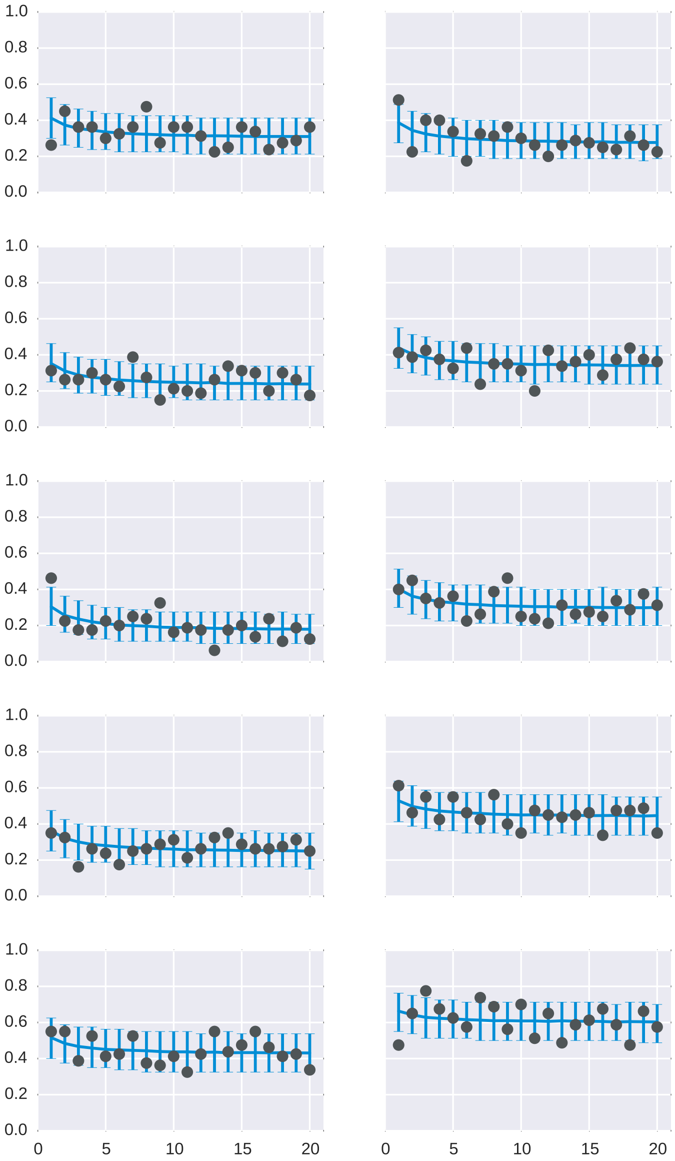


Figure 13. The posterior predictions fit the simulated data sets. The plots show the posterior predictive fit for one of the data sets from the lattice simulation for each of 10 marker loci. The gray dots represent the proportion of IIS pairs for each distance class (1–20). The blue lines represent the mean (horizontal curve) and the 95% credible intervals (vertical lines) for the distribution of hypothetical values that would be likely given the posterior distribution.

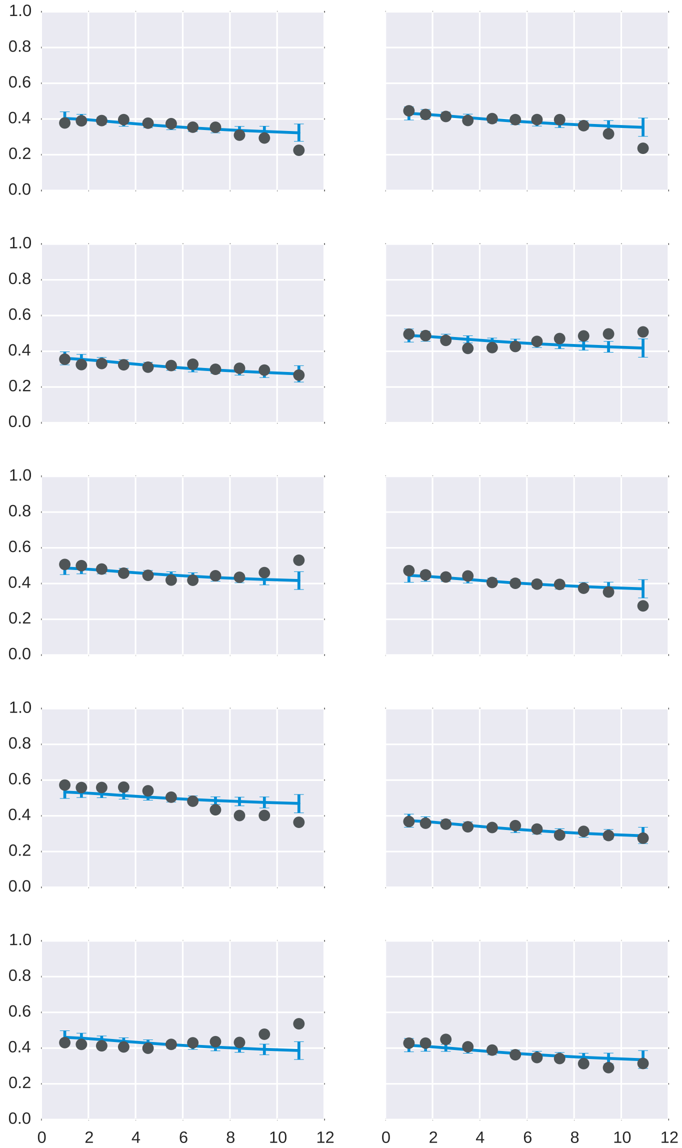


Figure 14. The posterior predictions fit the simulated pairwise data sets. The plots show the posterior predictive fit for one of the data sets from the lattice simulation for each of 10 marker loci. The gray dots represent the proportion of IIS pairs for each distance class. The blue lines represent the mean (horizontal curve) and the 95% credible intervals (vertical lines) for the distribution of hypothetical values that would be likely given the posterior distribution.

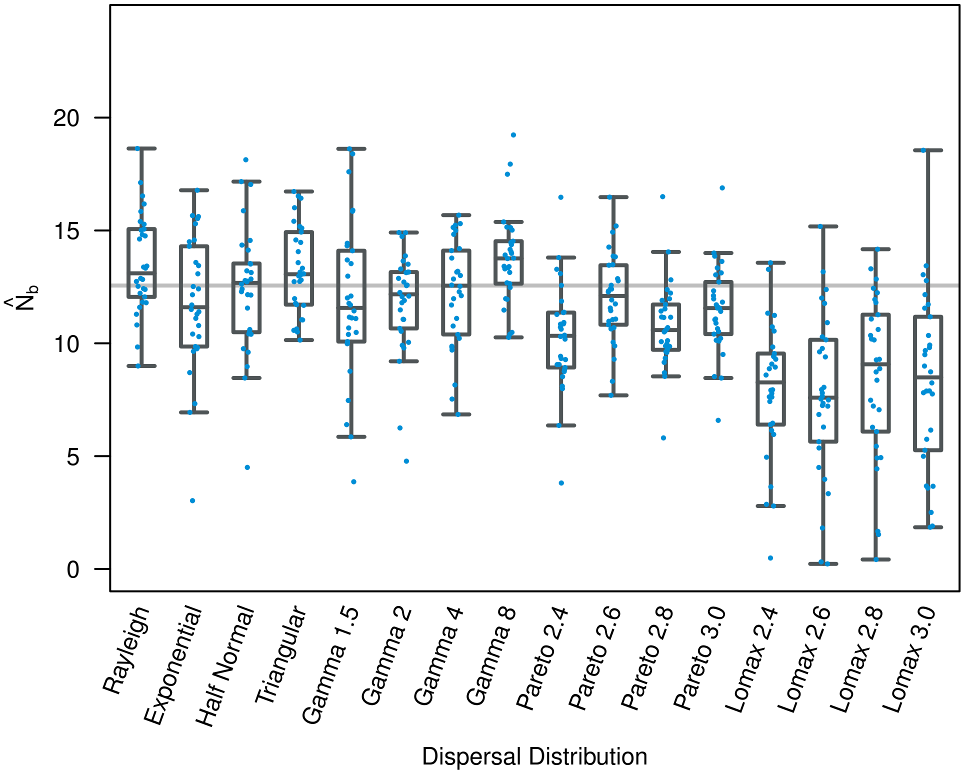


Figure 15. Neighborhood size estimates are close to the expected value for most dispersal distributions. Thirty data sets were simulated with different dispersal distributions. The parameters of the dispersal distributions were set so that the average squared parent-offspring distance (s^2) would be approximately 1 for an expected neighborhood size of 12.56 (gray horizontal line). Each data set contained 20 pairs of individuals with 20 independent markers for each of 20 distance classes (1–20). The box-whisker plots summarize the distribution of neighborhood size estimates (blue dots) for each data set. The features of the box-whisker plot are the same as in Figure 9.

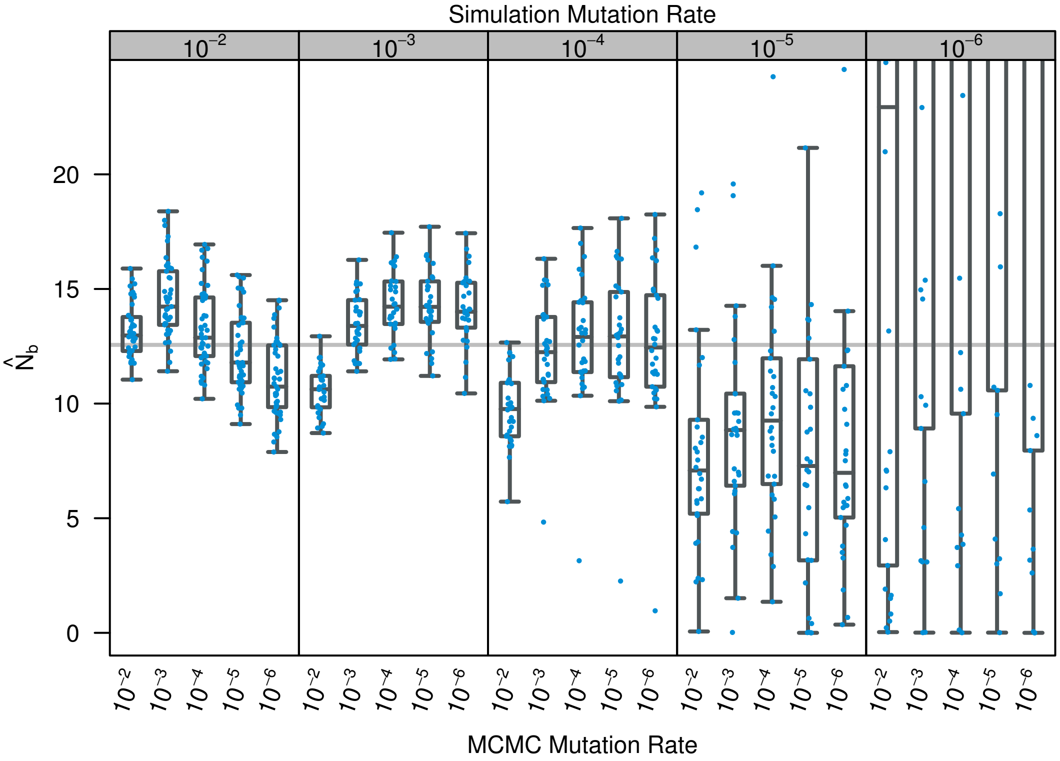


Figure 16. The mutation rate provided to the model has a small impact on the neighborhood size estimate. Thirty data sets were simulated with different mutation rates μ = 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, and 10^{-6}, $\sigma = 1$, and an expected neighborhood size of 12.56 (gray horizontal line). Each data set contained 20 pairs of individuals with 20 independent markers for each of 20 distance classes (1–20). For each data set, the MCMC algorithm was run with a different mutation rate parameter μ = 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, or 10^{-6}. The box-whisker plots summarize the distribution of neighborhood size estimates (blue dots) for each data set. The features of the box-whisker plot are the same as in Figure 9.

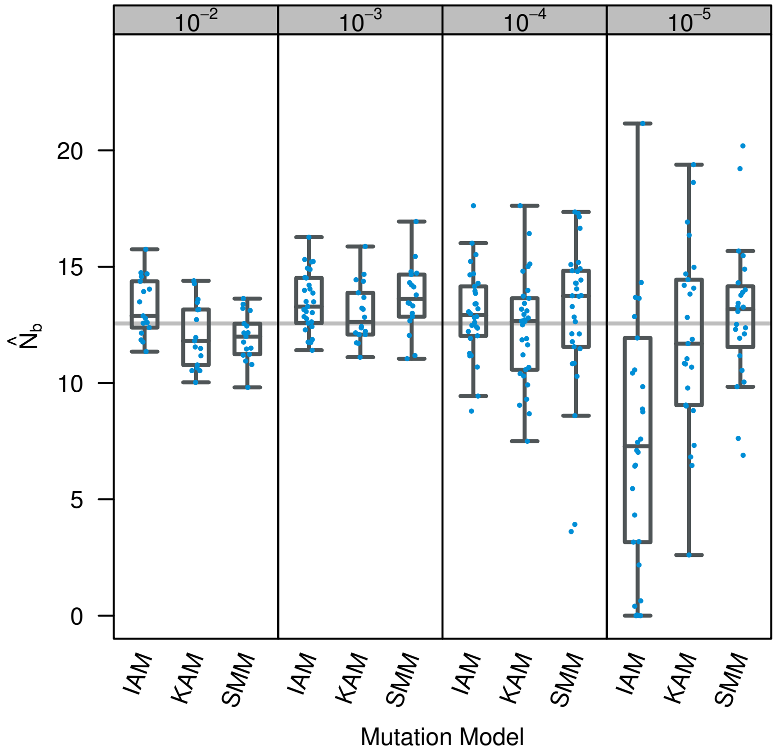


Figure 17. Neighborhood size estimates are similar for different mutation models. Thirty data sets were simulated with three different mutation models: infinite alleles model (IAM), K-alleles model (KAM), and step-wise mutation model (SMM); and four mutation rates, $\mu = 10^{-2}$, 10^{-3}, 10^{-4}, and 10^{-5}. The expected neighborhood size is 12.56 (gray horizontal line). Each data set contained 20 pairs of individuals with 30 independent markers for each of 20 distance classes (1–20). The box-whisker plots summarize the distribution of the neighborhood size estimates for each data set (blue dots).

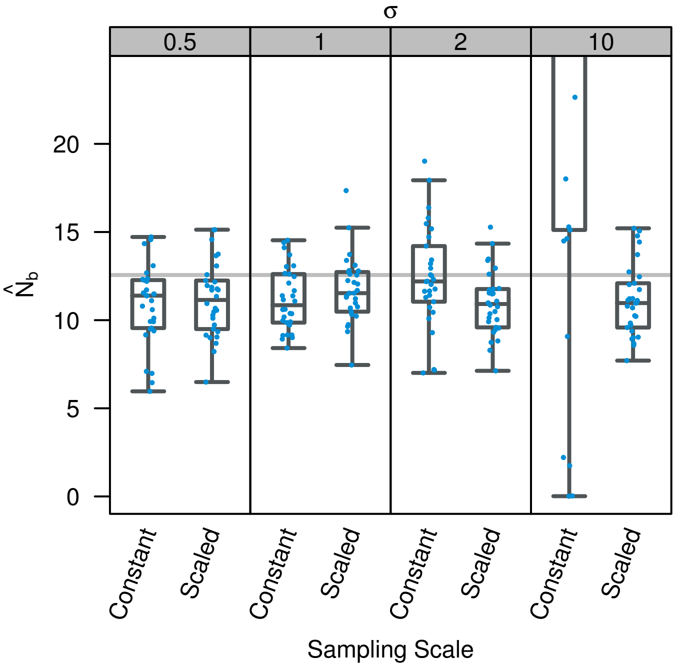


Figure 18. When dispersal distances are large, scaling the distance classes improves neighborhood size estimates. Thirty data sets were generated from the model with $\sigma = 0.5, 1.0, 2.0$, or 10.0 and density was set so that the expected neighborhood size would 12.56 in each case. Each data set contained 20 pairs of individuals with 10 independent markers for each of 20 distance classes from either 1–20 (constant) or 1σ–20σ (scaled). The box-whisker plots summarize the distribution of the neighborhood size estimates (blue dots).

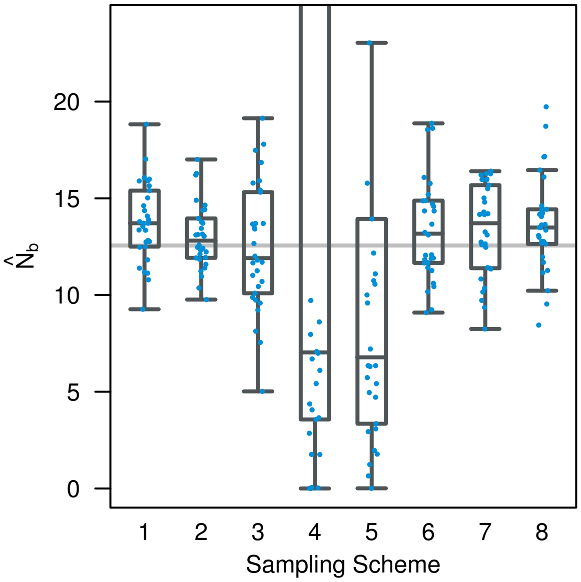


Figure 19. Neighborhood size estimates are less precise when fewer distance classes are sampled and estimates are not accurate when small distances are not sampled (4 and 5). Thirty populations were simulated with $\sigma = 1$ for an expected neighborhood size of 12.56 (gray horizontal line). Each data set contained 20 pairs of individuals with 30 independent markers for each of 40 distance classes (1–40). Neighborhood size estimates were made using different subsets of the 40 distance classes. The box-whisker plots summarize the distribution of neighborhood size estimates (blue dots). The following sampling schemes were used: (1) all 40 distance classes, (2) the first 20 distance classes, (3) the first 10 distance classes, (4) each distance class from 11–20, (5) each distance class from 6–15, (6) every other distance class from 1–19, (7) every third distance class from 1–28, and (8) every fourth distance class from 1–37.

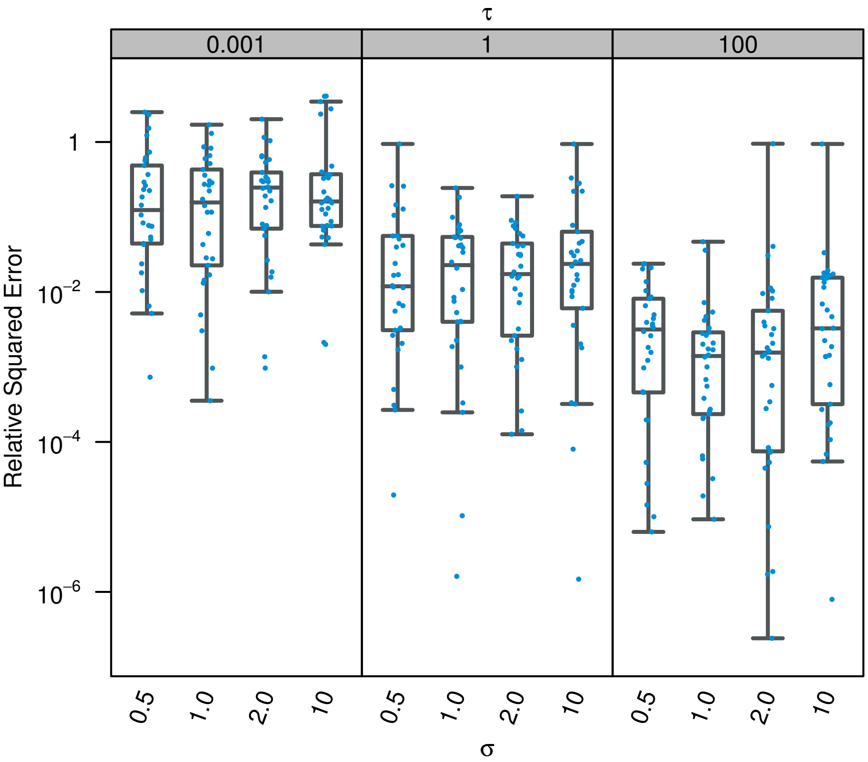


Figure 20. Estimates of the dispersal parameter are closer to the true value when the density prior has high precision (τ). Thirty data sets were generated from the model with σ = 0.5, 1.0, 2.0, or 10.0 and density was set so that the expected neighborhood size was 12.56 in each case. Each data set contained 20 pairs of individuals with 10 independent markers for each of 20 distance classes (1σ–20σ). For each data set, the prior distribution for the density parameter had a mean equal to the true value and a precision of $\tau = 0.001$, 1, or 100. The box-whisker plots summarize the distribution of the relative squared error of the dispersal parameter estimates (blue dots).

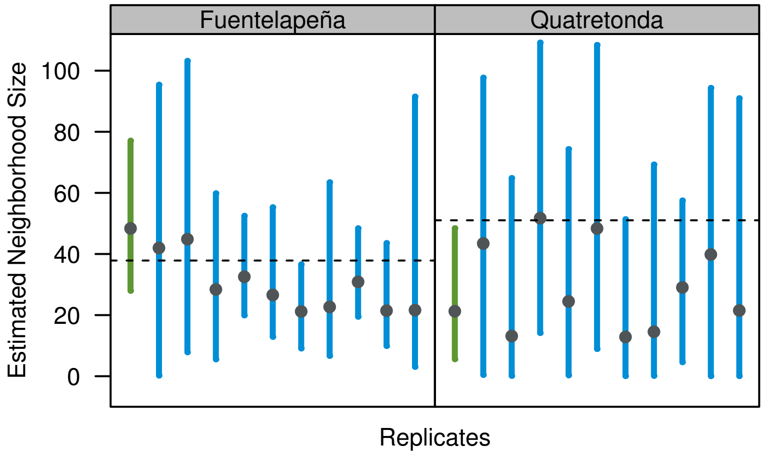


Figure 21. Estimates for the Fuentelapeña population were similar to the previously published value but the estimates for the Quatretonda population were more variable. The plot shows the mean of the neighborhood size posterior (gray dots) and the 95% credible interval (vertical lines). The blue lines indicate estimates for 10 different samples where pairs of individuals were randomly drawn from the data set independently, without replacement. The green lines indicate estimates using all pairwise comparisons of the samples. The black dashed line represents the point estimates from De-Lucas et al. (2009a).

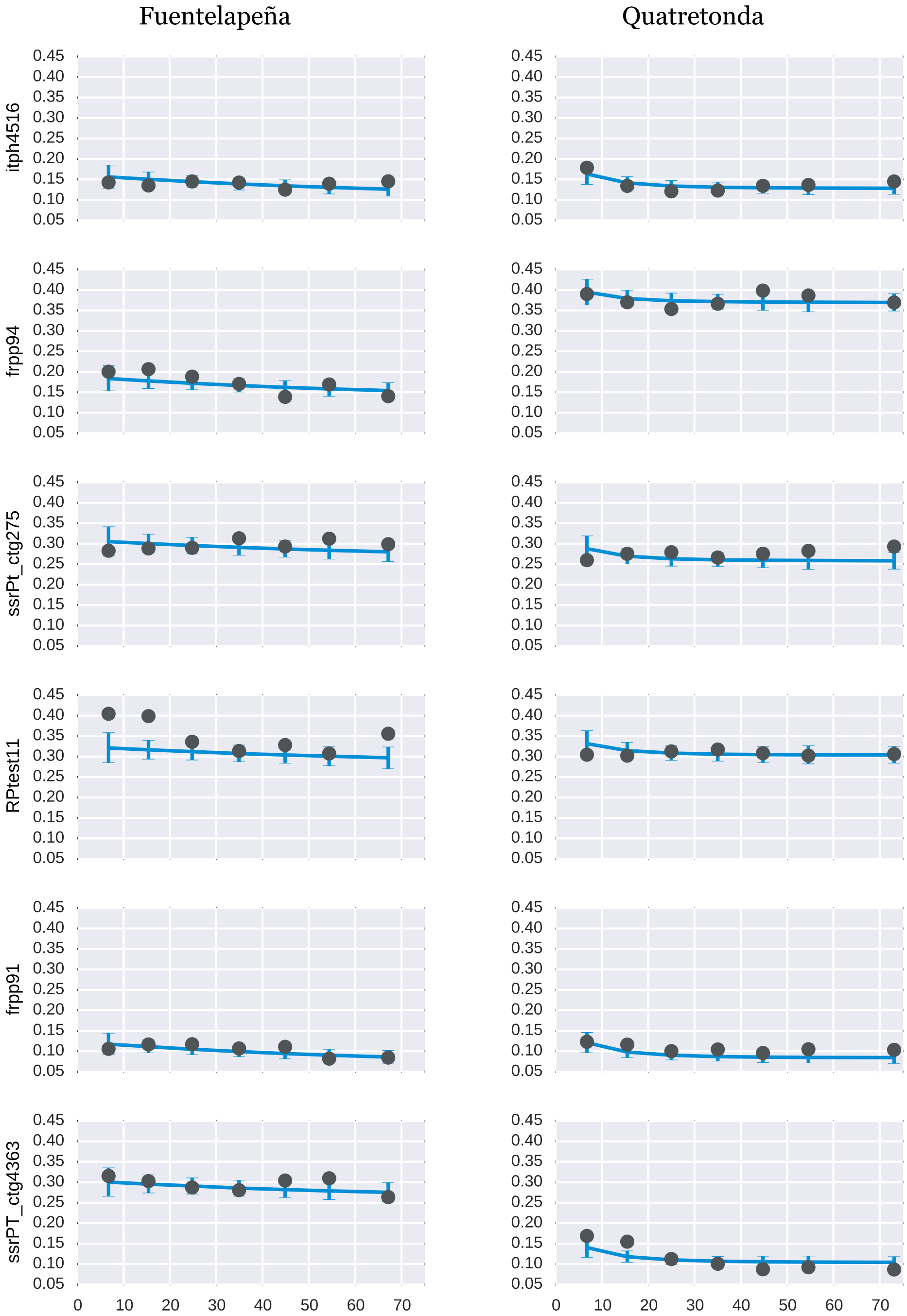


Figure 22. The posterior predictions fit the observed pairwise data. The plots show the posterior predictive fit for pairwise data from both populations. The gray dots represent the proportion of IIS pairs for each distance class. The blue lines represent the mean (horizontal curve) and the 95% credible intervals (vertical lines) for the distribution of hypothetical values that would be likely given the posterior distribution. The rows correspond to the six different microsatellite loci.

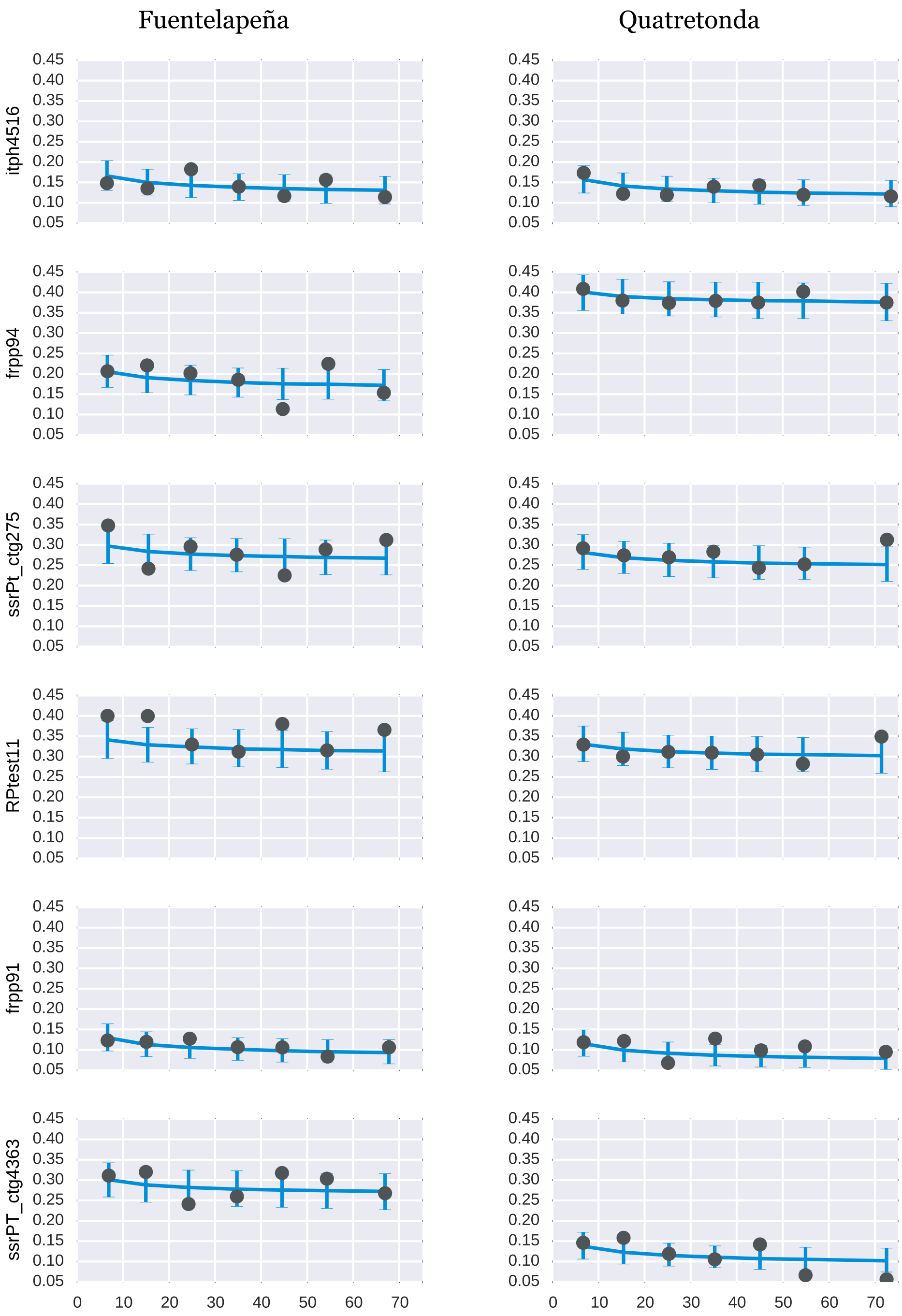


Figure 23. The posterior predictions fit the observed data. The plots show the posterior predictive fit for the first independent data set shown in Figure 21 for each population. The gray dots represent the proportion of IIS pairs for each distance class estimated from the data. The blue lines represent the mean (horizontal curve) and the 95% credible intervals (vertical lines) for the distribution of hypothetical values that would be likely given the posterior distribution. The rows correspond to the six different microsatellite loci.

Chapter 4

THE ROLE OF SELF-INCOMPATIBILITY SYSTEMS IN THE PREVENTION OF BIPARENTAL INBREEDING

Abstract

Hermaphroditic plants can experience inbreeding in two ways: self-fertilization, when they mate with themselves, or bi-parental inbreeding, when they mate with close relatives. Under isolation-by-distance when pollen and seed dispersal are limited, plants experience greater bi-parental inbreeding. Many plant species have evolved physical and genetic self-incompatibility (SI) systems which limit self-fertilization, but only the genetic SI systems can also limit bi-parental inbreeding. Genetic SI species are prevalent across the angiosperms and it is often assumed that the additional reduction in bi-parental inbreeding may be a factor in their success. To test this assumption, I developed a spatially-explicit, individual-based simulation of plant populations with either physical SI or one of three different types of genetic SI, and compared the amount of inbreeding in the populations. I found that the amount of inbreeding in the genetic SI populations was significantly lower than the physical SI populations and this reduction is due to bi-parental inbreeding avoidance. However, compared to the overall reduction in inbreeding this was relatively small. Genetic SI populations also suffered reduced female fecundity and had smaller census population sizes. Overall, I found little evidence that the success of genetic SI systems is due to bi-parental inbreeding avoidance because the effect is small compared to the reduction in self-fertilization and there would need to be a strong selective advantage to outweigh the cost of reduced female fecundity.

Introduction

Due to the sessile nature of angiosperms, offspring dispersal occurs only through the movement of pollen and seed, and in many plant species, pollen and seed dispersal distances rarely exceed a few meters from the parent (Fenster, 1991; Levin, 1981). Plants are therefore more likely to become established near their parent's location where there is a higher concentration of related individuals. Under these conditions, populations become spatially structured due to isolation-by-distance. If pollen dispersal is also limited, these related individuals are more likely to interbreed. This type of inbreeding is referred to as bi-parental inbreeding to distinguish it from the more extreme inbreeding that occurs when hermaphroditic plants self-fertilize. In many plant species, crosses between close neighbors have been shown to produce offspring that are less fit than average, and because the reduction in fitness is associated with spatial proximity, this is likely evidence of inbreeding depression resulting from isolation-by-distance (Heywood, 1991).

Both bi-parental inbreeding and self-fertilization (selfing) can increase homozygosity within a genome. Offspring that are produced through inbreeding are more likely to express recessive deleterious alleles and suffer reduced viability and fecundity (Charlesworth and Charlesworth, 1987; Charlesworth et al., 1990). Self-fertilizing species have more opportunities to purge highly deleterious alleles, but they tend to maintain a large number of slightly deleterious alleles (Charlesworth et al., 1990; Wang et al., 1999). Out-crossing species tend to maintain recessive deleterious alleles in a heterozygous state which can lead to inbreeding depression. However, when bi-parental inbreeding is common, some of the segregating deleterious alleles can be purged in outcrossing populations (Heywood, 1991).

Presumably as a result of inbreeding depression, plants have evolved a variety of inbreeding avoidance mechanisms, many of which are specifically directed at reducing self-fertilization. In heteromorphic self-incompatibility (SI) systems, each plant expresses one of the two or more genetically determined flower morphologies and pollen from one morph can only fertilize flowers of a different morph. Homomorphic SI systems are also genetically determined but they result in a large number of different molecular phenotypes that do not affect flower morphology. Although both of these SI systems have a genetic basis, I will refer to heteromorphic SI as physical SI because the mechanism is based on physical morphology, and I will refer to homomorphic SI as genetic SI.

Genetic SI systems are typically controlled by two tightly linked genes at an S locus, which determine the molecular phenotype of the pollen and allow the female receptors to recognize and reject pollen with the same phenotype. Genetic SI systems can be split into two categories: gametophytic

SI and sporophytic SI. In gametophytic SI systems, the phenotype of the pollen is determined by the pollen's S locus haplotype; whereas, in sporophytic SI systems, the phenotype is determined by the diploid genotype of the pollen donor. In both systems, the female receptor discriminates against any pollen with a matching S phenotype, regardless of whether it originated from the same plant.

Physical SI strategies reduce self-fertilization, but they do not provide a mechanism to prevent bi-parental inbreeding. Genetic SI systems, on the other hand, prevent self-fertilization and prevent crosses with individuals that share similar S alleles. Because similar S alleles are often shared among related individuals, genetic SI systems also provide a mechanism to limit bi-parental inbreeding.

Genetic SI systems are prevalent across angiosperm families (Igic et al., 2008), and it is commonly assumed that the evolutionary success of these systems is tied to their ability to reduce bi-parental inbreeding in addition to preventing self-fertilization (Charlesworth and Charlesworth, 1987). However, because the genetic consequences of bi-parental inbreeding and self-fertilization are similar, it is difficult to distinguish between the two types of inbreeding (Griffin and Eckert, 2003). As a result, there are no studies that directly estimate how much genetic SI systems reduce bi-parental inbreeding compared to self-fertilization.

There is evidence that bi-parental inbreeding is reduced in regions of the genome that are linked to the S locus. The forced heterozygosity at the S locus extends to other linked loci and can reduce the expression of recessive deleterious alleles at those loci. Deleterious alleles can accumulate in this region because they are sheltered from selection (Llaurens et al., 2009). It remains unclear, however, whether genetic SI systems reduce bi-parental inbreeding at loci that are not linked to the S locus.

Cartwright (2009) presented results from a simulation study which compared the amount of inbreeding in populations with physical or genetic SI systems. He confirmed that there was a large decrease in bi-parental inbreeding in genetic SI simulations near the S locus compared to physical SI systems, but at unlinked loci, the reduction in bi-parental inbreeding was comparatively small. This suggests that at unlinked loci, genetic SI systems only have a small impact on the amount of bi-parental inbreeding; however, Cartwright points out that he did not model inbreeding depression which would have introduced a selective advantage to avoid inbreeding.

In this study, I test whether bi-parental inbreeding avoidance is a driving force behind the evolution of genetic self-incompatibility systems in angiosperms. I develop a spatially-explicit, individual-based simulation to model continuous populations of self-incompatible plants. To differentiate between the two types of inbreeding, I compare the amount of inbreeding observed in genetic SI populations to the amount of inbreeding in physical SI populations where individuals are prevented from selfing but are free to mate with related individuals. Any reduction in inbreeding in the genetic SI populations compared the physical SI populations will be a result of a reduction in bi-parental inbreeding. I predict that the reduction in bi-parental inbreeding in the genetic SI populations will be more dramatic when isolation-by-distance is strong.

In the simulation, I model three different genetic SI systems which vary in the way they discriminate against pollen with matching S alleles. The first system is modeled after the gametophytic SI system (GSI). The GSI system is the least stringent system because half of the pollen produced by a plant can successfully fertilize a plant that has one S allele in common. The second system is modeled after the sporophytic system that is common in the Brassicaceae (BSI). In the BSI system, dominance relationships exist between the S alleles, and the pollen phenotype will reflect the dominant S allele. In this case, pollen can successfully fertilize a plant that shares the recessive allele but it cannot fertilize a plant that shares the dominant allele. Finally, I modeled a codominant version of sporophytic SI (SSI), where both of the S alleles are equally expressed in the pollen phenotype. There is no known biological equivalent of this SI system, and a situation where all S alleles are equally codominant is highly unlikely. Nevertheless, the SSI system serves to model an extreme case of discrimination where pollen is prevented from fertilizing any plant that shares either S allele. I predict that more stringent SI systems will show a greater reduction in bi-parental inbreeding.

One consequence of genetic self-incompatibility is that female fecundity can suffer when pollen from compatible mates is limited (Larson and Barrett, 2000). In genetic SI populations, the S locus is under negative, frequency-dependent selection, and pollen with a rare S phenotype will be favored. For this reason, a large number of S alleles needs to be maintained in the population for mating to be successful. When isolation-by-distance is strong, the pollen pool is reduced and individuals may struggle to find a mate. There is evidence that suggests that effective dispersal at the S locus increases in SI populations (Cartwright, 2009; Leducq et al., 2011) but mate limitation will still likely lead to reduced seed set.

Methods

I developed a spatially-explicit, individual-based simulation to model discrete generations of a self-incompatible plant population. In the simulation, populations inhabit a toroidal lattice where each cell is occupied by a single, hermaphroditic individual. The plants are diploid and have several independently assorting genetic loci.

In natural SI systems, the S locus usually contains two genes that control the pollen and the receptor phenotypes. Due to repressed recombination, these genes are tightly linked and inherited together (Casselman et al., 2000; Castric et al., 2010; Charlesworth and Awadalla, 1998; Kamau et al., 2007; Kawabe et al., 2006; Vieira et al., 2003). Therefore, in the simulation, the S locus is treated as a single gene with one allele. Mutations occur at the S locus at rate $\mu_s = 10^{-5}$, and each mutation results in a completely new S haplotype according to the infinite alleles mutation model.

The M-locus is a marker locus where all alleles are selectively neutral. The M-locus mutates at rate $\mu_m = 10^{-4}$ under the infinite alleles model. This marker is used to measure the amount of inbreeding in the population. In the initial population, each S and M allele is unique so the simulation must run for a burn-in period before it reaches a drift-mutation equilibrium.

Individuals also carry a total of 10 independent D-loci that are not linked to each other and are not linked to the S or M loci. Each locus may carry two possible alleles: a wild-type allele and a recessive deleterious allele. Individuals start out with all wild-type alleles that may mutate into the deleterious allele. The D locus mutates into the deleterious allele at rate $\mu_d = 0.1$, and cannot mutate back to the wild-type state. Each homozygous recessive genotype at a D locus increases the probability that an individual will become sterile by 0.005. Affected individuals are viable but are unable to produce pollen or seed. The purpose of the D locus is to model inbreeding depression by simulating the segregation of slightly deleterious alleles in the population. Usually, the probability of a deleterious mutation at a single locus is rare, but the probability of a deleterious mutation is high when the whole genome is considered. To maintain a large enough penalty for inbreeding, I used a high mutation rate at each D locus so that, on average, there would be one new deleterious mutation per haplotype.

At the beginning of each generation, fertile parent plants produce gametes — 10 pollen grains and 5 ovules — through independent assortment of loci. Pollen grains are dispersed from the parent's location according to a normal distribution along each axis with standard deviation σ. The pollen is then checked for compatibility with the plant in the new location based on the rules of the assigned SI system. If compatible, it is randomly assigned to an ovule, otherwise it is discarded. When pollen dispersal is complete, some ovules will be remain unfertilized while some ovules will have a pool of pollen from which they will randomly choose one. Unfertilized ovules will be aborted and fertilized ovules will form seeds. Seeds are then dispersed from the parent's location in the same way as the pollen. When seed dispersal is complete, a single seed from each cell will be randomly selected to become a parent in the next generation. Random mutations occur in the germ line of these parents before they produce gametes so that all of their offspring will carry the mutation.

Mating Systems

Pollen and receptor compatibility is determined by four different SI systems. The first is a physical self-incompatibility system (PSI) where individuals are obligate out-crossers but no genetic mating system is in place to prevent bi-parental inbreeding. This synthetic PSI system is 100% efficient at preventing self-fertilization. In the gametophytic self-incompatibility (GSI) system, the S phenotype of the pollen is determined by the pollen's haplotype. Pollen is compatible if its S allele does not match either S allele in the pollen recipient. In the co-dominant sporophytic self-incompatibility (SSI) system, the S phenotype of the pollen is determined by the diploid genotype of the pollen donor. All S alleles are codominant, and pollen is compatible with any plant that does not share either of its parent's S alleles. Dominant sporophytic self-incompatibility (BSI), is similar to SSI; however, dominance relationships exist between S alleles. The S alleles in the population are randomly assigned into a dominance hierarchy and pollen is compatible with any plant that does not share its dominant S allele. For comparison, some simulations were run with no self-incompatibility system (NSI) where plants were allowed to self-fertilize. Self-fertilization occurred when self pollen did not disperse outside the parent cell.

Simulation and Analysis

Simulations were run for each of four different landscape sizes (50×50, 100×100, 200×200, and 400×400) and for each of the four different SI systems (PSI, GSI, BSI, and SSI). The pollen and seed dispersal parameters were both set to either $\sigma = 1$, 2, 4, or 6. For the 50×50 and 100×100 populations, a random sample of 500 individuals was collected from the population every N generations after a 10,000 generation burn-in period, where N is the size of the population. I collected a total of 500 replicate samples from these nearly independent populations. Simulations using the 200×200 and 400×400 require a much larger investment in computing time. To reduce the computation time, I collected samples from the population every 10,000 generations after an N generation burn-in period. Relative to the size of the population, the number of generations between the samples is small so they may not be independent; however, to reduce some of the correlation, I combined data from 5 different simulation runs for a total of 500 equilibrium samples.

To measure inbreeding in each sample, I calculated the average probability that an individual carried two alleles at the M locus that were identical-by-descent. The alleles were considered to be identical-by-descent if they both descended from the same allele in a grandparent, regardless of mutation. From each sample, I also recorded the average homozygosity, and the average number of alleles at the S and M locus, and the average squared parent-offspring dispersal distance (s^2). For pollen and seed dispersal, total s^2 is expected to be $\sigma^2 = \sigma_s^2 + \sigma_p^2/2$, where σ_s represents seed movement and σ_p represents pollen movement (Crawford, 1984). In this formula, seed dispersal contributes more than pollen dispersal because seeds carry gametes from both parents whereas pollen only carries gametes from the father. From the whole population, I recorded the total number of adults, the seed set, and the number of sterile individuals.

To analyze the results, I used the Anderson-Darling two-sample test (Scholz and Stephens, 1987) implemented in the kSamples R package (R Core Team, 2015; Scholz and Zhu, 2016). The test statistic was T.AD = $(AD - (k-1))/\sigma$, and the P-value estimation method was set to simulate the default

10,000 random rank permutations. The distribution of values for each measurement was compared between the different simulations under the null hypothesis that the values came from the same underlying distribution. The P-values from the pairwise comparisons were adjusted for multiple tests using the Holm correction (Holm, 1979) and the significance criterion was set at 0.05 for all tests.

Results

Effect of Inbreeding Depression

In the simulations, inbreeding individuals that are homozygous at certain loci were penalized with an increased probability of becoming sterile. To determine the impact of this imposed inbreeding depression, I compared the amount of inbreeding in simulations with and without the deleterious effect. The probability of grandparental identity-by-descent at the M locus was used as a measure of recent inbreeding in the population. Figure 24 shows the pairwise comparisons for the results from simulations on a 100×100 landscape with the pollen and seed dispersal parameter $\sigma = 1$ for the NSI, PSI, GSI, BSI, and SSI systems. The dark blue squares represent comparisons that are not significantly different and the values along the right side of the figure represent the average probability of grandparental identity-by-descent across all population samples. The different mating systems are arranged from left to right in descending order based on the average probability. For the genetic SI systems, there is not a significant difference between the simulations with inbreeding depression (1) and without inbreeding depression (0).

To compare the stationary level of inbreeding in the populations, I also measured the average frequency of homozygotes (Fig. 24). Here, the physical and genetic SI systems all have a significantly lower frequency of homozygotes when there is a penalty for inbreeding compared to simulations where there is no penalty. The simulations with no SI system (NSI) had the highest frequency of homozygotes and there was not a significant difference between simulations with or without inbreeding depression.

Probability of Grandparental Identity-By-Descent

Figure 26A shows the results of pairwise comparisons for the 100×100 simulations with $\sigma = 1$ for the NSI, PSI, GSI, BSI, and SSI systems. The dark blue squares represent comparisons that are not significantly different and the values along the right side of the figure represent the average probability of grandparental identity-by-descent across all population samples. The different mating systems are arranged from left to right in descending order based on the average. The probability of identity-by-descent is significantly different between all SI systems except for GSI and BSI, and it decreases as the SI systems become more stringent. Figure 26B shows the empirical distribution of the probabilities. The distributions for the genetic SI systems are all significantly different than the PSI distribution; however, the difference is small when compared to the NSI distribution.

To ensure that this pattern is consistent for different population sizes, I compared results from 50×50, 100×100, 200×200, and 400×400 simulations. Figure 27 shows a plot similar to Fig. 26A that includes comparisons for each of these simulations. Because the simulations are arranged based on the average probability, the different SI mating systems group together such that the PSI simulations are at the high end and SSI simulations are at the low end. The GSI and BSI simulations group together in the middle with none of them showing a significant difference in inbreeding. The overall pattern is consistent for the different population sizes.

Isolation-by-Distance

I predicted that the greatest reduction in bi-parental inbreeding would occur when isolation-by-distance is strong. To test this, I compared the amount of inbreeding in 100×100 simulations with different seed and pollen dispersal abilities. Figure 28 shows the pairwise comparisons between these simulations. In this case, the simulations group together based on the dispersal parameter with $\sigma = 1$ (strong isolation-by-distance) at the high end and $\sigma = 6$ (weak isolation-by-distance) at the low end. When $\sigma = 1$, 2, or 4, we see the same pattern as before. PSI has the highest inbreeding, and it is significantly different from the other SI systems, BSI and GSI have lower inbreeding and are not significantly different from each other, and finally, SSI has the lowest inbreeding and is significantly different from the other SI systems. However, when $\sigma = 6$, SSI is not significantly different from GSI. The reduction in the average probability of identity-by-descent in the genetic SI systems compared to PSI is largest when isolation-by-distance is strong.

Homozygosity and Allele Diversity

Because M alleles mutate according to the infinite alleles model, any two alleles that are identical must have shared a common ancestor at some time in the past. Homozyotes are therefore a result of inbreeding that potentially occurred further back than the grandparent generation. Figure 29 shows the results of pairwise comparisons of the frequency of homozygotes at the M locus for each mating system and for each level of dispersal. The frequency is highest when dispersal is limited ($\sigma = 1$) and significantly different from the other dispersal levels. At each dispersal level, the different SI systems are not significantly different from each other.

I analyzed allele diversity at both the S locus and the M locus. Figure 30 shows the comparisons of allele counts at the M locus. When dispersal is low, the average number of alleles maintained in the population is higher and significantly different from most other simulations. For each dispersal level, the different SI types are not significantly different. Figure 31 shows the comparisons of allele counts for the S locus. Here, the simulations group together by SI type with SSI simulations maintaining the highest number of S alleles followed by GSI, BSI, and PSI. The simulations with $\sigma = 1$ maintain a significantly higher number of alleles in each SI system.

Population Demographics

In the 100×100 simulations, a maximum of 10,000 individuals can exist in the population, however, it is possible that some lattice cells will remain unfilled. Table 14 shows the average census number of individuals for each simulation. The PSI simulations have the largest average population sizes and the population size is not largely affected by dispersal. The genetic SI simulations have a higher average number of individuals when the dispersal parameter is larger. When $\sigma = 1$, the genetic SI simulations all have significantly reduced population sizes with the greatest reduction seen in the SSI simulations (Fig. A32).

Each individual carries a number of deleterious D loci and for each homozygous recessive genotype they carry, they will have an increased probability of becoming sterile. The average number of sterile individuals in the population across all simulations was 484.5 which is around 5% of the population. Table 14 shows averages for each simulation but none of the simulations were significantly different (Fig. A33).

In the simulation, each plant produces 5 ovules and a maximum of 50,000 seeds can be produced per generation. To achieve maximum seed set, there needs to be enough compatible pollen to fertilize all of the ovules. Table 14 shows the average seed set for each simulation. Seed set is lowest for the SSI simulations, and it is highest for the PSI simulations. For each SI system, seed set is lowest when there is strong isolation-by-distance (see Fig. A34 for pairwise comparisons).

The expected mean-squared parent-offspring dispersal distances are 1.5, 6, 24, and 54 for dispersal parameters 1, 2, 4, and 6, respectively. The observed s^2 values for each simulation are listed in Table 14. In all cases, the observed values are slightly higher than the expected values but when isolation-by-distance is strong, the relative difference is much larger. Between the different SI groups, the s^2 values are not significantly different when $\sigma = 2$, 4, or 6. When $\sigma = 1$, the SSI simulations show significantly higher dispersal than GSI and PSI (Fig. A35).

Discussion

For this study, I assumed that any reduction in inbreeding below the level observed in the PSI simulations indicated a reduction in bi-parental inbreeding. I found that there was a significant decrease in identity-by-descent in the genetic SI systems compared to the physical PSI systems and the difference was consistent for a range of population sizes. The strictest SI system that was simulated was the SSI system and it showed the greatest reduction in bi-parental inbreeding. The amount of inbreeding was not significantly different between the GSI and BSI populations and the amount of inbreeding was lower than what was observed for the PSI simulations. This result suggests that, in addition to preventing inbreeding associated with self-fertilization, genetic SI systems are responsible for reducing bi-parental inbreeding. This supports the hypothesis that bi-parental inbreeding avoidance offers an additional selective advantage in genetic SI species. However, when looking at the total reduction in inbreeding from a self-compatible system (NSI), the reduction due to bi-parental inbreeding is relatively small compared the reduction due to preventing self-fertilization.

The amount of bi-parental inbreeding is expected to be greater when the dispersal of both seeds and pollen is reduced. When seeds are dispersed locally, offspring are more likely to become established near the parent plant and will be surrounded by related individuals. If pollen dispersal is limited, there will be a large proportion of related individuals in the mating pool. Therefore, genetic SI systems should be most advantageous in populations structured by isolation-by-distance and they should show the greatest reduction in bi-parental inbreeding in this situation. As, expected, the results of simulations run with different dispersal parameters supports this. The genetic SI systems showed a greater reduction in inbreeding compared to the PSI system when isolation-by-distance was strong and this decreased as the dispersal parameter increased. Again, the BSI and GSI systems were not significantly different from each other and the SSI system showed the greatest reduction. The frequency of homozygotes in the population was not significantly different between the different SI systems within the same level of dispersal, but the frequency of homozygotes was significantly higher when isolation-by-distance was strong.

These results are consistent with the results from Cartwright (2009). However, here I included a selective advantage for avoiding inbreeding by introducing several deleterious loci. In the simulation, inbreeding lead to the expression of recessive deleterious alleles at the 10 D loci, and as more deleterious alleles were expressed there was a higher chance of inbreeding depression induced sterility. The probability of a deleterious mutation at each loci was 0.1 which resulted in a genome-wide recessive mutation rate close to 1. When comparing the amount of inbreeding to simulations without inbreeding depression, I found that the amount of recent inbreeding, indicated by the probability of grandparental identity-by-descent, was not significantly different. However, the amount of inbreeding at equilibrium, indicated by the average frequency of homozygotes, was significantly lower when inbreeding depression was applied. This suggests that there is little selection against inbreeding in the very recent past but overall there is selection against homozygotes when there is a penalty for inbreeding.

Among the simulations with inbreeding depression, the average number of sterile individuals was not significantly different for any of the different SI systems. For these simulations, I used only a single fixed selection coefficient of 0.005 per loci. This selection coefficient was selected because it was high enough to produce a large effect in the population but it was small enough that the slightly deleterious alleles would not be quickly purged from the population. However, using a wider range of selection coefficients may reveal different relationships.

In genetic SI systems, the S-locus experiences negative frequency-dependent selection which favors low frequency alleles (Wright, 1939). This type of selection allows a large number of S alleles to be maintained in population which is necessary to keep the number of available mates high (Byers and Meagher, 1992). This is especially true under isolation-by-distance because the number of potential mates is already restricted to a local region. Here I found that at equilibrium, the SSI simulations maintained the highest number of alleles, especially when dispersal was restricted. The GSI populations had the next highest number followed by the BSI populations. The number of alleles was much lower for the BSI populations and this is most likely an effect of the dominance relationships between the S alleles. In the BSI system, recessive S alleles are masked by dominant alleles allowing the pollen to be compatible with other plants that share the recessive S allele. Ultimately, fewer S alleles need to be maintained in this system because more crosses are compatible (Hiscock and Tabah, 2003).

Due to reduced recombination and linkage disequilibrium near the S locus, it has been shown that the maintenance of many alleles (Cartwright, 2009) and heterozygosity (Uyenoyama, 1997) seen at the S locus extends to linked loci. Enforced heterozygosity near the S locus can lead to the accumulation of deleterious alleles that are sheltered from selection. This leads to an increase in the genetic load because these deleterious alleles are difficult to purge. In this study, the deleterious alleles were not linked to the S locus, and I did not consider the effect of linked deleterious alleles.

Reducing bi-parental inbreeding came with a cost in the genetic SI simulations. As expected, the total seed set and therefore female fecundity suffered under the genetic SI systems (Vekemans et al., 1998). Fecundity selection in the simulation was modeled by limiting the number of pollen grains produced by each plant. After pollen dispersal, each plant has a finite pollen pool that is further reduced when a high proportion of the pollen grains are incompatible. If the number of compatible pollen grains is less than the number of ovules, there will be a reduction in seed set. The lowest seed set was observed in the SSI simulations which had the strictest rules for compatibility. Seed set was also lowest when dispersal was limited; this is likely because the pollen pool consisted of a high proportion of close neighbors which were more likely to be related and thus incompatible. The reduction in seed set also translated into a reduction the census population size which followed a similar pattern.

In summary, there is some evidence that genetic SI systems reduce the amount of inbreeding in a population more than the physical SI system, and this is due to a reduction in bi-parental inbreeding. However, the effect is small and approximately the same number of individuals experienced inbreeding depression for each of the SI systems. Under the genetic SI systems, female fecundity and population size is reduced. In the conditions simulated here, it is possible that any benefit received from bi-parental inbreeding avoidance is negated by a reduction in fecundity. Further studies could be carried out to better understand the evolutionary dynamics of genetic SI systems versus physical SI systems. Competition or invasion simulations may help determine if genetic SI systems can rescue plant populations suffering from inbreeding depression and to better understand the adaptive dynamics.

Table 14. Seed set and population size is reduced when the SI system is more stringent. The table provides the average number of individuals (N), the average number of sterile individuals, the average seed set, and the average squared parent-offspring dispersal distance (s^2) for simulations with different SI systems and different dispersal parameters (σ). The maximum possible number of individuals in the population is 10,000 and the maximum number of seeds is 50,000.

	σ	N	Sterile	Seed Set	s^2
PSI	1	9906.1	484.8	45678.8	1.72
	2	9905.2	486.4	46344.5	6.20
	4	9905.0	485.5	46460.8	24.22
	6	9905.2	486.3	46474.6	54.23
GSI	1	9886.6	483.8	44134.1	1.72
	2	9901.8	486.1	45965.9	6.22
	4	9902.7	483.6	46291.9	24.26
	6	9902.8	483.6	46341.2	54.20
BSI	1	9878.2	484.4	43545.9	1.72
	2	9900.2	484.2	45750.6	6.21
	4	9901.7	484.8	46109.6	24.22
	6	9901.3	483.1	46169.4	54.23
SSI	1	9846.7	483.3	41549.9	1.74
	2	9896.6	484.2	45447.2	6.23
	4	9900.8	484.7	46071.9	24.18
	6	9901.2	484.2	46169.1	54.16

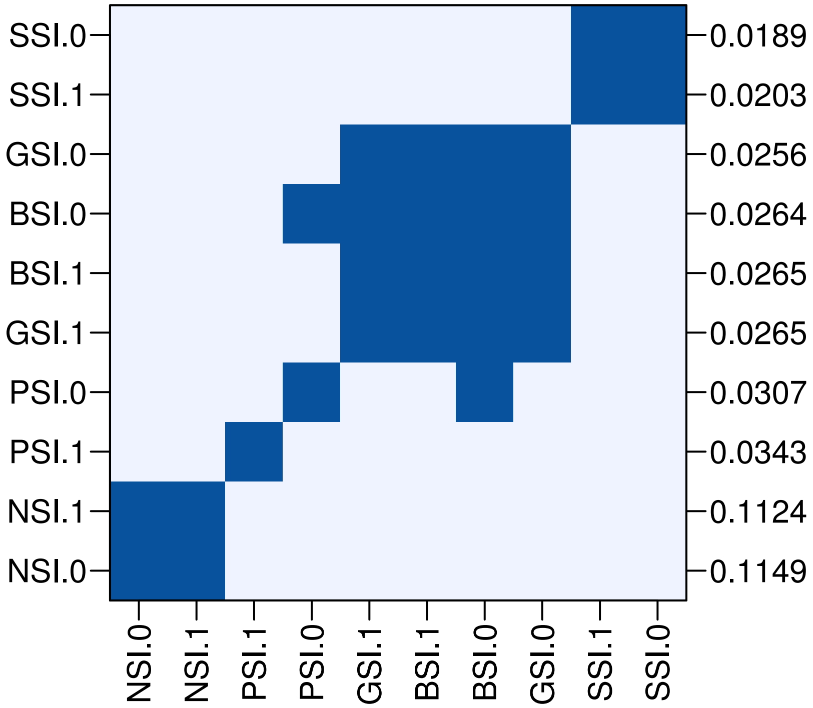


Figure 24. The probability of grandparental identity-by-descent for the genetic SI simulations is not significantly different when there is inbreeding depression compared to when there is no inbreeding depression. The plot shows the pairwise comparisons for each SI system with inbreeding depression (1) or without inbreeding depression (0). The simulations were run on a 100×100 landscape with the pollen and seed dispersal parameter $\sigma = 1$. The dark blue squares represent comparisons that were not significantly different. The values along the right side represent the average probability of grandparental identity-by-descent across all population samples, and the simulations are arranged from right to left in decreasing order based on the average probability.

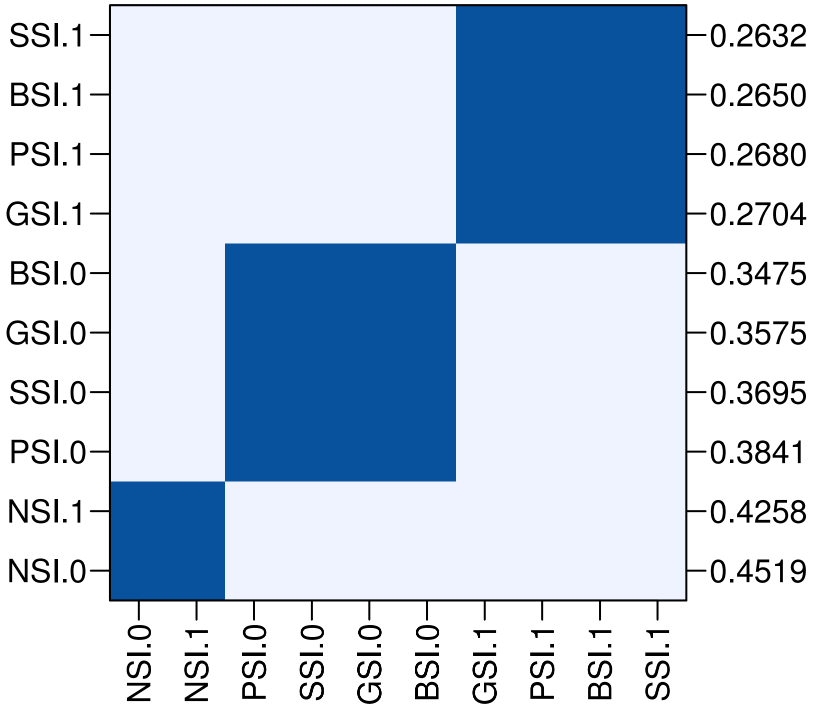


Figure 25. The frequency of homozygotes significantly decreases when inbreeding depression is introduced in the simulation. The plot shows the pairwise comparisons for each SI system with inbreeding depression (1) or without inbreeding depression (0). The simulations were run on a 100×100 landscape with the pollen and seed dispersal parameter $\sigma = 1$. The dark blue squares represent comparisons that were not significantly different. The values along the right side represent the average frequency of homozygotes across all population samples, and the simulations are arranged from right to left in decreasing order based on the average frequency.

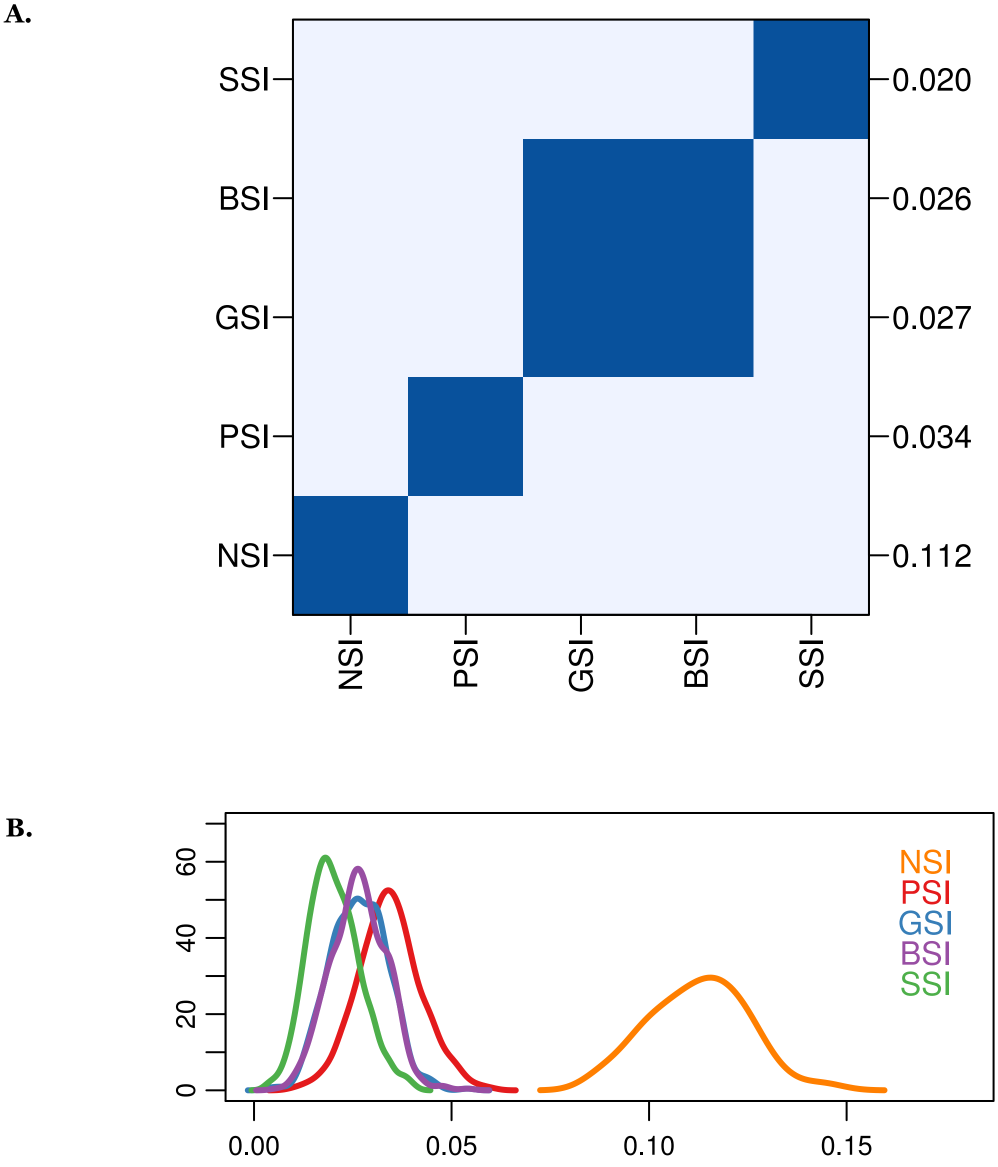


Figure 26. The probability of grandparental identity-by-descent decreases when the SI system is more stringent but BSI and GSI are not significantly different. A. The results of pairwise comparisons of the 100×100 simulations with $\sigma = 1$. The dark blue squares represent comparisons that were not significantly different. The values along the right side represent the average probability of grandparental identity-by-descent across all population samples. The different mating systems are arranged from left to right in descending order based on the average. B. The empirical density plot of the average probabilities of identity-by-descent for 500 samples from each mating system.

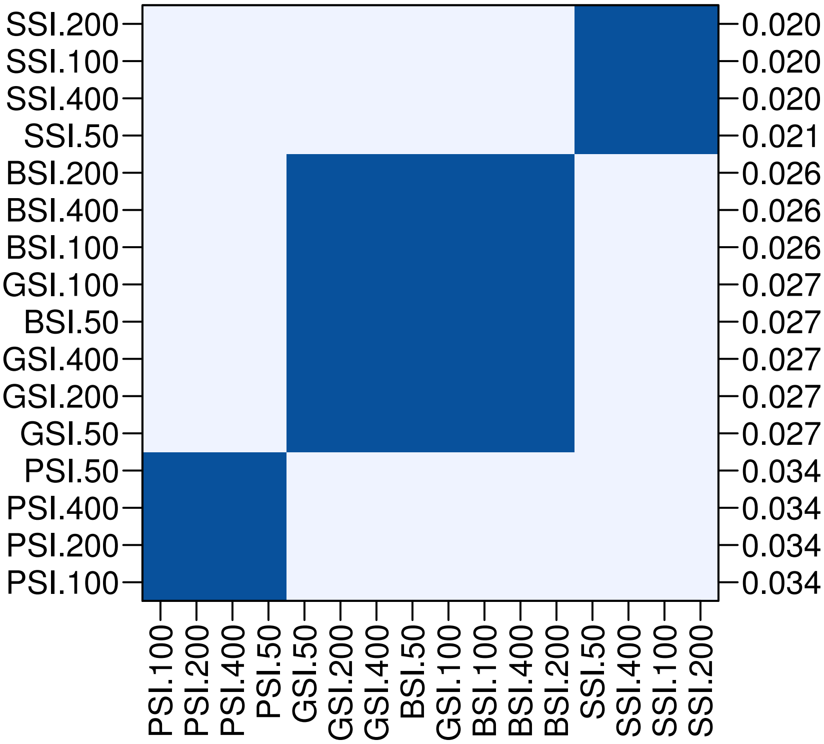


Figure 27. The probability of grandparental identity-by-descent follows a similar pattern for all population sizes. The plot shows the results of pairwise comparisons for each SI system and for each population size with $\sigma = 1$. The dark blue squares represent comparisons that were not significantly different. The values along the right side represent the average probability of grandparental identity-by-descent across all population samples. The labels along the left and bottom axes indicate the SI system and the width of the landscape and they are arranged from left to right in descending order based on the average probability.

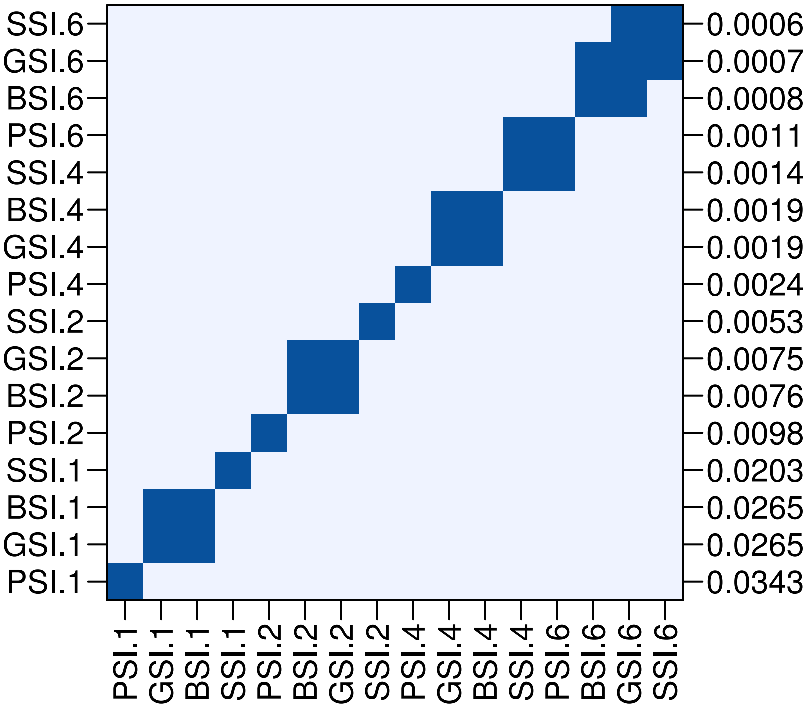


Figure 28. The decrease in inbreeding between PSI and the genetic SI systems is larger when there is strong isolation-by-distance. The plot shows the results of pairwise comparisons for each mating system at different dispersal levels. The features of the plot are the same as in Fig. 27. The labels along the left and bottom axes indicate the SI system and the dispersal parameter, σ, for both pollen and seed dispersal.

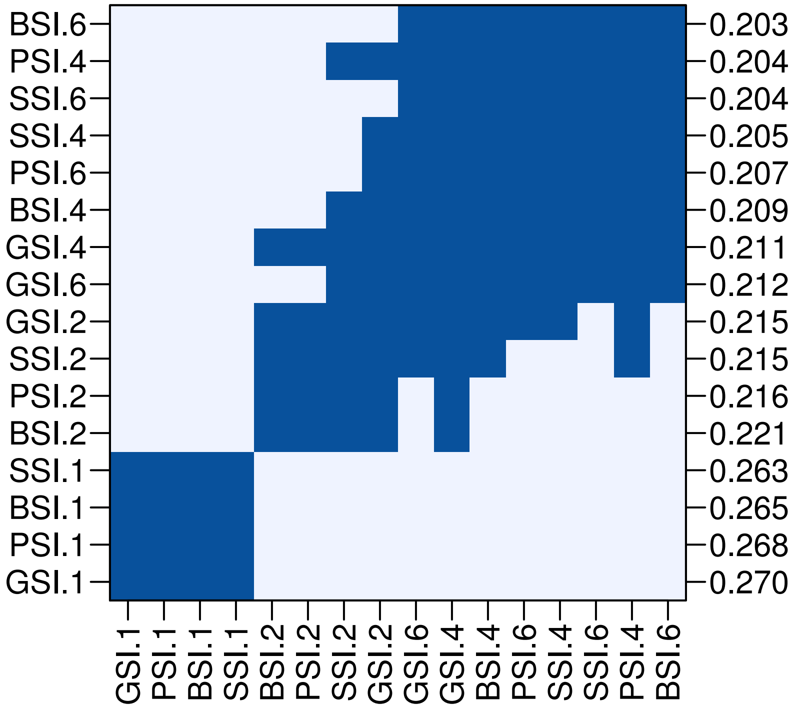


Figure 29. The frequency of homozygotes is higher when dispersal is limited but there is not a significant difference between the SI systems. The plot shows the results of the pairwise comparisons of the frequency of homozygotes for each SI system at different dispersal levels.

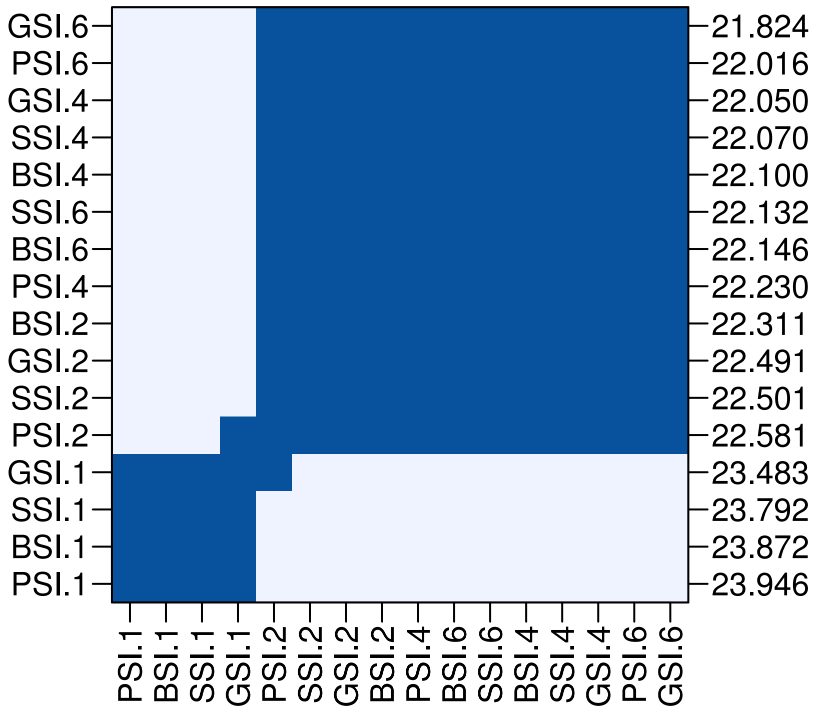


Figure 30. When isolation-by-distance is strong, more alleles are maintained at the M locus. The plot shows the results of the pairwise comparisons of the number of M alleles for each SI system at different dispersal levels.

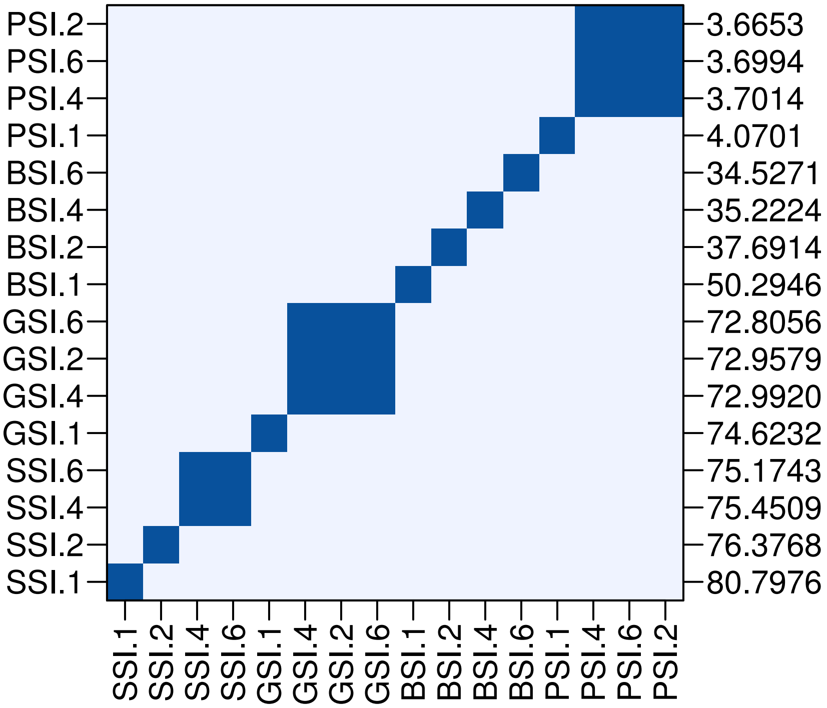


Figure 31. The SSI simulations maintain the highest number of S alleles followed by GSI, BSI, and PSI. More alleles are maintained when dispersal is limited. The plot shows the results of the pairwise comparisons of the number of S alleles for each SI system at different dispersal levels.

Chapter 5

CONCLUSION

For this dissertation I explored several different aspects of isolation-by-distance. In the first chapter I confirmed that the pattern of isolation-by-distance was similar for different dispersal distributions as long as they have the same second moment. I then argue the merits of modeling isolation-by-distance using the triangular distribution which produces uniform dispersal over the neighborhood area. I argued that this is more in line with the theory that the neighborhood represents a local panmictic unit and I provided an efficient algorithm for simulating triangular dispersal.

In the second chapter I presented a method I developed for the estimation of neighborhood size; the first such method to take a Bayesian approach. In this chapter I analyzed the performance of this method on data generated from the model, data from a lattice based simulation and a data set from two populations of *Pinus pinaster* Aiton. I found that when using independent data, the method had high coverage and low error but it was biased when fewer marker loci were used. I demonstrated that when using a composite marginal likelihood, the width of the credible intervals are artificially narrow when pairwise data is used and this results in reduced coverage. The method is robust to several violations of the model assumptions including the use of different dispersal distributions and different mutation models. Finally, I compared neighborhood size estimates using my model to the estimates published for the *Pinus pinaster* populations. For one population, the estimates agreed but for the second population my method provided a lower estimate.

In the third chapter, I examined bi-parental inbreeding, a consequence of isolation-by-distance, in simulated populations of self-incompatible plants. Genetic self-incompatiblity (SI) systems are extremely common across the angiosperms and I wanted to determine if this is a result of their ability to reduce bi-parental inbreeding. Surprisingly, I found reduced bi-parental inbreeding only accounts for a small portion of the total reduction in inbreeding demonstrated in the genetic SI systems. In addition, genetic SI populations produced fewer seeds and had a smaller population size, putting them at a disadvantage compared to physical SI systems.

REFERENCES

Andrew, R. L., Ostevik, K. L., Ebert, D. P., and Rieseberg, L. H. (2012). Adaptation with gene flow across the landscape in a dune sunflower. *Molecular Ecology*, 21:2078–2091.

Barluenga, M., Austerlitz, F., Elzinga, J. A., Teixeira, S., Goudet, J., and Bernasconi, G. (2011). Fine-scale spatial genetic structure and gene dispersal in *Silene latifolia*. *Heredity*, 106:13–24.

Barton, N. H., Etheridge, A. M., Kelleher, J., and Véber, A. (2013). Inference in two dimensions: Allele frequencies versus lengths of shared sequence blocks. *Theoretical Population Biology*, 87:105–119.

Bateman, A. J. (1950). Is gene dispersal normal? *Heredity*, 4:353–363.

Berdahl, A., Torney, C. J., Schertzer, E., and Levin, S. A. (2015). On the evolutionary interplay between dispersal and local adaptation in heterogeneous environments. *Evolution*, 69:1390–1405.

Bialozyt, R., Ziegenhagen, B., and Petit, R. J. (2006). Contrasting effects of long distance seed dispersal on genetic diversity during range expansion. *Journal of Evolutionary Biology*, 19:12–20.

Brent, R. P. (2007). Some long-period random number generators using shifts and xors. *ANZIAM Journal*, 48:C118–C202.

Bullock, J. M. and Clarke, R. T. (2000). Long distance seed dispersal by wind: Measuring and modelling the tail of the curve. *Oecologia*, 124:506–521.

Byers, D. L. and Meagher, T. R. (1992). Mate availability in small populations of plant species with homomorphic sporophytic self-incompatiblity. *Heredity*, 68:353–359.

Caine, M. L., Milligan, B. G., and Strand, A. E. (2000). Long-distance seed dispersal in plant populations. *American Journal of Botany*, 87:1217–1227.

Cartwright, R. A. (2009). Antagonism between local dispersal and self-incompatibility systems in a continuous plant population. *Molecular Ecology*, 18:2327–2336.

Casselman, A. L., Vrebalov, J., Conner, J. A., Singhal, A., Giovannoni, J., Nasrallah, M. E., and Nasrallah, J. B. (2000). Determining the physical limits of the brassica s locus by recombinational analysis. *The Plant Cell*, 12:23–33.

Castric, V., Bechsgaard, J. S., Grenier, S., Noureddine, R., Schierup, M. H., and Vekemans, X. (2010). Molecular evolution within and between self-incompatibility specificities. *Molecular Biology and Evolution*, 27:11–20.

Charlesworth, D. and Awadalla, P. (1998). The molecular population genetics of flowering plant self-incompatibility polymorphisms. *Heredity*, 91:1–9.

Charlesworth, D. and Charlesworth, B. (1987). Inbreeding depression and its evolutionary consequences. *Annual Review of Ecology and Systematics*, 18:237–268.

Charlesworth, D., Morgan, M. T., and Charlesworth, B. (1990). Inbreeding depression, genetic load, and the evolution of outcrossing rates in a multilocus system with no linkage. *Evolution*, 44:1469–1489.

Chesson, P. and Lee, C. T. (2005). Families of discrete kernels for modeling dispersal. *Theoretical Population Biology*, 67:241–256.

Chipperfield, J. D., Holland, E. P., Dytham, C., Thomas, C. D., and Hovestadt, T. (2011). On the approximation of continuous dispersal kernels in discrete-space models. *Methods in Ecology and Evolution*, 2:668–681.

Clark, C. J., Poulsen, J. R., Bolker, B. M., Connor, E. F., and Parker, V. T. (2005). Comparative seed shadows of bird-, monkey-, and wind-dispersed trees. *Ecology*, 86:2684–2694.

Clark, J. S. (1998). Why trees migrate so fast: Confronting theory with dispersal biology and the paleorecord. *The American Naturalist*, 152:204–224.

Clark, J. S., Lewis, M., and Horvath, L. (2001). Invasion by extremes: Population spread with variation in dispersal and reproduction. *The American Naturalist*, 157:537–554.

Crawford, T. J. (1984). The estimation of neighbourhood parameters for plant populations. *Heredity*, 52:273–283.

De-Lucas, A. I., González-Martínez, S. C., Vendramin, G. G., Hidalgo, E., and Heuertz, M. (2009a). Spatial genetic structure in continuous and fragmented populations of *Pinus pinaster* Aiton. *Molecular Ecology*, 18:4564–4576.

De-Lucas, A. I., González-Martínez, S. C., Vendramin, G. G., Hidalgo, E., and Heuertz, M. (2009b). Spatial genetic structure in continuous and fragmented populations of *Pinus pinaster* Aiton. *Dryad Digital Repository*.

Epperson, B. K. (1995). Spatial distributions of genotypes under isolation by distance. *Genetics*, 140:1431–1440.

Epperson, B. K. (2003). *Geographical Genetics*. Princeton University Press, Princeton, New Jersey.

Epperson, B. K. (2005). Estimating dispersal from short distance spatial autocorrelation. *Heredity*, 95:7–15.

Epperson, B. K. (2007). Plant dispersal, neighbourhood size and isolation by distance. *Molecular Ecology*, 16:3854–3865.

Epperson, B. K. and Li, T.-Q. (1997). Gene dispersal and spatial genetic structure. *Evolution*, 51:672–681.

Epperson, B. K., Mcrae, B. H., Scribner, K., Cushman, S. A., Rosenerg, M. S., Forin, M.-J., James, P. M. A., Murphy, M., Manel, S., Legendre, P., and Dale, M. R. T. (2010). Utility of computer simulations in landscape genetics. *Molecular Ecology*, 19:3549–3564.

Ewens, W. J. (2004). *Mathematical Population Genetics I. Theoretical Introduction*. Springer Science+Business Media, Inc., New York, New York, 2 edition.

Fenster, C. B. (1991). Gene flow in *Chamaecrista* fasciculata (leguminosae) i. gene dispersal. *Evolution*, 45:398–409.

Furstenau, T. N. and Cartwright, R. A. (2016). The effect of the dispersal kernel on isolation-by-distance in a continuous population. *PeerJ*, 4:e1848.

Gonzàlez-Martìnez, S. C., Burczyk, J., Nathan, R., Nanos, N., Gil, L., and Alìa, R. (2006). Effective gene dispersal and female reproductive success in Mediterranean maritime pine (*Pinus pinaster* Aiton). *Molecular Ecology*, 15:4577–4588.

Griffin, C. A. M. and Eckert, C. G. (2003). Experimental analysis of biparental inbreeding in a self-fertilizing plant. *Evolution*, 57:1513–1519.

Hardy, O. and Vekemans, X. (2002). SPAGeDi: A versatile computer program to analyse spatial genetic structure at the individual or population levels. *Molecular Ecology Notes*, 2:618–620.

Hardy, O. J. (2003). Estimation of pairwise relatedness between individuals and characterization of isolation-by-distance processes using dominant genetic markers. *Molecular Ecology*, 12(6):1577–1588.

Hardy, O. J. and Vekemans, X. (1999). Isolation by distance in a continuous population: Reconciliation between spatial autocorrelation analysis and population genetics models. *Heredity*, 83:145–154.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57:97–109.

Heywood, J. S. (1991). Spatial analysis of genetic variation in plant populations. *Annual Review of Ecology, Evolution, and Systematics*, 22:335–355.

Hiscock, S. J. and Tabah, D. A. (2003). The different mechanisms of sporophytic self-incompatiblity. *Philosophical Transactions of the Royal Society B*, 358:1037–1045.

Hoelzer, G. A., Drewes, R., Meier, J., and Doursat, R. (2008). Isolation-by-distance and outbreeding depression are sufficient to drive parapatric speciation in the absence of environmental influences. *PLoS Computational Biology*, 4:e1000126.

Holm, D. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6:65–70.

Houtan, K. S. V., Pimm, S. L., Halley, J. M., Jr., R. O. B., and Lovejoy, T. E. (2007). Dispersal of amazonian birds in continuous and fragmented forest. *Ecology Letters*, 10:219–229.

Howe, H. F., Schupp, E. W., and Westley, L. C. (1985). Early consequences of seed dispersal for a neotropical tree (*Virola surinamensis*). *Ecology*, 66:781–791.

Howe, H. F. and Smallwood, J. (1982). Ecology of seed dispersal. *Annual Review of Ecology and Systematics*, 13:201–228.

Ibrahim, K. M., Nichols, R. A., and Hewitt, G. M. (1996). Spatial patterns of genetic variation generated by different forms of dispersal during range expansion. *Heredity*, 77:282–291.

Igic, B., Lande, R., and Kohn, J. (2008). Loss of self-incompatibility and its evolutionary consequences. *International Journal of Plant Sciences*, 169:93–104.

Kamau, E., Charlesworth, B., and Charlesworth, D. (2007). Linkage disequilibrium and recombination rate estimates in the self-incompatibility region of *Arabidopsis lyrata*. *Genetics*, 176:2357–2369.

Kawabe, A., Hansson, B., Forrest, A., Hagenblad, J., and Charlesworth, D. (2006). Comparative gene mapping in *Arabidopsis lyrata* chromosomes 6 and 7 and *A. thaliana* chromosome iv: evolutionary history, rearrangements and local recombination rates. *Genetical Research*, 88:45–56.

Kimura, M. (1968). Genetic variability maintained in a finite population due to mutational production of neutral and nearly neutral isoalleles. *Genetical Research*, 11:247–269.

Kimura, M. and Crow, J. F. (1964). The number of alleles that can be maintained in a finite population. *Genetics*, 49:725–738.

Klein, E. K., Lavigne, C., Picault, H., Renard, M., and Gouyon, P.-H. (2006). Pollen dispersal of oilseed rape: Estimation of the dispersal function and effects of field dimension. *Journal of Applied Ecology*, 43:141–151.

Kot, M., Lewis, M. A., and van den Driessche, P. (1996). Dispersal data and the spread of invading organisms. *Ecology*, 77:2027–2042.

Larson, B. M. H. and Barrett, S. C. H. (2000). A comparative analysis of pollen limitation in flowering plants. *Biological Journal of the Linnean Society*, 69:503–520.

Leblois, R., Estoup, A., and Rousset, F. (2003). Influence of mutational and sampling factors on the estimation of demographic parameters in a “continuous” population under isolation by distance. *Molecular Biology and Evolution*, 20:491–502.

Leblois, R., Rousset, F., and Estoup, A. (2004). Influence of spatial and temporal heterogeneities on the estimation of demographic parameters in a continuous population using individual microsatellite data. *Genetics*, 166:1081–1092.

L’Ecuyer, P. and Simard, R. (2007). TestU01: A C library for empirical testing of random number generators. *ACM Transactions on Mathematical Software*, 33(4).

Leducq, J.-B., Llaurens, V., Castric, V., Saumitou-Laprade, P., Hardy, O. J., and Vekemans, X. (2011). Effect of balancing selection on spatial genetic structure within populations: Theoretical investigations on the self-incompatibility locus and empirical studies in *arabidopsis halleri*. *Heredity*, 106:319–329.

Leonardi, S., Piovani, P., Scalfi, M., Piotti, A., Giannini, R., and Menozzi, P. (2012). Effect of habitat fragementation on the genetic diversity and structure of peripheral populations of beech in central italy. *Journal of Heredity*, 103:408–417.

Levin, D. A. (1981). Dispersal versus gene flow in plants. *Annals of the Missouri Botanical Garden*, 68:233–253.

Llaurens, V., Gonthier, L., and Billiard, S. (2009). The sheltered genetic load linked to the *S* locus in plants: New insights from theoretical and empirical approaches in sporophytic self-incompatibility. *Genetics*, 183:1105–1118.

Lynch, M. and Ritland, K. (1999). Estimation of pairwise relatedness with molecular markers. *Genetics*, 152:1753–1766.

Malécot, G. (1969). *The Mathematics of Heredity*. W. H. Freeman and Company, Freeman, San Francisco.

Marsaglia, G. (2003). Xorshift RNGs. *Journal of Statistical Software*, 8.

Marsaglia, G. and Tsang, W. W. (2000a). A simple method for generating gamma variables. *ACM Transactions on Mathematical Software*, 26:363–372.

Marsaglia, G. and Tsang, W. W. (2000b). The ziggurat method for generating random variables. *Journal of Statistical Software*, 5.

Martìnez, I. and Gonzàlez-Taboada, F. (2009). Seed disperal patterns in a temperate forest during a mast event: Performance of alternative dispersal kernels. *Oecologia*, 159:389–400.

Maruyama, T. (1970). Effective number of alleles in a subdivided population. *Theoretical Population Biology*, 1:273–306.

Maruyama, T. (1972). Rate of decrease of genetic variability in a two-dimensional continuous population of finite size. *Genetics*, 70(4):639–651.

Meirmans, P. G. (2012). The trouble with isolation by distance. *Molecular Ecology*, 21:2839–2846.

Menéndez, P., Fan, Y., Garthwaite, P. H., and Sisson, S. A. (2014). Simultaneous adjustment of bias and coverage probabilities for confidence intervals. *Computational Statistics & Data Analysis*, 70:35–44.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equations of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21:1087–1091.

Moyle, L. C. (2006). Correlates of genetic differentiation and isolation by distance in 17 congeneric *Silene* species. *Molecular Ecology*, 15:1067–1081.

Nathan, R., Horvitz, N., He, Y., Kuparinen, A., Schurr, F. M., and Katul, G. G. (2011). Spread of North American wind-dispersed trees in future environments. *Ecology Letters*, 14:211–219.

Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A. R., Auton, A., Indap, A., King, K. S., Bergmann, S., Nelson, M. R., Stephens, M., and Bustamante, C. D. (2008). Genes mirror geography within europe. *Nature*, 456:98–101.

Novembre, J. and Slatkin, M. (2009). Likelihood-based inference in isolation-by-distance models using the spatial distribution of low frequency alleles. *Evolution*, 63:2914–2925.

Ohta, T. and Kumura, M. (1973). A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. *Genetical Research*, 22:201–204.

Patil, A., Huard, D., and Fonnesbeck, C. J. (2010). PyMC: Bayesian stochastic modelling in python. *Journal of Statistical Software*, 35.

Pauli, F., Racugno, W., and Ventura, L. (2011). Bayesian composite marginal likelihoods. *Statistica Sinica*, 21:149–164.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Raymond, M. and Rousset, F. (1995). Genepop (version 1.2): population genetics software for exact tests and ecumenicism. *Journal of Heredity*, 86:248–249.

Rousset, F. (1996). Equilibrium values of measures of population subdivision for stepwise mutation processes. *Genetics*, 142:1357–1362.

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

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

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

Rousset, F. (2004). *Genetic Structure and Selection in Subdivided Populations*. Princeton University Press, Princeton, New Jersey.

Rousset, F. (2008a). Demystifying Moran's I. *Heredity*, 100:231–232.

Rousset, F. (2008b). Genepop'007: a complete reimplementation of the genepop software for windows and linux. *Molecular Ecology Resources*, 8:103–106.

Sawyer, S. (1977). Asymptotic properties of the equilibrium probability of identity in a geographically structured population. *Advances in Applied Probability*, 9:268–282.

Scholz, F. and Zhu, A. (2016). *kSamples: K-Sample Rank Tests and their Combinations*. R package version 1.2-3.

Scholz, F. W. and Stephens, M. A. (1987). K-sample anderson-darling tests. *Journal of the American Statistical Association*, 82:918–924.

Slatkin, M. (1987). Gene flow and the geographic structure of natural populations. *Science*, 236:797–792.

Slatkin, M. (1993). Isolation by distance in equilibrium and non-equalibrium populations. *Evolution*, 47:264–279.

Spiegelhalter, D., Best, N. G., Carlin, B. P., and van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64:583–639.

Uyenoyama, M. K. (1997). Genealogical structure among alleles regulating self-incompatibility in natural populations of flowering plants. *Genetics*, 147:1389–1400.

Vekemans, X. and Hardy, O. J. (2004). New insights from fine-scale spatial genetic structure analyses in plant populations. *Molecular Ecology*, 13:921–935.

Vekemans, X., Schierup, M. H., and Christiansen, F. B. (1998). Mate availability and fecundity selection in multi-allelic self-incompatibility systems in plants. *Evolution*, 52:19–29.

Vieira, C. P., Charlesworth, D., and Vieira, J. (2003). Evidence for rare recombination at the gametophytic self-incompatibility locus. *Heredity*, 91:262–267.

Vitter, J. S. (1985). Random sampling with a reservoir. *ACM Transactions on Mathematical Software*, 11:37–57.

Vose, M. D. (1991). A linear algorithm for generating random numbers with a given distribution. *IEEE Transactions on Software Engineering*, 17:972–975.

Wang, J. (2014). Marker-based estimates of relatedness and inbreeding coefficients: An assessment of current methods. *Journal of Evolutionary Biology*, 27:518–530.

Wang, J., Hill, W., Charlesworth, D., and Charlesworth, B. (1999). Dynamics of inbreeding depression due to deleterious mutations in small populations: mutaiton parameters and inbreeding rate. *Genetical Research*, 74:165–178.

Wright, S. (1939). The distribution of self-sterility alleles in populations. *Genetics*, 24:539–552.

Wright, S. (1943). Isolation by distance. *Genetics*, 23:114–138.

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

Xu, X. and Reid, N. (2011). On the robustness of maximum composite likelihood estimate. *Journal of Statistical Planning and Inference*, 141:3047–3054.

Zhao, Y., Vrieling, K., Liao, H., Xiao, M., Zhu, Y., Rong, J., Zhang, W., Wang, Y., Yang, J., Chen, J., and Song, Z. (2013). Are habitat fragmentation, local adaptation and isolation-by-distance driving population divergence in wild rice *Oryza rufipogon*? *Molecular Ecology*, 22:5531–5547.

APPENDIX A

TRIANGULAR DISPERSAL KERNEL

Xorshift Random Number Generator

Xorshift is a type of pseudo-random number generator that relies on exclusive-or and bitshift operators (Marsaglia, 2003). Xorshift is one of the most efficient, high-quality random-number generators known. My implementation is a 64-bit xorshift with shift parameters (5, 15, 27) added to a Weyl series to decrease bit correlations (Brent, 2007). It passes the BigCrush tests in the TestU01 suite (L'Ecuyer and Simard, 2007).

Triangular Distributed Distances Produce a Uniform Distribution on a Disk

Proof

The probability density of a uniform distribution over a finite two-dimensional shape is defined as:

$$f(x,y) = \begin{cases} \frac{1}{\text{area of } S} & \text{if } (x,y) \in S \\ 0 & \text{otherwise} \end{cases}$$

where (x, y) are coordinates on the Cartesian plane and S is the set of all points within the shape. A uniform distribution on the region bounded by a circle is defined by $\frac{1}{\pi R^2}$ where R is the radius of the circle. I am interested in a circle with radius $R = 2\sigma$ and area $A = 4\pi\sigma^2$ so the non-zero part of the joint probability distribution is given by:

$$f(x,y;\sigma) = \frac{1}{4\pi\sigma^2} \quad \text{when } x^2 + y^2 \leq 4\sigma^2$$

Using the change of variables theorem for polar coordinates, (r, θ), I have:

$$\begin{aligned} \iint_D f(x,y)\ \mathrm{d}x\ \mathrm{d}y &= \iint_{D^*} f(r\cos\theta, r\sin\theta) r\ \mathrm{d}r\ \mathrm{d}\theta \\ &= \iint_{D^*} \frac{r}{4\pi\sigma^2}\ \mathrm{d}r\ \mathrm{d}\theta \\ &= \iint_{D^*} \frac{1}{2\pi}\frac{r}{2\sigma^2}\ \mathrm{d}r\ \mathrm{d}\theta \end{aligned}$$

I then integrate out the angle θ to isolate the distribution of distance, $f(r; \sigma)$.

$$f(r;\sigma) = \int_0^{2\pi} \frac{r}{4\pi\sigma^2}\ \mathrm{d}\theta = \frac{r}{4\pi\sigma^2}\theta\Big|_0^{2\pi} = \frac{r}{2\sigma^2} \quad \text{for } 0 \leq r \leq 2\sigma$$

The distribution of distances is equivalent to a special case of the triangular distribution. The probability density function for the triangular distribution is

$$f(r;a,b,c) = \begin{cases} 0 & \text{for } r < a \\ \frac{2(r-a)}{(b-a)(c-a)} & \text{for } a \leq r \leq c \\ \frac{2(b-r)}{(b-a)(b-c)} & \text{for } c \leq r \leq b \\ 0 & \text{for } r > b \end{cases}$$

where a is the lower limit, b is the upper limit, and c is the mode. In the special case I set $a = 0$ and $b = c = 2\sigma$. The probability density function for the special case of the triangular distribution simplifies to

$$f(r;\sigma) = \begin{cases} \frac{r}{2\sigma^2} & \text{for } 0 \leq r \leq 2\sigma \\ 0 & \text{otherwise} \end{cases}$$

Generating from a Triangular Distribution

Inverse sampling can be used to generate values from a triangular distribution. — Note that I am only working with monotonically increasing triangular distributions and not more general formulations. — If u is uniformly distributed in $(0, 1)$, the value $d = 2s\sqrt{u}$ has a triangular distribution with parameter s. However, a modified rejection sampling algorithm is faster. If u_1 and u_2 are independent and uniformly distributed in $(0, 1)$, then $d = 2s \max(u_1, u_2)$ also has a triangular distribution. Because I can generate 32-bit values for both u_1 and u_2 from a single 64-bit random number, this second algorithm is more efficient than the first. While it is possible to construct a ziggurat algorithm (Marsaglia and Tsang, 2000b) for a triangular distribution, my second algorithm is more efficient because it involves fewer steps and never rejects.

I compared the speed of these algorithms and a naive rejection sampler using the medium Crush tests (L'Ecuyer and Simard, 2007). This allowed us to compare the speeds of these algorithms in a data-intensive application as well as verify that the algorithms produced independent and identically distributed values from the correct distribution. The 'maximum' algorithm took 1656 seconds to complete, while the 'sqrt' took 1700s and the rejection sampler took 1911s. The maximum algorithm produced faster execution, but only sped up the tests by 3% over sqrt.

Generating Discrete Two-Dimensional Dispersal from a Triangular Distribution

I can use the maximum algorithm above to generate the values in polar coordinates and convert them to Cartesian coordinates; however, this requires calculating sine and cosine functions, which I would rather not do. When modeling dispersal on a lattice, the bounded nature of the triangular distribution allows dispersal to be modeled discretely. To discretize this distribution on a rectangular lattice I must determine the probabilities for each cell which are proportional to the area of the cell that is covered by a disk of radius $r = 2\sigma$ (centered on a focal cell). The algorithm described here produces probability tables by calculating the appropriate area for each cell and dividing by the total area. I assume that cells are squares with unit area.

Since the disk is symmetrical, this algorithm may be simplified by calculating areas for quadrant I of the disk and mirroring those values to the other quadrants. I further simplify by calculating approximately half of the areas for quadrant I and mirroring those as well. — Note that this results in cells along the x and y axes having an area of 1/2. — Starting at the center of the focal cell ($y_0 = 0$), I record the top/bottom boundary of each cell along the y-axis up to the radius: $y_1 = 0.5, y_2 = 1.5, \ldots, y_n = n - 0.5$ where $n = \sup_{n \in \mathbb{Z}} y_n \leq r$.

Next I calculate the area of the first column of cells which has a left boundary at $x_0 = 0$ and a right boundary at $x_1 = \min(0.5, r)$:

$$A = \int_{x_0}^{x_1} \sqrt{r^2 - x^2}\, dx$$

Starting with the bottom cell, I check if the area of a cell is less than the area of the column. If so, the cell is completely contained in the disk, and the cell is assigned a weight equal to its area. Its area is then subtracted from the area of the column. I continue this procedure until the the area of last cell is less than the remaining area of the column and assign the final cell a weight equal to the remaining area in the column.

I then move to the next column by setting $x_0 = 0.5$ and $x_1 = 1.5$. However, before I calculate the area, I must check if the edge of the disk passes through the bottom of the top cell. This occurs if $x_1^2 + y^2 > r^2$, where y is the value of the bottom boundary of the cell. When this occurs, I split the column into two smaller columns and each column is processed just like before. I continue calculating the area of subsequent columns until I reach the column that contains the point $\{x, y\} = \left\{r/\sqrt{2}, r/\sqrt{2}\right\}$, which marks the point where the edge of the disk intersects the diagonal. After this column is processed, the weights for these cells can be copied symmetrically. The weight of each cell is divided by the total area of the disk and becomes a probability. These probabilities are then copied symmetrically to the other three quadrants. The completed table of probabilities can then be passed into the alias algorithm for discrete sampling (Vose, 1991).

My implementation of a discretized triangular kernel can be found in src/disk.h and src/disk.cpp in the source code. Code for generating an alias table can be found in src/aliastable.h.

APPENDIX B

PAIRWISE COMPARISONS OF DEMOGRAPHIC PARAMETERS FROM SELF-INCOMPATIBILITY SIMULATIONS

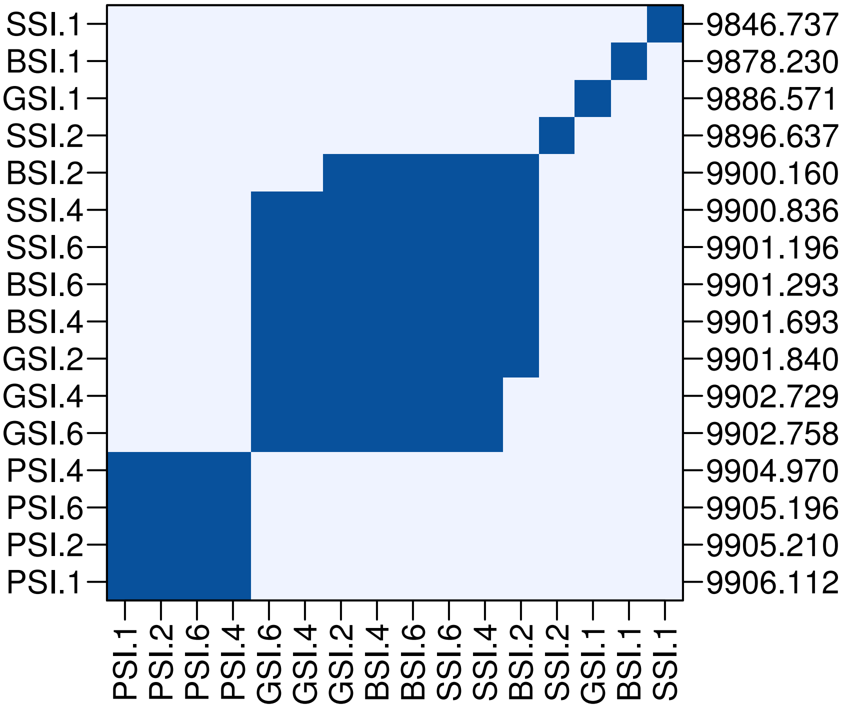


Figure 32. Population size is higher for PSI simulations compared to genetic SI simulations. The plot shows results of pairwise comparisons of the 100×100 simulations. The dark blue squares represent comparisons that were not significantly different. The values along the right side represent the average census population size across all populations for each type of simulation. The labels along the left and bottom axes indicate the SI system and the dispersal parameter (σ). The simulations are arranged from left to right in descending order based on the average population size.

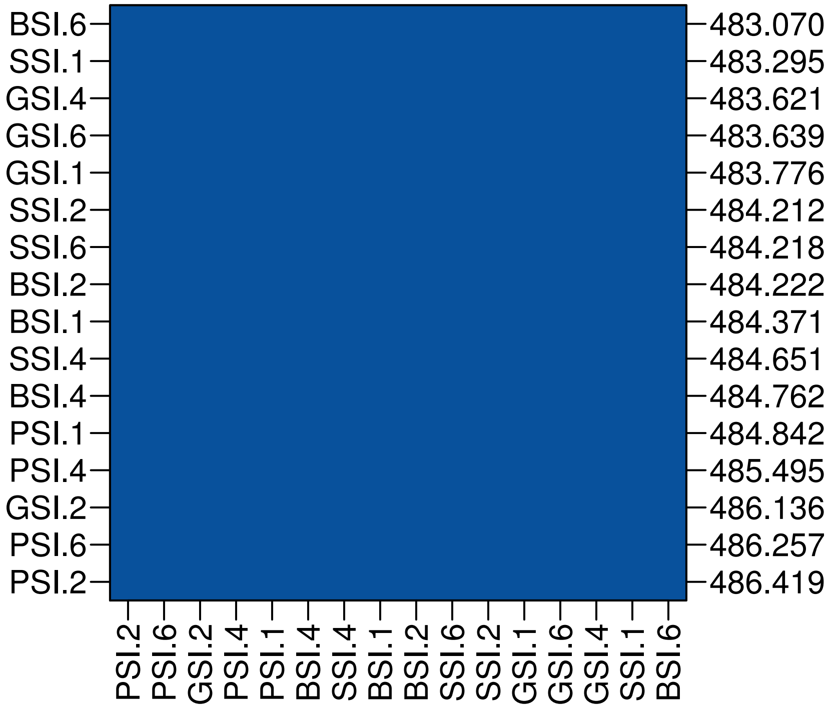


Figure 33. The average number of sterile individuals in the population is similar in all simulations. The plot shows results of pairwise comparisons of the 100×100 simulations. The dark blue squares represent comparisons that were not significantly different. The values along the right side represent the number of sterile individuals per population averaged across all populations for each type of simulation. The labels along the left and bottom axes indicate the SI system and the dispersal parameter (σ). The simulations are arranged from left to right in descending order based on the average number of sterile individuals.

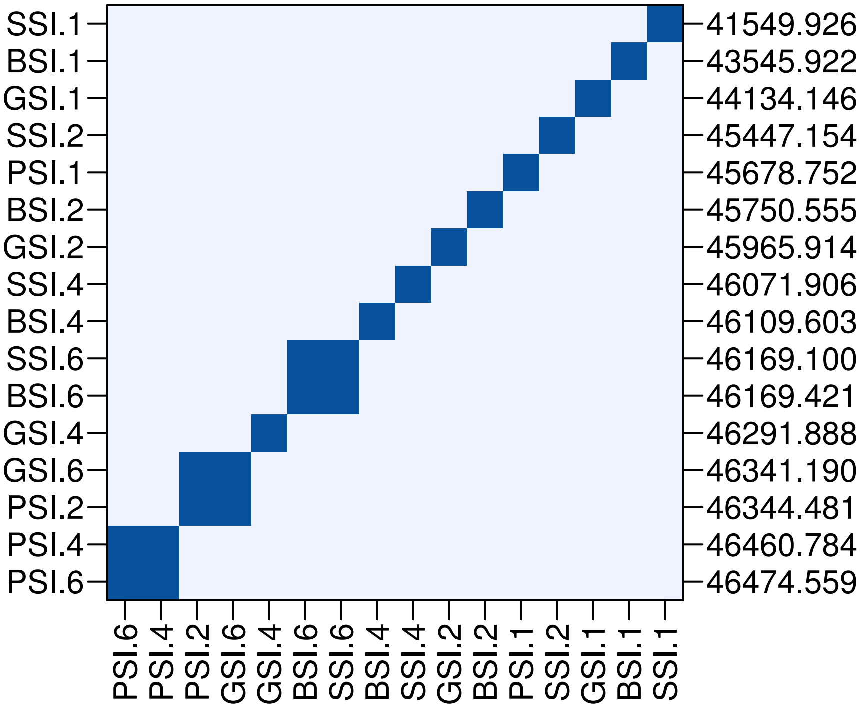


Figure 34. The seed set is lower for the genetic SI simulations and when dispersal is limited. The plot shows results of pairwise comparisons of the 100×100 simulations. The dark blue squares represent comparisons that were not significantly different. The values along the right side represent the total number seeds produced per population averaged across all populations for each type of simulation. The labels along the left and bottom axes indicate the SI system and the dispersal parameter (σ). The simulations are arranged from left to right in descending order based on the average number of seeds produced.

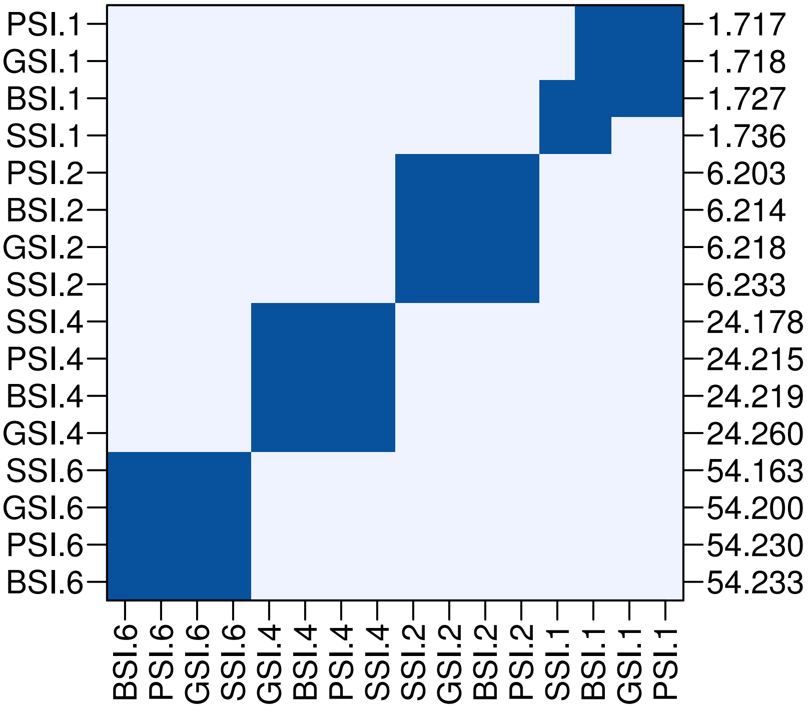


Figure 35. Observed dispersal is slightly higher than expected. The plot shows results of pairwise comparisons of the 100×100 simulations. The dark blue squares represent comparisons that were not significantly different. The values along the right side represent the average squared parent-offspring dispersal distances per population averaged across all populations for each type of simulation. The labels along the left and bottom axes indicate the SI system and the dispersal parameter (σ). The simulations are arranged from left to right in descending order based on the dispersal distance. The expected values are 1.5, 6, 24, and 54 for dispersal parameters 1, 2, 4, and 6, respectively.

APPENDIX C

PERMISSION TO USE PUBLISHED ARTICLES

My co-author, Reed Cartwright, has granted permission to use the previously published work that appears in Chapter 2 of this dissertation.