Conserv Genet (2014) 15:981–992
DOI 10.1007/s10592-014-0593-0

RESEARCH ARTICLE

Evaluating sample allocation and effort in detecting population differentiation for discrete and continuously distributed individuals

Erin L. Landguth · Michael K. Schwartz

Received: 24 December 2013 / Accepted: 7 March 2014 / Published online: 26 March 2014

Abstract One of the most pressing issues in spatial genetics concerns sampling. Traditionally, substructure and gene flow are estimated for individuals sampled within discrete populations. Because many species may be continuously distributed across a landscape without discrete boundaries, understanding sampling issues becomes paramount. Given large-scale, geographically broad conservation efforts, researchers are looking for guidance as to the trade-offs between sampling more individuals within a population versus few individuals scattered across more populations. Here, we conducted simulations that address these issues. We first established two archetypical patterns of dispersion: (1) individuals within discrete populations, and (2) continuously distributed individuals with limited dispersal. We used genotypes generated from a spatially-explicit, individual-based program and simulated genetic structure in individuals from nine different population sizes across a landscape that either had barriers to movement (defining discrete populations) or isolation-by-distance patterns (defining continuously distributed individuals). Then, given each pattern of dispersion, we allocated samples across four different sampling strategies for each of the nine population sizes in various configurations for sampling more individuals within a population versus fewer individuals scattered across more populations. We assessed the population genetic substructure with both the population-based metric, F_{ST}, and an individual-based metric, D_{PS} regardless of the true pattern of dispersion to allow us to better understand the effect of incorrectly matching the metric and the distribution (e.g., F_{ST} with continuously distributed individuals, and vice versa). We show that sampling many subpopulations (or sampling areas), thus sampling fewer individuals per subpopulation, overestimates measures of population subdivision with the population-based metric for both patterns of dispersion. In contrast, using the individual-based metric gives the opposite results: sampling too few subpopulations, and many individuals per subpopulation, produces an underestimate of the strength of isolation-by-distance. By comparing all results, we were able to suggest a strong predictive model of a chosen genetic structure metric for elucidating the sampling design trade-offs given each pattern of dispersion and configuration on the landscape.

Keywords CDPOP · F_{ST} · Isolation-by-distance · Isolation-by-barrier · Sampling optimization · Simulation modeling

E. L. Landguth (✉)
Division of Biological Sciences, University of Montana, 32 Campus Drive, Missoula, MT 59812, USA
e-mail: erin.landguth@mso.umt.edu

M. K. Schwartz
U.S.D.A. Forest Service, Rocky Mountain Research Station, 800 E. Beckwith Ave, Missoula, MT 59801, USA

Introduction

One of the most pressing questions facing researchers conducting either a population or spatial genetics study in a natural setting is how to allocate sampling effort. While costs of analyzing samples has declined in recent years (Seeb et al. 2011), demands on researchers to design studies with better inference across broader geographic ranges has increased (Schwartz and Vucetich 2009). Furthermore, for the study of many remote, endangered, or difficult to sample species, field and associated sampling costs are substantial.

Traditional sampling advice for population genetic studies is to collect enough samples in each subpopulation to accurately characterize the allele frequencies in the subpopulation. When number of allelic states per population is high, sampling effort needs to be high, as well (Ott 1992). Ott (1992) recommends sampling 80–100 individuals to have a high probability of detecting most alleles for human gene mapping. Other simulations reported that for an isolated population of 8,000 individuals, samples of 20 individuals could produce an accurate allele frequency distribution characterized at six microsatellite markers (Siniscalco et al. 1999). Sampling demands increased as highly variable microsatellites became the dominant tool in molecular ecology studies (Selkoe and Toonen 2006) and up to 30–40 alleles were identified in some natural populations (Hoffman et al. 2005; Purcell et al. 2006). Rao's (2001) simulations of an outbred population produced a useful "rule of thumb", where a sample size of four times the number of alleles was adequate to find all nearly equally frequent alleles with a high probability when the number of alleles is less than 35; and five times the number of alleles was adequate to find all alleles when there were between 35 and 100 alleles (per locus) in the population. Software has been developed to calculate the minimum sample size of genotypes required to detect all alleles with a given frequency at a locus with a given confidence (MINSAGE; http://www.imbs-luebeck.de/imbs/de/node/39), although a review of the literature suggests that this is rarely used in most molecular ecology studies. For population substructure and gene flow estimation some of these sampling requirements may be relaxed as most metrics are not sensitive to alleles at very low frequencies.

As molecular ecology and conservation genetics increased its repertoire of approaches from mostly studying discrete populations in a population genetics framework to methods that allowed the study of continuously distributed individuals and populations in a landscape genetics framework, advice on sampling has been lacking. Several studies have shown that sampling can have strong impacts on interpreting landscape genetic fresults (Murphy et al. 2008; Novembre et al. 2008; Schwartz and McKelvey 2009). For example, Landguth et al. (2012a) investigated the effect of study design on landscape genetics inference using a spatially-explicit, individual-based program to simulate genetic differentiation in a spatially continuous population inhabiting a landscape with gradual changes in resistance to movement. They found that while all three variables of interest (number of loci, alleles per locus, and individuals sampled from the population) influenced power on successfully identifying the generating process, generally increasing the number of loci used had the largest effect. However, their study used a spatially random sample drawn from a continuously distributed underlying population to test their ability to correctly identify the generating process. In reality, a truly random sampling design may be very difficult to achieve in the field. Furthermore, their results were derived from an idiosyncratic landscape that models a species specific movement for an organism with mid-range dispersal capabilities (e.g., American black bear, *Ursus americanus*).

In this study, we extend the simulations of Landguth et al. (2012b) in a more generalized framework to provide guidance as to optimal sampling allocation between individuals and "populations" for a particular study-wide sample size. Given archetypical distributions and varying degrees of gene flow, we then assess the population genetic substructure with population- and individual-based metrics, regardless of the true distribution of simulated individuals to allow us to better understand the effect of incorrectly matching the metric and the distribution.

Methods

Simulation program

We used CDPOP v1.1 (Landguth and Cushman 2010; Landguth et al. 2012a), an individual-based, spatially-explicit, landscape genetics program that models genetic divergence through time on a landscape surface and spatially located individuals as a function of individual-based movement through mating and dispersal, incorporating vital dynamics and all the factors that affect the frequency of an allele in a population (mutation, gene flow, genetic drift, and selection). It can simulate different movement behavior of individuals which allows the emergence of spatial genetic structure.

In CDPOP, individual movement (i.e., mating and dispersal) is modeled as probabilistic functions of the cumulative cost between individual locations across the landscape resistance surfaces (e.g., shortest-path of cumulative summed resistance values between locations). These movement cost functions are scaled to a user-specified truncated distance that is a proxy for species specific movement strategies (i.e., short-range versus long-range dispersers). This truncated value constrains all mate choices and dispersal distances to be less than or equal to a threshold value with probability of mating or dispersal distance within that limit specified by a user-defined probability function (e.g., negative binomial).

Simulation scenarios

We used nine different population sizes (N) that were a factorial combination of three subpopulation sizes ($S = [16, 36, 64]$) with three different numbers of individuals per

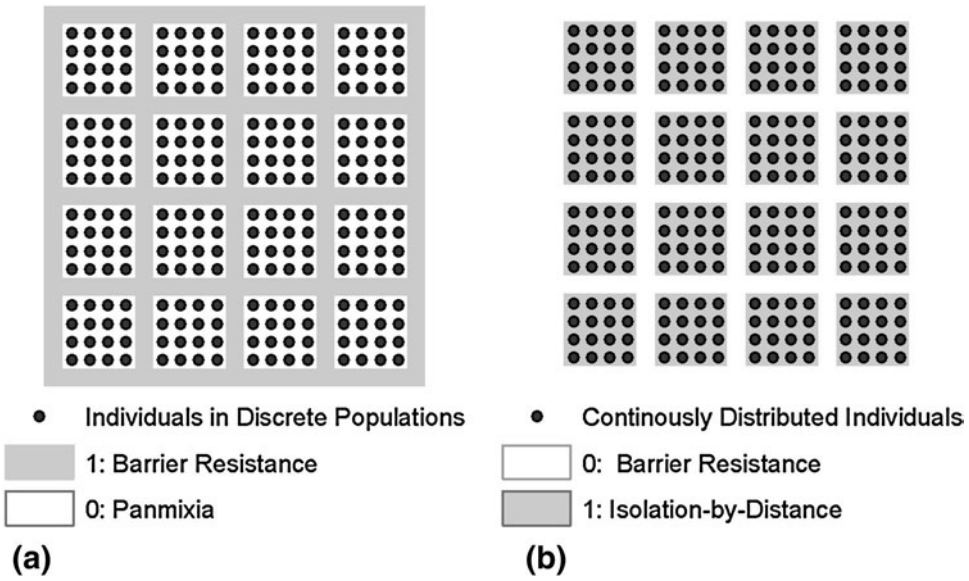

Fig. 1 An example of one of the nine population designs; $S = 16$ subpopulations, $I/S = 16$ individuals per subpopulation, and $n = 256$ total individuals for both the **a** discrete and **b** continuously distributed population scenarios. In the discrete population scenario, each subpopulation was separated by complete barrier of strength one, while mating was a random process within each discrete subpopulation. For the continuously distributed scenario, isolation-by-distance (IBD) controlled movement of individuals and the barrier was removed by setting the resistance value of the barrier to zero

subpopulation ($I/S = [16, 36, 64]$). For each of the nine population sizes, we ran CDPOP v1.1 to simulate individual genetic exchange across 50 non-overlapping (i.e., discrete) generations as functions of individual-based movement (mating and dispersal) for two different patterns of dispersion: discrete and continuously distributed populations. Sex was assigned at random initially. It is important to note that in the cases of the continuously distributed populations we kept the nomenclature for "subpopulations", although in a continuously distributed environment this is simply a "sampling area" or "sampling region" and not a true subpopulation. See Fig. 1 for an example of a simulated population where the total sample size (N) equals 256 individuals, achieved by collected samples from 16 subpopulations (S), where 16 individuals per subpopulation (I/S) were sampled in both the discrete (e.g., fish in a small pond) and continuously distributed (e.g., lynx across the boreal forest) population scenarios. In the discrete population scenario, each subpopulation was separated by complete barriers and by setting the dispersal (movement) to be less than the barrier strength, individuals were not allowed to disperse to another subpopulation (Landguth et al. 2010). Within each subpopulation, mating was random.

For the continuously distributed scenario, we conducted an IBD simulation modeling experiment by removing the barrier resistance between each subpopulation (i.e., setting the resistance value of the barrier to zero). Both mating and dispersal were controlled by a Moore neighborhood, where individuals could only mate with the surrounding eight nearest neighbors and offspring were only allowed to disperse to the surrounding eight nearest neighbor locations.

In both natural history scenarios, mating parameters were set in CDPOP to represent a population that was dioecious with both females and males mating with replacement. Offspring parameters were set, such that each female had a mean of three offspring following a Poisson process, with random sex assignment. This guaranteed an excess of offspring that ensured that all spatial locations in the population were filled through dispersal movement at each generational time step and avoided empty locations that require immigrants from an outside population. The remaining offspring were discarded once all the spatial locations were occupied by a dispersing individual and maintained a constant population size at every generation. This is equivalent to forcing emigrants out of the study area once all available home ranges are occupied (Balloux 2001; Landguth and Cushman 2010).

All simulated populations contained 20 neutral loci and 10 initial starting alleles per locus with no mutation rate (the latter of which is reasonable considering the short simulation time period), free recombination, and no initial linkage disequilibrium. As the program simulates stochastic processes, we ran ten Monte Carlo replicates for each scenario to quantify the mean and variability of the genetic structure.

Table 1 Sample allocation design across nine population configurations (simulated scenarios) as follows: (a) number of subpopulations (*S*), (b) the total individuals in each subpopulation (*I/S*), (c) the global population size (*N*), (d) the four sample allocation sizes applied to each population (*n*), and (e) the corresponding subpopulation allocation size for each sample allocation size in (d)

(a) Number of subpopulations (*S*) Simulated scenarios	(b) Individuals per subpopulation (*I/S*)	(c) Global population size (*N*) $S \times I/S$	(d) Sample allocation size (*n*) Total samples drawn from simulated scenarios	(e) Subpopulation allocation size Allocation of drawn samples into subpopulations of size ×
$4 \times 4 = 16$	16	256	16, 32, 64, 128	16, 8, 4, 2
$4 \times 4 = 16$	36	576	16, 32, 64, 128	16, 8, 4, 2
$4 \times 4 = 16$	64	1,024	16, 32, 64, 128	16, 8, 4, 2
$6 \times 6 = 36$	16	576	36, 72, 144, 288	36, 18, 12, 9, 6, 4, 3, 2
$6 \times 6 = 36$	36	1,296	36, 72, 144, 288	36, 18, 12, 9, 6, 4, 3, 2
$6 \times 6 = 36$	64	2,304	36, 72, 144, 288	36, 18, 12, 9, 6, 4, 3, 2
$8 \times 8 = 64$	16	1,024	64, 128, 256, 512	64, 32, 16, 8, 4, 2
8 x 8 = 64	**36**	**2,304**	**64, 128, 256, 512**	**64, 32, 16, 8, 4, 2**
$8 \times 8 = 64$	64	4,096	64, 128, 256, 512	64, 32, 16, 8, 4, 2

The bold sample allocation design is further illustrated for samples per subpopulation in Table 2

Table 2 Example sample allocation design for (a) $S = 64$ subpopulations and (b) $I/S = 36$ individuals per subpopulation with (c) sample allocation size of $n = 64$, 128, 256, and 512, across the (d) subpopulation allocation sizes

(a) Number of subpopulations (*S*)	(b) Individuals per subpopulation (*I/S*)	(c) Sample allocation size (*n*)	(d) Subpopulation allocation size	(e) Samples per subpopulation (e.g., 1 sample in 64 subpopulations)
$4 \times 4 = 64$	36	64	64, 32, 16, 8, 4, 2	1/64, 2/32, 4/16, 8/8, 16/4, 32/2
$4 \times 4 = 64$	36	128	64, 32, 16, 8, 4	2/64, 4/32, 8/16, 16/8, 32/4
$4 \times 4 = 64$	36	256	64, 32, 16, 8	4/64, 8/32, 16/16, 32/8
$4 \times 4 = 64$	36	512	64, 32, 16	8/64, 16/32, 32/16

Then, the corresponding sample per subpopulation gets drawn in (e)

Sample allocation scenarios

For each of nine global population size scenarios ($N = 256$ to $N = 4{,}096$) we drew four different sample sizes and divided these samples across varying subpopulation sizes resulting in 36 total sampling designs for each different pattern of dispersion (Tables 1, 2). This produced a total of 169 sampling schemes as follows: For the simulated scenarios with $S = 16$ subpopulations, we drew sample sizes (*n*) of 16, 32, 64, and 128 collected across 16, 8, 4, and 2 subpopulations (44 sampling scenarios). For the simulated scenarios with $S = 36$ subpopulations, we drew sample sizes of 36, 72, 144, and 288 collected across 36, 18, 12, 9, 6, 4, 3, and 2 subpopulations (72 sampling scenarios). For the simulated scenarios with $S = 64$ subpopulations, we drew sample sizes of 64, 128, 256, and 512 collected across 64, 32, 16, 8, 4, and 2 subpopulations (53 sampling scenarios). For example, Table 2 shows how the samples were drawn for the $S = 64$ subpopulation and $I/S = 36$ individuals per subpopulation scenario, resulting in 18 different sampling scenarios for that population (e.g., for a sample of $n = 128$, we drew samples per subpopulation of 2/64, 4/32, 8/16, 16/8, and 32/4). For each Monte Carlo simulation replicate, we randomly selected subpopulations and then individuals within the subpopulation to sample.

Assessing genetic structure

Our goal was not to assess metric sensitivity to genetic differentiation, rather to understand the optimal sampling allocation effort with limited resources for detecting differences in population genetic structure among scenarios. Therefore, for each of the nine population sizes and patterns of dispersion, we calculated the most widely used population-based metric, F_{ST} (Nei 1973; Nei and Chesser 1983), as well as the most commonly used individual-based metric, D_{PS} (proportion of shared alleles; Bowcock et al. 1994). We used F_{ST} to measure genetic structure. In addition, we performed a Mantel test (Mantel 1967) to correlate genetic distance (using D_{PS}) to the log transformed Euclidean distance (Rousset 1997; Graves et al. 2013) among individuals using the library 'ecodist' version 1.2.2 (Goslee and Urban 2007) in the statistical software package R (R Development Core Team 2012). Each calculation considered the total population size, which we defined as the 'true' population genetic structure. From the

total population, we then conducted the sampling scenarios. For the sample scenarios, we considered the (x,y)-locations to either group individuals into their designated subpopulations and estimate F_{ST} among groups or for comparing the genetic distance matrix (D_{PS}) and the log transformed Euclidean distance with the Mantel statistic). We note that for the population-based metric F_{ST}, sampling 1–2 individuals in a subpopulation is not statistically valid. Therefore, we did not consider these sample allocation designs and the 169 total sampling allocation designs were reduced to 142 in our analysis of F_{ST}. We calculated the 'true' and sampled values at each generation across the 50 generations and for each Monte Carlo replicate, while the sampled metrics were calculated across the 50 generations, with a random draw for each Monte Carlo replicate.

Determining optimal sampling strategy

For each generation and scenario we plotted the 'true' and sampled values. Since the 'true' value is a known constant at each generation (i.e., all subpopulations are sampled from), we hypothesized that each sample design would have some monotonic function of subpopulation size around the known metric value. Then using a spline interpolation (Python's SciPy interpolate function) at each generation, we extracted the subpopulation for each sample design that produced the closest value to the 'true' value (denoted as $\hat{S}$) and produced 1800 $\hat{S}$ values (50 generations * 36 sample allocation scenarios). We then asked what values of S, I/S, N, and n predict the optimal subpopulation sample size, $\hat{S}$. We modeled the response variable, $\hat{S}$, as a linear combination of the four covariates. Multi-model inference (information theoretic methods; Burnham and Anderson 2002) was conducted to produce candidate models as a linear combination of all possible combinations of these variables using the library 'MuMIn' version 1.7.7 (Barton 2012) in the statistical software package R (R Development Core Team 2012). We minimized Akaike's Information Criterion (AIC) and used AIC model weights to select candidate top models, and reported adjusted R^2 criteria for comparison purposes for each natural history strategy and metric value.

Results

Sampling based on population-based genetic differentiation

Figure 2 shows an example of $S = 64$ subpopulations and $I/S = 36$ individuals per subpopulation with higher and lower bounds of our sample allocation sizes of $n = 64$ and $n = 512$ shown for both discrete subpopulations and continuously-distributed individuals. Similar patterns are seen across the remaining eight simulated scenarios. Comparing patterns of dispersion (continuous versus discrete), the population-based metric F_{ST}, intended for discrete groups, is much lower in the continuously distributed scenario (Fig. 2; column 1 – $\sim$ 0.6 in discrete verses column 2 – $\sim$ 0.1 in continuous).

Allocating samples to too few individuals per subpopulation, for the benefit of sampling more subpopulations, produces an overestimate compared to 'true estimate' in all scenarios with F_{ST} (e.g., Fig. 2a; dashed-dotted line sample four individuals at 16 subpopulations). In contrast, allocating too many samples per subpopulation at the cost of sampling fewer subpopulations tends to underestimate the true population-based metric (e.g., Fig. 2a; diamond line sample 32 individuals at two subpopulations). This over- and under-estimating pattern is observed for both patterns of dispersion simulated. However, increasing n closer to the true population size reduces the observed bias caused by sample allocation as expected and observed in Fig. 2c, d.

Bias caused by sample allocation is also a function of time (generation) or the genetic variability of the population. To illustrate this, Fig. 3 shows results from a simulation with $S = 64$ subpopulations and $I/S = 36$ individuals per subpopulation with sample allocation size of $n = [64, 128, 256, 512]$. The figure is a snapshot of generation 10 (or $F_{ST} = 0.18$ and $F_{ST} = 0.06$ for discrete and continuous scenarios, respectively) and generation 50 (or $F_{ST} = 0.64$ and $F_{ST} = 0.10$ for discrete and continuous scenarios, respectively). When n is allocated to only a few subpopulations F_{ST} is underestimated. Conversely, overestimates of the 'true' F_{ST} occur at the cost of sampling less individuals per subpopulation (i.e., n is distributed across more subpopulations). As the genetic differentiation of the population increases over time in our simulations due to drift, sample allocation strategy also changes. For example, observing Fig. 3a compared to Fig. 3c with discrete subpopulations and $n = 512$ the optimal subpopulation to distribute 512 samples across would be $\hat{S} \cong 23$ for 10 generations ($F_{ST} = 0.18$) and and $\hat{S} \cong 28$ for 50 generations ($F_{ST} = 0.64$). In addition, the observed bias for discrete subpopulations is reduced as the population becomes more differentiated (Fig. 3c compared to Fig. 3a, b, d). Yet, the observed bias remains consistent for continuously distributed individuals (Fig. 3b, d).

We produced 1800 optimal subpopulation strategies for each of the respective patterns of dispersion that allowed us to predict $\hat{S}$ (optimal subpopulation sample size) as a function of the four covariates; S, I/S, N, and n. The most supported linear models for both discrete and continuously distributed populations are shown in Table 3. Based on the

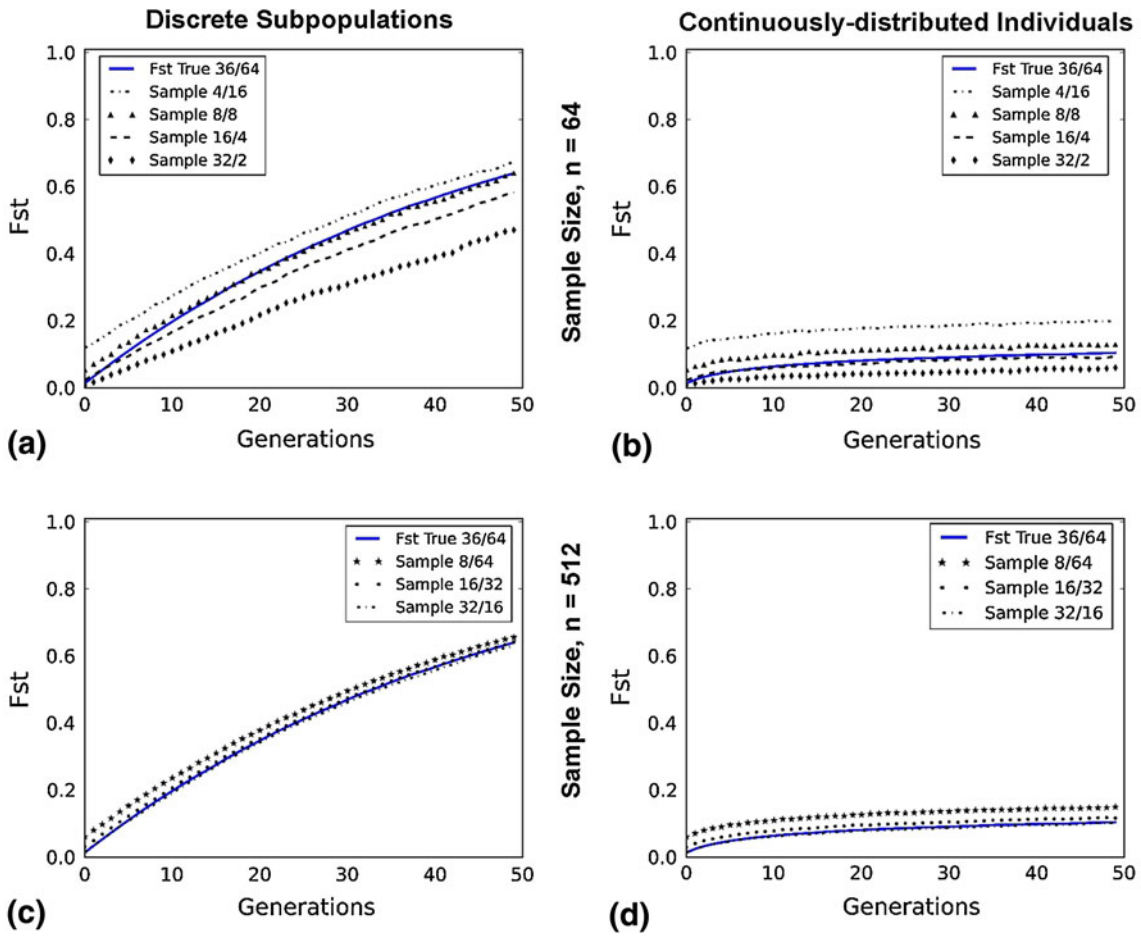

Fig. 2 F_{ST} for $S = 64$ subpopulations and $I/S = 36$ individuals per subpopulation with sample allocation size of $n = 64$ for **a** discrete subpopulations and **b** continuously-distributed individuals, and sample allocation size of $n = 512$ for **c** discrete subpopulations and **d** continuously-distributed individuals. *Solid blue line* is the 'true' F_{ST} metric for the $n = 2{,}304$ total population size. *Legend lines* refer to sample size (n)/subpopulations (S). Note that confidence intervals are too small to be viewed at this scale and unrealistic sample designs of 1–2 individuals per subpopulation not included. (Color figure online)

beta values, I/S and N were the weakest predictors of $\hat{S}$, while S and n were the strongest predictors for both discrete and continuously distributed populations.

Sampling based on individual-based genetic differentiation

For each of the nine populations, the 36 sample allocation designs, and the two patterns of dispersion, we calculated D_{PS} (Bowcock et al. 1994) as a measure of genetic dissimilarity and subsequently, tested for isolation-by-distance strength using the Mantel r statistic as the individual-based genetic differentiation metric across the 50 generations (i.e., genetic distance correlated to log transformed Euclidean distance). Figure 4 shows an example of $S = 64$ subpopulations and $I/S = 36$ individuals per subpopulation with the higher and lower bounds of our sample allocation sizes of $n = 64$ and $n = 512$ shown for both discrete subpopulations and continuously-distributed individuals. Similar patterns are seen across the remaining eight populations. The 'true' individual-based metric is roughly the same for both patterns of dispersion (Fig. 4; solid blue lines in column 1 verses column 2). Interestingly, opposite sampling results are observed with the individual-based metrics than with the population-based metric. Sampling too few subpopulations with more individuals produces an overestimate compared to the 'true estimate' in the discrete population scenarios (e.g., comparing diamond lines with a sample of 32 individuals at two subpopulations in Fig. 4a seen above the 'true' value compared to Fig. 2a seen below the 'true' value). Less obvious but in contrast, sampling too few individuals per subpopulation in order to sample more subpopulations tends to underestimate the true individual-based metric (e.g., Fig. 4d; all lines seen below the 'true' value). Note that in Fig. 4a, a sample of one individual across 64 subpopulations produces $r \sim 0$ and samples of 1 individual at every subpopulation for all sample sizes are

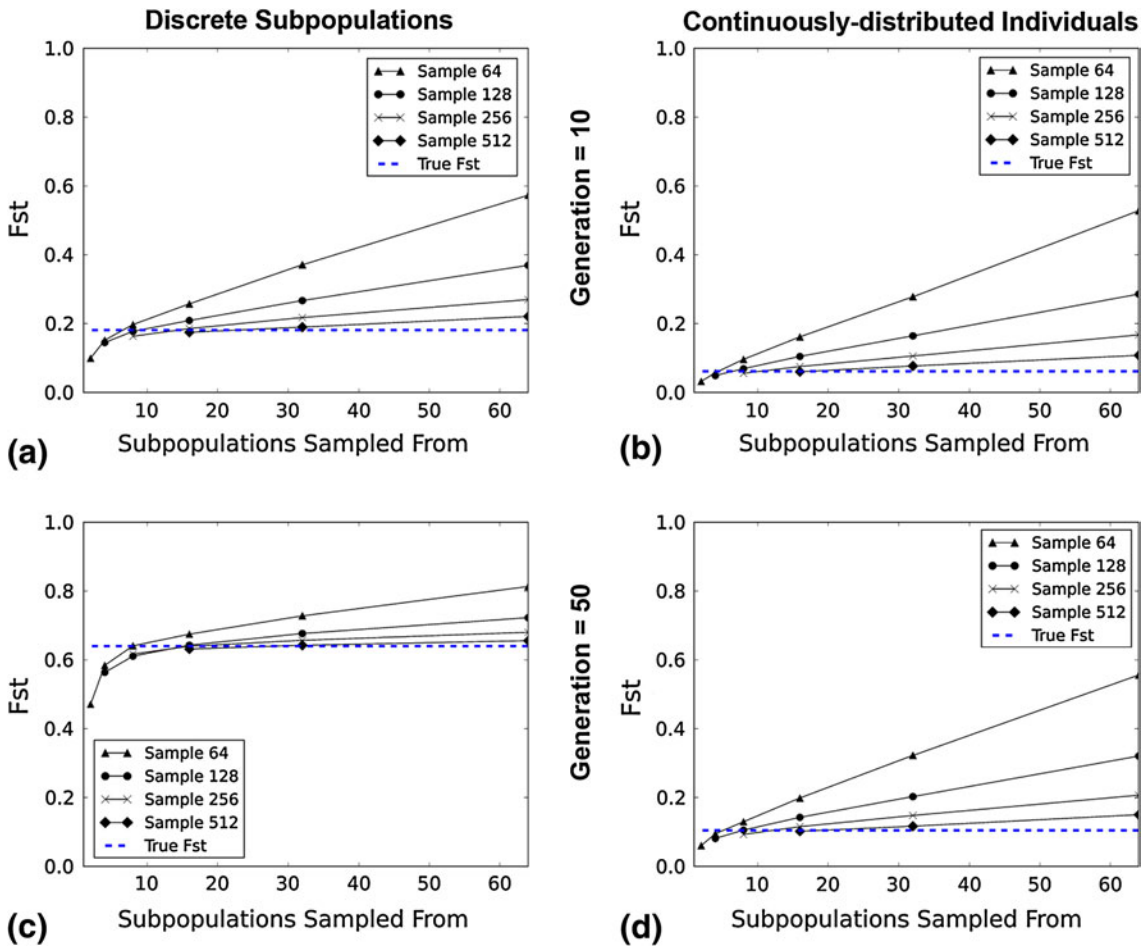

Fig. 3 F_{ST} versus the number of subpopulations sampled (S) for the example $S = 64$ subpopulations and $I/S = 36$ individuals per subpopulation with sample allocation size of $n = [64, 128, 256, 512]$ for the patterns of dispersion of **a** individuals within discrete populations at generation 10, **b** continuously distributed individuals at generation 10, **c** individuals within discrete populations at generation 50, and **d** continuously distributed individuals at generation 50. The *dashed blue horizontal line* is the 'true' F_{ST} value for that generation. Node locations along each line represent a simulated experiment and indicate the number of subpopulations that n was allocated across. (Color figure online)

Table 3 Top linear models (chosen using AIC and model weight) for optimal subpopulations to sample from, $\hat{S}$, for the population-based metric F_{ST}

Pattern of dispersal	S	I/S	N	n	Int	ΔAIC	R^2	w
Discrete	0.158	–	−0.003	0.047	3.338	0.00	0.81	0.73
	0.158	0.001	−0.003	0.047	3.358	2.00[a]	0.81	0.27
Continuous	0.081	−0.008	−0.001	0.024	2.133	0.00	0.92	0.59
	0.089	–	−0.001	0.024	1.718	0.73[b]	0.92	0.41

S—number of total subpopulations within the population, I/S—number of individuals within each subpopulation, N—total number of individuals in population, n—sample allocation size, Int—intercept, R^2—adjusted R^2 value, w—model selection weight

[a] Next model ΔAIC = 199.61

[b] Next model ΔAIC = 132.50

the only scenarios that produce non-significant p values. This is in part due to fact that if a sample of one individual from every discrete subpopulation is taken, then each individual is likely to be as different as possible ($D_{PS} \sim 0.96$) resulting in a non-significant correlation with spatial data. This over- and under-estimating pattern is observed primarily for the discrete population scenarios. For the continuously distributed scenarios, most all sampling designs tend to underestimate the true Mantel r value consistently after 10 generations.

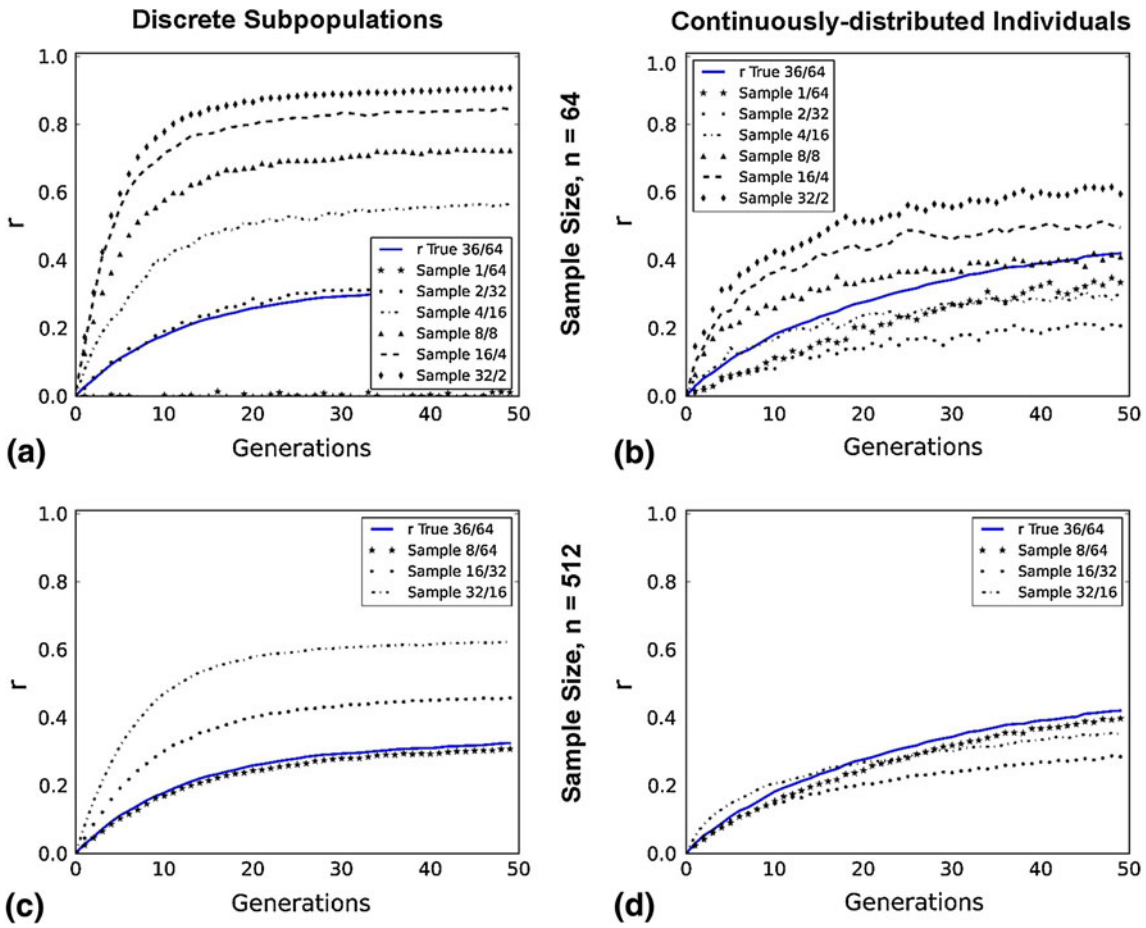

Fig. 4 Mantel r for $S = 64$ subpopulations and $I/S = 36$ individuals per subpopulation with sample allocation size of $n = 64$ for **a** discrete subpopulations and **b** continuously-distributed individuals, and sample allocation size of $S = 512$ for **c** discrete subpopulations and **d** continuously-distributed individuals. *Solid blue line* is the 'true' r metric for the $n = 2{,}304$ total population size. *Legend lines* refer to sample size (n)/subpopulations (S). Note that confidence intervals are too small to be viewed at this scale. (Color figure online)

Bias caused by sample allocation is also a function of time (generation) or the genetic variability of the population. Similar to the population-based metric approach we illustrate this bias with the example simulation of $S = 64$ subpopulations and $I/S = 36$ individuals per subpopulation with sample allocation size of $n = [64, 128, 256, 512]$. Figure 5 shows a snapshot of generation 10 (or $r = 0.16$ and $r = 0.17$ for discrete and continuous scenarios, respectively). We also show a snapshot of generation 50 (or $r = 0.33$ and $r = 0.42$ for discrete and continuous scenarios, respectively). In the discrete population scenarios, Mantel r using D_{PS} estimations show again a clear pattern that is contrary to the population-based metric estimations shown in Fig. 3. When S is allocated to only a few subpopulations the correlation between genetic distance (D_{PS}) and geographic distance is overestimated. Conversely, underestimates of the 'true' r occur more often when n is distributed across more subpopulations at the cost of sampling more individuals per subpopulation. Interestingly, most sampling designs trying to estimate the 'true' r-value produce underestimates, particularly in later generations. As the simulated genetic structure of the population increases, sample allocation strategy suggests allocating samples across more subpopulations. True r's increase across generations does not show the sensitivity with pattern of dispersion, as does the population-based metric.

Given that the optimal subpopulation to sample from, $\hat{S}$, is a function of genetic structure or generational time, we extracted $\hat{S}$ at every generation via linear spline interpolation that produced the closest 'true' r value for each sample allocation size and for each scenario (unless a scenario interpolation value did not cross the 'true' value as with the continuously distributed scenarios estimation of Mantel r). This resulted in 1800 optimal subpopulation strategies for each of the respective patterns of dispersion that allowed us to predict $\hat{S}$ as a function of the four covariates; S, I/S, N, and n. The most supported linear models based on AIC included all four predictors, with some of the top models lacking I/S and N (Table 4). Based

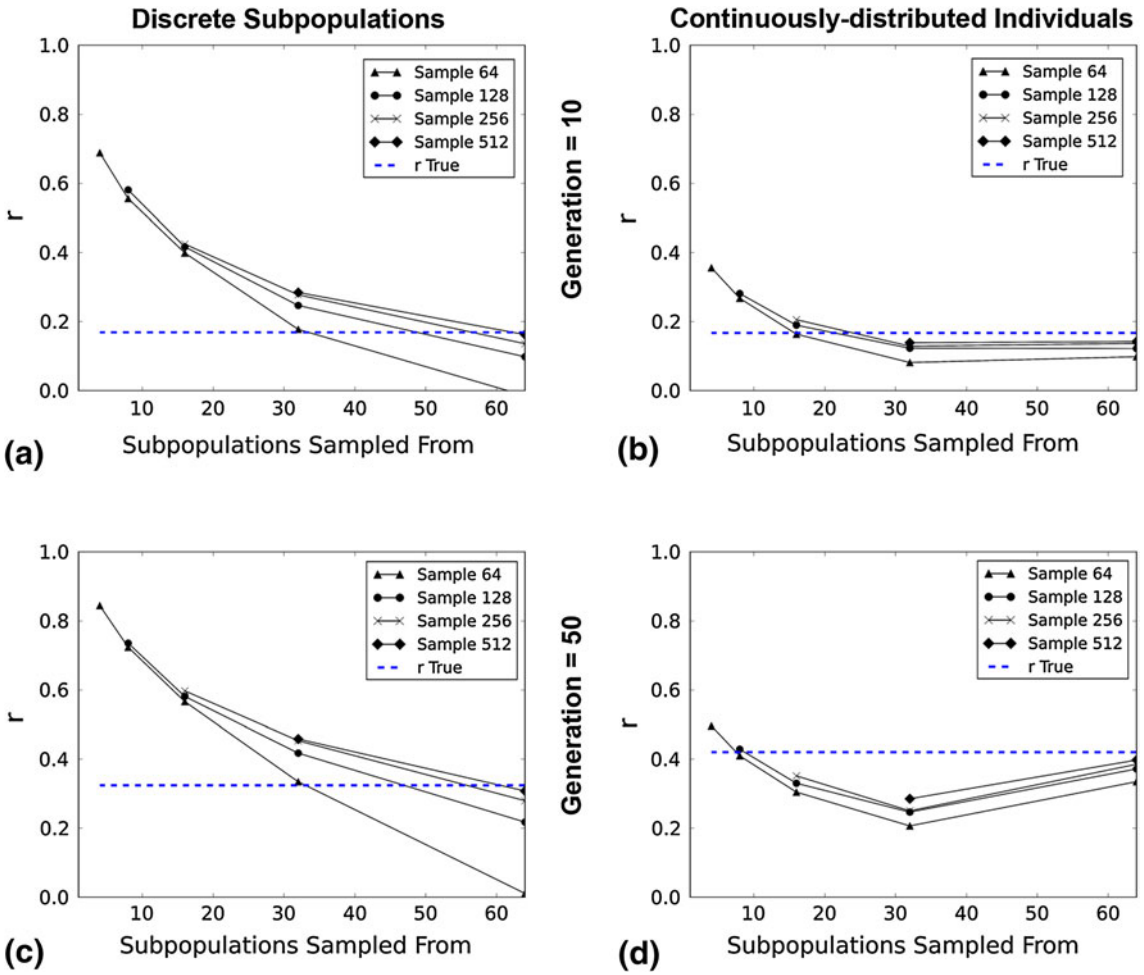

Fig. 5 Mantel *r* versus the number of subpopulations sampled (*S*) for the example $S = 64$ subpopulations and $I/S = 36$ individuals per subpopulation with sample allocation size of $n = [64, 128, 256, 512]$ for the patterns of dispersion of **a** individuals within discrete populations at generation 10, **b** continuously distributed individuals at generation 10, **c** individuals within discrete populations at generation 50, and **d** continuously distributed individuals at generation 50. The *dashed horizontal blue line* is the 'true' *r* value for that generation. Node locations along each line represent a simulated experiment and indicate the number of subpopulations that *n* was allocated across. (Color figure online)

Table 4 Top linear models (chosen using AIC and model weight) for optimal subpopulations to sample from, $\hat{S}$, for the individual-based metric D_{PS} and Mantel *r* calculation)

Pattern of dispersion	*S*	*I/S*	*N*	*n*	Int	ΔAIC	R^2	w
Discrete	0.628	–	−0.000	0.046	−1.102	0.00	0.97	0.52
	0.641	0.013	–	0.046	−1.551	1.76	0.97	0.22
	0.625	−0.003	−0.000	0.046	−0.988	1.94	0.97	0.20
	0.901	–	–	0.046	−1.104	4.24[a]	0.97	0.06
Continuous	0.089	−0.063	0.002	0.016	2.044	0.00	0.69	0.85
	0.145	–	0.001	0.016	−1.173	3.63[b]	0.69	0.14

S—number of total subpopulations within the population, *I/S*—number of individuals within each subpopulation, *N*—total number of individuals in population, *n*—sample allocation size, Int—intercept, R^2—adjusted R^2 value, w—model selection weight. The first grouping corresponds to five covariates, while the second grouping did not include *r*.true (NA)

[a] Next model ΔAIC = 834.00

[b] Next model ΔAIC = 9.49

on the beta values, *N* and *I/S* were the weakest predictors of $\hat{S}$, while *n* and *S* were the moderate to strongest predictors for both discrete and continuously distributed populations.

Discussion

This analysis formally evaluates the effect of sample allocation design on the ability to estimate genetic differentiation. One of the most important results from this work is the demonstration of the importance of matching the pattern of dispersion (continuous or grouped) and the metric chosen to analyze these data. It is clear that choice of metric and pattern of dispersion greatly influences sampling allocation design. For example, the population-based metric F_{ST}, intended for discrete groups, is much lower in the continuously distributed scenario (Fig. 2; column 1 – $\sim$ 0.6 in discrete verses column 2 – $\sim$ 0.1 in continuous), most likely due to continuous gene flow given no formal barriers in this scenario. However, it is important to note that most species will not fit precisely into one of the two extreme patterns of dispersion simulated here, but rather will fall somewhere on a continuum (e.g., areas of continuous distribution with interspersed barriers to dispersal, i.e., landscape heterogeneity). Researchers will often not know where their study species falls on this continuum, but that the data presented here are worth consideration during the interpretation of empirical data results.

When attempting to estimate population genetic structure with F_{ST}, there is a risk of over estimation when few individuals are collected across all subpopulations or sampling areas. In contrast, when samples are allocated such that all the sampling effort is from only a few subpopulations or sampling areas there is a risk of under estimating genetic structure (Figs. 2, 3). In the former case, when too few individuals are collected across subpopulations, we cannot accurately capture the true frequency of even common alleles. Thus, two small samples that are drawn from a population with nearly the same allele frequency distribution will have a positive F_{ST} due to sampling variation. However, in the latter case when sampling effort is concentrated on a few subpopulations, it is possible to miss pairs of populations that are more genetically distant, thus underestimating genetic structure as measured by F_{ST}.

When attempting to estimate population genetic structure with Mantel *r* using D_{PS}, the opposite results are observed; there is a risk of under estimating substructure when few individuals are collected across all subpopulations or areas in order to sample more locations. Here, we believe that collecting data from individuals at many locations but not accurately capturing the variability at each location causes most inter-individual comparisons to suggest no or little relationship among samples, as few alleles between samples will be the same, in turn underestimating genetic structure with a metric like D_{PS} (i.e., a decrease in power). The relationship and increase in power should improve with an increase in independent markers, as observed by Oyler-McCance et al. (2013). In contrast, when samples are collected from only a few subpopulations or sampling areas there is a risk of over estimating genetic structure (Figs. 4, 5). We believe this is an artifact of the sample allocation strategy (i.e., clustered design) that is heightened with the Mantel test: a few highly clustered sampling areas will act like they are separated by barriers and thus, the individual-based metric is missing many interstitial areas. Thus, when samples are on a continuum or on opposite ends of a gradient (e.g., isolation by distance), as in our simulations, genetic structure as estimated by D_{PS} is overestimated. We believe this result is similar to Meirmans (2012); i.e., patterns of isolation by distance can easily be mistaken for a hierarchical population structure. So when there are few highly clustered samples on a landscape (defining a hierarchical population structure) governed by a dispersal pattern of isolation by distance, we will see an even higher gradient in allele frequencies that are divided up into the respective sampling areas.

To illustrate the contrasting results obtained when the sampling design and the analytical method do not match, we consider the example $S = 25$, $I/S = 20$, and a desired sample size of $n = 100$. With a pattern of dispersion that mimics discrete subpopulations, our predictive models (Tables 3, 4) suggests allocating the 100 samples to 10.5 subpopulations when attempting to estimate F_{ST} for the entire population or 19.0 sampling areas when attempting to get the best estimate of Mantel *r* using D_{PS}. Increasing gene flow and looking at a pattern of dispersion that mimics a continuously distributed population, our predictive models suggests allocating the 100 samples to 5.9 subpopulations when attempting to estimate F_{ST} for the entire population or 5.6 sampling areas when attempting to estimate Mantel *r* using D_{PS}.

Overall, our simulations provide information on the potential risks sampling designs can have on reporting overall population structure for conservation and ecology. This risk is in addition to issues generated by misuse of the test statistic. Mantel tests have recently been criticized (Guillot and Rousset 2011; Meirmans 2012; Graves et al. 2013), despite many studies that have shown its usefulness under certain scenarios (e.g., Legendre and Fortin 2010; Landguth and Cushman 2010). Similar findings to that of Meirmans (2012), our variable results with the individual-based metric and Mantel test compared to that of the population-based metric, also points to problems with this approach if not used appropriately. Currently, there is no alternative individual-based distance test for assessing

population structure (but see Bradburd et al. 2013 for SNP datasets and ecological distance). However, the underlying metric in this study was D_{PS} and other correlative approaches could be considered (e.g., dbMEM; Legendre and Legendre 2012). Furthermore, very few studies have considered multiple (or alternative) distance-based metric choices to D_{PS} (but see Dyer et al. 2010 and Rousset and Leblois 2012), which could also impact sensitivity to population structure.

Overall, understanding the optimal sampling scheme to detect patterns of gene flow and substructure is an underappreciated element in molecular ecology and conservation genetics studies. There are several examples of where biases in sampling have impacted results and subsequent practical interpretation of these results. Tucker et al. (Tucker et al. 2013a, b) has shown that sampling an endangered carnivore species, the fisher (*Pekania* [*Martes*] *pennanti*), as if it were in discrete populations and not continuously distributed had strong impacts on population results and interpretation. Prior research, which had individuals in clusters, suggested strong population subdivision; yet when a sampling scheme that considered the species as continuously distributed was used, very little subdivision was evident (Tucker et al. 2013a). This example highlights the importance of understanding the complex interaction between the organism's natural history, the sampling scheme of the study, and ultimately the results. We hope that our work here can help those who are currently designing population genetic studies remove bias, and optimize sampling.

By using an individual-based simulation program, we were able to control for natural history strategies and thus, gene flow processes and resulting genetic structure as observed by two popular population and landscape genetics metrics. This factorial simulation provided a robust means to comprehensively evaluate the interactive effects of sampling design given each pattern of dispersion, and population structure and configuration on the landscape. However, these simulations provide a snapshot into the effects of sampling design and resulting genetic structure for conservation and ecology. Future studies should investigate the interaction of marker choice in this framework (e.g., SNP versus microsatellites; Willing et al. 2012) or estimates of neutral and adaptive differentiation. Such studies could also account for varying and fluctuating population sizes, overlapping generations, simulate more complex landscape resistance scenarios, and include source-sink dynamics. Such studies will not only help to determine the relative utility of sample design optimization efforts in conservation genetic studies, but also provide vital insights into fine-scale processes in heterogeneous environments.

Acknowledgments We thank two anonymous reviewers for comments on this manuscript.

References

Balloux F (2001) EASYPOP (Version 1.7): a computer program for population genetic simulations. J Hered 92:301–302

Barton K (2012) http://mumin.r-forge.r-project.org/MuMIn-manual.pdf

Bowcock AM, Ruiz-Linares A, Tomfohrde J, Minch E, Kidd JR, Cavalli-Sforza LL (1994) High resolution of human evolutionary trees with polymorphic microsatellites. Nature 368:455–457

Bradburd GS, Ralph PL, Graham MC (2013) Disentangling the effects of geographic and ecological isolation on genetic differentiation. Evolution. doi:10.1111/evo.12195

Burnham KP, Anderson DR (2002) Model selection and multimodel inference: a practical information-theoretic approach, 2nd edn. Springer, New York

Dyer RJ, Nason JD, Garrick RC (2010) Landscape modelling of gene flow: improved power using conditional genetic distance derived from the topology of population networks. Mol Ecol 19:3746–3759

Goslee SC, Urban DL (2007) The ecodist package for dissimilarity-based analysis of ecological data. J Stat Softw 22:1–19

Graves TA, Beier P, Royle A (2013) Current approaches using genetic distances produce poor estimates of landscape resistance to interindividual dispersal. Mol Ecol. doi:10.1111/mec.12348

Guillot G, Rousset F (2011) On the use of simple and partial Mantel tests in the presence of spatial auto-correlation, arXiv:1112.0651v1

Hoffman EA, Kolm N, Berglund A, Arguello JR, Jones AG (2005) Genetic structure in the coral-reef-associated *Banggai cardinalfish*, *Pterapogon kauderni*. Mol Ecol 14:1367–1375

Landguth EL, Cushman SA (2010) CDPOP: a spatially-explicit cost distance population genetics program. Mol Ecol Resour 10:156–161

Landguth EL, Cushman SA, Murphy M, Luikart G (2010) Relationships between migration rates and landscape resistance assessed using individual-based simulations. Mol Ecol Resour 10:854–862

Landguth EL, Cushman SA, Johnson NJ (2012a) Simulating natural selection in landscape genetics. Mol Ecol Resour 12:363–368

Landguth EL, Fedy BC, Garey A, Mumma M, Emel S, Oyler-McCance S, Cushman SA, Wagner HH, Fortin M-J (2012b) Effects of sample size, number of markers, and allelic richness on the detection of spatial genetic pattern. Mol Ecol Resour 12:276–284

Legendre P, Fortin M-J (2010) Comparison of the Mantel test and alternative approaches for detecting complex multivariate relationships in the spatial analysis of genetic data. Mol Ecol Resour 10:831–844

Legendre P, Legendre L (2012) Numerical ecology, 3rd edn. Elsevier, Amsterdam

Mantel N (1967) The detection of disease clustering and a generalized regression approach. Cancer Res 27:209–220

Meirmans PG (2012) The trouble with isolation by distance. Mol Ecol 21:2839–2846

Murphy M, Evans J, Cushman SA, Storfer A (2008) Representing genetic variation as continuous surfaces: an approach for identifying spatial dependency in landscape genetic studies. Ecography 31:685–697

Nei M (1973) Analysis of gene diversity in subdivided populations. Proc Natl Acad Sci USA 70:3321

Nei M, Chesser R (1983) Estimation of fixation indices and gene diversities. Ann Hum Genet 47:253–259

Novembre J et al (2008) Genes mirror geography within Europe. Nature 456:98–101

Ott J (1992) Strategies for characterizing highly polymorphic markers in human gene mapping. Am J Hum Genet 51:283–290

Oyler-McCance SJ, Fedy BC, Landguth EL (2013) Sample design effects in landscape genetics. Conserv Genet 14:275–285

Purcell JFH, Cowen RK, Hughes CR, Willams DA (2006) Weak genetic structure indicates strong dispersal limits: a tale of two coral reef fish. Proc Royal Soc B 273:1483–1490

Rao C (2001) Sample size considerations in genetic polymorphism studies. Hum Hered 52:191–200

R Development Core Team (2012) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. http://www.R-project.org

Rousset F (1997) Genetic differentiation and estimation of gene flow from F-statistics under isolation by distance. Genetics 145:1219–1228

Rousset F, Leblois R (2012) Likelihood-based inferences under a coalescent model of isolation by distance: two-dimensional habitats and confidence intervals. Mol Biol Evol 29:957–973

Schwartz MK, McKelvey KS (2009) Why sampling scheme matters: the effect of sampling scheme on landscape genetic results. Conserv Genet 10:441–452

Schwartz MK, Vucetich JA (2009) Molecules and beyond: assessing the distinctness of the Great Lakes wolf. Mol Ecol 18:2307–2309

Seeb JE, Carvalho G, Hauser L, Naish K, Roberts S, Seeb LW (2011) Single-nucleotide polymorphism (SNP) discovery and applications of SNP genotyping in nonmodel organisms. Mol Ecol Resour 11:1–8

Selkoe KA, Toonen RJ (2006) Microsatellites for ecologists: a practical guide to using and evaluating microsatellite markers. Ecol Lett 9:615–629

Siniscalco MR, Robledo PK, Bender C, Carcassi L, Contu L, Beck JC (1999) Population genomics in Sardinia: a novel approach to hunt for genomic combinations underlying complex traits and diseases. Cytogenet Cell Genet 86:148–152

Tucker JM, Schwartz MK, Truex RL, Pilgrim KL, Allendorf FW (2013a) Historical and contemporary DNA indicate fisher decline and isolation occurred prior to the European settlement of California. PLoS One. doi:10.1371/journal.pone.0052803

Tucker JM, Schwartz MK, Truex RL, Wisely SM, Allendorf FW (2013b) Sampling affects the detection of genetic subdivision and conservation implications for fisher in the Sierra Nevada. Conserv Genet. doi:10.1007/s10592-013-0525-4

Willing EM, Dreyer C, Oosterhout C (2012) Estimates of genetic differentiation measured by Fst do not necessarily require large sample sizes when using many SNP markers. PLoS One. doi:10.1371/journal.pone.0042649