Space is the Place: How Dispersal Shapes Genetic Variation in Continuous Space

C.J. Battey[*,1], **Peter Ralph**[†,1] **and Andrew Kern**[‡,2]
[*]University of Oregon Dept. Biology, Institute for Ecology Evolution

ABSTRACT Individuals exist in continuous space, but standard models in population genetics are based on discrete randomly-mating populations exchanging migrants. As the availability of population-level, whole-genome data allows inference of increasingly fine-scale patterns of ancestry in many species, models incorporating realistic demographic and spatial processes are needed to accurately describe spatial structure and control for its impacts on analyses of selection and demography. Here we implement a forward-time continuous-space model of molecular evolution and use it to study the impacts of limited dispersal in continuous landscapes on population genetic summary statistics, demographic inference, and association studies. Low dispersal slows the geographic spread of ancestry and inflates branch lengths for mid-frequency alleles because of slow coalescence among distant individuals. This inflation can bias demographic inference from the site frequency spectrum towards inferring a recent spike in population sizes. In a GWAS setting spatially correlated phenotypes cause spurious signals of genetic association when principal components analyses fail to capture the relevant spatial scale of phenotypic variation, which can occur under either very limited or very high dispersal. Last we use a machine learning regression approach to evaluate summary statistics as estimators of dispersal distance and find that correlation coefficients between spatial distance and moments of the distribution of pairwise haplotype block lengths retain strong signal even at high dispersal distances.

KEYWORDS Space; Population Structure; Demography; Haplotype block sharing

Introduction

The inescapable reality that biological organisms live, move, and reproduce in continuous spatial landscapes has been all but ignored in population genetic models.

In most sexual organisms individuals mate with other nearby individuals, leading to a positive correlation between genetic and geographic distances. This pattern of "isolation by distance" (**?**) is one of the most widely replicated empirical findings in population genetics (**???**), but is only heuristically approximated by existing forward-time models of molecular evolution based on discrete demes exchanging migrants. As the availability of population-level whole-genome data allows inference of increasingly fine-scale patterns of ancestry in many species, models incorporating biologically realistic demographic and spatial processes are needed to accurately describe continuous spatial structure and control for its impacts on analyses of selection and demography. This situation is particularly clear in human population genetics, where individual ancestries are often interpreted as mixtures of discrete populations in part because of modeling assumptions underlying discrete clustering algorithms and association analyses are often strongly dependent on statistical corrections for population stratification (**?**).

The best-studied approaches to population genetics in continuous space were developed by (**?**) and (**?**), who derived expressions for pairwise genetic differentiation in continuous space assuming Poisson distributed numbers of offspring and independent dispersal among individuals. However, (**?**) showed that these assumptions are incompatible – over time, a population meeting them will clump into a small number of geographic clusters occupying only a part of the available range. (NOTE: where to put Rousset et al. lit?)

NOTE: probably cut this paragraph???

This process is analogous to genetic drift, with spatial locations and population densities taking the place of sites and allele frequencies. Because individuals disperse from their natal location, new offspring are likely to be close to their parents. If by chance an individual has no offspring the local population density is then lowered, which decreases the chance of new in-

Manuscript compiled: Tuesday 5th March, 2019
[3]University of Oregon Dept. Biology, Institute for Ecology and Evolution.

dividuals dispersing into that area. This process compounds over time until only one or two patches of habitat are occupied at any given time. The resulting spatial clustering clearly does not describe spatial dynamics in most real populations, because declining growth rates in regions of high population density are common in both lab ( ) and wild populations ( ).

One method for modeling populations in continuous space is then to assume the existence of a grid of discrete randomly-mating populations connected by migration, which prevents clustering by forcing all regions to be occupied in every generation. Though good approximations of continuous structure given high dispersal (Barton(?)), these models are not truly continuous and limit investigation of spatial structure below the level of the deme. An alternative approach is to model the geographic spread of ancestry backwards in time through a diffusion approximation – an approach that has recently made significant progress in modeling both dispersal and demographic parameters (**????**). However, these models currently lack a well-described forward-time analog that produces biologically realistic spatial patterns.

A direct approach to the clustering problem of classical forward-time models is to incorporate density dependence by scaling individual fitness to local population density. This shifts reproductive output towards regions of low-density, preventing populations from clustering in space over time. This approach was taken previously by (**?**) who used an individual based model with continuous space and density dependent fitness to study the probability of speciation along continuous environmental gradients. (**Sentence referencing other continuous IBM's with density dependence**). However to our knowledge all previous implementations of continuous space models have focused on individuals or at most a few genotypes as the unit of analysis, which limits our ability to investigate the impacts of continuous space on genome-wide genetic variation as is now routinely sampled from real organisms.

Here we describe an implementation of an individual-based model in continuous space that incorporates overlapping generations, a Gaussian dispersal kernel, and density-dependent fitness, and scales to chromosome-scale alignments across tens of thousands of individuals. We use this model to describe the impacts of variation in dispersal distance on a variety of summary statistics typically analyzed under discrete population models, and apply a machine-learning regression approach to evaluate a new class of summary statistics based on correlations between geographic distance and moments of the distribution of pairwise haplotype block lengths as estimators of dispersal distance. We then examine how the fine-scale spatial structures occurring under limited dispersal impact demographic inference from the site frequency spectrum; finding that inflation of branch lengths for mid-frequency alleles associated with slow coalescence of distant individuals **TBD**.

Last, we use our model to study the impacts of isolation by distance in continuous space on genome-wide association studies (GWAS). Variation in ancestry proportions between case and control cohorts will cause inflation of test statistics at SNPs with different allele frequencies among populations (CITE). In addition, because most phenotypes are influenced by the environment and environmental factors are often spatially correlated, subtle spatial structure within and among populations can lead to inflated effect size estimates in GWAS of quantitative traits conducted on samples from large geographic regions (**?**). For example, imagine a population of flowering plants occurring over a large latitudinal range with populations subject to isolation by distance. Flowering time is strongly dependent on photoperiod and temperature, both of which typically vary with latitude. If we were to conduct a GWAS for flowering time without correcting for population structure we would find that all SNPs with allele frequencies varying clinally with latitude were significantly associated with flowering time, even if flowering time was determined purely by the environment.

A common method of correcting for population structure in GWAS is to run a principal components analysis on the genotype matrix and include PC coordinates as covariates in a linear regression for each SNP (**?**). Mixed model GWAS employs a different statistical framework but takes a similar approach by incorporating a kinship matrix describing pairwise genetic relatedness among samples as a random effect in a linear mixed model (**?**). In both approaches the key question is whether or not the ancestry components of the model adequately describe "background" levels of genetic variation in the sample relative to any confounding effects of ancestry and phenotype. Recent studies in European and British populations (**???**) have found evidence of residual population stratification in effect size estimates from large-scale GWAS employing both linear and mixed model approaches, suggesting that existing structure corrections may be inadequate. For traits controlled by a few large effect loci this confounding effect is likely to be relatively small, but for analyses of complex traits with thousands of putatively associated SNPs the potential for confounding is high. Here we simulate a range of spatially correlated phenotypes for simulated individuals and seek to identify conditions under which existing GWAS structure corrections fail.

Materials and Methods

A Forward-Time Model of Evolution in Continuous Space

We implemented our model using the non-Wright-Fisher module in the program SLiM v3.0 (**?**). Each time step consists of three stages: reproduction, dispersal, and competition. To reduce the parameter space we use the same parameter, denoted σ, to modulate the spatial scale of interactions at all three stages by adjusting the standard deviation of the corresponding Gaussian functions. As in previous work (**??**), σ as applied in our dispersal step is approximately equal to the mean parent-offspring distance.

At the beginning of the simulation individuals are distributed randomly on a continuous landscape. Mates are selected proportional to distances among individuals weighted by a Gaussian function with mean $1/(2\pi\sigma^2)$ and standard deviation σ. Wright's **?** "Neighborhood Size", defined as $4\pi\sigma^2 K$ where K is the population density, is then the approximate number of individuals available for mating in our simulation. The number of offspring is drawn from a Poisson distribution with mean $1/L$, where L is the average lifespan. If new offspring are produced, we simulate dispersal by taking two draws from a normal distribution with mean 0 and standard deviation σ and adding these to the x and y coordinates of the first parent.

Density-dependent competition is simulated by adjusting the probability of survival of each individual proportional to the local density of individuals up to a distance 3σ away, scaled according to the same Gaussian function used for mate selection. Given a per-unit carrying capacity K and average fecundity F, the probability of survival d for individual i after adjustment for density is:

$$d_i = min(0.95, \frac{1}{1 + \rho * \sum_i G(x)}) \quad (1)$$

where

$$\rho = F/((1 + F) * K) \quad (2)$$

and $G(x)$ is the Gaussian function used to weight distances to other individuals.

**NOTE: better ways to write this down? Should we write out the normal function too?

A major challenge in all spatial models is dealing with range edges. When local population density is used to model competition, edge or corner populations can be assigned artificially high fitness values because they lack neighbors within their interaction radius but outside the bounds of the simulation. We approximate a decline in habitat suitability near edges by decreasing the probability of survival proportional to the square root of distance to edges in units of σ. The final probability of survival for individuals within one σ of an edge is

$$s_i = d_i min(1, \sqrt{x_i/\sigma}) min(1, \sqrt{y_i/\sigma}) \\ min(1, \sqrt{(x_{max} - x_i)/\sigma}) min(1, \sqrt{(y_{max} - y_i)/\sigma}) \quad (3)$$

where x and y are spatial coordinates.

The full history of the genomes of final-generation individuals are stored after each timestep by SLiM and recorded as tree sequences (**?**). To speed up our simulations we set the mutation rate to 0 in SLiM simulations and later added mutations to the resulting tree sequences with msprime (**?**). Note that because time in SLiM is measured in timesteps rather than generation time, the desired per-generation rate must be scaled by generation time in this procedure (see "Genealogical Parameters" below for further information).

We compared our model output to a two-dimensional stepping-stone model and a one-population coalescent model implemented in msprime (**?**). To isolate spatial effects from other components of the model such as overlapping generations and density-dependent probability of survival, we also ran a second version of the SLiM model with random mate selection and dispersal.

We ran 400 simulations for the spatial and random-mating SLiM models on a square landscape of size 50 with per-unit carrying capacity $K = 5$ (census $N \approx 10,000$), average lifetime $L = 4$, genome size $1e8$, and recombination rates $1e-9$. A mutation rate of $1e-8 mutations/site/generation$ was then applied to the coalesced tree sequence with msprime. σ values were drawn from a uniform distribution bounded by 0.2 and 4. Simulations were run until all extant individuals shared a common ancestor within the simulation (i.e. the tree sequence had coalesced). Coalescent stepping-stone simulations were run on a 5x5 grid with a total N_e of 10,000 (N_e per population = 400) and 10 haploid samples per population; allowing migration between all adjacent populations. We ran 500 simulations drawing the number of migrants exchanged by adjacent demes from a uniform distribution between 0 and 40.

Genealogical Parameters

To study the geographic spread of genealogical ancestry in our simulations we output the full pedigree of all individuals over 50 generations with σ set to 0.2, 0.35, 0.5, 1, 2, and 3.5 (corresponding to Wright's Neighborhood Sizes of 2.5, 7.7, 15.7, 62.8, 251.3, and 769.7, respectively). We then selected the individual closest to the center of the landscape in the final generation of each simulation, pruned the pedigree to include only its genealogical ancestors (that is, regardless of the amount of genetic material actually inherited from a given ancestor) and plotted the age of its most recent ancestor in each region of the full landscape using the 2-dimensional summary feature of ggplot2 in R (**?**). We calculated the rate of spread of genealogical ancestry by estimating the mean and maximum distances between the focal individual and its ancestors in each generation and then fitting a linear model for distance as a function of generations using ordinary least squares regression.

Because generation time in our model is not fixed and may vary both across simulations we also calculated "effective" generation times for all simulations. To do this we output the full tree sequence of all individuals in every time step for 200 time steps. Starting at the final time step we then calculated the average age of parents in each previous generation, sampling past individuals in proportion to their genetic contribution to the final generation, which can be extracted from the tree sequence. When adding mutations to the tree sequences produced by SLiM, we scaled the mutation rate by the generation time because msprime's "mutate()" function assumes that time is measured in units of generations. So to achieve a $1e-8 mutations/site/generation$ mutation rate for a simulation in which the generation time was 5 time steps, we entered $1e-8/5$ for the mutation rate in msprime.mutate(). We then verified that this procedure produced the correct number of mutations by comparing a subset simulations with SLiM-generated mutations (which are applied only at mating events) with mutations added by msprime, finding that they were nearly identical in all cases.

Summary Statistics

We calculated a set of 20 summary statistics (Supplementary Table 1) from 60 diploid individuals sampled randomly from the final generation of each simulation using the python package scikit-allel (**?**). Statistics included common single-population summaries such as mean pairwise divergence (π), inbreeding coefficient (F_{is}), and Tajima's D, as well as the classic isolation-by-distance regression of genetic distance (D_{xy}) against the logarithm of geographic distances (**?**), which we summarized as the correlation coefficient between $log10(spatialdistance)$ and the proportion of identical base pairs among individuals.

Following recent studies that showed strong signals for dispersal and demography in the distribution of shared haplotype block lengths (**??**), we also calculated various summaries of the distribution of pairwise identical-by-state (IBS) haplotype block sharing among samples. Intuitively, haplotype block length distributions should contain more information about the genealogy of a set of samples than allele frequency measures alone because they reflect the influence of two Poisson processes (mutation and recombination) acting along its branches, rather than just one.

The full distribution of lengths of IBS tracts for each pair of individuals was first calculated with a custom python function. We then calculated the first three moments of this distribution (mean, variance, and skew) and the number of blocks over $1e6$ base pairs both for each pair of individuals and for the full distribution across all pairwise comparisons. To assess the degree to which the IBS tract length distribution shifts with spatial distance, we then estimated correlation coefficients between spatial

distance and each moment of the pairwise IBS tract distribution. Because more closely related individuals on average share longer haplotype blocks we expect that spatial distance will be negatively correlated with mean haplotype block length, and that this correlation will be strongest (i.e. most negative) when dispersal is low. The variance, skew, and count of long haplotype block statistics are meant to reflect the relative length of the right (upper) tail of the distribution, which represents the frequency of long haplotype blocks and is thus most informative of recent demographic events (**?**) (cite Coop & Ralph here & above).

Demographic Modeling

We fit single-population demographic models to the site frequency spectra of 60 randomly sampled individuals from each simulation with the program Stairwayplot (**?**). Site frequency spectra used for input data were calculated in scikit-allel. 100 bootstrap replicates were generated for each simulation by resampling genotype matrices with replacement over sites and recalculating the site frequency spectra. Models were then fit across all bootstrap replicates using default settings in Stairwayplot and the median estimate of Ne per time unit across bootstrap replicates was used to represent the output of each simulation. The resulting inferred population size histories were plotted with line colors scaled to σ to visualize how dispersal scaling impacts estimates of demographic history.

Association Studies

To assess the degree to which spatial structure confounds GWAS we simulated four types of phenotypic variation for 1000 randomly sampled final-generation individuals in each SLiM simulation and conducted a linear-regression GWAS with PC covariates in PLINK (**?**). SNPs with a minor allele frequency less than 1% were excluded from the analysis. Phenotype values were meant to roughly reflect the distribution of height across Europe, which has recently been found to be confounded with population structure in large scale GWAS (**??**). Conceptually our approach is similar to that taken in (**?**), though here we model continuous spatial variation and focus on genome-wide effects rather than low-frequency alleles specifically.

In the first simulation phenotypes for all individuals were drawn from a normal distribution with mean 110 and standard deviation 10. Next we simulated clinal environmental influences on phenotype by drawing the phenotypes from independent normal distributions in which the mean was scaled by an individual's x position such that it varied by 2 standard deviations across the map. Third, we approximate concentrated environmental effects by drawing phenotypes for individuals with x and y coordinates below 20 from a normal distribution with mean 2 standard deviations above the rest of the map. Last, we simulated a "patchy" environmental influence on phenotypes by selecting 10 random points on the map and drawing phenotypes for all individuals within two map units of any selected point from a normal distribution with mean 2 standard deviations above the rest of the map.

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

We ran linear-model GWAS both with and without the first 10 principal components as covariates in PLINK and summarized results across simulations by counting the number of significant SNPs with an expected false positive rate of less than 5% after adjusting p values with the R function p.adjust(...,method="fdr"). We also examined p values for systemic inflation by estimating the expected values from a uniform distribution (because no SNPs were used when generating phenotypes), plotting observed against expected values for all simulations, and using the two-dimensional binning feature of ggplot2 (**?**) to find the mean σ value in each region of quantile-quantile space.

Results

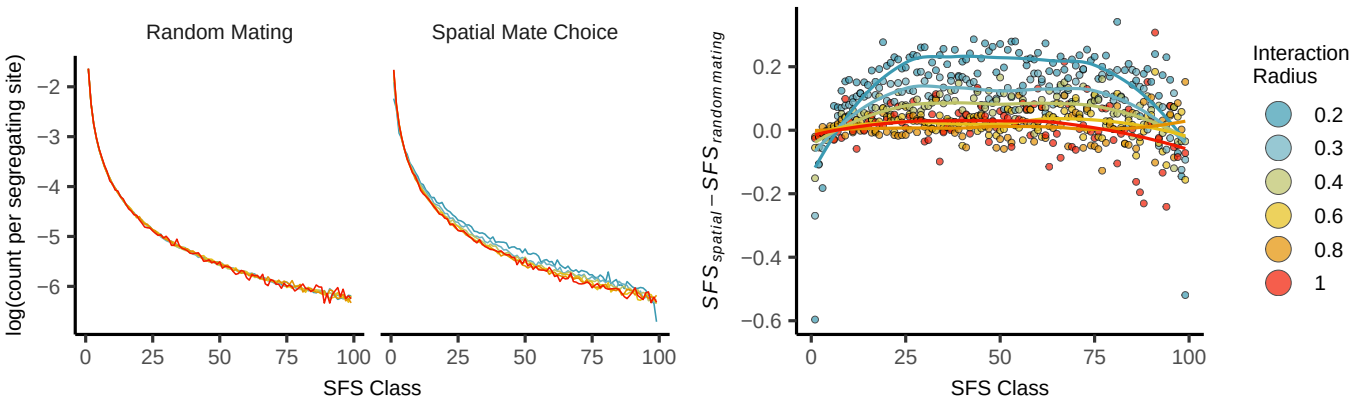

Figure 1 Scaled allele frequency distributions for random mating and spatial SLiM models under varying σ. The rightmost panel is the difference between the spatial and random-mating distributions, showing an inflation of mid-frequency sites for neighborhood sizes of less than 50. Lines are a local regression (loess) model at span 2.

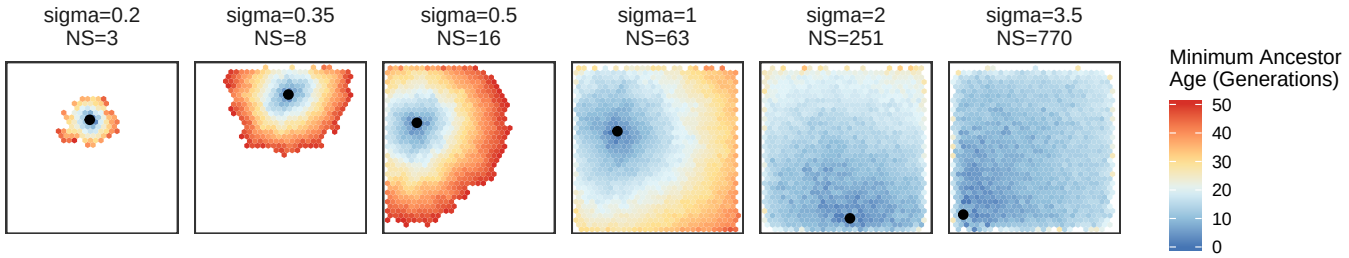

Figure 2 Geographic spread of genealogical ancestors from a random individual over 50 generations, by Neighborhood size and σ. Colors are scaled by the age of the most recent ancestor.

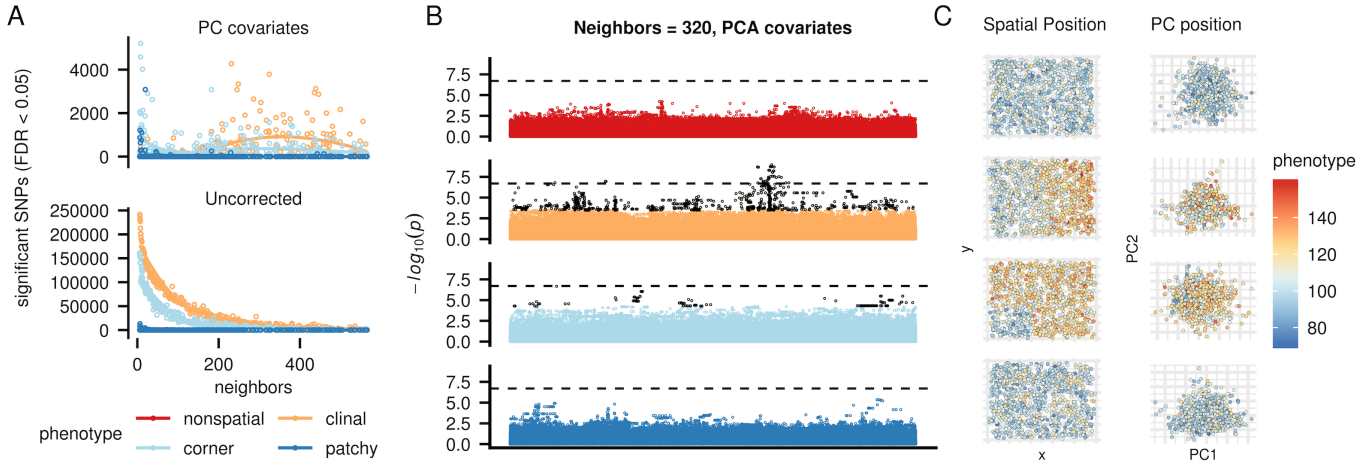

Figure 3 Impacts of spatially correlated phenotypes and isolation by distance on GWAS. Colors represent different phenotype distributions as indicated in the figure legend. A: numbers of significant SNPs after FDR correction for linear-model GWAS conducted with (top) or without (bottom) PC covariates. B: Example Manhattan plots for a single simulation with varying phenotype distributions. Points in black are significant at an estimated FDR cutoff of 0.05, while the dotted line shows the Bonferroni-corrrected p value of 0.05. In practice most human GWAS studies use a significance cutoff between these points. C and D show the corresponding spatial and principal component positions, colored by phenotype.