PREPRINT

Space is the Place: Effects of Continuous Spatial Structure on Analysis of Population Genetic Data

C.J. Battey[*,1], **Peter L. Ralph**[*,†] **and Andrew D. Kern**[*,†]
[*]University of Oregon Dept. Biology, Institute for Ecology Evolution

ABSTRACT Real geography is continuous, but standard models in population genetics are based on discrete, well-mixed populations. As a result many methods of analyzing genetic data assume that samples are a random draw from a well-mixed population, but are applied to clustered samples from populations that are structured clinally over space. Here we use simulations of populations living in continuous geography to study the impacts of dispersal and sampling strategy on population genetic summary statistics, demographic inference, and genome-wide association studies. We find that most common summary statistics have distributions that differ substantially from that seen in well-mixed populations, especially when Wright's neighborhood size is less than 100 and sampling is spatially clustered. The combination of low dispersal and clustered sampling causes demographic inference from the site frequency spectrum to infer more turbulent demographic histories, but averaged results across multiple simulations were surprisingly robust to isolation by distance. We also show that the combination of spatially autocorrelated environments and limited dispersal causes genome-wide association studies to identify spurious signals of genetic association with purely environmentally determined phenotypes, and that this bias is only partially corrected by regressing out principal components of ancestry. Last, we discuss the relevance of our simulation results for inference from genetic variation in real organisms.

KEYWORDS Space; Population Structure; Demography; Haplotype block sharing; GWAS

Introduction

The inescapable reality that biological organisms live, move, and reproduce in continuous geography is usually omitted from population genetic models. However, mates tend to live near to one another and to their offspring, leading to a positive correlation between genetic differentiation and geographic distance. This pattern of "isolation by distance" (**?**) is one of the most widely replicated empirical findings in population genetics (**???**). Despite a long history of analytical work describing the genetics of populations distributed across continuous geography (e.g., **????????**), much modern work still describes geographic structure as a set of discrete populations connected by migration (e.g., **??????**). For this reason, most population genetics statistics are interpreted with reference to discrete, well-mixed populations, and most empirical papers analyze variation within clusters of genetic variation inferred by programs like *STRUCTURE* (**?**) with methods that assume these are randomly mating units.

Manuscript compiled: Friday 8th November, 2019
[1]301 Pacific Hall, University of Oregon Dept. Biology, Institute for Ecology and Evolution. cbattey2@uoregon.edu.
[†]these authors co-supervised this project

The assumption that populations are "well-mixed" has important implications for downstream inference of selection and demography. Methods based on the coalescent (**??**) assume that the sampled individuals are a random draw from a well-mixed population that is much larger than the sample (**?**). The key assumption is that the individuals of each generation are *exchangeable*, so that there is no correlation between the fate or fecundity of a parent and that of their offspring (**?**). If dispersal or mate selection is limited by geographic proximity, this assumption can be violated in many ways. For instance, if mean viability or fecundity is spatially autocorrelated, then limited geographic dispersal will lead to parent–offspring correlations. Furthermore, nearby individuals will be more closely related than an average random pair, so drawing multiple samples from the same area of the landscape will represent a biased sample of the genetic variation present in the whole population (**?**).

Two areas in which spatial structure may be particularly important are demographic inference and genome-wide association studies (GWAS). Previous work has found that discrete population structure can create false signatures of population bottlenecks when attempting to infer demographic histories from microsatellite variation (**?**), statistics summarizing the site frequency spectrum (SFS) (**???**), or runs of homozygosity in a single individual (**?**). The increasing availability of whole-genome data has led to the development of many methods that attempt to infer detailed trajectories of population sizes through time based on a variety of summaries of genetic data (**????**). Because all of these methods assume that the populations being modeled are approximately randomly mating, they are likely affected by spatial biases in the genealogy of sampled individuals (**?**), which may lead to incorrect inference of population changes over time (**?**). However, previous investigations of these effects have focused on discrete rather than continuous space models, and the level of isolation by distance at which inference of population size trajectories become biased by structure is not well known. Here we test how two methods suitable for use with large samples of individuals – stairwayplot (**?**) and SMC++ (**?**) – perform when applied to populations evolving in continuous space with varying sampling strategies and levels of dispersal.

Spatial structure is also a major challenge for interpreting the results of genome-wide association studies (GWAS). This is because many phenotypes of interest have strong geographic differences due to the (nongenetic) influence of environmental or socioeconomic factors, which can therefore show spurious correlations with spatially patterned allele frequencies (**??**). Indeed, two recent studies found that previous evidence of polygenic selection on human height in Europe was confounded by subtle population structure (**??**), suggesting that existing methods to correct for population structure in GWAS are insufficient. However we have little quantitative idea of the population and environmental parameters that can be expected to lead to biases in GWAS.

Last, some of the most basic tools of population genetics are summary statistics like F_{IS} and Tajima's D, which are often interpreted as reflecting the influence of selection or demography on sampled populations (**?**). Statistics like Tajima's D are essentially summaries of the site frequency spectrum, which itself reflects variation in branch lengths and tree structure of the underlying genealogies of sampled individuals. Geographically limited mate choice distorts the distribution of these genealogies (**??**), which can affect the value of Tajima's D (**?**). Similarly, the distribution of tract lengths of identity by state among individuals contains information about not only historical demography (**??**) and selection (**?**), but also dispersal and mate choice (**??**). We are particularly keen to examine how such summaries will be affected by models that incorporate continuous space, both to evaluate the assumptions underlying existing methods and to identify where the most promising signals of geography lie.

To study this, we have implemented an individual-based model in continuous geography that incorporates overlapping generations, local dispersal of offspring, and density-dependent survival. We simulate chromosome-scale genomic data in tens of thousands of individuals from parameter regimes relevant to common subjects of population genetic investigation such as humans and *Drosophila*, and output the full genealogy and recombination history of all final-generation individuals. We use these simulations to test how sampling strategy interacts with geographic population structure to cause systematic variation in population genetic summary statistics typically analyzed assuming discrete population models. We then examine how the fine-scale spatial structures occurring under limited dispersal impact demographic inference from the site frequency spectrum. Last, we examine the impacts of continuous geography on genome-wide association studies (GWAS) and identify regions of

parameter space under which the results from GWAS may be misleading.

Materials and Methods

Modeling Evolution in Continuous Space

The degree to which genetic relationships are geographically correlated depends on the chance that two geographically nearby individuals are close relatives – in modern terms, by the tension between migration (the chance that one is descended from a distant location) and coalescence (the chance that they share a parent). A key early observation by Wright (**?**) is that this balance is often nicely summarized by the "neighborhood size", defined to be $N_W = 4\pi\rho\sigma^2$, where σ is the mean parent–offspring distance and ρ is population density. This can be thought of as proportional to the average number of potential mates for an individual (those within distance 2σ), or the number of potential parents of a randomly chosen individual. Empirical estimates of neighborhood size vary hugely across species – even in human populations, estimates range from 40 to over 5,000 depending on the population and method of estimation (Table 1).

The first approach to modeling continuously distributed populations was to endow individuals in a Wright-Fisher model with locations in continuous space. However, since the total size of the population is constrained, this introduces interactions between arbitrarily distant individuals, which (aside from being implausible) was shown by **?** to eventually lead to unrealistic population clumping if the range is sufficiently large. Another method for modeling spatial populations is to assume the existence of a grid of discrete randomly mating populations connected by migration, thus enforcing regular population density by edict. Among many other important results drawn from this class of "lattice" or "stepping stone" models, **?** showed that the slope of the linear regression of genetic differentiation (F_{ST}) against the logarithm of spatial distance is an estimate of neighborhood size. Although these grid models may be good approximations of continuous geography in many situations, they do not model demographic fluctuations, and limit investigation of spatial structure below the level of the deme, assumptions whose impacts are unknown. An alternative method for dealing with continuous geography is a new class of coalescent models, the Spatial Lambda Fleming-Viot models (**??**).

To avoid questionable assumptions, we here used forward-time, individual-based simulations. By scaling the probability of survival in each timestep to local population density we shift reproductive output towards low-density regions, which prevents populations from clustering. Such models have been used extensively in ecological modeling but rarely in population genetics, where to our knowledge previous implementations of continuous space models have focused on a small number of genetic loci, which limits the ability to investigate the impacts of continuous space on genome-wide genetic variation as is now routinely sampled from real organisms. By simulating chromosome-scale sequence alignments and complete population histories we are able to treat our simulations as real populations and replicate the sampling designs and analyses commonly conducted on real genomic data.

A Forward-Time Model of Evolution in Continuous Space

We simulated populations using the non-Wright-Fisher module in the program SLiM v3.1 (**?**). Each time step consists of three stages: reproduction, dispersal, and mortality. To reduce the parameter space we use the same parameter, denoted σ, to modulate the spatial scale of interactions at all three stages by adjusting the standard deviation of the corresponding Gaussian functions. As in previous work (**??**), σ is equal to the mean parent-offspring distance.

At the beginning of the simulation individuals are distributed uniformly at random on a continuous, square landscape. Individuals are hermaphroditic, and each time step, each produces a Poisson number of offspring with mean $1/L$ where L is the expected lifespan. Offspring disperse a Gaussian-distributed distance away from the parent with mean zero and standard deviation σ in both the x and y coordinates. Each offspring is produced with a mate selected randomly from those within distance 3σ, with probability of choosing a neighbor at distance x proportional to $\exp(-x^2/2\sigma^2)$.

To maintain a stable population, mortality increases with local population density. To do this we say that individuals at distance d have a competitive interaction with strength $g(d)$, where g is the Gaussian density with mean zero and standard deviation σ. Then, the sum of all competitive interactions with

individual i is $n_i = \sum_j g(d_{ij})$, where d_{ij} is the distance between individuals i and j and the sum is over all neighbors within distance 3σ. Since g is a probability density, n_i is an estimate of the number of nearby individuals per unit area. Then, given a per-unit carrying capacity K, the probability of survival until the next time step for individual i is

$$p_i = \min\left(0.95, \frac{1}{1 + n_i/(K(1+L))}\right). \quad (1)$$

We chose this functional form so that the equilibrium population density per unit area is close to K, and the mean lifetime is around L.

An important step in creating any spatial model is dealing with range edges. Because local population density is used to model competition, edge or corner populations can be assigned artificially high fitness values because they lack neighbors within their interaction radius but outside the bounds of the simulation. We approximate a decline in habitat suitability near edges by decreasing the probability of survival proportional to the square root of distance to edges in units of σ. The final probability of survival for individual i is then

$$s_i = p_i \min(1, \sqrt{x_i/\sigma}) \min(1, \sqrt{y_i/\sigma}) \min(1, \sqrt{(W - x_i)/\sigma}) \min(1, \sqrt{(W - y_i)/\sigma}) \quad (2)$$

where x_i and y_i are the spatial coordinates of individual i, and W is the width (and height) of the square habitat. This buffer roughly counteracts the increase in fitness individuals close to the edge would otherwise have.

To isolate spatial effects from other components of the model such as overlapping generations, increased variance in reproductive success, and density-dependent fitness, we also implemented simulations identical to those above except that mates are selected uniformly at random from the population, and offspring disperse to a uniform random location on the landscape. We refer to this model as the "random mating" model, in contrast to the first, "spatial" model.

We stored the full genealogy and recombination history of final-generation individuals as tree sequences (**?**), as implemented in SLiM (**?**). Scripts for figures and analyses are available at https://github.com/petrelharp/spaceness.

We ran 400 simulations for the spatial and random-mating models on a square landscape of width $W = 50$ with per-unit carrying capacity $K = 5$ (census $N \approx 10,000$), average lifetime $L = 4$, genome size = 10^8, recombination rate = 10^{-9}, and drawing σ values from a uniform distribution between 0.2 and 4. To speed up the simulations and limit memory overhead we used a mutation rate of 0 in SLiM and later applied mutations to the tree sequence with msprime's `mutate` function (**?**). Because msprime applies mutations proportionally to elapsed time, we divided the mutation rate of 10^{-8} mutations per site per generation by the average generation time estimated for each value of σ (see 'Demographic Parameters' below) to convert the rate to units of mutations per site per unit time. We show that this procedure produces the same site-frequency spectrum as applying mutations directly in SLiM in Figure S2. Simulations were run for 1.6 million timesteps (approximately $30N$ generations).

To check that our model produces reasonable results, we compared its output to that of a stepping-stone model implemented in msprime (**?**). These results are discussed in detail in Appendix 1, but in general we find that our model produces many of patterns generated by stepping stone models while avoiding some artifacts associated with discretization of the landscape.

Demographic Parameters

Our demographic model includes parameters that control population density (K), mean life span (L), and dispersal distance (σ). However, nonlinearity of local demographic stochasticity causes actual realized averages of these demographic quantitites to deviate from the specified values in a way that depends on the neighborhood size. Therefore, to properly compare to theoretical expectations, we empirically calculated these demographic quantities in simulations. We recorded the census population size in all simulations. To estimate generation times, we stored ages of the parents of every new individual born across 200 timesteps, after a 100 generation burn-in, and took the mean. To

estimate variance in offspring number, we tracked the number of offspring for all individuals for 100 timesteps following a 100-timestep burn-in period, subset the resulting table to include only the last timestep recorded for each individual, and calculated the variance in number of offspring across all individuals in timesteps 50-100. All calculations were performed with information recorded in the tree sequence, using pyslim (https://github.com/tskit-dev/pyslim).

Sampling

Our model records the genealogy and sequence variation of the complete population, but in real data, genotypes are only observed from a relatively small number of sampled individuals. We modeled three sampling strategies similar to common data collection methods in empirical genetic studies (Figure 1). "Random" sampling selects individuals at random from across the full landscape, "point" sampling selects individuals proportional to their distance from four equally spaced points on the landscape, and "midpoint" sampling selects individuals in proportion to their distance from the middle of the landscape. Downstream analyses were repeated across all sampling strategies.

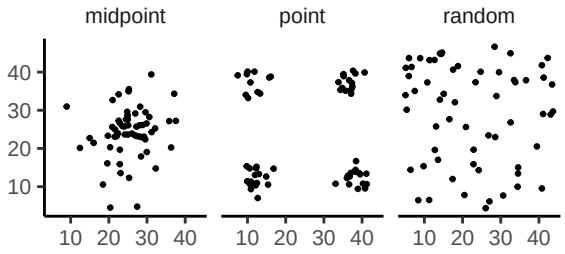

Figure 1 Example sampling maps for 60 individuals on a 50×50 landscape for midpoint, point, and random sampling strategies, respectively.

Summary Statistics

We calculated the site frequency spectrum and a set of 18 summary statistics (Table S1) from 60 diploid individuals sampled from the final generation of each simulation using the python package scikit-allel (**?**). Statistics included common single-population summaries including mean pairwise divergence (π), inbreeding coefficient (F_{IS}), and Tajima's D, as well as an isolation-by-distance regression of genetic distance (D_{xy}) against the logarithm of geographic distance analogous to **?**'s approach, which we summarized as the correlation coefficient between the logarithm of the spatial distance and the proportion of identical base pairs across pairs of individuals.

Following recent studies that showed strong signals for dispersal and demography in the distribution of shared haplotype block lengths (**??**), we also calculated various summaries of the distribution of pairwise identical-by-state (IBS) block lengths among sampled chromosomes. The full distribution of lengths of IBS tracts for each pair of chromosomes was first calculated with a custom python function. We then calculated the first three moments of this distribution (mean, variance, and skew) and the number of blocks over 10^6 base pairs both for each pair of individuals and for the full distribution across all pairwise comparisons.

We then estimated correlation coefficients between spatial distance and each moment of the pairwise IBS tract distribution. Because more closely related individuals on average share longer haplotype blocks we expect that spatial distance will be negatively correlated with mean haplotype block length, and that this correlation will be strongest (i.e., most negative) when dispersal is low. The variance, skew, and count of long haplotype block statistics are meant to reflect the relative length of the right (upper) tail of the distribution, which represents the frequency of long haplotype blocks, and so should reflect recent demographic events (**?**). For a subset of simulations, we also calculated cumulative distributions for IBS tract lengths across pairs of distant (> 48 map units) and nearby (< 2 map units) individuals. Last, we examined the relationship between allele frequency and the spatial dispersion of

an allele by calculating the average distance among individuals carrying each derived allele in a set of simulations representing a range of neighborhood sizes.

The effects of sampling on summary statistic estimates were summarized by testing for differences in mean (ANOVA, (**?**)) and variance (Levene's test, (**?**)) across sampling strategies for each summary statistic.

Last, we used a subset of summary statistics to compare how our continuous model performs relative to a reverse-time stepping stone model (Appendix 1).

Demographic Inference

To assess the impacts of continuous spatial structure on demographic inference we inferred population size histories for all simulations using two approaches: stairwayplot (**?**) and SMC++ (**?**). Stairwayplot fits its model to a genome-wide estimate of the SFS, while SMC++ also incorporates linkage information. For both methods we sampled 20 individuals from all spatial simulations using random, midpoint, and point sampling strategies.

As recommended by its documentation, we used stairwayplot to fit models with multiple bootstrap replicates drawn from empirical genomic data, and took the median inferred N_e per unit time as the best estimate. We calculated site frequency spectra with scikit-allel (**?**), generated 100 bootstrap replicates per simulation by resampling over sites, and fit models for all bootstrap samples using default settings.

For SMC++, we first output genotypes as VCF with msprime and then used SMC++'s standard pipeline for preparing input files assuming no polarization error in the SFS. We used the first individual in the VCF as the "designated individual" when fitting models, and allowed the program to estimate the recombination rate during optimization. We fit models using the 'estimate' command rather than the now recommended cross-validation approach because our simulations had only a single contig.

To evaluate the performance of these methods we binned simulations by neighborhood size, took a rolling median of inferred N_e trajectories across all model fits in a bin for each method and sampling strategy. We also examined how varying levels of isolation by distance impacted the variance of N_e estimates by calculating the standard deviation of N_e from each best-fit model and plotting these against neighborhood size.

Association Studies

To assess the degree to which spatial structure confounds GWAS we simulated four types of nongenetic phenotype variation for 1000 randomly sampled individuals in each spatial SLiM simulation and conducted a linear regression GWAS with principal components as covariates in PLINK (**?**). SNPs with a minor allele frequency less than 0.5% were excluded from this analysis. Phenotype values were set to vary by two standard deviations across the landscape in a rough approximation of the variation seen in height across Europe (**???**). Conceptually our approach is similar to that taken by **?**, though here we model fully continuous spatial variation and compare GWAS output across a range of dispersal distances.

In all simulations, the phenotype of each individual is determined by adding independent Gaussian noise with mean zero to a mean that may depend on spatial position such that the mean of spatially-correlated phenotypes varies by two standard deviations across the landscape. We then adjust the geographic pattern of mean phenotype to create four spatially autocorrelated environmental influences on phenotype. In the first simulation of *nonspatial* environments, the mean did not change, so that all individuals' phenotypes were drawn independently from a Gaussian distribution with mean 110 and standard deviation 10. Next, to simulate *clinal* environmental influences on phenotype, we increased the mean phenotype from 100 on the left edge of the range to 120 on the right edge (two phenotypic standard deviations). Concretely, an individual at position (x, y) in a 50×50 landscape has mean phenotype $100 + 2x/5$. Third, we simulated a more concentrated "*corner*" environmental effect by setting the mean phenotype for individuals with both x and y coordinates below 20 to 120 (two standard deviations above the rest of the map). Finally, in "*patchy*" simulations we selected 10 random points on the map and set the mean phenotype of all individuals within three map units of each of these points to 120.

We performed principal components analysis (PCA) using scikit-allel (**?**) on the matrix of derived allele counts by individual for each simulation. SNPs were first filtered to remove strongly linked sites by calculating LD between all pairs of SNPs in a 200-SNP moving window and dropping one of each pair of sites with an R^2 over 0.1. The LD-pruned allele count matrix was then centered and all sites scaled to unit variance when conducting the PCA, following recommendations in **?**.

We ran linear-model GWAS both with and without the first 10 principal components as covariates in PLINK and summarized results across simulations by counting the number of SNPs with p-value below 0.05 after adjusting for an expected false positive rate of less than 5% (**?**). We also examined p values for systematic inflation by estimating the expected values from a uniform distribution (because no SNPs were used when generating phenotypes), plotting observed against expected values for all simulations, and summarizing across simulations by finding the mean σ value in each region of quantile-quantile space. Results from all analyses were summarized and plotted with the "ggplot2" (**?**) and "cowplot" (**?**) packages in R (**?**).

Results

Demographic Parameters and Run Times

Adjusting the spatial dispersal and interaction distance, σ, has a surprisingly large effect on demographic quantities that are usually fixed in Wright-Fisher models – the generation time, census population size, and variance in offspring number, shown in Figure 2. Because our simulation is paramaterized on an individual level, these population parameters emerge as a property of the interactions among individuals rather than being directly set . Variation across runs occurs because, even though the parameters K and L that control population density and mean lifetime respectively were the same in all simulations, the strength of stochastic effects depends strongly on the spatial interaction distance σ. For instance, the population density near to individual i (denoted n_i above) is computed by averaging over roughly $N_W = 4\pi K\sigma^2$ individuals, and so has standard deviation proportional to $1/\sqrt{N_W}$ – it is more variable at lower densities. (Recall that N_W is Wright's neighborhood size.) Since the probability of survival is a nonlinear function of n_i, actual equilibrium densities and lifetimes differ from K and L. This is the reason that we included *random mating* simulations – where mate choice and offspring dispersal are both nonspatial – since this should preserve the random fluctuations in local population density while destroying any spatial genetic structure. We verified that random mating models retained no geographic signal by showing that summary statistics did not differ significantly between sampling regimes (Table S2), unlike in spatial models (discussed below).

There are a few additional things to note about Figure 2. First, all three quantities are non-monotone with neighborhood size. Census size largely declines as neighborhood size increases for both the spatial and random mating models. However, for spatial models this decline only begins for neighborhood size $\geq$ 10. By a neighborhood sizes larger than 100, the spatial and random mating models are indistinguishable from one another, a sign that our simulations are performing as expected. Census sizes range from $\approx 14,000$ at low σ in the random mating model to $\approx 10,000$ for both models when neighborhood sizes approach 1,000.

Generation time similarly shows complex behavior with respect to neighborhood sizes, and varies between 5.2 and 4.9 timesteps per generation across the parameter range explored. Under both the spatial and random mating models, generation time reaches a minimum at a neighborhood size of around 50. Interestingly, under the range of neighborhood sizes that we examined, generation times between the random mating and spatial models are never quite equivalent – presumably this would cease to be the case at neighborhood sizes higher than we simulated here.

Last, we looked at the variance in number of offspring – a key parameter determining the effective population size. Surprisingly, the spatial and random mating models behave quite differently: while the variance in offspring number increases nearly monotonically under the spatial model, the random mating model actually shows a decline in the variance in offspring number until a neighborhood size ≈ 10 before it increases and eventually equals what we observe in the spatial case.

Run times for our model scale approximately linearly with neighborhood size (Figure 3), with the lowest neighborhood sizes reaching 30N generations in around a day and those with neighborhood

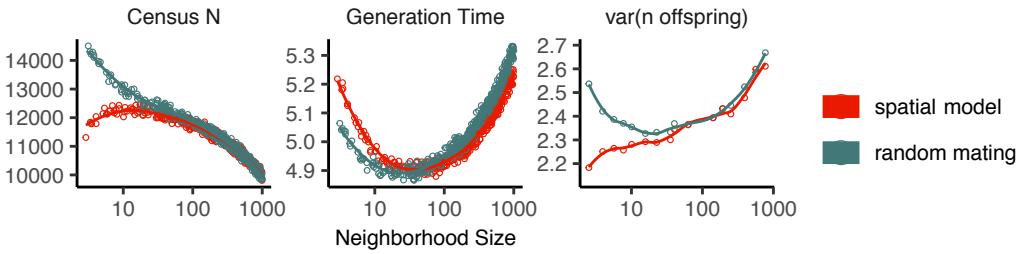

Figure 2 Genealogical parameters from spatial and random mating SLiM simulations, by neighborhood size.

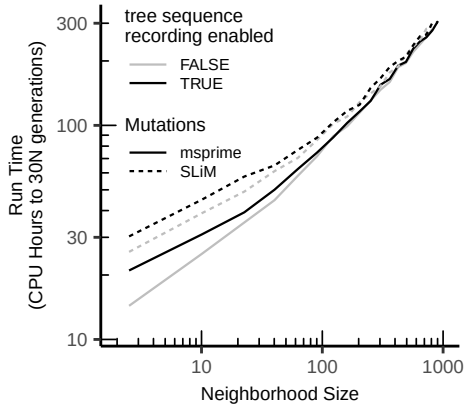

Figure 3 Run times of continuous space simulations with landscape width 50 and expected density 5 under varying neighborhood size. Times are shown for simulations run with mutations applied directly in SLiM (dashed lines) or later applied to tree sequences with msprime (solid lines). Times for simulations run with tree sequence recording disabled are shown in grey.

size approaching 1,000 requiring up to two weeks of computation. At small neighborhood sizes the overhead caused by processing mutations and simplifying tree sequences can add $\approx 30\%$ to the total run time, but above neighborhood sizes of ≈ 100 the effects of neighborhood size dominate because our simulation requires pairwise calculations for sets of individuals within a 3σ radius. As currently implemented running simulations at neighborhood sizes > 1000 to coalescence is likely impractical, though running these models for more limited timescales is possible.

Impacts of Continuous Space on Population Genetic Summary Statistics

Even though certain aspects of population demography depend on the scale of spatial interactions, it still could be that population genetic variation is well-described by a well-mixed population model. Indeed, mathematical results suggest that genetic variation in some spatial models should be well-approximated by a Wright-Fisher population if neighborhood size is large and all samples are geographically widely separated (**??**). However, the behavior of most common population genetic summary statistics other than Tajima's D (**?**) has not yet been described in realistic geographic models. Moreover, as we will show, spatial sampling strategies can affect summaries of genetic variation at least as strongly as the underlying population dynamics .

Site Frequency Spectra and Summaries of Diversity Figure 4 shows the effect of varying neighborhood size and sampling strategy on the site frequency spectrum (Figure 4A) and several standard population genetic summary statistics (Figure 4B). Consistent with findings in island and stepping stone simulations (**?**), the SFS shows a significant enrichment of intermediate frequency variants in comparison to the nonspatial expectation. This bias is most pronounced below neighborhood sizes ≤ 100 and is exacerbated by midpoint and point sampling of individuals (depicted in Figure 1). Reflecting this, Tajima's D is quite positive in the same situations (Figure 4B). Notably, the point at which Tajima's D approaches 0 differs strongly across sampling strategies – varying from a neighborhood size of roughly 50 for random sampling to at least 1000 for midpoint sampling.

One of the most commonly used summaries of variation is Tajima's summary of nucleotide divergence, θ_π, calculated as the mean density of nucleotide differences averaged across pairs of samples. As can be seen in Figure 4B, θ_π in the spatial model is inflated by up to three-fold relative to the random mating model. This pattern is opposite the expectation from census population size (Figure 2), because the spatial model has *lower* census size than the random mating model at neighborhood sizes less than 100. Differences between these models likely occur because θ_π is a measure of mean time to most recent common ancestor between two samples, and at small values of σ, the time for dispersal to mix ancestry across the range exceeds the mean coalescent time under random mating. (For instance, at the smallest value of $\sigma = 0.2$, the range is 250 dispersal distances wide, and since the location of a diffusively moving lineage after k generations has variance $k\sigma^2$, it takes around $250^2 = 62500$ generations to mix across the range, which is roughly ten times larger than the random mating effective population size). θ_π using each sampling strategy approaches the random mating expectation at its own rate, but by a neighborhood size of around 100 all models are roughly equivalent. Interestingly, the effect of sampling strategy is reversed relative to that observed in Tajima's D – midpoint sampling reaches random mating expectations around neighborhood size 50, while random sampling is inflated until around neighborhood size 100.

Values of observed heterozygosity and its derivative F_{IS} also depend heavily on neighborhood size under spatial models as well as the sampling scheme. F_{IS} is inflated above the expectation across most of the parameter space examined and across all sampling strategies. This effect is caused by a deficit of heterozygous individuals in low-dispersal simulations – a continuous-space version of the Wahlund effect (**?**). Indeed, for random sampling under the spatial model, F_{IS} does not approach the random mating equivalent until neighborhood sizes of nearly 1000. On the other hand, the dependency of raw observed heterozygosity on neighborhood size is not monotone. Under midpoint sampling observed heterozygosity is inflated even over the random mating expectation, as a result of the a higher proportion of heterozygotes occurring in the middle of the landscape (Figure S5). This echoes a report from **?** who observed a similar excess of heterozygosity in the middle of the landscape when simulating under a lattice model.

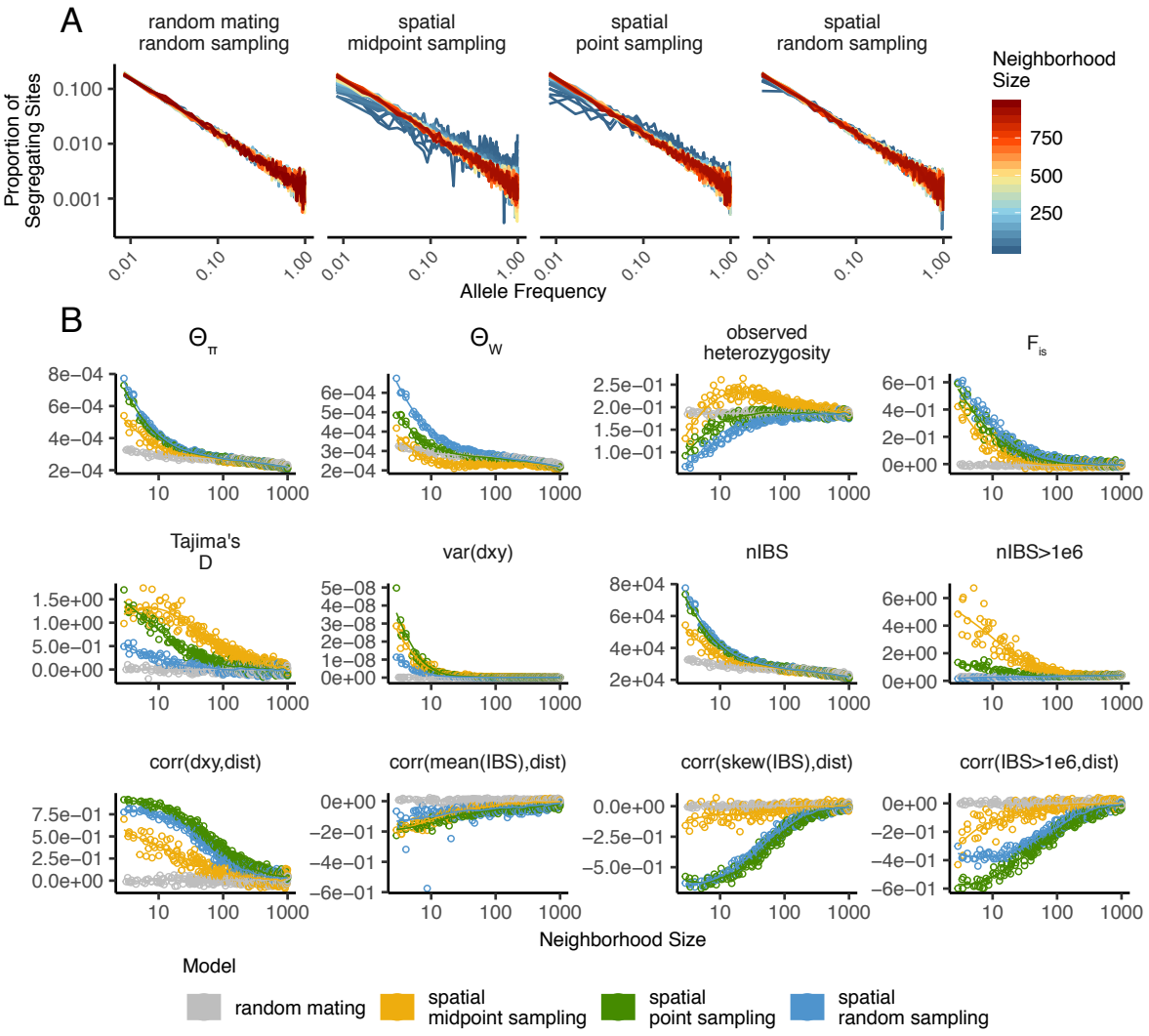

Figure 4 Log-scaled site frequency spectrum (A) and summary statistic distributions (B) by sampling strategy and neighborhood size.

IBS tracts and correlations with geographic distance We next turn our attention to the effect of geographic distance on haplotype block length sharing, summarized for sets of nearby and distant individuals in Figure 5. There are two main patterns to note. First, nearby individuals share more long IBS tracts than distant individuals (as expected because they are on average more closely related). Second, the difference in the number of long IBS tracts between nearby and distant individuals decreases as neighborhood size increases. This reflects the faster spatial mixing of populations with higher dispersal, which breaks down the correlation between the IBS tract length distribution and geographic distance. This can also be seen in the bottom row of Figure 4B, where the correlation coefficients between the summaries of the IBS tract length distribution (the mean, skew, and count of tracts over 10^6bp) and geographic distance approaches 0 as neighborhood size increases.

The patterns observed for correlations of IBS tract lengths with geographic distance are similar to those observed in the more familiar regression of allele frequency measures such as D_{xy} (i.e., "genetic distance") or F_{ST} against geographic distance (**?**). D_{xy} is positively correlated with the geographic distance between the individuals, and the strength of this correlation declines as dispersal increases (Figure 4B), as expected (**??**). This relationship is very similar across random and point sampling strategies, but is weaker for midpoint sampling, perhaps due to a dearth of long-distance comparisons. In much of empirical population genetics a regression of genetic differentiation against spatial distance is a de-facto metric of the significance of isolation by distance. The similar behavior of moments of the pairwise distribution of IBS tract lengths shows why haplotype block sharing has recently emerged as a promising source of information on spatial demography through methods described in **?** and **?**.

Spatial distribution of allele copies Mutations occur in individuals and spread geographically over time. Because low frequency alleles generally represent recent mutations (**??**), the geographic dispersion of an allele may covary along with its frequency in the population. To visualize this relationship we calculated the average distance among individuals carrying a focal derived allele across simulations with varying neighborhood sizes, shown in Figure 6. On average we find that low frequency alleles are the most geographically restricted, and that the extent to which geography and allele frequency are related depends on the amount of dispersal in the population. For populations with large neighborhood sizes we found that even very low frequency alleles can be found across the full landscape, whereas in populations with low neighborhood sizes the relationship between distance among allele copies and their frequency is quite strong. This is the basic process underlying **?**'s (**?**) method for estimating dispersal distances based on the distribution of low frequency alleles, and also generates the greater degree of bias in GWAS effect sizes for low frequency alleles identified in **?**.

Effects of Space on Demographic Inference

One of the most important uses for population genetic data is inferring demographic history of populations. As demonstrated above, the site frequency spectrum and the distribution of IBS tracts varies across neighborhood sizes and sampling strategies. Does this variation lead to different inferences of past population sizes? To ask this we inferred population size histories from samples drawn from our simulated populations with two approaches: stairwayplot (**?**), which uses a genome-wide estimate of the SFS, and SMC++ (**?**), which incorporates information on both the SFS and linkage disequilibrium across the genome.

Figure 7A shows the median inferred population size histories from each method across all simulations, grouped by neighborhood size and sampling strategy. In general these methods tend to slightly overestimate ancient population sizes and infer recent population declines when neighborhood sizes are below 20 and sampling is spatially clustered (Figure 7A, Figure S6). The overestimation of ancient population sizes however is relatively minor, averaging around a two-fold inflation at 10,000 generations before present in the worst-affected bins. For stairwayplot we found that many runs infer dramatic population bottlenecks in the last 1,000 generations when sampling is spatially concentrated, resulting in ten-fold or greater underestimates of recent population sizes. However SMC++ appeared more robust to this error, with runs on point- and midpoint-sampled simulations at the lowest neighborhood sizes underestimating recent population sizes by roughly half and those on randomly sampled simulations showing little error. Above neighborhood sizes of around 100, both

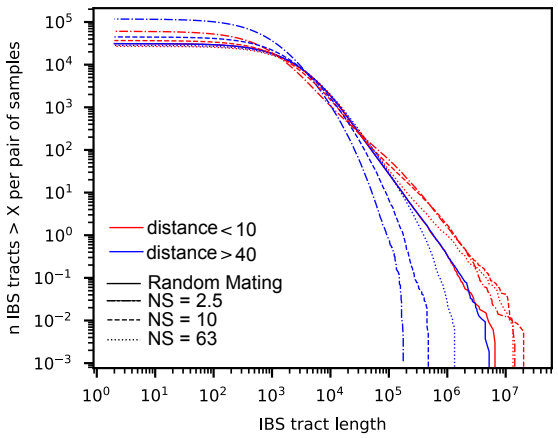

Figure 5 Cumulative distributions for IBS tract lengths per pair of individuals at different geographic distances, across three neighborhood sizes (NS).

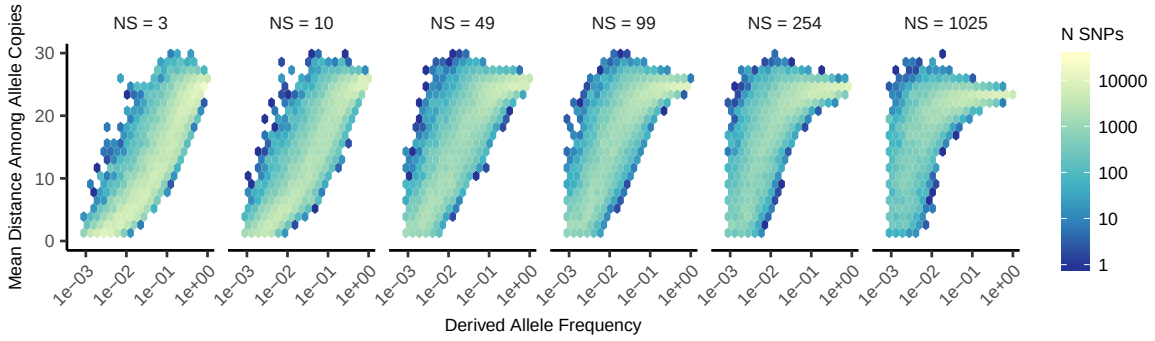

Figure 6 Trends in the distance among allele copies at varying derived allele frequencies and neighborhood sizes (NS).

methods performed relatively well when averaging across results from multiple simulations.

However, individual model fits from both methods frequently reflected turbulent demographic histories (Figure S6), with the standard deviation of inferred N_e across time points often exceeding the expected N_e for both methods (Figure 7B). That is, despite the constant population sizes in our simulations, both methods tended to infer large fluctuations in population size over time, which could potentially result in incorrect biological interpretations. On average the variance of inferred population sizes was elevated at the lowest neighborhood sizes and declines as dispersal increases, with the strongest effects seen in stairwayplot model fits with for clustered sampling and neighborhood sizes less than 20 (Figure 7B).

GWAS

To ask what confounding effects spatial genetic variation might have on genome-wide association studies we performed GWAS on our simulations using phenotypes that were determined solely by the environment – so, any SNP showing statistically significant correlation with phenotype is a false positive. As expected, spatial autocorrelation in the environment causes spurious associations across much of the genome if no correction for genetic relatedness among samples is performed (Figures 8 and S7). This effect is particularly strong for clinal and corner environments, for which the lowest dispersal levels cause over 60% of SNPs in the sample to return significant associations. Patchy environmental distributions, which are less strongly spatially correlated (Figure 8A), cause fewer false positives overall but still produce spurious associations at roughly 10% of sites at the lowest neighborhood sizes. Interestingly we also observed a small number of false positives in roughly 3% of analyses on simulations with nonspatial environments, both with and without PC covariates included in the regression.

The confounding effects of geographic structure are well known, and it is common practice to control for this by including principal components (PCs) as covariates to control for these effects. This mostly works in our simulations – after incorporating the first ten PC axes as covariates, the vast majority of SNPs no longer surpass a significance threshold chosen to have a 5% false discovery rate (FDR). However, a substantial number of SNPs – up to 1.5% at the lowest dispersal distances – still surpass this threshold (and thus would be false positives in a GWAS), especially under "corner" and "patchy" environmental distributions (Figure 8C). At neighborhood sizes larger than 500, up to 0.31% of SNPs were significant for corner and clinal environments. Given an average of 132,000 SNPs across simulations after MAF filtering, this translates to up to 382 false-positive associations; for human-sized genomes, this number would be much larger. In most cases the p values for these associations were significant after FDR correction but would not pass the threshold for significance under the more conservative Bonferroni correction (see example Manhattan plots in figure S7).

Clinal environments cause an interesting pattern in false positives after PC correction: at low neighborhood sizes the correction removes nearly all significant associations, but at neighborhood sizes above roughly 250 the proportion of significant SNPs increases to up to 0.4% (Figure 8). This may be due to a loss of descriptive power of the PCs – as neighborhood size increases, the total proportion of variance explained by the first 10 PC axes declines from roughly 10% to 4% (Figure 8B). Essentially, PCA seems unable to effectively summarize the weak population structure present in large-neighborhood simulations, but these populations continue to have enough spatial structure to create significant correlations between genotypes and the environment. A similar process can also be seen in the corner phenotype distribution, in which the count of significant SNPs initially declines as neighborhood size increases and then increases at approximately the point at which the proportion of variance explained by PCA approaches its minimum.

Figure 8D shows quantile-quantile plots that show the degree of genome-wide inflation of test statistics in PC-corrected GWAS across all simulations and environmental distributions. For clinal environments, $-\log_{10}(p)$ values are most inflated when neighborhood sizes are large, consistent with the pattern observed in the count of significant associations after PC regression. In contrast corner and patchy environments cause the greatest inflation in $-\log_{10}(p)$ at neighborhood sizes less than 100, which likely reflects the inability of PCA to account for fine-scale structure caused by very limited dispersal. Finally, we observed that PC regression appears to overfit to some degree for all phenotype

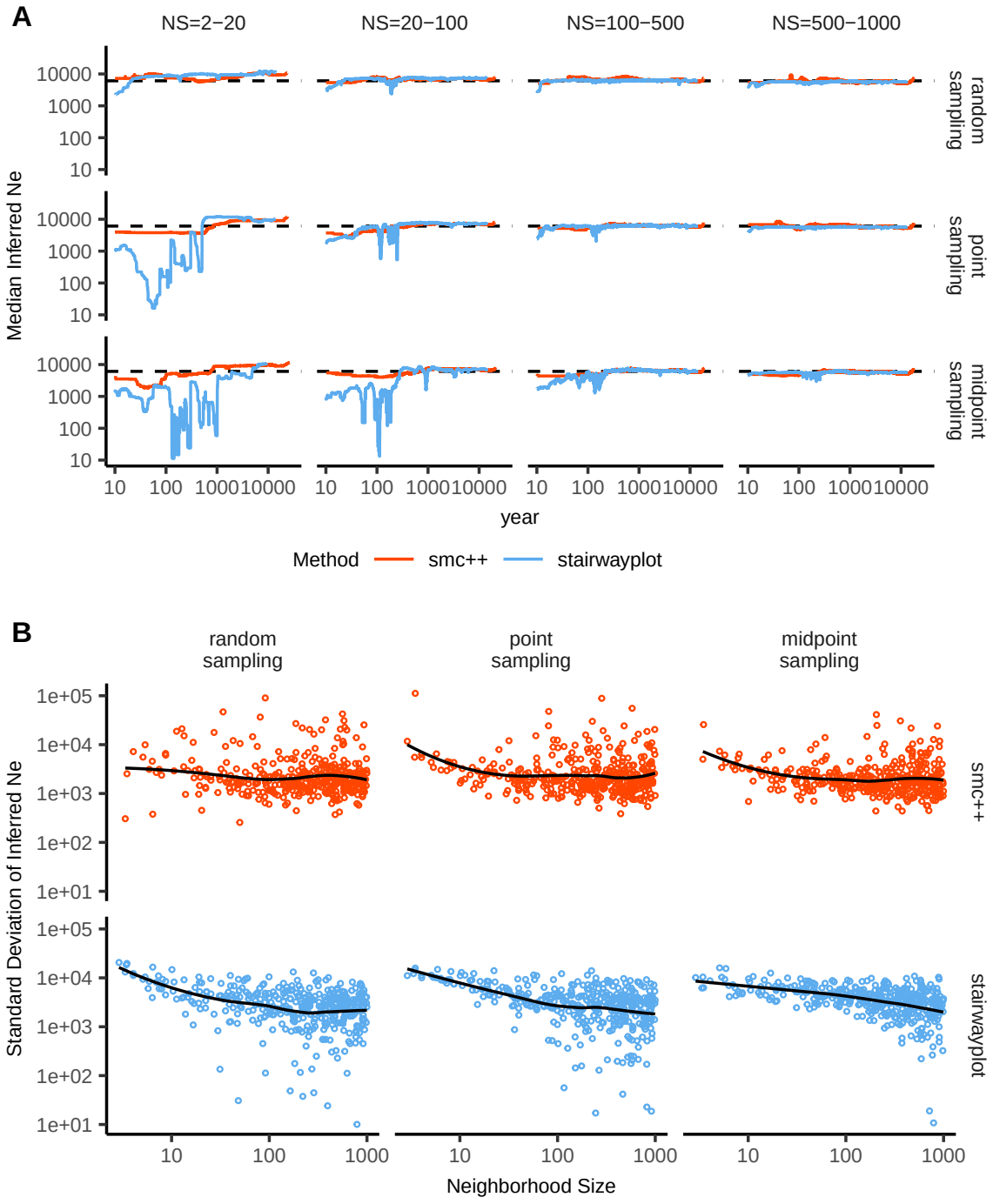

Figure 7 A: Rolling median inferred N_e trajectories for stairwayplot and smc++ across sampling strategies and neighborhood size bins. The dotted line shows the mean N_e of random-mating simulations. B: Standard deviation of individual inferred N_e trajectories, by neighborhood size and sampling strategy. Black lines are loess curves. Plots including individual model fits are shown in Figure S6.

distributions, visible in Figure 8D as points falling below the 1:1 line.

Discussion

In this study, we have used efficient forward time population genetic simulations to describe the myriad influence of continuous geography on genetic variation. In particular, we examine how three main types of downstream empirical inference are affected by unmodeled spatial population structure – 1) population genetic summary statistics, 2) inference of population size history, and 3) genome-wide association studies (GWAS). As discussed above, space often matters (and sometimes dramatically), both because of how samples are arranged in space, and because of the inherent patterns of relatedness established by geography.

Effects of Dispersal

Limited dispersal inflates effective population size, creates correlations between genetic and spatial distances, and introduces strong distortions in the site frequency spectrum that are reflected in a positive Tajima's D (Figure 4). At the lowest dispersal distances, this can increase genetic diversity threefold relative to random-mating expectations. These effects are strongest when neighborhood sizes are below 100, but in combination with the effects of nonrandom sampling they can persist up to neighborhood sizes of at least 1000 (e.g., inflation in Tajima's D and observed heterozygosity under midpoint sampling). If samples are chosen uniformly from across space, the general pattern is similar to expectations of the original analytic model of **?**, which predicts that populations with neighborhood sizes under 100 will differ substantially from random mating, while those above 10,000 will be nearly indistinguishable from panmixia.

The patterns observed in sequence data reflect the effects of space on the underlying genealogy. Nearby individuals coalesce rapidly under limited dispersal and so are connected by short branch lengths, while distant individuals take much longer to coalesce than they would under random mating. Mutation and recombination events in our simulation both occur at a constant rate along branches of the genealogy, so the genetic distance and number of recombination events separating sampled individuals simply gives a noisy picture of the genealogies connecting them. Tip branches (i.e., branches subtending only one individual) are then relatively short, and branches in the middle of the genealogy connecting local groups of individuals relatively long, leading to the biases in the site frequency spectrum shown in Figure 4.

The genealogical patterns introduced by limited dispersal are particularly apparent in the distribution of haplotype block lengths (Figure 4). This is because identical-by-state tract lengths reflect the impacts of two processes acting along the branches of the underlying genealogy – both mutation and recombination – rather than just mutation as is the case when looking at the site frequency spectrum or related summaries. This means that the pairwise distribution of haplotype block lengths carries with it important information about genealogical variation in the population, and correlation coefficients between moments of the this distribution and geographic location contain signal similar to the correlations between F_{ST} or D_{xy} and geographic distance (**?**). Indeed this basic logic underlies two recent studies explicitly estimating dispersal from the distribution of shared haplotype block lengths (**??**). Conversely, because haplotype-based measures of demography are particularly sensitive to variation in the underlying genealogy, inference approaches that assume random mating when analyzing the distribution of shared haplotype block lengths are likely to be strongly affected by spatial processes.

Effects of Sampling

One of the most important differences between random mating and spatial models is the effect of sampling: in a randomly mating population the spatial distribution of sampling effort has no effect on estimates of genetic variation (Table S1) , but when dispersal is limited sampling strategy can compound spatial patterns in the underlying genealogy and create pervasive impacts on all downstream genetic analyses (see also **?**). In most species, the difficulty of traveling through all parts of a species range and the inefficiency of collecting single individuals at each sampling site means that most studies follow something closest to the "point" sampling strategy we simulated, in which multiple individuals are

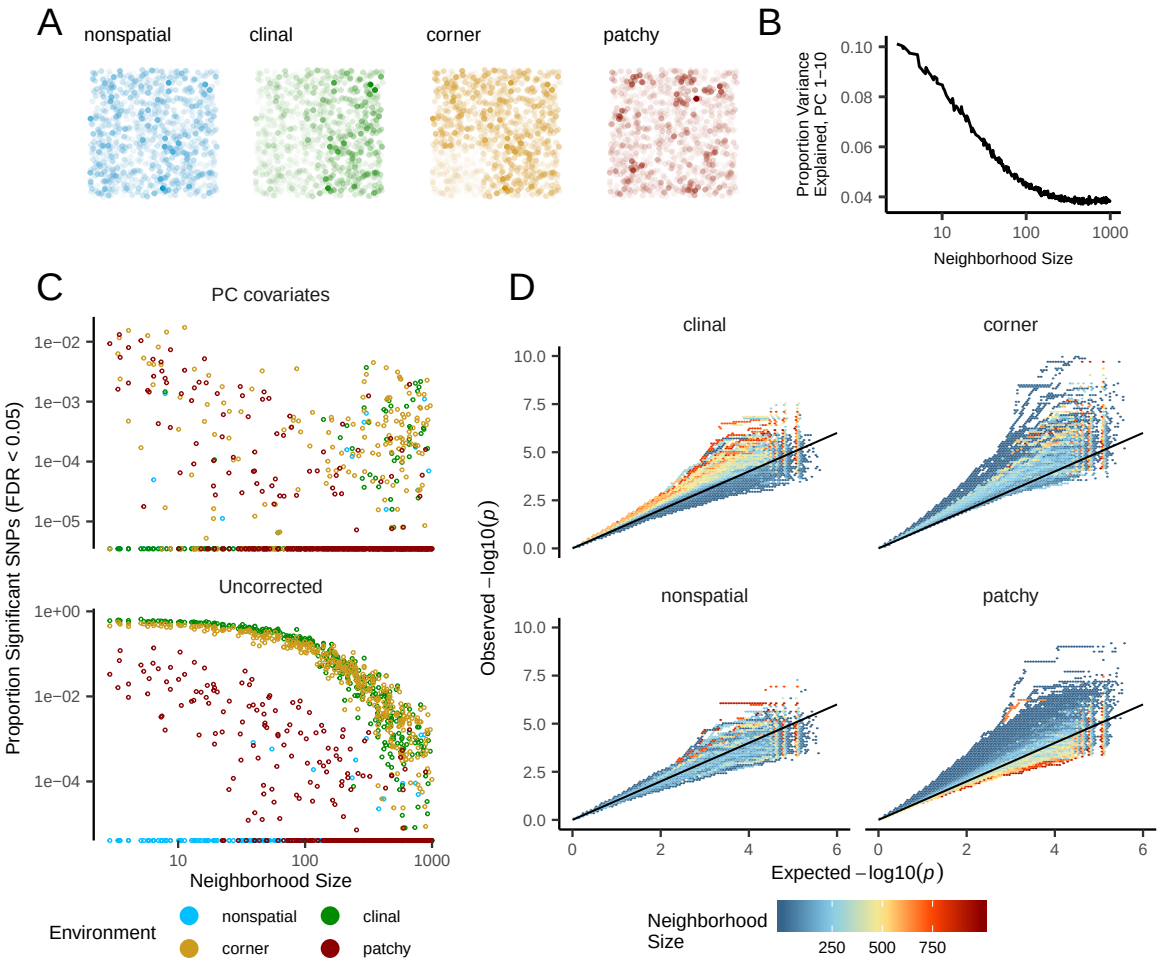

Figure 8 Impacts of spatially varying environments and isolation by distance on linear regression GWAS. Simulated quantitative phenotypes are determined only by an individual's location and the spatial distribution of environmental factors. In **A** we show the phenotypes and locations of sampled individuals under four environmental distributions, with transparency scaled to phenotype. As neighborhood size increases a PCA explains less of the total variation in the data (**B**). Spatially correlated environmental factors cause false positives at a large proportion of SNPs, which is partially but not entirely corrected by adding the first 10 PC coordinates as covariates (**C**). Quantile-quantile plots in **D** show inflation of $-\log_{10}(p)$ after PC correction across all simulations and environmental distributions, with colors scaled by the median neighborhood size in each region of q-q space.

sampled from nearby points on the landscape. For example, in ornithology a sample of 10 individuals per species per locality is a common target when collecting for natural history museums. In classical studies of *Drosophila* variation the situation is considerably worse, in which a single orchard might be extensively sampled.

When sampling is clustered at points on a landscape and dispersal is limited, the sampled individuals will be more closely related than a random set of individuals. Average coalescence times of individuals collected at a locality will then be more recent and branch lengths shorter than expected by analyses assuming random mating. This leads to fewer mutations and recombination events occurring since their last common ancestor, causing a random set of individuals to share longer average IBS tracts and have fewer nucleotide differences. For some data summaries, such as Tajima's D, Watterson's Θ, or the correlation coefficient between spatial distance and the count of long haplotype blocks, this can result in large differences in estimates between random and point sampling (Figure 4). Inferring underlying demographic parameters from these summary statistics – unless the nature of the sampling is somehow taken into account – will be subject to bias if sampling is not random across the landscape.

However, we observed the largest sampling effects using "midpoint" sampling. This model is meant to reflect a bias in sampling effort towards the middle of a species' range. In empirical studies this sampling strategy could arise if, for example, researchers choose to sample the center of the range and avoid range edges to maximize probability of locating individuals during a short field season. Because midpoint sampling provides limited spatial resolution it dramatically reduces the magnitude of observed correlations between spatial and genetic distances. More surprisingly, midpoint sampling also leads to strongly positive Tajima's D and an inflation in the proportion of heterozygous individuals in the sample – similar to the effect of sampling a single deme in an island model as reported in (**?**). This increase in observed heterozygosity appears to reflect the effects of range edges, which are a fundamental facet of spatial genetic variation. If individuals move randomly in a finite two-dimensional landscape then regions in the middle of the landscape receive migrants from all directions while those on the edge receive no migrants from at least one direction. The average number of new mutations moving into the middle of the landscape is then higher than the number moving into regions near the range edge, leading to higher heterozygosity and lower inbreeding coefficients (F_{IS}) away from range edges. Though here we used only a single parameterization of fitness decline at range edges we believe this is a general property of non-infinite landscapes as it has also been observed in previous studies simulating under lattice models (**??**).

In summary, we recommend that empirical researchers collect individuals from across as much of the species' range as practical, choosing samples separated by a range of spatial scales. Many summary statistics are designed for well-mixed populations, and so provide different insights into genetic variation when applied to different subsets of the population. Applied to a cluster of samples, summary statistics based on segregating sites (e.g., Watterson's Θ and Tajima's D), heterozygosity, or the distribution of long haplotype blocks, can be expected to depart significantly from what would be obtained from a wider distribution of samples. Comparing the results of analyses conducted on all individuals versus those limited to single individuals per locality can provide an informative contrast. Finally we wish to point out that the bias towards intermediate allele frequencies that we observe may mean that the importance of linked selection, at least as is gleaned from the site frequency spectrum, may be systematically underestimated currently.

Demography

Previous studies have found that population structure and nonrandom sampling can create spurious signals of population bottlenecks when attempting to infer demographic history with microsatellite variation, summary statistics, or runs of homozygosity (**????**). Here we found that methods that infer detailed population trajectories through time based on the SFS and patterns of LD across the genome are also subject to this bias, with some combinations of dispersal and sampling strategy systematically inferring deep recent population bottlenecks and overestimating ancient N_e by a around a factor of 2. We were surprised to see that both stairwayplot and SMC++ can tolerate relatively strong isolation by distance – i.e., neighborhood sizes of 20 – and still perform well when averaging results across multiple simulations. Inference in populations with neighborhood sizes over 20 was relatively

unbiased unless samples were concentrated in the middle of the range (Figure 7). Although median demography estimates across many independent simulations were fairly accurate, empirical work has only a single estimate to work with, and individual model fits (Figure S6) suggest that spuriously inferred population size changes and bottlenecks are common, especially at small neighborhood sizes. As we will discuss below, most empirical estimates of neighborhood size, including all estimates for human populations, are large enough that population size trajectories inferred by these approaches should not be strongly affected by spatial biases created by dispersal in continuous landscapes. In contrast, **?** found that varying migration rates through time could create strong biases in inferred population trajectories from an *n*-island model with parameters relevant for human history, suggesting that changes in migration rates through time are more likely to drive variation in inferred N_e than isolation by distance.

We found that SMC++ was more robust to the effects of space than stairwayplot, underestimating recent populations by roughly half in the worst time periods rather than nearly 10-fold as with stairwayplot. Though this degree of variation in population size is certainly meaningful in an ecological context, it is relatively minor in population genetic terms. In general methods directly assessing haplotype structure in phased data (for example, **?**) are thought to provide increased resolution for recent demographic events, but in this case the error we observed was essentially an accurate reflection of underlying genealogies in which terminal branches are anomalously short. Combined with our analysis of IBS tract length variation (Figure 5) this suggests that haplotype-based methods are likely to be affected by similar biases.

A more worrying pattern was the high level of variance in inferred N_e trajectories for individual model fits using these methods, which was highest in simulations with the smallest neighborhood size (Figure 7, Figure S6). This suggests that, at a minimum, researchers working with empirical data should replicate analyses multiple times and take a rolling average if model fits are inconsistent across runs. Splitting samples and running replicates on separate subsets – the closest an empirical study can come to our design of averaging the results from multiple simulations – may also alleviate this issue.

Our analysis suggests that many empirical analyses of population size history using methods like SMC++ are robust to error caused by spatial structure within continuous landscapes. Inferences drawn from static SFS-based methods like stairwayplot should be treated with caution when there are signs of isolation by distance in the underlying data (for example, if a regression of F_{ST} against the logarithm of geographic distance has a significantly positive slope), and in particular an inference of population bottlenecks in the last 1000 years should be discounted if sampling is clustered, but estimates of deeper time patterns are likely to be fairly accurate. The biases in the SFS and haplotype structure identified above (see also **???**) are apparently small enough that they fall within the range of variability regularly inferred by these approaches, at least on datasets of the size we simulated.

GWAS

Spatial structure is particularly challenging for genome-wide association studies, because the effects of dispersal on genetic variation are compounded by spatial variation in the environment (**?**). Spatially restricted mate choice and dispersal causes variation in allele frequencies across the range of a species. If environmental factors affecting the phenotype of interest also vary over space, then allele frequencies and environmental exposures will covary over space. In this scenario an uncorrected GWAS will infer genetic associations with a purely environmental phenotype at any site in the genome that is differentiated over space, and the relative degree of bias will be a function of the degree of covariation in allele frequencies and the environment (i.e., Figure 8C, bottom panel). This pattern has been demonstrated in a variety of simulation and empirical contexts (**?????????**).

Incorporating PC positions as covariates in a linear-regression GWAS (**?**) is designed to address this challenge by regressing out a baseline level of "average" differentiation. In essence, a PC-corrected GWAS asks "what regions of the genome are more associated with this phenotype than the average genome-wide association observed across populations?" In our simulations, we observed that this procedure can fail under a variety of circumstances. If dispersal is limited and environmental variation is clustered in space (i.e., corner or patchy distributions in our simulations), PCA positions fail to capture the fine-scale spatial structure required to remove all signals of association. Conversely, as

dispersal increases, PCA loses power to describe population structure before spatial mixing breaks down the relationship between genotype and the environment. These effects were observed with all spatially correlated environmental patterns, but were particularly pronounced if environmental effects are concentrated in one region, as was also found by **?**. Though increasing the number of PC axes used in the analysis may reduce the false-positive rate, this may also decrease the power of the test to detect truly causal alleles (**?**).

In this work we simulated a single chromosome with size roughly comparable to one human chromosome. If we scale the number of false-positive associations identified in our analyses to a GWAS conducted on whole-genome data from humans, we would expect to see several thousand weak false-positive associations after PC corrections in a population with neighborhood sizes up to at least 1000 (which should include values appropriate for many human populations). Notably, very few of the spurious associations we identified would be significant at a conservative Bonferroni-adjusted p-value cutoff (see Figure S7). This suggests that GWAS focused on finding strongly associated alleles for traits controlled by a limited number of variants in the genome are likely robust to the impacts of continuous spatial structure. However, methods that analyze the combined effects of thousands or millions of weakly associated variants such as polygenic risk scores (**?**) are likely to be affected by subtle population structure. Indeed as recently identified in studies of genotype associations for human height in Europe (**??**), PC regression GWAS in modern human populations do include residual signal of population structure in large-scale analyses of polygenic traits. When attempting to make predictions across populations with different environmental exposures, polygenic risk scores affected by population structure can be expected to offer low predictive power, as was shown in a recent study finding lower performance outside European populations (**?**).

In summary, spatial covariation in population structure and the environment confounds the interpretation of GWAS p-values, and correction using principal components is insufficient to fully separate these signals for polygenic traits under a variety of environmental and population parameter regimes. Other GWAS methods such as mixed models (**?**) may be less sensitive to this confounding, but there is no obvious reason that this should be so. One approach to estimating the degree of bias in GWAS caused by population structure is LD score regression (**?**). Though this approach appears to work well in practice, its interpretation is not always straightforward and it is likely biased by the presence of linked selection (**?**). In addition, we observed that in many cases the false-positive SNPs we identified appeared to be concentrated in LD peaks similar to those expected from truly causal sites (Figure S7), which may confound LD score regression.

We suggest a straightforward alternative for species in which the primary axes of population differentiation is space (note this is likely not the case for some modern human populations): run a GWAS with spatial coordinates as phenotypes and check for p-value inflation or significant associations. If significant associations with sample locality are observed after correcting for population structure, the method is sensitive to false positives induced by spatial structure. This is essentially the approach taken in our "clinal" model (though we add normally distributed noise to our phenotypes). This approach has recently been taken with polygenic scores for UK Biobank samples in **?**, finding that scores are correlated with birth location even in this relatively homogenous sample . Of course, it is possible that genotypes indirectly affect individual locations by adjusting organismal fitness and thus habitat selection across spatially varying environments, but we believe that this hypothesis should be tested against a null of stratification bias inflation rather than accepted as true based on GWAS results.

Where are natural populations on this spectrum?

For how much of the tree of life do spatial patterns circumscribe genomic variation? In Table 1 we gathered estimates of neighborhood size from a range of organisms to get an idea of how likely dispersal is to play an important role in patterns of variation. These values should be compared to our simulation results with some caution, as though we expect neighborhood size to have similar effects across species of varying global N_e (**?**), in our study we evaluated only a relatively small population of $\approx 10,000$. In addition, these empirical examples are likely biased towards small-neighborhood species (because few studies have quantified neighborhood size in species with very high dispersal or population density). However, from the available data we find that neighborhood sizes in the range we

Table 1 Neighborhood size estimates from empirical studies.

Species	Description	Neighborhood Size	Method	Citation
Ipomopsis aggregata	flowering plant	12.60 - 37.80	Genetic	(**?**)
Borrichia frutescens	salt marsh plant	20 - 30	Genetic+Survey	(**?**)
Oreamnos americanus	mountain goat	36 - 100	Genetic	(**?**)
Homo sapiens	Gainj- and Kalam-speaking people, Papua New Guinea	40 - 213	Genetic	(**?**)
Formica sp.	colonial ants	50 - 100	Genetic	(**?**)
Astrocaryum mexicanum	palm tree	102 - 895	Genetic+survey	(**?**)
Spermophilus mollis	ground squirrel	204 - 480	Genetic+Survey	(**?**)
Sceloperus olivaceus	lizard	225 - 270	Survey	(**?**)
Dieffenbachia longispatha	beetle-pollinated colonial herb	227 - 611	Survey	(**?**)
Aedes aegypti	Yellow-fever mosquito	268	Genetic	(**?**)
Homo sapiens	Gainj- and Kalam-speaking people, Papua New Guinea	410	Survey	(**?**)
Quercus laevis	Oak tree	> 440	Genetic	(**?**)
Drosophila pseudoobscura	fruit fly	500 - 1,000	Survey+Crosses	(**?**)
Homo sapiens	POPRES data NE Europe	1,342 - 5,425	Genetic	(**?**)
Bebicium vittatum	intertidal snail	240,000	Survey	(**?**)
Bebicium vittatum	intertidal snail	360,000	Genetic	(**?**)

simulated are fairly common across a range of taxa. At the extreme low end of empirical neighborhood size estimates we see some flowering plants, large mammals, and colonial insects like ants. Species such as this have neighborhood size estimates small enough that spatial processes are likely to strongly influence inference. These include some human populations such as the Gainj- and Kalam-speaking people of Papua New Guinea, in which the estimated neighborhood sizes in (**?**) range from 40 to 410 depending on the method of estimation. Many more species occur in a middle range of neighborhood sizes between 100 and 1000 – a range in which spatial processes play a minor role in our analyses under random spatial sampling but are important when sampling of individuals in space is clustered. Surprisingly, even some flying insects with huge census population sizes fall in this group, including fruit flies (*D. melanogaster*) and mosquitoes (*A. aegypti*). Last, many species likely have neighborhood sizes much larger than we simulated, including modern humans in northeastern Europe (**?**). For these species demographic inference and summary statistics are likely to reflect minimal bias from spatial effects as long as dispersal is truly continuous across the landscape. While that is so we caution that association studies in which the effects of population structure are confounded with spatial variation in the environment are still sensitive to dispersal even at these large neighborhood sizes.

Future Directions and Limitations

As we have shown, a large number of population genetic summary statistics contain information about spatial population processes. We imagine that combinations of such summaries might be sufficient for the construction of supervised machine learning regressors (e.g., **?**) for the accurate estimation of dispersal from genetic data. Indeed, **?** found that inverse interpolation on a vector of summary statistics provided a powerful method of estimating dispersal distances. Expanding this approach to include the haplotype-based summary statistics studied here and applying machine learning regressors built for general inference of nonlinear relationships from high-dimensional data may allow precise estimation of spatial parameters under a range of complex models.

One facet of spatial variation that we did not address in this study is the confounding of dispersal and population density implicit in the definition of Wright's neighborhood size. Our simulations were run under constant densities, but **?** and **?** have shown that these parameters are identifiable under some continuous models. Similarly, though the scaling effects of dispersal we show in Figure 4 should occur in populations of any total size, other aspects such as the number of segregating sites are also likely affected by the total landscape size (and so total census N). Much additional work remains to be done to better understand how these parameters interact to shape genetic variation in continuous space, which we leave to future studies.

Though our simulation allows incorporation of realistic demographic and spatial processes, it is inevitably limited by the computational burden of tracking tens or hundreds of thousands of individuals in every generation. In particular, computations required for mate selection and spatial competition scale approximately with the product of the total census size and the neighborhood size and so increase rapidly for large populations and dispersal distances. The reverse-time model of continuous space evolution described by **?** and implemented by **?** allows exploration of parameter regimes with population and landscape sizes more directly comparable to empirical cases like humans. Alternatively, implementation of parallelized calculations may allow progress with forward-time simulations.

Finally, we believe that the difficulties in correcting for population structure in continuous populations using principal components analysis or similar decompositions is a difficult issue, well worth considering on its own. How can we best avoid spurious correlations while correlating genetic and phenotypic variation without underpowering the methods? Perhaps optimistically, we posit that process-driven descriptions of ancestry and/or more generalized unsupervised methods may be able to better account for carry out this task.

Data Availability

Scripts used for all analyses and figures are available at https://github.com/petrelharp/spaceness.

Acknowledgements

We thank Brandon Cooper, Matt Hahn, Doc Edge, and others for reading and thinking about this manuscript. CJB and ADK were supported by NIH award R01GM117241.

Literature Cited

Appendix 1

Comparisons with Stepping-Stone Models

We checked that our model produces reasonable results by comparing it to a reverse-time stepping-stone model implemented in msprime (**?**). In this class of models we imagine an $n \times n$ grid of populations exchanging migrants with neighboring populations at rate m. If these models are good approximations of the continuous case we expect that results will converge as $n \to \infty$, so we ran simulations while varying n from 5 to 50 (Table A1). To compare with continuous models we first distributed the the same "effective" number of individuals across the landscape as in our continuous-space simulations ($\approx$ 6100, estimated from θ_π of random-mating continuous-space simulations). We then approximate the mean per-generation dispersal distance σ given a total landscape width W as the product of the probability of an individual being a migrant and the distance traveled by migrants: $\sigma \approx 4m(W/n)$. We ran 500 simulations per n while sampling σ from $U(0.2, 4)$. We then randomly selected 60 diploid individuals from each simulation (approximating diploidy by combining pairs of chromosomes with contiguous indices within demes) and calculated a set of six summary statistics using the scripts described in the summary statistics portion of the main text.

demes per side (n)	N_e per deme	samples per deme
5	244	20
10	61	10
20	15.25	2
50	2.44	1

Table A1 stepping-stone simulation parameters

In general we find many of the qualitative trends are similar among continuous and stepping-stone models and that, in most cases, statistics from stepping-stone models approach the continuous model as the resolution of the grid increases. For example, θ_π is inflated at low neighborhood sizes (i.e. low m), and the extent of the inflation increases to approach the continuous case as the resolution of the landscape increases. Similar pattterns are observed for F_{is} and observed heterozygosity. However, θ_W behaves differently, with increased grid resolution leading to lower values. This in turn drives an even more positive Tajima's D in grid simulations at small neighborhood sizes.

These differences relative to our continuous model mainly reflect two shortcomings of the reverse-time stepping stone model. If we simulate a coarse grid with relatively large populations in each deme, we cannot accurately capture the dynamics of small neighborhood sizes because mating within each deme remains random regardless of the migration rate connecting demes. This likely explains the trends in π, observed heterozygosity, and F_{is}. However increasing the number of demes while holding the total number of individuals constant results in small within-deme populations for which even the minimum sample size of 1 approaches the local N_e (Table A1). This results in an excess of short terminal branches in the coalescent tree, which decreases the total branch length and leads to fewer segregating sites, deflated θ_W, and inflated Tajima's D. Overall then our continuous model reproduces important features of spatial structure approximated by reverse-time stepping-stone models

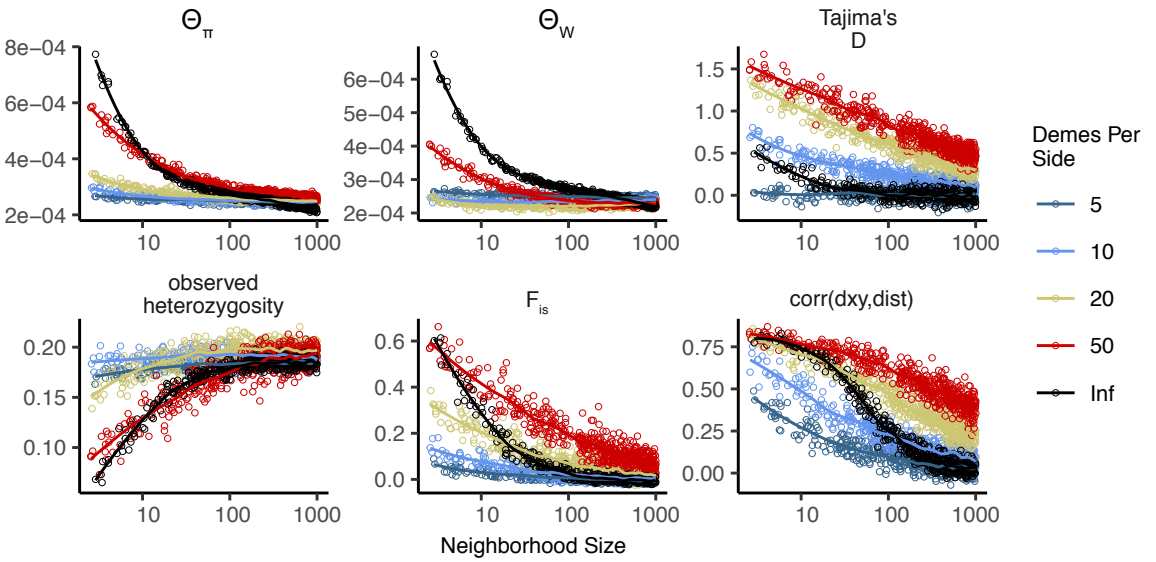

Figure A1 Summary statistics for 2-dimensional coalescent stepping-stone models with fixed total N_e and varying numbers of demes per side. The black "infinite" points are from our forward-time continuous space model. Inter-deme migration rates are related to σ as described above.

at moderate neighborhood sizes while avoiding some artifacts caused binning the landscape into discrete demes.

Supplementary Figures and Tables

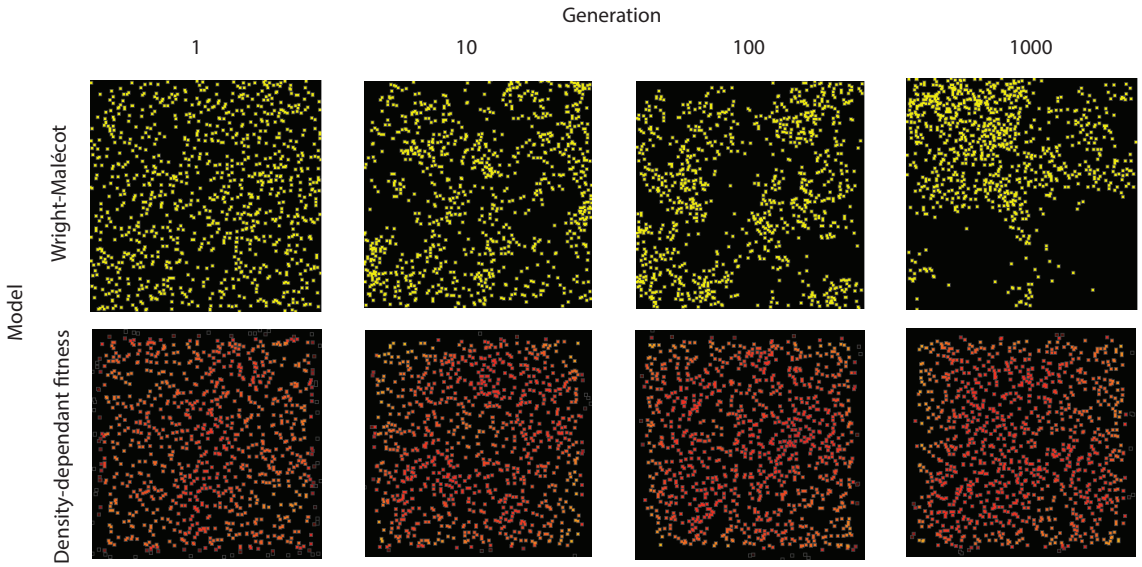

Figure S1 Maps of individual locations in a continuous-space Wright-Malécot model with independent dispersal of all individuals (top) and under our continuous space model incorporating density-dependant fitness (bottom). The clustering seen in the top row is the "Pain in the Torus" described by **?**.

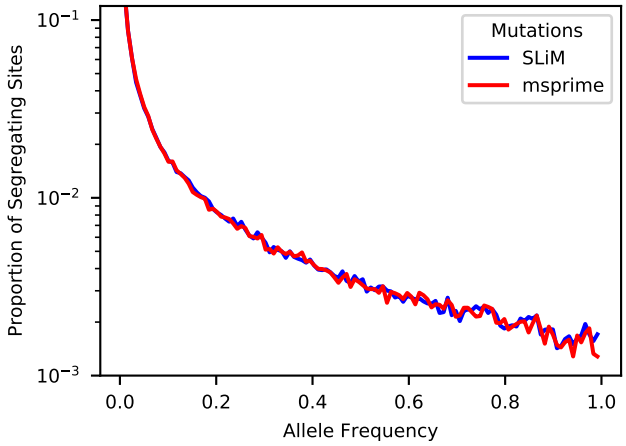

Figure S2 Site frequency spectra from a simulation with neighborhood size = 12.5 when mutations are recorded directly in SLiM (blue line) or applied later in msprime (red line).

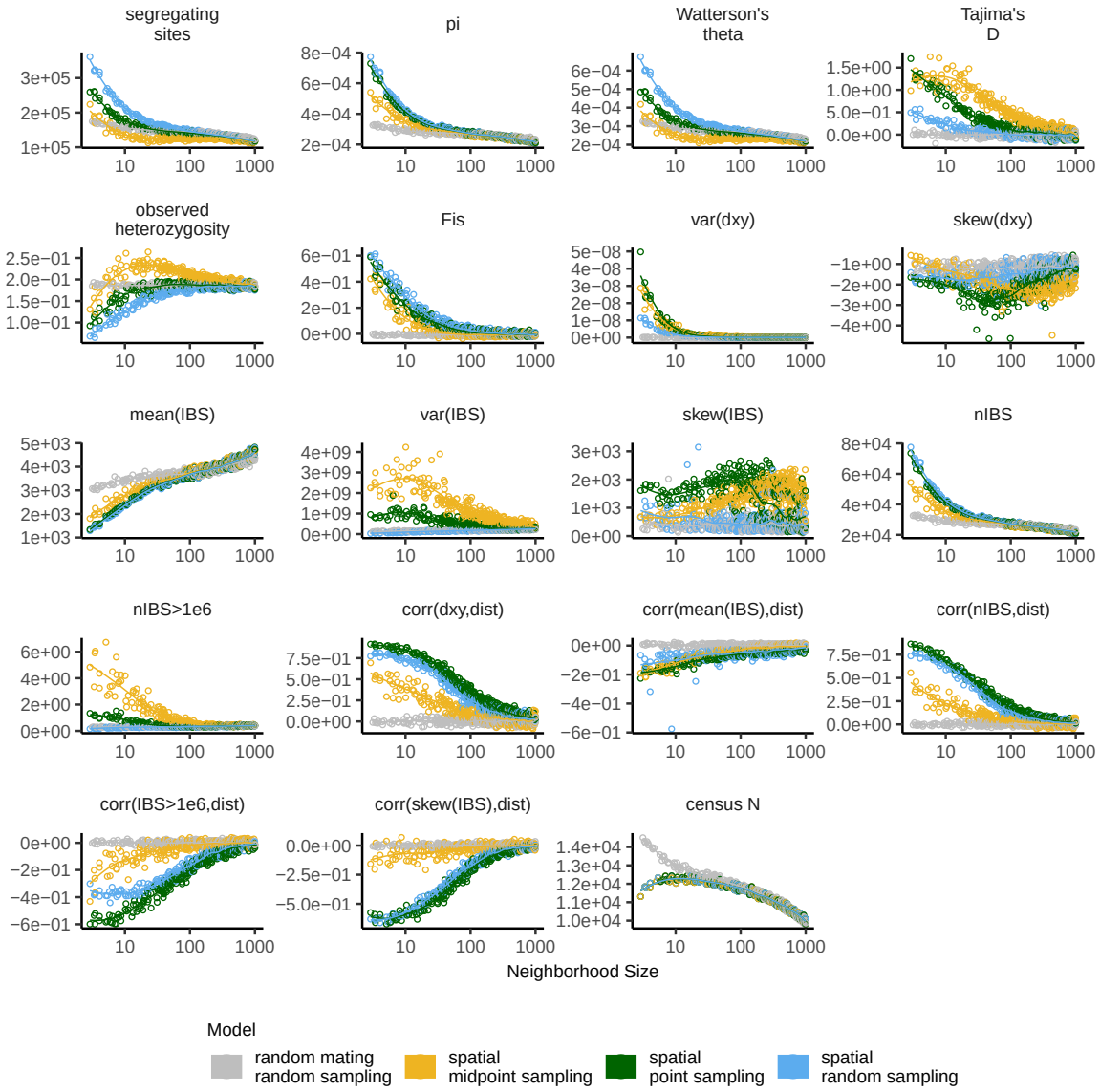

Figure S3 Change in summary statistics by neighborhood size and sampling scheme calculated from simulated sequence data of 60 individuals.

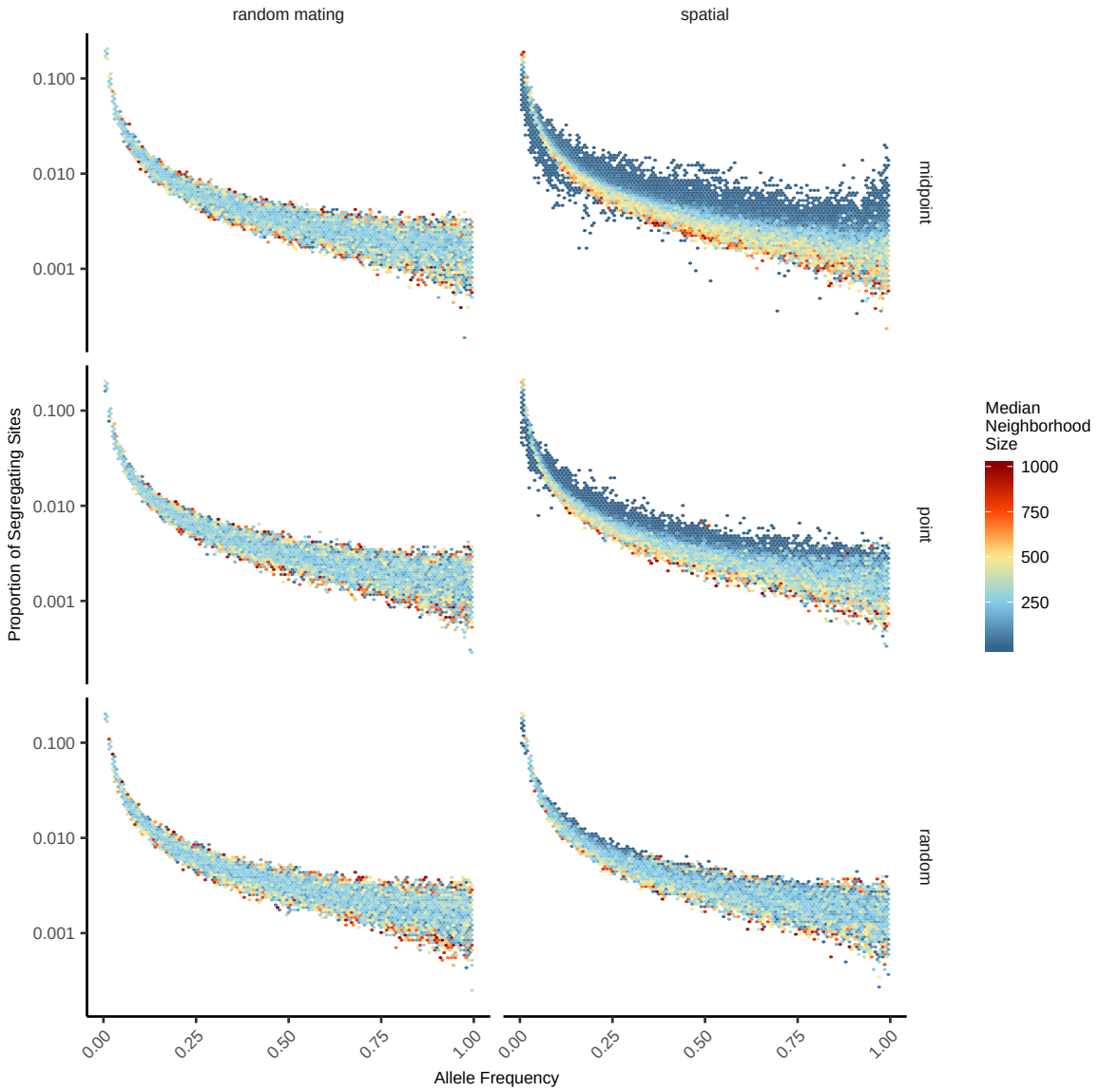

Figure S4 Site frequency spectra for random mating and spatial SLiM models under all sampling schemes.

774

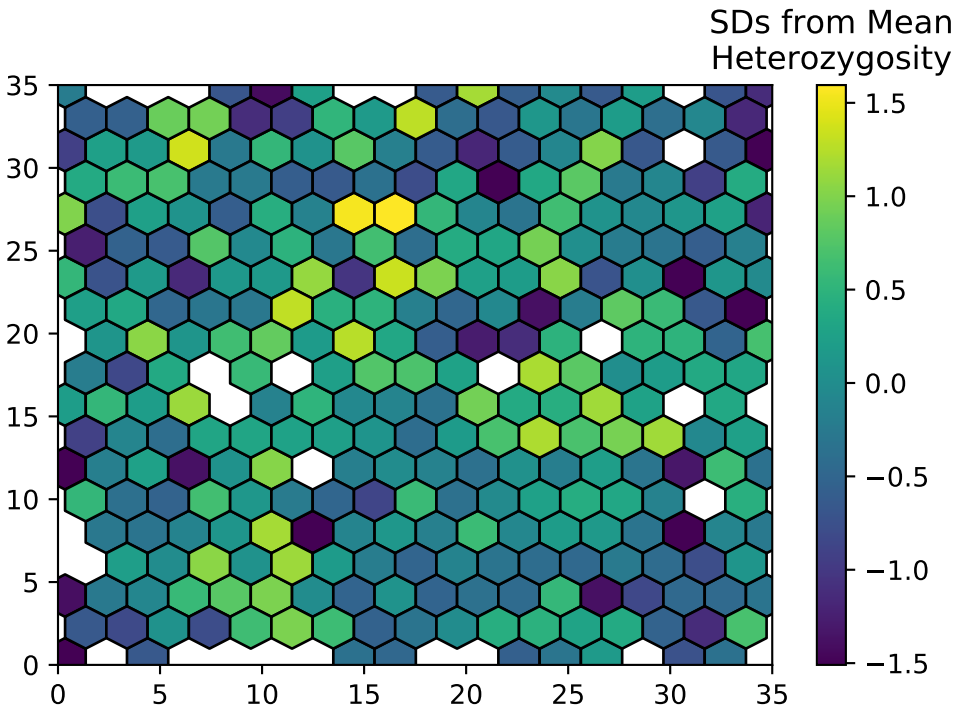

Figure S5 Variation in observed heterozygosity (i.e. proportion of heterozygous individuals) in hexagonal bins across the landscape, estimated from a random sample of 200 individuals from the final generation of a simulation with neighborhood size ≈25. Values were Z-normalized for plotting.

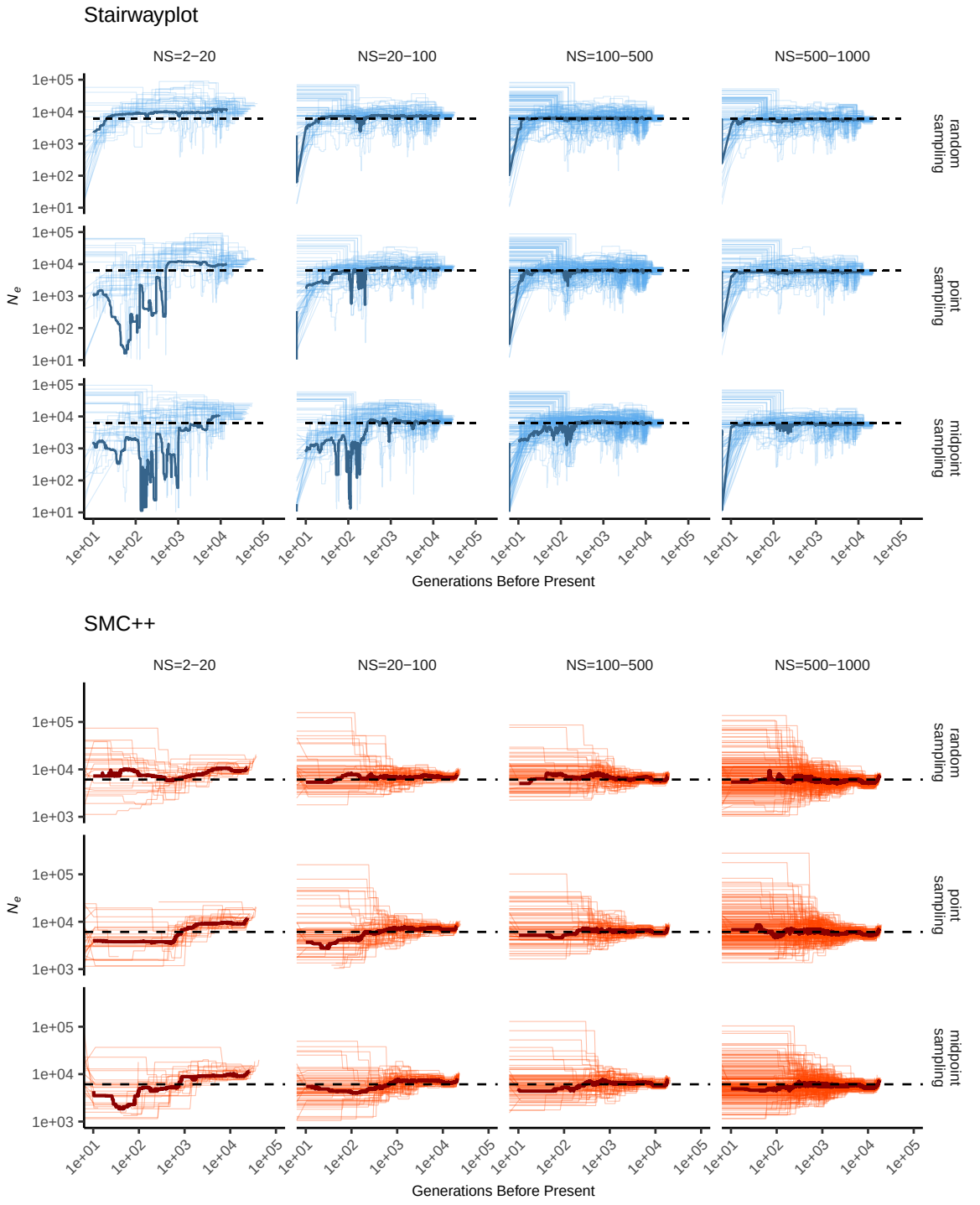

Figure S6 Inferred demographic histories for spatial SLiM simulations, by sampling scheme and neighborhood size (NS) range. Thick lines are rolling medians across all simulations in a bin and thin lines are best fit models for each simulation. Dashed horizontal lines are the average N_e across random-mating SLiM models estimated from θ_π.

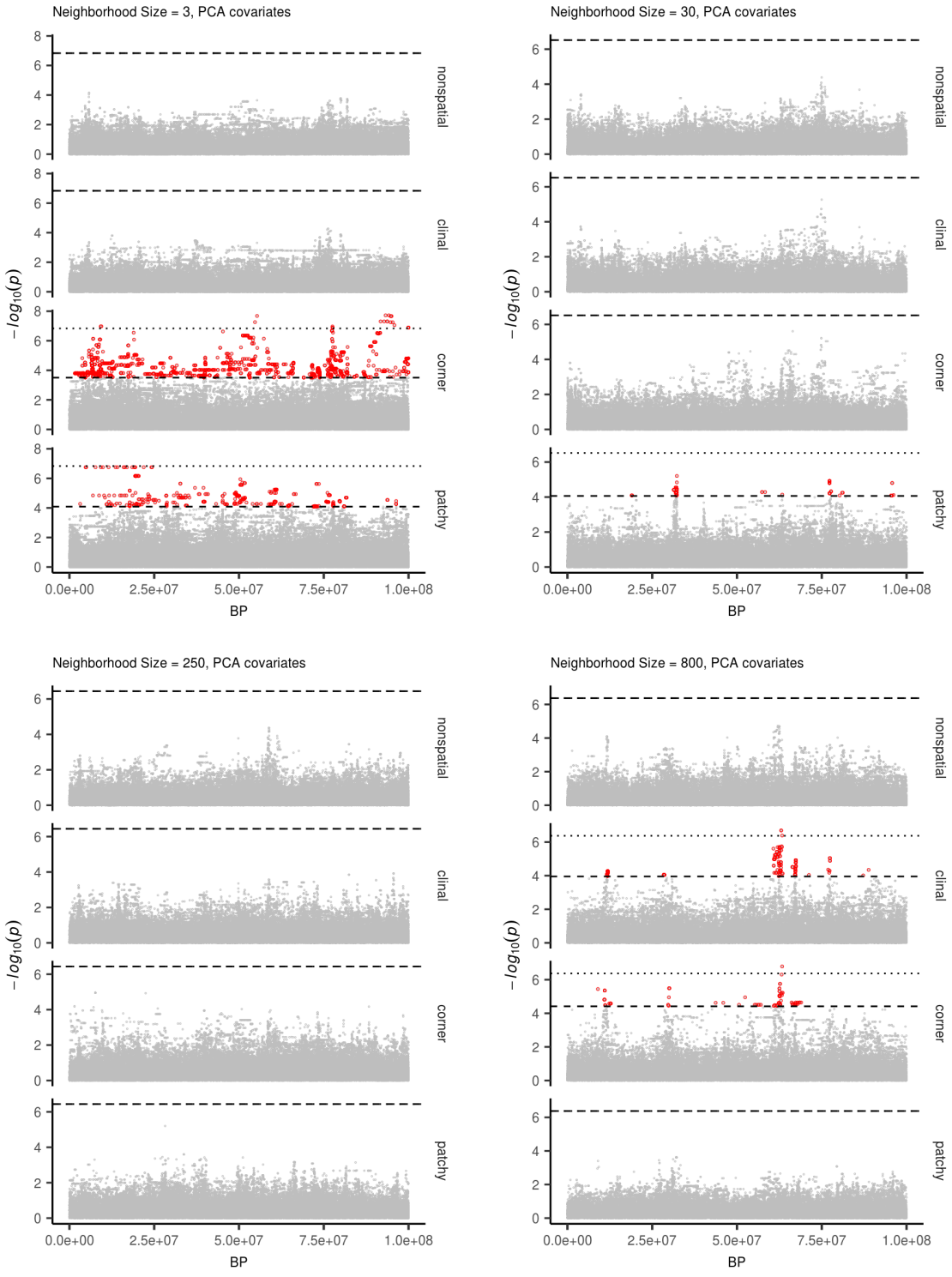

Figure S7 Manhattan plots for a sample of simulations at varying neighborhood sizes. Labels on the right of each plot describes the spatial distribution of environmental factors (described in the methods section of the main text). Points in red are significantly associated with a nongenetic phenotype using a 5% FDR threshold (dashed line). For runs with significant associations the dotted line is a Bonferroni-adjusted cutoff for $p = 0.05$.

Table S1 Summary statistics calculated on simulated genotypes.

Statistic	Description
$\Theta_{/pi}$	Mean of the distribution of pairwise genetic differences
$\Theta_{/W}$	Effective population size based on segregating sites
Segregating Sites	Total number of segregating sites in the sample
Tajima's D	Difference in $\Theta_{/pi}$ and $\Theta_{/W}$ over its standard deviation
Observed Heterzygosity	Proportion of heterozygous individuals in the sample
F_{IS}	Wright's inbreeding coefficient $1 - H_e / H_o$
$var(D_{xy})$	Variance in the distribution of pairwise genetic distances
$skew(D_{xy})$	Skew of the distribution of pairwise genetic distances
$mean(IBS)$	Mean of the distribution of pairwise identical-by-state (IBS) tract lengths taken over all pairs.
$var(IBS)$	Variance of the distribution of pairwise identical-by-state (IBS) tract lengths taken over all pairs.
$skew(IBS)$	Skew of the distribution of pairwise identical-by-state (IBS) tract lengths taken over all pairs.
$nIBS$	Mean number of IBS tracts with length > 2bp across all pairs in the sample.
$nIBS > 1e6$	Mean number of IBS tracts over 1×10^6bp per pair across all pairs in the sample.
$corr(D_{xy}, dist)$	Pearson correlation between genetic distance and $\log_{10}(spatial\ distance)$
$corr(mean(IBS), dist)$	Pearson correlation between the mean of the IBS tract distribution for each pair of samples and $\log_{10}(spatial\ distance)$
$corr(nIBS, dist)$	Pearson correlation between the number of IBS tracts for each pair of samples and $\log_{10}(spatial\ distance)$
$corr(IBS > 1e6, dist)$	Pearson correlation between the number of IBS tracts $> 1 \times 10^6$bp for each pair of samples and $\log_{10}(spatial\ distance)$
$corr(skew(IBS), dist)$	Pearson correlation between the skew of the distribution of pairwise haplotype block lengths for each pair of samples and $\log_{10}(spatial\ distance)$

Table S2 Anova and Levene's test p values for differences by sampling strategy. Bolded values are rejected at $\alpha = 0.05$

variable	model	p(equal means)	p(equal variance)
segsites	random mating	0.998190	0.980730
Θ_π	random mating	0.997750	0.996450
Θ_W	random mating	0.998190	0.980730
Tajima's D	random mating	0.879690	0.188770
observed heterozygosity	random mating	0.531540	0.433230
F_{IS}	random mating	0.474790	0.785730
$mean(D_{xy})$	random mating	0.997770	0.996510
$var(D_{xy})$	random mating	0.283630	0.647240
$skew(D_{xy})$	random mating	0.958320	0.260750
$corr(Dxy, dist)$	random mating	0.601980	**0.000000**
$mean(IBS)$	random mating	0.997960	0.997730
$var(IBS)$	random mating	0.486450	0.399490
$skew(IBS)$	random mating	0.117980	0.069770
$nIBS$	random mating	0.997680	0.996570
$nIBS > 1e6$	random mating	0.834870	0.888730
$corr(mean(IBS), dist)$	random mating	0.073270	0.308420
$corr(IBS > 1e6, dist)$	random mating	0.268440	**0.002100**
$corr(skew(IBS), dist)$	random mating	0.396920	**0.000620**
$corr(nIBS, dist)$	random mating	0.581090	**0.000000**
segsites	spatial	**0.000000**	**0.000000**
Θ_π	spatial	**0.026510**	**0.013440**
Θ_W	spatial	**0.000000**	**0.000000**
Tajima's D	spatial	**0.000000**	**0.000000**
observed heterozygosity	spatial	**0.000000**	**0.000000**
F_{IS}	spatial	**0.000000**	**0.000120**
$mean(D_{xy})$	spatial	**0.025390**	**0.012910**
$var(D_{xy})$	spatial	**0.004970**	**0.006230**
$skew(D_{xy})$	spatial	**0.000000**	**0.000000**
$corr(Dxy, dist)$	spatial	**0.000000**	**0.000000**
$mean(IBS)$	spatial	0.272400	0.114250
$var(IBS)$	spatial	**0.000000**	**0.000000**
$skew(IBS)$	spatial	**0.000000**	**0.000000**
$nIBS$	spatial	**0.033920**	**0.016640**
$nIBS > 1e6$	spatial	**0.000000**	**0.000000**
$corr(mean(IBS), dist)$	spatial	**0.000000**	0.590540
$corr(IBS > 1e6, dist)$	spatial	**0.000000**	**0.000000**
$corr(skew(IBS), dist)$	spatial	**0.000000**	**0.000000**
$corr(nIBS, dist)$	spatial	**0.000000**	**0.000000**

C. J. Battey,
Peter Ralph,
and Andrew Kern

Resubmission Cover Letter
Genetics

Friday 8th November, 2019

To the Editor(s) –

We are writing to submit a revised version of our manuscript, "Space is the Place: Effects of Continuous Spatial Structure on Analysis of Population Genetic Data".

Sincerely,

C. J. Battey, Peter Ralph, and Andrew Kern

Reviewer AE:

> The manuscript admirably explores a lot of consequences of isolation-by-distance in the context of a novel model that is easily amenable to forward simulation; however, given that this model may be used in a lot of future studies based on the precedent set here, there is some concern about the model and its support. Reviewers 2 and 3 highlight this in particular (it underlies the main 2 points of reviewer 2's review, and the core of Reviewer 3's comment), and I agree. Whatever can be done to strengthen the standing of this model, and/or connect it to more thoroughly studied models, will be helpful for the manuscript. The concern would be that there are peculiarities of this model that do not generalize well. A new supplemental section or opener to the results section establishing the model more thoroughly would make the strongest response.

(I would generally cut down the quoted bits like the above to only what's essential, but haven't done that yet.)
(IMPORTANT: don't reorder or delete "points" below - it messes up the automatic numbering!)

(AE.1) Line 35: *Also cite Wilkins and Wakeley, Genetics 2002; Wilkins 2004*

Reply: Done.

(AE.2) (p. 3, l. 124) *"Such models have been used extensively in ecological modeling but rarely in population genetics " Detailing these previous uses via citations and elaboration may help alleviate the major concern about the provenenance of this model and its unique behaviors (see general comments above and R2 and R3 comments).*
(Peter - did you have any in mind here? I can come up with a couple but don't know the ecology lit on this very well) **Reply**:

(AE.3) (p. 4, l. 166) *Please describe computation time needed per replicate*

Reply: We have added a figure (Figure 3) and short discussion (p. 9, l. 331) of run times.

(AE.4) (p. 22, l. 735) *I read the acknowledgement to the Hearth and Creative Sky Brewing with a sense of familiarity in feeling of gratitude to my own favorite cafes and breweries, but I it's not a great precedent for Acknowledgements to be filled this way. Please cut.*

Reply: Good point; we have done this.

(AE.5) Figure 4: *Show random-mating expectation*

Reply: *(working on this)*

(AE.6) Figure 3A, S2: *Perhaps more revealing to show on log-log scale?*

Reply: Good suggestion – the SFS in Fig 3A is now on a log-log scale, which shows the slight decrease in low frequency SNPs a little better.

(AE.7) Figure S3: *Caption seems to be missing detail*

Reply: Thanks for catching this - we have revised this caption to add details including the simulation parameters. (p. 29, l. 773)

Reviewer 1:

We thank the reviewer for their very constructive comments. Responses follow below:

> This study explores biases arising in population-based inference when 1) real population samples are coming from spatial habitat with various degree of structuring while inference is made assuming random mating population; 2) imperfect sampling in practice that fails to represent full diversity across entire population habitat; 3) phenotypes that vary across geography and create spurious associations with genotypes. While earlier studies explored the effect of strong structure on population genetic inference and GWAS, this work focuses on less extreme scenarios of structuring that arises in populations evolving in continuous habitat. By using non-Wright-Fisher model, authors simulated chromosome-scale samples from populations that evolved in continuous space, and that can model environmental factors to create phenotypes varying over space. As a result, this study identified spatial structuring scenarios (small neighbourhood size 10-100) that coupled with imperfect sampling strategies lead to a biased inference of widely used population genetic statistics (altogether 18 statistics) such as pi (average pairwise sequence differences), heterozygosity (and inbreeding coefficient), and IBS tract sharing. Accordingly, inference of the effective population size history was also strongly affected under these parameter ranges. Finally, the authors use their spatial modelling to demonstrate that typical GWAS with PC-based correction cannot entirely remove spurious signals of genotype-phenotype associations arising from purely environmental factors. Overall, the authors explore an important but often neglected source of bias that can affect inference in many population-based studies (in medical genetics, evolutionary biology and ecology). This study can be of interest to a broader audience of readership, and I have only minor comments to improve clarity and increase accessibility for readers:

(1.1) *When neighbourhood size is small (10-100), the mean number of IBS tracts > 2bp (nIBS as in Table S1) is elevated similar to Wright's inbreeding coefficient, but mean of the distribution of pairwise IBS (mean(IBS)) is decreased. What could be the source of this discrepancy? How exactly mean(IBS) was calculated?*

Reply: The mean of the IBS tract distribution is calculated by building a (very large) vector of lengths of all IBS tracts between all pairs of individuals and taking the mean over this entire vector. This distribution is quite strongly skewed (i.e. see figure 4 and the skew(IBs) panel of fig S1) such that the highest density is in low values. We explain the depressed mean value by noting that at low neighborhood sizes the coalescence time of distant (in space) individuals is inflated, which allows more time for recombination and mutation to break up IBS tracts. The flip side of this is that there are then more IBS tracts in total across a random sample of individuals, explaining the corresponding inflation in the number of tracts. We also see an increase in the number of very long IBS tracts at low neighborhood sizes, which corresponds to an excess of very recent coalescent events for groups of nearby individuals. *(need to think of something to say we did about this in the text)*

(1.2) *The authors use K to denote both carrying capacity (p. 4, l. 144) and population density (p. 4, l. 147). It might be better to use a different notation for these quantities since carrying capacity is fixed while density is an emergent quantity in the non-Wright-Fisher model. Use of K to denote carrying capacity and density is a bit confusing. For example, on (p. 7, l. 295) it is said that ' the "population density" (K) and "mean lifetime" (L) parameters were the same in all simulations'. Here K seems to indicate carrying capacity rather than density? The latter is an emergent quantity and varies across simulation runs?*

Reply: We agree that this distinction is worth emphasizing! We've adjusted our language to hopefully remind the reader that K is a parameter that controls population density, rather than being equal to it, at (p. 4, l. 147) and (p. 4, l. 180) and (p. 7, l. 295).

(1.3) *Concerning the non-Wright-Fisher model used, it would be helpful to emphasize that some of the parameters are emergent in contrast to Wright-Fisher model. For example, on Page 11, lines 306-308, the author's goal was to look at census size variation and variation in other quantities. This would be better understood if to emphasize that these parameters are emergent properties in the non-Wright-Fisher model used.*

Reply: We have added to the text at the beginning of the results to emphasize that this analysis is necessary because these parameters are emergent rather than fixed (p. 7, l. 294).

(1.4) *Page 9, line 242, Perhaps 'Demographic Inference' might better reflect the content of this section.*

Reply: Good suggestion – we have changed the section heading to "demographic inference"

(1.5) (p. 6, l. 262) *This sentence with 'Gaussian noise with mean zero and standard deviation 10' is confusing since it was mentioned earlier that the modelled phenotype must vary as human height across Europe, and human height varies 2 standard deviations. Only after reading the whole paragraph it becomes clear that 'standard deviation 10' here referes to unit of height. Please consider rephrasing this sentence.*

Reply: We have revised this sentence to clarify that we aim to produce a variation in mean phenotype of two standard deviations across the landscape (p. 6, l. 263).

(1.6) (p. 7, l. 283) *In the sentence, 'We also examined p values for systemic inflation' I think the authors meant 'systematic inflation'.*

Reply: Whoops; thanks. Fixed.

(1.7) *Page 11, Please correct the legend in Figure 2: must be 'spatial model' and 'random mating' model.*

Reply: Thanks for catching our confusing legend title placement! We have moved "model" to after "spatial" as suggested.

(1.8) *Optional: a dashed line in Figure 2 that shows the total carrying capacity of 50*50*5=12500 would be helpful.*

Reply: This is a good suggestion, but we decided to not include this as we don't have straightforward expectations for the other parameters shown.

(1.9) *Page 13, line 349, The phrase 'affect summaries of variation' is better to replace with 'summaries of genetic variation'.*

Reply: done. (p. 9, l. 340)

(1.10) *Please add or correct references to supplementary figures: For example, Figure S2 was probably meant to accompany Figure 3A, while Figure S1 Figure 3B, but references in the text are absent. In fact, the first reference is made to Figure S3 on page 15.*

Reply: *(check on final pass)*

(1.11) *There are also several typos and errors in the text. For example, on Page 12, lines 309; Page 27, line 655.*

Reply: Thank you for noting these – they have been corrected.

Reviewer 2:

Battey et al. use spatially explicit population genetic simulations to analyze the effects of spatial structure on (i) the estimation of key population genetic parameters, in turn used to (ii) make inferences about population history, and on (iii) confounding in genome-wide association studies (GWAS). I Liked the paper a lot. It's interesting, well-written and addresses an important question - the effect of spatial population structure on population genetic statistics and inference-and I enjoyed reading it. The most positive aspects were:

1. It nice to actually see spatially explicit simulations and I'm happy that forward simulation is now fast enough that you can do this sort of thing.
2. The paper is very clear and well-written, easy to understand the motivation and most of the details. That's not always the case for this sort of paper.
3. I felt that the section about the effect on GWAS was the most interesting and novel part of the paper and gave me some intuition that I hadn't had before.

I don't have any major criticisms. There were a few aspects that I thought might warrant some additional discussion, and a few specific questions below. The general questions I had after reading it were:

(2.1) *To what extent are any of the results dependent on the exact method of simulation. There are a number of choices about the exact details of the simulations (e.g. the way the overlapping generations are handled, the edge effects and, particularly, the form of Equation 1 - see below). It's not so much that these are non-standard (since I don't think there is a standard) and they all sort of make sense heuristically, and I was left wondering whether these sorts of choices actually make a difference. Do the authors have some thoughts/intuition/results about that? Given that the results in Fig. 3 seem quite consistent with expectations, I suspect that on some level it doesn't make much difference but then there are intermediate results like Fig. 2 which seem a bit counter-intuitive and I wonder if those aspects depend on the simulation scheme.*

Reply: This is a good point also raised by other reviewers. We have added an appendix section comparing our model to a reverse-time stepping stone simulation on several relevant summary statistics (p. 22, l. 738). We show that for θ_π and several other statistics the stepping stone model approaches the continuous model as the resolution of the landscape increases. We also see some interesting differences that seem to reflect artifacts from discretization. We are also curious how all the other analyses we test would be affected using other simulation schemes, but hope to explore that aspect in future work.

(2.2) *Related to the first point, to what extent are the results qualitatively different to those that would be obtained in a stepping-stone model? My interpretation is that they are actually very similar, but I didn't see whether that was explicitly discussed. In some sense, it's still easier to do large simulations in a stepping-stone model so it would be nice to be reassured that that's still ok.*

Reply: See above. It is certainly easier and faster, but comes with significant issues either when neighborhood size is lower than the population of a deme, or when demes are small enough that the sample size approaches local N_e, at least in reverse-time simulations.

(2.3) *The source of equation (1) is not obvious to me. I sort of see how it makes sense, but a little but more intuition or a brief derivation or an illuminating either in the main text or the supplement, would be helpful.*

Reply: *(peter - any suggestions here?)*

(2.4) *The authors use a scaling factor in equation (2) to counteract the increase in fitness of individuals at the edges. Can they provide a figure showing that this is the case. What does "roughly" mean on line 164. Perhaps a heatmap of the fitness of individuals across the grid with and without the scaling factor?*

Reply: Good suggestion – we have added a supplemental figure to show the distribution of individual fitness across the landscape with and without our edge-scaling approach. MAKE THIS FIG AND POINT TO IT HERE.

(2.5) *It would be helpful provide the figure showing that generating mutations during the forward simulations in SLIM is equivalent to applying mutations using msprime on pre-generated trees (line 185)? It sounds like this procedure would underestimate the variance in the number of mutations, since you remove the effect of random generation time. Is this effect small?*

Reply: We have added a figure showing sample site-frequency-spectra generated from a subset of simulations run with SLiM mutations, and then using msprime to apply mutations to the same tree sequences with our generation-time scaling approach S2. These approaches yield extremely similar spectra so we believe any error in the approximation is quite small. We also now show variation in simulation times with different mutation and tree sequence recording schemes in Figure 3, and included a brief discussion of this in (p. 9, l. 331)

(2.6) *Can the authors provide a bit more intuition behind the patterns of variation seen in generation time, census population size, and variance in the number of offspring with respect to neighborhood size seen in Figure 2? For example, it is not obvious to me why the census population size, for example, should decline systematically with respect to neighborhood size. Presumably this isn't just due to the local demographic stochasticity. Could the authors briefly interpret the observed patterns or cite appropriate literature?*

Reply:

(2.7) *Fig. 7D: I am surprised by the extent to which the observed values of -log10(p) fall below the y=x line. Particularly in the lower right panel for large neighbourhood sizes. I would expect that to be close to panmictic - why are the P-values underdispersed? That seems like a potential bug, or else something weird is going on.*

Reply: We have checked the code to the best of our abilities and did not find a bug causing the underdispersion. It seems to reflect overcorrection in the regression when using PC coordinates as covariates – the PCA is capturing some information about the spatial genetic variation which itself covaries only weakly with the phenotype, and as a result we see anomalously low $-log10(p)$ when regressing genotype against phenotype. *(Peter or Andy – any advice here?)*

(2.8) *Lines 706-716, It might be worth citing Haworth et al Nature Communications 2019 (https://doi.org/10.1038/s41467-018-08219-1) who do the proposed test (GWAS for birth location) in UK Biobank to illustrate the population structure.*

Reply: Done - thank you for pointing us to this study. (p. 19, l. 671)

(2.9) *The analysis and discussion around the effect of GWAS is focused on PCA correction. Do mixed models help at all?*

Reply: We are also very interested to know how mixed models perform here, but think that adding a second GWAS method would make this section too large for the current paper. We have added a note to (p. 19, l. 657) specifically citing mixed models as alternate methods that may perform better.

(2.10) *The github link to the code didn't work for me. I assume it will be made public later, but at this point I can't tell whether the code is available/useable.*

Reply: Apologies, I had forgotten to make the repo public. The link should work for everyone now.

Reviewer 3:

> The present study deals with a "hot topic" in spatial population genetics. Most inferential and descriptive methods in statistical spatial population genetic rely on a discrete approximation of space and it is not clear what impact this approximation may have when individuals migrate along a continuum instead. Spatial patterns in sampling is also another major issue which is often simply dismissed, mainly because of the paucity of statistical methods to deal with it. This work touches on these important issues in a timely manner.
>
> Although I was enthusiastic about the topic, I was quite disappointed with the core of the study, i.e., the forward-in-time simulation of populations in continuous space. The field has been struggling with this issue for decades – examples of spectacular failures like the Wright-Malecot model (see Felsenstein's "pain in the torus' article, 1975) or, more recently, the "mugration' or "discrete trait analysis' model in phylodynamics (see De Maio et al. 2015) have probably mostly harmed our research field – that one cannot make the economy of using a sound probabilistic model for generating geo-referenced genetic data. It does not seem to be the case here unfortunately.

(3.1) *First, the simulation starts with individuals distributed uniformly at random in space. Is there any indication that the three-step algorithm used here maintains this distribution during the course of evolution? If it does not, then is there any stationary regime and how many generations does one need to wait before reaching it? I do appreciate that the competitive interaction term was introduced in order to avoid seeing the "clumping' of individuals that hampers the Wright-Malecot model. Yet, just because there are no such clusters does not mean that the spatial distribution of individuals reaches a stable regime and that the distribution reached, if any, is reasonable from a biological perspective.*

Reply: This is a good point. We evaluated this while developing the simulation using the built-in visualization tools of the SLiM GUI, and have now added a supplementary figure (Figure **??**) showing the distribution of individuals in our density-dependant spatial model and a continuous-space Wright-Fisher simulation without density dependance.

Second, the demographic process used here involves birth and death of individuals. Does the population survive asymptotically or, like any birth-death process, eventually dies with probability one? In fact, one needs to know a little about the dynamics of the population size to decide whether the corresponding process is reasonable from a biological standpoint.

(3.2)

Reply: Similarly here we monitored the population size over time while running SLiM. All simulations maintained asymptotic populations, and none of the runs we started ended because populations crashed. *(is there something else we should show here?)*

(3.3) *Third, it is not clear what the relationship between the expected lifespan and the probability of survival is. The expected lifespan, L, is first defined as the inverse of the expected number of offspring produced by a parent. The authors also define the probability of survival of a given individual at a given point in space, p_i. Hence, the expected lifespan at a point in space (and time) is the mean of a geometric distribution with parameter p_i, i.e., $1/p_i$. Now, it is far from being obvious what the relationship between these two approaches for defining the expected lifespan actually is.*

Reply: *(Peter – ideas here?)*

(3.4) *Also, the web page https://github.com/petrelharp/spaceness does not seem to exist so that I was not able to experiment with the forward-in-time generator used here unfortunately.*

Reply: I apologize – I had forgotten to set the repo to public. It should work now.

(3.5) *All in all, more efforts need to be made here in my opinion to show that the forward-in-time simulations generate sensible outcomes. Sensible in terms of the behavior of the population demography at equilibrium (provided such equilibrium indeed exists) along with that of the spatial distribution of individuals. The authors could provide some guarantee of the good behavior of their model as evidenced from simulations using a broad range of parameter values for generating data. Alternatively, they could elect to use the spatial-Lambda-Fleming-Viot model for their simulations, which, in my opinion would seem the most sensible option given that (1) it is possible to run backward-in-time simulations under this model, thereby saving a lot of computation time and (2) it is a well-studied model with good mathematical and biological properties and (3) it is implemented in a publicly available software program (https://github.com/jeromekelleher/discsim)*

Reply: This is a good point also raised by other reviewers. We have added an Appendix comparing our model to a stepping-stone, rather than the spatial-Labmda-Fleming-Viot, because we think these are the most familiar and widely used class of spatial models. See Appendix 1. We find that many features of our model are well approximated by stepping-stone models, and that for statistics like θ_π the stepping stone model results approach our continuous space model as the number of demes used to describe the landscape increases.

(3.6) *Figure 2: I do not understand why the neighborhood size varies to the same extent in the random mating model as it does for the spatial model. For the random mating model, I would have expected the neighborhood size to be equal to the census size since all individuals have the same probability of being a parent of any given offspring. From lines 166->171, it is clear that the spatial model would converge to the random mating model when the mean parent-offspring distance tends to infinity only if we were to ignore the impact of range edges. I am thus wondering whether the variation of neighborhood size one observes in Fig 2 for the random mating model is just a consequence of border effects. If that is the case, then the authors should state it clearly and try to justify it from a biological perspective.*

Reply: Census population in the random mating model scales with "neighborhood size" because of two effects – the edge scaling and the competition
(need a better answer here...)

(3.7) *Line 729-731: "Many more species occur in a middle range of neighborhood sizes between 100 and 1000 - a range in which spatial processes play a minor role in our analyses [...]". Do the authors think that the spatial processes would still play a minor role when neighborhood sizes exceed 100-1000 if the habitat was larger than that taken in the present simulations? It would also probably be useful to mention that neighborhood sizes given in Table 1 should be compared with extreme caution since the size of the corresponding habitats vary across species. More generally, I suspect that the size of the habitat has a substantial impact on the vast majority of statistics examined in this study. Indeed, the mean parent-offspring distance, which is at the core of the definition of Wright's neighborhood size, is only small or large relative to the size of the habitat.*

Reply: This is a good point. Wright's work (**?**) suggests some aspects of genetic variation such as variance in allele frequencies and inbreeding coefficients can be estimated by looking only at what he would later (**?**) call "neighborhood size", but certainly other aspects like the number of segregating sits will also depend on total landscape size. We now note on (p. 19, l. 681) that we have evaluated

only one landscape size, and have added a sentence to the discussion noting that exploration of these patterns in varying landscape sizes is an important avenue for further research (p. 21, l. 713).

(3.8) *Line 753-757: please add a reference to Guindon, Guo and Welch (2016). This study clearly shows that population density and dispersal parameters are identifiable and can indeed be estimated in practice under the spatial Lambda-Fleming-Viot model.*

Reply: Done. Thank you for pointing us to this study.

Reviewer 4:

> The manuscript by Battey et al explores the consequence of a well-known violation to population genetic models: the fact that populations are spatially structured and mate along a geographical cline, rather than randomly. This topic is important, particularly in light of recent working describing how spatially correlated genetic and environmental impacts can confound some population genetic insights, such as positive selection for height in Europe. The analyses and investigations presented here are thorough and sensible, and my comments are primarily intended to broaden accessibility for this interesting topic.

(4.1) *Introduction. The discussion is very clear, articulating the three primary goals of the project: the impact of failing to model spatial population structure on 1) population genetic summary statistics, 2) inference on demographic history from population genetic data, and 3) impacts on GWAS summary statistics. I found the discussion a bit easier to follow than the introduction and would suggest streamlining and introducing the topic a bit more. Since the paper follows the flow described in the discussion, it might help orient readers by introducing these topics in the same order.*

Reply: *(working on this)* Thank you for this suggestion. We have slightly revised the introduction and hope it is now clearer; however since we want to cover a little history and motivation for our continuous model vs stepping-stone approaches in the intro it does have a different flow from the discussion.

(4.2) *I agree that most modern work describes structure as discrete populations connected by migration. However, some methods/studies have explicitly modeled spatial structure, e.g. especially in ecology or using methods like dadi (diffusion approximations). Highlighting some examples of previously identified structure not possible to infer without modeling geography would be helpful to contextualize this work.*

Reply: *(working on this)*

(4.3) *There is some reference to spatial models using grids (e.g. Rousset 1997). Some additional discussion contextualizing more recent methods like EEMS that also construct demes and model migration through divergence between neighboring demes would be helpful and interesting.*

Reply: Good point on EEMS. We have added the most recent EEMS paper to the citations on (p. 1, l. 34)

(4.4) *Demographic modeling. Both approaches tested, stairwayplot and SMC++, are most sensitive to older demographic events, and consequently are very noisy and underestimate effect population sizes, especially in smaller neighborhood sizes. Models that consider haplotype structure are much better suited to this time period. It would be helpful to either 1) discuss the varying time sensitivities of different classes of demographic inference methods and how spatial patterns of genetic variation would influence these inferences, or 2) apply a method of this class (many options, e.g. DoRIS, IBDNe, Tracts, Globetrotter, etc) and show how it performs.*

Reply: We now discuss haplotype methods in the relevant discussion section (p. 18, l. 594). However though these methods should be more accurate for recent events it is not clear that this will improve performance per se. The dips in recent inferred Ne from stairwayplot are not just prediction noise, but actually reflect an underlying genealogy in which terminal branches are shorter than expected from a constant-size random-mating population (see e.g. figure 4A and 5). The interpretation error is that these short branches are generated by spatial structure rather than changes in population size over time – a point also made in the (**?**) paper we discuss in the introduction and discussion.

(4.5) *GWAS mixed models. To what extent can spatial signals (e.g. corner, patchy) be corrected with mixed models, e.g. with PCs and PC-adjusted GRM as in Conomos et al, 2016 using PC-AiR and PC-Relate)? Is patchiness related to dispersal? I'm curious how this relates to the predictive ability of GWAS phenotypes with some spatial association that may or may not be associated with environmental effects.*

Reply: *(is this an ok response? two reviewers asked about mixed models but I feel like including them would require expanding the gwas section quite a bit and I'd rather now. But if either of you feel like it's critical I could probably get it done next week)* Good question – we are also interested to know how mixed models perform here, but think that to properly test that we would want to change our design to generate phenotypes from simulated genotypes. This would allow us to evaluate false-negatives in addition to false-positives. This is important because, if mixed models do provide stronger control for stratification they are also likely to remove true signals of causal SNPs if those SNPs covary with spatial structure. We now point to these methods explicitly in the discussion (p. 19, l. 657), but think that incorporating that study here would make this paper too long. We also think the PC results are still quite relevant as the method is still seen in many studies.

(4.6) *Code availability. This github link doesn't work, but is important to be able to evaluate for review: https://github.com/petrelharp/spaceness*

Reply: Apologies, it was accidentally set to private. The link should work now.

(4.7) *Definitions and interpretations. There are quite a large number of metrics discussed in Figure 3B, and it's a lot to take in. It might be helpful to have a table with a reminder of what the metric is, its interpretation, and how it is computed.*

Reply: We have included a table describing the summary statistics in Figure S1.

(4.8) *Notation: "Offspring disperse a Gaussian-distributed distance away from the parent with mean zero and standard deviation σ in both the x and y coordinates. Each offspring is produced with a mate selected randomly from those within distance 3σ, with probability of choosing a neighbor at distance x proportional to $\exp(-x^2/2\sigma^2)$." I think x may be overloaded here, or I'm confused. Clarify?*

Reply: *(Peter and Andy – ideas here?)*

(4.9) *When introducing the "spatial model" as opposed to this "random model," the more concrete illustration in Figure 1 hasn't yet been referenced, which makes it harder to follow. It would be helpful to introduce this figure with the model. Additionally, when Figure 1 is introduced, the order is from right to left (random, then point, then midpoint). It would be helpful to rearrange the figure to mirror what's in the text.*

Reply: We have rearranged the figure as suggested.

(4.10) *Not sure I follow this example: "Concretely, an individual at position (x, y) in a 50 × 50 landscape has mean phenotype 100 + 2x/5."*

Reply:

(4.11) *Minor typo (through vs though): "This occurs because, even through the "population density" (K) and "mean lifetime" (L) parameters..."*

Reply: Thanks, this sentence has been revised and fixed.

(4.12) *Define NS abbreviation in Figure 5.*

Reply: Done.