JOURNAL OF Evolutionary Biology

doi: 10.1111/j.1420-9101.2012.02497.x

Machine learning identifies specific habitats associated with genetic connectivity in *Hyla squirella*

T. D. HETHER[1] & E. A. HOFFMAN

The Department of Biology, University of Central Florida, Orlando, FL, USA

Keywords:

amphibian;
clusters;
ensemble learning;
gene flow;
landscape genetics;
Random Forest;
southeastern USA.

Summary

The goal of this study was to identify and differentiate the influence of multiple habitat types that span a spectrum of suitability for *Hyla squirella*, a widespread frog species that occurs in a broad range of habitat types. We collected microsatellite data from 675 samples representing 20 localities from the southeastern USA and used machine-learning methodologies to identify significant habitat features associated with genetic structure. In simulation, we confirm that our machine-learning algorithm can successfully identify landscape features responsible for generating between-population genetic differentiation, suggesting that it can be a useful hypothesis-generating tool for landscape genetics. In our study system, we found that *H. squirella* were spatially structured and models including specific habitat types (i.e. upland oak forest and urbanization) consistently explained more variation in genetic distance (median $pR^2 = 47.78$) than spatial distance alone (median $pR^2 = 23.81$). Moreover, we estimate the relative importance that spatial distance, upland oak and urbanized habitat have in explaining genetic structure of *H. squirella*. We discuss how these habitat types may mechanistically facilitate dispersal in *H. squirella*. This study provides empirical support for the hypothesis that habitat-use can be an informative correlate of genetic differentiation, even for species that occur in a wide range of habitats.

Introduction

Understanding how genetic variation is partitioned among populations is of fundamental importance in evolutionary biology. One pattern that has emerged, isolation by distance (IBD; Wright 1943), is nearly ubiquitous in natural populations as spatial restrictions facilitate mating between more proximate individuals (Rousset, 2004). Depending upon factors such as an individual's intrinsic dispersal ability and physical dispersal barriers, however, the correlation of genetic distance and spatial distance can vary greatly. The unexplained variance in the genetic distance/spatial distance correlation (i.e. low Pearson correlation coefficients of IBD) has prompted researchers to identify additional factors that may explain patterns of genetic connectivity among populations (Jenkins *et al.*, 2010). Indeed, within the past decade, there have been a number of empirical studies that have shown that landscape features, such as roads (Balkenhol & Waits, 2009), anthropogenic modified habitat (Pavlacky *et al.*, 2009), gaps in habitat (Pierson *et al.*, 2010) and species-specific corridors (Banks *et al.*, 2005; Spear & Storfer, 2010), facilitate or inhibit genetic connectivity among natural populations (for review see Manel *et al.*, 2003; Storfer *et al.*, 2007; Holderegger & Wagner, 2008).

Genetic connectivity depends on landscape connectivity (Stevens *et al.*, 2006; Wang *et al.*, 2008). Landscape connectivity, or the functional relationship among habitat patches (With & Crist, 1995), can be broken down into its *structural connectivity* and *functional connectivity*

Correspondence: Eric Hoffman, The Department of Biology, University of Central Florida, 4000 Central Florida Blvd. Orlando, FL 32816, USA.
Tel.: +1 407 823 4007; fax: +1 407 823 5769;
e-mail: eric.hoffman@ucf.edu
[1]Present address: Bioinformatics and Computational Biology Program, Department of Biology, University of Idaho, Life Sciences South 441D, PO Box 443051, Moscow, ID 83844-3051, USA.

components (Stevens *et al.*, 2006; Baguette & Van Dyck, 2007). The former refers to the habitat abundance and configuration throughout the landscape and is often the landscape features directly quantified in landscape genetic studies (for review see Storfer *et al.*, 2007); the latter refers to habitat specificity and the intrinsic dispersal ability of the focal taxon (With & Crist, 1995). Clearly both of these components are important for effectively connecting or disconnecting populations throughout the landscape; however, as functional connectivity is largely species-specific, it can be more difficult to directly quantify its contribution to genetic connectivity. Indeed, recent landscape genetic studies have focussed on how focal species behaviour and life-history attributes link together aspects of species ecology to help explain genetic structure. For example, Sacks *et al.* (2008) examined population genetic structure in a wide-ranging, highly mobile generalist, *Canis latrans*. Individuals of this species display a behavioural phenomenon whereby preference exists for dispersing through areas that are similar to their natal habitat. Using autosomal loci (microsatellites), Y chromosome and mitochondrial sequence data, Sacks *et al.* (2008) found genetic structure concordant with general habitat subdivisions of the heterogeneous California Floristic Province – indicating that this 'natal-habitat-biased dispersal' behaviour, a component of functional connectivity, affects spatial genetic structure.

Theory predicts that another species-specific behavioural phenomenon, habitat affinity, has the potential to affect the distribution of organisms (With *et al.*, 1997), and empirical data suggest that a species' mobility is enhanced by the presence of preferred habitat types (reviewed in Stevens *et al.*, 2006). Thus, we would predict that the greater the percentage of the landscape between two populations that is perceived as 'suitable' the more dispersal will occur between those populations. Two comparative studies, each among closely related species, show the importance of habitat suitability in facilitating gene flow among populations. Vandergast *et al.* (2004) compared genetic structure for three spider species of the genus *Tetragnatha*, two forest specialists and one generalist, within and among fragmented forests of the Island of Hawaii. The matrix separating remnant forests consisted mainly of a recent (< 200 year old) lava flow. Data based on mtDNA and allozymes showed that restricted habitat (i.e. only forested areas) has resulted in significant genetic structure for the two habitat specialists, whereas no evidence of global genetic differentiation or IBD was present in the generalist. Similarly, Brouat *et al.* (2003) assessed levels of IBD at fine spatial scales (up to 13.6 km) for two carabid species (again, one was a forest specialist and one was a generalist). Evidence of IBD based on microsatellite data was identified for the specialist species, but the generalist species only displayed a weak pattern of IBD. It is important to note that the forest/nonforest distinction of habitat suitability made in both of the above examples provided clear interpretation of the results – the abundance of suitable habitat affects genetic connectivity – but the binary 'suitable/nonsuitable' distinction is unlikely to be generalized in every study system (Fahrig, 2003; Ouin *et al.*, 2004). For most species, it is more likely that a spectrum of suitability exists (Stewart *et al.*, 2010), especially when the species utilizes any of several habitats within a given landscape.

Study system

Hyla squirella, the squirrel tree frog, is one of the most abundant tree frogs found along the Atlantic and Gulf Coastal Plains of the United States, occurring from Virginia to eastern Texas and south to the Florida Keys (Lannoo, 2005). As a terrestrial frog, it possesses many of the attributes that make amphibians ideal study organisms for population genetic studies. For example, the species is philopatric, returning to wetlands to breed (Duellman & Trueb, 1986; Blaustein *et al.*, 1994, 2003; Funk *et al.*, 2005; Manier & Arnold, 2006; Giordano *et al.*, 2007). Carr (1940) described this species as showing little discrimination in terms of habitat selection. *Hyla squirella* occurs in a wide range of habitats: fields and urbanized areas (Deckert, 1915; Wright, 2002); swamps (Lannoo, 2005); pine and oak groves (Wright & Wright, 1995); open wooded areas (Carr, 1940; Wright & Wright, 1995); and almost anywhere adjacent to food, moisture and shelter (Conant & Collins, 1998). Breeding habitats include grassy, ephemeral pools free of predatory fish, such as roadside ditches (Wright & Wright, 1995; Babbitt & Tanner, 1997; Jensen *et al.*, 2008), or open canopy ponds (Binckley & Resetarits, 2007), such as those found in modified pasture habitat. These studies on *H. squirella* ecology and literature on anuran ecology and genetics enabled us to predict that seven general habitat types, derived from the 2001 Southeast Gap Analysis Project (SEGAP; Comer & Schulz, 2007), would influence gene flow among *H. squirella* populations (Table 1; detailed description of habitats can be found in Appendix S1).

The overarching goal of our study was to identify and differentiate the influence of multiple habitat types that span a spectrum of suitability for *H. squirella*. Our first aim was to test for spatial genetic structure among *H. squirella* populations by testing for genetic clustering and isolation by distance. Such results provided us with a baseline 'spatial genetic landscape' on which we overlaid habitat data. If genetic clusters were identified, we would conduct our landscape genetic analyses within a given cluster because we assume contemporary habitats (e.g. agriculture or urbanization) are likely too recent to create the observed genetic clusters (see Discussion). Next, we asked whether the inclusion of the habitat data (Table 1) increased the explanatory power of IBD. Because we have eight dependent variables, some of which may interact nonlinearly or be redundant with one another,

Table 1 Southeast regional GAP (SEGAP) data set-derived variables used to assess habitat permeability in this study. For each habitat, the name, abbreviation, brief description, ecological justification, general genetic prediction and reference(s) are given. A detailed list of these habitats can be found in Appendix S1.

Name	Abbreviation	Brief Description	Ecological Justification	Genetic Prediction	Reference
Urbanization land cover†	urban	Developed urbanized land of varying intensity	Houses and buildings provide various degrees of shelter	Increase connectivity	Reviewed in Storfer *et al.* (2007), Holderegger & Wagner (2008) and Balkenhol *et al.* (2009)
Silviculture†	sil	Forests established by planting and/or seeding in and can include dense forest canopy cover	Many pond-breeding amphibians require upland forested habitats for foraging and overwintering	Increase connectivity	Wright (2002) and Jensen *et al.* (2008)
Pastures and Crop land†	pas	Agricultural land for livestock grazing or the production of seed or hay crops	Pasture land often comprises a mosaic of ephemeral, open canopy ponds suitable for breeding *H. squirella*	Increase connectivity	Semlitsch (1998) and Babbitt *et al.* (2006)
Mesic flatwoods	flat	Forested systems characterized by Pinus spp. with frequent, low-intensity fires and subject to seasonally high water tables	Fire regime in this system concomitant with hydroperiod allows for relatively high occurrence of suitable breeding habitats	Increase connectivity	Babbitt & Tanner (1997, 2000), Babbitt *et al.* (2006) and Binckley & Resetarits (2007)
Swamp	swamp	Hardwood/deciduous canopy dominants and hydrology dominated by rainfall and sheetflow	Preferred habitat for *H. squirella*	Increase connectivity	Binckley & Resetarits (2007)
Upland oak hammock	oak	Upland oak–dominated habitat with infrequent fire frequency	Many pond-breeding amphibians require upland forested habitats for foraging and overwintering	Increase connectivity	Jensen *et al.* (2008)
Water and Floodplain forest	rff	Open water and forested systems associated with lotic environments	Flooding (from nearby rivers) is a key ecological factor in this system, which can increase the density of ponds containing predatory fish	Decrease connectivity	Semlitsch (1998) and Babbitt *et al.* (2006)

†Descriptions of nonecological systems provided by T. Earnhardt from the Biodiversity and Spatial Information Center at North Carolina State University.

we used machine-learning methodology adapted from the study conducted by Murphy *et al.* (2010) to identify features associated with our genetic data. If the full model explained more of the variation in pairwise genetic distance than the 'null' IBD-only model, we used a novel model selection technique to retain the fewest predictors that, with their inclusion, did not significantly differ from the full model in terms of model fit. We used the 'variable importance' measure generated from our machine-learning analysis to rank those features that best explain genetic distance among *H. squirella* populations. Under a spatially explicit, coalescent theory-based simulation framework, we confirmed that our model selection algorithm successfully identifies the landscape features that were mechanistically responsible for the observed patterns of genetic differentiation. We conclude by discussing how population connectivity in *H. squirella* matched our *a priori* predictions and how this study further advances the growing field of landscape genetics.

Materials and methods

Sampling of molecular data

We collected molecular data from 675 *H. squirella* representing 20 localities (e.g. breeding ponds) throughout Florida and southern Georgia (Appendix S2). Individual frogs were collected by hand from breeding choruses; the longest toe on the right hind leg was surgically removed with sterile scissors; tissue was placed in anhydrous calcium sulphate for preservation; and frogs were returned to their site of capture as per our University of Central Florida IACUC protocol (09-21W). Habitat types were derived from the 2001 SEGAP, which consists of ecological systems (natural or semi-natural), human-modified land (e.g. pastures and urbanized regions) and nonterrestrial land cover type (e.g. lakes) at a resolution of 30 m × 30 m. Ecological systems are defined as US National Vegetation Classification plant community associations that tend to co-occur in areas with similar

ecological dynamics (e.g. flooding, fire regime) and similar environmental settings (Comer & Schulz, 2007).

We extracted DNA using the standard phenol–chloroform method (Sambrook & Russel, 2001). Individuals were genotyped at nine microsatellite loci (*hsq103*, *hsq107*, *hsq111*, *hsq116*, *hsq126*, *hsq131*, *hsq133*, *hsq135* and *hsq136*) that were specifically developed for *H. squirella* (for reaction conditions see Abdoullaye *et al.*, 2010). Typical PCRs used 40 ng of DNA as template, 1× PCR buffer, 0.2 μM each dNTP, 0.2 units of *Taq* polymerase, between 1.56–1.85 mM $MgCl_2$, 0.1 μM forward primer, 0.4 μM reverse primer and 0.4 μM M13 tag. Each forward primer was given a M13 (-21) tail (5′-TGTAAACGACGGCCAGT-3′) for fluorescent labelling (Schuelke, 2000; Abdoullaye *et al.*, 2010). PCR amplicons were visualized on a 2% agarose gel to verify amplification, and fragments were analysed on a CEQ 8000 genetic analysis system and software (Beckman-Coulter, Fullerton, CA, USA). Genotypes were initially checked for high null allele frequencies (> 0.09), allelic dropout, and scoring errors with Micro-checker v 2.23 (Van Oosterhout *et al.*, 2004). We tested for significant deviations from Hardy–Weinberg and linkage equilibria (HWLE) using Fisher's exact test in the program Genepop v. 4.0.7 (Raymond & Rousset, 1995; Rousset, 2008). Markov chain parameters for all tests included a dememorization of 10 000 and 500 batches (10 000 iterations per batch). We accounted for multiple comparisons by applying a sequential Bonferroni correction (Rice, 1989). Pairwise genetic distance was estimated using D_{ps}′ (hereafter *D*ps; Bowcock *et al.*, 1994) in the program Microsat v 1.5b (Minch *et al.*, 1996).

Genetic clusters

Uncharacterized genetic clusters can be problematic for landscape genetic analyses because pairwise genetic distance should be higher for population pairs that occur between, rather than within, genetic clusters. Because we were interested in contemporary processes and patterns, we first tested for the occurrence of genetic clustering (using two methods) and then restricted our analyses to pairwise comparisons for populations within genetic clusters. First, we used the r package Geneland v 3.1.4 (Guillot *et al.*, 2005a,b; Guillot, 2008; Guillot *et al.*, 2008) to infer the number of genetic clusters and to obtain estimates of population membership in a geographical context. Geneland is useful in identifying areas of high landscape resistance or identifying discrete boundaries where gene flow is reduced (Guillot *et al.*, 2005b). In Geneland, we used the spatial model and assumed uncorrelated allele frequencies between subpopulations to estimate genetic clusters. We allowed the number of HWLE populations to be an unknown parameter and allowed for joint updates of population membership and allele frequencies (Guillot, 2008). We considered the minimum and maximum number of clusters, *K*, to range from 1 to 10. The maximum rate of the Poisson process was set to the number of individuals ($n = 676$); the maximum number of nuclei in the free Voronoi tessellation was set to three times the number of individuals as recommended by the program's authors. The number of MCMC iterations was 3×10^5 (recording every 50 iterations; post-process burnin = 2000 saved iterations). Unlike other genetic assignment tests to date, Geneland is unique in that it can explicitly account for the presence of null alleles. Therefore, for Geneland analyses, we included all nine microsatellite loci; for any locus-population combination that showed evidence of high null allele frequencies (see Results; null allele frequency > 0.09), we filtered out null alleles (i.e. set filter.NA = TRUE). We performed 10 independent runs using the above parameters and used the mean posterior density to choose the best run given the data. Second, we used the program Structure as an additional test for the occurrence of genetic clustering. For all 675 individuals, representing 20 collecting localities, we performed a short pilot run in Structure v 2.3.1 (Pritchard *et al.*, 2000) for each $K = 1$–20. Likelihood values for each *K* increased to a point then decreased noticeably after around $K = 10$ (data not shown). Therefore, we performed 10 independent runs for each $K = 1$–10 using the admixture model with correlated allele frequencies among subpopulations and allowed the degree of admixture, α, to be inferred from the data. We collected data for 5×10^5 iterations (allowing the first 2×10^5 iteration to be discarded as burnin). All other parameters were set to their default values. We inferred the number of true clusters using the Δ*K* criterion (Evanno *et al.*, 2005).

Random Forest and landscape genetic analyses

To identify potentially important habitats, we used the nonlinear classification and regression tree analysis Random Forest (RF; Breiman, 2001). RF has been used in a range of disciplines including bioinformatics (Bulger *et al.*, 2003), chemoinformatics (Svetnik *et al.*, 2003, 2005), ecology (Cutler *et al.*, 2007) and landscape ecology (Rustigian *et al.*, 2003) and has recently been introduced to landscape genetics (Murphy *et al.*, 2010). Briefly, RF builds upon a procedure whereby an ensemble of regression trees are grown, each on a bootstrap sample of the training data (Svetnik *et al.*, 2005). The predictions of each bootstrap tree are then averaged (for regression) to give a final prediction (Svetnik *et al.*, 2003). Regression trees notoriously have high variance; small changes in the data can result in different series of splits down the tree. Averaging the predictions of many bootstrapped trees reduces this variance (Hastie *et al.*, 2009). The RF procedure adds another layer of randomness to the regression tree building procedure to further reduce variance by reducing the correlation among trees; this reduction is accomplished by randomly selecting only a subset of the total predictors as candidates for node

splitting (Hastie *et al.*, 2009). RF is an appealing technique in landscape genetics as it has the ability to handle data sets with a relatively large number of predictors compared to number of observations. RF can also handle redundant and (or) irrelevant predictors, provide a type of cross-validation in parallel with the training step and provide variable importance that can be used along with partial dependence plots to aid in biological interpretation (Svetnik *et al.*, 2003; Laikre *et al.*, 2005; Cutler *et al.*, 2007; Hastie *et al.*, 2009).

We used RF to determine the relative importance of spatial distance and to identify the habitat types that may have contributed to observed genetic structure in this system. Here, our methodology was adapted from the study conducted by Murphy *et al.* (2010), with changes mentioned in the following paragraphs. This approach can be broken down into four general steps: (i) combine genetic and landscape data, (ii) run full RF model and calculate model improvement ratio (MIR) for each variable, (iii) perform model selection algorithm and (iv) run final RF for chosen submodel to obtain final variable importance, predicted response and overall model significance.

To combine genetic and landscape data, we first constructed a network of hypothetical (linear) pairwise corridors among populations within a genetic cluster. For each population pair, we combined pairwise estimates of genetic differentiation, pairwise spatial distance (km) and percentage of each habitat. The per cent of each habitat depended upon the 'corridor width', the width of the landscape that extends outwards from each corridor. We performed the RF methodology described below for varying corridor widths (diameter = 500 m, 2 and 10 km).

We used the R package randomForest v 4.5-28 (Liaw & Wiener, 2002) to run RF with all predictors (i.e. the 'full' model) in regression mode with 5000 trees. Measures of the importance of each predictor (I_p) are generated automatically in RF and were converted to model improvement ratios (MIRs) by dividing each I_p by the maximum I_p (MIR = I_p/I_{max}).

Our model selection criteria differed from the methodology of Murphy *et al.* (2010) in two ways. First, we created submodels via iteratively removing each predictor starting with the one with the lowest MIR, until only the predictor with the largest MIR was retained. Second, we ran 30 independent runs of regression trees (or 'forests') for each of these submodels and obtained 95% confidence intervals around their means. We considered two submodels to have significantly different pseudo-R^2 values (pR^2; see Liaw & Wiener, 2002) if they had nonoverlapping confidence intervals. We selected the submodel with the fewest retained predictors whose mean was not significantly different from that of the best fitting submodel. In this way, our chosen RF model was not significantly different than the submodel with the best fit, yet it contained fewer predictors. For the chosen RF model, we determined overall direction of each predictor while averaging out the effects of other predictors using partial dependence plots (Cutler *et al.*, 2007; Hastie *et al.*, 2009). Significance was estimated by randomizing the response of the chosen submodel (i.e. genetic distance) 9999 times, obtaining model fit (pR^2) for each iteration, and estimating the tail probability of the Monte Carlo null distribution ($\alpha = 0.05$) as in Murphy *et al.*'s (2010) study.

Lastly, we used two methods to compare competing models of population connectivity. First, we qualitatively compared confidence intervals of RF-derived model fit (pR^2) for the chosen landscape genetics model and for a model containing only spatial distance (i.e. the 'IBD-only' model) using the same RF procedure as above. Second, we used linear regression and an information-theoretic approach (Burnham & Anderson, 2002) to compare four models of genetic connectivity: (i) the IBD-only model, (ii) a landscape genetics model that contained predictors chosen using the above RF algorithm, (iii) a binary model with all habitats predicted to increase connectivity combined into one metric and (iv) a full additive model with spatial distance and all seven habitat types. We transformed all predictors to better meet parametric assumptions and used Akaike weights (w_i) to infer the best model from the candidate pool.

Simulation

We performed population genetic simulations to verify the efficacy of the above RF analysis and habitat selection algorithm. Specifically, in each of three hypothetical landscape scenarios, we tested whether our RF methodology could identify and differentiate between (i) habitat responsible for influencing variability in pairwise genetic differentiation and (ii) irrelevant predictors.

General simulation parameters

We ran all simulations using the computer program splatche2 (Ray *et al.*, 2010), a spatially explicit population genetic simulation framework based on coalescent theory. In all three scenarios, the landscape consisted of a 100×100 raster where each raster cell corresponded to a deme with a specific habitat code. Demographic expansion of the virtual world occurred in two steps. First, at the beginning of the simulation, a single founding population of 200 diploid simulated individuals – arbitrarily located at coordinates East 93, North 41 – was allowed to expand and colonize the virtual world. In this phase, all cells of the raster contained sufficient ecological resources to support demes of a carrying capacity k of 50 individuals each. The actual process of expansion was achieved by this carrying capacity as well as migration and population growth rates. For simplicity and brevity, we only consider per-generation emigration out of a given deme to occur in the four cardinal directions at a

rate of $m = 30\%$ of the total number of individuals in that deme; further, we assume population growth of each deme is logistically regulated ($r = 0.5$). Five hundred generations under this regime was more than enough to successfully colonize each landscape. The second phase of the demographic expansion followed and differed from the first by allowing different habitats to form. These habitats vary in their carrying capacity and differed among the landscape scenarios. Using the same demographic processes as in the first phase, the metapopulation was allowed to evolve for another 9500 generations. Following termination of this demographic simulation, 16 demes that contained 50 individuals were chosen at random and 500 unlinked microsatellite loci ($\mu = 5.0 \times 10^{-5}$) were simulated for each individual. As these microsatellite loci followed a strict stepwise mutation model, we calculated pairwise R_{ST} for each population pair in Arlequin 3.5 (Excoffier & Lischer, 2010). For each population pair, we calculated the per cent of each habitat type and the pairwise Euclidean distance connecting each pair. For simplicity, we assumed that the 'corridor width' connecting two populations was one cell wide. Below we provide specific details for each landscape scenario we investigated.

The Barrier landscape

In the first scenario (Fig. 1a), each cell was coded as a binary suitable/unsuitable metric. Here, 'suitable' habitat serves two functions: (i) it provides sufficient ecological resources to support $k = 50$ individuals and (ii) individuals can move about adjacent suitable habitats. In contrast, the carrying capacity of unsuitable habitat is zero and individuals cannot move across unsuitable habitat. In this scenario, a swath of unsuitable habitat bisects the landscape. Such habitat could be thought of as a river, a roadway, or any other linear patch of unsuitable habitat. Following completion of the two demographic phases, 16 demes were chosen at random from the suitable habitat. Pairwise R_{ST} data, combined with Euclidean distance and per cent suitable and unsuitable habitat were collected for RF analysis.

The Patch-Mosaic landscape

As the name implies, this landscape (Fig. 1b) is characterized by patches of sufficient resource to support $k = 50$ individuals per demes with a mosaic of habitats that lack said resources. Thus, similar to the Barrier landscape, the Patch-Mosaic can be distilled to a binary landscape; however, the former differs from the latter in the carrying capacity of the unsuitable/mosaic habitat. In the Barrier landscape, the unsuitable habitat eliminated any potential for gene flow between the populations spanning the barrier. In contrast, gene flow is possible through the mosaic: the carrying capacity of each mosaic habitat was set to 5; indeed, on average, the number of effective emigrants leaving a mosaic cell is $Nm = k*m = 1.5$. Therefore, in addition to demes within patches exchanging migrants, demes in disparate patches can achieve some level of gene flow. The 16 demes used to simulate genetic data were chosen randomly from 'patch' habitat. Pairwise R_{ST} data, combined with Euclidean distance and per cent patch and mosaic habitat were collected for analysis.

Multiple Habitat landscape

Rarely do natural systems exist in a binary landscape such as those outlined above. Rather, a spectrum of suitability habitats is available (Stewart *et al.*, 2010). We simulated such a scenario by building upon the Patch-Mosaic landscape above. Here, we increased the types of non-Patch habitat (Fig. 1c) by altering the carrying capacity of mosaic cells. In this landscape, some cells are conducive to higher migration (i.e. $k = 10$; $Nm = 3$), others have lower migration ($k = 5$; $Nm = 1.5$), whereas others have complete barrier to gene flow ($k = 0$; $Nm = 0$). To make a direct comparison between the

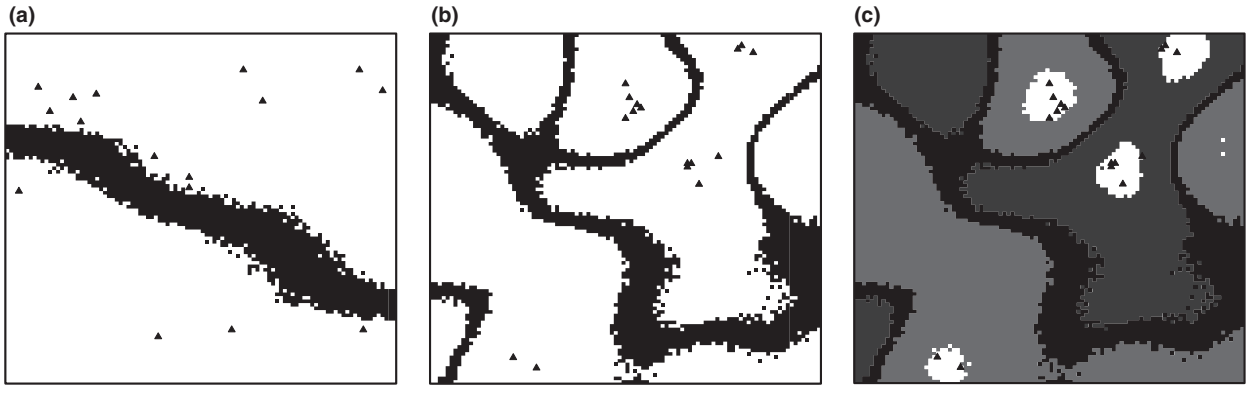

Fig. 1 Three landscape scenarios used in the simulation of genetic data. (a) Barrier (b) Patch-Mosaic (c) Multiple habitats. Shaded area corresponds to carrying capacity (lighter shading denotes higher carrying capacity). Demes in which genetic data were simulated are denoted by triangles.

Patch-Mosaic and the Multiple Habitat landscape scenarios, the same set of 16 demes were analysed. Once again, pairwise R_{ST}, Euclidean distance, and per cent of each habitat were collected for analysis.

RF analysis on simulated data sets

We were interested in testing the reliability of the above RF method in distinguishing signal from noise. Thus, in addition to running RF with the pattern-generating features (i.e. habitats types and spatial distance), which influence the variability of pairwise genetic differentiation, we added noisy predictors generated from the standard normal distribution. For each data set, the number of noisy predictors was such that the total number of predictors did not exceed 10. For each of the three data sets, we performed 1000 pseudoreplicates of the RF methodology described above and averaged the results for each landscape scenario. We considered the RF method reliable if it could identify the pattern-generated features responsible for variability in pairwise R_{ST} values.

Results

Overall, we found that *H. squirella* are spatially structured, with evidence of IBD and genetic clusters, that the RF-based model that included habitat types was better than the IBD-only model at explaining patterns of genetic connectivity, and that our RF analysis successfully identified those landscape features that were mechanistically responsible for variable genetic differentiation in simulation. As predicted, spatial distance was negatively associated with gene flow, whereas upland *oak* and *urban* habitats were positively associated with gene flow in *H. squirella*. These findings are detailed in the following paragraphs.

The spatial genetic landscape of *H. squirella*

We identified two loci, *hsq131* and *hsq136*, that consistently showed evidence of high-frequency null alleles and, with the exception of the Geneland analysis (see Materials and Methods), were removed from all downstream analyses. After applying a sequential Bonferroni correction for the remaining seven loci for all populations, five comparisons remained significantly out of HWE and no population had more than one locus out of HWE. Our data set consisted of a large range of expected heterozygosities [0.00–0.95; mean = 0.64 ± 0.33 standard deviation (SD)]. Allelic richness per locus, rarified to 14 diploid individuals per population, ranged from 1 to 15.2. Despite the large differences in these extreme values, the mean allelic richness (average ± SD across loci) for each population was similar throughout the study domain (7.4 ± 0.6).

The Geneland and Structure results were congruent with both analyses identifying the same two genetic clusters (Fig. 2). Isoclines in Fig. 2 show the posterior probabilities of genetic cluster membership for the 'eastern' cluster (given in Geneland). The western cluster consisted of only four sampling localities. The relatively low number of localities in this cluster made it impractical to obtain reliable estimates of landscape genetic patterns. Therefore, the western cluster was not used in any of the following landscape genetic analyses. In contrast, the eastern cluster consisted of 16 localities ranging throughout Florida's peninsula to southern Georgia (Fig. 2).

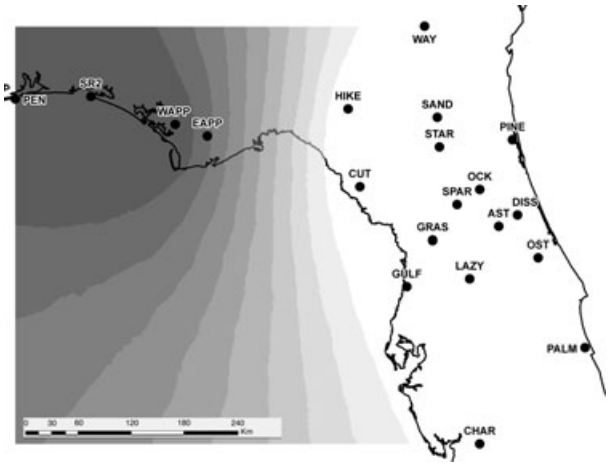

Fig. 2 Genetic clusters identified from both Geneland and Structure. Presented here are posterior probabilities (Geneland) of eastern cluster membership. Posterior probabilities of the western cluster are defined as one minus the posterior probability of the eastern cluster. Black dots represent populations sampled. Population abbreviations cross-reference with locality names and spatial reference (Appendix S2). Print edition of article: The posterior probabilities of belonging to the eastern cluster are denoted by the 10 levels of shading: darkest denotes the lowest probability (0–0.1) and lightest (white) the highest (0.9–1.0).

Landscape genetics analysis

Our landscape genetic analysis revealed a strong pattern of model improvement when habitat variables were incorporated into models of IBD. Models including habitat types consistently explained more of the variation in genetic distance (median $pR^2 = 47.78$) than spatial distance alone (median $pR^2 = 23.81$; Table 2). With regard to the three distinct corridor widths analysed, we found that for a given pairwise observation, metrics were highly correlated among different corridor widths (Appendix S3). Also, models with different corridor widths had similar prediction error (MSE), overlapping model fit (i.e. pR^2) confidence intervals and similar predictors retained. Therefore, we report only the data from the smallest corridor width (500 m).

Our RF-based model selection algorithm identified only two habitats, upland oak forest (*oak*) and urbanization

Table 2 Features associated with genetic connectivity among *Hyla squirella* populations using Random Forest. Presented here are the chosen predictors following model selection for the 500-m corridor width (detailed summary statistics for all corridor widths can be found in Appendix S2). pR^2 is a pseudo-R^2; MSE denotes mean-squared error. Summary statistics, based on constructing 30 forests for each submodel (see Materials and Methods), include median and 95% confidence intervals of pR^2. *P*-value of the chosen submodel is provided (see text). Model denotes the chosen variables (names cross-reference with Table 1) following model selection. These variables are ordered starting with the most important variable (in terms of MIR values) to the least important.

Model	Median pR^2	95% CI	Median MSE	*P*-value	Model
Spatial distance only	23.81	23.57–24.06	1.33E-01	< 0.001	km
Landscape genetics	47.78	47.53–48.16	9.12E-04	< 0.001	km, oak, urban

Table 3 Competing models [full, Random Forest (RF), binary and isolation by distance (IBD)] of population connectivity in *Hyla squirella*. k = number of parameters, AIC_c = AIC adjusted for small sample size, Δ_i = AIC_c differences from minimum AIC_c value, w_i = Akaike weights. km represents geographical distance, see Table 1 for definitions of other habitat variables.

Model: variables used	k	AIC_c	Δ_i	w_i	Adjusted R^2
Full: km + oak + urban + pas + sil + rff + swamp + flat	9	−455.22	0.00	0.56	0.5269
RF: km + oak + urban	4	−454.77	0.45	0.44	0.4846
Binary	4	−443.05	12.56	0.00	0.4395
IBD: km	2	−430.96	24.26	0.00	0.3639

(*urban*), that were significantly correlated with genetic distance. When including these two habitats (and spatial distance), the model error (MSE) decreased by approximately three orders of magnitude compared to the IBD-only model (Table 2). Similarly, the pR^2 for the chosen RF model was nearly twice that of the IBD-only model. Akaike weights, a measure of relative model performance, confirm that our RF-derived model was better than the either the IBD-only model or the binary model and was neither better nor worse than the full saturated model (but contained fewer parameters; Table 3).

Partial dependence plots (Fig. 3) display the individual variables included in the best supported landscape genetics model. These plots show the effect of a given predictor on the model after accounting for the average effects of the other retained predictors. From the plots and MIR values, the effect of spatial distance (km) is large compared to the next most important predictor – per cent upland oak – and showed the characteristically positive relationship with pairwise genetic distance (Fig. 3). On the other hand, the two retained habitats showed a negative relationship with genetic distance. Interestingly, the correlation between *D*ps and per cent urbanization appears to have a nonlinear relationship (Fig. 3), but this effect may be due to an outlier (or small sample sizes at higher urbanization densities). One data point consisted of over 37% urban habitat (more than three times the average value). Substituting this datum with the median

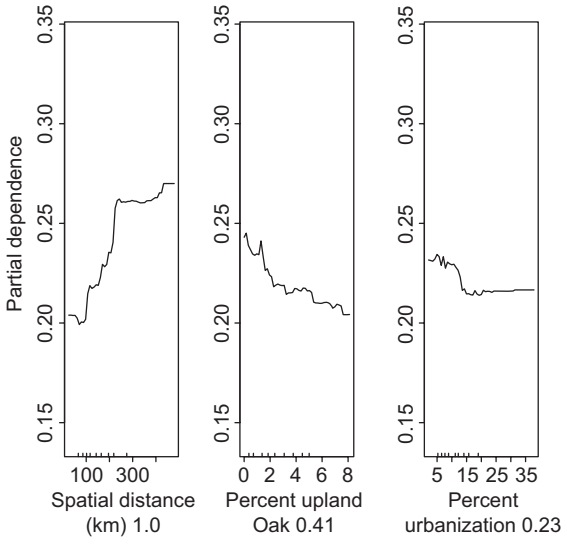

Fig. 3 Factors correlated with genetic differentiation in *Hyla squirella*. Graphs here are partial dependence plots for the chosen Random Forest model. Values following variables denote model improvement ratios (MIRs), that is, the importance of a given predictor relative to the most important predictor (far left; MIR = 1.0). These plots show the predictive function of (log-transformed) D_{ps} on a given predictor while accounting for the average effects of other retained predictors.

value and rerunning the above analyses yielded a more linear relationship without appreciably changing the above results (data not shown).

RF identifies pattern-generating landscape features

Finally, population genetic simulations verified the efficacy of the RF analysis and habitat selection algorithm used in this study (Table 4). Averaging over pseudoreplicates, the Barrier and Patch-Mosaic results were similar in terms of the amount of variation explained (mean pR^2 = 48.9 and 47.8, respectively) and mean pR^2 of the Multiple Habitat scenario was considerably higher, 89.6%. With the exception of the 'complete barrier' habitat in the Multiple Habitat scenario, in which 0.5% of the time RF excluded it in the final model, relevant predictors (i.e. spatial distance, per cent suitable, lower

Table 4 Summary of simulation results for each landscape scenario examined. For each landscape scenario below, the per cent variation explained (pR^2) is given as well as the percentage of time a given variable was included in the final model [SD = spatial distance; suitable = suitable habitat (see Materials and Methods); unsuitable/Mosaic = unsuitable habitat (Barrier scenario) or mosaic habitat (Patch-Mosaic scenario), N*m* = three habitats used in the Multiple Habitat scenario with varying effective migration rates]. Note that the results for 'Random' reflect the average (among the random predictors) percentage of time Random Forest retained a random predictor in the model selection algorithm.

Landscape scenario	pR^2	SD	Suitable	Unsuitable/Mosaic	N*m* = 0, 1.5, 3	Random
Barrier	48.9	100	100	100	–	3.8
Patch-Mosaic	47.8	100	100	100	–	13.8
Multiple Habitat	89.6	100	100	–	99.5, 100, 100	2.5

migration and higher migration habitat) were always found to be important explainers of variation in pairwise genetic differentiation (Table 4). Although the model selection algorithm shows both high sensitivity (100%, 100% and 99% for Barrier, Patch-Mosaic and Multiple Habitat, respectively) and high specificity (94.6%, 80.3% and 95% for Barrier, Patch-Mosaic and Multiple Habitat, respectively), RF was more likely to include irrelevant predictors in the final model than it was to exclude relevant predictors (i.e. sensitivity > specificity).

Discussion

We identified discrete habitat types that correlated with genetic differentiation, a proxy for gene flow (Wright 1943), for an abundant species that occupies multiple putatively suitable habitats. As such, this study provides empirical data to support the hypothesis that habitat-use can be an informative correlate of genetic differentiation, even for species that occur in a wide range of habitats (e.g. Sacks *et al.*, 2008; Munoz-Fuentes *et al.*, 2009).

Simulating landscape genetic variation

One of the major goals of landscape genetics is to identify landscape features that are associated with patterns of gene flow and population genetic connectivity (Storfer *et al.*, 2007). To examine whether RF could identify features responsible for affecting gene flow, we used a spatially explicit population genetic simulator, SPLATCHE2, to simulate dynamic two-phase demographic scenarios. These simulations served to test the effectiveness of the RF analysis implemented in this study to identify 'important' landscape genetic features. Importantly, we examined three landscapes with varying complexity and in each landscape, RF tended to discern signal from noise.

It is appropriate here to note that our methodology identified those habitats responsible for varying between-population genetic differentiation in a controlled, simulated system. Implicit assumptions used here, and indeed many LG analyses, are that the process-generating features (in this case habitat types) have had long enough time to influence gene flow patterns, that these features can be accurately measured at the time of genetic sampling, that the underlying processes are 'stationary' (Fortin & Dale, 2008) and that the molecular markers used provide sufficient resolution to detect genetic differentiation. Any of these assumptions, if violated, can lead to erroneous inferences. Furthermore, natural systems likely have several process-generating features acting across temporal scales (for review see Anderson *et al.*, 2010), a problem we did not encounter in simulation. Therefore, we stress that the LG methods, including RF, are correlative and can be used as a hypothesis-generating tool. We use such a framework here and suggest casual mechanisms to explain the observed genetic data in *H. squirella.*

The spatial genetic landscape

The 'spatial genetic landscape' for *H. squirella* comprised a mixture of expected and unexpected results. First, the southeastern United States is a region that has been thoroughly studied with regard to many taxa including amphibians (Avise, 2000; Soltis *et al.*, 2006). The extent of our study area spans east and west of the Apalachicola River, an area that demarcates a phylogeographic break for many taxa (Lemmon *et al.*, 2007; Pauly *et al.*, 2007; Degner *et al.*, 2010). Thus, we suspected that historic gene flow between populations that span the Apalachicola River may have been reduced if rivers served as dispersal barriers in *H. squirella*. Although this study identified two genetic clusters (Fig. 2) and these clusters split near the Apalachicola River, two populations found on opposing sides of the river (WAPP and EAPP; Appendix S2) group together into the western group. This inconsistency may be the result of recent admixture of the populations on either side of the river. While testing this river-barrier hypothesis is beyond the scope of the current study, it is worth noting that genetic clustering was identified in this general region, but the genetic pattern does not indicate a long-standing break occurring at the river as has been found for other amphibian species (Pauly *et al.*, 2007).

Second, owing to high abundance of individuals (large N_e, T.D.H., unpublished data) and the abundance of habitats upon which *H. squirella* occupy, we predicted

that spatial distance would have a minimal effect on genetic differentiation as genetic drift would be a weak force in differentiating populations and structural connectivity would be high. Unexpectedly, we found that spatial distance was the most predictive of the variables used in this study (i.e. highest model improvement ratio). This result differs from the findings of Murphy *et al.* (2010), which also used RF to assess amphibian population connectivity in a landscape genetics context. However, this discrepancy may reflect differences in spatial scale (extent) used in Murphy *et al.*'s (2010) study, and this study, as the former was considerably smaller, or it may suggest that different mechanisms are driving connectivity depending on species and location. Future studies with *H. squirella* conducted at a finer spatial scale and additional landscape genetic studies of amphibians in general will elucidate the generality of our results.

Habitats associated with genetic structuring in the squirrel tree frog

With regard to the second aim of this study, we found that incorporating habitat data increased the predictive power of our analyses (Tables 2 and 3). Our mean-squared error in the landscape genetics model – which included spatial distance, upland oak hammock, and urbanization – was roughly three orders of magnitude smaller than the model that excluded habitat data. Correspondingly, the (pseudo) per cent variation explained for the final landscape genetics model ($pR^2 = 47.72$) was twice that of the IBD-only model (Table 2). Our information-theoretic analysis provides similar results as our Random Forest analysis, although the magnitude of difference between the IBD only and final landscape genetic model is greater for the latter. Given that we know that incorporating landscape improves our estimate of connectivity, we can now address how these landscape features may have influenced the observed patterns of genetic structure.

While individuals of a species may utilize several habitats to varying degrees, some habitats may be of particular importance in connecting populations. *Hyla squirella* can be found in several habitats within its range but we found two, oak and urban areas, to be particular important in explaining population genetic connectivity. When comparing the model improvement ratios, we found that oak was roughly twice as important as urban areas (Fig. 3). The importance of upland oak habitat identified here is consistent with a recent mark–recapture study that examined landscape features correlated with *H. squirella* survival and recapture rates. Windes (2010) found that *H. squirella* display strong site fidelity (typical of many anurans), but recapture rates generally decreased with increased size of surrounding upland woodlot area. In addition, woodlot area was positively associated with *H. squirella* survival. Taken together, the mark–recapture (Windes, 2010) and landscape genetics research (this study) underscores the necessity of upland terrestrial habitats for *H. squirella*. In general, upland habitats are believed to be important for amphibian survival, and our data corroborate this supposition by providing evidence of gene flow through upland oak habitat.

Indeed, an appealing aspect of utilizing genetic data and landscape composition in concert is that one can test ecological expectations for the taxon of interest when detailed natural history information is unavailable. The approach itself can provide insight into which habitats are important for the focal taxon. For example, semi-aquatic amphibians spend some part of their life history in upland habitats surrounding breeding ponds, but the use of these habitats by amphibians remains poorly understood because reliable sampling in upland terrestrial environments is difficult (Windes, 2010). However, this study identified that upland oak habitat increases connectivity without extensive sampling of populations in said habitat. Moreover, because upland oak habitat is important for *H. squirella*, it is likely important in maintaining gene flow among populations of more specialized members of the cinerea clade (Dodd & Cade, 1998; Semlitsch, 1998; Bulger *et al.*, 2003; Semlitsch & Bodie, 2003; Trenham & Shaffer, 2005).

The influence of urbanization on survival has been previously investigated for a number of anuran species. Studies tend to show that increasing road density can act as a barrier to gene flow (e.g. *H. femoralis* and *H. gratiosa*, Jensen *et al.*, 2008) and has a negative impact on anuran density (Reh & Seitz, 1990; Fahrig *et al.*, 1995). However, with regard to *H. squirella*, we found that urbanized habitats did indeed correlate with decreased pairwise genetic distance (Fig. 3). This is likely because *H. squirella* readily find refuge in anthropogenic features such as houses (Wright, 2002; Jensen *et al.*, 2008) and other types of human-made habitat (Wright, 2002; Jensen *et al.*, 2008). Therefore, gene flow is likely enhanced through mildly urbanized areas because of the increased frequency of man-made temporary ponds and roadside ditches. *Hyla squirella* breed exclusively in temporary ponds (Boughton *et al.*, 2000; Windes, 2010), and numerous individuals can be found in breeding choruses along ephemeral roadside ditches. Owing to the Coastal Plains' low topographic relief, these roadside ditches are abundant and effectively 'connect' the region. Moreover, these temporary water bodies can be structurally complex with vegetation, which is known to increase *H. squirella* tadpole survival rates (Babbitt & Tanner, 2000). Lastly, oviposition site preference for *H. squirella* females is towards temporary ponds with open forest canopy cover (Babbitt & Tanner, 1997), which is characteristic of many roadside ditches. Indeed, the ability of anurans to utilize a broad range of habitat types may be what enables certain species to thrive in face of urbanization. The anuran assemblage in Florida's urban-rich regions is largely

represented by relatively generalist species such as *B. terrestris*, *Osteopilus septentrionalis*, *H. cinerea* and *H. squirella* (T.D.H. personal observation).

Knowledge of the types and abundance of useable habitats are critical factors in understanding functional and genetic connectivity of natural populations; however, in addition to multiple habitats of varying suitability occurring in the landscape, the ability for species to access multiple habitats may be required for population success. At the extreme end of the spectrum, *landscape complementation* indicates that two 'nonsubstitutable' resources, partitioned between two distinct habitats, may be required for a given population to persist (Binckley & Resetarits, 2007). In our study, we found evidence that *H. squirella* was able to use habitats cooperatively. Although we found that *H. squirella* tend to have high gene flow in urbanized habitats, caution should be taken when considering urban habitat, itself, as 'good' for *H. squirella*. This is because success in urbanized areas was likely tied to the availability of other habitats. We found that the relationship between genetic distance and urbanization depends upon the amount of upland habitat that is intermixed with urban habitat. Per cent urbanization only goes so far to reduce genetic distance between populations. For example, we found that for a given per cent urbanization, pairwise genetic distances were higher as per cent upland oak habitat approached zero (Fig. S1). Additionally, we did not adequately sample populations of *H. squirella* separated by high densities of urbanization. With regard to the intensity of urbanized landscape, we suspect that populations of *H. squirella* separated by lower and higher levels of urbanization will show the lowest effective migration (via fewer man-made temporary ponds and fewer green spaces, respectively), and populations separated by intermediate levels of urbanization (and consequently higher levels of habitat heterogeneity) show the highest effective gene flow (*sensu* Connell, 1978).

Conclusion

In natural populations, spatial genetic structure is the rule rather than the exception and we found evidence of this in *H. squirella*. However, the physical proximity of populations only weakly explained the pattern of genetic differentiation. Our study corroborates a growing body of literature that suggests landscape features influence rates of gene flow among populations (Jenkins *et al.*, 2010) and that incorporating these landscape variables into models of IBD increase the explanatory power of population connectivity. Gene flow depends upon both functional and structural components of landscape connectivity. We predicted that functional connectivity among *H. squirella* populations would be high due to the species' catholic habitat preferences. Despite the apparent use of several habitats, however, our Random Forest methodology identified a subset of those habitats as important in terms of genetic connectivity. We highlighted how these two habitats, upland oak hammock and urbanized areas, may mechanistically facilitate dispersal although we stress that future (experimental) work is needed to more fully understand the relative importance of each. Finally, our general methodology is applicable to a range of study systems as it can provide a 'first step' in identifying habitats associated with population connectivity when detailed natural history of the study organism is absent or when extensive sampling of individuals is lacking in cases of rare or endangered species.

Acknowledgments

We would like to thank those individuals that greatly contributed to this project. Rosanna Tursi, Sarah May, Juan Daza, Allyson Fenwick, Haakon Kalkvik, Genevieve Metzger, Greg Territo, Ocean Cohen, Erica Bree Rosenblum, Luke Harmon, Denim Jochimsen, Steve Spear, Jonathan Eastmen, Matt Pennell, Kayla Hardwick and Samuel Beam provided insightful comments to earlier versions of this manuscript. Andrew Storfer and Melanie Murphy from the Department of Zoology at Washington State University and Richard Cutler from the Department of Mathematics and Statistics at Utah State University kindly provided theoretical and statistical comments. Lisa McCauley, Kim Medley and James Angelo from the Department of Biology at the University of Central Florida and Todd Earnhardt from the Biodiversity and Spatial Information Center at North Carolina State University helped with GIS and provided anthropogenic habitat classification descriptions. We thank Genevieve Metzger, Christian Metzger, Allyson Fenwick, Will Fenwick, Greg Territo and Jessica Hightower for help collecting tissue samples. This project was supported by Sigma Xi Grants-in-Aid research #G200810150636 and the University of Central Florida.

References

Abdoullaye, D., Acevedo, I., Adebayo, A.A., Behrmann-Godel, J., Benjamin, R.C., Bock, D.G. *et al.* 2010. Permanent Genetic Resources added to Molecular Ecology Resources Database 1 August 2009–30 September 2009. *Mol. Ecol. Resour.* **10**: 232–236.

Anderson, C.D., Epperson, B.K., Fortin, M.J., Holderegger, R., James, P., Rosenberg, M.S. *et al.* 2010. Considering spatial and temporal scale in landscape'Äêgenetic studies of gene flow. *Mol. Ecol.* **19**: 3565–3575.

Avise, J.C. 2000. *Phylogeography: The History and Formation of Species*. Harvard University Press, Cambridge, MA.

Babbitt, K.J. & Tanner, G.W. 1997. Effects of cover and predator identity on predation of *Hyla squirella* tadpoles. *J. Herpetol.* **31**: 128–130.

Babbitt, K.J. & Tanner, G.W. 2000. Use of temporary wetlands by anurans in a hydrologically modified landscape. *Wetlands* **20**: 313–322.

Babbitt, K.J., Baber, M.J. & Brandt, L.A. 2006. The effect of woodland proximity and wetland characteristics on larval anuran assemblages in an agricultural landscape. *Can. J. Zool.-Rev. Can. Zool.* **84**: 510–519.

Baguette, M. & Van Dyck, H. 2007. Landscape connectivity and animal behavior: functional grain as a key determinant for dispersal. *Landscape Ecol.* **22**: 1117–1129.

Balkenhol, N. & Waits, L.P. 2009. Molecular road ecology: exploring the potential of genetics for investigating transportation impacts on wildlife. *Mol. Ecol.* **18**: 4151–4164.

Balkenhol, N., Gugerli, F., Cushman, S.A., Waits, L.P., Coulon, A., Arntzen, J.W. *et al.* 2009a. *Identifying Future Research Needs in Landscape Genetics: Where to From Here?* Springer, Landscape Ecology **24**: 455–463.

Banks, S.C., Lindenmayer, D.B., Ward, S.J. & Taylor, A.C. 2005. The effects of habitat fragmentation via forestry plantation establishment on spatial genotypic structure in the small marsupial carnivore, *Antechinus agilis*. *Mol. Ecol.* **14**: 1667–1680.

Binckley, C.A. & Resetarits, W.J. 2007. Effects of forest canopy on habitat selection in treefrogs and aquatic insects: implications for communities and metacommunities. *Oecologia* **153**: 951–958.

Blaustein, A.R., Wake, D.B. & Sousa, WP 1994. Amphibian declines: judging stability, persistence, and susceptibility of populations to local and global extinctions. *Conserv. Biol.* **8**: 60–71.

Blaustein, A.R., Romansic, J.M., Kiesecker, J.M. & Hatch, A.C. 2003. Ultraviolet radiation, toxic chemicals and amphibian population declines. *Divers. Distrib.* **9**: 123–140.

Boughton, R.G., Staiger, J. & Franz, R. 2000. Use of PVC pipe refugia as a sampling technique for hylid treefrogs. *Am. Midl. Nat.* **144**: 168–177.

Bowcock, A.M., Ruizlinares, A., Tomfohrde, J., Minch, E., Kidd, J.R. & Cavallisforza, L.L. 1994. High-resolution of human evolutionary trees with polymorphic microsatellites. *Nature* **368**: 455–457.

Breiman, L. 2001. Random Forests. *Mach. Learn.* **45**: 5–32.

Brouat, C., Sennedot, F., Audiot, P., Leblois, R. & Rasplus, J.Y. 2003. Fine-scale genetic structure of two carabid species with contrasted levels of habitat specialization. *Mol. Ecol.* **12**: 1731–1745.

Bulger, J.B., Scott, N.J. & Seymour, R.B. 2003. Terrestrial activity and conservation of adult California red-legged frogs Rana aurora draytonii in coastal forests and grasslands. *Biol. Conserv.* **110**: 85–95.

Burnham, K.P. & Anderson, D.R. 2002. *Model Selection and Multimodel Inference: A Practical Information-theoretic Approach*. Springer Verlag.

Carr, A.F. 1940. Dates of Frog Choruses in Florida. *Copeia* **1940**: 55.

Comer, P.J. & Schulz, K.A. 2007. Standardized ecological classification for mesoscale mapping in the southwestern United States. *Rangel. Ecol. Manage.* **60**: 324–335.

Conant, R. & Collins, J.T. 1998. *A Field Guide to Reptiles and Amphibians of Eastern and Central North America*, 3rd edn. Houghton Mifflin Company, New York.

Connell, J.H. 1978. Diversity in tropical rain forests and coral reefs. *Science* **199**: 1302.

Cutler, D.R., Edwards, T.C., Beard, K.H., Cutler, A. & Hess, K.T. 2007. Random Forests for classification in ecology. *Ecology* **88**: 2783–2792.

Deckert, R.F. 1915. Further Notes on the Salientia of Jacksonville, FLA. *Copeia* **18**: 3–5.

Degner, J.F., Silva, D.M., Hether, T.D., Daza, J.M. & Hoffman, E.A. 2010. Fat frogs, mobile genes: unexpected phylogeographic patterns for the ornate chorus frog (*Pseudacris ornata*). *Mol. Ecol.* **19**: 2501–2515.

Dodd, C.K. & Cade, B.S. 1998. Movement patterns and the conservation of amphibians breeding in small, temporary wetlands. *Conserv. Biol.* **12**: 331–339.

Duellman, W.E. & Trueb, L. 1986. *Biology of Amphibians*. The Johns Hopkins University Press, Baltimore.

Evanno, G., Regnaut, S. & Goudet, J. 2005. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Mol. Ecol.* **14**: 2611–2620.

Excoffier, L. & Lischer, H.E.L. 2010. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Mol. Ecol. Resour.* **10**: 564–567.

Fahrig, L. 2003. Effects of habitat fragmentation on biodiversity. *Annu. Rev. Ecol. Evol. Syst.* **34**: 487–515.

Fahrig, L., Pedlar, J.H., Pope, S.E., Taylor, P.D. & Wegner, J.F. 1995. Effect of road traffic on amphibian density. *Biol. Conserv.* **73**: 177–182.

Fortin, M.J. & Dale, M. 2008. *Spatial Analysis: A Guide for Ecologists*. Cambridge University Press, Cambridge, UK.

Funk, W.C., Greene, A.E., Corn, P.S. & Allendorf, F.W. 2005. High dispersal in a frog species suggests that it is vulnerable to habitat fragmentation. *Biol. Lett.* **1**: 13–16.

Giordano, A.R., Ridenhour, B.J. & Storfer, A. 2007. The influence of altitude and topography on genetic structure in the long-toed salamander (Abystoma macrodactulym). *Mol. Ecol.* **16**: 1625–1637.

Guillot, G. 2008. Inference of structure in subdivided populations at low levels of genetic differentiation-the correlated allele frequencies model revisited. *Bioinformatics* **24**: 2222–2228.

Guillot, G., Estoup, A., Mortier, F. & Cosson, J.F. 2005a. A spatial statistical model for landscape genetics. *Genetics* **170**: 1261–1280.

Guillot, G., Mortier, F. & Estuop, A. 2005b. Geneland: a computer package for landscape genetics. *Mol. Ecol.* **5**: 712–715.

Guillot, G., Santos, F. & Estoup, A. 2008. Analysing georeferenced population genetics data with Geneland: a new algorithm to deal with null alleles and a friendly graphical user interface. *Bioinformatics* **24**: 1406–1407.

Hastie, T., Tibshirani, R. & Friedman, J. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd edn. Springer, New York, NY.

Holderegger, R. & Wagner, H.H. 2008. Landscape genetics. *Bioscience* **58**: 199–207.

Jenkins, D.G., Carey, M., Czerniewska, J., Fletcher, J., Hether, T., Jones, A. *et al.* 2010. A meta analysis of isolation by distance: relic or reference standard for landscape genetics? *Ecography* **33**: 315–320.

Jensen, J.B., Camp, C.D., Gibbons, W. & Elliott, M., eds. 2008. *Amphibians and Reptiles of Georgia*. The University of Georgia Press, Athens, GA.

Laikre, L., Miller, L.M., Palme, A., Palm, S., Kapuscinski, A.R., Thoresson, G. *et al.* 2005. Spatial genetic structure of northern pike (Esox lucius) in the Baltic Sea. *Mol. Ecol.* **14**: 1955–1964.

Lannoo, M., ed. 2005. *Amphibian Declines: The Conservation Status of United States Species*. University of California Press, Berkeley and Los Angeles, CA.

Lemmon, E.M., Lemmon, A.R. & Cannatella, D.C. 2007. Geological and climatic forces driving speciation in the continentally distributed trilling chorus frogs (pseudacris). *Evolution* **61**: 2086–2103.

Liaw, A. & Wiener, M. 2002. Classification and Regression by randomForest. *R News* **2**: 18–22.

Manel, S., Schwartz, M.K., Luikart, G. & Taberlet, P. 2003. Landscape genetics: combining landscape ecology and population genetics. *Trends Ecol. Evol.* **18**: 189–197.

Manier, M.K. & Arnold, S.J. 2006. Ecological correlates of population genetic structure: a comparative approach using a vertebrate metacommunity. *Proc. R. Soc. Lond. B. Biol. Sci.* **273**: 3001–3009.

Minch, E., Ruiz-Linares, A., Goldstein, D., Feldman, M. & Cavalli-Sforza, L.L. 1996. *MICROSAT (Version 1.5b): A Computer Program for Calculating Various Statistics on Microsatellite Allele Data*. program available from: http://hpgl.stanford.edu/projects/microsat/.

Munoz-Fuentes, V., Darimont, C.T., Wayne, R.K., Paquet, P.C. & Leonard, J.A. 2009. Ecological factors drive differentiation in wolves from British Columbia. *J. Biogeogr.* **36**: 1516–1531.

Murphy, M.A., Evans, J.S. & Storfer, A. 2010. Quantifying Bufo boreas connectivity in Yellowstone National Park with landscape genetics. *Ecology* **91**: 252–261.

Ouin, A., Aviron, S., Dover, J. & Burel, F. 2004. Complementation/supplementation of resources for butterflies in agricultural landscapes. *Agric. Ecosyst. Environ.* **103**: 473–479.

Pauly, G.B., Piskurek, O. & Shaffer, H.B. 2007. Phylogeographic concordance in the southeastern United States: the flatwoods salamander, Ambystoma cingulatum, as a test case. *Mol. Ecol.* **16**: 415–429.

Pavlacky, D.C., Goldizen, A.W., Prentis, P.J., Nicholls, J.A. & Lowe, A.J. 2009. A landscape genetics approach for quantifying the relative influence of historic and contemporary habitat heterogeneity on the genetic connectivity of a rainforest bird. *Mol. Ecol.* **18**: 2945–2960.

Pierson, J.C., Allendorf, F.W., Saab, V., Drapeau, P. & Schwartz, M.K. 2010. Do male and female black-backed woodpeckers respond differently to gaps in habitat? *Evol. Appl.* **3**: 263–278.

Pritchard, J.K., Stephens, M. & Donnelly, P. 2000. Inference of population structure using multilocus genotype data. *Genetics* **155**: 945–959.

Ray, N., Currat, M., Foll, M. & Excoffier, L. 2010. SPLATCHE2: a spatially explicit simulation framework for complex demography, genetic admixture and recombination. *Bioinformatics* **26**: 2993.

Raymond, M. & Rousset, F. 1995. GENEPOP (Version-1.2) – population-genetics software for exact tests and ecumenicism. *J. Hered.* **86**: 248–249.

Reh, W. & Seitz, A. 1990. The influence of land use on the genetic structure of populations of the common frog Rana temporaria. *Biol. Conserv.* **54**: 239–249.

Rice, W.R. 1989. Analyzing tables of statistical tests. *Evolution* **43**: 223–225.

Rousset, F. 2004. *Genetic Structure and Selection in Subdivided Populations*. Princeton University Press, Princeton, NJ.

Rousset, F. 2008. GENEPOP '007: a complete re-implementation of the GENEPOP software for Windows and Linux. *Mol. Ecol. Resour.* **8**: 103–106.

Rustigian, H.L., Santelmann, M.V. & Schumaker, N.H. 2003. Assessing the potential impacts of alternative landscape designs on amphibian population dynamics. *Landscape Ecol.* **18**: 65–81.

Sacks, B.N., Bannasch, D.L., Chomel, B.B. & Ernest, H.B. 2008. Coyotes demonstrate how habitat specialization by individuals of a generalist species can diversify populations in a heterogeneous ecoregion. *Mol. Biol. Evol.* **25**: 1384–1394.

Sambrook, J. & Russel, D. 2001. *Molecular Cloning: A Laboratory Manual*, 3rd edn. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.

Schuelke, M. 2000. An economic method for the fluorescent labeling of PCR fragments. *Nat. Biotechnol.* **18**: 233–234.

Semlitsch, R.D. 1998. Biological delineation of terrestrial buffer zones for pond-breeding salamanders. *Conserv. Biol.* **12**: 1113–1119.

Semlitsch, R.D. & Bodie, J.R. 2003. Biological criteria for buffer zones around wetlands and riparian habitats for amphibians and reptiles. *Conserv. Biol.* **17**: 1219–1228.

Soltis, D.E., Morris, A.B., McLachlan, J.S., Manos, P.S. & Soltis, P.S. 2006. Comparative phylogeography of unglaciated eastern North America. *Mol. Ecol.* **15**: 4261–4293.

Spear, S.F. & Storfer, A. 2010. Anthropogenic and natural disturbance lead to differing patterns of gene flow in the Rocky Mountain tailed frog, Ascaphus montanus. *Biol. Conserv.* **143**: 778–786.

Stevens, V.M., Verkenne, C., Vandewoestijne, S., Wesselingh, R.A. & Baguette, M. 2006. Gene flow and functional connectivity in the natterjack toad. *Mol. Ecol.* **15**: 2333–2344.

Stewart, A., Komers, P.E. & Bender, D.J. 2010. Assessing landscape relationships for habitat generalists. *Ecoscience* **17**: 28–36.

Storfer, A., Murphy, M.A., Evans, J.S., Goldberg, C.S., Robinson, S., Spear, S.F. *et al.* 2007. Putting the 'landscape' in landscape genetics. *Heredity* **98**: 128–142.

Svetnik, V., Liaw, A., Tong, C., Culberson, J.C., Sheridan, R.P. & Feuston, B.P. 2003. Random Forest: a classification and regression tool for compound classification and QSAR modeling. *J. Chem. Inf. Comput. Sci.* **43**: 1947–1958.

Svetnik, V., Wang, T., Tong, C., Liaw, A., Sheridan, R.P. & Song, Q.H. 2005. Boosting: an ensemble learning tool for compound classification and QSAR modeling. *J. Chem. Inf. Model.* **45**: 786–799.

Trenham, P.C. & Shaffer, H.B. 2005. Amphibian upland habitat use and its consequences for population viability. *Ecol. Appl.* **15**: 1158–1168.

Van Oosterhout, C., Hutchinson, W.F., Wills, D.P.M. & Shipley, P. 2004. MICRO-CHECKER: software for identifying and correcting genotyping errors in microsatellite data. *Mol. Ecol. Notes* **4**: 535–538.

Vandergast, A.G., Gillespie, R.G. & Roderick, G.K. 2004. Influence of volcanic activity on the population genetic structure of Hawaiian Tetragnatha spiders: fragmentation, rapid population growth and the potential for accelerated evolution. *Mol. Ecol.* **13**: 1729–1743.

Wang, Y.H., Yang, K.C., Bridgman, C.L. & Lin, L.K. 2008. Habitat suitability modelling to correlate gene flow with landscape connectivity. *Landscape Ecol.* **23**: 989–1000.

Windes, K. 2010. Treefrog (*Hyla squirella*) responses to rangeland management in semi-tropical Florida, USA. *Biology*. The University of Central Florida, Orlando. MS, pp. 110.

With, K.A. & Crist, T.O. 1995. Critical Thresholds in Species Responses to Landscape Structure. *Ecology* **76**: 2446–2459.

With, K.A., Gardner, R.H. & Turner, M.G. 1997. Landscape connectivity and population distributions in heterogeneous environments. *Oikos* **78**: 151–169.
Wright, A.H. 2002. *Life-Histories of the Frogs of Okefinokee Swamp, Georgia*. Cornell University Press, Ithaca, NY.
Wright, A.H. & Wright, A.A. 1995. *Handbook of Frogs and Toads of the United States and Canada*. Comstock Publishing Company, Inc, Ithaca, NY.
Wright, S. 1943. Isolation by Distance. *Genetics* **28**: 139–156.

Supporting information

Additional Supporting Information may be found in the online version of this article:
Appendix S1 Detailed list of predictors used to assess habitat permeability in this study.
Appendix S2 Spatial reference of sampled localities used in this study.
Appendix S3 Linear correlation coefficients among habitat types and corridor widths.
Figure S1 Bivariate partial dependence plot for *Hyla squirella* genetic differentiation.

As a service to our authors and readers, this journal provides supporting information supplied by the authors. Such materials are peer-reviewed and may be re-organized for online delivery, but are not copy-edited or typeset. Technical support issues arising from supporting information (other than missing files) should be addressed to the authors.

Received 22 August 2011; revised 20 February 2012; accepted 26 February 2012