

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

C.J. Battey^{*,†}, 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. Stepping-stone models reproduce some of these effects, but discretizing the landscape introduces artifacts which in some cases are exacerbated at higher resolutions. 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 revealed surprisingly little systematic bias. 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., ?????) or as an average over such discrete models (??). 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: Tuesday 18th February, 2020

[†]301 Pacific Hall, University of Oregon Dept. Biology, Institute for Ecology and Evolution. cbattey2@uoregon.edu.

[†]these authors co-supervised this project

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

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

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

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

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

⁹³ GWAS may be misleading.

⁹⁴ Materials and Methods

⁹⁵ Modeling Evolution in Continuous Space

⁹⁶ The degree to which genetic relationships are geographically correlated depends on the chance that
⁹⁷ two geographically nearby individuals are close relatives – in modern terms, by the tension between
⁹⁸ migration (the chance that one is descended from a distant location) and coalescence (the chance that
⁹⁹ they share a parent). A key early observation by ? is that this balance is often nicely summarized by
¹⁰⁰ the “neighborhood size”, defined in two dimensions to be $N_W = 4\pi\rho\sigma^2$, where σ^2 is one half of the
¹⁰¹ mean squared parent–offspring distance and ρ is population density (see ? for further discussion of
¹⁰² parameter definitions in one- and two-dimensional habitats). This can be thought of as proportional
¹⁰³ to the average number of potential mates for an individual (those within distance 2σ), or the number
¹⁰⁴ of potential parents of a randomly chosen individual. Empirical estimates of neighborhood size vary
¹⁰⁵ hugely across species – even in human populations, estimates range from 40 to over 5,000 depending
¹⁰⁶ on the population and method of estimation (Table 1).

¹⁰⁷ The first approach to modeling continuously distributed populations was to endow individuals
¹⁰⁸ in a Wright-Fisher model with locations in continuous space. However, since the total size of the
¹⁰⁹ population is constrained, this introduces interactions between arbitrarily distant individuals, which
¹¹⁰ (aside from being implausible) was shown by ? to eventually lead to unrealistic population clumping
¹¹¹ if the range is sufficiently large. Another method for modeling spatial populations is to assume the
¹¹² existence of a grid of discrete randomly mating populations connected by migration, thus enforcing
¹¹³ regular population density by edict. Among many other results drawn from this class of “lattice” or
¹¹⁴ “stepping stone” models (?), ? showed that the slope of the linear regression of genetic differentiation
¹¹⁵ (F_{ST}) against the logarithm of spatial distance is an estimate of neighborhood size. Although these
¹¹⁶ grid models may be good approximations of continuous geography in many situations, they do not
¹¹⁷ model demographic fluctuations, and limit investigation of spatial structure below the level of the
¹¹⁸ deme, assumptions whose impacts are unknown. An alternative method for dealing with continuous
¹¹⁹ geography is a new class of coalescent models, the Spatial Lambda Fleming-Viot models (??).

¹²⁰ To avoid hard-to-evaluate approximations, we here used forward-time, individual-based simulations
¹²¹ across continuous geographical space. The question of what regulates real populations has a long
¹²² history and many answers (e.g., ???), but it is clear that populations must at some point have density-
¹²³ dependent feedback on population size, or else they would face eventual extinction or explosion. In the
¹²⁴ absence of unrealistic global population regulation, this regulation must be local, and there are many
¹²⁵ ways to achieve this (?). In our simulations, each individual’s probability of survival is a decreasing
¹²⁶ function of local population density, which shifts reproductive output towards low-density regions,
¹²⁷ and produces total census sizes that fluctuate around an equilibrium. This also prevents the population
¹²⁸ clumping seen by ? (Supplemental Figure S1)). Such models have been used extensively in ecological
¹²⁹ modeling (?????) but rarely in population genetics, where to our knowledge implementations of
¹³⁰ continuous space models before their availability through SLiM (?) have focused on a small number
¹³¹ of genetic loci (e.g., ?????), which limits the ability to investigate the impacts of continuous space
¹³² on genome-wide genetic variation as is now routinely sampled from real organisms. By simulating
¹³³ chromosome-scale sequence alignments and complete population histories we are able to treat our
¹³⁴ simulations as real populations and replicate the sampling designs and analyses commonly conducted
¹³⁵ on real genomic data.

¹³⁶ A Forward-Time Model of Evolution in Continuous Space

¹³⁷ We simulated populations using the program SLiM v3.1 (?). Each time step consists of three stages:
¹³⁸ reproduction, dispersal, and mortality. To reduce the number of parameters we use the same parameter,
¹³⁹ denoted σ , to modulate the spatial scale of interactions at all three stages by adjusting the standard
¹⁴⁰ deviation of the corresponding Gaussian functions. Informally, we think of σ as the “dispersal distance”,
¹⁴¹ although only one of those stages is dispersal.

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 $1/L$. 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 d proportional to the Gaussian density with mean zero and standard deviation σ , which is $g(d) = \exp(-d^2/2\sigma^2)/(2\sigma^2)$.

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 $g(d)$. 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 ; for more description see the Appendix.

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, though the effect is relatively subtle (Figure S2).

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/kern-lab/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 bp, recombination rate 10^{-9} per bp per generation, 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 verify that this procedure produces the same site frequency spectrum as applying mutations directly in SLiM in Figure S3, in agreement with theory (?). Simulations were run for 1.6 million timesteps (approximately $30N$ generations).

We also compared our model’s output to a commonly-used approximation of continuous space, the stepping-stone model, which we simulated with `msprime` (?). These results are discussed in detail in the Appendix, but in general we find that the demographic structure of a stepping-stone model

188 can depend strongly on the chosen discretization, and some artifacts of discretization seem to become
189 stronger in the limit of a fine grid. For many summary statistics, finer discretizations (we used a 50×50
190 grid) produced similar results to the continuous model, but this was not true for others (e.g., F_{IS} and
191 Tajima's D), which differed from the continuous model *more* at finer discretizations.

192 **Demographic Parameters**

193 Our demographic model includes parameters that control population density (K), mean life span (L),
194 and dispersal distance (σ). However, nonlinearity of local demographic stochasticity causes actual
195 realized averages of these demographic quantities to deviate from the specified values in a way that
196 depends on the neighborhood size. Therefore, to properly compare to theoretical expectations, we
197 empirically calculated these demographic quantities in simulations. We recorded the census population
198 size in all simulations, and used mean population density (ρ , census size divided by total area) to
199 compute neighborhood size as $N_W = 4\pi\rho\sigma^2$. To estimate generation times, we stored ages of the
200 parents of every new individual born across 200 timesteps, after a 100 generation burn-in, and took the
201 mean. To estimate variance in offspring number, we tracked the lifetime total number of offspring for
202 all individuals for 100 timesteps following a 100-timestep burn-in period, and calculated the variance
203 in number of offspring across all individuals in timesteps 50-100. All calculations were performed with
204 information recorded in the tree sequence, using `pyslim` (?).

205 Also note that σ controls the dispersal of offspring away from only *one* parent (e.g., the seed parent
206 for plants), while the usual population genetics definition of dispersal distance involves the distance
207 to a randomly chosen one of the two parents. (thus taking into account the distance from pollen
208 parent as well). This second component – the distance between mates – has in our simulations a
209 distribution that is of order σ but that depends on the population's patchiness. If both between-mate
210 distance and dispersal distance has variance σ^2 along each axis, then the mean squared distance to
211 a randomly chosen parent along that axis would be $(\sigma^2 + 2\sigma^2)/2 = 3\sigma^2/2$. Also note that Wright's
212 definition of neighborhood size should rightly use the *effective* dispersal distance, i.e., the mean
213 squared displacement along an axis between parent-child pairs found moving back along a lineage (?).
214 Nonetheless, we use σ and ρ as defined here to compute N_W because these quantities are more easily
215 observable in practice than their "effective" versions. (Also, note that effective population density is
216 lower than the actual population density, while the σ we use is an underestimate, so our numbers may
217 be close to the "effective" values used in theory.

218 **Sampling**

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

226 **Summary Statistics**

227 We calculated the site frequency spectrum and a set of 18 summary statistics (Table S1) from 60 diploid
228 individuals sampled from the final generation of each simulation using the python package `scikit-allel`
229 (?). Statistics included common single-population summaries including mean pairwise divergence
230 (π), inbreeding coefficient (F_{IS}), and Tajima's D , as well as (motivated by ?'s results) the correlation
231 coefficient between the logarithm of the spatial distance and the proportion of identical base pairs
232 across pairs of individuals.

233 Following recent studies that showed strong signals for dispersal and demography in the dis-
234 tribution of shared haplotype block lengths (e.g., ??), we also calculated various summaries of the
235 distribution of pairwise identical-by-state (IBS) block lengths among sampled chromosomes, defined
236 to be the set of distances between adjacent sites that differ between the two chromosomes. The full

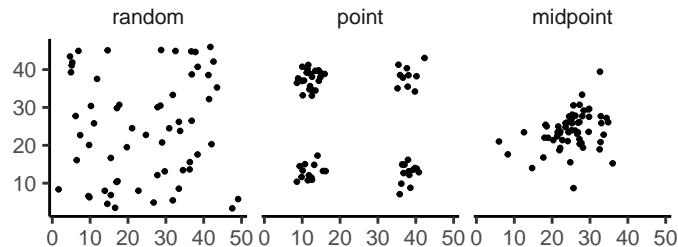


Figure 1 Example sampling maps for 60 individuals on a 50×50 landscape for midpoint, point, and random sampling strategies, respectively.

237 distribution of lengths of IBS tracts for each pair of chromosomes was first calculated with a custom
 238 python function. We then calculated the first three moments of this distribution (mean, variance, and
 239 skew) and the number of blocks over 10^6 base pairs both for each pair of individuals and for the
 240 full distribution across all pairwise comparisons. We then calculated correlation coefficients between
 241 spatial distance and each moment of the pairwise IBS tract distribution. Because more closely related
 242 individuals on average share longer haplotype blocks we expect that spatial distance will be negatively
 243 correlated with mean haplotype block length, and that this correlation will be strongest (i.e., most
 244 negative) when dispersal is low. The variance, skew, and count of long haplotype block statistics are
 245 meant to reflect the relative length of the right (upper) tail of the distribution, which represents the
 246 frequency of long haplotype blocks, and so should reflect recent demographic events (?). For a subset
 247 of simulations, we also calculated cumulative distributions for IBS tract lengths across pairs of distant
 248 (more than 40 map units) and nearby (less than 10 map units) individuals. Last, we examined the
 249 relationship between allele frequency and the spatial dispersion of an allele by calculating the average
 250 distance between individuals carrying each derived allele.

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

254 **Demographic Inference**

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

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

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

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

274 **Association Studies**

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

283 In all simulations, the phenotype of each individual is determined by drawing from a Gaussian
284 distribution with standard deviation 10 and a mean that may depend on spatial position. In spatially
285 varying models, the mean phenotype differs by two standard deviations across the landscape. We
286 then adjust the geographic pattern of mean phenotype to create four types of spatially autocorrelated
287 environmental influences on phenotype. In the first simulation of *nonspatial* environments, the mean
288 did not change, so that all individuals' phenotypes were drawn independently from a Gaussian
289 distribution with mean 110 and standard deviation 10. Next, to simulate *clinal* environmental influences
290 on phenotype, we increased the mean phenotype from 100 on the left edge of the range to 120 on the
291 right edge (two phenotypic standard deviations). Concretely, the mean phenotype p for an individual
292 at position (x, y) is $p = 100 + 2x/5$. Third, we simulated a more concentrated "corner" environmental
293 effect by setting the mean phenotype to 120 for individuals with both x and y coordinates below 20
294 (two standard deviations above the rest of the map). Finally, in "patchy" simulations we selected 10
295 random points on the map and set the mean phenotype of all individuals within three map units of
296 each of these points to 120.

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

302 We ran linear-model GWAS both with and without the first 10 principal components as covariates in
303 PLINK and summarized results across simulations by counting the number of SNPs with p -value below
304 0.05 after adjusting for an expected false positive rate of less than 5% (?). We also examined p values
305 for systematic inflation by comparing to the values expected from a uniform distribution (because no
306 SNPs were used when generating phenotypes, well-calibrated p -values should be uniform).

307 Results from all analyses were summarized and plotted with the "ggplot2" (?) and "cowplot" (?)
308 packages in R (?).

309 **Results**

310 **Demographic Parameters and Run Times**

311 Adjusting the spatial dispersal and interaction distance, σ , has a surprisingly large effect on de-
312 mographic quantities that are usually fixed in Wright-Fisher models – the generation time, census
313 population size, and variance in offspring number, shown in Figure 2. Because our simulation is
314 parameterized on an individual level, these population parameters emerge as a property of the inter-
315 actions among individuals rather than being directly set. Variation across runs occurs because, even
316 though the parameters K and L that control population density and mean lifetime respectively were
317 the same in all simulations, the strength of stochastic effects depends strongly on the spatial interaction
318 distance σ . For instance, the population density near to individual i (denoted n_i above) is computed
319 by averaging over roughly $N_W = 4\pi K\sigma^2$ individuals, and so has standard deviation proportional to
320 $1/\sqrt{N_W}$ – it is more variable at lower densities. (Recall that N_W is Wright's neighborhood size.) Since
321 the probability of survival is a nonlinear function of n_i , actual equilibrium densities and lifetimes differ
322 from K and L . This is the reason that we included *random mating* simulations – where mate choice and
323 offspring dispersal are both nonspatial – since this should preserve the random fluctuations in local
324 population density while destroying any spatial genetic structure. We verified that random mating

325 models retained no geographic signal by showing that summary statistics did not differ significantly
 326 between sampling regimes (Table S2), unlike in spatial models (discussed below).

327 There are a few additional things to note about Figure 2. First, all three quantities are non-monotone
 328 with neighborhood size. Census size largely declines as neighborhood size increases for both the spatial
 329 and random mating models. However, for spatial models this decline only begins for neighborhood
 330 size ≥ 10 . Spatial and random mating models are indistinguishable from one another for neighborhood
 331 sizes larger than 100. Census sizes range from around 14,000 at low σ in the random mating model
 332 to 10,000 for both models when neighborhood sizes approach 1,000. The scaling of census sizes in
 333 both random-mating and spatial models appears to be related to two consequences of the spatial
 334 competition function: the decline of fitness at range edges, which effectively reduces the habitable area
 335 by one σ around the edge of the map and so results in a smaller habitable area at high σ values; and
 336 variation in the equilibrium population density given varying competition radii. Furthermore, census
 337 size increases in spatial models as neighborhood size increases from 2 to 10. This may reflect an Allee
 338 effect (?) in which some individuals are unable to find mates when the mate selection radius is very
 339 small.

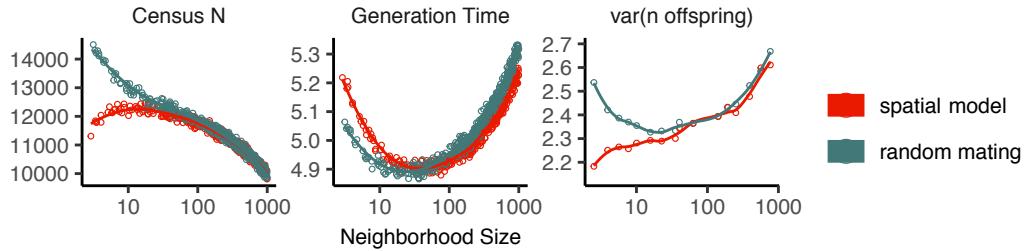


Figure 2 Genealogical parameters from spatial and random mating SLiM simulations, by neighborhood size.

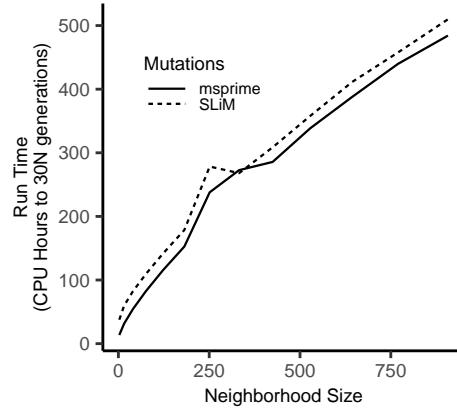


Figure 3 Run times of continuous space simulations with landscape width 50 and expected density 5 under varying neighborhood size. Times are shown for simulations run with mutations applied directly in SLiM (dashed lines) or later applied to tree sequences with msprime (solid lines). Times for simulations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

340 Generation time similarly shows complex behavior with respect to neighborhood sizes, and varies
 341 between 5.2 and 4.9 timesteps per generation across the parameter range explored. Under both the
 342 spatial and random mating models, generation time reaches a minimum at a neighborhood size of
 343 around 50. Interestingly, under the range of neighborhood sizes that we examined, generation times

344 between the random mating and spatial models are never quite equivalent – presumably this would
345 cease to be the case at neighborhood sizes higher than we simulated here.

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

351 Run times for our model scale approximately linearly with neighborhood size (Figure 3), with the
352 lowest neighborhood sizes reaching 30N generations in around a day and those with neighborhood size
353 approaching 1,000 requiring up to three weeks of computation. As currently implemented, running
354 simulations at neighborhood sizes more than 1,000 to coalescence is likely impractical, though running
355 these models for more limited timescales and then “recapitulating” the simulation using reverse-time
356 simulation from the resulting tree sequence in msprime is possible (?).

357 **Impacts of Continuous Space on Population Genetic Summary Statistics**

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

366 **Site Frequency Spectra and Summaries of Diversity** Figure 4 shows the effect of varying neighbor-
367 hood size and sampling strategy on the site frequency spectrum (Figure 4, Figure S5) and several
368 standard population genetic summary statistics (Figure 4B; additional statistics are shown in Figure
369 S4). Consistent with findings in island and stepping stone simulations (?), the SFS shows a significant
370 enrichment of intermediate frequency variants in comparison to the nonspatial expectation. This bias
371 is most pronounced below a neighborhood size of 100 and is exacerbated by midpoint and point
372 sampling of individuals (depicted in Figure 1). Reflecting this, Tajima’s D is quite positive in the same
373 situations (Figure 4B). Notably, the point at which Tajima’s D approaches 0 differs strongly across
374 sampling strategies – varying from a neighborhood size of roughly 50 for random sampling to at least
375 1000 for midpoint sampling.

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

392 Values of observed heterozygosity and its derivative F_{IS} also depend heavily on neighborhood size
393 under spatial models as well as the sampling scheme. F_{IS} is inflated above the expectation across
394 most of the parameter space examined and across all sampling strategies. This effect is caused by a



Figure 4 Site frequency spectrum (A; note axes are log-scaled) and summary statistic distributions (B) by sampling strategy and neighborhood size.

395 deficit of heterozygous individuals in low-dispersal simulations – a continuous-space version of the
396 Wahlund effect (?). Indeed, for random sampling under the spatial model, F_{IS} does not approach the
397 random mating equivalent until neighborhood sizes of nearly 1000. On the other hand, the dependency
398 of raw observed heterozygosity on neighborhood size is not monotone. Under midpoint sampling
399 observed heterozygosity is inflated even over the random mating expectation, as a result of the a
400 higher proportion of heterozygotes occurring in the middle of the landscape (Figure S6). This echoes a
401 report from ? who observed a similar excess of heterozygosity in the middle of the landscape when
402 simulating under a lattice model.

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

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

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

435 **Effects of Space on Demographic Inference**

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

443 Figure 7A shows rolling medians of inferred population size histories from each method across all
444 simulations, grouped by neighborhood size and sampling strategy. In general these methods tend to

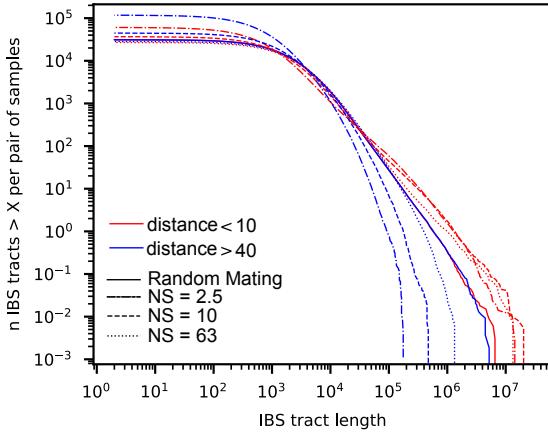


Figure 5 Cumulative distributions for IBS tract lengths per pair of individuals at different geographic distances, across three neighborhood sizes (NS). Nearby pairs (red curves) share many more long IBS tracts than do distant pairs (blue curves), except in the random mating model. The distribution of long IBS tracts between nearby individuals are very similar across neighborhood sizes, but distant individuals are much more likely to share long IBS tracts at high neighborhood size than at low neighborhood size.

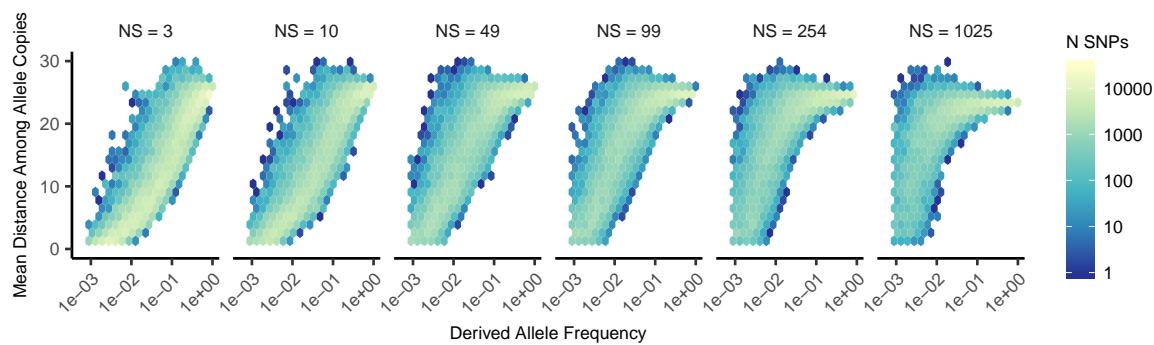


Figure 6 Spatial spread of rare alleles by neighborhood size (NS): Each plot shows the distribution (across derived alleles and simulations) of average pairwise distance between individuals carrying a focal derived allele and derived allele frequency.

445 slightly overestimate ancient population sizes and infer recent population declines when neighborhood
446 sizes are below 20 and sampling is spatially clustered. The overestimation of ancient population sizes
447 however is relatively minor, averaging around a two-fold inflation at 10,000 generations before present
448 in the worst-affected bins. For stairwayplot we found that many runs infer dramatic population
449 bottlenecks in the last 1,000 generations when sampling is spatially concentrated, resulting in ten-fold
450 or greater underestimates of recent population sizes. However SMC++ appeared more robust to
451 this error, with runs on point- and midpoint-sampled simulations at the lowest neighborhood sizes
452 underestimating recent population sizes by roughly half and those on randomly sampled simulations
453 showing little error. Above neighborhood sizes of around 100, both methods performed relatively well
454 when averaging across results from multiple simulations.

455 However, individual simulations were often inferred to have turbulent demographic histories, as
456 shown by the individually inferred histories (shown in Figure S7). Indeed, the standard deviation of
457 inferred N_e across time points (shown in Figure 7B) often exceeds the expected N_e for both methods.
458 That is, despite the nearly constant population sizes in our simulations, both methods tended to infer
459 large fluctuations in population size over time, which could potentially result in incorrect biological
460 interpretations. On average the variance of inferred population sizes was elevated at the lowest
461 neighborhood sizes and declines as dispersal increases, with the strongest effects seen in stairwayplot
462 results with clustered sampling and neighborhood sizes less than 20 (Figure 7B).

463 GWAS

464 To ask what confounding effects spatial genetic variation might have on genome-wide association
465 studies we performed GWAS on our simulations using phenotypes that were determined solely by
466 the environment – so, any SNP showing statistically significant correlation with phenotype is a false
467 positive. As expected, spatial autocorrelation in the environment causes spurious associations across
468 much of the genome if no correction for genetic relatedness among samples is performed (Figures 8 and
469 S8). This effect is particularly strong for clinal and corner environments, for which the lowest dispersal
470 levels cause over 60% of SNPs in the sample to return significant associations. Patchy environmental
471 distributions, which are less strongly spatially correlated (Figure 8A), cause fewer false positives
472 overall but still produce spurious associations at roughly 10% of sites at the lowest neighborhood
473 sizes. Interestingly we also observed a small number of false positives in roughly 3% of analyses
474 on simulations with nonspatial environments, both with and without PC covariates included in the
475 regression.

476 The confounding effects of geographic structure are well known, and it is common practice to
477 control for this by including principal components (PCs) as covariates to control for these effects. This
478 mostly works in our simulations – after incorporating the first ten PC axes as covariates, the vast
479 majority of SNPs no longer surpass a significance threshold chosen to have a 5% false discovery rate
480 (FDR). However, a substantial number of SNPs – up to 1.5% at the lowest dispersal distances – still
481 surpass this threshold (and thus would be false positives in a GWAS), especially under “corner” and
482 “patchy” environmental distributions (Figure 8C). At neighborhood sizes larger than 500, up to 0.31%
483 of SNPs were significant for corner and clinal environments. Given an average of 132,000 SNPs across
484 simulations after MAF filtering, this translates to up to 382 false-positive associations; for human-sized
485 genomes, this number would be much larger. In most cases the p values for these associations were
486 significant after FDR correction but would not pass the threshold for significance under the more
487 conservative Bonferroni correction (see example Manhattan plots in figure S8).

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

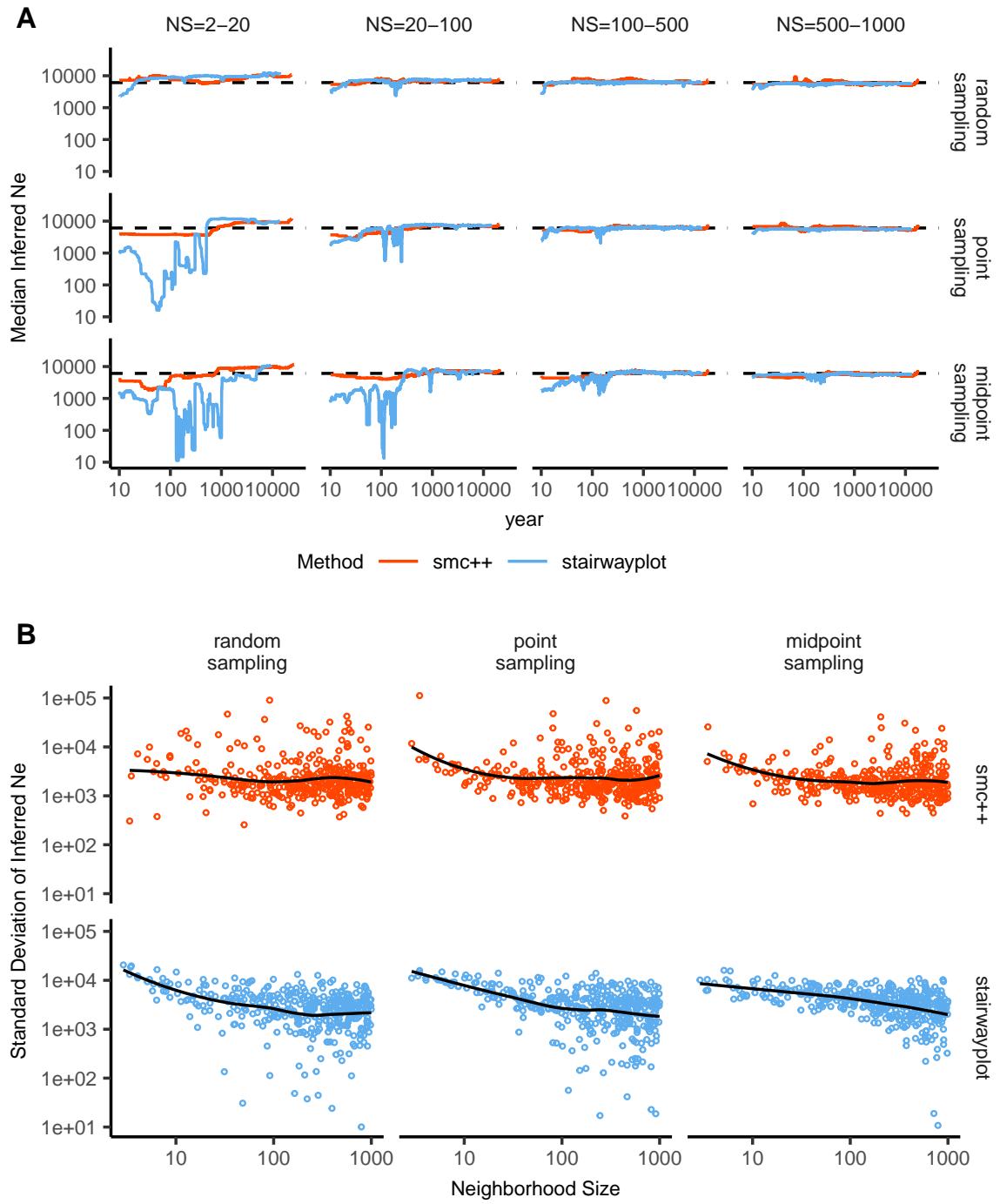


Figure 7 A: Rolling median inferred N_e trajectories for stairwayplot and smc++ across sampling strategies and neighborhood size bins. The dotted line shows the mean N_e of random-mating simulations. B: Standard deviation of individual inferred N_e trajectories, by neighborhood size and sampling strategy. Black lines are loess curves. Plots including individual model fits are shown in Figure S7.

497 declines as neighborhood size increases and then increases at approximately the point at which the
498 proportion of variance explained by PCA approaches its minimum.

499 Figure 8D shows quantile-quantile plots for a subset of simulations that show the degree of genome-
500 wide inflation of test statistics in PC-corrected GWAS across all simulations and environmental distri-
501 butions. An alternate visualization is also included in figure S9. For clinal environments, $-\log_{10}(p)$
502 values are most inflated when neighborhood sizes are large, consistent with the pattern observed in
503 the count of significant associations after PC regression. In contrast corner and patchy environments
504 cause the greatest inflation in $-\log_{10}(p)$ at neighborhood sizes less than 100, which likely reflects
505 the inability of PCA to account for fine-scale structure caused by very limited dispersal. Finally, we
506 observed that PC regression appears to overfit to some degree for all phenotype distributions, visible
507 in Figure 8D as points falling below the 1:1 line.

508 Discussion

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

516 Effects of Dispersal

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

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

536 The genealogical patterns introduced by limited dispersal are particularly apparent in the distribu-
537 tion of haplotype block lengths (Figure 4). This is because identical-by-state tract lengths reflect the
538 impacts of two processes acting along the branches of the underlying genealogy – both mutation and
539 recombination – rather than just mutation as is the case when looking at the site frequency spectrum or
540 related summaries. This means that the pairwise distribution of haplotype block lengths carries with
541 it important information about genealogical variation in the population, and correlation coefficients
542 between moments of the this distribution and geographic location contain signal similar to the corre-
543 lations between F_{ST} or D_{xy} and geographic distance (?). Indeed this basic logic underlies two recent
544 studies explicitly estimating dispersal from the distribution of shared haplotype block lengths (??).
545 Conversely, because haplotype-based measures of demography are particularly sensitive to variation

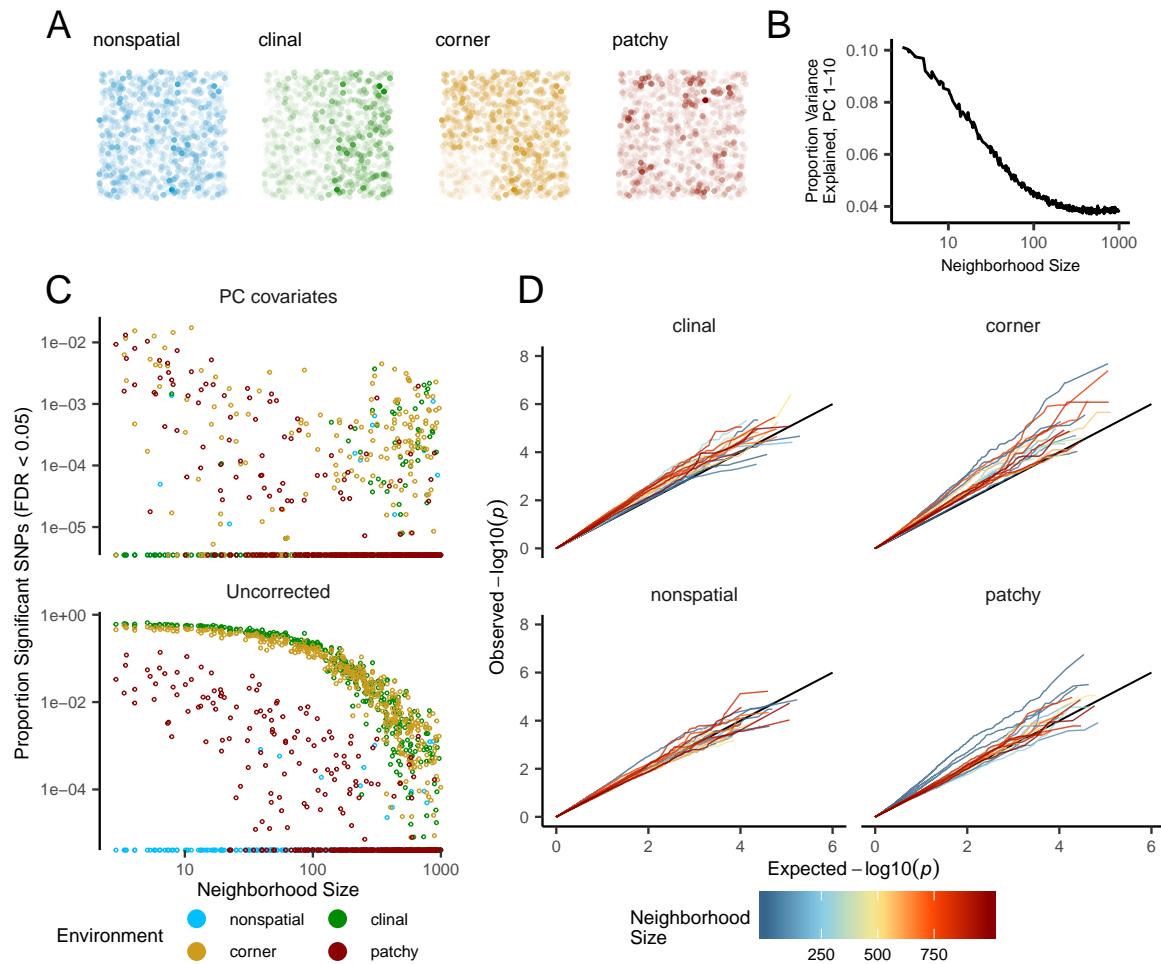


Figure 8 Impacts of spatially varying environments and isolation by distance on linear regression GWAS. Simulated quantitative phenotypes are determined only by an individual's location and the spatial distribution of environmental factors. In **A** we show the phenotypes and locations of sampled individuals 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 for simulations with spatially structured environments, with line colors showing the neighborhood size of each simulation.

546 in the underlying genealogy, inference approaches that assume random mating when analyzing the
547 distribution of shared haplotype block lengths are likely to be strongly affected by spatial processes.

548 **Effects of Sampling**

549 One of the most important differences between random mating and spatial models is the effect of
550 sampling: in a randomly mating population the spatial distribution of sampling effort has no effect on
551 estimates of genetic variation (Table S1), but when dispersal is limited sampling strategy can compound
552 spatial patterns in the underlying genealogy and create pervasive impacts on all downstream genetic
553 analyses (see also ?). In most species, the difficulty of traveling through all parts of a species range and
554 the inefficiency of collecting single individuals at each sampling site means that most studies follow
555 something closest to the “point” sampling strategy we simulated, in which multiple individuals are
556 sampled from nearby points on the landscape. For example, in ornithology a sample of 10 individuals
557 per species per locality is a common target when collecting for natural history museums. In classical
558 studies of *Drosophila* variation the situation is considerably worse, in which a single orchard might be
559 extensively sampled.

560 When sampling is clustered at points on a landscape and dispersal is limited, the sampled indi-
561 viduals will be more closely related than a random set of individuals. Average coalescence times of
562 individuals collected at a locality will then be more recent and branch lengths shorter than expected by
563 analyses assuming random mating. This leads to fewer mutations and recombination events occurring
564 since their last common ancestor, causing a random set of individuals to share longer average IBS tracts
565 and have fewer nucleotide differences. For some data summaries, such as Tajima’s D , Watterson’s
566 θ , or the correlation coefficient between spatial distance and the count of long haplotype blocks, this
567 can result in large differences in estimates between random and point sampling (Figure 4). Inferring
568 underlying demographic parameters from these summary statistics – unless the spatial locations of the
569 sampled individuals are somehow taken into account – will likely be subject to bias.

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

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

598 **Demography**

599 Previous studies have found that population structure and nonrandom sampling can create spurious
600 signals of population bottlenecks when attempting to infer demographic history with microsatellite
601 variation, summary statistics, or runs of homozygosity (????). Here we found that methods that infer
602 detailed population trajectories through time based on the SFS and patterns of LD across the genome
603 are also subject to this bias, with some combinations of dispersal and sampling strategy systematically
604 inferring deep recent population bottlenecks and overestimating ancient N_e by around a factor of 2.
605 We were surprised to see that both stairwayplot and SMC++ can tolerate relatively strong isolation by
606 distance – i.e., neighborhood sizes of 20 – and still perform well when averaging results across multiple
607 simulations. (However, note the high amount of between-simulation variance seen in Figure S7.)
608 Inference in populations with neighborhood sizes over 20 was relatively unbiased unless samples were
609 concentrated in the middle of the range (Figure 7). Although median demography estimates across
610 many independent simulations were fairly accurate, empirical work has only a single estimate to work
611 with, and individual model fits (Figure S7) suggest that spuriously inferred population size changes
612 and bottlenecks are common, especially at small neighborhood sizes. As we will discuss below, most
613 empirical estimates of neighborhood size, including all estimates for human populations, are large
614 enough that population size trajectories inferred by these approaches should not be strongly affected
615 by spatial biases created by dispersal in continuous landscapes. In contrast, ? found that varying
616 migration rates through time could create strong biases in inferred population trajectories from an
617 n -island model with parameters relevant for human history, suggesting that changes in migration rates
618 through time are more likely to drive variation in inferred N_e than isolation by distance.

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

628 A more worrying pattern was the high level of variance in inferred N_e trajectories for individual
629 model fits using these methods, which was highest in simulations with the smallest neighborhood
630 size (Figure 7, Figure S7). This suggests that, at a minimum, researchers working with empirical data
631 should replicate analyses multiple times and take a rolling average if model fits are inconsistent across
632 runs. Splitting samples and running replicates on separate subsets – the closest an empirical study can
633 come to our design of averaging the results from multiple simulations – may also alleviate this issue.

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

643 **GWAS**

644 Spatial structure is particularly challenging for genome-wide association studies, because the effects of
645 dispersal on genetic variation are compounded by spatial variation in the environment (?). Spatially
646 restricted mate choice and dispersal causes variation in allele frequencies across the range of a species.
647 If environmental factors affecting the phenotype of interest also vary over space, then allele frequencies
648 and environmental exposures will covary over space. In this scenario an uncorrected GWAS will

649 infer genetic associations with a purely environmental phenotype at any site in the genome that is
650 differentiated over space, and the relative degree of bias will be a function of the degree of covariation
651 in allele frequencies and the environment (i.e., Figure 8C, bottom panel). This pattern has been
652 demonstrated in a variety of simulation and empirical contexts (?????????).

653 Incorporating PC positions as covariates in a linear-regression GWAS (?) is designed to address this
654 challenge by regressing out a baseline level of “average” differentiation. In essence, a PC-corrected
655 GWAS asks “what regions of the genome are more associated with this phenotype than the average
656 genome-wide association observed across populations?” In our simulations, we observed that this
657 procedure can fail under a variety of circumstances. If dispersal is limited and environmental variation
658 is clustered in space (i.e., corner or patchy distributions in our simulations), PC positions fail to capture
659 the fine-scale spatial structure required to remove all signals of association. Conversely, as dispersal
660 increases, PCA loses power to describe population structure before spatial mixing breaks down the
661 relationship between genotype and the environment. These effects were observed with all spatially
662 correlated environmental patterns, but were particularly pronounced if environmental effects are
663 concentrated in one region, as was also found by ?. Though increasing the number of PC axes used in
664 the analysis may reduce the false-positive rate, this may also decrease the power of the test to detect
665 truly causal alleles (?).

666 In this work we simulated a single chromosome with size roughly comparable to one human
667 chromosome. If we scale the number of false-positive associations identified in our analyses to a
668 GWAS conducted on whole-genome data from humans, we would expect to see several thousand
669 weak false-positive associations after PC corrections in a population with neighborhood sizes up to at
670 least 1000 (which should include values appropriate for many human populations). Notably, very few
671 of the spurious associations we identified would be significant at a conservative Bonferroni-adjusted
672 *p*-value cutoff (see Figure S8). This suggests that GWAS focused on finding strongly associated alleles
673 for traits controlled by a limited number of variants in the genome are likely robust to the impacts of
674 continuous spatial structure. However, methods that analyze the combined effects of thousands or
675 millions of weakly associated variants such as polygenic risk scores (??) are likely to be affected by
676 subtle population structure. Indeed as recently identified in studies of genotype associations for human
677 height in Europe (??), PC regression GWAS in modern human populations do include residual signal
678 of population structure in large-scale analyses of polygenic traits. In addition to error associated with
679 varying patterns of linkage disequilibrium and allele frequency among populations, the confounding
680 of environmental and genetic effects on phenotypes introduced by population structure is expected to
681 lead to low predictive power when polygenic scores are generated for populations outside the original
682 GWAS cohort, as was shown in a recent study finding lower polygenic score predictive power outside
683 European populations (?).

684 In summary, spatial covariation in population structure and the environment confounds the inter-
685 pretation of GWAS *p*-values, and correction using principal components is insufficient to fully separate
686 these signals for polygenic traits under a variety of environmental and population parameter regimes.
687 Other GWAS methods such as mixed models (?) may be less sensitive to this confounding, but there
688 is no obvious reason that this should be so. One approach to estimating the degree of bias in GWAS
689 caused by population structure is LD score regression (?). Though this approach appears to work well
690 in practice, its interpretation is not always straightforward and it is likely biased by the presence of
691 linked selection (?). In addition, we observed that in many cases the false-positive SNPs we identified
692 appeared to be concentrated in LD peaks similar to those expected from truly causal sites (Figure S8),
693 which may confound LD score regression.

694 We suggest a straightforward alternative for species in which the primary axes of population
695 differentiation are associated with geography (note this is likely not the case for some modern human
696 populations): run a GWAS with spatial coordinates as phenotypes and check for *p*-value inflation or
697 significant associations. If significant associations with sample locality are observed after correcting
698 for population structure, the method is sensitive to false positives induced by spatial structure. This
699 is essentially the approach taken in our “clinal” model (though we add normally distributed noise
700 to our phenotypes). This approach has recently been taken with polygenic scores for UK Biobank
701 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.

Table 1 Neighborhood size estimates from empirical studies.

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

706 **Where are natural populations on this spectrum?**

707 For how much of the tree of life do spatial patterns circumscribe genomic variation? In Table 1 we
 708 gathered estimates of neighborhood size from a range of organisms to get an idea of how strongly
 709 local geographic dispersal affects patterns of variation. This is an imperfect measure: some aspects
 710 of genetic variation are most strongly determined by neighborhood size (?), others (e.g., number of
 711 segregating sites) by global N_e , or the ratio of the two. In addition, definitions of "population density"
 712 in genetic versus ecological studies may lead to varying estimates of neighborhood size for a given
 713 species , and these empirical examples may be biased towards small-neighborhood species because few
 714 studies have quantified neighborhood size in species with very high dispersal or population density.

715 However, from the available data we find that neighborhood sizes in the range we simulated are
 716 fairly common across a range of taxa. At the extreme low end of empirical neighborhood size estimates
 717 we see some flowering plants, large mammals, and colonial insects like ants with neighborhood sizes
 718 less than roughly 100. Species such as this have neighborhood size estimates small enough that spatial
 719 processes are likely to strongly influence inference. These include some human populations such as the
 720 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 the recent ancestors of 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.

732 **Other demographic models**

733 Any simulation of a population of reproducing organisms requires some kind of control on population
734 sizes, or else the population will either die out or grow very large after a sufficiently long period of time.
735 The usual choice of population regulation for population genetics – a constant size, as in the Wright–
736 Fisher model – implies biologically unrealistic interactions between geographically distant parts of the
737 species range. Our choice to regulate population size by including a local density-dependent control on
738 mortality is only one of many possible ways to do this. We could have instead regulated fecundity, or
739 recruitment, or both; this general class of models is sometimes referred to as the “Bolker–Pacala model”
740 (?). It is not currently clear how much different choices of demographic parameters, or of functional
741 forms for the regulation, might quantitatively affect our results, although the general predictions
742 should be robust to similar forms of regulation. As is usual in population genetics, the populations
743 are entirely *intrinsically* regulated. Alternatively, population size could be regulated by interactions
744 with other species (e.g., a Lotka–Volterra model), or extrinsically specified by local resource availability
745 (e.g., by food or nest site availability). Indeed, our model could be interpreted as a caricature of such a
746 model: as local density increases, good habitat is increasingly occupied, pushing individuals into more
747 marginal habitat and increasing their mortality. Many such models should behave similarly to ours,
748 but others (especially those with local population cycling), may differ dramatically.

749 Population genetic simulations often use grids of discrete demes, which are assumed to approximate
750 continuous space. However, there are theoretical reasons to expect that increasingly fine grids of
751 discrete demes do not approach the continuous model (?). If continuous space can be approximated by
752 a limit of discrete models, this should be true regardless of the precise details of the discrete model.
753 Although we carefully chose parameters to match our continuous models, we found that some aspects
754 of genetic variation diverged from the continuous case as the discretization got finer. This suggests that
755 these models do not converge in the limit. However, many populations may indeed be well-modeled
756 as a series of discrete, randomly-mating demes if, for example, suitable habitats are patchily distributed
757 across the landscape. There is a clear need for greater exploration of the consequences for population
758 genetics of ecologically realistic population models.

759 **Future Directions and Limitations**

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

768 One facet of spatial variation that we did not address in this study is the confounding of dispersal
769 and population density implicit in the definition of Wright’s neighborhood size. Our simulations
770 were run under constant densities, but ? and ? have shown that these parameters are identifiable
771 under some continuous models. Similarly, though the scaling effects of dispersal we show in Figure 4

772 should occur in populations of any total size, other aspects such as the number of segregating sites
773 are also likely affected by the total landscape size (and so total census size). Indeed, our finding that
774 stepping-stone and continuous-space models match in only certain aspects of genetic variation (Figure
775 A1) shows that qualitatively similar models can produce different results dependent on the specific
776 parameterizations used. While we believe our continuous model is a more appropriate depiction
777 of many species' demographies than lattice models, it is likely that some populations and breeding
778 systems do more closely resemble a series of interconnected random-mating populations. As with
779 all population models, the best approximation for any empirical system will depend on the natural
780 history of the species in question. Much additional work remains to be done to better understand how
781 life history, range size, and habitat geometry interact to shape genetic variation in continuous space,
782 which we leave to future studies.

783 Though our simulation allows incorporation of realistic demographic and spatial processes, it
784 is inevitably limited by the computational burden of tracking tens or hundreds of thousands of
785 individuals in every generation. In particular, computations required for mate selection and spatial
786 competition scale approximately with the product of the total census size and the neighborhood size
787 and so increase rapidly for large populations and dispersal distances. The reverse-time spatial Lambda–
788 Fleming–Viot model described by ? and implemented by ? allows exploration of larger population and
789 landscape sizes, but the precise connection of these models to forward-time demography is not yet
790 clear. Alternatively, implementation of parallelized calculations may allow progress with forward-time
791 simulations.

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

798 **Data Availability**

799 Scripts used for all analyses and figures are available at <https://github.com/kern-lab/spaceness>.

800 **Acknowledgements**

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

803 **Literature Cited**

804 **Comparisons with Stepping-Stone Models**

805 We also compared our model results to a regular grid of discrete populations, which is commonly
806 used as an approximation of continuous geography. An important reason that this approximation
807 is often made is that it allows more efficient, coalescent simulations; we implemented these using
808 msprime (?). In this class of models we imagine an $n \times n$ grid of populations exchanging migrants
809 with neighboring populations at rate m . If these models are good approximations of the continuous
810 case we expect that results will converge as $n \rightarrow \infty$ (while scaling m appropriately and keeping total
811 population size fixed), so we ran simulations while varying n from 5 to 50 (Table A1). To compare
812 with continuous models we first distributed the same “effective” number of individuals across
813 the landscape as in our continuous-space simulations (≈ 6100 , estimated from θ_π of random-mating
814 continuous-space simulations). We then approximate the mean per-generation dispersal distance σ
815 given a total landscape width W as the product of the probability of an individual being a migrant and
816 the distance traveled by migrants: $\sigma = 4m(W/n)$. This means that m in different simulations with the
817 same σ scales with \sqrt{n} . We ran 500 simulations for each value of n while sampling σ from $U(0.2, 4)$.

819 We then randomly selected 60 diploid individuals from each simulation (approximating diploidy by
 820 combining pairs of chromosomes with contiguous indices within demes) and calculated a set of six
 821 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

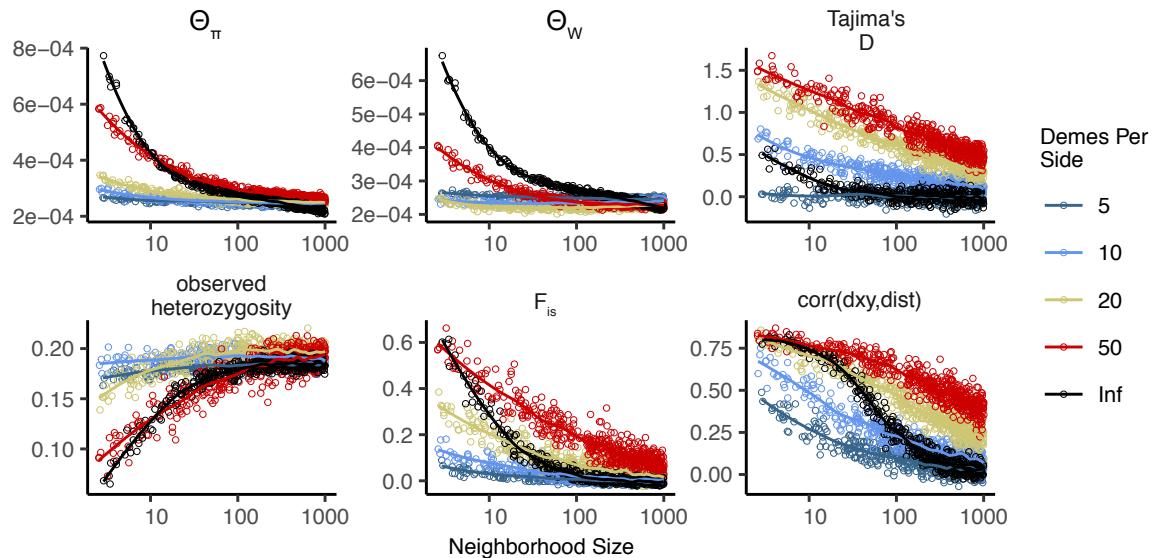


Figure A1 Summary statistics for 2-dimensional coalescent stepping-stone models with fixed total N_e and varying numbers of demes per side. The black “infinite” points are from our forward-time continuous space model. Inter-deme migration rates are related to σ as described above.

822 In general we find many of the qualitative trends are similar among continuous and stepping-
 823 stone models and that, at low neighborhood sizes, many (but not all) statistics from stepping-stone
 824 models approach the continuous model as the resolution of the grid increases. For example, θ_π is
 825 lower in stepping-stone models at low neighborhood sizes (i.e., low m), but increases to approach the
 826 continuous case as the resolution of the landscape increases. Similar patterns are observed for observed
 827 heterozygosity. However, θ_W behaves differently, showing a non-monotonic relationship with grid
 828 resolution. This results in an increasingly positive Tajima’s D in grid simulations at small neighborhood
 829 sizes, to a much greater extent than seen in a continuous model. In contrast to θ_π , increasing the
 830 resolution of the grid causes Tajima’s D to deviate *more* from what is seen in the continuous model.
 831 Similarly, although F_{IS} approaches the continuous case as the resolution of the grid increases at
 832 very small neighborhood sizes, at intermediate neighborhood sizes the continuous case best matches
 833 intermediate grid resolutions.

834 These differences relative to our continuous model mainly reflect two shortcomings of the reverse-
 835 time stepping stone model. If we simulate a coarse grid with relatively large populations in each

836 deme, we cannot accurately capture the dynamics of small neighborhood sizes because mating within
 837 each deme remains random regardless of the migration rate connecting demes. This likely explains
 838 the trends in θ_π , observed heterozygosity, and F_{IS} . However increasing the number of demes while
 839 holding the total number of individuals constant results in small within-deme populations for which
 840 even the minimum sample size of 1 approaches the local N_e (Table A1). This results in an excess
 841 of short terminal branches in the coalescent tree, which decreases the total branch length and leads
 842 to fewer segregating sites, deflated θ_W , and inflated Tajima's D . Overall, the stepping-stone model
 843 reproduces important features of spatial structure in our continuous space model, such as a decline in
 844 θ_π and correlations between spatial and genetic distance with increasing migration, but introduces
 845 artifacts caused by binning the landscape into discrete demes.

846 **Demographic model**

847 Local population regulation is controlled by two parameters, L , and K . Here, we show that these
 848 should be close to the average lifespan of an individual and the average number of individuals per
 849 unit area, respectively. We chose our demographic model so that every individual has on average $1/L$
 850 offspring each time step, and if the local population density of an individual is n , then their probability
 851 of survival until the next time step is (equation (1)):

$$p = \min \left(0.95, \frac{1}{1 + n/(K(1 + L))} \right). \quad (3)$$

852 We capped survival at 0.95 so that we would not have exceptionally long-lived individuals in sparsely
 853 populated areas – otherwise, an isolated individual might live for a very long time. Since $1 - p \approx$
 854 $n/(K(1 + L))$, mortality goes up roughly linearly with the number of neighbors (on a scale given by
 855 K), as would be obtained if, for instance, mortality is due to agonistic interactions. Ignoring migration,
 856 a region is at demographic equilibrium if the per-capita probability of death is equal to the birth rate,
 857 i.e., if $1 - p = 1/L$. (Note that there is no effect of age in the model, which would make the analysis
 858 more complicated.) Solving this for n , we get that in a well-mixed population, the equilibrium density
 859 should be around

$$n = K \frac{L + 1}{L - 1} \quad (4)$$

860 individuals per unit area. At this density, the per-capita death rate is $1/L$, so the mean lifetime is L .
 861 This equilibrium density is *not* K , but (since $L = 4$) is two-thirds larger. However, in practice this
 862 model leads to a total population size which is around K multiplied by total geographic area (but
 863 which depends on σ , as discussed above). The main reason for this is that since offspring tend to
 864 be near their parents, individuals tend to be “clumped”, and so experience a higher average density
 865 than the “density” one would compute by dividing census size by geographic area (?). To maintain a
 866 constant expected total population size would require making (say) K depend on σ ; however, typical
 867 local population densities might then be more dissimilar.

868 **Supplementary Figures and Tables**

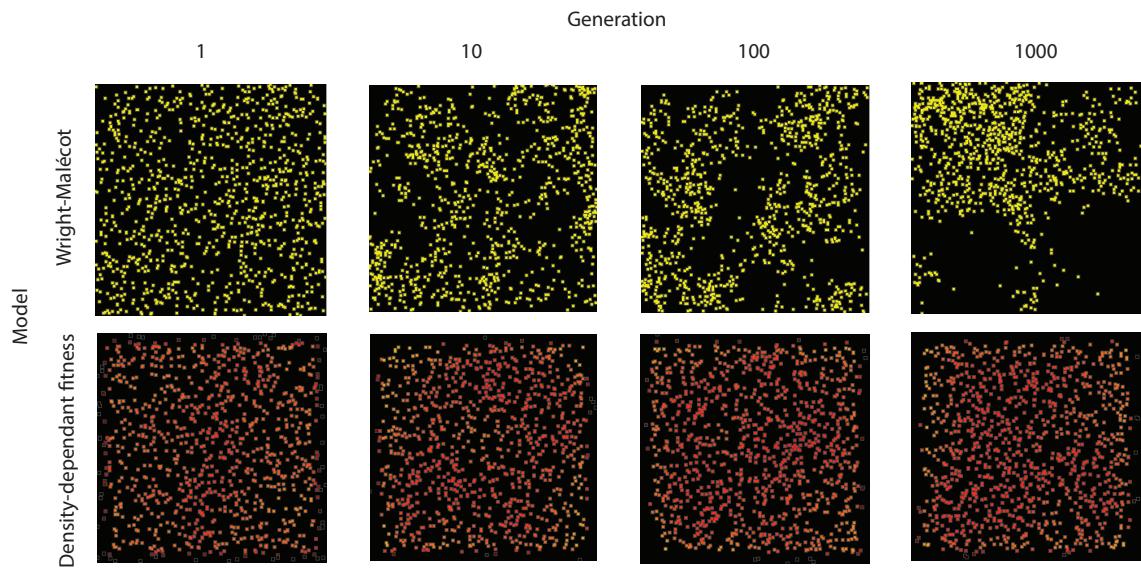


Figure S1 Maps of individual locations in a continuous-space Wright-Malécot model with independent dispersal of all individuals (top) and under our continuous space model incorporating density-dependant fitness (bottom). The clustering seen in the top row is the “Pain in the Torus” described by ?.

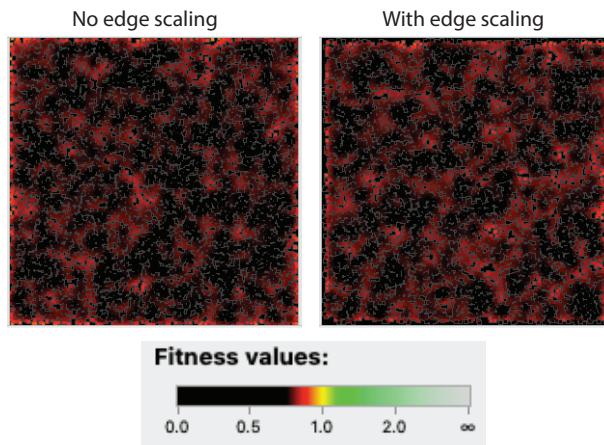


Figure S2 Comparison of individual fitness across the landscape in simulations with (right) and without (left) a decline in fitness approaching range edges. Note the slight excess of high-fitness individuals at edges on the left, which is (partially) counteracted by the scaling procedure.

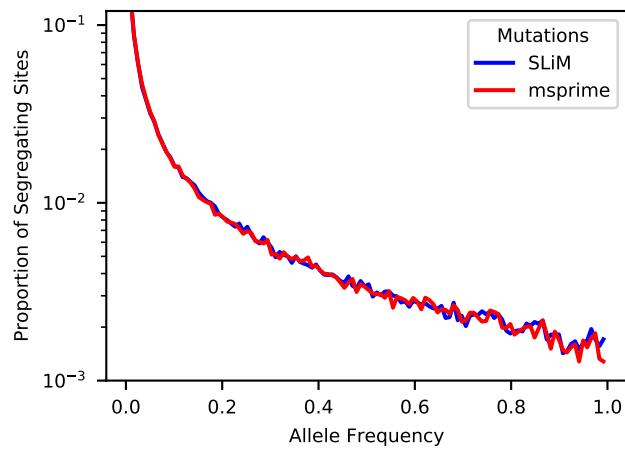


Figure S3 Site frequency spectra from a simulation with neighborhood size = 12.5 when mutations are recorded directly in SLiM (blue line) or applied later in msprime (red line).

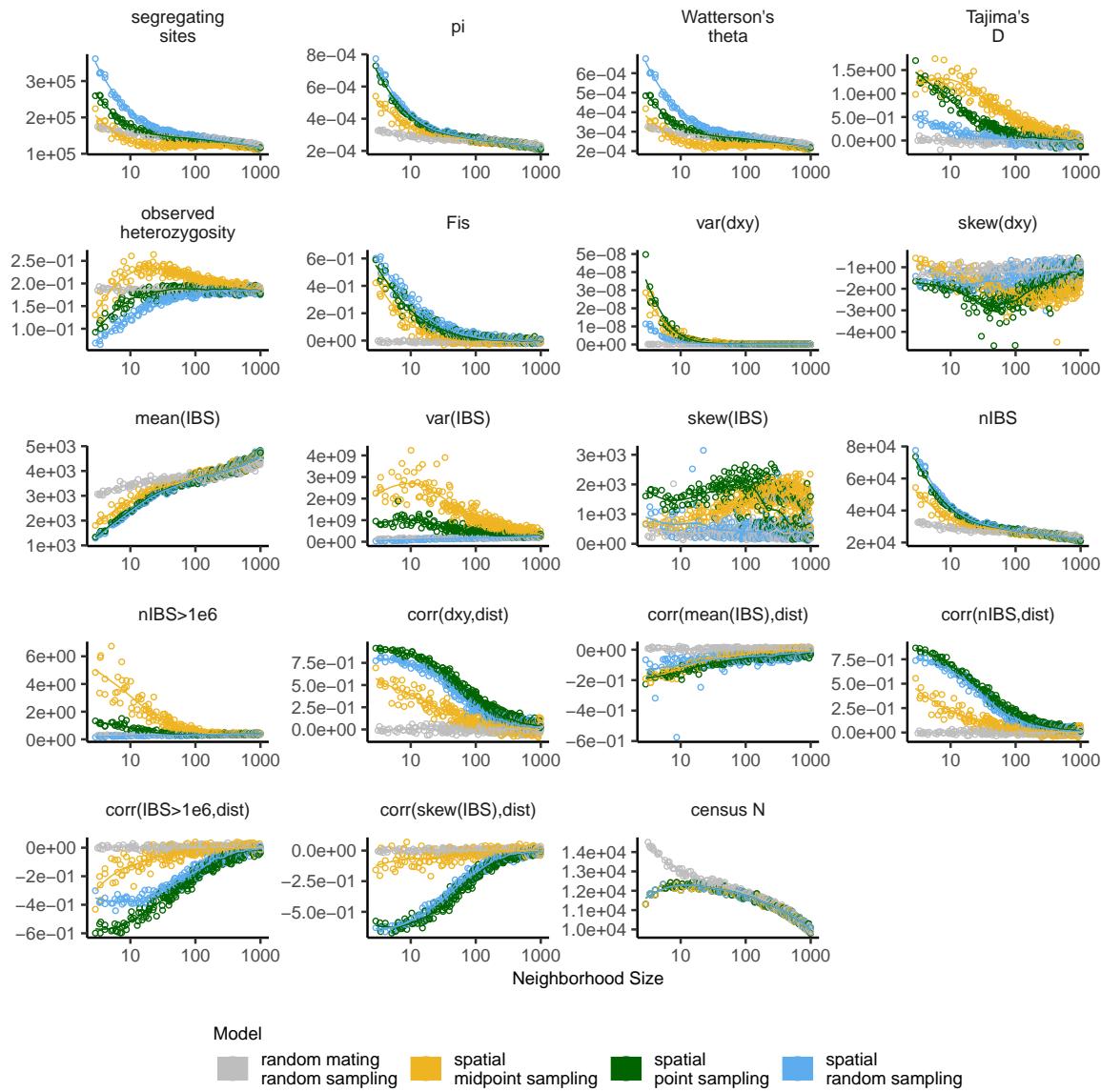


Figure S4 Change in summary statistics by neighborhood size and sampling scheme calculated from simulated sequence data of 60 individuals.

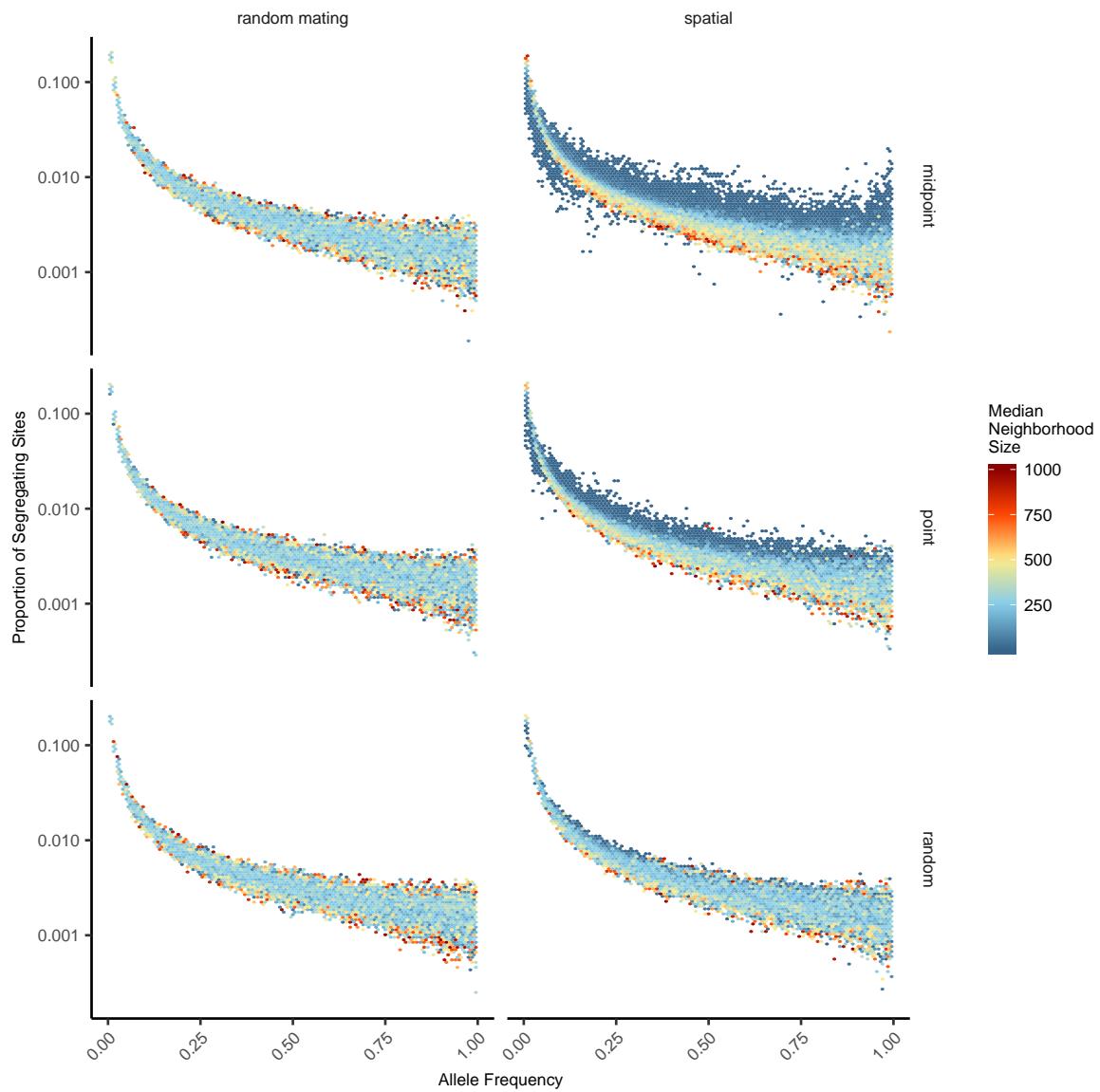


Figure S5 Site frequency spectra for random mating and spatial SLiM models under all sampling schemes.

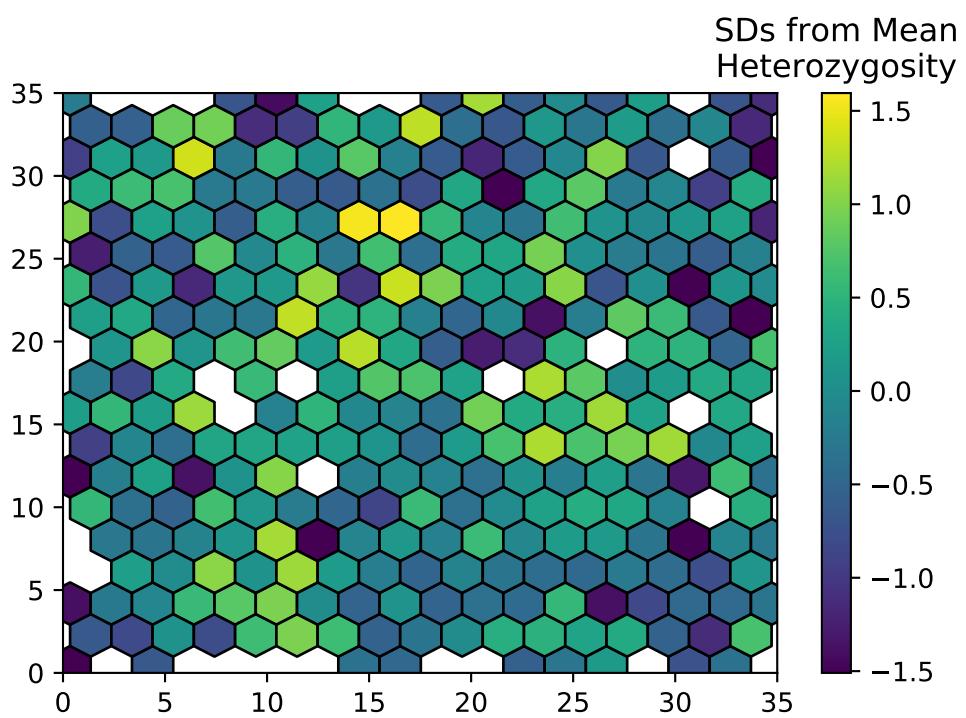


Figure S6 Variation in observed heterozygosity (i.e. proportion of heterozygous individuals) in hexagonal bins across the landscape, estimated from a random sample of 200 individuals from the final generation of a simulation with neighborhood size ≈ 25 . Values were Z-normalized for plotting.

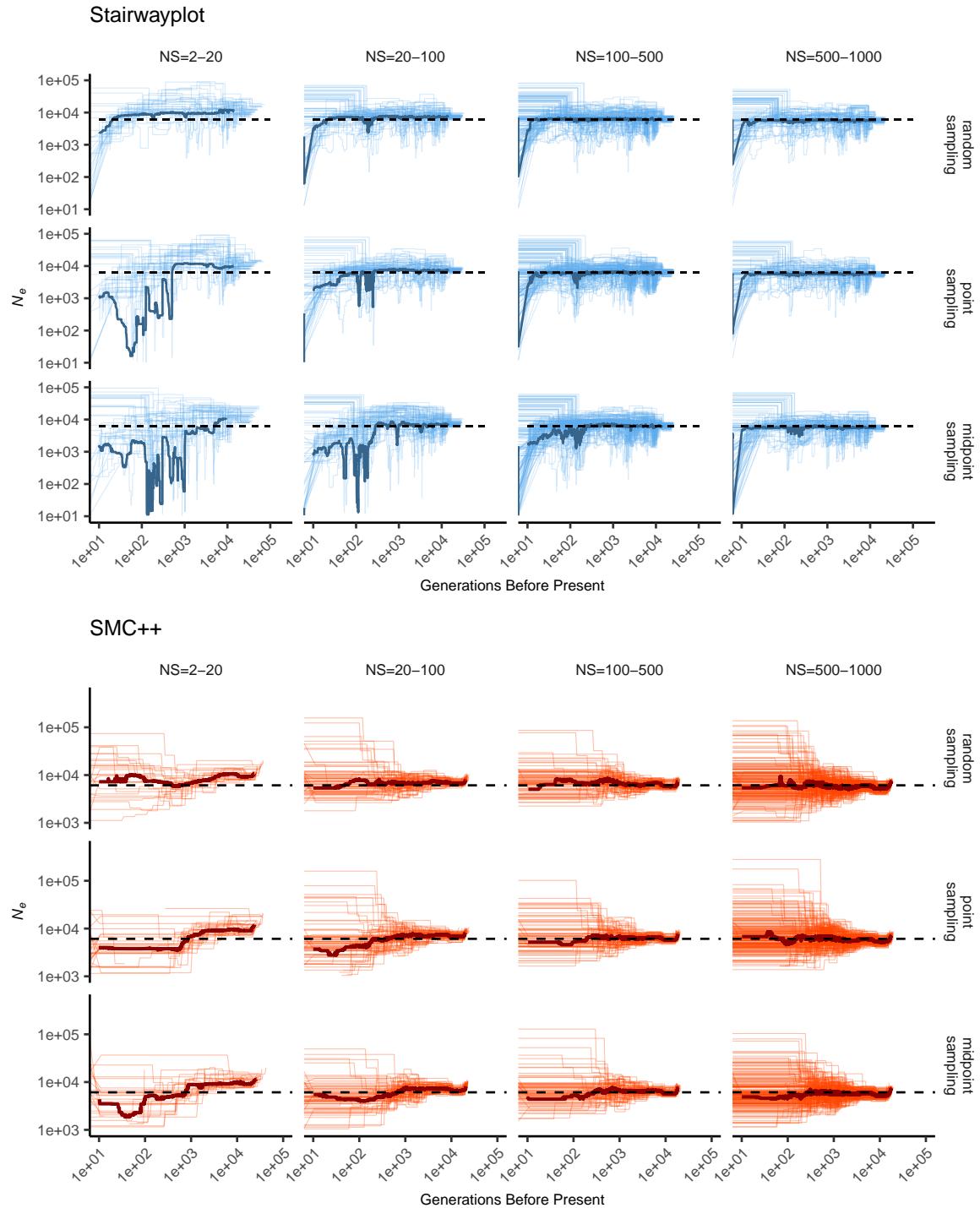


Figure S7 Inferred demographic histories for spatial SLiM simulations, by sampling scheme and neighborhood size (NS) range. Thick lines are rolling medians across all simulations in a bin and thin lines are best fit models for each simulation. Dashed horizontal lines are the average N_e across random-mating SLiM models estimated from θ_π .

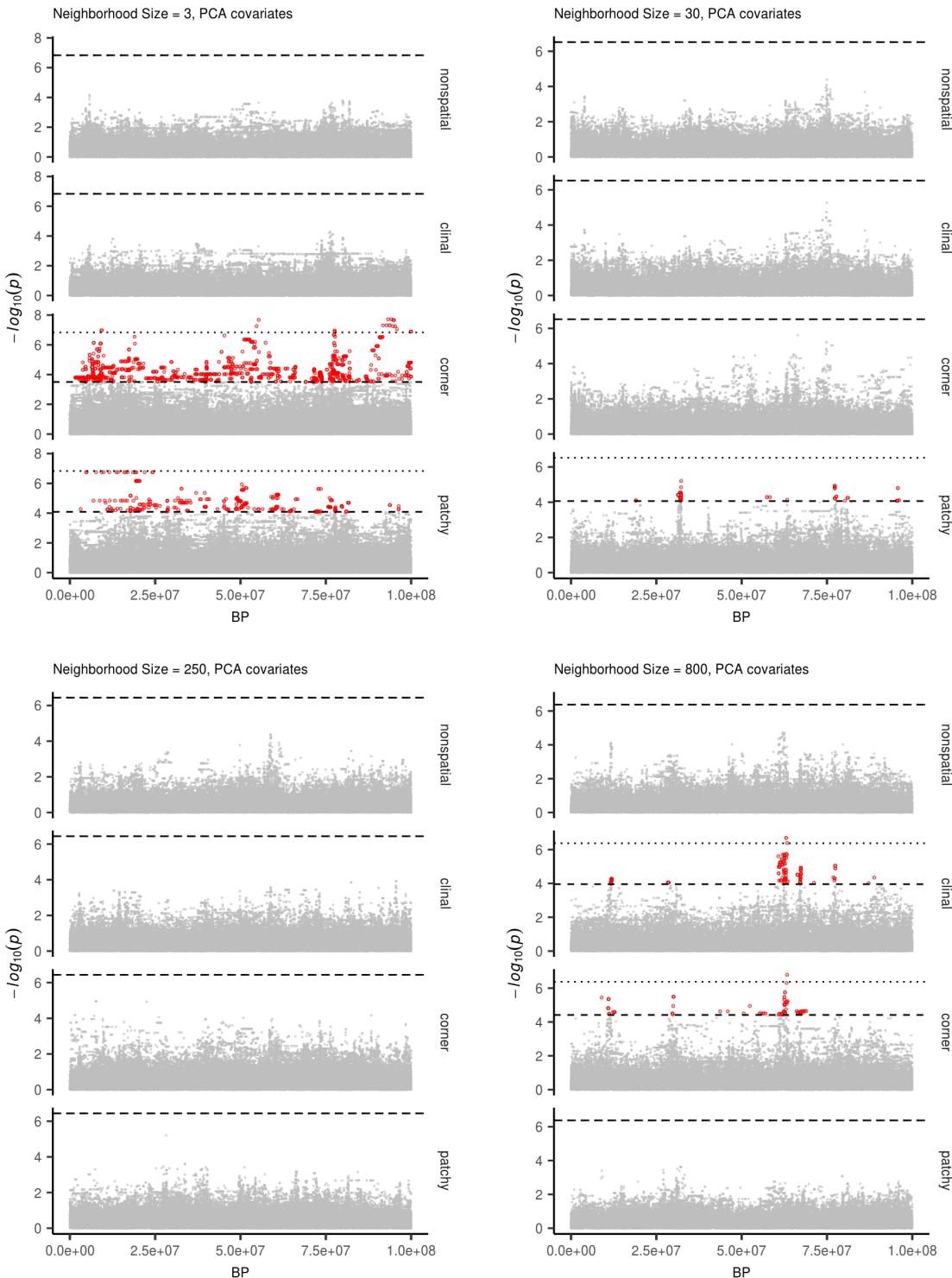


Figure S8 Manhattan plots for a sample of simulations at varying neighborhood sizes. Labels on the right of each plot describes the spatial distribution of environmental factors (described in the methods section of the main text). Points in red are significantly associated with a nongenetic phenotype using a 5% FDR threshold (dashed line). For runs with significant associations the dotted line is a Bonferroni-adjusted cutoff for $p = 0.05$.

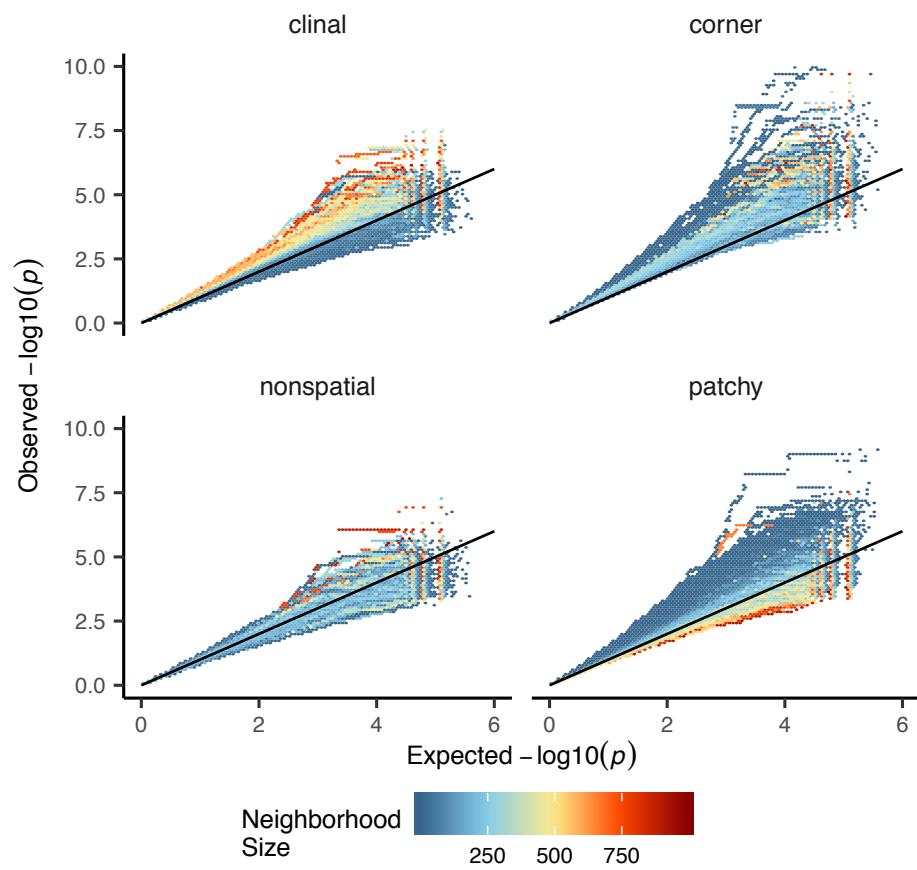


Figure S9 Quantile-quantile plots showing observed $-\log_{10}(p)$ for PC-corrected GWAS run on simulations with varying neighborhood sizes and environmental distributions. Hexagonal bins are colored by the average neighborhood size of simulations with points falling in a given region of quantile-quantile space. Qqplots for a subset of these simulations are shown as lines in Figure 8D.

Table S1 Summary statistics calculated on simulated genotypes.

Statistic	Description
Θ_{pi}	Mean of the distribution of pairwise genetic differences
Θ_W	Effective population size based on segregating sites
Segregating Sites	Total number of segregating sites in the sample
Tajima's D	Difference in Θ_{pi} and Θ_W over its standard deviation
Observed Heterozygosity	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
$mean(IVS)$	Mean of the distribution of pairwise identical-by-state (IBS) tract lengths taken over all pairs.
$var(IVS)$	Variance of the distribution of pairwise identical-by-state (IBS) tract lengths taken over all pairs.
$skew(IVS)$	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 1×10^6 bp per pair across all pairs in the sample.
$corr(D_{xy}, dist)$	Pearson correlation between genetic distance and $\log_{10}(spatial\ distance)$
$corr(mean(IVS), dist)$	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 $\log_{10}(spatial\ distance)$
$corr(IVS > 1e6, dist)$	Pearson correlation between the number of IBS tracts > 1×10^6 bp for each pair of samples and $\log_{10}(spatial\ distance)$
$corr(skew(IVS), dist)$	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
$\Theta\pi$	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
$corr(D_{xy}, dist)$	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
$\Theta\pi$	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
$corr(D_{xy}, dist)$	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

Resubmission Cover Letter
Genetics

C. J. Battey,
Peter Ralph,
and Andrew Kern
Tuesday 18th February, 2020

To the Editor(s) –

We are writing to submit another 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 Decision Letter:

It is most important that you address the following in a revised manuscript:

1. in response to reviewer 3, add text to help clarify how the reproduction scheme / dispersal model differs from models that do not induce the bimodality in the distance distribution the reviewer points out (see my comments below as well);
2. Respond to the concern about the definitions of sigma and neighborhood size;
3. Reviewer 2 has several minor comments that would be straightforward and helpful to address/add depth into your discussion (e.g., comments 2, 4, 5).
4. The other minor comments from each reviewer.

Thanks very much for the useful input. The inaccuracy in describing dispersal under the model was our mistake, and we have added more careful discussion of this point (see below). We have modified the manuscript to incorporate all remaining feedback, as enumerated below, and think this has improved the manuscript.

Reviewer AE Comments:

The point of R3 on the bimodality in the realized dispersal distance in your model is interesting. You can make this more explicit in the paper and describe how there are realistic biological scenarios that motivate such a model (many/most plants) and demarcate how well you think you can generalize to others (the new appendix already helps in that regard). It strikes me your model has relevance for patrilocal/matrilocal systems where the paternal/maternal parent-offspring distances differ - though your model is hermaphroditic so it's an imperfect analogy. Perhaps, to explore more typical animal mating patterns, you can quickly check the impact for one or two outcomes of changing the simulation such that offspring are dispersed around the midpoint of the parental origins, as this would remove the bimodality. That said, please be sure we respect the work done is extensive and don't mean to invite/require a complete re-working. The core tension is between ensuring your results are generalizable if that is what you wish to claim; versus specific to a particular model that is biologically motivated. It's particularly key for the material on GWAS, where you are explicitly trying to address human GWAS results: those sections will be read and taken at face value by a large community even if this simulation scheme introduces some discrepancies from what might more realistically emerge. So at a minimum, some caveats directly in the GWAS discussion section would help.

Thank you for your comments. We have responded to each reviewer below. In response to reviewer 2 we have added several caveats to the GWAS section noting other sources of error such as allele frequency and LD variation among populations. Reviewer 3 raised two good points on our definition of neighborhood size and the interpretation of σ – first, that σ refers to one- rather than two-dimensional distances between parent and offspring (this is consistent with previous theoretical work, and we have corrected our definition in the text), and second, that the use of separate draws for mate selection and dispersal means the total distance traveled by a locus along a pedigree link is greater for “paternal” than “maternal” material (taking “maternal” to mean the parent from which dispersal is calculated). Because our simulated genome is autosomal and individuals hermaphroditic and non-selfing, a locus will experience a mix of “maternal” and “paternal” transmission events along its genealogy. So we see this as a scaling issue rather than one of a “bimodal” dispersal kernel per se – the expected axial displacement of a locus along a pedigree link is $\sqrt{3/2}\sigma$, rather than σ . (Note that although the parent-offspring displacement is a mixture of mean-zero Gaussians with respect to a bimodal distribution of standard deviations, the norm of any such mixtures is never itself bimodal.) We have added a paragraph discussing these points (while trying to keep out of the weeds), to (p. 5, l. 205).

Reviewer 1:

Only minor changes were needed, and they were made. This study neatly summarises many biases arising from unmodelled continuous space (pertinent for real studies), and I would not extend it by further analyses. When polishing the final version, the authors can make some of their messages stronger. For example, Figure 7 and the accompanying text may give the impression that we are doing pretty well with smc++. It is only when panel B is seen one can appreciate the degree of uncertainty. So, in a way, Figure S7 is more telling as to how much we can trust our results in real analysis. After all, in reality, as authors mention, we only have one sample set, not replicates.

Thank you for your comments. While we agree that our analysis found problems with demographic inference under isolation by distance, since this manifests as higher variance in predictions rather than a straightforward bias as we expected (i.e. a reduction in recent inferred Ne), we have chosen to keep the somewhat more nuanced presentation of results in that section used in a previous version, although we added an additional note about the between-replicate variance to the discussion. (p. 18, l. 608)

Reviewer 2:

The authors have done a commendable job of responding to the reviewers' comments. The extended discussion and additional supplemental figures helps with the clarity of the manuscript. The results for the comparison with the discrete stepping-stone model were useful - though see below. The authors have also provided sufficient detail justifying their choice of model parameters.

I don't think I have any critical remaining issues or suggestions, but I did have a couple of thoughts. The authors might just want to think about them.

Thanks for the encouraging words and helpful points, as well as for your mindfulness of the burden of additional analyses.

(2.1) I appreciate the appendix comparing summary statistics in the continuous and stepping-stone models, although I must confess I am now even more uncertain than I was before about which models are more appropriate. I get that the coalescent models make assumptions that are not directly interpretable in terms of observable parameters (although are “interactions between geographically distant parts of the species range” really biologically unrealistic?). This forward model is more interpretable, in that sense, but does it actually produce more realistic data? I guess that’s a question for future work, but that question was one of the main things I came away from this paper with.

Reply: This is a good point. We have added additional discussion noting that while our continuous model has some advantages over discrete models, the best model for any empirical system will inevitably depend on the life history of the organism in question. (p. 22, l. 780)

(2.2) (p. 13, l. 494) “PCA is unable to capture”. Is this a sample size effect? i.e. the precision of the estimates of the PCs depends on sample size, so it’s possible that you just need a larger sample to estimate the PCs when you have weak structure. Or is it something else?

Reply: We have added the caveat “given the sample sizes tested” to the above statement. (p. 13, l. 494)

(2.3) (p. 19, l. 676) Is Khera et al (2018) really the right reference for the idea of polygenic scores? They have been around for a lot longer than that.

Reply: Yes this was an oversight. We now cite Wray et al. 2007 for individual disease risk prediction from GWAS estimates and the first, to our knowledge, PGS application (International Schizophrenia Consortium, 2009).

(2.4) (p. 19, l. 677) *I don't know that the results of Berg and Sohail et al. are indicating a fundamental issue with PC correction, rather than a problem with meta-analysis.*

Reply: We think ? shows that population structure leaves signals in GWAS that resemble those of polygenic selection, and given that this was most apparent in GIANT data employing PCA corrections, indicates a problem with the GWAS that make up the GIANT study at least as much as any issues with meta-analysis.

(2.5) (p. 19, l. 682) *Is the result of Martin et al. really due to population stratification? There seem to be a number of factors affecting transferability of PRS (LD differences, effect size heterogeneity, differences in causal variants etc.), but I don't know that anyone thinks residual structure is the major one, particularly for eg. UKB summary stats.*

Reply: This is a good point – differences in allele frequency and LD structure certainly have significant effects on PRS transferability (at least when the GWAS is based on SNP assays rather than whole-genome sequencing). We think the population structure issue is important to point to here because it will not be solved by simply generating more detailed genomic data from existing samples, but have added a caveat noting these other important potential sources of error at the head of that sentence. (p. 19, l. 682)

(2.6) "population genetics flavor" (p. 21, l. 742) *I don't understand what this means?*

Reply: We meant to contrast this sort of "single-species" model, common in population genetics, to a more "ecological" model involving explicit modeling of resources and/or other species. We've reworded this. (p. 21, l. 742)

(2.7) (p. 22, l. 796) *"process-driven descriptions of ancestry and/or more generalized unsupervised methods". Aren't these two things opposite? That's ok, but it sounds a bit like you have no idea what would help...*

Reply: They certainly entail research in somewhat different directions, but you are correct that we don't know which is the best way to go.

(2.8) Fix citation: Peter L. Ralph and Ashander ??? on pg. 5, line 216

Reply: corrected - thanks.

(2.9) 'there' written twice on pg. 27, line 1033

Reply: thanks for catching this, corrected.

Reviewer 3:

I am fairly pleased overall with the revised version of this article. I commend the authors for the substantial amount of work conducted. This contribution now convincingly demonstrates how important it is to properly account for population structure when individuals are distributed along a spatial continuum.

Thank you for your comments.

(3.1) A couple of minor points about the forward model are still puzzling me though. First of all, in one dimension, the variance of $L(o)$, the random variable corresponding to the offspring's spatial position, given $l(p)$, the parent's position, is equal to $E((L(o) - l(p))^2) := \sigma^2$. Hence, σ is here the square root of the expected squared Euclidean distance between parents and offspring. In general, assuming isotropic migrations, σ is the square root of the expected squared Euclidean distance taken along any dimension. It seems to me that this definition is distinct from the mean parent-offspring distance, as stated by the authors.

Reply: Good point here. We did intend to define σ as is usual in the literature, so have corrected the wording and now direct readers to (?) for discussion of how these parameters are defined in one- and two-dimensional habitats. (p. 3, l. 102)

(3.2) More importantly, I am still a bit confused about the reproduction scheme. My understanding is that, once mating takes place, the two parents involved (noted as $P(1)$ and $P(2)$) each produce a Poisson number of offspring. The location of each offspring is determined by a Gaussian distribution of covariance matrix $\sigma^2 I$, centered on the location of the parent that "produced" this offspring. Let $O(1, i)$ and $O(2, j)$ be the i -th and j -th offspring produced by $P(1)$ and $P(2)$ respectively. Now, if I am not mistaken, half of the genetic material of $O(2, j)$ comes from $P(1)$ (the other half coming from $P(2)$) while $O(1, i)$ receives one chromosome from $P(2)$ (and another one from its "true" parent, i.e., $P(1)$). Hence, when it comes to the parent-offspring distance as measured from their genetic material, half of the chromosomes have a "small" dispersal distance, equal to that of the parent-offspring distance, while the other half have a "large" distance roughly equal to the parent-offspring distance plus the distance between mates.

It is important that the authors clarify this last point because most results presented in this study rely on Wright's neighborhood size, which derives from a model where dispersal of individuals and chromosomes follow well-defined spatial dynamics. A sensible comparison of neighborhood sizes requires that dispersal (along with effective population density) is quantified in a coherent way across different models.

Additionally, the neighborhood size as defined by the authors (see (p. 5, l. 199)) is a function of the census density rather than the effective density of individuals. I am therefore not fully convinced that the range of values for this parameter is comparable to that given in Table 1.

Reply: You are entirely correct here; apologies for the confusion. To clarify the definition of σ and discuss these issues somewhat, we have added a new paragraph at (p. 5, l. 205). As we discuss there, we still calculate $N_W = 4\sigma^2\rho$, where σ is the mean mother-child dispersal distance (not including the distance between mates) and ρ is the census population density. This is not the quantity that is predicted to best describe geographic patterns of genetic variation, which would use *effective* versions of both σ and ρ . This does not bother us for a few reasons. First, we do not attempt to match statistics in our simulations to theoretical results (e.g., the Wright–Malecot formula). Instead, we use N_W as a proxy for "strength of spatial structure", and in these simulations, the "effective" N_W certainly has a monotonic relationship with the N_W we compute. The remaining question is whether, as the reviewer asks, the actual values of N_W are comparable to those computed in the wild. This depends, of course, on how those values were estimated in wild populations. In our opinion, a straightforward estimate of N_W using observed dispersal distances and census density or reproductive adults (similar to how we compute N_W for simulations) is likely more robust and reflective of modern demography than most methods to estimate "effective" versions.

Furthermore, our imprecisions roughly cancel out. As now noted in the text, an average of one and two dispersal distances is a better model for the mean parent-offspring distance; this would give us a value of $\sqrt{3/2}\sigma$, so by using only σ we have lowered N_W by a factor of about 2/3. However, effective population density is lower than census density by a factor of also around 2/3, which would cancel

the first. So, we are fairly confident that our values of N_W match the theoretical versions to within the precision we require (which is only order-of-magnitude, really).

(3.3) Furthermore, the neighborhood size estimates obtained for *Bebicium vittatum* derive from the analysis of a one-dimensional habitat. The product $\sigma^2 \rho$ is thus expressed in this particular case as a number of individuals *per unit of space* (in two dimensions, this product is expressed as a number of individuals). Hence, the unit of space matters a lot here, prohibiting the direct comparison with other figures in the same table or with the simulations conducted in this study.

Reply: Good point, we have removed *Bebicium vittatum* from table 1.

(3.4) Finally, Leblois, Estoup and Streiff (2006) *Molecular Ecology* conducted a large simulation studies with a focus very close to that of the present study. In particular, they investigated the impact of sampling on the inference of summary statistics commonly used in population genetics. It would probably be relevant to compare some of the results presented in Leblois et al. to that put forward here.

Reply: Thank you for bringing this study to our attention. We have added it to the citations in the demographic inference section.

Reviewer 4:

The authors have thoroughly addressed nearly all of my comments in their timely manuscript, and I am satisfied with their responses.

Thank you very much!

(4.1) Regarding Table S1, I would still suggest adding a column about interpretation for those variables and evaluated with respect to the simulated results.

Reply: Thank you for your comments. We chose to evaluate a few key statistics in detail in the main text and then provide an interpretation rooted in how the marginal genealogies of the tree sequence interact with space, and hope that this will assist readers when examining other statistics shown.