

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

C.J. Battey*,¹, Peter L. Ralph*,[†] and Andrew D. Kern*,[†]

*University of Oregon Dept. Biology, Institute for Ecology Evolution

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

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

Introduction

The inescapable reality that biological organisms live, move, and reproduce in continuous geography is usually omitted from population genetic models. However, mates tend to live near to one another and to their offspring, leading to a positive correlation between genetic differentiation and geographic distance. This pattern of "isolation by distance" (Wright 1943) is one of the most widely replicated empirical findings in population genetics (Aguillon *et al.* 2017; Jay *et al.* 2012; Sharbel *et al.* 2000). Despite a long history of analytical work describing the genetics of populations distributed across continuous geography (e.g., Wright (1943); Rousset (1997); Barton *et al.* (2002, 2010); Ringbauer *et al.* (2017); Robledo-Arnuncio and Rousset (2010); Wilkins and Wakeley (2002); Wilkins (2004)), much modern work still describes geographic structure as a set of discrete populations connected by migration (e.g., Wright 1931; Epperson 2003; Rousset and Leblois 2011; Shirk and Cushman 2014; Lundgren and Ralph 2019) or as an average over such discrete models (Petkova *et al.* 2015; Al-Asadi *et al.* 2019). For this reason, most population genetics statistics are interpreted with reference to discrete, well-mixed populations,

Manuscript compiled: Thursday 5th December, 2019

¹301 Pacific Hall, University of Oregon Dept. Biology, Institute for Ecology and Evolution. cbattey2@uoregon.edu.

[†]these authors co-supervised this project

40 and most empirical papers analyze variation within clusters of genetic variation inferred by programs
41 like *STRUCTURE* (Pritchard *et al.* 2000) with methods that assume these are randomly mating units.

42 The assumption that populations are “well-mixed” has important implications for downstream
43 inference of selection and demography. Methods based on the coalescent (Kingman 1982; Wakeley
44 2009) assume that the sampled individuals are a random draw from a well-mixed population that is
45 much larger than the sample (Wakeley and Takahashi 2003). The key assumption is that the individuals
46 of each generation are *exchangeable*, so that there is no correlation between the fate or fecundity of a
47 parent and that of their offspring (Huillet and Möhle 2011). If dispersal or mate selection is limited by
48 geographic proximity, this assumption can be violated in many ways. For instance, if mean viability or
49 fecundity is spatially autocorrelated, then limited geographic dispersal will lead to parent–offspring
50 correlations. Furthermore, nearby individuals will be more closely related than an average random
51 pair, so drawing multiple samples from the same area of the landscape will represent a biased sample
52 of the genetic variation present in the whole population (Städler *et al.* 2009).

53 Two areas in which spatial structure may be particularly important are demographic inference and
54 genome-wide association studies (GWAS). Previous work has found that discrete population structure
55 can create false signatures of population bottlenecks when attempting to infer demographic histories
56 from microsatellite variation (Chikhi *et al.* 2010), statistics summarizing the site frequency spectrum
57 (SFS) (Ptak and Przeworski 2002; Städler *et al.* 2009; St. Onge *et al.* 2012), or runs of homozygosity in a
58 single individual (Mazet *et al.* 2015). The increasing availability of whole-genome data has led to the
59 development of many methods that attempt to infer detailed trajectories of population sizes through
60 time based on a variety of summaries of genetic data (Liu and Fu 2015; Schiffels and Durbin 2014;
61 Sheehan *et al.* 2013; Terhorst *et al.* 2016). Because all of these methods assume that the populations
62 being modeled are approximately randomly mating, they are likely affected by spatial biases in the
63 genealogy of sampled individuals (Wakeley 1999), which may lead to incorrect inference of population
64 changes over time (Mazet *et al.* 2015). However, previous investigations of these effects have focused on
65 discrete rather than continuous space models, and the level of isolation by distance at which inference
66 of population size trajectories become biased by structure is not well known. Here we test how two
67 methods suitable for use with large samples of individuals – stairwayplot (Liu and Fu 2015) and
68 SMC++ (Terhorst *et al.* 2016) – perform when applied to populations evolving in continuous space
69 with varying sampling strategies and levels of dispersal.

70 Spatial structure is also a major challenge for interpreting the results of genome-wide association
71 studies (GWAS). This is because many phenotypes of interest have strong geographic differences due
72 to the (nongenetic) influence of environmental or socioeconomic factors, which can therefore show
73 spurious correlations with spatially patterned allele frequencies (Bulik-Sullivan *et al.* 2015; Mathieson
74 and McVean 2012). Indeed, two recent studies found that previous evidence of polygenic selection on
75 human height in Europe was confounded by subtle population structure (Sohail *et al.* 2018; Berg *et al.*
76 2018), suggesting that existing methods to correct for population structure in GWAS are insufficient.
77 However we have little quantitative idea of the population and environmental parameters that can be
78 expected to lead to biases in GWAS.

79 Last, some of the most basic tools of population genetics are summary statistics like F_{IS} and
80 Tajima’s D , which are often interpreted as reflecting the influence of selection or demography on
81 sampled populations (Tajima 1989). Statistics like Tajima’s D are essentially summaries of the site
82 frequency spectrum, which itself reflects variation in branch lengths and tree structure of the underlying
83 genealogies of sampled individuals. Geographically limited mate choice distorts the distribution of
84 these genealogies (Maruyama 1972; Wakeley 1999), which can affect the value of Tajima’s D (Städler
85 *et al.* 2009). Similarly, the distribution of tract lengths of identity by state among individuals contains
86 information about not only historical demography (Harris and Nielsen 2013; Ralph and Coop 2013)
87 and selection (Garud *et al.* 2015), but also dispersal and mate choice (Ringbauer *et al.* 2017; Baharian
88 *et al.* 2016). We are particularly keen to examine how such summaries will be affected by models that
89 incorporate continuous space, both to evaluate the assumptions underlying existing methods and to
90 identify where the most promising signals of geography lie.

91 To study this, we have implemented an individual-based model in continuous geography that
92 incorporates overlapping generations, local dispersal of offspring, and density-dependent survival. We

93 simulate chromosome-scale genomic data in tens of thousands of individuals from parameter regimes
94 relevant to common subjects of population genetic investigation such as humans and *Drosophila*, and
95 output the full genealogy and recombination history of all final-generation individuals. We use these
96 simulations to test how sampling strategy interacts with geographic population structure to cause
97 systematic variation in population genetic summary statistics typically analyzed assuming discrete
98 population models. We then examine how the fine-scale spatial structures occurring under limited
99 dispersal impact demographic inference from the site frequency spectrum. Last, we examine the
100 impacts of continuous geography on genome-wide association studies (GWAS) and identify regions of
101 parameter space under which the results from GWAS may be misleading.

102 Materials and Methods

103 *Modeling Evolution in Continuous Space*

104 The degree to which genetic relationships are geographically correlated depends on the chance that
105 two geographically nearby individuals are close relatives – in modern terms, by the tension between
106 migration (the chance that one is descended from a distant location) and coalescence (the chance that
107 they share a parent). A key early observation by Wright (1946) is that this balance is often nicely
108 summarized by the “neighborhood size”, defined to be $N_W = 4\pi\rho\sigma^2$, where σ is the mean parent–
109 offspring distance and ρ is population density. This can be thought of as proportional to the average
110 number of potential mates for an individual (those within distance 2σ), or the number of potential
111 parents of a randomly chosen individual. Empirical estimates of neighborhood size vary hugely
112 across species – even in human populations, estimates range from 40 to over 5,000 depending on the
113 population and method of estimation (Table 1).

114 The first approach to modeling continuously distributed populations was to endow individuals in a
115 Wright-Fisher model with locations in continuous space. However, since the total size of the population
116 is constrained, this introduces interactions between arbitrarily distant individuals, which (aside from
117 being implausible) was shown by Felsenstein (1975) to eventually lead to unrealistic population
118 clumping if the range is sufficiently large. Another method for modeling spatial populations is to
119 assume the existence of a grid of discrete randomly mating populations connected by migration, thus
120 enforcing regular population density by edict. Among many other results drawn from this class of
121 “lattice” or “stepping stone” models (Epperson 2003), Rousset (1997) showed that the slope of the linear
122 regression of genetic differentiation (F_{ST}) against the logarithm of spatial distance is an estimate of
123 neighborhood size. Although these grid models may be good approximations of continuous geography
124 in many situations, they do not model demographic fluctuations, and limit investigation of spatial
125 structure below the level of the deme, assumptions whose impacts are unknown. An alternative
126 method for dealing with continuous geography is a new class of coalescent models, the Spatial Lambda
127 Fleming-Viot models (Barton *et al.* 2010; Kelleher *et al.* 2014).

128 To avoid hard-to-evaluate approximations, we here used forward-time, individual-based simulations
129 across continuous geographical space. The question of what regulates real populations has a
130 long history and many answers (e.g., Lloyd 1967; Antonovics and Levin 1980; Crawley 1990), but it is
131 clear that populations must at some point have density-dependent feedback on population size, or
132 else they would face eventual extinction or explosion. In the absence of unrealistic global population
133 regulation, this regulation must be local, and there are many ways to achieve this (Bolker *et al.* 2003). In
134 our simulations, each individual’s probability of survival is a decreasing function of local population
135 density, which shifts reproductive output towards low-density regions, and produces total census sizes
136 that fluctuate around an equilibrium. This also prevents the population clumping seen by Felsenstein
137 (1975) (Supplemental Figure S1)). Such models have been used extensively in ecological modeling
138 (Durrett and Levin 1994; Bolker and Pacala 1997; Law *et al.* 2003; Fournier and Méléard 2004; Champer
139 *et al.* 2019) but rarely in population genetics, where to our knowledge implementations of continuous
140 space models before their availability through SLiM (Haller and Messer 2019) have focused on a
141 small number of genetic loci (e.g., Slatkin and Barton 1989; Barton *et al.* 2002; Robledo-Arnuncio and
142 Rousset 2010; Rossine 2014), which limits the ability to investigate the impacts of continuous space
143 on genome-wide genetic variation as is now routinely sampled from real organisms. By simulating

144 chromosome-scale sequence alignments and complete population histories we are able to treat our
145 simulations as real populations and replicate the sampling designs and analyses commonly conducted
146 on real genomic data.

147 **A Forward-Time Model of Evolution in Continuous Space**

148 We simulated populations using the program SLiM v3.1 (Haller and Messer 2019). Each time step
149 consists of three stages: reproduction, dispersal, and mortality. To reduce the parameter space we
150 use the same parameter, denoted σ , to modulate the spatial scale of interactions at all three stages
151 by adjusting the standard deviation of the corresponding Gaussian functions. As in previous work
152 (Wright 1943; Ringbauer *et al.* 2017), σ is equal to the mean parent-offspring distance.

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

159 To maintain a stable population, mortality increases with local population density. To do this we say
160 that individuals at distance d have a competitive interaction with strength $g(d)$, where g is the Gaussian
161 density with mean zero and standard deviation σ . Then, the sum of all competitive interactions with
162 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
163 all neighbors within distance 3σ . Since g is a probability density, n_i is an estimate of the number of
164 nearby individuals per unit area. Then, given a per-unit carrying capacity K , the probability of survival
165 until the next time step for individual i is

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

166 We chose this functional form so that the equilibrium population density per unit area is close to K ,
167 and the mean lifetime is around L ; for more description see the Appendix.

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

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

174 where x_i and y_i are the spatial coordinates of individual i , and W is the width (and height) of the
175 square habitat. This buffer roughly counteracts the increase in fitness individuals close to the edge
176 would otherwise have, though the effect is relatively subtle (Figure S2).

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

182 We stored the full genealogy and recombination history of final-generation individuals as tree
183 sequences (Kelleher *et al.* 2018), as implemented in SLiM (Haller *et al.* 2019). Scripts for figures and
184 analyses are available at <https://github.com/kern-lab/spaceness>.

185 We ran 400 simulations for the spatial and random-mating models on a square landscape of width
186 $W = 50$ with per-unit carrying capacity $K = 5$ (census $N \approx 10,000$), average lifetime $L = 4$, genome
187 size 10^8 bp, recombination rate 10^{-9} per bp per generation, and drawing σ values from a uniform
188 distribution between 0.2 and 4. To speed up the simulations and limit memory overhead we used a

189 mutation rate of 0 in SLiM and later applied mutations to the tree sequence with msprime's `mutate`
 190 function (Kelleher *et al.* 2016). Because msprime applies mutations proportionally to elapsed time,
 191 we divided the mutation rate of 10^{-8} mutations per site per generation by the average generation
 192 time estimated for each value of σ (see 'Demographic Parameters' below) to convert the rate to units
 193 of mutations per site per unit time. We verify that this procedure produces the same site frequency
 194 spectrum as applying mutations directly in SLiM in Figure S3, in agreement with theory (Ralph *et al.*
 195 2019). Simulations were run for 1.6 million timesteps (approximately $30N$ generations).

196 We also compared our model's output to a commonly-used approximation of continuous space,
 197 the stepping-stone model, which we simulated with msprime (Kelleher *et al.* 2016). These results are
 198 discussed in detail in the Appendix, but in general we find that the demographic structure of a stepping-
 199 stone model can depend strongly on the chosen discretization, and some artifacts of discretization do
 200 not seem to disappear in the limit of a fine grid. For many summary statistics, finer discretizations (we
 201 used a 50×50 grid) produced similar results to the continuous model, but this was not true for others
 202 (e.g., F_{IS} and Tajima's D), which differed from the continuous model *more* at finer discretizations.

203 Demographic Parameters

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

216 Sampling

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

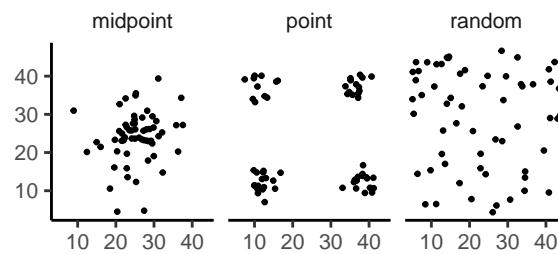


Figure 1 Example sampling maps for 60 individuals on a 50×50 landscape for midpoint, point, and random sampling strategies, respectively.

224 **Summary Statistics**

225 We calculated the site frequency spectrum and a set of 18 summary statistics (Table S1) from 60 diploid
226 individuals sampled from the final generation of each simulation using the python package scikit-
227 allel (Miles and Harding 2017). Statistics included common single-population summaries including
228 mean pairwise divergence (π), inbreeding coefficient (F_{IS}), and Tajima's D , as well (motivated by
229 Rousset (1997)'s results) the correlation coefficient between the logarithm of the spatial distance and
230 the proportion of identical base pairs across pairs of individuals.

231 Following recent studies that showed strong signals for dispersal and demography in the distri-
232 bution of shared haplotype block lengths (e.g., Ringbauer *et al.* 2017; Baharian *et al.* 2016), we also
233 calculated various summaries of the distribution of pairwise identical-by-state (IBS) block lengths
234 among sampled chromosomes, defined to be the set of distances between adjacent sites that differ
235 between the two chromosomes. The full distribution of lengths of IBS tracts for each pair of chromo-
236 somes was first calculated with a custom python function. We then calculated the first three moments
237 of this distribution (mean, variance, and skew) and the number of blocks over 10^6 base pairs both
238 for each pair of individuals and for the full distribution across all pairwise comparisons. We then
239 calculated correlation coefficients between spatial distance and each moment of the pairwise IBS tract
240 distribution. Because more closely related individuals on average share longer haplotype blocks we
241 expect that spatial distance will be negatively correlated with mean haplotype block length, and that
242 this correlation will be strongest (i.e., most negative) when dispersal is low. The variance, skew, and
243 count of long haplotype block statistics are meant to reflect the relative length of the right (upper) tail
244 of the distribution, which represents the frequency of long haplotype blocks, and so should reflect
245 recent demographic events (Chapman and Thompson 2002). For a subset of simulations, we also
246 calculated cumulative distributions for IBS tract lengths across pairs of distant (more than 48 map
247 units) and nearby (less than 2 map units) individuals. Last, we examined the relationship between
248 allele frequency and the spatial dispersion of an allele by calculating the average distance between
249 individuals carrying each derived allele.

250 The effects of sampling on summary statistic estimates were summarized by testing for differences
251 in mean (ANOVA, (R Core Team 2018)) and variance (Levene's test, (Fox and Weisberg 2011)) across
252 sampling strategies for each summary statistic.

253 **Demographic Inference**

254
255 To assess the impacts of continuous spatial structure on demographic inference we inferred pop-
256 ulation size histories for all simulations using two approaches: stairwayplot (Liu and Fu 2015) and
257 SMC++ (Terhorst *et al.* 2016). Stairwayplot fits its model to a genome-wide estimate of the SFS, while
258 SMC++ also incorporates linkage information. For both methods we sampled 20 individuals from all
259 spatial simulations using random, midpoint, 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 (Miles and Harding 2017),
263 generated 100 bootstrap replicates per simulation by resampling over sites, and fit models for all
264 bootstrap samples using 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 (Purcell *et al.*
278 2007). SNPs with a minor allele frequency less than 0.5% were excluded from this analysis. Phenotype
279 values were set to vary by two standard deviations across the landscape in a rough approximation
280 of the variation seen in height across Europe (Turchin *et al.* 2012; Garcia and Quintana-Domeque
281 2006, 2007). Conceptually our approach is similar to that taken by Mathieson and McVean (2012),
282 though here we model fully continuous spatial variation and compare GWAS output across a range of
283 dispersal distances.

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

298 We performed principal components analysis (PCA) using scikit-allel (Miles and Harding 2017) on
299 the matrix of derived allele counts by individual for each simulation. SNPs were first filtered to remove
300 strongly linked sites by calculating LD between all pairs of SNPs in a 200-SNP moving window and
301 dropping one of each pair of sites with an R^2 over 0.1. The LD-pruned allele count matrix was then
302 centered and all sites scaled to unit variance when conducting the PCA, following recommendations
303 in Patterson *et al.* (2006).

304 We ran linear-model GWAS both with and without the first 10 principal components as covariates
305 in PLINK and summarized results across simulations by counting the number of SNPs with p -value
306 below 0.05 after adjusting for an expected false positive rate of less than 5% (Benjamini and Yekutieli
307 2001). We also examined p values for systematic inflation by comparing to the values expected from
308 a uniform distribution (because no SNPs were used when generating phenotypes, well-calibrated
309 p -values should be uniform).

310 Results from all analyses were summarized and plotted with the "ggplot2" (Wickham 2016) and
311 "cowplot" (Wilke 2019) packages in R (R Core Team 2018).

312 **Results**

313 **Demographic Parameters and Run Times**

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

325 from K and L . This is the reason that we included *random mating* simulations – where mate choice and
 326 offspring dispersal are both nonspatial – since this should preserve the random fluctuations in local
 327 population density while destroying any spatial genetic structure. We verified that random mating
 328 models retained no geographic signal by showing that summary statistics did not differ significantly
 329 between sampling regimes (Table S2), unlike in spatial models (discussed below).

330 There are a few additional things to note about Figure 2. First, all three quantities are non-monotone
 331 with neighborhood size. Census size largely declines as neighborhood size increases for both the spatial
 332 and random mating models. However, for spatial models this decline only begins for neighborhood
 333 sizes ≥ 10 . Spatial and random mating models are indistinguishable from one another for neighborhood
 334 sizes larger than 100. Census sizes range from around 14,000 at low σ in the random mating model
 335 to 10,000 for both models when neighborhood sizes approach 1,000. The scaling of census sizes in
 336 both random-mating and spatial models appears to be related to two consequences of the spatial
 337 competition function: the decline of fitness at range edges, which effectively reduces the habitable area
 338 by one σ around the edge of the map and so results in a smaller habitable area at high σ values; and
 339 variation in the equilibrium population density given varying competition radii. Furthermore, census
 340 size increases in spatial models as neighborhood size increases from 2 to 10. This may reflect an Allee
 341 effect (Allee *et al.* 1949) in which some individuals are unable to find mates when the mate selection
 342 radius is very small.

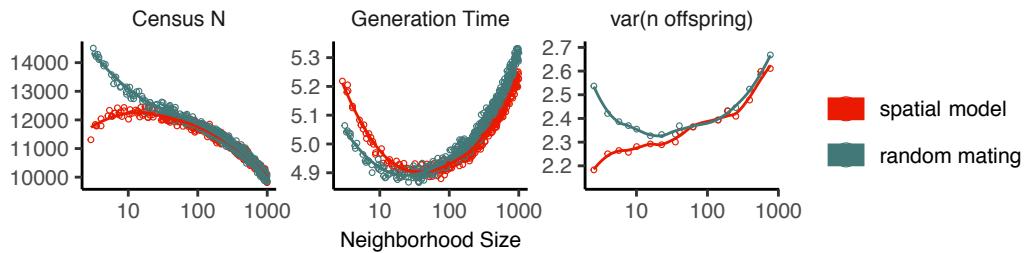


Figure 2 Genealogical parameters from spatial and random mating SLiM simulations, by neighborhood size.

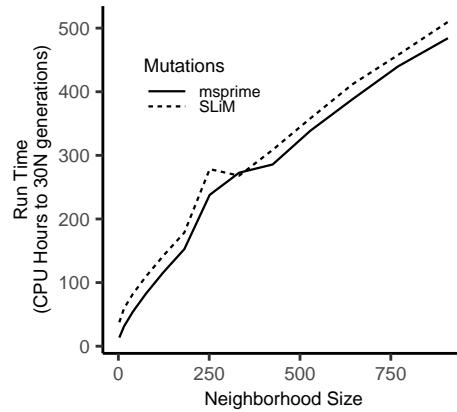


Figure 3 Run times of continuous space simulations with landscape width 50 and expected density 5 under varying neighborhood size. Times are shown for simulations run with mutations applied directly in SLiM (dashed lines) or later applied to tree sequences with msprime (solid lines). Times for simulations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

343 Generation time similarly shows complex behavior with respect to neighborhood sizes, and varies

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

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

354 Run times for our model scale approximately linearly with neighborhood size (Figure 3), with the
355 lowest neighborhood sizes reaching 30N generations in around a day and those with neighborhood
356 size approaching 1,000 requiring up to three weeks of computation. As currently implemented running
357 simulations at neighborhood sizes more than 1,000 to coalescence is likely impractical, though running
358 these models for more limited timescales and then “recapitulating” the simulation using reverse-time
359 simulation from the resulting tree sequence in msprime is possible (Haller *et al.* 2019).

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

361 Even though certain aspects of population demography depend on the scale of spatial interactions, it
362 still could be that population genetic variation is well-described by a well-mixed population model.
363 Indeed, mathematical results suggest that genetic variation in some spatial models should be well-
364 approximated by a Wright-Fisher population if neighborhood size is large and all samples are ge-
365 ographically widely separated (??). However, the behavior of most common population genetic
366 summary statistics other than Tajima’s *D* (Städler *et al.* 2009) has not yet been described in realistic
367 geographic models. Moreover, as we will show, spatial sampling strategies can affect summaries of
368 genetic variation at least as strongly as the underlying population dynamics.

369 **Site Frequency Spectra and Summaries of Diversity** Figure 4 shows the effect of varying neighbor-
370 hood size and sampling strategy on the site frequency spectrum (Figure 4, Figure S5) and several
371 standard population genetic summary statistics (Figure 4B; additional statistics are shown in Figure
372 S4). Consistent with findings in island and stepping stone simulations (Städler *et al.* 2009), the SFS
373 shows a significant enrichment of intermediate frequency variants in comparison to the nonspatial
374 expectation. This bias is most pronounced below a neighborhood size of 100 and is exacerbated by
375 midpoint and point sampling of individuals (depicted in Figure 1). Reflecting this, Tajima’s *D* is quite
376 positive in the same situations (Figure 4B). Notably, the point at which Tajima’s *D* approaches 0 differs
377 strongly across sampling strategies – varying from a neighborhood size of roughly 50 for random
378 sampling to at least 1000 for midpoint sampling.

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



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

395 Values of observed heterozygosity and its derivative F_{IS} also depend heavily on neighborhood size
396 under spatial models as well as the sampling scheme. F_{IS} is inflated above the expectation across
397 most of the parameter space examined and across all sampling strategies. This effect is caused by
398 a deficit of heterozygous individuals in low-dispersal simulations – a continuous-space version of
399 the Wahlund effect (Wahlund 1928). Indeed, for random sampling under the spatial model, F_{IS} does
400 not approach the random mating equivalent until neighborhood sizes of nearly 1000. On the other
401 hand, the dependency of raw observed heterozygosity on neighborhood size is not monotone. Under
402 midpoint sampling observed heterozygosity is inflated even over the random mating expectation, as a
403 result of the a higher proportion of heterozygotes occurring in the middle of the landscape (Figure S6).
404 This echoes a report from Shirk and Cushman (2014) who observed a similar excess of heterozygosity
405 in the middle of the landscape when simulating under a lattice model.

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

416 The patterns observed for correlations of IBS tract lengths with geographic distance are similar
417 to those observed in the more familiar correlation of allele frequency measures such as D_{xy} (i.e.,
418 “genetic distance”) or F_{ST} against geographic distance (Rousset 1997). D_{xy} is positively correlated
419 with the geographic distance between the individuals, and the strength of this correlation declines
420 as dispersal increases (Figure 4B), as expected (Wright 1943; Rousset 1997). This relationship is very
421 similar across random and point sampling strategies, but is weaker for midpoint sampling, perhaps
422 due to a dearth of long-distance comparisons. In much of empirical population genetics a regression
423 of genetic differentiation against spatial distance is a de-facto metric of the significance of isolation
424 by distance. The similar behavior of moments of the pairwise distribution of IBS tract lengths shows
425 why haplotype block sharing has recently emerged as a promising source of information on spatial
426 demography through methods described in Ringbauer *et al.* (2017) and Baharian *et al.* (2016).

427 **Spatial distribution of allele copies** Mutations occur in individuals and spread geographically over
428 time. Because low frequency alleles generally represent recent mutations (Sawyer 1977; Griffiths *et al.*
429 1999), the geographic spread of an allele may covary along with its frequency in the population. To
430 visualize this relationship we calculated the average distance among individuals carrying a focal
431 derived allele across simulations with varying neighborhood sizes, shown in Figure 6. On average
432 we find that low frequency alleles are the most geographically restricted, and that the extent to which
433 geography and allele frequency are related depends on the amount of dispersal in the population.
434 For populations with large neighborhood sizes we found that even very low frequency alleles can be
435 found across the full landscape, whereas in populations with low neighborhood sizes the relationship
436 between distance among allele copies and their frequency is quite strong. This is the basic process
437 underlying Novembre and Slatkin’s (2009) method for estimating dispersal distances based on the
438 distribution of low frequency alleles, and also generates the greater degree of bias in GWAS effect sizes
439 for low frequency alleles identified in Mathieson and McVean (2012).

440 **Effects of Space on Demographic Inference**

441 One of the most important uses for population genetic data is inferring demographic history of popu-
442 lations. As demonstrated above, the site frequency spectrum and the distribution of IBS tracts varies
443 across neighborhood sizes and sampling strategies. Does this variation lead to different inferences of
444 past population sizes? To ask this we inferred population size histories from samples drawn from our

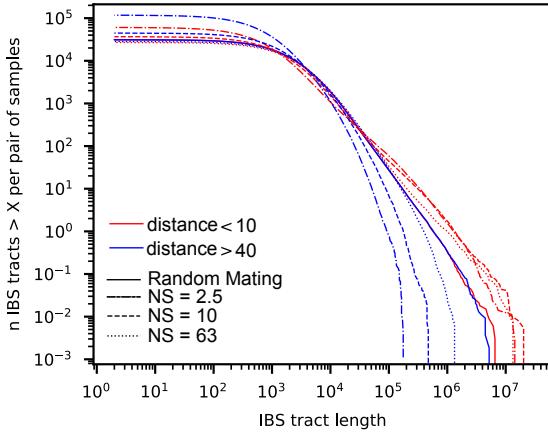


Figure 5 Cumulative distributions for IBS tract lengths per pair of individuals at different geographic distances, across three neighborhood sizes (NS). Nearby pairs (red curves) share many more long IBS tracts than do distant pairs (blue curves), except in the random mating model. The distribution of 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.

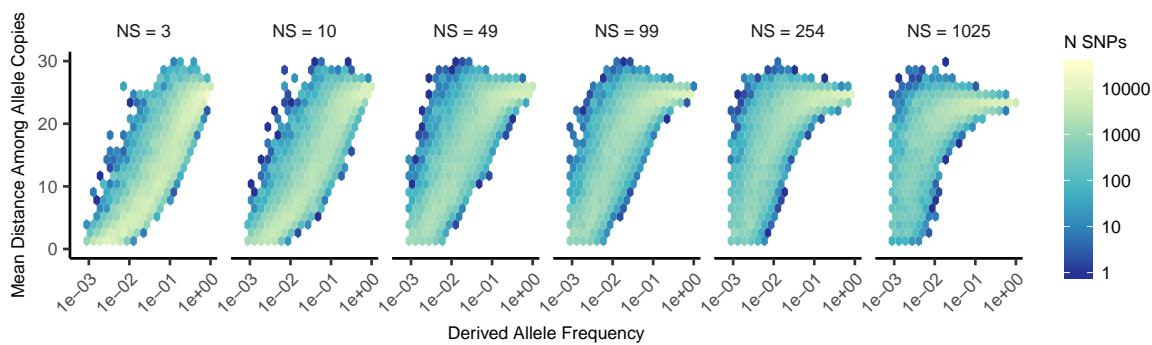


Figure 6 Spatial spread of rare alleles by neighborhood size (NS): Each plot shows the distribution (across derived alleles and simulations) of average pairwise distance between individuals carrying a focal derived allele and derived allele frequency.

445 simulated populations with two approaches: stairwayplot (Liu and Fu 2015), which uses a genome-
446 wide estimate of the SFS, and SMC++ (Terhorst *et al.* 2016), which incorporates information on both the
447 SFS and linkage disequilibrium across the genome.

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

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

468 **GWAS**

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

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

493 Clinal environments cause an interesting pattern in false positives after PC correction: at low
494 neighborhood sizes the correction removes nearly all significant associations, but at neighborhood
495 sizes above roughly 250 the proportion of significant SNPs increases to up to 0.4% (Figure 8). This
496 may be due to a loss of descriptive power of the PCs – as neighborhood size increases, the total

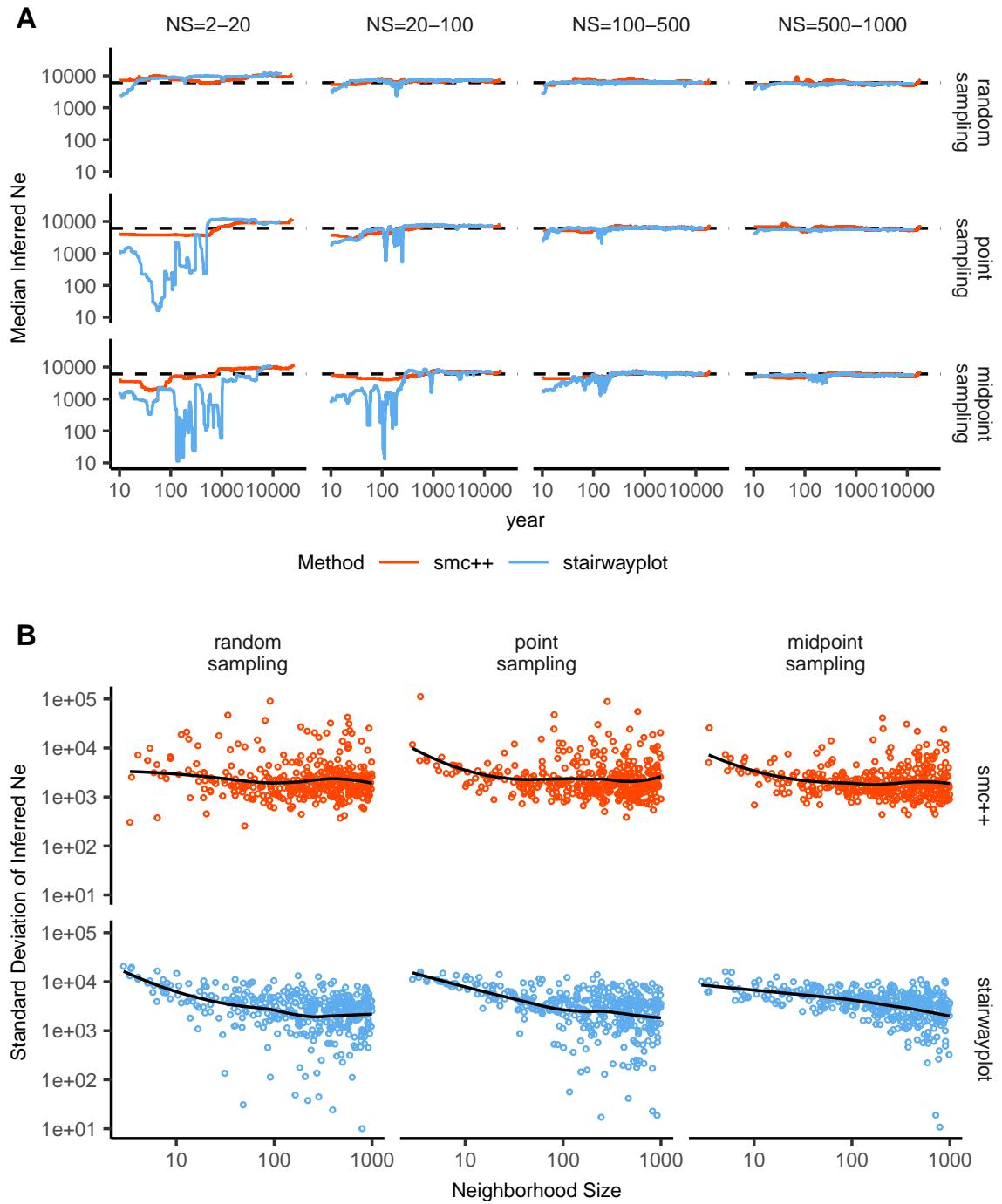


Figure 7 A: Rolling median inferred N_e trajectories for stairwayplot and smc++ across sampling strategies and neighborhood size bins. The dotted line shows the mean N_e of random-mating simulations. B: Standard deviation of individual inferred N_e trajectories, by neighborhood size and sampling strategy.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

497 proportion of variance explained by the first 10 PC axes declines from roughly 10% to 4% (Figure
498 8B). Essentially, PCA seems unable to effectively summarize the weak population structure present in
499 large-neighborhood simulations, but these populations continue to have enough spatial structure to
500 create significant correlations between genotypes and the environment. A similar process can also be
501 seen in the corner phenotype distribution, in which the count of significant SNPs initially declines as
502 neighborhood size increases and then increases at approximately the point at which the proportion of
503 variance explained by PCA approaches its minimum.

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

513 Discussion

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

521 Effects of Dispersal

522 Limited dispersal inflates effective population size, creates correlations between genetic and spatial
523 distances, and introduces strong distortions in the site frequency spectrum that are reflected in a
524 positive Tajima's D (Figure 4). At the lowest dispersal distances, this can increase genetic diversity
525 threefold relative to random-mating expectations. These effects are strongest when neighborhood
526 sizes are below 100, but in combination with the effects of nonrandom sampling they can persist up to
527 neighborhood sizes of at least 1000 (e.g., inflation in Tajima's D and observed heterozygosity under
528 midpoint sampling). If samples are chosen uniformly from across space, the general pattern is similar
529 to expectations of the original analytic model of Wright (1943), which predicts that populations with
530 neighborhood sizes under 100 will differ substantially from random mating, while those above 10,000
531 will be nearly indistinguishable from panmixia.

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

541 The genealogical patterns introduced by limited dispersal are particularly apparent in the distribu-
542 tion of haplotype block lengths (Figure 4). This is because identical-by-state tract lengths reflect the
543 impacts of two processes acting along the branches of the underlying genealogy – both mutation and
544 recombination – rather than just mutation as is the case when looking at the site frequency spectrum or
545 related summaries. This means that the pairwise distribution of haplotype block lengths carries with
546 it important information about genealogical variation in the population, and correlation coefficients

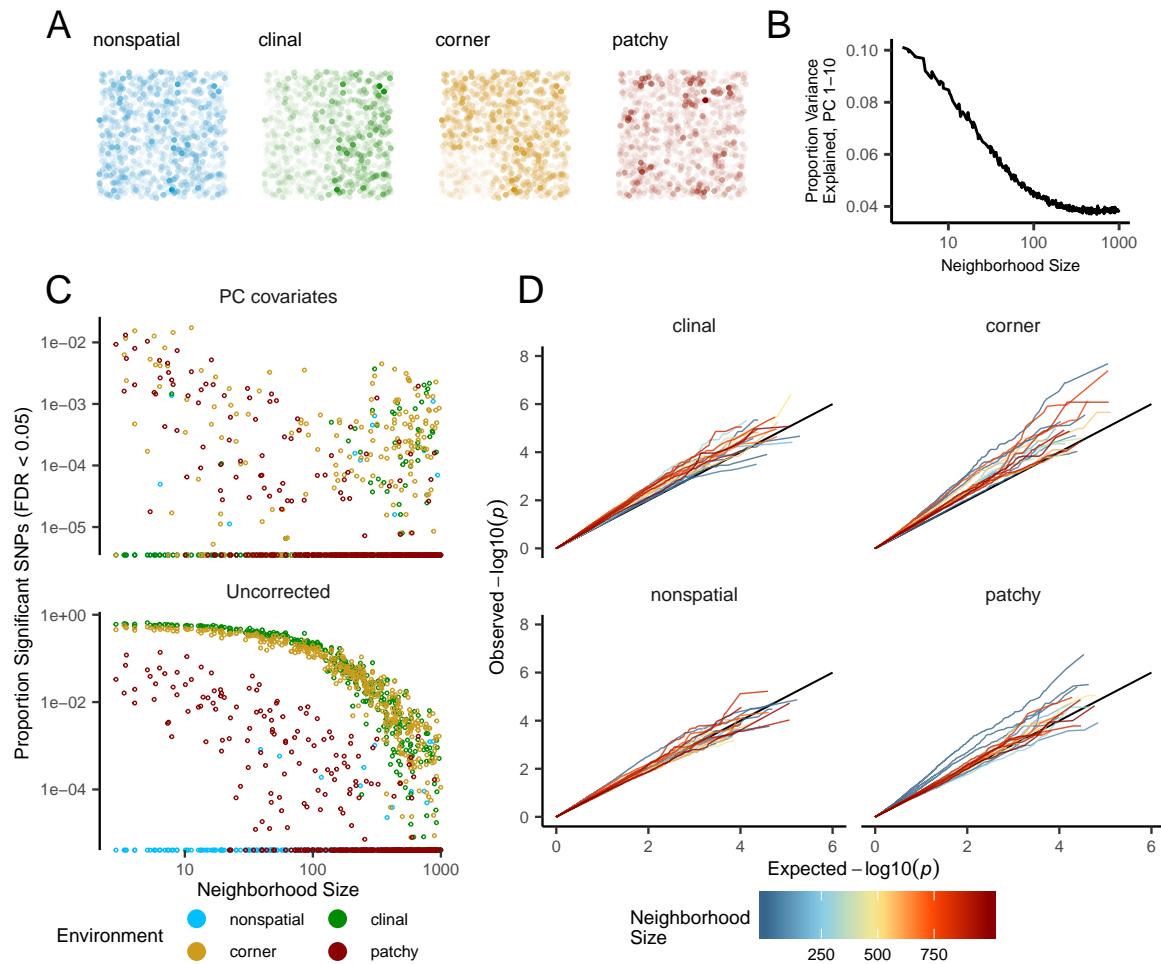


Figure 8 Impacts of spatially varying environments and isolation by distance on linear regression GWAS. Simulated quantitative phenotypes are determined only by an individual's location and the spatial distribution of environmental factors. In **A** we show the phenotypes and locations of sampled individuals 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.

547 between moments of the this distribution and geographic location contain signal similar to the correlations
548 between F_{ST} or D_{xy} and geographic distance (Rousset 1997). Indeed this basic logic underlies
549 two recent studies explicitly estimating dispersal from the distribution of shared haplotype block
550 lengths (Ringbauer *et al.* 2017; Baharian *et al.* 2016). Conversely, because haplotype-based measures of
551 demography are particularly sensitive to variation in the underlying genealogy, inference approaches
552 that assume random mating when analyzing the distribution of shared haplotype block lengths are
553 likely to be strongly affected by spatial processes.

554 **Effects of Sampling**

555 One of the most important differences between random mating and spatial models is the effect of
556 sampling: in a randomly mating population the spatial distribution of sampling effort has no effect on
557 estimates of genetic variation (Table S1), but when dispersal is limited sampling strategy can compound
558 spatial patterns in the underlying genealogy and create pervasive impacts on all downstream genetic
559 analyses (see also Städler *et al.* (2009)). In most species, the difficulty of traveling through all parts
560 of a species range and the inefficiency of collecting single individuals at each sampling site means
561 that most studies follow something closest to the “point” sampling strategy we simulated, in which
562 multiple individuals are sampled from nearby points on the landscape. For example, in ornithology a
563 sample of 10 individuals per species per locality is a common target when collecting for natural history
564 museums. In classical studies of *Drosophila* variation the situation is considerably worse, in which a
565 single orchard might be extensively sampled.

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

576 We observed the largest sampling effects using “midpoint” sampling. This model is meant to
577 reflect a bias in sampling effort towards the middle of a species’ range. In empirical studies this
578 sampling strategy could arise if, for example, researchers choose to sample the center of the range
579 and avoid range edges to maximize probability of locating individuals during a short field season.
580 Because midpoint sampling provides limited spatial resolution it dramatically reduces the magnitude
581 of observed correlations between spatial and genetic distances. More surprisingly, midpoint sampling
582 also leads to strongly positive Tajima’s D and an inflation in the proportion of heterozygous individuals
583 in the sample – similar to the effect of sampling a single deme in an island model as reported in Städler
584 *et al.* (2009). This increase in observed heterozygosity appears to reflect the effects of range edges,
585 which are a fundamental facet of spatial genetic variation. If individuals move randomly in a finite
586 two-dimensional landscape then regions in the middle of the landscape receive migrants from all
587 directions while those on the edge receive no migrants from at least one direction. The average number
588 of new mutations moving into the middle of the landscape is then higher than the number moving
589 into regions near the range edge, leading to higher heterozygosity and lower inbreeding coefficients
590 (F_{IS}) away from range edges. Though here we used only a single parameterization of fitness decline at
591 range edges we believe this is a general property of non-infinite landscapes as it has also been observed
592 in previous studies simulating under lattice models (Neel *et al.* 2013; Shirk and Cushman 2014).

593 In summary, we recommend that empirical researchers collect individuals from across as much
594 of the species’ range as practical, choosing samples separated by a range of spatial scales. Many
595 summary statistics are designed for well-mixed populations, and so provide different insights into
596 genetic variation when applied to different subsets of the population. Applied to a cluster of samples,
597 summary statistics based on segregating sites (e.g., Watterson’s θ and Tajima’s D), heterozygosity, or
598 the distribution of long haplotype blocks, can be expected to depart significantly from what would be

599 obtained from a wider distribution of samples. Comparing the results of analyses conducted on all
600 individuals versus those limited to single individuals per locality can provide an informative contrast.
601 Finally we wish to point out that the bias towards intermediate allele frequencies that we observe may
602 mean that the importance of linked selection, at least as is gleaned from the site frequency spectrum,
603 may be systematically underestimated currently.

604 **Demography**

605 Previous studies have found that population structure and nonrandom sampling can create spurious
606 signals of population bottlenecks when attempting to infer demographic history with microsatellite
607 variation, summary statistics, or runs of homozygosity (Chikhi *et al.* 2010; Städler *et al.* 2009; Ptak
608 and Przeworski 2002; Mazet *et al.* 2015). Here we found that methods that infer detailed population
609 trajectories through time based on the SFS and patterns of LD across the genome are also subject to this
610 bias, with some combinations of dispersal and sampling strategy systematically inferring deep recent
611 population bottlenecks and overestimating ancient N_e by around a factor of 2. We were surprised
612 to see that both stairwayplot and SMC++ can tolerate relatively strong isolation by distance – i.e.,
613 neighborhood sizes of 20 – and still perform well when averaging results across multiple simulations.
614 Inference in populations with neighborhood sizes over 20 was relatively unbiased unless samples were
615 concentrated in the middle of the range (Figure 7). Although median demography estimates across
616 many independent simulations were fairly accurate, empirical work has only a single estimate to work
617 with, and individual model fits (Figure S7) suggest that spuriously inferred population size changes
618 and bottlenecks are common, especially at small neighborhood sizes. As we will discuss below, most
619 empirical estimates of neighborhood size, including all estimates for human populations, are large
620 enough that population size trajectories inferred by these approaches should not be strongly affected by
621 spatial biases created by dispersal in continuous landscapes. In contrast, Mazet *et al.* (2015) found that
622 varying migration rates through time could create strong biases in inferred population trajectories from
623 an n -island model with parameters relevant for human history, suggesting that changes in migration
624 rates through time are more likely to drive variation in inferred N_e than isolation by distance.

625 We found that SMC++ was more robust to the effects of space than stairwayplot, underestimating
626 recent populations by roughly half in the worst time periods rather than nearly 10-fold as with
627 stairwayplot. Though this degree of variation in population size is certainly meaningful in an ecological
628 context, it is relatively minor in population genetic terms. Methods directly assessing haplotype
629 structure in phased data example, (e.g., Browning and Browning 2015) are thought to provide increased
630 resolution for recent demographic events, but in this case the error we observed was essentially an
631 accurate reflection of underlying genealogies in which terminal branches are anomalously short.
632 Combined with our analysis of IBS tract length variation (Figure 5) this suggests that haplotype-based
633 methods are likely to be affected by similar biases.

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

640 Our analysis suggests that many empirical analyses of population size history using methods like
641 SMC++ are robust to error caused by spatial structure within continuous landscapes. Inferences drawn
642 from static SFS-based methods like stairwayplot should be treated with caution when there are signs
643 of isolation by distance in the underlying data (for example, if a regression of F_{ST} against the logarithm
644 of geographic distance has a significantly positive slope), and in particular an inference of population
645 bottlenecks in the last 1000 years should be discounted if sampling is clustered, but estimates of deeper
646 time patterns are likely to be fairly accurate. The biases in the SFS and haplotype structure identified
647 above (see also Wakeley 1999; Chikhi *et al.* 2010; Städler *et al.* 2009) are apparently small enough that
648 they fall within the range of variability regularly inferred by these approaches, at least on datasets of
649 the size we simulated.

650 **GWAS**

651 Spatial structure is particularly challenging for genome-wide association studies, because the effects of
652 dispersal on genetic variation are compounded by spatial variation in the environment (Mathieson
653 and McVean 2012). Spatially restricted mate choice and dispersal causes variation in allele frequencies
654 across the range of a species. If environmental factors affecting the phenotype of interest also vary over
655 space, then allele frequencies and environmental exposures will covary over space. In this scenario an
656 uncorrected GWAS will infer genetic associations with a purely environmental phenotype at any site
657 in the genome that is differentiated over space, and the relative degree of bias will be a function of the
658 degree of covariation in allele frequencies and the environment (i.e., Figure 8C, bottom panel). This
659 pattern has been demonstrated in a variety of simulation and empirical contexts (Price *et al.* 2006; Yu
660 *et al.* 2005; Young *et al.* 2018; Mathieson and McVean 2012; Kang *et al.* 2008, 2010; Bulik-Sullivan *et al.*
661 2015; Berg *et al.* 2018; Sohail *et al.* 2018).

662 Incorporating PC positions as covariates in a linear-regression GWAS (Price *et al.* 2006) is designed
663 to address this challenge by regressing out a baseline level of “average” differentiation. In essence, a
664 PC-corrected GWAS asks “what regions of the genome are more associated with this phenotype than
665 the average genome-wide association observed across populations?” In our simulations, we observed
666 that this procedure can fail under a variety of circumstances. If dispersal is limited and environmental
667 variation is clustered in space (i.e., corner or patchy distributions in our simulations), PC positions fail
668 to capture the fine-scale spatial structure required to remove all signals of association. Conversely, as
669 dispersal increases, PCA loses power to describe population structure before spatial mixing breaks
670 down the relationship between genotype and the environment. These effects were observed with all
671 spatially correlated environmental patterns, but were particularly pronounced if environmental effects
672 are concentrated in one region, as was also found by Mathieson and McVean (2012). Though increasing
673 the number of PC axes used in the analysis may reduce the false-positive rate, this may also decrease
674 the power of the test to detect truly causal alleles (Lawson *et al.* 2019).

675 In this work we simulated a single chromosome with size roughly comparable to one human
676 chromosome. If we scale the number of false-positive associations identified in our analyses to a
677 GWAS conducted on whole-genome data from humans, we would expect to see several thousand weak
678 false-positive associations after PC corrections in a population with neighborhood sizes up to at least
679 1000 (which should include values appropriate for many human populations). Notably, very few of the
680 spurious associations we identified would be significant at a conservative Bonferroni-adjusted *p*-value
681 cutoff (see Figure S8). This suggests that GWAS focused on finding strongly associated alleles for traits
682 controlled by a limited number of variants in the genome are likely robust to the impacts of continuous
683 spatial structure. However, methods that analyze the combined effects of thousands or millions of
684 weakly associated variants such as polygenic risk scores (Khera *et al.* 2018) are likely to be affected
685 by subtle population structure. Indeed as recently identified in studies of genotype associations for
686 human height in Europe (Berg *et al.* 2018; Sohail *et al.* 2018), PC regression GWAS in modern human
687 populations do include residual signal of population structure in large-scale analyses of polygenic traits.
688 When attempting to make predictions across populations with different environmental exposures,
689 polygenic risk scores affected by population structure can be expected to offer low predictive power, as
690 was shown in a recent study finding lower performance outside European populations (Martin *et al.*
691 2019).

692 In summary, spatial covariation in population structure and the environment confounds the in-
693 terpretation of GWAS *p*-values, and correction using principal components is insufficient to fully
694 separate these signals for polygenic traits under a variety of environmental and population parameter
695 regimes. Other GWAS methods such as mixed models (Kang *et al.* 2008) may be less sensitive to
696 this confounding, but there is no obvious reason that this should be so. One approach to estimating
697 the degree of bias in GWAS caused by population structure is LD score regression (Bulik-Sullivan
698 *et al.* 2015). Though this approach appears to work well in practice, its interpretation is not always
699 straightforward and it is likely biased by the presence of linked selection (Berg *et al.* 2018). In addition,
700 we observed that in many cases the false-positive SNPs we identified appeared to be concentrated in
701 LD peaks similar to those expected from truly causal sites (Figure S8), which may confound LD score

702 regression.

703 We suggest a straightforward alternative for species in which the primary axes of population
 704 differentiation is space (note this is likely not the case for some modern human populations): run a
 705 GWAS with spatial coordinates as phenotypes and check for *p*-value inflation or significant associations.
 706 If significant associations with sample locality are observed after correcting for population structure,
 707 the method is sensitive to false positives induced by spatial structure. This is essentially the approach
 708 taken in our “clinal” model (though we add normally distributed noise to our phenotypes). This
 709 approach has recently been taken with polygenic scores for UK Biobank samples in Haworth *et al.*
 710 (2019), finding that scores are correlated with birth location even in this relatively homogenous sample.
 711 Of course, it is possible that genotypes indirectly affect individual locations by adjusting organismal
 712 fitness and thus habitat selection across spatially varying environments, but we believe that this
 713 hypothesis should be tested against a null of stratification bias inflation rather than accepted as true
 714 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	(Campbell and Dooley 1992)
<i>Borreria frutescens</i>	salt marsh plant	20 - 30	Genetic+Survey	(Antlfinger 1982)
<i>Oreamnos americanus</i>	mountain goat	36 - 100	Genetic	(Shirk and Cushman 2014)
<i>Homo sapiens</i>	Gainj- and Kalam-speaking people, Papua New Guinea	40 - 213	Genetic	(Rousset 1997)
<i>Formica sp.</i>	colonial ants	50 - 100	Genetic	(Pamilo 1983)
<i>Astrocaryum mexicanum</i>	palm tree	102 - 895	Genetic+survey	(Eguiarte <i>et al.</i> 1993)
<i>Spermophilus mollis</i>	ground squirrel	204 - 480	Genetic+Survey	(Antolin <i>et al.</i> 2001)
<i>Sceloporus olivaceus</i>	lizard	225 - 270	Survey	(Kerster 1964)
<i>Dieffenbachia longispatha</i>	beetle-pollinated colonial herb	227 - 611	Survey	(Young 1988)
<i>Aedes aegypti</i>	Yellow-fever mosquito	268	Genetic	(Jasper <i>et al.</i> 2019)
<i>Homo sapiens</i>	Gainj- and Kalam-speaking people, Papua New Guinea	410	Survey	(Rousset 1997)
<i>Quercus laevis</i>	Oak tree	> 440	Genetic	(Berg and Hamrick 1995)
<i>Drosophila pseudoobscura</i>	fruit fly	500 - 1,000	Survey+Crosses	(Wright 1946)
<i>Homo sapiens</i>	POPRES data NE Europe	1,342 - 5,425	Genetic	(Ringbauer <i>et al.</i> 2017)
<i>Bebicium vittatum</i>	intertidal snail	240,000	Survey	(Rousset 1997)
<i>Bebicium vittatum</i>	intertidal snail	360,000	Genetic	(Rousset 1997)

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

716 For how much of the tree of life do spatial patterns circumscribe genomic variation? In Table 1 we
 717 gathered estimates of neighborhood size from a range of organisms to get an idea of how strongly

718 local geographic dispersal affects patterns of variation. This is an imperfect measure: some aspects
719 of genetic variation are most strongly determined by neighborhood size (Wright 1946), others (e.g.,
720 number of segregating sites) are more strongly determined by global N_e or by the ratio of the two. In
721 addition, these empirical examples are likely biased towards small-neighborhood species (because few
722 studies have quantified neighborhood size in species with very high dispersal or population density).
723 However, from the available data we find that neighborhood sizes in the range we simulated are fairly
724 common across a range of taxa. At the extreme low end of empirical neighborhood size estimates
725 we see some flowering plants, large mammals, and colonial insects like ants. Species such as this
726 have neighborhood size estimates small enough that spatial processes are likely to strongly influence
727 inference. These include some human populations such as the Gainj- and Kalam-speaking people of
728 Papua New Guinea, in which the estimated neighborhood sizes in Rousset (1997) range from 40 to 410
729 depending on the method of estimation. Many more species occur in a middle range of neighborhood
730 sizes between 100 and 1000 – a range in which spatial processes play a minor role in our analyses
731 under random spatial sampling but are important when sampling of individuals in space is clustered.
732 Surprisingly, even some flying insects with huge census population sizes fall in this group, including
733 fruit flies (*D. melanogaster*) and mosquitoes (*A. aegypti*). Last, many species likely have neighborhood
734 sizes much larger than we simulated, including the recent ancestors of modern humans in northeastern
735 Europe (Ringbauer *et al.* 2017). For these species demographic inference and summary statistics are
736 likely to reflect minimal bias from spatial effects as long as dispersal is truly continuous across the
737 landscape. While that is so we caution that association studies in which the effects of population
738 structure are confounded with spatial variation in the environment are still sensitive to dispersal even
739 at these large neighborhood sizes.

740 **Other demographic models**

741 Any simulation of a population of reproducing organisms for a reasonably long amount of time
742 requires some kind of control on population sizes, or else the population will either die out or grow
743 very large. The usual choice of population regulation for population genetics – a constant size, as
744 in the Wright–Fisher model – implies biologically unrealistic interactions between geographically
745 distant parts of the species range. Our choice to regulate population size by including a local density-
746 dependent control on mortality is only one of many possible ways to do this. We could have instead
747 regulated fecundity, or recruitment, or both; this general class of models is sometimes referred to as the
748 “Bolker–Pacala model” (Bolker and Pacala 1997). It is not currently clear how much different choices of
749 vital rates, or of functional forms for the regulation, might quantitatively affect our results, although the
750 general predictions should be robust to similar forms of regulation. Since populations are still entirely
751 *intrinsically* regulated, our model still has a very strong “population genetics” flavor. Alternatively,
752 population size could be regulated by interactions with other species (e.g., a Lotka–Volterra model),
753 or extrinsically specified by local resource availability (e.g., by food or nest site availability). Indeed,
754 our model could be interpreted as a caricature of such a model: as local density increases, good
755 habitat is increasingly occupied, pushing individuals into more marginal habitat and increasing their
756 mortality. Many such models should behave similarly to ours, but others (especially those with
757 local population cycling), may differ dramatically. There is a clear need for greater exploration of the
758 consequences for population 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., Schrider and Kern 2018) for the
763 accurate estimation of dispersal from genetic data. Indeed, Ashander *et al.* (2018) found that inverse
764 interpolation on a vector of summary statistics provided a powerful method of estimating dispersal
765 distances. Expanding this approach to include the haplotype-based summary statistics studied here
766 and applying machine learning regressors built for general inference of nonlinear relationships from
767 high-dimensional data may allow precise estimation of spatial parameters under a range of complex
768 models.

769 One facet of spatial variation that we did not address in this study is the confounding of dispersal
770 and population density implicit in the definition of Wright's neighborhood size. Our simulations were
771 run under constant densities, but Guindon *et al.* (2016) and Ringbauer *et al.* (2017) have shown that
772 these parameters are identifiable under some continuous models. Similarly, though the scaling effects
773 of dispersal we show in Figure 4 should occur in populations of any total size, other aspects such as
774 the number of segregating sites are also likely affected by the total landscape size (and so total census
775 size). Much additional work remains to be done to better understand how these parameters interact to
776 shape genetic variation in continuous space, which we leave to future studies.

777 Though our simulation allows incorporation of realistic demographic and spatial processes, it
778 is inevitably limited by the computational burden of tracking tens or hundreds of thousands of
779 individuals in every generation. In particular, computations required for mate selection and spatial
780 competition scale approximately with the product of the total census size and the neighborhood
781 size and so increase rapidly for large populations and dispersal distances. The reverse-time spatial
782 Lambda–Fleming–Viot model described by Barton *et al.* (2010) and implemented by Kelleher *et al.*
783 (2014) allows exploration of larger population and landscape sizes, but the precise connection of these
784 models to forward-time demography is not yet clear. Alternatively, implementation of parallelized
785 calculations may allow progress with forward-time simulations.

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

792 Data Availability

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

794 Acknowledgements

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

797 Literature Cited

- 798 Aguillon, S. M., J. W. Fitzpatrick, R. Bowman, S. J. Schoech, A. G. Clark, *et al.*, 2017 Deconstructing
799 isolation-by-distance: The genomic consequences of limited dispersal. *PLOS Genetics* **13**: 1–27.
- 800 Al-Asadi, H., D. Petkova, M. Stephens, and J. Novembre, 2019 Estimating recent migration and
801 population-size surfaces. *PLoS genetics* **15**: e1007908.
- 802 Allee, W. C., O. Park, A. E. Emerson, T. Park, K. P. Schmidt, *et al.*, 1949 Principles of animal ecology.
803 Technical report, Saunders Company Philadelphia, Pennsylvania, USA.
- 804 Antlfinger, A. E., 1982 Genetic neighborhood structure of the salt marsh composite, *Borrichia frutescens*.
805 *Journal of Heredity* **73**: 128–132.
- 806 Antolin, M. F., B. V. Horne, M. D. Berger, Jr., A. K. Holloway, J. L. Roach, *et al.*, 2001 Effective population
807 size and genetic structure of a piute ground squirrel (*Spermophilus mollis*) population. *Canadian
808 Journal of Zoology* **79**: 26–34.
- 809 Antonovics, J. and D. A. Levin, 1980 The ecological and genetic consequences of density-dependent
810 regulation in plants. *Annual Review of Ecology and Systematics* **11**: 411–452.
- 811 Ashander, J., P. Ralph, E. McCartney-Melstad, and H. B. Shaffer, 2018 Demographic inference in a
812 spatially-explicit ecological model from genomic data: a proof of concept for the mojave desert
813 tortoise. *bioRxiv* .
- 814 Baharian, S., M. Barakatt, C. R. Gignoux, S. Shringarpure, J. Errington, *et al.*, 2016 The great migration
815 and African-American genomic diversity. *PLOS Genetics* **12**: 1–27.

- 816 Barton, N. H., F. Depaulis, and A. M. Etheridge, 2002 Neutral evolution in spatially continuous
817 populations. *Theoretical Population Biology* **61**: 31–48.
- 818 Barton, N. H., J. Kelleher, and A. M. Etheridge, 2010 A new model for extinction and recolonization in
819 two dimensions: Quantifying phylogeography. *Evolution* **64**: 2701–2715.
- 820 Benjamini, Y. and D. Yekutieli, 2001 The control of the false discovery rate in multiple testing under
821 dependency. *The Annals of Statistics* **29**: 1165–1188.
- 822 Berg, E. E. and J. L. Hamrick, 1995 Fine-scale genetic structure of a turkey oak forest. *Evolution* **49**:
823 110–120.
- 824 Berg, J. J., A. Harpak, N. Sinnott-Armstrong, A. M. Joergensen, H. Mostafavi, *et al.*, 2018 Reduced
825 signal for polygenic adaptation of height in UK Biobank. *bioRxiv* .
- 826 Bolker, B. and S. W. Pacala, 1997 Using moment equations to understand stochastically driven spatial
827 pattern formation in ecological systems. *Theoretical Population Biology* **52**: 179 – 197.
- 828 Bolker, B. M., S. W. Pacala, and C. Neuhauser, 2003 Spatial dynamics in model plant communities:
829 What do we really know? *The American Naturalist* **162**: 135–148, PMID: 12858259.
- 830 Browning, S. R. and B. L. Browning, 2015 Accurate non-parametric estimation of recent effective
831 population size from segments of identity by descent. *The American Journal of Human Genetics* **97**:
832 404–418.
- 833 Bulik-Sullivan, B. K., P.-R. Loh, H. K. Finucane, S. Ripke, J. Yang, *et al.*, 2015 LD score regression
834 distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genetics* **47**:
835 291 EP –.
- 836 Campbell, D. R. and J. L. Dooley, 1992 The spatial scale of genetic differentiation in a hummingbird-
837 pollinated plant: Comparison with models of isolation by distance. *The American Naturalist* **139**:
838 735–748.
- 839 Champer, J., I. Kim, S. E. Champer, A. G. Clark, and P. W. Messer, 2019 Suppression gene drive in
840 continuous space can result in unstable persistence of both drive and wild-type alleles. *bioRxiv* .
- 841 Chapman, N. H. and E. A. Thompson, 2002 The effect of population history on the lengths of ancestral
842 chromosome segments. *Genetics* **162**: 449–458.
- 843 Chikhi, L., V. C. Sousa, P. Luisi, B. Goossens, and M. A. Beaumont, 2010 The confounding effects of
844 population structure, genetic diversity and the sampling scheme on the detection and quantification
845 of population size changes. *Genetics* **186**: 983–995.
- 846 Crawley, M. J., 1990 The population dynamics of plants. *Philosophical Transactions of the Royal Society
847 of London. Series B: Biological Sciences* **330**: 125–140.
- 848 Durrett, R. and S. Levin, 1994 The importance of being discrete (and spatial). *Theoretical Population
849 Biology* **46**: 363–394.
- 850 Eguiarte, L. E., A. Búrquez, J. Rodríguez, M. Martínez-Ramos, J. Sarukhán, *et al.*, 1993 Direct and
851 indirect estimates of neighborhood and effective population size in a tropical palm, *Astrocaryum
852 mexicanum*. *Evolution* **47**: 75–87.
- 853 Epperson, B., 2003 *Geographical Genetics*. Monographs in Population Biology, Princeton University
854 Press.
- 855 Felsenstein, J., 1975 A pain in the torus: Some difficulties with models of isolation by distance. *The
856 American Naturalist* **109**: 359–368.
- 857 Fournier, N. and S. Méléard, 2004 A microscopic probabilistic description of a locally regulated
858 population and macroscopic approximations. *The Annals of Applied Probability* **14**: 1880–1919.
- 859 Fox, J. and S. Weisberg, 2011 *An R Companion to Applied Regression*. Sage, Thousand Oaks CA, second
860 edition.
- 861 Garcia, J. and C. Quintana-Domeque, 2006 The evolution of adult height in europe: A brief note.
862 Working Paper .
- 863 Garcia, J. and C. Quintana-Domeque, 2007 The evolution of adult height in Europe: A brief note.
864 Economics & Human Biology **5**: 340 – 349.
- 865 Garud, N. R., P. W. Messer, E. O. Buzbas, and D. A. Petrov, 2015 Recent selective sweeps in North
866 American *Drosophila melanogaster* show signatures of soft sweeps. *PLOS Genetics* **11**: 1–32.
- 867 Griffiths, R., S. Tavaré, *et al.*, 1999 The ages of mutations in gene trees. *The Annals of Applied Probability*
868 **9**: 567–590.

- 869 Guindon, S., H. Guo, and D. Welch, 2016 Demographic inference under the coalescent in a spatial
870 continuum. *Theoretical population biology* **111**: 43–50.
- 871 Haller, B. C., J. Galloway, J. Kelleher, P. W. Messer, and P. L. Ralph, 2019 Tree-sequence recording
872 in SLiM opens new horizons for forward-time simulation of whole genomes. *Molecular Ecology
873 Resources* **19**: 552–566.
- 874 Haller, B. C. and P. W. Messer, 2019 Slim 3: Forward genetic simulations beyond the Wright–Fisher
875 model. *Molecular biology and evolution* **36**: 632–637.
- 876 Harris, K. and R. Nielsen, 2013 Inferring demographic history from a spectrum of shared haplotype
877 lengths. *PLOS Genetics* **9**: 1–20.
- 878 Haworth, S., R. Mitchell, L. Corbin, K. H. Wade, T. Dudding, *et al.*, 2019 Apparent latent structure within
879 the UK Biobank sample has implications for epidemiological analysis. *Nature communications* **10**:
880 333.
- 881 Huillet, T. and M. Möhle, 2011 On the extended Moran model and its relation to coalescents with
882 multiple collisions. *Theoretical Population Biology* pp. –.
- 883 Jasper, M., T. Schmidt, N. Ahmad, S. Sinkins, and A. Hoffmann, 2019 A genomic approach to inferring
884 kinship reveals limited intergenerational dispersal in the yellow fever mosquito. *bioRxiv* .
- 885 Jay, F., P. Sjödin, M. Jakobsson, and M. G. Blum, 2012 Anisotropic isolation by distance: The main
886 orientations of human genetic differentiation. *Molecular Biology and Evolution* **30**: 513–525.
- 887 Kang, H. M., J. H. Sul, S. K. Service, N. A. Zaitlen, S.-y. Kong, *et al.*, 2010 Variance component model to
888 account for sample structure in genome-wide association studies. *Nature Genetics* **42**: 348 EP –.
- 889 Kang, H. M., N. A. Zaitlen, C. M. Wade, A. Kirby, D. Heckerman, *et al.*, 2008 Efficient control of
890 population structure in model organism association mapping. *Genetics* **178**: 1709–1723.
- 891 Kelleher, J., A. Etheridge, and N. Barton, 2014 Coalescent simulation in continuous space: Algorithms
892 for large neighbourhood size. *Theoretical Population Biology* **95**: 13 – 23.
- 893 Kelleher, J., A. M. Etheridge, and G. McVean, 2016 Efficient coalescent simulation and genealogical
894 analysis for large sample sizes. *PLoS Comput Biol* **12**: 1–22.
- 895 Kelleher, J., K. R. Thornton, J. Ashander, and P. L. Ralph, 2018 Efficient pedigree recording for fast
896 population genetics simulation. *PLOS Computational Biology* **14**: 1–21.
- 897 Kerster, H. W., 1964 Neighborhood size in the rusty lizard, *Sceloporus olivaceus*. *Evolution* **18**: 445–457.
- 898 Khera, A. V., M. Chaffin, K. G. Aragam, M. E. Haas, C. Roselli, *et al.*, 2018 Genome-wide polygenic
899 scores for common diseases identify individuals with risk equivalent to monogenic mutations.
900 *Nature Genetics* **50**: 1219–1224.
- 901 Kingman, J., 1982 The coalescent. *Stochastic Processes and their Applications* **13**: 235 – 248.
- 902 Law, R., D. J. Murrell, and U. Dieckmann, 2003 Population growth in space and time: Spatial logistic
903 equations. *Ecology* **84**: 252–262.
- 904 Lawson, D. J., N. M. Davies, S. Haworth, B. Ashraf, L. Howe, *et al.*, 2019 Is population structure in the
905 genetic biobank era irrelevant, a challenge, or an opportunity? *Human Genetics* .
- 906 Liu, X. and Y.-X. Fu, 2015 Exploring population size changes using SNP frequency spectra. *Nature
907 Genetics* **47**: 555 EP –.
- 908 Lloyd, M., 1967 ‘Mean crowding’. *Journal of Animal Ecology* **36**: 1–30.
- 909 Lundgren, E. and P. L. Ralph, 2019 Are populations like a circuit? Comparing isolation by resistance to
910 a new coalescent-based method. *Molecular Ecology Resources* **19**: 1388–1406.
- 911 Martin, A. R., M. Kanai, Y. Kamatani, Y. Okada, B. M. Neale, *et al.*, 2019 Clinical use of current polygenic
912 risk scores may exacerbate health disparities. *Nature Genetics* **51**: 584–591.
- 913 Maruyama, T., 1972 Rate of decrease of genetic variability in a two-dimensional continuous population
914 of finite size. *Genetics* **70**: 639–651.
- 915 Mathieson, I. and G. McVean, 2012 Differential confounding of rare and common variants in spatially
916 structured populations. *Nature Genetics* **44**: 243 EP –.
- 917 Mazet, O., W. Rodríguez, S. Grusea, S. Boitard, and L. Chikhi, 2015 On the importance of being
918 structured: instantaneous coalescence rates and human evolution—lessons for ancestral population
919 size inference? *Heredity* **116**: 362 EP –.
- 920 Miles, A. and N. Harding, 2017 *cgh/scikit-allel*: v1.1.8.
- 921 Neel, M. C., K. McKelvey, N. Ryman, M. W. Lloyd, R. Short Bull, *et al.*, 2013 Estimation of effective

- 922 population size in continuously distributed populations: there goes the neighborhood. *Heredity* **111**:
 923 189 EP –.
- 924 Novembre, J. and M. Slatkin, 2009 Likelihood-based inference in isolation-by-distance models using
 925 the spatial distribution of low-frequency alleles. *Evolution* **63**: 2914–2925.
- 926 Pamilo, P., 1983 Genetic differentiation within subdivided populations of formica ants. *Evolution* **37**:
 927 1010–1022.
- 928 Patterson, N., A. L. Price, and D. Reich, 2006 Population structure and eigenanalysis. *PLOS Genetics* **2**:
 929 1–20.
- 930 Peter L Ralph, J. G., Jerome Kelleher and J. Ashander, ???? .
- 931 Petkova, D., J. Novembre, and M. Stephens, 2015 Visualizing spatial population structure with esti-
 932 mated effective migration surfaces. *Nature Genetics* **48**: 94 EP –.
- 933 Price, A. L., N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, *et al.*, 2006 Principal
 934 components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*
 935 **38**: 904 EP –.
- 936 Pritchard, J. K., M. Stephens, and P. Donnelly, 2000 Inference of population structure using multilocus
 937 genotype data. *Genetics* **155**: 945–959.
- 938 Ptak, S. E. and M. Przeworski, 2002 Evidence for population growth in humans is confounded by
 939 fine-scale population structure. *Trends in Genetics* **18**: 559–563.
- 940 Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, *et al.*, 2007 Plink: A tool set for
 941 whole-genome association and population-based linkage analyses. *The American Journal of Human
 942 Genetics* **81**: 559 – 575.
- 943 R Core Team, 2018 *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical
 944 Computing, Vienna, Austria.
- 945 Ralph, P. and G. Coop, 2013 The geography of recent genetic ancestry across Europe. *PLoS Biol* **11**:
 946 e1001555.
- 947 Ralph, P., K. Thornton, and J. Kelleher, 2019 Efficiently summarizing relationships in large samples: a
 948 general duality between statistics of genealogies and genomes. *bioRxiv* .
- 949 Ringbauer, H., G. Coop, and N. H. Barton, 2017 Inferring recent demography from isolation by distance
 950 of long shared sequence blocks. *Genetics* **205**: 1335–1351.
- 951 Robledo-Arnuncio, J. J. and F. Rousset, 2010 Isolation by distance in a continuous population under
 952 stochastic demographic fluctuations. *Journal of Evolutionary Biology* **23**: 53–71.
- 953 Rossine, F. W. S., 2014 *Espaço e diversificação: uma perspectiva teórica*. Master's dissertation in ecologia:
 954 Ecossistemas terrestres e aquáticos, University of São Paulo, São Paulo : Instituto de Biociências.
- 955 Rousset, F., 1997 Genetic differentiation and estimation of gene flow from F-statistics under isolation
 956 by distance. *Genetics* **145**: 1219–1228.
- 957 Rousset, F. and R. Leblois, 2011 Likelihood-based inferences under isolation by distance: Two-
 958 dimensional habitats and confidence intervals. *Molecular Biology and Evolution* **29**: 957–973.
- 959 Sawyer, S., 1977 On the past history of an allele now known to have frequency p. *Journal of Applied
 960 Probability* **14**: 439–450.
- 961 Schiffels, S. and R. Durbin, 2014 Inferring human population size and separation history from multiple
 962 genome sequences. *Nature Genetics* **46**: 919 EP –.
- 963 Schrider, D. R. and A. D. Kern, 2018 Supervised machine learning for population genetics: A new
 964 paradigm. *Trends in Genetics* **34**: 301 – 312.
- 965 Sharbel, T. F., B. Haubold, and T. Mitchell-Olds, 2000 Genetic isolation by distance in *Arabidopsis
 966 thaliana*: biogeography and postglacial colonization of Europe. *Molecular Ecology* **9**: 2109–2118.
- 967 Sheehan, S., K. Harris, and Y. S. Song, 2013 Estimating variable effective population sizes from multiple
 968 genomes: A sequentially Markov conditional sampling distribution approach. *Genetics* **194**: 647–662.
- 969 Shirk, A. J. and S. A. Cushman, 2014 Spatially-explicit estimation of Wright's neighborhood size in
 970 continuous populations. *Frontiers in Ecology and Evolution* **2**: 62.
- 971 Slatkin, M. and N. H. Barton, 1989 A comparison of three indirect methods for estimating average
 972 levels of gene flow. *Evolution* **43**: 1349–1368.
- 973 Sohail, M., R. M. Maier, A. Ganna, A. Bloemendal, A. R. Martin, *et al.*, 2018 Signals of polygenic
 974 adaptation on height have been overestimated due to uncorrected population structure in genome-

- wide association studies. bioRxiv .
- St. Onge, K. R., A. E. Palmé, S. I. Wright, and M. Lascoux, 2012 Impact of sampling schemes on demographic inference: An empirical study in two species with different mating systems and demographic histories. *G3: Genes, Genomes, Genetics* **2**: 803–814.
- Städler, T., B. Haubold, C. Merino, W. Stephan, and P. Pfaffelhuber, 2009 The impact of sampling schemes on the site frequency spectrum in nonequilibrium subdivided populations. *Genetics* **182**: 205–216.
- Tajima, F., 1989 Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**: 585–595.
- Terhorst, J., J. A. Kamm, and Y. S. Song, 2016 Robust and scalable inference of population history from hundreds of unphased whole genomes. *Nature Genetics* **49**: 303 EP –.
- Turchin, M. C., C. W. Chiang, C. D. Palmer, S. Sankararaman, D. Reich, *et al.*, 2012 Evidence of widespread selection on standing variation in Europe at height-associated SNPs. *Nature Genetics* **44**: 1015 EP –.
- Wahlund, S., 1928 Zusammensetzung von populationen und korrelationserscheinungen vom standpunkt der vererbungslehre aus betrachtet. *Hereditas* **11**: 65–106.
- Wakeley, J., 1999 Nonequilibrium migration in human history. *Genetics* **153**: 1863–1871.
- Wakeley, J., 2009 *Coalescent Theory, an Introduction*. Roberts and Company, Greenwood Village, CO.
- Wakeley, J. and T. Takahashi, 2003 Gene genealogies when the sample size exceeds the effective size of the population. *Mol Biol Evol* **20**: 208–213.
- Wickham, H., 2016 *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.
- Wilke, C. O., 2019 *cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2'*. R package version 0.9.4.
- Wilkins, J. F., 2004 A separation-of-timescales approach to the coalescent in a continuous population. *Genetics* **168**: 2227–2244.
- Wilkins, J. F. and J. Wakeley, 2002 The coalescent in a continuous, finite, linear population. *Genetics* **161**: 873–888.
- Wright, S., 1931 Evolution in Mendelian populations. *Genetics* **16**: 97.
- Wright, S., 1943 Isolation by distance. *Genetics* **28**: 114–138.
- Wright, S., 1946 Isolation by distance under diverse systems of mating. *Genetics* **31**: 336.
- Young, A. I., M. L. Frigge, D. F. Gudbjartsson, G. Thorleifsson, G. Bjornsdottir, *et al.*, 2018 Relatedness disequilibrium regression estimates heritability without environmental bias. *Nature Genetics* **50**: 1304–1310.
- Young, H. J., 1988 Neighborhood size in a beetle pollinated tropical aroid: effects of low density and asynchronous flowering. *Oecologia* **76**: 461–466.
- Yu, J., G. Pressoir, W. H. Briggs, I. Vroh Bi, M. Yamasaki, *et al.*, 2005 A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics* **38**: 203 EP –.

Comparisons with Stepping-Stone Models

We also compared our model results to a regular grid of discrete populations, which is commonly used as an approximation of continuous geography. An important reason that this approximation is often made is that it allows more efficient, coalescent simulations; we implemented these using `msprime` (Kelleher *et al.* 2016). In this class of models we imagine an $n \times n$ grid of populations exchanging migrants with neighboring populations at rate m . If these models are good approximations of the continuous case we expect that results will converge as $n \rightarrow \infty$ (while scaling m appropriately and keeping total population size fixed), so we ran simulations while varying n from 5 to 50 (Table A1). To compare with continuous models we first distributed the same “effective” number of individuals across the landscape as in our continuous-space simulations (≈ 6100 , estimated from θ_π of random-mating continuous-space simulations). We then approximate the mean per-generation dispersal distance σ given a total landscape width W as the product of the probability of an individual

1025 being a migrant and the distance traveled by migrants: $\sigma = 4m(W/n)$. This means that m in different
 1026 simulations with the same σ scales with \sqrt{n} . We ran 500 simulations for each value of n while
 1027 sampling σ from $U(0.2, 4)$. We then randomly selected 60 diploid individuals from each simulation
 1028 (approximating diploidy by combining pairs of chromosomes with contiguous indices within demes)
 1029 and calculated a set of six summary statistics using the scripts described in the summary statistics
 1030 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

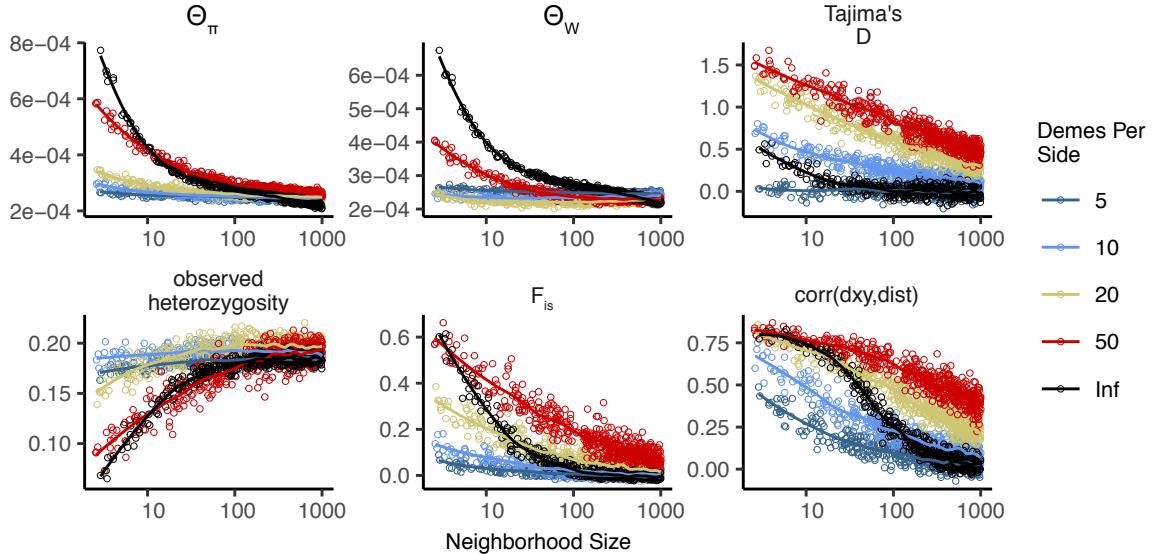


Figure A1 Summary statistics for 2-dimensional coalescent stepping-stone models with fixed total N_e and varying numbers of demes per side. The black “infinite” points are from our forward-time continuous space model. Inter-deme migration rates are related to σ as described above.

1031 In general we find many of the qualitative trends are similar among continuous and stepping-stone
 1032 models and that, in most cases, statistics from stepping-stone models approach the continuous model
 1033 as the resolution of the grid increases. For example, θ_π is inflated at low neighborhood sizes (i.e. low
 1034 m), and the extent of the inflation increases to approach the continuous case as the resolution of the
 1035 landscape increases. Similar patterns are observed for F_{is} and observed heterozygosity. However, θ_W
 1036 behaves differently, with increased grid resolution leading to lower values. This in turn drives an even
 1037 more positive Tajima’s D in grid simulations at small neighborhood sizes.

1038 These differences relative to our continuous model mainly reflect two shortcomings of the reverse-
 1039 time stepping stone model. If we simulate a coarse grid with relatively large populations in each
 1040 deme, we cannot accurately capture the dynamics of small neighborhood sizes because mating within
 1041 each deme remains random regardless of the migration rate connecting demes. This likely explains

1042 the trends in π , observed heterozygosity, and F_{is} . However increasing the number of demes while
 1043 holding the total number of individuals constant results in small within-deme populations for which
 1044 even the minimum sample size of 1 approaches the local N_e (Table A1). This results in an excess of
 1045 short terminal branches in the coalescent tree, which decreases the total branch length and leads to
 1046 fewer segregating sites, deflated θ_W , and inflated Tajima's D . Overall then our continuous model
 1047 reproduces important features of spatial structure approximated by reverse-time stepping-stone models
 1048 at moderate neighborhood sizes while avoiding some artifacts caused binning the landscape into
 1049 discrete demes.

1050 Demographic model

1051 Local population regulation is controlled by two parameters, L , and K . Here, we show that these
 1052 should be close to the average lifespan of an individual and the average number of individuals per
 1053 unit area, respectively. We chose our demographic model so that every individual has on average $1/L$
 1054 offspring each time step, and if the local population density of an individual is n , then their probability
 1055 of survival until the next time step is (equation (1)):

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

1056 We capped survival at 0.95 so that we would not have exceptionally long-lived individuals in sparsely
 1057 populated areas – otherwise, an isolated individual might live for a very long time. Since $1 - p \approx$
 1058 $n/(K(1+L))$, mortality goes up roughly linearly with the number of neighbors (on a scale given by
 1059 K), as would be obtained if, for instance, mortality is due to agonistic interactions. Ignoring migration,
 1060 a region is at demographic equilibrium if the per-capita probability of death is equal to the birth rate,
 1061 i.e., if $1 - p = 1/L$. (Note that there is no effect of age in the model, which would make the analysis
 1062 more complicated.) Solving this for n , we get that in a well-mixed population, the equilibrium density
 1063 should be around

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

1064 individuals per unit area. At this density, the per-capita death rate is $1/L$, so the mean lifetime is L .
 1065 This equilibrium density is *not* K , but (since $L = 4$) is two-thirds larger. However, in practice this model
 1066 leads to a total population size which is around K multiplied by total geographic area (but which
 1067 depends on σ , as discussed above). The main reason for this is that since offspring tend to be near
 1068 their parents, individuals tend to be “clumped”, and so experience a higher average density than the
 1069 “density” one would compute by dividing census size by geographic area (Lloyd 1967). To maintain a
 1070 constant expected total population size would require making (say) K depend on σ ; however, typical
 1071 local population densities might then be more dissimilar.

¹⁰⁷² **Supplementary Figures and Tables**

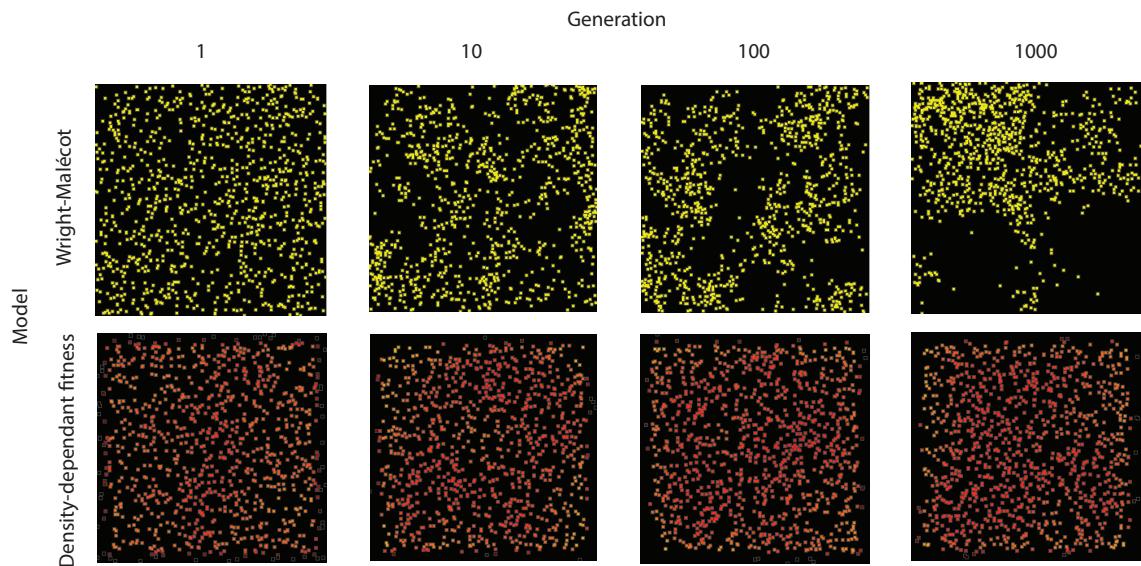


Figure S1 Maps of individual locations in a continuous-space Wright-Malécot model with independent dispersal of all individuals (top) and under our continuous space model incorporating density-dependant fitness (bottom). The clustering seen in the top row is the “Pain in the Torus” described by Felsenstein (1975).

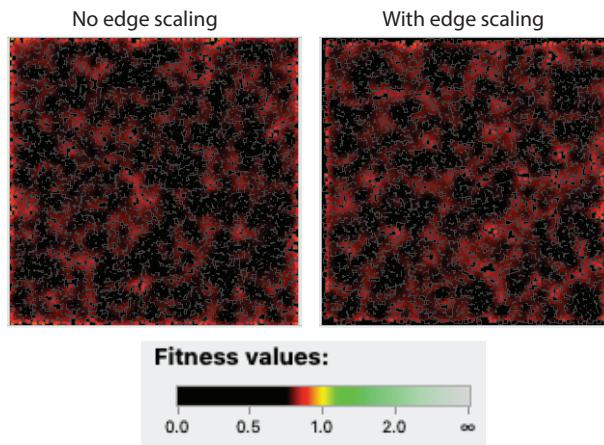


Figure S2 Comparison of individual fitness across the landscape in simulations with (right) and without (left) a decline in fitness approaching range edges. Note the slight excess of high-fitness individuals at edges on the left, which is (partially) counteracted by the scaling procedure.

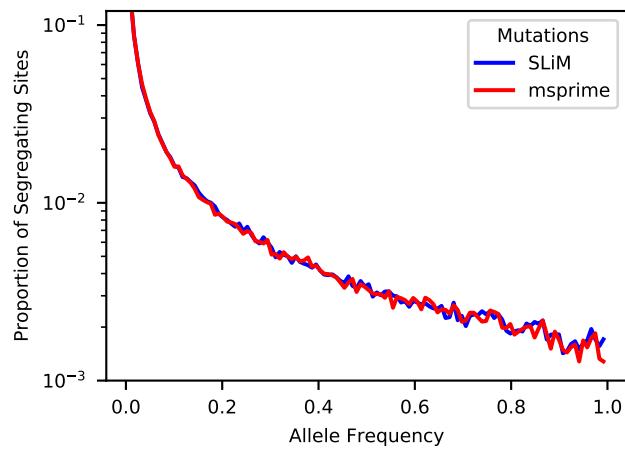


Figure S3 Site frequency spectra from a simulation with neighborhood size = 12.5 when mutations are recorded directly in SLiM (blue line) or applied later in msprime (red line).

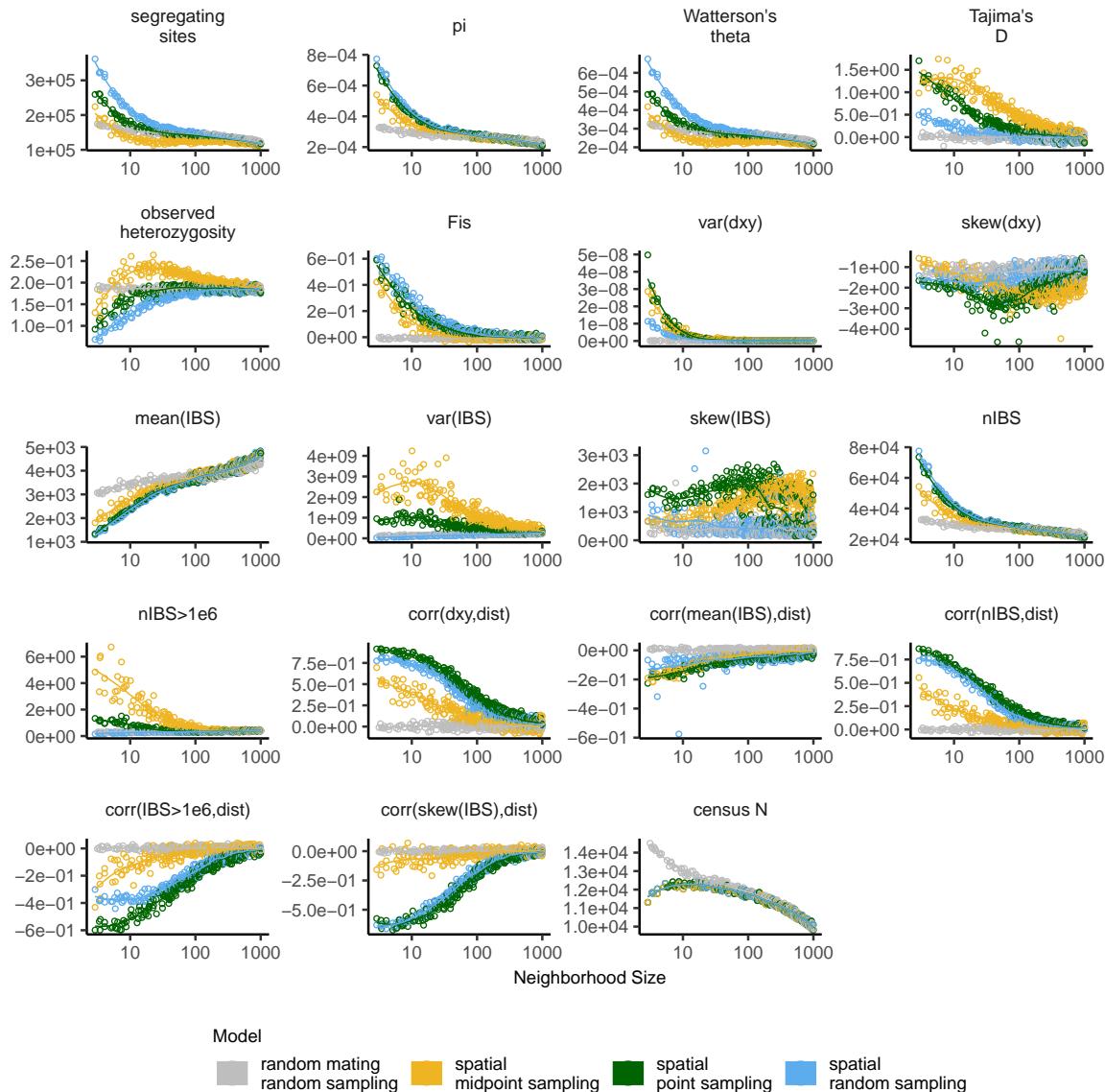


Figure S4 Change in summary statistics by neighborhood size and sampling scheme calculated from simulated sequence data of 60 individuals.

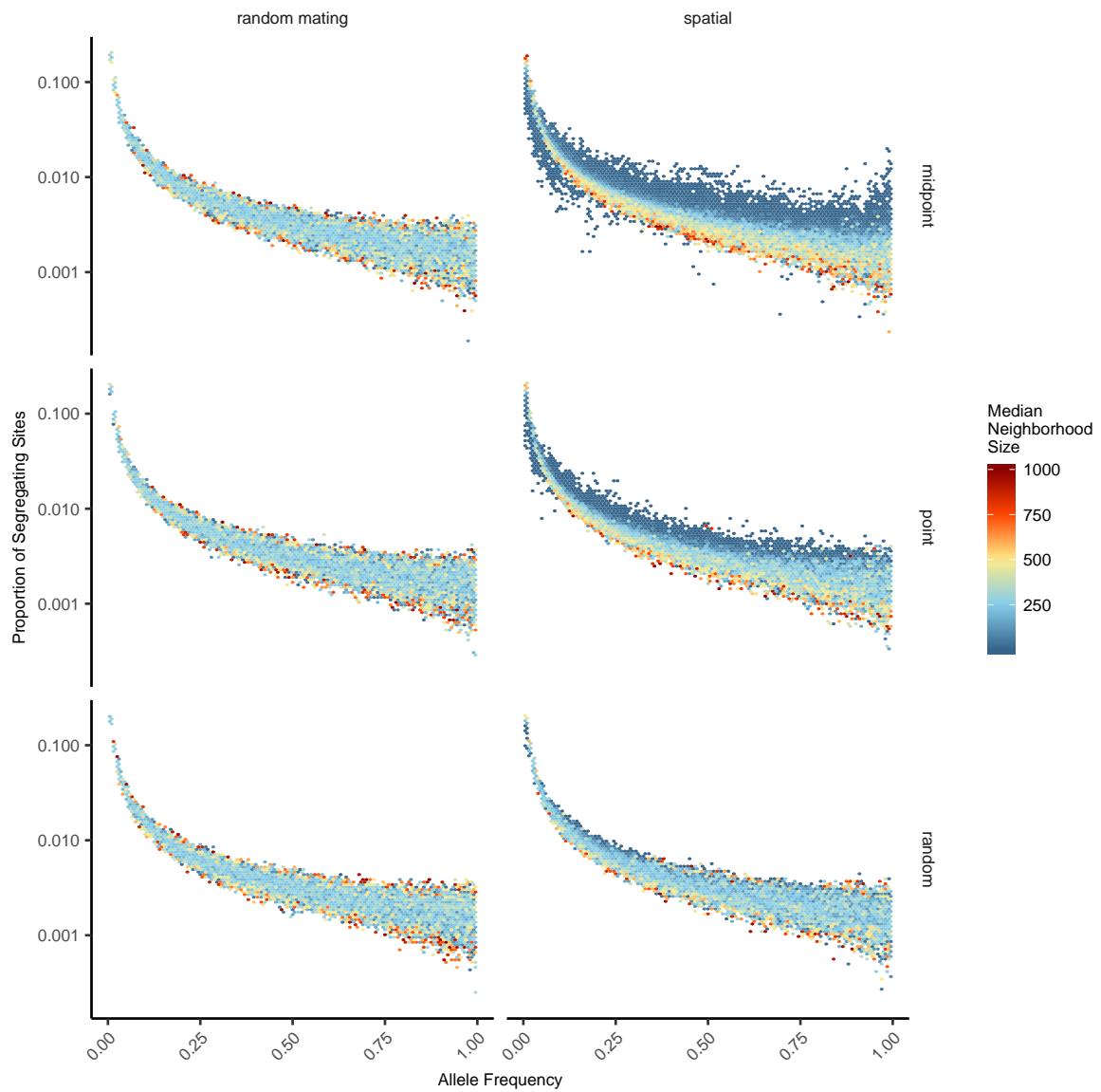


Figure S5 Site frequency spectra for random mating and spatial SLiM models under all sampling schemes.

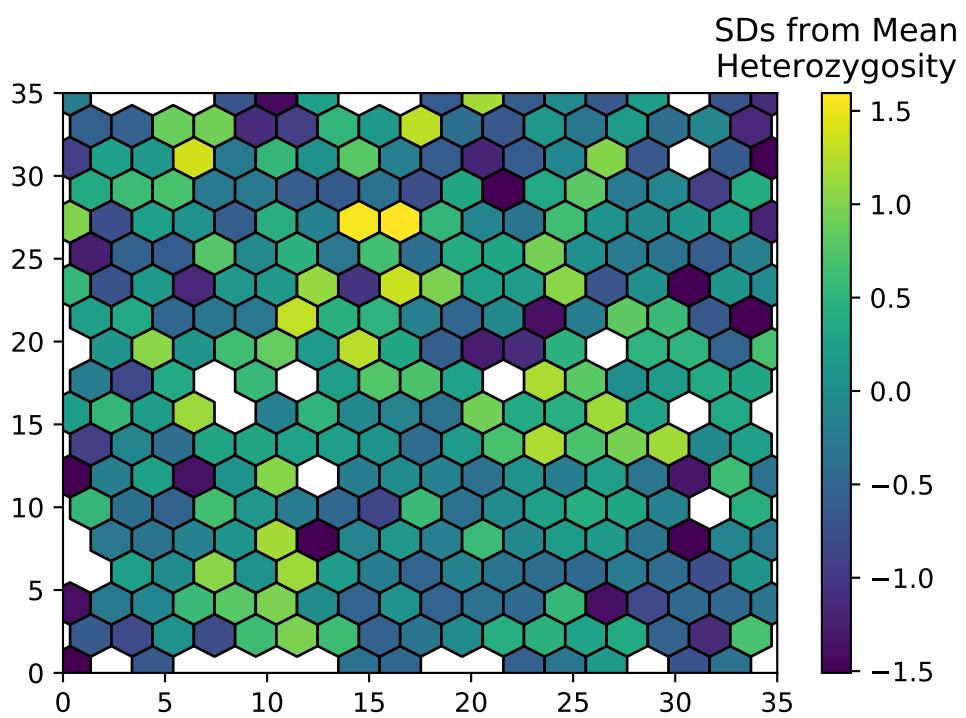


Figure S6 Variation in observed heterozygosity (i.e. proportion of heterozygous individuals) in hexagonal bins across the landscape, estimated from a random sample of 200 individuals from the final generation of a simulation with neighborhood size ≈ 25 . Values were Z-normalized for plotting.

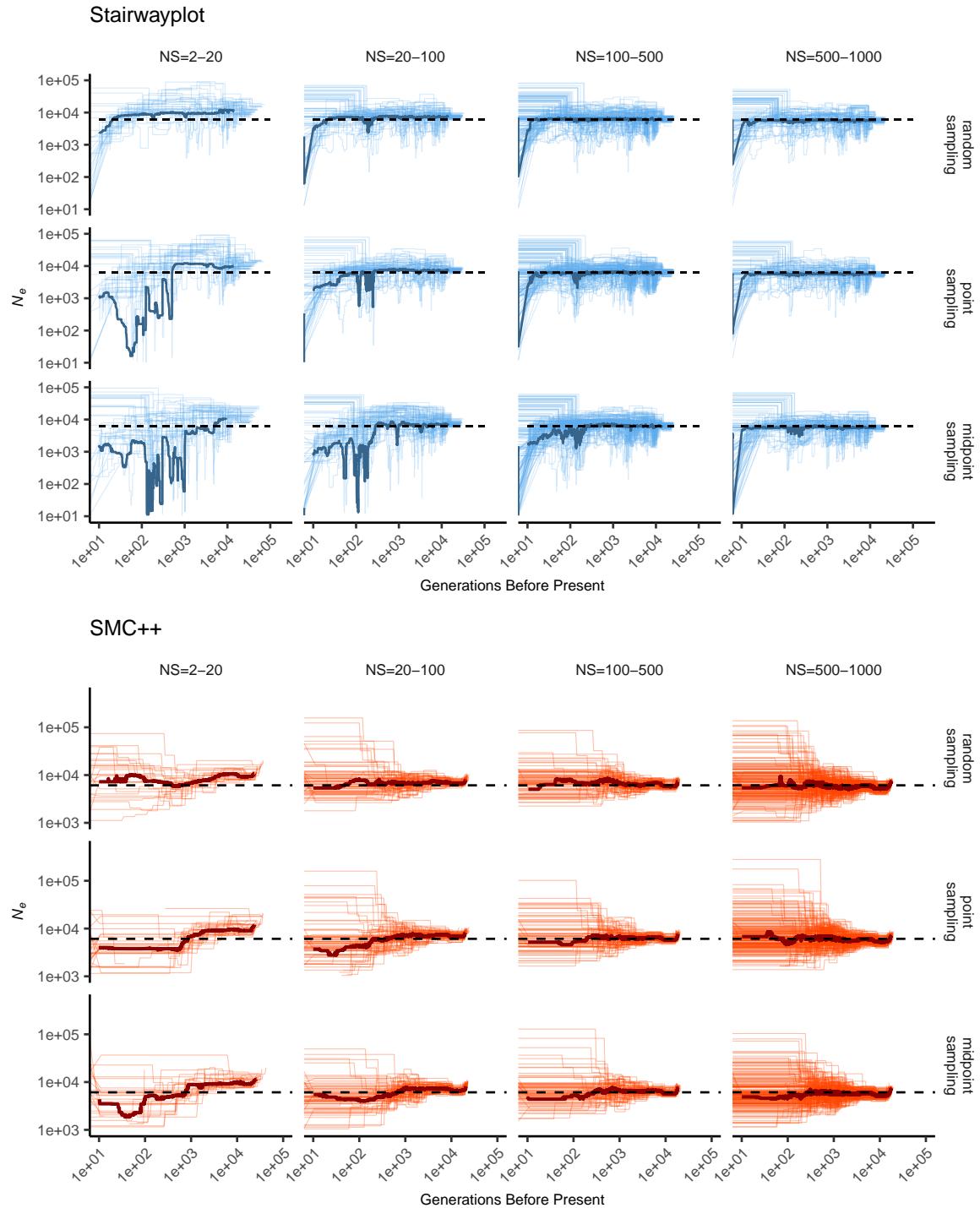


Figure S7 Inferred demographic histories for spatial SLiM simulations, by sampling scheme and neighborhood size (NS) range. Thick lines are rolling medians across all simulations in a bin and thin lines are best fit models for each simulation. Dashed horizontal lines are the average N_e across random-mating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

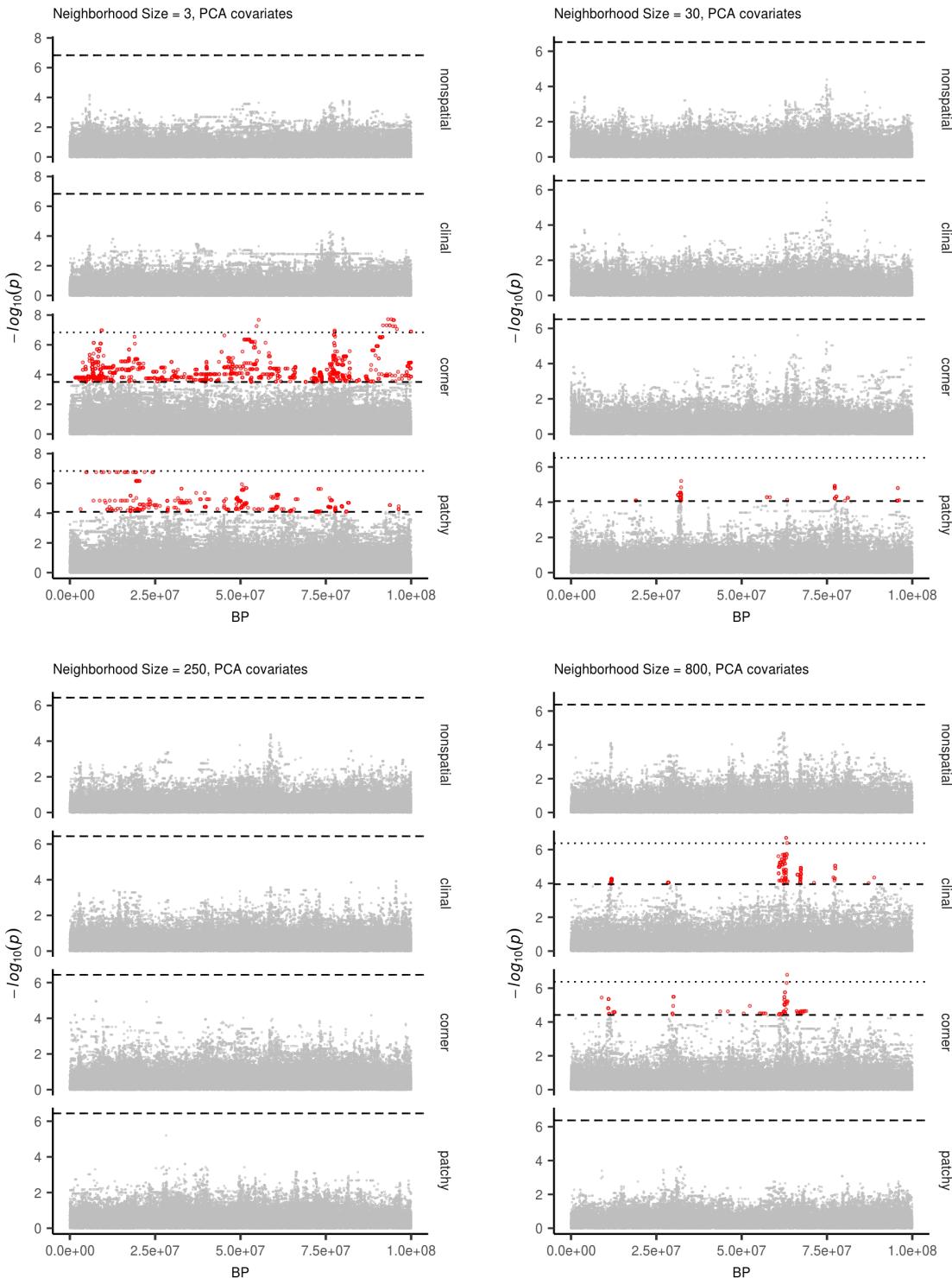


Figure S8 Manhattan plots for a sample of simulations at varying neighborhood sizes. Labels on the right of each plot describes the spatial distribution of environmental factors (described in the methods section of the main text). Points in red are significantly associated with a nongenetic phenotype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

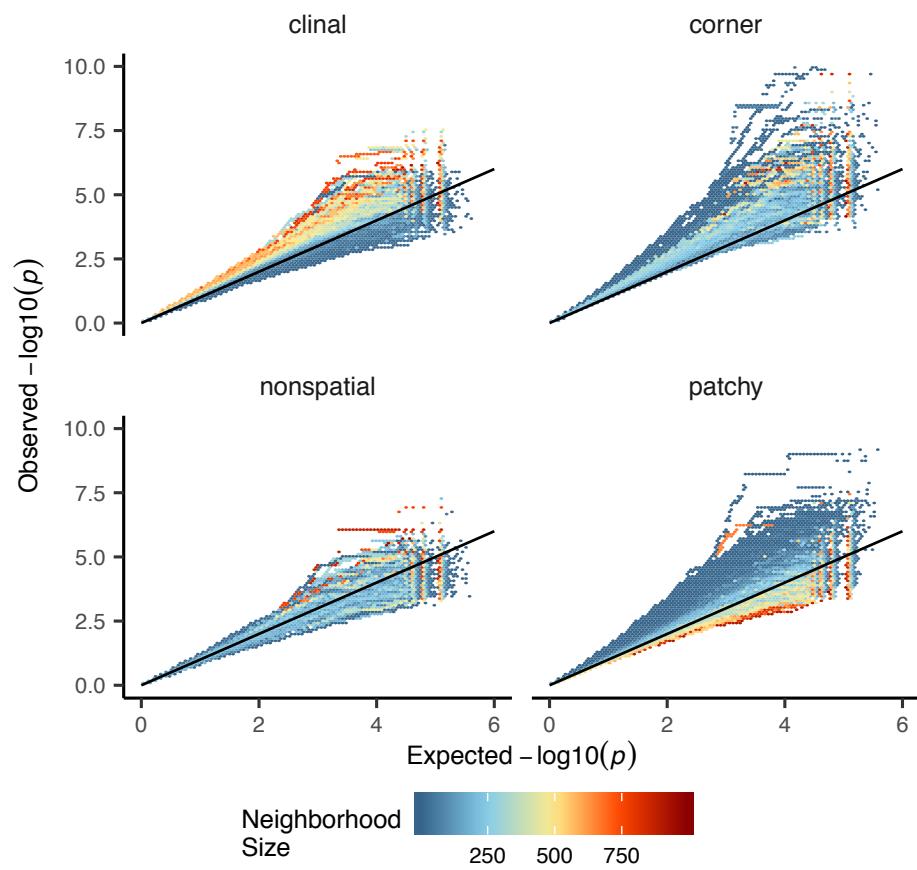


Figure S9 Quantile-quantile plots showing observed $-\log_{10}(p)$ for PC-corrected GWAS run on simulations with varying neighborhood sizes and environmental distributions. Hexagonal bins are colored by the average neighborhood size of simulations with points falling in a given region of quantile-quantile 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
Thursday 5th December, 2019

To the Editor(s) –

We are writing to submit a revised version of our manuscript, "Space is the Place: Effects of Continuous Spatial Structure on Analysis of Population Genetic Data". We thank the AE and reviewers for constructive comments, which we have addressed in detail, below. The main changes we have made have to do with our choice of model. We have added some more discussion of the biological motivation for the particular model of continuous space we use (with references to the ecological literature, where it appears) to the Methods (around (p. 3, l. 130)), and in the Discussion (at (p. 21, l. 743)). We have included more explanation of some of the more opaque consequences of a more realistic demographic model, around (p. 8, l. 342). Finally, we have added a substantial comparison to the discrete stepping-stone model, most of which is in the Appendix.

Sincerely,

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

Reviewer AE:

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

This is a good point. There are many ways to go in the quest for demographic realism, and the fundamental question – what regulates population density in real populations – is still very much an open question (or rather, a question with a great many answers). As highlighted below, we have included further discussion of our particular choice of local, density-mediated control on mortality, along with pointers into the literature, at several places in the manuscript; see in particular in the Methods when we introduce the model (p. 3, l. 139), and (p. 21, l. 743).

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

Reply: Done. (p. 1, l. 35)

(AE.2) (p. 3, l. 139) “Such models have been used extensively in ecological modeling but rarely in population genetics”: Detailing these previous uses via citations and elaboration may help alleviate the major concern about the provenance of this model and its unique behaviors (see general comments above and R2 and R3 comments).

Reply: Good idea – we have added some historical discussion and a few more citations to this section, (p. 3, l. 139). Also see the additional discussion in the Discussion (p. 21, l. 743).

(AE.3) (p. 4, l. 186) Please describe computation time needed per replicate

Reply: We have added a figure (Figure 3) and short discussion (p. 9, l. 359) of run times.

(AE.4) (p. 22, l. 796) I read the acknowledgement to the Hearth and Falling Sky Brewing with a sense of familiarity in feeling of gratitude to my own favorite cafes and breweries, but I it's not a great precedent for Acknowledgements to be filled this way. Please cut.

Reply: Good point; we have done this.

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

Reply: This is done.

(AE.6) Figure 4A, S5: Perhaps more revealing to show on log-log scale?

Reply: Good suggestion – the SFS in Fig 4A is now on a log-log scale, which shows the slight decrease in low frequency SNPs a little better. We've left Fig S5 as-is.

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

Reply: Thanks for catching this - we have revised this caption to add details including the simulation parameters. (Supplemental Figure S6).

Reviewer 1:

Overall, the authors explore an important but often neglected source of bias that can affect inference in many population-based studies (in medical genetics, evolutionary biology and ecology). This study can be of interest to a broader audience of readership, and I have only minor comments to improve clarity and increase accessibility for readers:

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

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

Reply: The mean IBS is simply the mean of the vector of lengths of IBS tracts (p. 6, l. 235). The short answer is that if one splits a chromosome up into more pieces (increasing $n\text{IBS}$), the mean length of those pieces must necessarily be smaller. At low neighborhood sizes, spatially distant individuals tend to have longer coalescence times, leading to more (and hence smaller) IBS tracts. This can be seen in Supplemental Figure S4. Note that $\text{skew}(IBS)$ and number of long ($>1\text{e}6$) IBS tracts reflect other aspects of the distribution.

(1.2) The authors use K to denote both carrying capacity (p. 4, l. 164) and population density (p. 4, l. 167). It might be better to use a different notation for these quantities since carrying capacity is fixed while density is an emergent quantity in the non-Wright-Fisher model. Use of K to denote carrying capacity and density is a bit confusing. For example, on (p. 7, l. 319) it is said that 'the "population density" (K) and "mean lifetime" (L) parameters were the same in all simulations'. Here K seems to indicate carrying capacity rather than density? The latter is an emergent quantity and varies across simulation runs?

Reply: We agree and have adjusted our language to underscore the point that K is a parameter that controls population density, rather than being equal to it, at (p. 4, l. 167) and (p. 5, l. 204) and (p. 7, l. 319).

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

Reply: We have added to the text at the beginning of the results to emphasize that this analysis is necessary because these parameters are emergent rather than fixed (p. 7, l. 318).

(1.4) (p. 6, l. 254) Perhaps 'Demographic Inference' might better reflect the content of this section.

Reply: Good suggestion – we have changed the section heading to "demographic inference" (p. 6, l. 254).

(1.5) (p. 7, l. 285) This sentence with 'Gaussian noise with mean zero and standard deviation 10' is confusing since it was mentioned earlier that the modelled phenotype must vary as human height across Europe, and human height varies 2 standard deviations. Only after reading the whole paragraph it becomes clear that 'standard deviation 10' here refers to unit of height. Please consider rephrasing this sentence.

Reply: We have revised this sentence to clarify that we aim to produce a variation in mean phenotype of two standard deviations across the landscape (p. 7, l. 285).

(1.6) (p. 7, l. 307) In the sentence, 'We also examined p values for systemic inflation' I think the authors meant 'systematic inflation'.

Reply: Whoops; thanks. Fixed.

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

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

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

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

(1.9) (p. 9, l. 368) The phrase 'affect summaries of variation' is better to replace with 'summaries of genetic variation'.

Reply: Done. (p. 9, l. 368)

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

Reply: Thanks for catching this, should be fixed now.

(1.11) There are also several typos and errors in the text. For example, (p. 7, l. 319); (p. 19, l. 655).

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

Reviewer 2:

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

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

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

Thanks very much for the encouraging words!

(2.1) To what extent are any of the results dependent on the exact method of simulation. There are a number of choices about the exact details of the simulations (e.g. the way the overlapping generations are handled, the edge effects and, particularly, the form of Equation 1 - see below). It's not so much that these are non-standard (since I don't think there is a standard) and they all sort of make sense heuristically, and I was left wondering whether these sorts of choices actually make a difference. Do the authors have some thoughts/intuition/results about that? Given that the results in Figure 4 seem quite consistent with expectations, I suspect that on some level it doesn't make much difference but then there are intermediate results like Figure 2 which seem a bit counter-intuitive and I wonder if those aspects depend on the simulation scheme.

Reply: This is a good point also raised by other reviewers. We have added some more discussion of the choice of demographic model ((p. 3, l. 139) and (p. 21, l. 743)).

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

Reply: We have added an appendix comparing our model to a reverse-time stepping stone simulation on several relevant summary statistics (p. 26, l. 1013). For some of these statistics (e.g., θ_π), the behavior of the stepping stone model becomes more similar (although not identical) to the continuous model as the resolution of the landscape increases, but other statistics are more strongly affected by discretization. We are also curious how all the other analyses we perform would be affected using other simulation schemes, but hope to explore that aspect in future work. Furthermore, discrete-deme coalescent simulations are certainly faster, but come with significant issues either when neighborhood size is lower than the population of a deme, or when demes are small enough that the sample size approaches local N_e .

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

Reply: Additional discussion of the model is now provided in the appendix (see "Demographic model") (p. 28, l. 1051)

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

Reply: Good suggestion – we have added a supplemental figure (Figure S2) to show the distribution of individual fitness across the landscape with and without our edge-scaling approach.

(2.5) It would be helpful provide the figure showing that generating mutations during the forward simulations in SLIM is equivalent to applying mutations using msprime on pre-generated trees (p. 5, l. 193)? It sounds like this procedure would underestimate the variance in the number of mutations, since you remove the effect of random generation time. Is this effect small?

Reply: Theory says that any difference here would be extremely small and would affect only the variance, not the mean. Nonetheless, we have added a figure showing sample site frequency spectra generated from a subset of simulations run with SLiM mutations, and then using msprime to apply

mutations to the same tree sequences with our generation-time scaling approach (Figure S3). These approaches yield extremely similar spectra.

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

Reply: We agree that this is hard to intuit: there are a lot of factors at work. We have added some more discussion of these phenomena (p. 8, l. 342).

(2.7) Figure 8D: I am surprised by the extent to which the observed values of $-\log_{10}(p)$ fall below the $y = x$ line. Particularly in the lower right panel for large neighbourhood sizes. I would expect that to be close to panmictic - why are the P-values underdispersed? That seems like a potential bug, or else something weird is going on.

Reply: We have checked the code to the best of our abilities and did not find a bug causing the underdispersion, but we have switched our visualization in Figure 8D to now show a subset of qqplots as lines, rather than the hexagonal binning over all simulations. This lets us see the density in addition to the range of values, and shows that relatively few points fall below the 1:1 lines. Those that do seem to reflect overcorrection in the regression when using PC coordinates as covariates – the PCA is capturing some information about the spatial genetic variation which itself covaries only weakly with the phenotype, and as a result we see anomalously low $-\log_{10}(p)$ when regressing genotype against phenotype.

(2.8) (p. 20, l. 711) It might be worth citing (Haworth et al. 2019), who do the proposed test (GWAS for birth location) in UK Biobank to illustrate the population structure.

Reply: Done - thank you for pointing us to this study. (p. 20, l. 711)

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

Reply: We are also very interested to know how mixed models perform here, but think that adding a second GWAS method would make this section too large for the current paper. We have added a note to (p. 19, l. 695) specifically citing mixed models as alternate methods that may perform better.

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

Reply: We apologize – we had forgotten to make the repository public. This has been corrected.

Reviewer 3:

The present study deals with a “hot topic” in spatial population genetics. Most inferential and descriptive methods in statistical spatial population genetic rely on a discrete approximation of space and it is not clear what impact this approximation may have when individuals migrate along a continuum instead. Spatial patterns in sampling is also another major issue which is often simply dismissed, mainly because of the paucity of statistical methods to deal with it. This work touches on these important issues in a timely manner.

We are glad you agree this is an important topic to explore!

Although I was enthusiastic about the topic, I was quite disappointed with the core of the study, i.e., the forward-in-time simulation of populations in continuous space. The field has been struggling with this issue for decades – examples of spectacular failures like the Wright-Malecot model (see Felsenstein’s “pain in the torus” article, 1975) or, more recently, the “mugration” or “discrete trait analysis” model in phylodynamics (see De Maio et al. 2015) have probably mostly harmed our research field – that one cannot make the economy of using a sound probabilistic model for generating geo-referenced genetic data. It does not seem to be the case here unfortunately.

We were surprised that the reviewer seems to be saying that local, density-mediated control of mortality doesn’t constitute a “sound probabilisitic model”, but hope that our additional discussion of the choice, and citations to the history of this sort of model, help to better motivate our choice.

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

Reply: This is a natural question that we asked ourselves while developing the simulation using the built-in visualization tools of the SLiM GUI. We have now added a supplementary figure (Figure S1) showing the distribution of individuals in our density-dependant spatial model and a continuous-space Wright-Fisher simulation without density dependance. We encourage the reviewer to run our SLiM recipe using the GUI at different parameter values to see for themselves how quickly equilibrium is attained – we are tempted to include movies of this as supplementary material, but that seems excessive. Also note that we did *not* include competitive interaction to avoid clumping – we included it to make the model biologically realistic (so that, for instance, the population does not grow without bound), and lack of clumping is a *consequence* of this choice. We have clarified this in the text. (p. 3, l. 137)

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

Reply: As in the last point, as part of sanity checking our models we monitored the population size over time while with SLiM. All simulations maintained asymptotic populations, and none of the runs we started ended because populations crashed. We agree that the population model will die out eventually with probability 1 (it is a Markov process with a single accessible absorbing state (death), and since it isn’t a branching process, doesn’t grow without bound). However, it will take an extremely long time to hit that state, so that extinction is not a practical consideration. We have added a clarifying note about this. (p. 3, l. 136)

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

Reply: We explained the relationship here, in a new Appendix ('Demographic Model'; (p. 28, l. 1051)).

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

Reply: We apologize – we had forgotten to make the repository public. Now it is.

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

Reply: Hopefully, the additional plots and discussion around demographic modeling (notably, the discussion at (p. 21, l. 743)) helps to increase the transparency in this model. We agree that demographic realism is a very important consideration, and due to the often hard-to-intuit nature of spatial demography, simulations need to be carefully sanity checked. (Furthermore, what is “realistic” for one species will not be for another!) We have added an Appendix comparing our model to a stepping-stone (p. 26, l. 1013), because we think these are the most familiar and widely used class of spatial models. We find that many features of our model are well approximated by stepping-stone models, and that for statistics like θ_{π} the stepping stone model results approach our continuous space model as the number of demes used to describe the landscape increases.

We definitely think that a comparison to the spatial Lambda-Fleming-Viot would be an extremely useful thing to do, for the reasons the reviewer mentions, but this is beyond the scope of the current study. The SLFV model is only known to be an approximation of one particular biologically explicit model (of patchy extinction-recolonization) and is only conjectured to be a good approximation for spatial models more generally. To reiterate, the definition of the SLFV was explicitly motivated by a desire to approximate spatial models, especially Bolker-Pacala models like ours. Therefore, our study does important preliminary work for simulation study of the spatial Lambda-Fleming-Viot, since it works through some of the issues in simulating biologically realistic spatial models, and describes some of the patterns.

(3.6) Figure 2: I do not understand why the neighborhood size varies to the same extent in the random mating model as it does for the spatial model. For the random mating model, I would have expected the neighborhood size to be equal to the census size since all individuals have the same probability of being a parent of any given offspring. From [the paragraph at (p. 5, l. 210)], it is clear that the spatial model would converge to the random mating model when the mean parent-offspring distance tends to infinity only if we were to ignore the impact of range edges. I am thus wondering whether the variation of neighborhood size one observes in Fig 2 for the random mating model is just a consequence of border effects. If that is the case, then the authors should state it clearly and try to justify it from a biological perspective.

Reply: It’s a good point that neighborhood size, as defined to be “number of possible parents”, should indeed be the census size. However, we are using N_W defined in the same way for both spatial and random mating simulations as a way of comparing spatial to “nonspatial” models keeping other aspects of demography (almost) the same. This is indeed part of what’s going on. We have added some discussion of the census size scaling at (p. 5, l. 210) and (p. 8, l. 342), but haven’t added a detailed discussion of whether or not most species experience decreased habitat quality at range edges, since that seems like it is getting too far into the weeds.

(3.7) (p. 21, l. 729) “Many more species occur in a middle range of neighborhood sizes between 100 and 1000 - a range in which spatial processes play a minor role in our analyses [...]. Do the authors think that the spatial processes would still play a minor role when neighborhood sizes exceed 100-1000 if the habitat was larger than that taken in the present simulations? It would also probably be useful to mention that neighborhood sizes given in Table 1 should be compared with extreme caution since the size of the corresponding habitats vary across species. More generally, I suspect that the size of the habitat has a substantial impact on the vast majority of statistics examined in this study. Indeed, the mean parent-offspring distance, which is at the core of the definition of Wright’s neighborhood size, is only small or large relative to the size of the habitat.

Reply: This is a good point. Wright’s work (Wright 1943) suggests some aspects of genetic variation such as variance in allele frequencies and inbreeding coefficients can be estimated by looking only at what he would later (Wright 1946) call “neighborhood size”, but certainly other aspects like the number of segregating sites will also depend on total landscape size. We now note on (p. 21, l. 720) that we have evaluated only one landscape size, and have added a sentence to the discussion noting that exploration of these patterns in varying landscape sizes is an important avenue for further research (p. 22, l. 775).

(3.8) (p. 22, l. 771) Please add a reference to Guindon, Guo and Welch (2016). This study clearly shows that population density and dispersal parameters are identifiable and can indeed be estimated in practice under the spatial Lambda-Fleming-Viot model.

Reply: Done. Thank you for pointing us to this study. (p. 22, l. 771)

Reviewer 4:

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

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

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

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

Reply: We have expanded our citations of some of the relevant ecology literature (p. 3, l. 130), which we hope helps to contextualize the study better. We haven’t cited dadi, which works with discrete populations (the “diffusion” is in frequency space).

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

Reply: We agree that not noting EEMS was an oversight. We have added the most recent EEMS paper to the citations at (p. 1, l. 38), and have added an Appendix giving a more thorough comparison with stepping-stone models.

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

Reply: We now discuss haplotype methods in the relevant discussion section (p. 18, l. 628). However though these methods should be more accurate for recent events it is not clear that this will improve performance per se. The dips in recent inferred Ne from stairwayplot are not just prediction noise, but actually reflect an underlying genealogy in which terminal branches are shorter than expected from a constant-size random-mating population (see e.g. figure 4A and 5). The interpretation error is that these short branches are generated by spatial structure rather than changes in population size over time – a point also made in the (Mazet *et al.* 2015) paper we discuss in the introduction and discussion.

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

Reply: Good question – we are also interested to know how mixed models perform here, but think that to properly test that we would want to change our design to generate phenotypes from simulated genotypes. This would allow us to evaluate false-negatives in addition to false-positives. This is important because, if mixed models do provide stronger control for stratification they are also likely to remove true signals of causal SNPs if those SNPs covary with spatial structure. We now point to these methods explicitly in the discussion (p. 19, l. 695), but think that incorporating that study here would make this paper too long and we plan to explore this timely issue in the future. We also think the PC results are still quite relevant as the method is quite population in current studies.

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

Reply: Apologies, it was accidentally set to private. The link should work now (and has been replaced by the primary link, <https://github.com/kern-lab/spaceness>).

(4.7) Definitions and interpretations. There are quite a large number of metrics discussed in Figure 4, and it's a lot to take in. It might be helpful to have a table with a reminder of what the metric is, its interpretation, and how it is computed.

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

(4.8) (p. 4, l. 156) Notation: "Offspring disperse a Gaussian-distributed distance away from the parent with mean zero and standard deviation σ in both the x and y coordinates. Each offspring is produced with a mate selected randomly from those within distance 3σ , with probability of choosing a neighbor

at distance x proportional to $\exp(-x^2/2\sigma^2)$." I think x may be overloaded here, or I'm confused. Clarify?

Reply: We have clarified the notation for the second instance (referring to the distance among individuals) to d . (p. 4, l. 156)

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

Reply: We have rearranged the figure as suggested.

(4.10) (p. 7, l. 293) *Not sure I follow this example: "Concretely, an individual at position (x, y) in a 50×50 landscape has mean phenotype $100 + 2x/5$."*

Reply: We have clarified by switching to "the phenotype p for an individual at location (x,y) is then $p = 100 + 2x/5$ ". (p. 7, l. 293)

(4.11) (p. 7, l. 319) *Minor typo (through vs though): This occurs because, even through the "population density" (K) and "mean lifetime" (L) parameters...*

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

(4.12) *Define NS abbreviation in Figure 6.*

Reply: Done.