Received: 7 September 2016 | Revised: 21 March 2017 | Accepted: 14 April 2017

DOI: 10.1111/1755-0998.12684

RESOURCE ARTICLE

WILEY MOLECULAR ECOLOGY RESOURCES

A comparison of individual-based genetic distance metrics for landscape genetics

A. J. Shirk[1] | E. L. Landguth[2] | S. A. Cushman[3]

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

[2]Computational Ecology Laboratory, Division of Biological Sciences, University of Montana, Missoula, MT, USA

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

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

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

Abstract

A major aim of landscape genetics is to understand how landscapes resist gene flow and thereby influence population genetic structure. An empirical understanding of this process provides a wealth of information that can be used to guide conservation and management of species in fragmented landscapes and also to predict how landscape change may affect population viability. Statistical approaches to infer the true model among competing alternatives are based on the strength of the relationship between pairwise genetic distances and landscape distances among sampled individuals in a population. A variety of methods have been devised to quantify individual genetic distances, but no study has yet compared their relative performance when used for model selection in landscape genetics. In this study, we used population genetic simulations to assess the accuracy of 16 individual-based genetic distance metrics under varying sample sizes and degree of population genetic structure. We found most metrics performed well when sample size and genetic structure was high. However, it was much more challenging to infer the true model when sample size and genetic structure was low. Under these conditions, we found genetic distance metrics based on principal components analysis were the most accurate (although several other metrics performed similarly), but only when they were derived from multiple principal components axes (the optimal number varied depending on the degree of population genetic structure). Our results provide guidance for which genetic distance metrics maximize model selection accuracy and thereby better inform conservation and management decisions based upon landscape genetic analysis.

KEYWORDS
genetic distance, isolation by distance, isolation by resistance, landscape genetics, model selection, principal components analysis

1 | INTRODUCTION

Landscape genetics is an emerging field of ecology that integrates population genetics, landscape ecology and spatial statistics (Manel, Schwartz, Luikart, & Taberlet, 2003; Balkenhol, Cushman, Storfer, & Waits, 2015). A major goal of landscape genetic analysis is a spatial understanding of how geographic distance and landscape heterogeneity limit gene flow. If gene flow is low relative to genetic drift, populations become differentiated, creating a spatial genetic pattern that can be used to infer landscape resistance to gene flow. Resistance models provide an empirical basis to inform conservation efforts such as designing wildlife movement corridors (Beier, Majka, Newell, & Garding, 2008; Cushman, Wasserman, Landguth, & Shirk, 2013; Epps, Wehausen, Bleich, Torres, & Brashares, 2007) and assessing the potential impacts of landscape and climate change on population genetic and demographic viability (Manel & Holderegger, 2013; Sommer, McDevitt, & Balkenhol, 2013; Wasserman et al., 2012). Indeed, the need for landscape genetic analysis and

connectivity conservation is becoming increasingly important in an era of rapid habitat conversion, fragmentation and range shifts driven by climate change (Segelbacher et al., 2010; Vos et al., 2008).

Landscape genetic analysis begins by forming hypotheses of how a landscape resists gene flow. A commonly tested hypothesis is based on the concept of isolation by distance (IBD; Wright, 1943), which posits that genetic differentiation is a function of the Euclidean distance between individuals or subpopulations. Alternatively, hypotheses (modelled as raster grids called resistance surfaces; Spear, Balkenhol, Fortin, McRae, & Scribner, 2010) may reflect the concept of isolation by resistance (IBR; McRae, 2006), where heterogeneous landscapes variably resist gene flow. The interaction between gene flow and genetic drift may create complex patterns of genetic differentiation on the landscape. In complex landscapes, distances are not measured in Euclidean units, but rather, in effective distances (e.g., least-cost or circuit distances; Adriaensen et al., 2003; McRae, 2006) given the variable resistance of landscape features that affect the focal species' movement (e.g., roads, forests, agricultural fields, cities). Collectively, we refer to Euclidean and effective distances calculated from landscape models of IBD or IBR, respectively, as landscape distance (LD).

Hypotheses of how landscapes resist gene flow are generally evaluated by statistical tests relating LD to empirical genetic distances (GD), which quantify pairwise genetic dissimilarity based on the multilocus genotypes of individuals sampled across the population of interest. In cases where populations form discrete, internally panmictic subpopulations, landscape and genetic distances are calculated between sampled subpopulations. However, most wild populations are more continuously distributed at varying densities across complex landscapes. Under these circumstances, distances are more appropriately calculated between individuals than subpopulations, because violating the assumption of discrete subpopulations can lead to bias and error in the estimate of genetic parameters (Shirk & Cushman, 2011, 2014). Sampling many individuals (often several hundred) rather than a few subpopulations provides individual-based analyses with much greater power to detect landscape genetic patterns (e.g., Landguth, Cushman, Murphy, & Luikart, 2010). However, there are a number of different individual-based GD metrics currently being used in landscape genetic analyses, and to date, there has been no comparison of their relative ability to quantify genetic differentiation in a way that maximizes the likelihood of identifying the true model among competing alternatives. Without a rigorous comparison, the choice of GD metric has largely been arbitrary, with unknown consequences for landscape genetic inferences.

To better understand differences in performance among existing individual-based GD metrics, we used population genetic simulations under both IBD and IBR to evaluate 16 metrics (Table 1) in terms of their accuracy in identifying the true model during model selection. These metrics included a variety of coefficients of kinship (*Kc*), fraternity (*Fc*) and relatedness (*Rc*), all of which are based on probabilities of alleles being identical by descent relative to a reference population (in this case, the sampled population). We also evaluated several metrics based only on the mathematical dissimilarity of genotypes, rather than the probabilities of identity by descent, including

TABLE 1 Individual-based genetic distance metrics evaluated, including abbreviation, assumption (+) of Hardy–Weinberg equilibrium (HWE) and the ploidy appropriate for the method

	Assumptions			
Metric	**Abbreviation**	**HWE**	**Ploidy**	**Reference**
Kinship coefficient	*Kc.Lo*		Any[a]	Loiselle, Sork, Nason, & Graham (1995)
	Kc.R		Any[a]	Ritland (1996)
Relatedness coefficient	*Rc.L&R*	+	2	Lynch & Ritland (1999)
	Rc.Q&G		Any[b]	Queller & Goodnight (1989)
	Rc.W	+	2	Wang (2002)
	Rc.Li	+	2	Li et al. (1993)
Fraternity coefficient	*Fc.L&R*	+	2	Lynch & Ritland (1999)
	Fc.W	+	2	Wang (2002)
Rousset's *a*	$\hat{a}$		≥2[c]	Rousset (2000)
Proportion of shared alleles	D_{PS}		Any	Bowcock et al., (1994)
Bray–Curtis distance	*BC*		Any	Bray & Curtis (1957)
Euclidean distance	*Euc*		Any	–
Principle component analysis	*PCA.1axis*		Any	Shirk et al. (2010)
	PCA.4axis		Any	–
	PCA.16axis		Any	–
	PCA.64axis		Any	–

This table was partially based on a similar table in the manual for the program SPAGEDI (Hardy & Vekemans, 2002).
[a]Ploidy of 1 to 8 allowed in the program SPAGEDI.
[b]Ploidy of 2 allowed in the program SPAGEDI.
[c]Ploidy of 2 to 8 allowed in the program SPAGEDI.

metrics based on principal components analysis (PCA), Euclidean distance, Bray–Curtis distance and the proportion of shared alleles (D_{PS}). Finally, we included a GD metric (Rousset's $\hat{a}$; Rousset, 2000) designed to reflect genetic differentiation among individuals in continuous populations under IBD (and by extension, IBR, if distances are given in effective rather than Euclidean units, as in Shirk & Cushman, 2011, 2014).

The simulations included two different levels of dispersal that created either a low or high degree of genetic structure within the population. We also evaluated GD metrics under two different samples sizes, including about 18% and 100% of the simulated population. We expected the accuracy of all metrics to be greatest when population genetic structure (i.e., the genetic signal) was high and sample size was high (i.e., the noise from sampling variation was low). Conversely, we expected accuracy to be lowest when sample size was small and genetic structure low. In addition, we hypothesized that GD metrics that assumed random mating (which is violated under IBD and IBR) would perform poorly relative to other metrics, particularly when population genetic structure was high. Finally, this comparison served as a test of a hypothesis we described in Shirk, Wallin, Cushman, Rice, and Warheit (2010), which proposed GD metrics based on PCA eigenvalues would have greater power to detect genetic patterns, particularly when population structure was low and sample size limited (Shirk et al., 2010).

2 | METHODS

2.1 | Population genetic simulation

We used CDPOP (Landguth & Cushman, 2010) to simulate 100 nonoverlapping generations of mating and dispersal among 1,085 diploid individuals arrayed in a hexagonal grid (Figure 1). CDPOP is an individual-based simulator of population genetic processes. It simulates mating and dispersal in a finite population 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 effective distances to the vacant locations. If all locations are occupied, any remaining offspring not yet assigned to a location are eliminated.

We ran 10 replicate sets of CDPOP simulations for each of two different maximum dispersal distances (d), including 20% or 100% of the maximum Euclidean distance between occupied locations. In each set, we simulated one population under IBD and 10 populations under IBR. With 11 simulations per set, 10 replicates and two levels of dispersal, there were a total of 220 simulations. All simulations were based on landscapes represented by a 1,024 × 1,024 cell raster

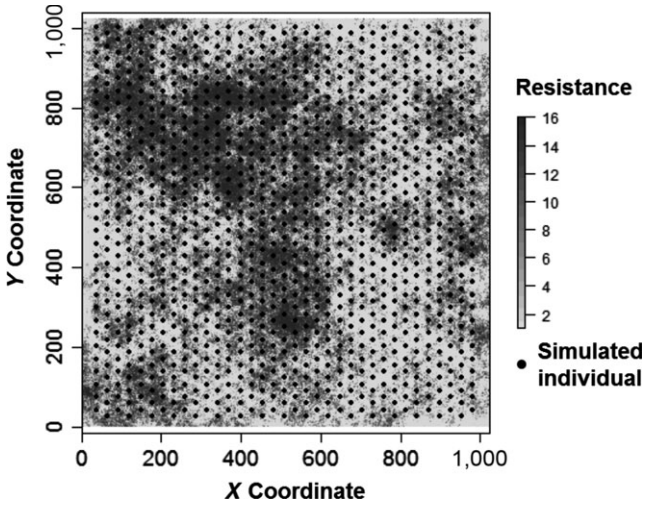

FIGURE 1 Simulated landscape and population. This is an example of an isolation-by-resistance raster (1,024 × 1,024 grid cells), with resistance varying from 1 (white) to 16 (black) and low resistance cells clustered with a cohesion value of 0.5. The population (n = 1,085; black circles) was arrayed in a hexagonal grid in all simulations

grid, with each side of the square cells measuring 1 unit of Euclidean distance. We used the program QRULE (Gardner, 1999) to create the IBR landscapes, with resistance values ranging from 1 to 16 in equal proportions. These values represented resistance to movement such that the "cost" to traverse a grid cell was equal to the cell size (1 unit) times the resistance. Thus, a value of one implies that Euclidean and effective distances are equal, and is the raster equivalent of IBD. In the IBR landscapes, values greater than one indicate additional cost to movement beyond the effect of distance alone. The lowest value (1) was distributed with a cohesion parameter of 0.5, which produced clumped distributions of resistance that are commonly observed in real landscapes (Figure 1). Because this pattern of resistance varied between the 10 IBR landscapes (Fig. S1), the effective landscape distance between individuals varied in the IBR simulations.

In the IBD simulations, dispersal and mating probabilities were a function of the inverse square of the Euclidean distance between individuals, with a threshold maximum distance (d) set to either 20% or 100% of the largest pairwise Euclidean distance between individuals in the landscape (1,345 units). In the IBR simulations, dispersal and mating probabilities were a function of the inverse square of the cost-weighted distance between individuals, with a threshold maximum distance set to 20% or 100% of the largest pairwise cost-weighted distance between individuals in the landscape (mean = 3,200, SD = 438). We calculated cost distances (i.e., accumulated cost of the least-cost path between all pairs of individuals, given the raster resistance surface) with the "GDISTANCE" package (van Etten, 2014) in the R statistical environment (R Core Team 2013). We expect very low genetic differentiation in simulations where dispersal was allowed up to 100% of the maximum pairwise distances in both the IBD and IBR simulations. Conversely, we expected high genetic differentiation when dispersal was limited to 20% of the maximum pairwise distances.

Except for varying mating and dispersal probability as a function of the landscape (IBD or one of the 10 IBR landscapes), the CDPOP parameters were the same in all simulations and based on a constant population of 1,085 individuals meeting Wright–Fisher assumptions (Wright, 1931). 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. This provided 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. In each generation for 100 generations, CDPOP recorded the genotypes of all individuals in the population. Previous simulations using similar landscapes and populations have shown that landscape genetic patterns emerge and equilibrate within 100 simulated generations (e.g., Landguth et al., 2010).

To track the formation of genetic structure over time within the simulations, we calculated the inbreeding coefficient (F_{IS}) from observed (H_o) and expected (H_e) heterozygosity recorded for all individuals each generation in the CDPOP output file ($F_{IS} = 1 - H_o/H_e$). F_{IS} quantifies the reduction in heterozygosity due to nonrandom mating within the sampling extent. Positive F_{IS} values indicate the local extent of mating and dispersal is small relative to the full extent of the sampling area (Shirk & Cushman, 2014). We used F_{IS} to track the formation of population genetic structure over time in the simulation and to confirm that the degree of genetic structure matches our expectations (i.e., population genetic structure should be greater for IBR simulations compared to IBD and when dispersal is low compared to when dispersal is high).

2.2 | Genetic distances

We calculated individual-based GD metrics from the genotypes of each individual in the population recorded by CDPOP after 100 simulated generations. All GD metrics were represented as $N \times N$ distance matrices, where N equals the sample size (either 200 or 1,085). We calculated individual-based GD using 16 different methods listed in Table 1. We used the program SPAGEDI (Hardy & Vekemans, 2002) to calculate the kinship and relatedness coefficients as well as Rousset's $\hat{a}$. We used the R package "ADEGENET" (Jombart, 2008) to calculate proportion of shared alleles. We calculated the PCA-based GD metrics in the R statistical environment (R Core Team 2013) by first calculating principal components (PC) from allele usage (0, 1 or 2) for all alleles in the population and then creating distance matrices from the Euclidean distance among varying numbers of PC axes (1, 4, 16 or 64; Shirk et al., 2010). Finally, we calculated Euclidean and Bray–Curtis GD using the R package "ECODIST" (Goslee & Urban, 2007).

To assess the relationships among individual-based GD metrics, we computed a dissimilarity matrix among all metrics based on the complement of the Pearson correlation averaged across all 220 simulations (20 IBD and 200 IBR simulations). We then used the "stats" package in R to perform hierarchical clustering on the dissimilarity matrix and plotted the results with a dendrogram.

2.3 | Model selection

For each of the 10 replicate CDPOP runs, there were 10 IBR simulations (based on the 10 IBR resistance surfaces described above) and 1 IBD simulation, for a total of 11 simulations per replicate. We evaluated each GD metric in terms of its ability to correctly identify the true model among the 11 competing alternatives. We used the MLPE.lmm() function of the R package "RESISTANCEGA" (Peterman, Connette, Semlitsch, & Eggert, 2014) to fit linear mixed-effects models with maximum-likelihood population effects (MLPE) for each competing model (Clarke, Rothery, & Raybould, 2002). The mixed-effects model implemented in this function accounts for the nonindependence inherent in pairwise distance matrices, avoiding problems associated with other common landscape genetic model selection methods (Clarke et al., 2002). The response variable was GD (based on the multilocus genotypes of the simulation corresponding to the true resistance model), and the predictor variable was log-transformed LD values (pairwise cost distances calculated between all individuals given the resistance model, or Euclidean distance if the model was IBD). We used the Akaike information criterion (AIC, obtained from the fitted MLPE model) to determine which resistance hypothesis among the 11 candidates was most related to the pattern of genetic differentiation quantified by the GD matrix (based on the lowest AIC score). We then calculated the per cent accuracy for each GD metric based on the proportion of times it identified the true model among competing alternatives during the 110 iterations (10 replicates of 11 resistance models). For comparison, a random model selection would have an accuracy of 1/11 = 9.1%. In addition to calculating accuracy, we also used the sem.model.fits() function in the R package PIECEWISESEM (Lefcheck, 2016) to calculate the top model's conditional R^2, a measure of model fit for mixed-effects models that includes both fixed and random effects.

The accuracy assessment described above reflects a binary correct/incorrect classification of accuracy. However, a GD metric could identify an incorrect model that was similar in resistance values to the true model, and therefore be, in a sense, partially correct. To quantify the degree to which the identified model matched the true model, we calculated the mean absolute difference between the resistance of the true model and the top model.

3 | RESULTS

3.1 | Simulations

In the high dispersal (d = 100%) simulations, under both IBD and IBR, population genetic structure (i.e., nonrandom mating measured by F_{IS} calculated over the entire simulated population) arose quickly within the first five generations and then plateaued at a low level with $F_{IS} < 0.01$ for the remainder of the simulation (Figure 2). In the low dispersal (d = 20%) simulations, under both IBD and IBR,

population structure arose more slowly and the increase in F_{IS} did not plateau until about 40–50 generations. The IBR simulations with low dispersal exhibited slightly higher genetic structure after the plateau ($F_{IS} \sim 0.05$) compared to the low dispersal IBD simulations ($F_{IS} \sim 0.04$).

3.2 | GD metric evaluation

The two fraternity coefficients (Fc.L&R and Fc.W) and the four related PCA-based GD metrics formed distinct branches in the dendrogram, reflecting low correlation to other GD metrics (Figure 3). The other metrics were all highly correlated ($r > 0.70$) and formed a distinct cluster in the dendrogram. The BC and D_{PS} metrics were perfectly correlated and henceforth will be treated as a single metric.

Across all metrics, accuracy was highest (mean = 92.3%) when sample size was high (n = 1,085) and dispersal was low (d = 20%). Conversely, accuracy was lowest (mean = 48.9%) when sample size was low (n = 200) and dispersal was high (d = 100%). Intermediate levels of accuracy (mean = 79.9%) occurred when either sample size or dispersal was low, but not both (Figure 4; Table S1). Under the most favourable conditions to identify the true model (high sample size and low dispersal), all but three of the GD metrics achieved near 100% accuracy. The three that underperformed under these conditions included the two fraternity coefficients (Fc.L&R and Fc.W) and the PCA metric using only 1 PC axis. Under the most difficult conditions to identify the true model (low sample size and high dispersal), the PCA metric using 64 PC axes was the most accurate metric. Accuracy of the PCA metric increased as more PC axes were used. The poorest performing metrics remained the same as above and included the two fraternity coefficients and the PCA metric with only 1-axis. All other metrics performed similarly. In intermediate conditions (low sample size or low dispersal, but not both), the most accurate metrics included three of the four relatedness coefficients (Rc.Q&G, Rc.W and Rc.Li), Rousset's $\hat{a}$, D_{PS}/BC, Euclidean and the PCA-based GD metrics with >1 axis (Figure 4 middle two panels). The least accurate metrics remained the same as above and included the two fraternity coefficients and the PCA metric with only 1-axis. The two kinship coefficients (Kc.Lo and Kc.R) as well as the relatedness coefficient of Lynch and Ritland (Rc.L&R) formed a middle tier of accuracy, generally underperforming the top tier and outperforming the bottom tier.

In general, GD metrics associated with high model selection accuracy also had high model fit (as measured by conditional R^2 of the mixed-effects model relating GD to LD; Fig. S2). For example, as the number of PC axes increased in the PCA-based GD metric, model selection accuracy generally increased and so did the R^2.

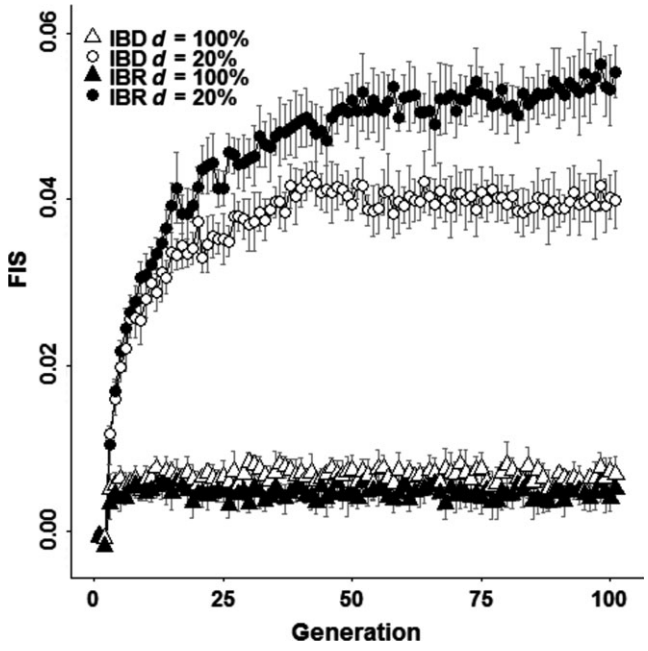

FIGURE 2 Establishment of population genetic structure. Population genetic structure, as measured by the mean inbreeding coefficient (F_{IS}) calculated from all simulated individuals, plateaued within the first few generations for both the isolation-by-distance (IBD; white symbols) and isolation-by-resistance (IBR; black symbols) simulations when maximum dispersal distance (d) was 20% of the maximum pairwise landscape distance (triangle symbols). When dispersal was higher (d = 100% of maximum pairwise landscape distance), genetic structure established more slowly, plateauing after about 40–50 generations. Values represent the mean of F_{IS} across 10 (for IBD) or 100 (for IBR) simulations, and error bars represent the standard deviation

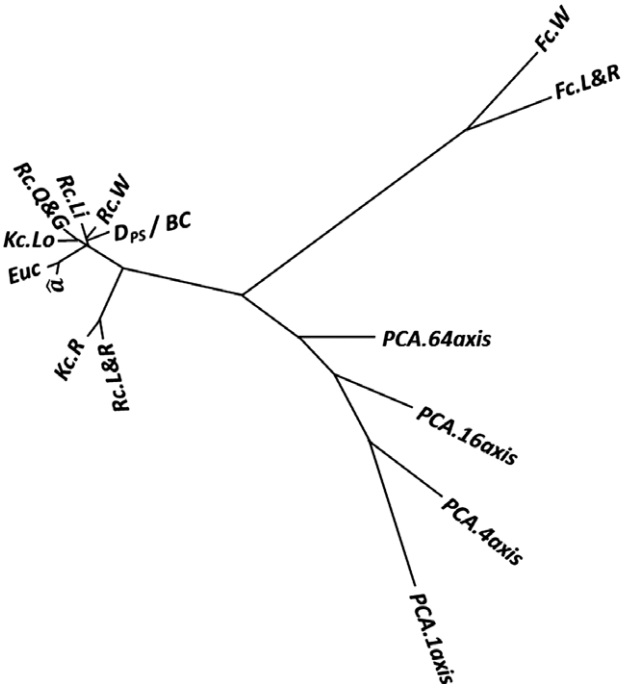

FIGURE 3 Hierarchical clustering of genetic distance metrics. The dendrogram branch distances correspond to the compliment of the Pearson correlation among the 16 individual-based genetic distance metrics, averaged across 220 simulations (20 IBD and 200 IBR)

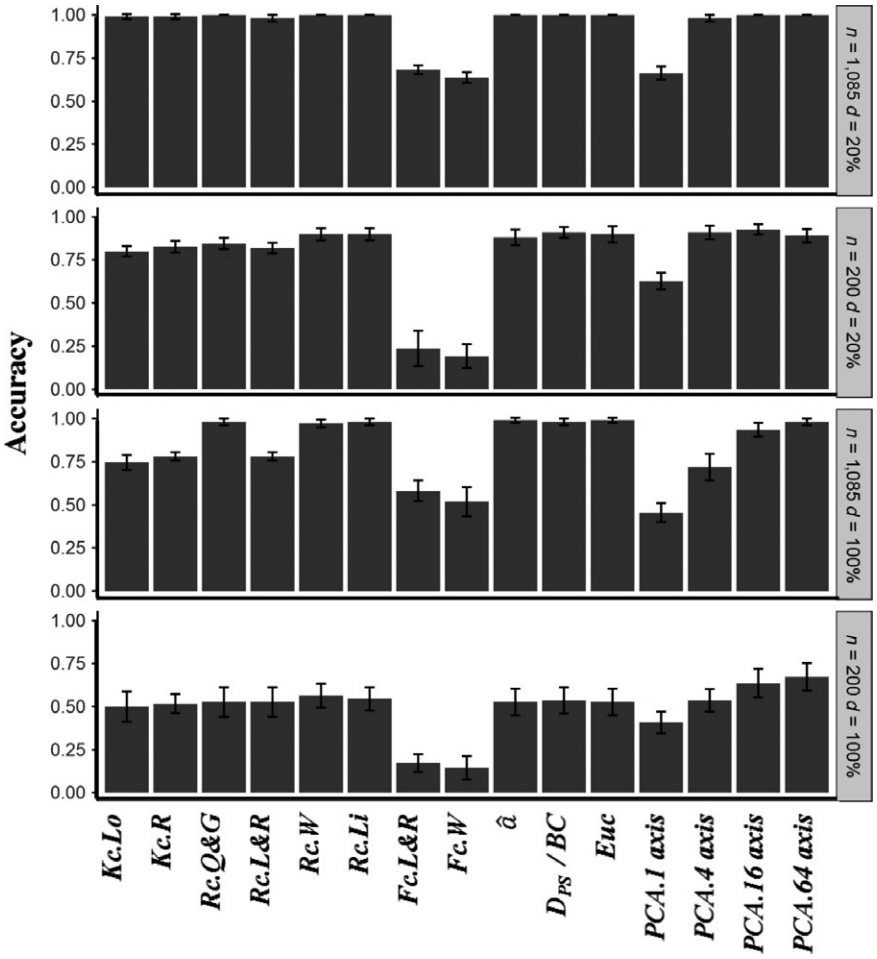

FIGURE 4 Model selection accuracy. The proportion of the time the correct model was identified among 11 competing hypotheses is shown for each of 16 individual-based genetic distance metrics, at two levels of sample size (n = 200 or 1,085) and two levels of maximum dispersal (d = 20% or 100% of the maximum pairwise landscape distance). Values represent the mean accuracy across 10 isolation-by-distance simulations and 100 isolation-by-resistance simulations, and error bars represent the standard deviation

Conversely, the two fraternity coefficients consistently had the lowest accuracy in model selection and had very low corresponding R^2 values. However, this relationship was not always consistent. For example, the Rousset's $\hat{a}$ and Euclidean GD metrics had R^2 values at least twice as high as the other metrics not based on PCA, but this higher degree of model fit did not translate into greater model selection accuracy.

The mean absolute error (MAE) in resistance (i.e., the average absolute difference between the true resistance model and the model identified during model selection) reflected the same relative differences among GD metrics, varying by metric, sample size and dispersal distance. Under optimal conditions (high sample size and low dispersal), all metrics except the two fraternity coefficients and the 1-axis PCA metric exhibited low MAE approaching zero (Fig. S3). Under the most difficult conditions (i.e., low sample size and/or high dispersal), the two fraternity coefficients had the highest MAE (almost 5) and the high axis (≥16) PCA metrics had the lowest MAE (approaching 2 as the number of axes increased).

The distribution of variance explained by the PC axes varied depending on the strength of dispersal (Figure 5). When dispersal was low (d = 20%), the first two PC axes explained significantly more variance than subsequent axes (based on the broken stick model criterion; McGarigal, Cushman, & Stafford, 2000). However, when dispersal was high (d = 100%), the variance explained by each subsequent axis was only slightly lower, and no PC axes were significant.

4 | DISCUSSION

We have demonstrated that the accuracy of model selection varies as a function of sample size, the degree of spatial structure in the population (a function of the dispersal parameter) and the metric used to quantify genetic distances among individuals. The importance of sample size and dispersal was expected based both on population genetic theory and on recent studies that have explored model selection accuracy in landscape genetics (Kierepka & Latch, 2015; Landguth et al., 2012; Zeller et al., 2016). However, the differences we observed in model selection accuracy among 16 individual-based genetic distance metrics are a novel finding that fills a key knowledge gap in the practice of landscape genetics. At present, there is no consensus on which genetic distance metrics to use and few studies provide justification for their choice of metric. Our results provide guidance regarding which metrics produce the most accurate inferences, particularly under the most challenging conditions that occur when sample sizes are limited and dispersal is high.

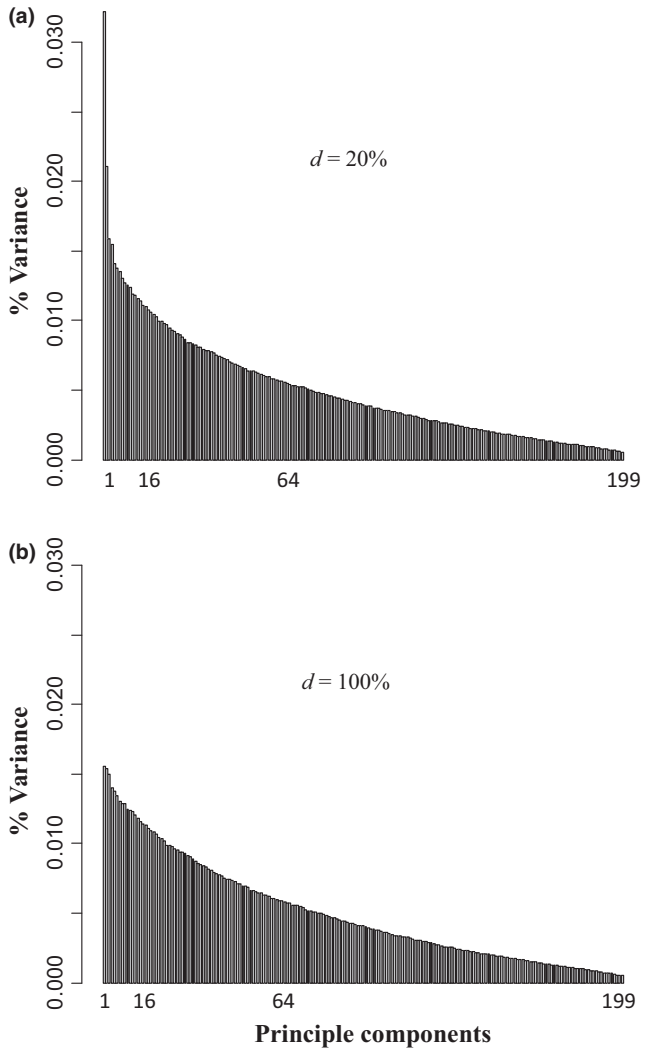

FIGURE 5 Variance explained by principle component (PC) axes. The proportion of the variance explained by each of 199 PC axes is shown for two levels of maximum dispersal (*d*); (a) *d* = 20% and (b) *d* = 100% of the maximum pairwise landscape distance. Values reflect the mean of 100 isolation-by-resistance simulations, with sample size = 200

Shirk et al. (2010) hypothesized that PCA-based metrics would perform best under these conditions because PCA concentrates variance among loci into composite gradients that should provide a stronger signal than metrics that weight all loci (even those with little variability) equally. Our results confirm this hypothesis, but also show that some of the simplest metrics, such as Euclidean distance, Bray–Curtis and D_{PS}, perform nearly as well under the conditions of our evaluation. Use of the best performing metrics should improve the reliability and consistency of landscape genetic analyses and better inform conservation and management decisions based on them.

As expected, detecting the true landscape model underlying the simulated pattern of genetic differentiation was most reliable when sample size was large and population genetic structure was high. This has been demonstrated in several comparative landscape genetic studies using simulations (Landguth et al., 2010, 2012; Zeller et al., 2016). Under these conditions in the present study, most GD metrics achieved nearly 100% accuracy. However, both fraternity coefficient metrics and the 1-axis PCA metric performed poorly by comparison. Fraternity coefficients have identical probabilities of identity by descent for most closely related individuals (except full-siblings and double first cousins). It may be that the inability to resolve closely related individuals limits the ability of fraternity coefficients to quantify genetic distances in a way that maximizes model selection accuracy in landscape genetics. We are aware of no published studies using fraternity coefficients in landscape genetic model selection (although fraternity coefficients have been used in landscape genetic analyses; e.g., Oddou-Muratorio, Demesure-Musch, Pélissier, & Gouyon, 2004). Our results support the continued avoidance of these metrics in landscape genetic model selection, regardless of population structure or sample size.

The PCA GD metric based on 1 PC axis also performed poorly relative to other metrics under all levels of sample size and dispersal. Unlike fraternity coefficients, however, this metric has been used previously in landscape genetic model selection (Shirk et al., 2010). Given that including more than one PC axis greatly improved model selection accuracy in our simulations under all conditions, our results suggest future landscape genetic studies should consider including multiple axes. Indeed, there are already examples of landscape genetic model selection (e.g., Castillo, Epps, Davis, & Cushman, 2014) and landscape genetic analyses (e.g., Grivet, Sork, Westfall, & Davis, 2008) that have employed multiple PC axes to quantify genetic distances.

The increase in model selection accuracy due to including multiple PC axes raises the question of exactly how many to include. Generally, PCA-based analyses involve quantifying the per cent of variation explained by each PC axis and then interpreting or summarizing only a relatively few axes that are significant (e.g., as determined by statistical tests such as the broken stick model; Jackson, 1993; Peres-Neto, Jackson, & Somers, 2005). In landscape genetics of discrete populations separated by strong barriers, these large eigenvectors have been shown to correspond to the major genetic clusters in a population (Jombart, 2008; Patterson, Price, & Reich, 2006); however, their meaning in continuous populations under IBD and IBR is less clear. In our simulations, when dispersal was low (*d* = 20%), the first several eigenvectors were generally significant (Figure 5, panel A). Including these first several axes (as opposed to just the first PC axis) resulted in a large improvement in model selection accuracy, but adding additional axes beyond the significant ones did not improve accuracy further. In fact, the PC axes explaining the lowest variance are likely driven by sampling noise and including them should reduce model selection accuracy. The effect of adding "noisy" axes may have been responsible for the slightly decreased accuracy we observed when 64 of the 199 PC axes (when *n* = 200) were included (although the differences were not significant).

When dispersal was high (*d* = 100%), no PC axes were significant using the broken stick test (Figure 5, panel B). Under these conditions, model selection accuracy increased as more axes were added up to the maximum we evaluated (64). This suggests the subtle population structure under these conditions is not captured by a few large eigenvectors. Thus, if the first few PC axes are not significant,

it indicates a more cryptic population structure that may require many PC axes to quantify. Under the most challenging conditions to identify the true landscape model (i.e., high dispersal and low sample size), the high-dimension PCA-based metric (with 64 PC axes) was the most accurate of all GD metrics. This supports the notion that PCA-based GD metrics are more sensitive compared to other methods because the largest eigenvectors are comprised of alleles that are more variable in the population and therefore more diagnostic of population genetic structure compared to common alleles (as proposed by Shirk et al., 2010).

Although our study highlights the potential advantages of PCA-based GD metrics, we note that the spatial configuration of the population on the landscape has been shown to affect the distribution of variance explained among the eigenvectors (Patterson et al., 2006). Thus, the appropriate number of PC axes to include may vary with the population distribution. For example, even though the 1-axis PCA-based GD metric performed poorly in this study, in Shirk et al. (2010), the 1-axis PCA-based GD metric exhibited a surprisingly strong correlation to several of the related landscape resistance models evaluated during model selection. It may be that differentiation along the linear distribution of the population (i.e., mountain goats distributed along a linear mountain range) was adequately quantified by a single dominant eigenvector. Exploration of the performance of PCA-based GD metrics under varying population configurations (not just the square uniform distribution used in this study), landscape configurations, number of eigenvectors and the proportion of variance explained by the included eigenvectors should provide insights to improve the reliability of this approach under diverse population structures.

Aside from the most accurate (multiaxis PCA) and least accurate (fraternity coefficients and 1-axis PCA) metrics, there was little difference among the other metrics under most conditions. All of these middle-tier metrics clustered together in the same tertiary branch of the dendrogram and were highly correlated with each other ($r > 0.70$), so it is not surprising they were similar in performance. Among them, three of the four relatedness coefficients, as well as Rousset's $\hat{a}$, D_{PS}, Bray–Curtis and Euclidean, all performed slightly better than the two kinship coefficients or the relatedness coefficient of Lynch and Ritland (1999) under most conditions, although the differences were not significant. Surprisingly, even though the assumption of a large random breeding population in HWE was violated in all simulations (particularly when dispersal was low), the relatedness coefficients of Wang (2002) and Li, Weeks, and Chakravarti (1993) performed as well as Rousset's $\hat{a}$, which was designed to quantify genetic differentiation in continuous populations isolated by distance. Violation of the HWE assumption may produce an upward bias in the estimate of relatedness for these coefficients (Wang, 2011). It appears that the relative differences among individuals were preserved, despite the bias, such that the pairwise genetic distances were still strongly related to the corresponding pairwise landscape distances. If so, meeting this assumption may be more critical for analyses where the absolute value of relatedness matters more than the relative values between individuals.

Graves, Beier, and Royle (2013) noted that the stochastic processes that produce genetic variation result in a high degree of variation in interindividual genetic distances. This produces a high signal-to-noise ratio that has the potential to confound landscape genetic analysis, particularly those that assume a linear relationship between genetic and landscape distances. To explore this issue, they used population genetic simulations to assess model selection accuracy in the context of stochastic population processes. Specifically, they systematically explored the parameter space using Mantel tests (Mantel, 1967) to relate simulated genetic distances (using the D_{PS} metric) to effective distances and found that the optimal parameters rarely matched the true parameters of the simulated resistance model. Ultimately, they concluded that "the picture looks quite glum for precisely or accurately estimating resistance values using interindividual genetic distances and Mantel correlations." Our study is not directly comparable because we used a different model selection approach (based on AIC values from mixed-effects models, rather than Mantel tests) and our resistance model parameter space was much more restricted (i.e., we only explored 11 univariate resistance hypotheses per model selection exercise, compared to a large multivariate parameter space). However, our results indicate that it is possible to detect landscape genetic relationships with high accuracy using linear models despite the low signal-to-noise ratio, particularly in highly structured populations (i.e., when dispersal is low relative to landscape distances between individuals).

Unlike the Mantel test used in Graves et al. (2013), the mixed-effects model we used for model selection includes random effects that account for the nonindependence inherent in the distance matrices. This alone may explain our contrasting observation that linear models can potentially yield highly accurate landscape genetic inferences. However, our results suggest the choice of GD metric is also important. It is common in landscape genetic analysis with linear models to observe poor measures of model fit because "noisy" genetic distances do not form strongly linear relationships to landscape distances, even with log transformation. Indeed, the conditional R^2 values in our study for many of the metrics was very low (<0.25), and the GD metrics with the lowest model fit were the least accurate in model selection. However, the high-dimension PCA-based GD metrics exhibited a much stronger linear relationship to landscape distances, with conditional R^2 values approaching 0.75. This suggests PCA-based GD metrics improve the linearity of landscape genetic relationships, perhaps explaining their greater accuracy relative to other GD metrics in our study.

For new landscape genetic studies, our results suggest PCA-based GD metrics offer the best prospects for model selection accuracy and linearizing the relationship between GD and LD, although the number of PC axes to include requires further study. GD metrics based on Euclidean distance, Rousset's $\hat{a}$, Bray–Curtis, D_{PS} and relatedness coefficients (except *Rc.L&R*) offer alternatives to PCA GD metrics with nearly the same accuracy. Model fit using these metrics was low, however, suggesting these metrics are noisier and less linearly related to LD compared to PCA-based metrics. The Euclidean, Bray–Curtis, D_{PS} and PCA-based GD metrics have the advantage of making no

biological assumptions, so they are appropriate for any population, at any level of ploidy or inbreeding. For previous landscape genetic studies that did not use the top-performing metrics, it may be informative to repeat them with better performing GD metrics. Indeed, several studies have already explored the congruence of two or more GD metrics to better understand the sensitivity of their results to the metric chosen (e.g., Castillo et al., 2014; Shirk et al., 2010). If genetic differentiation in the population under study is very high and sample size is large, our results suggest the choice of metric is much less important, as all metrics except the two fraternity coefficients and the 1-axis PCA metric had accuracy approaching 100%. However, in cases where distances between sampled individuals or subpopulations exceeds the local extent of mating and dispersal (e.g., in a highly dispersed and clumped population, or a highly dispersed and clumped sampling of a continuous population), the relationship between landscape and genetic distances becomes uncoupled. Despite high genetic differentiation under these circumstances, no genetic distance metric would be expected to perform well, regardless of sample size.

The focus of this study was to compare the performance of GD metrics in landscape genetic model selection. In this comparison, we used a limited set of landscape models, an idealized population continuously distributed across the landscape, and a single statistical metric (AIC derived from mixed-effects modelling with maximum-likelihood population effects) to evaluate the relationship between LD and GD. We have no reason to expect the relative performance among GD metrics to change in the context of different landscapes, assumptions or model selection criterion. However, it is possible our results are sensitive to these sources of variation. Understanding how the accuracy of landscape genetic inferences is affected by the way hypothesis space is searched and the properties of the statistical test remain key knowledge gaps in this emerging field, although recent studies are helping to fill that void (Balkenhol, Waits, & Dezzani, 2009; Cushman et al., 2013; Kierepka & Latch, 2015; Zeller et al., 2016). Future studies aimed at better understanding these aspects of landscape genetics may benefit from using the best performing GD metrics.

ACKNOWLEDGMENTS

This research was supported in part by funds provided by National Science Foundation grant #EF-1442597.

AUTHOR CONTRIBUTIONS

A.J.S. performed the analysis and wrote the manuscript. E.L.L. ran the CDPOP simulations and provided helpful comments on the manuscript. S.A.C. provided guidance on the analysis and helpful comments on the manuscript.

DATA ACCESSIBILITY

Simulated landscapes, distance matrices, R code and CDPOP simulation outputs are available on DRYAD. CDPOP software and user manual are available at http://cel.dbs.umt.edu/cms/CDPOP.

REFERENCES

Adriaensen, F., Chardon, J. P., De Blust, G., Swinnen, E., Villalba, S., Gulinck, H., & Matthysen, E. (2003). The application of "least-cost" modelling as a functional landscape model. *Landscape and Urban Planning*, *64*, 233–247.

Balkenhol, N., Cushman, S. A., Storfer, A. T., & Waits, L. P. (2015). *Landscape genetics: Concepts, methods, applications*. West Sussex, UK: John Wiley & Sons.

Balkenhol, N., Waits, L. P., & Dezzani, R. J. (2009). Statistical approaches in landscape genetics: An evaluation of methods for linking landscape and genetic data. *Ecography*, *32*, 818–830.

Beier, P., Majka, D., Newell, S., & Garding, E. (2008). Best management practices for wildlife corridors. *America*, *178*, 1–14.

Bowcock, A. M., Ruiz-Linares, A., Tomfohrde, J., Minch, E., Kidd, J. R., & Cavalli-Sforza, L. L. (1994). High resolution of human evolutionary trees with polymorphic microsatellites. *Nature*, *368*, 455–457.

Bray, J. R., & Curtis, J. T. (1957). An ordination of the upland forest communities of southern Wisconsin. *Ecological Monographs*, *27*, 325.

Castillo, J. A., Epps, C. W., Davis, A. R., & Cushman, S. A. (2014). Landscape effects on gene flow for a climate-sensitive montane species, the American pika. *Molecular Ecology*, *23*, 843–856.

Clarke, R. T., Rothery, P., & Raybould, A. F. (2002). Confidence limits for regression relationships between distance matrices: estimating gene flow with distance. *Journal of Agricultural, Biological, and Environmental Statistics*, *7*, 361–372.

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*, 51–72.

Epps, C. W., Wehausen, J. D., Bleich, V. C., Torres, S. G., & Brashares, J. S. (2007). Optimizing dispersal and corridor models using landscape genetics. *Journal of Applied Ecology*, *44*, 714–724.

Gardner, R. H. (1999). *Rule: Map generation and a spatial analysis program*.

Goslee, S. C., & Urban, D. L. (2007). The ecodist package for dissimilarity-based analysis of ecological data. *Journal of Statistical Software*, *22*, 1–19.

Graves, T. A., Beier, P., & Royle, J. A. (2013). Current approaches using genetic distances produce poor estimates of landscape resistance to interindividual dispersal. *Molecular Ecology*, *22*, 3888–3903.

Grivet, D., Sork, V. L., Westfall, R. D., & Davis, F. W. (2008). Conserving the evolutionary potential of California valley oak (*Quercus lobata* N, e): A multivariate genetic approach to conservation planning. *Molecular Ecology*, *17*, 139–156.

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*, 618–620.

Jackson, D. A. (1993). Stopping rules in principal component analysis: A comparison of heuristical and statistical approaches. *Ecology*, *74*, 2204–2214.

Jombart, T. (2008). ADEGENET: A R package for the multivariate analysis of genetic markers. *Bioinformatics*, *24*, 1403–1405.

Kierepka, E. M., & Latch, E. K. (2015). Performance of partial statistics in individual-based landscape genetics. *Molecular Ecology Resources*, *15*, 512–525.

Landguth, E. L., & Cushman, S. A. (2010). CDPOP: A spatially explicit cost distance population genetics program. *Molecular Ecology Resources*, *10*, 156–161.

Landguth, E. L., Cushman, S. A., Murphy, M. A., & Luikart, G. (2010). Relationships between migration rates and landscape resistance assessed using individual-based simulations. *Molecular Ecology Resources*, *10*, 854–862.

Landguth, E. L., Fedy, B. C., Oyler-Mccance, S. J., Garey, A. L., Emel, S. L., Mumma, M., … 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*, 276–284.

Lefcheck, J. S. (2016). PIECEWISESEM: Piecewise structural equation modelling in R for ecology, evolution, and systematics. *Methods in Ecology and Evolution*, *7*, 573–579.

Li, C. C., Weeks, D. E., & Chakravarti, A. (1993). Similarity of DNA fingerprints due to chance and relatedness. *Human Heredity*, *43*, 45–52.

Loiselle, B. A., Sork, V. L., Nason, J., & Graham, C. (1995). Spatial genetic structure of a tropical understory shrub, Psychotria officinalis (Rubiaceae). *American Journal of Botany*, *82*, 1420–1425.

Lynch, M., & Ritland, K. (1999). Estimation of pairwise relatedness with molecular markers. *Genetics*, *152*, 1753–1766.

Manel, S., & Holderegger, R. (2013). Ten years of landscape genetics. *Trends in Ecology and Evolution*, *28*, 614–621.

Manel, S., Schwartz, M. K., Luikart, G., & Taberlet, P. (2003). Landscape genetics: Combining landscape ecology and population genetics. *Trends in Ecology and Evolution*, *18*, 189–197.

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

McGarigal, K., Cushman, S., & Stafford, S. (2000). *Multivariate statistics for wildlife and ecology research*. New York, NY: Springer Science + Business Media.

McRae, B. H. (2006). Isolation by resistance. *Evolution*, *60*, 1551–1561.

Oddou-Muratorio, S., Demesure-Musch, B., Pélissier, R., & Gouyon, P. H. (2004). Impacts of gene flow and logging history on the local genetic structure of a scattered tree species, *Sorbus torminalis* L. Crantz. *Molecular Ecology*, *13*, 3689–3702.

Patterson, N., Price, A. L., & Reich, D. (2006). Population structure and eigenanalysis. *PLoS Genetics*, *2*, 2074–2093.

Peres-Neto, P. R., Jackson, D. A., & Somers, K. M. (2005). How many principal components? Stopping rules for determining the number of non-trivial axes revisited. *Computational Statistics and Data Analysis*, *49*, 974–997.

Peterman, W. E., Connette, G. M., Semlitsch, R. D., & Eggert, L. S. (2014). Ecological resistance surfaces predict fine-scale genetic differentiation in a terrestrial woodland salamander. *Molecular Ecology*, *23*, 2402–2413.

Queller, D. C., & Goodnight, K. F. (1989). Estimating relatedness using genetic markers. *Evolution*, *43*, 258–275.

R Core Team. (2013). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing, ISBN 3-900051-07-0, URL http://www.R-project.org/

Ritland, K. (1996). Estimators for pairwise relatedness and individual inbreeding coefficients. *Genetical Research*, *67*, 175–185.

Rousset, F. (2000). Genetic differentiation between individuals. *Journal of Evolutionary Biology*, *13*, 58–62.

Segelbacher, G., Cushman, S. A., Epperson, B. K., Fortin, M. J., Francois, O., Hardy, O., … Manel, S. (2010). Applications of landscape genetics in conservation biology: Concepts and challenges. *Conservation Genetics*, *11*, 375–385.

Shirk, A. J., & Cushman, S. A. (2011). SGD: Software for estimating spatially explicit indices of genetic diversity. *Molecular Ecology Resources*, *11*, 922–934.

Shirk, A. J., & Cushman, S. A. (2014). Spatially-explicit estimation of Wright's neighborhood size in continuous populations. *Frontiers in Ecology and Evolution*, *2*, 1–12.

Shirk, A. J., Wallin, D. O., Cushman, S. A., Rice, C. G., & Warheit, K. I. (2010). Inferring landscape effects on gene flow: A new model selection framework. *Molecular Ecology*, *19*, 3603–3619.

Sommer, S., McDevitt, A. D., & Balkenhol, N. (2013). Landscape genetic approaches in conservation biology and management. *Conservation Genetics*, *14*, 249–251.

Spear, S. F., Balkenhol, N., Fortin, M. J., McRae, B. H., & Scribner, K. (2010). Use of resistance surfaces for landscape genetic studies: Considerations for parameterization and analysis. *Molecular Ecology*, *19*, 3576–3591.

van Etten, J. (2014). *R package* GDISTANCE*: Distances and routes on geographical grids*.

Vos, C. C., Berry, P., Opdam, P., Baveco, H., Nijhof, B., O'Hanley, J., … Kuipers, H. (2008). Adapting landscapes to climate change: Examples of climate-proof ecosystem networks and priority adaptation zones. *Journal of Applied Ecology*, *45*, 1722–1731.

Wang, J. (2002). An estimator for pairwise relatedness using molecular markers. *Genetics*, *160*, 1203–1215.

Wang, J. (2011). Unbiased relatedness estimation in structured populations. *Genetics*, *187*, 887–901.

Wasserman, T., Cushman, S. A., Shirk, A., Landguth, E., & Littell, J. (2012). Simulating the effects of climate change on population connectivity of American marten (Martes americana) in the northern Rocky Mountains, USA. *Landscape Ecology*, *27*, 211–215.

Wright, S. (1931). Evolution in Mendelian populations. *Genetics*, *16*, 97–159.

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

Zeller, K. A., Creech, T. G., Millette, K. L., Crowhurst, R. S., Long, R. A., Wagner, H. H., … Landguth, E. L. (2016). Using simulations to evaluate Mantel-based methods for assessing landscape resistance to gene flow. *Ecology and Evolution*, *6*, 4115–4128.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

How to cite this article: Shirk AJ, Landguth EL, Cushman SA. A comparison of individual-based genetic distance metrics for landscape genetics. *Mol Ecol Resour*. 2017;00:1–10. https://doi.org/10.1111/1755-0998.12684