Received: 14 June 2020 | Revised: 2 September 2020 | Accepted: 11 September 2020

DOI: 10.1111/1755-0998.13267

RESOURCE ARTICLE

MOLECULAR ECOLOGY RESOURCES WILEY

The effect of gene flow from unsampled demes in landscape genetic analysis

Andrew J. Shirk[1] | Erin L. Landguth[2] | Samuel A. Cushman[3]

[1]Climate Impacts Group, College of the Environment, University of Washington, Seattle, WA, USA

[2]Computational Ecology Laboratory, School of Public and Community Health Sciences, University of Montana, Missoula, MO, USA

[3]USDA Forest Service, Rocky Mountain Research Station, Flagstaff, AZ, USA

Correspondence
Andrew J. Shirk, Climate Impacts Group, University of Washington, Seattle, Washington 98195, USA.
Email: ashirk@uw.edu

Funding information
National Science Foundation, Grant/Award Number: DEB-1340852 and EF-1442597

Abstract

An assumption of correlative landscape genetic methods is that genetic differentiation at neutral markers arises solely from the degree to which the intervening landscape between individuals or populations resists gene flow. However, this assumption is violated when gene flow occurs into the sampled population from an unsampled, differentiated deme. This may happen when sampling within only a portion of a population's extent or when closely related species hybridize with the sampled population. In both cases, violation of the modelling assumptions has the potential to reduce landscape genetic model selection accuracy and result in poor inferences. We used individual-based population genetic simulations in complex landscapes within a model selection framework to explore the potential confounding effect of gene flow from unsampled demes. We hypothesized that as gene flow from outside the sampling extent increased, model selection accuracy would decrease due to the formation of a hybrid zone where allele frequencies were perturbed in a way that was not correlated with effective distances between sampled individuals. Surprisingly, we found this expectation was unfounded, because the reduced accuracy due to admixture was counteracted by an increase in allelic diversity as alleles spread from the unsampled deme into the sampled population. These new alleles increased the power to detect landscape genetic relationships and even slightly improving model selection accuracy overall. This is a reassuring result, suggesting that sampling the full extent of a population or related species that may hybridize may be unnecessary, as long as other well-established sampling requirements are met.

KEYWORDS

accuracy, deme, gene flow, landscape genetics, model selection, sampling

1 | INTRODUCTION

A primary motivation for the field of landscape genetics has been to understand how complex landscapes influence gene flow among individuals and populations (Balkenhol et al., 2015). As the field matures, our knowledge of how landscape genetic approaches perform under varying evolutionary dynamics is improving (Shirk et al., 2018). Progress has been made in understanding sources of error in model selection (Dudaniec et al., 2016; Zeller et al., 2016), comparing different model selection approaches (Cushman et al., 2013; Shirk et al., 2018), quantifying environmental distances and landscape heterogeneity (McRae, 2006; Wang & Bradburd, 2014), selecting appropriate genetic distance metrics (Shirk et al., 2017), scales of selection (Zeller et al., 2017) and appropriate sampling designs (Landguth et al., 2012; Oyler-McCance et al., 2013). Despite this progress, a number of key challenges remain in landscape genetics, including issues of scale, sampling, statistical methods, theoretical frameworks and hypothesis testing (Balkenhol et al., 2015).

Among these remaining challenges, an issue that has garnered relatively little consideration to date is the potential source of model

selection error that may arise when gene flow occurs from outside the sampling extent into the sampled population. Many populations are more or less continuously distributed across heterogeneous landscapes that influence the degree to which individuals in one area are functionally connected to the rest of the population. The result is a population occurring at varying densities, with complex spatial patterns of genetic differentiation and local effective population size (Shirk & Cushman, 2014). Sampling neutral genetic markers in these populations is the empirical basis for understanding how landscapes influence gene flow. Most individual-based landscape genetic methods depend upon observing a correlation between genetic distances (based on neutral markers) that quantify genetic differentiation among sampled individuals, and landscape distances that measure the effective (e.g., cost-distances or resistance-distances) between sampled individuals given the resistance of the intervening landscape to gene flow. Typically, a researcher hypothesizes that one or more environmental variables are related to landscape resistance to gene flow for the sampled population. Next, a range of alternative parameterizations relating those variables to resistance are either proposed or a hypothesis space is searched by an algorithm (Peterman, 2018; Shirk et al., 2010). The best model among competing alternatives is then selected based on the strength of the correlation between genetic and landscape distances.

An underlying assumption of these correlative landscape genetic methods is that genetic differentiation at neutral markers arises solely from the degree to which the intervening landscape between individuals or populations resists gene flow. However, this assumption is violated when gene flow occurs into the sampled population from an unsampled, genetically differentiated deme. This results in admixture penetrating from the sampling extent boundary, resulting in altered allele frequencies and larger genetic distances between samples collected from the admixture zone compared to the rest of the sampled population. This means researchers are sampling admixed individuals near the periphery of the sampling boundary that contain alleles from both within and outside of the study's sampling extent. The result is that individuals near the population interface are more genetically divergent than less admixed individuals near the core of the sampling extent. Importantly, this increase in genetic distance is not related to landscape distances, but rather, the proximity of a differentiated and unsampled deme nearby and the magnitude of gene flow from it into the sampled population.

Unsampled demes are likely to be common in landscape genetic studies for several reasons. First, a priori, the pattern of genetic differentiation within a population is unknown before genetic sampling occurs, raising the possibility of unsampled demes within the study area extent in areas where sampling is sparse. Second, the full extent of a population may be unknown or impractical to sample due to the expense, logistical challenges, seasonal migrations and other constraints, leaving unsampled demes beyond the study area extent. Third, if the sampled population hybridizes with a closely related but unsampled species beyond the sampling extent, the resulting interspecies gene flow can influence genetic distances among sampled individuals in a way that is unrelated to the intervening landscape.

The first scenario above (lack of knowledge about population genetic structure prior to sampling) is an issue of within-population sampling that has been explored previously (Oyler-McCance et al., 2013; Schwartz & McKelvey, 2009). However, no study has yet addressed the potential confounding influence of the second and third scenarios above (inability to sample the full population extent or the presence of a related species that hybridizes with the sampled population), which have in common that gene flow occurs from unsampled demes beyond the study's sampling extent into the sampled population. Here, we used individual-based population genetic simulations to examine the effect of gene flow from outside the sampling extent on the accuracy of landscape genetic model selection. These simulations included a variety of environmental variables influencing gene flow (continuous, categorial and linear features) as well as a variety of patterns and magnitudes of resistance in the landscape. The 300 generations of the simulation also produced varying states of equilibria with the unsampled deme in which to explore the temporal dimension of this issue. We hypothesized that as gene flow from outside the sampling extent increased, model selection accuracy would decrease due to the formation of a hybrid zone where allele frequencies were altered in a way that was not correlated with effective distances between sampled individuals.

2 | MATERIALS AND METHODS

2.1 | Simulation and analysis overview

Our simulations occurred within rectangular 1,024 × 2,048 cell raster landscapes oriented on a north–south axis (Figure 1). The pattern and magnitude of resistance to gene flow within these landscapes varied depending on which environmental variables affected gene flow in the simulation (described below under "Environmental variables influencing gene flow") and a resistance parameter R_{max} that related rates of gene flow to the environmental variables (described below under "Resistance models"). We divided the landscape into northern and southern halves and placed a population of individuals in each half initially separated by a movement barrier that prevented mating and dispersal between populations. Later, we removed the barrier in some simulations (see "Population genetic simulations" below), allowing gene flow between populations. In other simulations, we kept the barrier in place to serve as a control for later comparison. We considered the northern population to be the unsampled deme in our simulations (hereafter the "unsampled population"). The southern population was sampled for landscape genetic analysis (hereafter the "sampled population") to determine if model selection accuracy was impacted by gene flow from the unsampled population. The variability in the pattern and magnitude of resistance in the total of 810 simulations we ran was designed to produce variability in rates of gene flow entering the sampled population, making it possible to robustly assess the impact of an unsampled deme on model selection accuracy (described below under "Calculating model selection

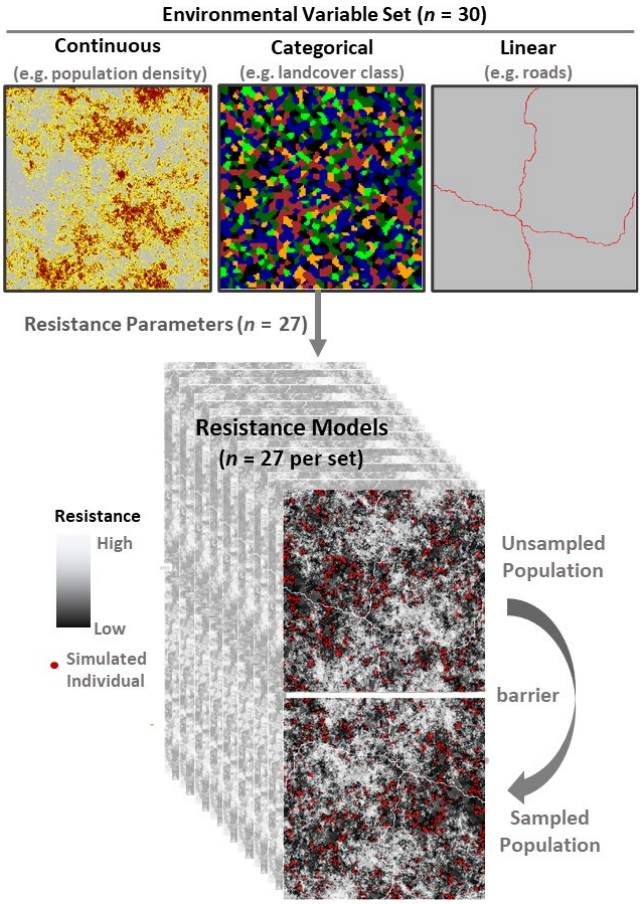

FIGURE 1 Simulated landscapes and populations. We created 30 "sets" of three environmental variables (continuous, categorical, and liner) potentially influencing gene flow in our simulations. Each variable was created within a 1,024 × 1,024 square grid and then mirror imaged to create a 1,024 × 2,048 rectangular grid. Within each set of three environmental variables, we produced 27 unique resistance models that varied according to two parameters: the number of environmental variables to include (1–3) and the maximum resistance value (R_{max}) for each variable. We created a population of 1,000 individuals in each half of the landscape (red dots) with a spatial distribution that was determined probabilistically as a function of the model in each set with the highest mean resistance. A movement barrier (white gap) separated the two populations at the beginning of the simulations [Colour figure can be viewed at wileyonlinelibrary.com]

accuracy") as well as to explore how various aspects of gene flow from an unsampled deme might influence model selection (see "Factors influencing model selection accuracy" below).

2.2 | Environmental variables influencing gene flow

We produced 30 "sets" of three environmental variables (continuous, categorical and linear) to serve as potential sources of landscape resistance to gene flow in our simulations. We produced each variable in a 1,024 × 1,024 grid placed in the northern half of our simulated landscape. We then extended a mirror image copy of each variable into the southern half of the landscape (Figure 1). The mirror image configuration created continuity at the boundary between the northern and southern halves, resulting in a seamless landscape pattern in the border region. This also removed differences in the pattern of resistance within the northern and southern populations as a confounding factor in our analysis.

We used the program QRULE version 4 (Gardner, 1999) to produce a continuous environmental variable for each of the 30 sets. This variable was intended to reflect gradients such as population density or slope. Approximately 30% of the landscape had a value of zero and the remaining cells were assigned values ranging from 1 to 128 in approximately equal proportions (i.e., a continuous variable ranging from 0 to 128). To produce a variety of patterns in the continuous variables, we varied the QRULE *H* parameter (the continuous variable for each set was produced with *H* set to 0.1, 0.5 or 0.9, with 10 sets at each level). Larger *H* values resulted in a greater aggregation of grid cells with similar values (see Figure S1 for an example).

We used the randomClusterNN function of the NLMPY version 0.1.5 python package (Etherington et al., 2015) to produce a categorical variable for each of the 30 sets. This variable was intended to reflect discrete landscape characteristics such as landcover classes or habitat patches. This function produces discrete patches of grid cells with identical values, which we reclassified such that 30% of the landscape had a value of zero and the remainder of the landscape was approximately equally apportioned into patches with values ranging from 1 to 8, representing levels of the categorical variable (e.g., landcover classes). Like the continuous variables, we created a variety of categorical variable patterns by varying the aggregation of these patches according to three levels of *H* (the categorical variable for each set was produced with *H* set to 0.1, 0.5 or 0.9, with 10 sets at each level).

We produced a linear variable (e.g., roads) for each of the 30 sets by first generating continuous QRULE landscapes as described above for continuous variables. Like the other variables, we created a variety of linear variable patterns by varying the aggregation of grid cell values according to three levels of *H* (the linear variable for each set was produced with *H* set to 0.1, 0.5 or 0.9, with 10 sets at each level). We then randomly selected a *y* coordinate ranging from 1 to 1,024 and connected the left edge grid cell at (0, *y*) to the right edge grid cell at (1,024, 1,024 − *y*) via the least-cost path given the QRULE landscape as a resistance model. With the same approach, we created a second least-cost path from the top edge to the bottom edge. This produced two irregular diagonal linear features bisecting the landscape horizontally and vertically and intersecting to form an X shape (see Figure 1 for an example). In general, high cohesion of the resistance model produced linear features that curved broadly around large patches of high resistance. Conversely, low cohesion produced linear features that were more direct, but locally circuitous as they weaved around smaller more fragmented patches of high resistance. Both paths were assigned a value of 1 and all other grid cells were set equal to zero.

2.3 | Resistance models

We created 30 "sets" of resistance models to use in our simulations described below. Each set was based on 27 unique parameterizations that we used to transform the environmental variables described above into resistance and then combine them together into a single surface. Each set contained one each of the continuous, categorical and linear landscape variables described above. Thus, the 27 models in each set represented alternative parameterizations of the same variables. The resistance parameters defined which of the three variables were included in the resistance model (varied from one to three) and the maximum resistance (R_{max}) for each variable (Supporting Information). With these parameters, we linearly rescaled each landscape variable to range from zero to R_{max}, summed the rescaled variables and then added 1 to produce the final resistance model. In this way, resistance models always had a minimum value of 1 regardless of how many variables the model contained. Producing 27 alternative parameterizations for each of 30 sets of environmental variables yielded a total of 810 unique resistance models. Each of these resistance models was used to determine dispersal and mating probabilities in a population genetic simulation described below.

2.4 | Population genetic simulations

We used CDPOP version 1.2.35 (Landguth & Cushman, 2010) to simulate the population genetic dynamics of gene flow between two populations (the northern unsampled and the southern sampled population) given a resistance model quantifying rates of gene flow. We ran one simulation for each of the 810 resistance models described above. The hierarchy of these simulations thus followed the hierarchy of the resistance models (i.e., 30 sets each containing 27 alternative resistance parameters that relate up to three environmental variables to rates of gene flow). CDPOP is an individual-based simulator of population genetic processes. It simulates mating and dispersal in a finite population of individuals assigned to fixed locations, recording allele usage by all individuals per generation. In each generation, adult individuals mate according to a user-specified mating system and probability function based on proximity in Euclidean or effective distance. Once mated, females give birth to a number of offspring determined by a user-specified probability function which can also control the sex ratio at birth. After birth, adult mortality occurs probabilistically based on user-specified demographic parameters. Finally, vacant locations where adults died are filled by dispersing offspring. Dispersal probabilities follow a user-specified function based on Euclidean or cost-distances to the vacant locations. If all locations are occupied, any remaining offspring not yet assigned to a location are eliminated.

The distribution of individuals within a population inhabiting the northern and southern halves of the simulated landscape varied across the 30 sets but was constant within each set. Because our analysis was based on a model selection exercise within each set, this removed spatial variability in population distribution as a confounding factor in our analysis. For each set, we created two populations of 1,000 diploid individuals each, initially separated by a movement barrier (Figure 1). The location of the individuals in the southern sampled population was a mirror image of the northern unsampled population, just as the landscapes they inhabited were created by joining mirror images of the environmental variables. Individuals were placed probabilistically as a function of the inverse square of resistance (i.e., individuals only occurred in low-resistance cells) according to the resistance model with the highest resistance in each set (this model always included all three environmental variables).

Each simulation began with 100 nonoverlapping generations of mating and dispersal with a complete mating and dispersal barrier separating the two populations. Within each population, dispersal and mating probabilities were a function of the inverse square of the pairwise landscape distances between individual locations (calculated from the resistance model used in the simulation), with a maximum distance set to the average maximum possible pairwise Euclidean distance between individuals in the landscape (1,309 units). If our simulations occurred under the assumptions of isolation by distance (Wright, 1943), the population would be panmictic with this setting. However, the resistance in the simulated landscapes reduced the distances over which mating and dispersal occur, resulting in spatial genetic structure and the populations followed the model of isolation by resistance (McRae, 2006). The CDPOP parameters were the same in all simulations and generally matched Wright–Fisher expectations. Generations were discrete and nonoverlapping (i.e., all adults died simultaneously at the end of each generation). There was no selection, mutation or immigration from outside the population. Individuals were diploid. Mating was sexual and with replacement for either sex. The number of offspring was based on a Poisson distribution with a mean of 4, providing ample offspring to fill all vacant locations. The simulation tracked alleles at 30 codominant marker loci, with 30 alleles randomly assigned per locus to the first generation. The alleles initially assigned to each population were completely different, creating two fully differentiated populations to begin the simulation. Links to the Mendeley repository containing all CDPOP inputs, outputs and parameters are provided in the Supporting Information and in the Data Accessibility section.

During the first 100 generations, population genetic structure was expected to form within each population as a function of the resistance model used to determine the probabilities of mating and dispersal. After the 100-generation "burn-in" period, we ran the simulations for an additional 200 generations in two alternate ways. In the "no barrier" simulations, we allowed mating to occur between the two populations (but not dispersal) only in one direction, from the unsampled northern deme into the sampled southern population. The simulation dispersal parameters and landscapes were designed such that removal of the mating barrier would initially create an admixture zone in the sampled southern population near the boundary with the unsampled population. After an initial formation of admixed offspring in this region, subsequent generations of mating and dispersal were expected to allow alleles to spread further

into the sampled population at a rate and pattern dictated by the simulation's resistance model. In parallel (i.e., starting with the same simulations from the initial 100-generation burn-in period), we also ran "control" simulations where the barrier remained in place for 200 generations.

2.5 | Calculating model selection accuracy

For each of the 810 simulations (representing 30 sets of environmental variables transformed into resistance models according to the 27 unique resistance parameterizations) at each of 21 selected generations during the simulations (generations 0, 5, 10, 15, 20, 25, 50, 75, 100, 105, 110, 115, 120, 125, 150, 175, 200, 225, 250, 275 and 300) we performed a model selection exercise where the resistance model used to determine mating and dispersal probabilities in the simulation was the "true" model and the other 26 resistance model parameterizations in the same set were considered alternative models. Previous studies (Shirk et al., 2018; Zeller et al., 2016) have shown that competing models with ≥ 90% correlation to the true model greatly decreases model selection accuracy in landscape genetic analyses, so we removed alternative models that exceeded this threshold, based on the Pearson correlation between pairwise effective distances among sampled individuals given the competing resistance models.

Rather than base our evaluation on all individuals in the sampled southern population extent, we used a stratified random sampling approach that is commonly used in field sampling for population genetic analyses. We divided the square southern landscape into a large 4 × 4 grid and randomly subsampled 16 individuals from each grid cell for a total of 256 samples. If there were fewer than 16 individuals in a grid cell, we increased the number of individuals sampled in other grid cells until the total of 256 was met. For each candidate resistance model, we calculated effective distances (i.e., the accumulated cost of the least-cost path between all pairs of sampled individuals, given the resistance model) among all 256 sampled individuals with the "gdistance" package version 1.3–6 (Van Etten, 2017) in the R statistical computing environment (R Core team., 2019). We used the program SPAGEDI version 1.5 (Hardy & Vekemans, 2002) to calculate genetic distances between all sampled individuals based on Rousset's *a* (an individual-based genetic distance metric that was designed to reflect the population structure of continuous populations; Rousset, 2000) from the multilocus genotypes corresponding to the CDPOP simulation based on the "true" resistance model.

To determine if the true model could be identified among the competing alternative models in each set, we used an Akaike Information Criterion score (AIC; Akaike, 1998) model selection approach. For each of the 27 candidate models in the set, we used the MLPE.lmm() function from the RESISTANCEGA package version 3.1–1 (Peterman, 2018) in R to fit a linear mixed effects model relating the response (pairwise genetic distances among sampled individuals) to pairwise effective distances between sampled individuals given the candidate resistance model, while accounting for the nonindependence of the pairwise observations with random effects. The top model was selected based on the lowest AIC score. This method of model selection has recently been shown to outperform other correlative landscape genetic model selection methods (Shirk et al., 2018). From these results, we created a binary variable reflecting the true/false outcome of the ability of the model selection to identify the true model among the competing alternatives. The code used to assess model selection accuracy and the results are provided in the Supporting Information.

2.6 | Factors influencing model selection accuracy

Once the simulations were complete, we analysed the model selection results and the CDPOP outputs over the 300 simulated generations in several ways, in both the "no barrier" and control simulations. Our primary objective was to assess model selection accuracy as a function of several factors we considered to be potentially influential in the context of gene flow from an unsampled deme. The response variable reflected the binary observation of whether the model selection process described above identified the true model among competing alternatives in the "no barrier" simulations where gene flow from the unsampled population was allowed to flow into the sampled population. The predictors (fixed effects) included several variables we hypothesized influenced model selection accuracy, scaled and centred to avoid convergence problems in the model fitting. These included *prevalence* (the proportion of the sampled population's alleles originating from the unsampled northern deme), the degree to which the unsampled deme's alleles spread into the sampled population (*alleleDist*), allelic richness (*A*) in the sampled population (we expected higher allelic diversity would provide higher power), the number of candidate models (*nhyp*) evaluated in the model selection process (we expected fewer models in the candidate set would increase the likelihood of selecting the true model by chance), the number of variables (*nvar*) present in the "true" resistance model (we expected more complex models may be more difficult to identify), and the mean cost-distance (*meanCD*) among sampled individuals given the true resistance model (we expected higher resistance models result in greater genetic structure and are easier to identify).

We used the LME4 package version 1.1–23 (Bates et al., 2015) in R to fit a mixed effects logistic regression model using the above response and predictor variables. We included a random intercept for the candidate model identity crossed with the "set" (every set contained each of the 27 resistance parameterizations used to produce the candidate models within the set) to account for the covariance structure of the hierarchical study design. We used the PRESENCEABSENCE package version 1.1.9 (Moisen & Freeman, 2008) in R to assess the performance of this model using a variety of metrics, including AUC (area under the receiver operator curve), per cent correctly classified (PCC), sensitivity, specificity, true skill statistic (TSS) and kappa. For metrics requiring binary classification of

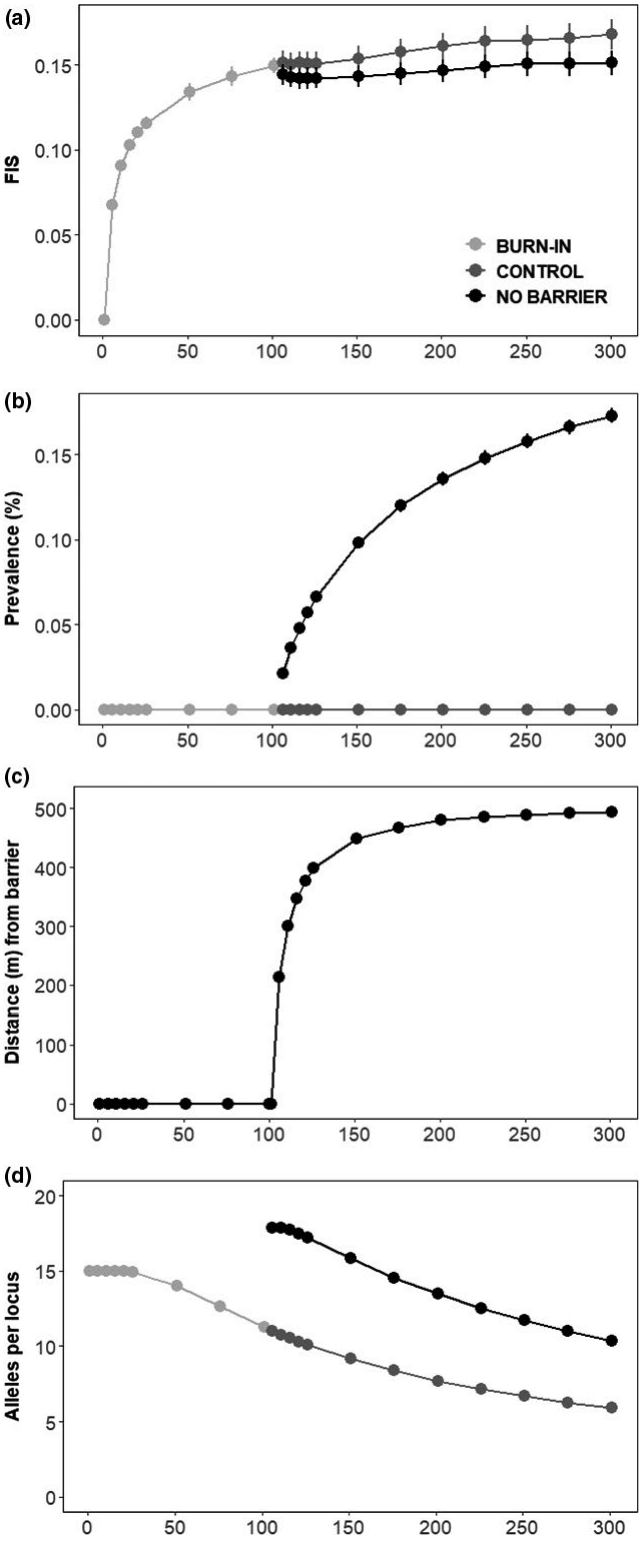

FIGURE 2 Population genetic dynamics of the sampled population over time. At the start of the simulation, the two populations (sampled and unsampled) were separated by a complete barrier to mating and dispersal movements, completely differentiated (i.e., shared no common alleles), and internally panmictic (i.e., random distribution of alleles within each population). During the first 100 generations (the "burn-in" period), mating and dispersal probabilities within each population were controlled by the resistance model specified in the simulation. During the burn-in period, the two populations each approached an equilibrium where the pattern of resistance was reflected in the population genetic structure, as indicated by the inbreeding coefficient (F_{IS}) plateauing over the first 100 generations (a; F_{IS} averaged across all simulations; $n = 810$). After the burn-in period, the mating barrier was removed in the "no barrier" simulations (black lines and circles) but not in the "control" simulations (dark grey lines and circles), and the simulation continued for another 200 generations. During this time, the prevalence of alleles from the unsampled population observed within the sampled population increased rapidly (b; % of total alleles averaged across all simulations, $n = 810$) and they increasingly spread from the border between the two populations (where the barrier was located) into the sampled population over time (c; distance [m] from barrier averaged across all simulations, $n = 810$). The influx of unique alleles from the unsampled population immediately increased allelic diversity in the sampled population relative to the burn-in period prior to admixture between the two populations (d; average alleles per locus, averaged across all simulations, $n = 810$). Over time, however, genetic drift resulted in loss of allelic diversity, but the "no barrier" simulations always maintained a higher level of allelic diversity compared to the control simulations. Error bars represent 95% confidence intervals

the model prediction, we used a threshold value where sensitivity equalled specificity, as we considered Type I and Type II errors equally important.

3 | RESULTS

The proportion of alleles from the unsampled population observed in the sampled population increased continuously from zero over time after the burn-in period, but only in the "no barrier" simulations (Figure 2). The novel alleles introduced from the unsampled population spread rapidly into the sampled population, reaching an average distance that approached the centre of the population (512 m from the barrier) within 150 generations before plateauing (Figure 2). This had the effect of rapidly increasing the allelic diversity of the sampled population within the first 10 generations after the barrier was removed, compared to the control simulations where the barrier remained in place (Figure 2). Because of the Wright–Fisher assumptions of no mutation or immigration in the simulations, genetic drift reduced allelic diversity thereafter, but it remained consistently higher in the "no barrier" simulations relative to the control simulations with the barrier present (Figure 2). In simulations based on landscapes with lower resistance, the distance alleles from the unsampled population spread into the sampled population was greater compared to landscapes with higher resistance (Figure S2).

During the 100-generation burn-in period, population genetic structure, as measured by the inbreeding coefficient (F_{IS}), increased rapidly and began to plateau as local mating within the sampled population created spatial heterogeneity in allele frequencies and reduced homozygosity relative to Hardy–Weinberg expectations (Figure 2). In the control simulations where the barrier remained in place, the increase in population structure continued to level off.

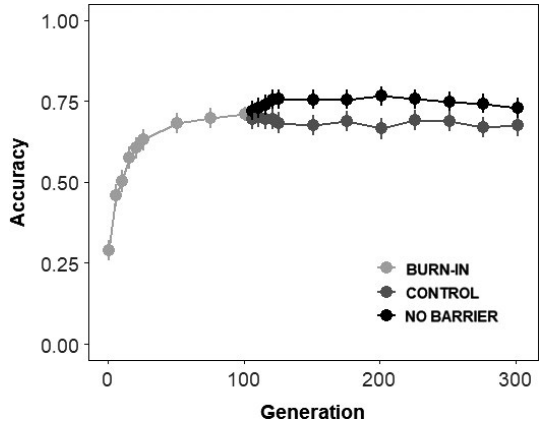

FIGURE 3 Model selection accuracy as a function of simulation time (averaged across all simulations, $n = 810$). Accuracy represents the proportion of simulations where the true (causal) resistance model governing mating and dispersal in the simulation was selected among a pool of 27 alternative models. During the "burn-in" period of the first 100 generations, model selection accuracy increased as the population transitioned from panmictic (i.e., a random starting allele distribution) to structured according to the pattern of resistance in the model used to govern rates of mating and dispersal in the simulation. At generation 100, the mating barrier was removed in the "no barrier" simulations (black line), allowing novel alleles from the unsampled population to spread into the sampled population, increasing the power to detect landscape genetic relationships, and therefore slightly improving model selection accuracy relative to the control simulations (dark grey line) where the barrier remained. Error bars reflect the 95% confidence intervals

TABLE 1 Effect of simulation features on landscape genetic model selection accuracy

Covariate	Coefficient	Coefficient SE	*z*-score	*p*-value
alleleDist	0.2466	0.0401	6.15	<.00000***
nhyp	−0.2391	0.0628	−3.81	.00014***
meanCD	0.3501	0.1097	3.19	.00142**
A	0.1217	0.0567	2.15	.03177*
nvar	−0.3499	0.2328	−1.50	.13282
prevalence	−0.0333	0.0552	−0.60	.54670

Note: We fitted a mixed effects logistic regression model to explore the potential impact of several covariates potentially related to model selection accuracy in our simulations (the response variable), including *alleleDist* (mean Euclidean distance from the barrier at which alleles originating in the unsampled deme were observed in the sampled population), *A* (alleles per locus from both the sampled population and unsampled deme observed in the sampled population), *meanCD* (mean cost-weighted distance between sampled individuals given the true resistance model), *nhyp* (number of hypotheses in the candidate model set), *nvar* (number of variables in the true resistance model) and *prevalence* (proportion of the sampled population's alleles originating from the unsampled deme). The coefficients, standard error, *z*-score, and *p*-value are shown.

Asterisks indicate significant at

***$\alpha = 0.001$,

**$\alpha = 0.01$ and

*$\alpha = 0.05$.

In the "no barrier" simulations where gene flow was allowed from the unsampled population, mating became more random in the admixture zone, reducing inbreeding and increasing disequilibrium (Figure 2). Within 100 generations after the mating barrier was removed, as alleles from the unsampled population spread throughout the sampled population, a new state of pseudo-equilibrium was reached, resulting in stabilizing F_{IS} values (Figure 2).

Model selection accuracy over the same time series increased rapidly during the burn-in period (Figure 3) as population genetic structure in the sampled population went from completely undifferentiated to approaching a state of pseudo-equilibrium. In the "no barrier" simulations where the barrier was removed after the burn-in period, model selection accuracy increased slightly during the first 25 generations of gene flow from the unsampled population, then decreased slightly between generations 200 and 300 (Figure 3). In the control simulations, model selection accuracy declined slightly after the burn-in period. Model selection accuracy was highest at intermediate levels of cost-distance among individuals given the true resistance model (Figure S3) and at intermediate levels of allelic diversity in the sampled population (Figure S4).

Mixed effects logistic regression modelling identified several variables significantly related to model selection accuracy in the simulations (Table 1), including *alleleDist* (the average distance from the barrier that alleles from the unsampled deme occurred in the sampled population), *nhyp* (the number of hypotheses in the model set), *meanCD* (the average cost-distance among sampled individuals given the true resistance model) and *A* (the number of alleles per locus in the sampled population, including alleles from the unsampled deme). The relationship to model selection accuracy was positive for *alleleDist*, *meanCD* and *A* but negative for *nhyp*. The variables *nvar* and *prevalence* were not significantly related to model selection accuracy. Model fit metrics showed a strong fit to the data, with AUC = 0.83, PCC = 0.74, TSS = 0.48, kappa = 0.41 and sensitivity = specificity = 0.74.

4 | DISCUSSION

Our a priori expectation of reduced model selection accuracy due to gene flow from an unsampled deme was based on the idea that a zone of admixture would form in the sampled population near the boundary with the unsampled population after the simulated mating barrier was removed. These admixed offspring would exhibit larger genetic distances with other sampled individuals due to admixture, not the resistance of the sampled population's landscape, as is assumed in correlative landscape genetic analysis. We expected that violation of this assumption would uncouple the relationship between effective and genetic distances, adding noise to the analysis, and reduce model selection accuracy. Contrary to our expectations, however, we found that model selection accuracy was largely unaffected, or even slightly improved, by gene flow from the unsampled

deme, depending on several factors we explored. This result was observed across a wide range of landscapes containing different environmental variables influencing gene flow (continuous, categorical or linear) at varying magnitudes of resistance and varying levels of genetic disequilibrium. This finding suggests other factors (sample size, number of loci, sampling locations, etc.) may be more important to consider when designing sampling schemes and conducting landscape genetic analyses than gene flow from unsampled demes.

Previous studies of sampling strategies and their impact on landscape genetic analysis have focused on within-population sampling issues and were based on simulations of a single continuously distributed population inhabiting a complex (Landguth et al., 2012; Oyler-McCance et al., 2013) or simple landscape (Schwartz & McKelvey, 2009) reflecting isolation by resistance (McRae, 2006) or isolation by distance (Wright, 1943), respectively. They addressed the relative performance of sample schemes (e.g., random, uniform, clustered), sampling density, numbers of sampled individuals, number of sampled loci and allelic richness. We are aware of no study that has addressed between-population dynamics that arise when gene flow occurs from an unsampled deme (either of the same species or of a similar species that hybridizes with the species of interest) into the sampled population.

The most important factor driving model selection accuracy in our simulations was not the prevalence of alleles from the unsampled population in the sampled population (indeed, this variable was found to have no significant relationship to accuracy), but rather, how far those alleles spread into the sampled population and the resulting increase in allelic diversity (including alleles originating from the sampled population as well as the novel alleles introduced by gene flow from the unsampled deme). Even in the most resistant landscapes, however, alleles from the unsampled population eventually spread substantially into the sampled population. Unlike in the admixture zone, the spread of these novel alleles into the sampled population was driven by the intervening landscape, thus meeting the assumptions of correlative landscape genetic analysis. As such, the spread of these new alleles increased the power to detect landscape genetic relationships and therefore slightly improved model selection accuracy.

A potential confounding factor in this dynamic is the relationship between the rate of spread of alleles from the unsampled population and the magnitude of resistance in the landscape. The magnitude of resistance has consistently been shown to be an important factor in model selection accuracy (Cushman et al., 2013), as higher resistance creates strong patterns of genetic differentiation within the population and therefore greater signal to noise ratios. We observed this in our simulations as well, as the magnitude of resistance was strongly related to model selection accuracy. Resistance does not only affect the strength of the landscape genetic pattern, however. It also controls the rate at which alleles spread from the unsampled population into the sampled population. In highly resistant landscapes, the reduced spread of the unsampled population's alleles limits the helpful increase in power that comes with greater allelic diversity, but also provides greater signal. Conversely, in lower resistance landscapes, allelic diversity increases with the greater spread of alleles into the sampled population, but this may be counteracted by the weaker signal of the underlying landscape.

In the event of gene flow from unsampled demes, our results suggest the highest model selection accuracy may come in the intermediate case, where resistance is sufficiently high to provide a strong signal but not so high as to impede the spread of novel alleles throughout the sampled population. Indeed, we observed the highest level of model selection accuracy at intermediate levels of allelic diversity in the sampled population and intermediate levels of resistance. If the pattern and magnitude of resistance in our simulated landscapes was sufficient to limit the spread of the unsampled deme's alleles beyond the initial zone of admixture, we believe our expectation of reduce model selection accuracy would have been met. However, this would require a near complete barrier to migration that would probably be easily detected given the strong signal, even with the added noise due to gene flow from the unsampled deme.

In addition to these spatial dynamics, we also observed a temporal dynamic related to model selection accuracy in our simulations. After the burn-in period, model selection accuracy increased rapidly for approximately 10 generations after the barrier was removed in the "no barrier" simulations, driven by the increase in allelic diversity afforded by the spread of alleles from the unsampled deme. Afterwards, however, model selection accuracy decreased slightly, coinciding with loss of allelic diversity in both the "no barrier" and control simulations due to genetic drift. Drift reduced allelic diversity because it was not countered by mutation and immigration given the Wright–Fisher assumptions of our simulations. Allelic diversity has been shown to be an important factor in model selection accuracy (Landguth et al., 2012) and was also significantly related to model selection accuracy in our study. By the end of the simulation, the model selection accuracy observed in the "no barrier" simulations was almost the same as it was at the time of barrier removal. Thus, the small benefit of gene flow from an unsampled deme may be ephemeral if genetic drift reduces allelic diversity over time.

A key finding from our study is that the magnitude of the effect of gene flow from unsampled demes was minor, regardless of how far the alleles had spread into the sampled population or how many generations had passed since gene flow began. In real populations, the effect may be even smaller, because, unlike in our simulations, the unsampled population is unlikely to be completely differentiated from the sampled population at the time gene flow begins. This is an important finding because, a priori, it is often difficult to identify all demes that may potentially provide gene flow into the sampling extent, or to know the degree to which the sampled population might hybridize with related species. Our results suggest other sampling considerations identified in previous studies are far more important than gene flow from differentiated demes. For example, sampling at an appropriate density, over an appropriate sample extent, along environmental gradients potentially affecting gene flow, at a sufficient number of (ideally highly variable) loci, and accounting for time lags have all been identified

as critical to maximizing model selection accuracy using correlative landscape genetic approaches (Cushman & Landguth, 2010a, 2010b; Cushman et al., 2012; Cushman, et al., 2013; Oyler-McCance et al., 2013).

In many ways, the limited impact of gene flow from an unsampled deme is similar to other simulation studies that have explored establishment of population genetic structure after an initial period of panmixia (Landguth et al., 2010; Shirk et al., 2018; Zeller et al., 2016). In those studies, rather than removal of a mating barrier between populations, the source of the initial disequilibrium was a change from random mating within a population to a complex landscape where mating probabilities were governed by a resistance surface. In all cases, a new equilibrium was quickly reached and the effect on model selection accuracy was modest compared to other factors evaluated (Shirk et al., 2018; Zeller et al., 2016).

Although we examined a broad range of landscape variables and resistance magnitudes over 300 generations of simulation time, there are many aspects of population genetic dynamics we did not vary or explore in the context of our study. Thus, there may be circumstances in which our inferences may not be valid and should therefore be the focus of future investigation. For example, gene flow from unsampled demes could come in repeated waves, rather than as a constant rate of introgression. The environmental gradients influencing gene flow could be directional, rather than omnidirectional and patchy as in our simulations. The density of individuals in the sampled and unsampled demes could be dramatically different than the even distribution we assumed. Similarly, the starting allele frequencies and allelic diversity could be very different from our assumptions of two completely differentiated populations with high allelic diversity. Also, the Wright–Fisher assumptions in our simulations could all be violated, with unknown impacts on the outcome. For these reasons, further exploration of the potential impacts of unsampled demes on model selection accuracy is needed.

Ideally, we would validate the inferences from our simulation study by field studies of real populations (Landguth et al., 2015) where the true resistance of the landscape was known, and varying degrees of gene flow occurred between the sampled and unsampled demes. However, the true resistance of a landscape is rarely, if ever, known with certainty (landscape genetic model selection only identifies the most well-supported model among those tested), and replication under different scenarios of gene flow is difficult to achieve. Simulations have the advantage of allowing explicit testing of hypotheses with complete control over the population genetic dynamics, allowing strong inferences about cause and effect and the magnitude of impacts from stipulated pattern–process relationships.

ACKNOWLEDGEMENTS

This research was supported in part by funds provided by National Science Foundation grants EF-1442597 and DEB-1340852.

AUTHOR CONTRIBUTIONS

A.S., E.L. and S.C. designed the study. A.S. produced the simulated landscapes and resistance models. A.S. and E.L. parameterized and ran the CDPOP simulations. A.S. performed the statistical analysis and wrote the manuscript.

DATA AVAILABILITY STATEMENT

Simulated landscape variables, resistance models, distance matrices, CDPOP simulation inputs/outputs, and R code we used in our analysis are available on Mendeley (https://doi.org/10.17632/hgw27zzch6.1). CDPOP software and user manual are available at http://github.com/ComputationalEcologyLab/CDPOP.

ORCID

Andrew J. Shirk https://orcid.org/0000-0001-7700-5334

REFERENCES

Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. (pp. 199–213). In 2nd International Symposium on Information Theory. https://doi.org/10.1007/978-1-4612-1694-0_15

Balkenhol, N., Cushman, S. A., Waits, L. P., & Storfer, A. (2015). Current Status, Future Opportunities, and Remaining Challenges in Landscape Genetics. *Landscape Genetics: Concepts, Methods, Applications*, *2*, 247–256. https://doi.org/10.1002/9781118525258.ch14

Bates, D., Mächler, M., Bolker, B. M., & Walker, S. C. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), https://doi.org/10.18637/jss.v067.i01

Cushman, S. A., & Landguth, E. L. (2010a). Scale dependent inference in landscape genetics. *Landscape Ecology*, *25*(6), 967–979. https://doi.org/10.1007/s10980-010-9467-0

Cushman, S. A., & Landguth, E. L. (2010b). Spurious correlations and inference in landscape genetics. *Molecular Ecology*, *19*(17), 3592–3602. https://doi.org/10.1111/j.1365-294X.2010.04656.x

Cushman, S. A., Shirk, A., & Landguth, E. L. (2012). Separating the effects of habitat area, fragmentation and matrix resistance on genetic differentiation in complex landscapes. *Landscape Ecology*, *27*(3), 369–380. https://doi.org/10.1007/s10980-011-9693-0

Cushman, S. A., Shirk, A. J., & Landguth, E. L. (2013). Landscape genetics and limiting factors. *Conservation Genetics*, *14*(2), 263–274. https://doi.org/10.1007/s10592-012-0396-0

Cushman, S. A., Wasserman, T. N., Landguth, E. L., & Shirk, A. J. (2013). Re-evaluating causal modeling with mantel tests in landscape genetics. *Diversity*, *5*(1), 51–72. https://doi.org/10.3390/d5010051

Dudaniec, R. Y., Worthington Wilmer, J., Hanson, J. O., Warren, M., Bell, S., & Rhodes, J. R. (2016). Dealing with uncertainty in landscape genetic resistance models: A case of three co-occurring marsupials. *Molecular Ecology*, *25*(2), 470–486. https://doi.org/10.1111/mec.13482

Etherington, T. R., Holland, E. P., & O"Sullivan, D. (2015). NLMpy: A python software package for the creation of neutral landscape models within a general numerical framework. *Methods in Ecology and Evolution*, *6*(2), 164–168. https://doi.org/10.1111/2041-210X.12308

Gardner, R. H. (1999). RULE: Map generation and a spatial analysis program. *Landscape Ecological Analysis*, https://doi.org/10.1007/978-1-4612-0529-6_13

Hardy, O. J., & Vekemans, X. (2002). SPAGeDI: A versatile computer program to analyse spatial genetic structure at the individual or population levels. *Molecular Ecology Notes*, *2*(4), 618–620. https://doi.org/10.1046/j.1471-8286.2002.00305.x

Landguth, E. L., & Cushman, S. A. (2010). Cdpop: A spatially explicit cost distance population genetics program. *Molecular Ecology Resources*, *10*(1), 156–161. https://doi.org/10.1111/j.1755-0998.2009.02719.x

Landguth, E., Cushman, S. A., & Balkenhol, N. (2015). Simulation modeling in landscape genetics. In N. Balkenhol, & L. P. Waits (Eds.), *Landscape genetics: concepts, methods, applications* (pp. 99–113). https://doi.org/10.1002/9781118525258.ch06

Landguth, E. L., Cushman, S. A., Schwartz, M. K., McKelvey, K. S., Murphy, M., & Luikart, G. (2010). Quantifying the lag time to detect barriers in landscape genetics. *Molecular Ecology*, *19*(19), 4179–4191. https://doi.org/10.1111/j.1365-294X.2010.04808.x

Landguth, E. L., Fedy, B. C., OYLER-McCANCE, S. J., Garey, A. L., Emel, S. L., Mumma, M., Wagner, H. H., Fortin, M.-J., & Cushman, S. A. (2012). Effects of sample size, number of markers, and allelic richness on the detection of spatial genetic pattern. *Molecular Ecology Resources*, *12*(2), 276–284. https://doi.org/10.1111/j.1755-0998.2011.03077.x

McRae, B. H. (2006). Isolation by resistance. *Evolution*, *60*(8), 1551–1561. https://doi.org/10.1554/05-321.1

Moisen, G., & Freeman, E. (2008). PresenceAbsence: An r package for presence absence analysis. *Journal of Statistical Software*, *23*(11), 1–31.

Oyler-McCance, S. J., Fedy, B. C., & Landguth, E. L. (2013). Sample design effects in landscape genetics. *Conservation Genetics*, *14*(2), 275–285. https://doi.org/10.1007/s10592-012-0415-1

Peterman, W. E. (2018). ResistanceGA: An R package for the optimization of resistance surfaces using genetic algorithms. *Methods in Ecology and Evolution*, *9*(6), 1638. https://doi.org/10.1111/2041-210X.12984

R Core team. (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing (Vol. *55*, pp. 275–286).

Rousset, F. (2000). Genetic differentiation between individuals. *Journal of Evolutionary Biology*, *13*(1), 58–62. https://doi.org/10.1046/j.1420-9101.2000.00137.x

Schwartz, M. K., & McKelvey, K. S. (2009). Why sampling scheme matters: The effect of sampling scheme on landscape genetic results. *Conservation Genetics*, *10*(2), 441–452. https://doi.org/10.1007/s10592-008-9622-1

Shirk, A. J., & Cushman, S. A. (2014). Spatially-explicit estimation of Wright"s neighborhood size in continuous populations. *Frontiers in Ecology and Evolution*, *2*(OCT), 1–12. https://doi.org/10.3389/fevo.2014.00062

Shirk, A. J., Landguth, E. L., & Cushman, S. A. (2017). A comparison of individual-based genetic distance metrics for landscape genetics. *Molecular Ecology Resources*, 1308–1317. https://doi.org/10.1111/1755-0998.12684

Shirk, A. J., Landguth, E. L., & Cushman, S. A. (2018). A comparison of regression methods for model selection in individual-based landscape genetic analysis. *Molecular Ecology Resources*, *18*(1), 55–67. https://doi.org/10.1111/1755-0998.12709

Shirk, A. J. J., Wallin, D. O. O., Cushman, S. A. A., Rice, C. G. G., & Warheit, K. I. I. (2010). Inferring landscape effects on gene flow: A new model selection framework. *Molecular Ecology*, *19*(17), 3603–3619. https://doi.org/10.1111/j.1365-294X.2010.04745.x

Van Etten, J. (2017). gdistance: Distances and Routes on Geographical Grids. R package version 1.1-9. *Journal of Statistical Software*, *76*(13).

Wang, I. J., & Bradburd, G. S. (2014). Isolation by environment. *Molecular Ecology*, *23*(23), 5649–5662. https://doi.org/10.1111/mec.12938

Wright, S. (1943). Isolation by Distance. *Genetics*, *28*(2), 114–138.

Zeller, K. A., Creech, T. G., Millette, K. L., Crowhurst, R. S., Long, R. A., Wagner, H. H., Balkenhol, N., & Landguth, E. L. (2016). Using simulations to evaluate Mantel-based methods for assessing landscape resistance to gene flow. *Ecology and Evolution*, *6*(12), 4115–4128. https://doi.org/10.1002/ece3.2154

Zeller, K. A., Vickers, T. W., Ernest, H. B., & Boyce, W. M. (2017). Multi-level, multi-scale resource selection functions and resistance surfaces for conservation planning: Pumas as a case study. *PLoS One*, *12*(6), e0179570. https://doi.org/10.1371/journal.pone.0179570

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Shirk AJ, Landguth EL, Cushman SA. The effect of gene flow from unsampled demes in landscape genetic analysis. *Mol Ecol Resour*. 2021;21:394–403. https://doi.org/10.1111/1755-0998.13267