PROCEEDINGS B

rspb.royalsocietypublishing.org

Research

Cite this article: Espíndola A, Ruffley M, Smith ML, Carstens BC, Tank DC, Sullivan J. 2016 Identifying cryptic diversity with predictive phylogeography. *Proc. R. Soc. B* **283**: 20161529.
http://dx.doi.org/10.1098/rspb.2016.1529

Received: 7 July 2016
Accepted: 27 September 2016

Subject Areas:
ecology, evolution

Keywords:
cryptic diversity, lineage discovery, Pacific Northwest rainforest, predictive phylogeography, random forest

Author for correspondence:
Anahí Espíndola
e-mail: anahi.espindola@gmail.com

Electronic supplementary material is available online at https://dx.doi.org/10.6084/m9.figshare.c.3512460.

THE ROYAL SOCIETY PUBLISHING

Identifying cryptic diversity with predictive phylogeography

Anahí Espíndola[1,2], Megan Ruffley[1,2], Megan L. Smith[3], Bryan C. Carstens[3], David C. Tank[1,2] and Jack Sullivan[1,2]

[1]Department of Biological Sciences, University of Idaho, 875 Perimeter Drive MS 3051, Moscow, ID 83844-3051, USA
[2]Biological Sciences, Institute for Bioinformatics and Evolutionary Studies (IBEST), 875 Perimeter Drive MS 3051, Moscow, ID 83844-3051, USA
[3]Department of Evolution, Ecology, and Organismal Biology, The Ohio State University, 318 W. 12th Avenue, 300 Aronoff Labs, Columbus, OH 43210-1293, USA

AE, 0000-0001-9128-8836

Identifying units of biological diversity is a major goal of organismal biology. An increasing literature has focused on the importance of cryptic diversity, defined as the presence of deeply diverged lineages within a single species. While most discoveries of cryptic lineages proceed on a taxon-by-taxon basis, rapid assessments of biodiversity are needed to inform conservation policy and decision-making. Here, we introduce a predictive framework for phylogeography that allows rapidly identifying cryptic diversity. Our approach proceeds by collecting environmental, taxonomic and genetic data from codistributed taxa with known phylogeographic histories. We define these taxa as a reference set, and categorize them as either harbouring or lacking cryptic diversity. We then build a random forest classifier that allows us to predict which other taxa endemic to the same biome are likely to contain cryptic diversity. We apply this framework to data from two sets of disjunct ecosystems known to harbour taxa with cryptic diversity: the mesic temperate forests of the Pacific Northwest of North America and the arid lands of Southwestern North America. The predictive approach presented here is accurate, with prediction accuracies placed between 65% and 98.79% depending of the ecosystem. This seems to indicate that our method can be successfully used to address ecosystem-level questions about cryptic diversity. Further, our application for the prediction of the cryptic/non-cryptic nature of unknown species is easily applicable and provides results that agree with recent discoveries from those systems. Our results demonstrate that the transition of phylogeography from a descriptive to a predictive discipline is possible and effective.

1. Background

Delimiting and identifying independent lineages is critical not only for taxonomy, but also for understanding the processes leading to the diversification of life, defining conservation strategies, and communicating among scientific and non-scientific communities [1]. The discovery of biodiversity can be impeded by the presence of cryptic diversity [2], where deep genetic divergence within a nominal species is present but not accompanied by known fixed morphological differences between sets of populations. Several factors have stimulated a growing literature that discusses the importance of such cryptic diversity. First, the prioritization of conservation efforts on a regional scale is often determined by species richness and endemism, and the identification of cryptic diversity is critical to estimating these parameters [3,4]. Second, the discovery of cryptic diversity establishes evolutionarily significant units for conservation [5,6]. Third, the discovery of cryptic diversity is critical to understanding anthropogenic biotic changes, invasions and ecosystem health [7]. Thus, cryptic diversity is a vital component of biological diversity [8,9], and

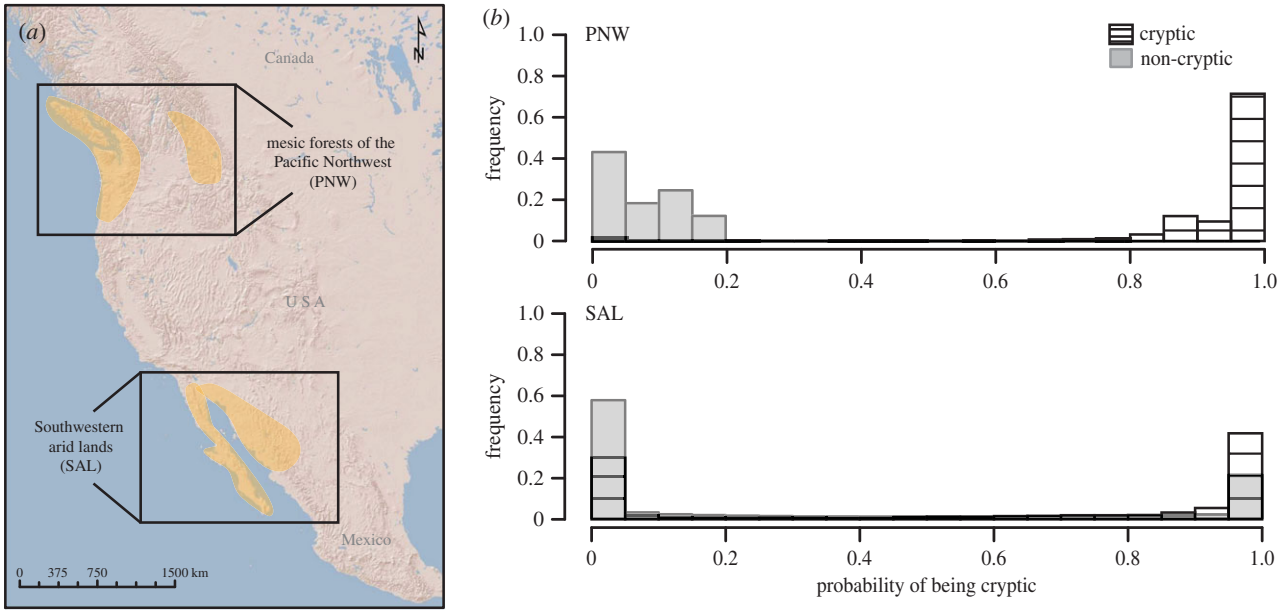

Figure 1. (*a*) Range of the biomes investigated in this study. Marked areas indicate extent of the biome. (*b*) Distribution of probability of harbouring cryptic diversity in the two datasets, and using the jackknife downsampling approach. Grey, non-cryptic taxa; dashed, cryptic taxa. PNW, Pacific Northwest; SAL, Southwestern arid lands. (Online version in colour.)

its discovery is central to documenting fundamental units in ecology [10], evolution and conservation [2].

Several novel methods have recently been introduced to assess the presence of independent lineages within taxa that were originally described as a single species [11–14]. While the assumptions and methodological details of these methods vary [15], each conducts lineage delimitation on a species-by-species basis and thus requires a detailed phylogeographic investigation into each taxon. Such investigations are time-consuming and expensive, and these costs limit the number of systems that can be investigated. Given the myriad threats to the Earth's biodiversity [16–19] and the difficulties inherent to the discovery of cryptic biodiversity [20], biodiversity discovery should no longer proceed on a taxon-by-taxon basis, but rather, rapid discovery and characterization of biodiversity is needed [21].

The transition of phylogeography from a largely descriptive to a predictive discipline will facilitate this goal. Here, we introduce a novel predictive framework that uses existing genetic data from previous phylogeographic studies (see [22]), open-access taxonomic and species occurrence data (e.g. GBIF) and climatic data (e.g. WorldClim). Cryptic diversity may be predictable from analysis of environmental data [8], and comparative phylogeographic research has demonstrated broadly congruent patterns of cryptic diversity in multiple species occupying the same fragmented biome [22–24]. In such conditions, cryptic divergence may be driven by random processes (e.g. genetic drift, mutations) or by divergent selection [2,25]. Longer periods of isolation are expected to increase ecological differentiation/divergence [26,27], and this has been demonstrated in several taxa [28,29].

The predictive approach to phylogeography proposed here uses random forest (RF) classifications, and integrates phylogeographic data with species occurrences, taxonomy and climatic data to define a predictive classifier for the rapid assessment of cryptic diversity. Our approach requires (i) the identification of two sets of taxa on which to adjust and train the classification: a set that contains cryptic diversity, and one that lacks it, (ii) a database of georeferenced occurrences and (iii) ecological and taxonomic data associated with the localities of species occurrences. Disjunct biomes in which several cryptic and non-cryptic species have been identified on either side of an isolating barrier are excellent models to develop and test this RF predictive approach.

2. Material and methods

(a) Two disjunct biomes as models

(i) The mesic forests of the Pacific Northwest of North America

The Pacific Northwest of North America (PNW) supports the world's greatest extent of temperate coniferous rainforests [30,31] that extends between 40° N and 53° N latitude along the Pacific coast and the inland Northern Rocky Mountains (figure 1*a*). This ecosystem is rich in endemic species, and includes at least 150 plant, animal and fungal species that are disjunct with over 300 km of arid Columbia Basin shrub-steppe between the coastal and inland portions of its distribution [30–32]. Because of the degree of isolation between these disjunct regions, conspecific populations from coastal and inland habitats have received much attention [33,34], with studies demonstrating that the system harbours substantial cryptic diversity (e.g. [35], reviewed in [36]).

Numerous hypotheses have been proposed to explain the origin of the disjunction (summarized by Brunsfeld *et al.* [37]). They posit either persistence of inland rainforests throughout the Pleistocene [30] or post-Pleistocene dispersal to the inland rainforests. The first hypothesis predicts high cryptic diversity, because of the hypothesized old age of the disjunction [37], and indeed, phylogeographic investigations have revealed cryptic diversity in taxa that were originally described as single, disjunct species (e.g. [33,38]; see electronic supplementary material, table S1). The alternative hypotheses deny the persistence of inland Pleistocene refugia, instead positing the inland dispersal of rainforest taxa after the Pleistocene (electronic supplementary material, table S1). Because the latter invoke the recent (post-Pleistocene) establishment of inland taxa, these recent dispersal models predict a lack of cryptic diversity, which has been shown in some studied taxa [33].

(ii) Arid lands of Southwestern North America

The arid lands of Southwestern North America (SAL) include a series of xeric areas that extend from the Southwestern United States to northern Mexico [39], and contain the Sonoran, Chihuahuan and the Baja California deserts. Although the area displays strong endemism, the Baja California–Sonoran areas share many species that are separated by the Colorado River and Gulf of California ([40,41]; figure 1*a*). Geological data indicate that the Baja Californian desert became isolated from the Sonoran desert after the separation of the Baja Peninsula from the mainland Sonoran region. Currently, the regions are isolated by the Gulf of California, an oceanic incursion of around 300 km [40].

Several hypotheses have been proposed to explain the similar composition of the two xeric areas. The first hypothesizes that peninsular populations became isolated from continental populations following the formation of the Gulf of California around 5 Ma [40]. Consistent with this, phylogeographic analyses of several disjunct species have shown strong divergence between disjunct populations (reviewed by [42]; electronic supplementary material, table S1), with cryptic species identified on each side of the Gulf. Alternatively, it has been proposed that some disjunct species were isolated in the Baja Californian deserts and have only recently colonized the Sonoran region; this hypothesis is consistent with the observation of several xeric-adapted species that do not show genetic differentiation between the two areas, and thus do not display cryptic diversity (e.g. [40,43]; electronic supplementary material, table S1 [44]).

(b) Occurrence datasets

For the PNW, we used several published datasets ([15,33,38, 45–47]; electronic supplementary material, table S1). They included tailed frogs (*Ascaphus*), Pacific giant salamanders (*Dicamptodon*), the Van Dyke's salamander complex (*Plethodon vandykei* and *P. idahonesis*), water voles (*Microtus richardsoni*), dusky willows (*Salix melanopsis*) and the blue-grey taildropper slug (*Prophysaon coeruleum*). To these datasets, we also added newly generated data for *Chonaphe armata*, a polydesmid millipede (GenBank accessions KX904729 - KX904806).

For the SAL, we focused on the compilation of 14 bird, mammal and amphibian taxa discussed by Zink [44]. Although not exhaustive for this system, this list (electronic supplementary material, table S1) is sufficient to demonstrate the general applicability of our predictive approach, and has two other salient features. First, approximately half of the taxa show cryptic diversity across the Colorado River and half do not (electronic supplementary material, table S1). Second, as in the PNW system, some taxa harbouring cryptic diversity have been elevated to species status based on the results of phylogeographic studies (e.g. *Peromyscus fraterculus*; [48]).

For each taxon, we compiled occurrence localities from GBIF, the primary literature and individual natural history collections (i.e. private and museum collections). We gathered 8228 observed localities of taxa from the PNW (average of 1175 localities per species). We gathered 487 735 localities from the SAL (average of 34 744 localities per species). Prior to further use, data were curated, with conspecific repeated, non-georeferenced localities or observations that fell clearly out of the range of the species excluded from the dataset.

(c) Taxa categories

We characterized each taxon as containing cryptic diversity (i.e. cryptic) or lacking cryptic diversity (i.e. non-cryptic) using two analytical approaches. First, we calculated the posterior probabilities of explicit phylogeographic models using an approximate Bayesian computation (ABC) approach. By doing this, we identified the most probable phylogeographic scenario given the data. Specifically, we evaluated three migration models (electronic supplementary material, figure S1), which represent different recent dispersal scenarios. The first two migration models consisted of post-Pleistocene divergence with subsequent gene flow either from east to west or from west to east. The third migration model consisted of pre-Pleistocene divergence with subsequent gene flow in both directions, and approximates a scenario in which there was divergence in the Pliocene followed by secondary contact. Finally, we compared the best of the migration models with a model of pre-Pleistocene divergence with no subsequent gene flow (ancient vicariance, AV; electronic supplementary material, table S1). We used simulations to determine the rejection method, and a combination of summary statistics that resulted in the correct model being selected consistently for the ABC analyses (see electronic supplementary material, S1 for more details). Data were simulated in ms [49] with 100 001 draws from the prior to match the actual data for each species under consideration, under the three migration models. Summary statistics for the observed data were calculated in DNAsp v. 5.1.0 [50]. A simple rejection step with a tolerance of 0.01 and the summary statistics π, Tajima's D, π within each population, and π between populations were used to approximate the posterior probabilities of the models.

The second approach used Bayesian molecular clock analyses to identify divergence times for the deepest nodes that span the disjunction. We assumed that relatively old divergences coupled with reciprocally monophyletic populations structured by geographical areas—i.e. inland versus coastal for the PNW; Baja California versus Mexico and the Sonoran Desert for the SAL—support the AV scenario. Alternatively, results pointing to relatively young divergence events—i.e. dispersal events—and non-reciprocally monophyletic populations suggest continuous gene flow and, therefore, recent dispersal. We used BEAST v. 1.8.2 [51] to infer topology and divergence times with mitochondrial or chloroplast sequence data for SAL and PNW taxa. A model of sequence evolution for each taxon was selected using DT-ModSel [52] (electronic supplementary material, table S5 and see electronic supplementary material, S1 for more details). The Markov chain Monte Carlo analysis was run for 100 million generations, sampling every 1000 generations, with a random starting tree and strict molecular clock. Convergence of all chains was visually assessed using TRACER v. 1.6 [53] and assumed to reach stationarity when effective sample size (ESS) values for all parameters were more than 200 (except for *Peromyscus*; ESS > 30). We used TreeAnnotator [51] to discard 'burn-in' states and summarize all remaining sampled trees (90 001 trees).

(d) Predictor variables

Because climate has been shown to be a good proxy for ecological preferences [54], for each locality, we extracted climatic data using the bioclimatic variables available in WorldClim [55], at a resolution of 30 arc-seconds (approx. 1 km^2). Owing to the potential for correlation among these climatic variables, we selected a subset of variables with low correlation ($r^2 < 0.7$). Thus, we conducted our analyses with eight climatic variables in the PNW dataset (annual mean temperature, bio1; mean diurnal range, bio2; isothermality, bio3; maximum temperature of warmest month, bio5; temperature annual range, bio7; annual precipitation, bio12; precipitation seasonality, bio15; and precipitation of driest quarter, bio17) and 10 in the SAL dataset (bio1; bio2; bio3; bio7; mean temperature of wettest quarter, bio8; mean temperature of driest quarter, bio9; bio12; precipitation of driest month, bio14; bio15; and precipitation of warmest quarter, bio18). These data manipulation steps were conducted in R using several functions from the packages *rgdal* [56], *dismo* [57], *biomod2* [58] and *adehabitat* [59]. Along with climatic variables, we also used major taxonomic ranks (e.g. classes, phyla; electronic

supplementary material, table S1) as input in our classification methodologies. For this, we associated each locality to the taxonomic rank to which the species belong. This coarse approach was developed as a proxy for the incorporation of major life-history traits that correlate with deep phylogeny into the predictive framework. The taxonomic ranks used here are very coarse, and as such, the life-history traits we refer to correspond to very general classifications of life, such as the ability to photosynthesize (e.g. rank 'plant'), or general classes in the animal kingdom that correlate with strongly divergent reproductive and developmental strategies (e.g. 'amphibian' versus 'bird' versus 'mammal').

(e) Random forest analyses

Several multivariate classification approaches can be applied to environmental and taxonomic variables. RF [60] is a powerful method that can be applied to predicting the presence or absence of cryptic lineages in species that co-occupy a disjunct biome. This machine-learning approach is based on the use of decision trees [61], which are used to classify and predict the assignment of observations into the response categories of interest, in this case, taxa that harbour cryptic diversity (which we will refer to as cryptic) versus those that do not (which we will refer to as non-cryptic). Each node in these decision trees represents a dichotomization of the data, defined by a condition in one of the predictor variables. The tree is grown (i.e. more nodes are added) with the addition of more conditions, and splitting the data until reaching the tips of the tree, the point where all observations are classified. When new data become available, novel observations can be classified, using the existing classification tree and the vector of predictor variables associated with those data [61]. While in a decision-tree analysis only one decision tree is built, in RF, the original full dataset is randomly bootstrapped and the variables randomly selected, and a decision tree is constructed for each bootstrapped dataset. The conditions contained in the final RF classification represent the modes of the conditions obtained in the full set of bootstrapped classification trees. The ensemble nature of the RF decision trees accommodates uncertainty and biases associated with the classification process that is used for the construction of each individual tree. Furthermore, the lack of distributional assumptions and the use of a subset of variables in each splitting node result in several advantages for classification trees over traditional methods such as discriminant function analyses or linear discriminant analyses (see for instance [62] for a discussion). RF has been applied in ecology [62,63], bioinformatics [64], the health sciences [65] and recently in statistical phylogeography [66].

We conducted the RFs analyses using the randomForest function from the *randomForest* package [67] in R. We allowed the RF function to construct 5000 decision trees using a random selection of our localities (with two-thirds of the dataset sampled with replacement to construct each decision tree) and setting the m value to 3 (three variables randomly sampled as candidates for each node). We assigned each specific locality datum to the category cryptic or non-cryptic (electronic supplementary material, table S1); this was the response variable and was assigned based on the results of the analytical approach described above (see §2c). As predictor variables, we used the climatic and taxonomic variables mentioned above (see §2d).

The RF method uses a subset of the data to train and construct the classification tree and then assesses the accuracy of the prediction on the 'out of bag' remainder of the observations to cross-evaluate model performance. Along with this measure, we apply a jackknife approach to test more explicitly the predictive power of the RF approach. This allowed us to evaluate how well the training dataset was able to predict presence/absence of cryptic diversity for a taxon that was initially not used to build the decision trees. For each run, we trained the model on all taxa except one. Predictions were then made for the omitted taxon. We then calculated the prediction accuracies of the method, as (i) $\text{accuracy}_{\text{overall}} = (n_{\text{true cryptic localities}} + n_{\text{true non-cryptic localities}})/n_{\text{total predicted localities}}$, (ii) $\text{accuracy}_{\text{cryptic}} = n_{\text{true cryptic localities}}/n_{\text{total cryptic localities}}$, (iii) $\text{accuracy}_{\text{non-cryptic}} = n_{\text{true non-cryptic localities}}/n_{\text{total non-cryptic localities}}$.

(i) Tests of data quality and its effect on results

One potential issue associated with biological data is the effect that an imbalanced representation of the different categories could have on the RF decision trees constructed. For example, in our case studies, the number of localities categorized as cryptic versus non-cryptic were unequally represented in the dataset (see electronic supplementary material, table S1). In this situation, it is possible that very heavily collected taxa/categories could drive the RF function and bias the overall predictions. Similarly, taxa/categories with few localities could also impinge on accuracy, as they would not contribute very strongly to the classification. To address this concern, we follow the recommendations by Chen *et al*. [68] and use two resampling strategies to evaluate how dataset balance and size affects the final predictive assignment. In the first strategy, all species in the dataset were resampled to obtain equal numbers of localities per species. Based on the characteristics of the PNW and SAL datasets, we resampled 100 times different numbers of localities: 141, 1500, 4500 and 9000 localities for the PNW dataset, and 1000, 5000, 13 500 and 25 000 for the SAL dataset. In the second strategy, we downsampled the number of observations in the majority category (i.e. cryptic taxa) to the number of observations in the minority (non-cryptic) taxa. To determine the feasibility of using existing data to predict the likelihood that a species harbours cryptic diversity, we then evaluated how the prediction accuracy changed as a function of each resampling approach. All dataset manipulations were done using custom scripts in R (see dryad repository).

(f) Predicting unknown taxa

For the PNW, we selected the red alder (*Alnus rubra*), the western red cedar (*Thuja plicata*) and the robust lancetooth snail (*Haplotrema vancouverense*). For the SAL, we selected the Gila woodpecker (*Melanerpes uropygialis*), Costa's hummingbird (*Calypte costae*) and the desert woodrat (*Neotoma lepida*). To predict the presence or absence of cryptic diversity in these species, we searched, downloaded and curated localities from GBIF, the bibliography and collection databases. To perform this selection, we used only localities that had been already georeferenced (i.e. geographical coordinates were already present), excluding those that fell outside of the range of the species (e.g. georeferencing errors) and those that were obviously wrong. We then applied the RF-generated classification described above for the appropriate (PNW or SAL) dataset, using the predict function of the R package *randomForest*. This provided a prediction for the presence or absence of cryptic diversity in the unknown taxa.

3. Results and discussion

(a) General overview

Phylogeography is a prolific discipline, with nearly 40 000 investigations published to date (Web of Science search of 'phylogeograph*' in title, abstract or keywords, 6 January 2016) on species collected from more than 4.8×10^6 localities [69]. These studies represent a tremendous resource for the

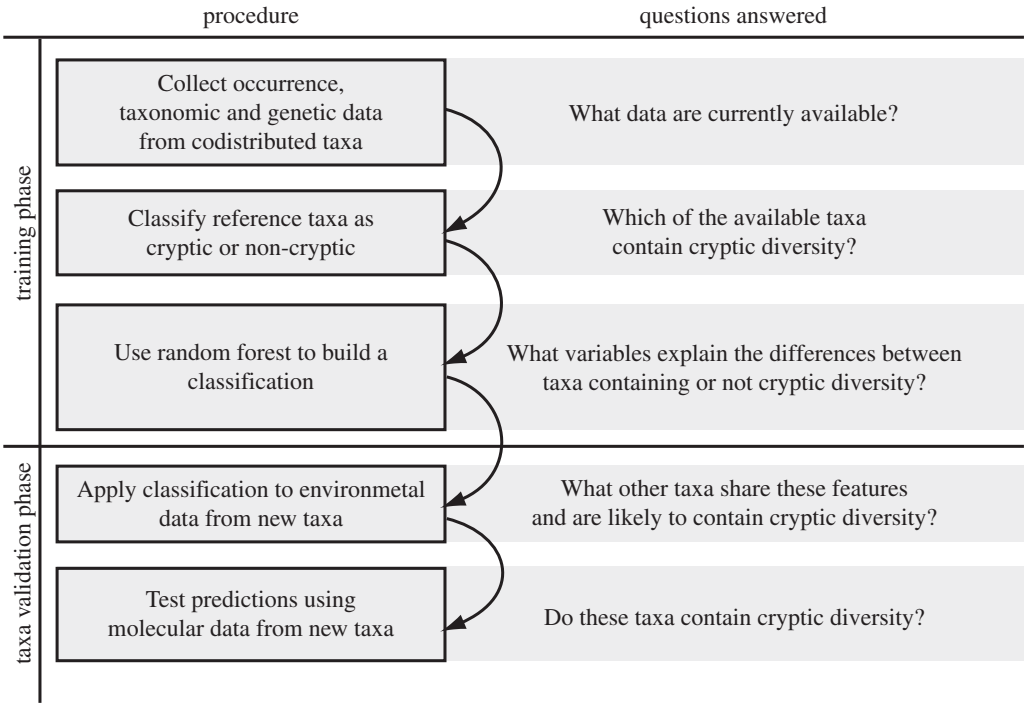

Figure 2. Conceptual overview of the steps and questions asked in the predictive procedure presented in this work.

biodiversity sciences because they include the samples of genetic diversity from across the range of the focal taxon. However, the information contained in these studies is difficult to access and cannot easily be generalized. Although comparative phylogeographic investigations are abundant and can effectively investigate regional diversification [33,40,70], they do not reflect global patterns.

In this study, we presented a new method that synthesizes data from multiple phylogeographic studies with other environmental and taxonomic information to predict the presence (or absence) of cryptic diversity within unstudied taxa. By testing the method with real data from two disjunct North American biomes, we demonstrate that climatic and taxonomic data can be used to predict the presence of cryptic diversity, and suggest that this method may be widely applicable to any ecosystem for which partial information about cryptic diversity (or the lack thereof) exists. Given the rapidly increasing number of phylogeographic studies, the potential application of this approach is high, and it may provide the means for phylogeography to transition into a predictive discipline.

Our predictive approach to phylogeography consists of several steps (figure 2): (i) compiling an occurrence dataset for taxa present in an ecosystem that either harbour or lack cryptic diversity, (ii) categorizing the taxa as cryptic or non-cryptic using phylogeographic methods, (iii) identifying appropriate variables (environmental, taxonomic, etc.) to construct classification trees with RF, (iv) adjusting an RF classification using resampled and downsampled datasets if the datasets are imbalanced, (v) using the resulting RF to predict presence/absence of cryptic diversity in taxa for which the genetic structure is unknown.

(b) Dataset compilation and data structure

Both the PNW and the SAL have a disjunct distribution (figure 1*a*) and in some taxa, the disjunction has been shown to correspond to the presence of cryptic diversity [37,42]. In both these ecosystems, the identification of an ancient divergence event has been associated with the recognition of cryptic diversity, whereas recent colonization of one of the disjunct areas has been linked with the absence of cryptic diversity, making these two biomes ideal for testing the utility of the here-presented method. Of all occurrence records, 1919 and 95 247 passed the quality test for the PNW and SAL taxa, respectively (electronic supplementary material, table S1). Not surprisingly, the occurrence data were very imbalanced in both datasets, with a minimum of 33 and 643 localities per taxon in the PNW and SAL datasets, respectively, and a maximum of 610 and 41 672 localities in the PNW and SAL datasets, respectively.

(c) Categorizing the training taxa

When using the demographic (ABC) and phylogenetic (Bayesian molecular clock) analyses to classify our reference set of taxa (i.e. the training dataset) into those that harbour and those that lack cryptic diversity, our phylogeographic results agreed with those obtained from previous studies [33,40,44]. Indeed, we identified several species that harbour cryptic diversity in each biome (electronic supplementary material, tables S2–S3). Specifically, our ABC analyses for these species identified the AV phylogeographic model as the most probable given the data (electronic supplementary material, table S2), and displayed old divergences and reciprocal monophyly among the disjunct regions (molecular clock analyses; electronic supplementary material, table S3). Therefore, the use of these phylogeographic methods appears to be an appropriate and objective approach to standardize the initial class assignment.

(i) Pacific Northwest taxa

Four of the seven species were identified in our phylogeographic analyses as non-cryptic. In those species, the ABC analyses identified a migration model as the one with the highest posterior probability (electronic supplementary

material, table S2) and the Bayesian molecular clock approach showed no reciprocal monophyly between disjunct populations in all species but *C. armata* (electronic supplementary material, table S3 and figures S6–S13). Three amphibian taxa, *Plethodon*, *Dicamptadon* and *Ascaphus*, were identified as cryptic. In these taxa, the AV model had the highest posterior probability (electronic supplementary material, table S2) and their disjunct populations were reciprocally monophyletic (electronic supplementary material, table S3).

(ii) Southwest arid lands taxa

Two of the four taxa analysed did not display cryptic diversity. In these datasets, the migration models with recent divergence events had the highest posterior probabilities (electronic supplementary material, table S2) and their disjunct populations were not reciprocally monophyletic (electronic supplementary material, table S3 and figures S2–S5). Both mammalian taxa were identified as cryptic by our phylogeographic analyses. For these taxa, the AV model had the highest posterior probability (electronic supplementary material, table S2) and the Bayesian molecular clock approach indicated reciprocal monophyly among disjunct populations (electronic supplementary material, table S3).

(d) Random forest data classification

In both datasets, we built an RF classification [60] of the cryptic and non-cryptic classes, using bioclimatic and taxonomic variables as classifiers, and we evaluated the classification accuracies with a jackknife approach. Because, in our datasets, the two classes were imbalanced, we tested the effect of that data structure by performing four resampling approaches and one downsampling approach. Our results demonstrate that the classifications are accurate, and that the accuracy improves when classifications are built using balanced datasets. This result agrees with previous studies on the use of RF [68], which indicated that data imbalance can have pervasive effects on the classification results. Our results strongly suggest that data balancing manipulation should be done when applying our approach.

(i) Random forest on the full datasets

In the PNW dataset, the RF approach successfully predicted the presence or absence of cryptic diversity for most taxa (overall accuracy of 98.79%; table 1). Further, the RF accuracies per category were balanced: 98.52% for species that harbour cryptic diversity and 100% for those that do not. However, although the RF was accurate overall, the prediction for one taxon, *C. armata*, did not agree with our expectations (electronic supplementary material, table S4). The most important variables contributing to the classification between the cryptic and non-cryptic groups were the taxonomic rank and annual mean temperature (bio1).

In the case of the SAL, the RF only reached an accuracy of 56.56% (table 1). Predictive accuracies were lower for the non-cryptic than the cryptic categories: 23.85% and 69.28%, respectively (table 1). Among all studied taxa, six were wrongly predicted (electronic supplementary material, table S4). The most important variables contributing to the classification between the cryptic and non-cryptic groups were the taxonomic rank, the mean annual temperature (bio1) and precipitation during the warmest quarter (bio18).

Table 1. Prediction accuracies (in %), based on the full, the downsampled and the resampled datasets. Values indicate overall and category-based (i.e. cryptic versus non-cryptic) accuracies.

dataset	overall	cryptic	non-cryptic
PNW			
full	98.78	98.52	100.00
downsampling	98.78	98.52	100.00
resampling 141	77.52	83.14	51.78
resampling 1500	98.78	98.52	100.00
resampling 4500	98.78	98.52	100.00
resampling 9000	98.78	98.52	100.00
SAL			
full	56.56	69.28	23.85
downsampling	64.44	62.38	69.96
resampling 1500	62.86	61.23	67.24
resampling 5000	64.72	63.68	67.49
resampling 13 500	64.10	62.57	68.23
resampling 25 000	65.38	64.36	68.13

(ii) Random forest on balanced datasets

In the PNW dataset, downsampling the data did not affect the results obtained with the full dataset (overall accuracy of 98.78%, table 1). Although the prediction accuracies generally increased for individual taxa (electronic supplementary material, table S4 and figure S15), the overall accuracy did not change between the full and most resampled datasets (table 1). In the case of the SAL dataset, the downsampling approach improved the overall prediction accuracy (64.44%) compared with the full dataset (table 1), and could correctly predict 11 of the 14 taxa (electronic supplementary material, table S4 and figure S16). The improvement was likely due to reduced variance in prediction accuracies between categories. In particular, the non-cryptic category saw a strong increase in the prediction accuracy (23.85% in the full dataset, 69.96% in the downsampled versus 68.23% in the resampled one; table 1), with only a small loss of accuracy in the cryptic category.

(iii) Overall evaluations

When the sampling was balanced and the sampling size increased, our analyses demonstrated moderate to high prediction accuracy (table 1 and electronic supplementary material, S4). This method is thus useful for predicting cryptic diversity and performs well on different types of datasets. Because the RF method does not have any distributional assumptions [60], it will be useful in biodiversity questions, especially when compared with other distribution-based approaches (see electronic supplementary material). Interestingly, the error rate in RF is known to increase with correlation of decision trees in the forest [60], and this behaviour is likely one characteristic driving the differences in the accuracies obtained when the resampling approaches are applied. When the number of localities is small or some taxa are overrepresented, the decision trees

are more likely to be correlated, because the same localities may be resampled multiple times (with small sample sizes) and/or the amount of information available will be much larger for some taxa than for others (with imbalanced samplings). Thus, balancing and increasing the number of samples per species directly increases the overall and category accuracies, and diminishes variation in the calculated accuracy (table 1, see also [68]). By using a combination of RF and resampling approaches, we were able to balance the predictive powers of the two categories and generate unbiased predictions (i.e. very high accuracy in the prediction of both categories).

Although the overall results are promising, some predictions disagreed with previous expectations. For example, in the PNW dataset, RF consistently predicted *C. armata* as cryptic (electronic supplementary material, table S4), and because there was disagreement in the results between the ABC (electronic supplementary material, table S2) and BEAST (electronic supplementary material, table S3) approaches, we cannot properly assess the accuracy of the RF prediction for that taxon. The differences between the two analytical methods may be due to a lack of genetic signal in the dataset, and/or to the fact that genetic data for this taxon included localities from regions absent from the other taxa (see electronic supplementary material, S1 for further discussion). To clarify this prediction, we plan to collect additional molecular data for that taxon. In the SAL dataset, three of the 14 species were not correctly predicted (electronic supplementary material, table S4). However, cryptic diversity in that dataset was mostly defined using previously published classifications by the authors (molecular data unavailable).

The results of the jackknife resampling suggest that our proposed method performs differently in the two biomes investigated here. These differences could be due to either methodological or biological grounds. Methodologically, it can be argued that the set of variables used to build the RF classification functions may contribute differently to sorting lineages harbouring or not harbouring cryptic diversity in the two biomes. For instance, while climate and taxonomy may be very appropriate to capture the differences between cryptic and non-cryptic species in the PNW, they may be less appropriate to do so for the SAL taxa. Using those variables in such a situation could lead to poor classifications, lack of model generalizability and eventually reduced prediction accuracy (see a thorough discussion in [71]). For this reason, the inclusion of other biologically relevant variables, such as modes of dispersal, generation times, population sizes or morphological characters, may strongly increase the predictive power of the approach, and this line will be explored in future studies. The differences among the results for the two biomes may also be due to biological grounds. Indeed, although the two biomes are disjunct because of an important dispersal barrier, these barriers may not be equivalent. While in the PNW the barrier is virtually continuous and broad (i.e. hundreds of kilometres of unsuitable habitat), this is not the case for the SAL, where the disjunct populations may come into contact along the Colorado River, a relatively permeable barrier. This may have implications for the number of dispersal events and potential for ecological differentiation among the disjunct taxa. A second explanation relates to this: in this work, we used climate and taxonomy as a proxy for the ecology and biology of taxa. This entails the assumption that cryptic taxa harbour ecological differences from their non-cryptic counterparts, and that it is possible to use those differences to classify and predict cryptic from non-cryptic entities. In this framework, it is possible that taxa from the PNW dataset are more ecologically different from those from the SAL. Such a situation would cause the RF approach to be more accurate in the PNW than the SAL dataset. An evaluation of ecological (i.e. climatic) differentiation and niche occupancy in the two datasets provides some support for this interpretation (see electronic supplementary material, S1 and figure S17).

(e) Predicting diversity in unknown taxa

To demonstrate the application of our method, we used the RF approach to predict the presence or absence of cryptic diversity in a set of taxa for which the presence of cryptic diversity has not been assessed with genetic data, so that we could prioritize future work. We assessed three taxa per biome; the three taxa from the PNW (i.e. red alder *Alnus rubra*, Western red cedar *Thuja plicata* and robust lancetooth *Haplotrema vancouverense*) were predicted to lack cryptic diversity with relatively high probabilities (98.06%, 97.91% and 98.24%, respectively). Two of the three taxa selected from the SAL (Costa's hummingbird, *C. costae* and the desert woodrat, *N. lepida*) were predicted to contain cryptic diversity, whereas the Gila woodpecker *M. uropygiales* was predicted to lack cryptic diversity (55.28%, 68.48% and 51.23%, respectively). Interestingly, *N. lepida* has been recently shown to possess cryptic diversity based on published revisionary data [72].

4. Conclusion

Our results represent the first attempt at predictive phylogeography as an explicit eco-evolutionary discipline. The RF-based approach introduced here holds a great deal of promise for predicting cryptic diversity in biomes where there has been at least some sampling effort, and where other unsampled lineages share those same distributional patterns. One of the direct applications of this approach is allowing the prioritization of additional efforts to discover and describe cryptic species. Further, because our method is computationally efficient, once the presence (or absence) of cryptic diversity of the unknown taxa has been predicted and verified with molecular means, these new taxa can be integrated into an updated training dataset and the RF classification rebuilt. This way, the RF classification is constantly informed when new data for the biome becomes available, which should increase its predictive accuracy.

Ethics. This research project did not require any ethical approval from the local research ethics committee.

Data accessibility. Data and scripts are available in the Dryad Digital Repository at: http://dx.doi.org/10.5061/dryad.ss7d6 [73].

Authors' contributions. A.E., M.R. and M.L.S. analysed the data. All authors participated in the design of the study and drafted the manuscript. All authors gave final approval for publication.

Competing interests. The authors declare having no conflict of interests.

Funding. This work was supported by the National Sciences Foundation through the grants no. DEB-1457519 and DEB-1457726.

Acknowledgements. The authors thank the BC Royal Museum and the Carnegie Museum of Natural History for allowing access to their specimen databases. We thank J. Degenhardt and A. Stevenson for collected samples and data from *C. armata*. We thank the editors and reviewers for helpful comments that improved this manuscript.

References

1. Costello MJ, May RM, Stork NE. 2013 Can we name Earth's species before they go extinct? *Science* **339**, 413–416. (doi:10.1126/science.1230318)
2. Bickford D, Lohman DJ, Sodhi NS, Ng PK, Meier R, Winker K, Ingram KK, Das I. 2007 Cryptic species as a window on diversity and conservation. *Trends Ecol. Evol.* **22**, 148–155. (doi:10.1016/j.tree.2006.11.004)
3. Riddle BR, Hafner DJ. 1999 Species as units of analysis in ecology and biogeography: time to take the blinders off. *Glob. Ecol. Biogeogr.* **8**, 433–441. (doi:10.1046/j.1365-2699.1999.00170.x)
4. Fusinatto LA, Alexandrino J, Haddad CFB, Brunes TO, Rocha CFD, Sequeira F. 2013 Cryptic genetic diversity is paramount in small-bodied amphibians of the genus *Euparkerella* (Anura: Craugastoridae) endemic to the Brazilian Atlantic forest. *PLoS ONE* **8**, pARTN e79504. (doi:10.1371/journal.pone.0079504)
5. Beheregaray LB, Caccone A. 2007 Cryptic biodiversity in a changing world. *J. Biol.* **6**, 9. (doi:10.1186/jbiol60)
6. Demos TC, Kerbis Peterhans JC, Agwanda B, Hickerson MJ. 2014 Uncovering cryptic diversity and refugial persistence among small mammal lineages across the Eastern Afromontane biodiversity hotspot. *Mol. Phylogenet. Evol.* **71**, 41–54. (doi:10.1016/j.ympev.2013.10.014)
7. Pérez-Portela R, Arranz V, Rius M, Turon X. 2013 Cryptic speciation or global spread? The case of a cosmopolitan marine invertebrate with limited dispersal capabilities. *Sci. Rep.-UK* **3**, 3197. (doi:10.1038/srep03197)
8. Meleg IN, Zaksek V, Fiser C, Kelemen BS, Moldovan OT. 2013 Can environment predict cryptic diversity? The case of *Niphargus* inhabiting Western Carpathian groundwater. *PLoS ONE* **8**, ARTN e76760. (doi:10.1371/journal.pone.0076760)
9. Cook BD, Page TJ, Hughes JM. 2008 Importance of cryptic species for identifying 'representative' units of biodiversity for freshwater conservation. *Biol. Conserv.* **141**, 2821–2831. (doi:10.1016/j.biocon.2008.08.018).
10. Prada C, McIlroy SE, Beltran DM, Valint DJ, Ford SA, Hellberg ME, Coffroth MA. 2014 Cryptic diversity hides host and habitat specialization in a gorgonian–algal symbiosis. *Mol. Ecol.* **23**, 3330–3340. (doi:10.1111/mec.12808)
11. Ence DD, Carstens BC. 2011 SpedeSTEM: a rapid and accurate method for species delimitation. *Mol. Ecol. Resour.* **11**, 473–480. (doi:10.1111/j.1755-0998.2010.02947.x)
12. Hausdorf B, Hennig C. 2010 Species delimitation using dominant and codominant multilocus markers. *Syst. Biol.* **59**, 491–503. (doi:10.1093/sysbio/syq039)
13. O'Meara BC. 2010 New heuristic methods for joint species delimitation and species tree inference. *Syst. Biol.* **59**, 59–73. (doi:10.1093/sysbio/syp077)
14. Yang ZH, Rannala B. 2010 Bayesian species delimitation using multilocus sequence data. *Proc. Natl Acad. Sci. USA* **107**, 9264–9269. (doi:10.1073/pnas.0913022107)
15. Carstens BC *et al.* 2013 Model selection as a tool for phylogeographic inference: an example from the willow *Salix melanopsis*. *Mol. Ecol.* **22**, 4014–4028. (doi:10.1111/mec.12347)
16. Newbold T *et al.* 2014 A global model of the response of tropical and sub-tropical forest biodiversity to anthropogenic pressures. *Proc. R. Soc. B* **281**, 20141371. (doi:10.1098/rspb.2014.1371).
17. Titeux N, Henle K, Mihoub JB, Regos A, Geijzendorffer IR, Cramer W, Verburg PH, Brotons L. 2016 Biodiversity scenarios neglect future land-use changes. *Glob. Chang Biol.* **22**, 2505–2515. (doi:10.1111/gcb.13272)
18. Heard MJ, Smith KF, Ripp KJ, Berger M, Chen J, Dittmeier J, Goter M, McGarvey ST, Ryan E. 2013 The threat of disease increases as species move toward extinction. *Conserv. Biol.* **27**, 1378–1388. (doi:10.1111/cobi.12143)
19. Joppa LN *et al.* 2016 Filling in biodiversity threat gaps. *Science* **352**, 416–418. (doi:10.1126/science.aaf3565)
20. Funk WC, Caminer M, Ron SR. 2012 High levels of cryptic species diversity uncovered in Amazonian frogs. *Proc. R. Soc. B* **279**, 1806–1814. (doi:10.1098/rspb.2011.1653)
21. Pimm SL, Jenkins CN, Abell R, Brooks TM, Gittleman JL, Joppa LN, Raven PH, Roberts CM, Sexton JO. 2014 The biodiversity of species and their rates of extinction, distribution, and protection. *Science* **344**, ARTN 1246752. (doi:10.1126/science.1246752)
22. Crisp MD, Cook LG. 2007 A congruent molecular signature of vicariance across multiple plant lineages. *Mol. Phylogenet. Evol.* **43**, 1106–1117. (doi:10.1016/j.ympev.2007.02.030)
23. Meegaskumbura M, Bossuyt F, Pethiyagoda R, Manamendra-Arachchi K, Bahir M, Milinkovitch MC, Schneider CJ. 2002 Sri Lanka: an amphibian hot spot. *Science* **298**, 379. (doi:10.1126/science.298.5592.379)
24. Sullivan J, Arellano E, Rogers DS. 2000 Comparative phylogeography of Mesoamerican highland rodents: concerted versus independent response to past climatic fluctuations. *Am. Nat.* **155**, 755–768. (doi:10.1086/303362)
25. Coyne JA, Orr HA. 2004 *Speciation*. Sunderland, MA: Sinauer Associates; xiii, 545, 542 p. of plates p.
26. Feder JL, Flaxman SM, Egan SP, Comeault AA, Nosil P. 2013 Geographic mode of speciation and genomic divergence. *Annu. Rev. Ecol. Evol. Syst.* **44**, 73–97. (doi:10.1146/annurev-ecolsys-110512-135825)
27. Peterson AT. 2011 Ecological niche conservatism: a time-structured review of evidence. *J. Biogeogr.* **38**, 817–827. (doi:10.1111/j.1365-2699.2010.02456.x)
28. Anacker BL, Strauss SY. 2014 The geography and ecology of plant speciation: range overlap and niche divergence in sister species. *Proc. R. Soc. B* **281**, 20132980. (doi:10.1098/rspb.2013.2980)
29. Rissler LJ, Apodaca JJ. 2007 Adding more ecology into species delimitation: ecological niche models and phylogeography help define cryptic species in the black salamander (*Aneides flavipunctatus*). *Syst. Biol.* **56**, 924–942. (doi:10.1080/10635150701703063)
30. DellaSala DA, Alaback P, Craighead L, Goward T, Paquet PC, Spribille T. 2011 Temperate and boreal rainforests of Inland Northwestern North America. In *Temperate and boreal rainforests of the world* (ed. DA DellaSala), pp. 82–110. Washington, DC: Island Press.
31. DellaSala DA, Moola F, Alaback P, Paquet PC, Schoen JW, Noss RF. 2011 Temperate and boreal rainforests of the Pacific Coast of North America. In *Temperate and boreal rainforests of the world* (ed. DA DellaSala), pp. 42–81. Washington, DC: Island Press.
32. Bjork CR. 2010 Distribution patterns of disjunct and endemic vascular plants in the interior wetbelt of northwest North America. *Botany* **88**, 409–428. (doi:10.1139/B10-030)
33. Carstens BC, Brunsfeld SJ, Demboski JR, Good JM, Sullivan J. 2005 Investigating the evolutionary history of the Pacific Northwest mesic forest ecosystem: hypothesis testing within a comparative phylogeographic framework. *Evolution* **59**, 1639–1652. (doi:10.1554/04-661.1)
34. Gavin DG. 2009 The coastal-disjunct mesic flora in the inland Pacific Northwest of USA and Canada: refugia, dispersal and disequilibrium. *Divers. Distrib.* **15**, 972–982. (doi:10.1111/j.1472-4642.2009.00597.x)
35. Nielson M, Lohman K, Daugherty CH, Allendorf FW, Knudsen KL, Sullivan J. 2006 Allozyme and mitochondrial DNA variation in the tailed frog (Anura: *Ascaphus*): the influence of geography and gene flow. *Herpetologica* **62**, 235–258. (doi:10.1655/0018-0831(2006)62[235:Aamdvi]2.0.Co;2)
36. Hendricks P, Maxell B, Lenard S, Currier C. 2008 Surveys and predicted distribution models for land mollusks on USFS Northern Region lands: 2007. (USDA Forest Service, Northern Region.
37. Brunsfeld SJ, Sullivan J, Soltis DE, Soltis PS. 2001 Comparative phylogeography of Northwestern North America: a synthesis. In *Integrating ecological and evolutionary concepts in a spatial context* (eds J Silverton, J Antonovics), pp. 319–339. Oxford, UK: Blackwell Science.
38. Nielson M, Lohman DJ, Sullivan J. 2001 Phylogeography of the tailed frog (*Ascaphus truei*): implications for the biogeography of the Pacific Northwest. *Evolution* **55**, 147–160. (doi:10.1111/j.0014-3820.2001.tb01280.x)
39. Ricketts TH *et al*. 1999 *Terrestrial ecoregions of North America. A conservation assessment*. Washington, DC: Island Press.
40. Riddle BR, Hafner DJ, Alexander LF, Jaeger JR. 2000 Cryptic vicariance in the historical assembly of a Baja California peninsular desert biota. *Proc. Natl*

Acad. Sci. USA **97**, 14 438–14 443. (doi:10.1073/pnas.250413397)

41. Zink RM, Manne L. 2014 Homage to Hutchinson, and the role of ecology in lineage divergence and speciation. *J. Biogeogr.* **41**, 999–1006. (doi:10.1111/jbi.12252)
42. Schierenbeck KA. 2014 *Phylogeography of California: an introduction*. Oakland, CA: University of California Press.
43. Garrick RC, Nason JD, Meadows CA, Dyer RJ. 2009 Not just vicariance: phylogeography of a Sonoran Desert euphorb indicates a major role of range expansion along the Baja peninsula. *Mol. Ecol.* **18**, 1916–1931. (doi:10.1111/j.1365-294X.2009.04148.x)
44. Zink RM. 2002 Methods in comparative phylogeography, and their application to studying evolution in the North American aridlands. *Integr. Comp. Biol.* **42**, 953–959. (doi:10.1093/icb/42.5.953)
45. Pelletier TA, Crisafulli C, Wagner S, Zellmer AJ, Carstens BC. 2015 Historical species distribution models predict species limits in Western *Plethodon* salamanders. *Syst. Biol.* **64**, 909–925. (doi:10.1093/sysbio/syu090)
46. Wilke T, Duncan N. 2004 Phylogeographical patterns in the American Pacific Northwest: lessons from the arionid slug *Prophysaon coeruleum*. *Mol. Ecol.* **13**, 2303–2315. (doi:10.1111/j.1365-294X.2004.02234.x)
47. Brunsfeld SJ, Miller TR, Carstens BC. 2007 Insights into the biogeography of the Pacific Northwest of North America: evidence from the phylogeography of *Salix melanopsis*. *Syst. Bot.* **32**, 129–139. (doi:10.1600/036364407780360094)
48. Riddle BR, Hafner DJ, Alexander LF. 2000 Phylogeography and systematics of the *Peromyscus eremicus* species group and the historical biogeography of North American warm regional deserts. *Mol. Phylogenet. Evol.* **17**, 145–160. (doi:10.1006/mpev.2000.0841)
49. Hudson RR. 2002 Generating samples under a Wright–Fisher neutral model of genetic variation. *Bioinformatics* **18**, 337–338. (doi:10.1093/bioinformatics/18.2.337)
50. Rozas J, Librado P, Sánchez-DelBarrio JC, Messeguer X, Rozas R. 2010 *DNAsp. (5.10.1 ed)*. Barcelona, Spain: Universitat de Barcelona.
51. Drummond AJ, Suchard MA, Xie D, Rambaut A. 2012 Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol. Biol. Evol.* **29**, 1969–1973. (doi:10.1093/molbev/mss075)
52. Minin V, Abdo Z, Joyce P, Sullivan J. 2003 Performance-based selection of likelihood models for phylogeny estimation. *Syst. Biol.* **52**, 674–683. (doi:10.1080/10635150390235494)
53. Rambaut A, Suchard MA, Xie D, Drummond AJ. 2014 Tracer v1.6. See http://beast.bio.ed.ac.uk/Tracer.
54. Peterson AT, Soberón J, Pearson RG, Anderson RP, Martínez-Meyer E, Nakamura M, Araújo MB. 2011 *Ecological niches and geographic distributions*, px, 314 p. Princeton, NJ: Princeton University Press.
55. Hijmans RJ, Cameron SE, Parra JL, Jones PJ, Jarvis A. 2005 Very high resolution interpolated climate surfaces for global land areas. *Int. J. Climatol.* **25**, 1965–1978. (doi:10.1002/joc.1276)
56. Bivand R, Keitt T, Rowlingson B. 2015 rgdal: Bindings for the Geospatial Data Abstraction Library. R package version 1.0-7. See http://CRAN.R-project.org/package=rgdal.
57. Hijmans RJ, Phillips S, Leathwick J, Elith J. 2015 dismo: Species Distribution Modeling. R package version 1.0-12. See http://CRAN.R-project.org/package=dismo.
58. Thuiller W, Georges D, Engler R. 2014 biomod2: Ensemble platform for species distribution modeling. R package version 3.1-64. See http://CRAN.R-project.org/package=biomod2.
59. Calenge C. 2006 The package adehabitat for the R software: a tool for the analysis of space and habitat use by animals. *Ecol. Model.* **197**, 516–519. (doi:10.1016/j.ecolmodel.2006.03.017)
60. Breiman L. 2001 Random forests. *Mach. Learn.* **45**, 5–32. (doi:10.1023/A:1010933404324)
61. Breiman L. 1984 *Classification and regression trees*. Belmont, CA: Wadsworth International Group.
62. Cutler DR, Edwards TCJ, Beard KH, Cutler A, Hess KT, Gibson J, Lawler JJ. 2007 Random forests for classification in ecology. *Ecology* **88**, 2783–2792. (doi:10.1890/07-0539.1)
63. Evans JS, Murphy MA, Holden ZA, Cushman SA. 2011 Modeling species distribution and change using random forest. In *Predictive species and habitat modeling in landscape ecology—concepts and applications* (eds CA Drew, YF Wiersma, F Huettmann), pp. 139–159. New York, NY: Springer.
64. Boulesteix A-L, Janitza S, Kruppa J, Konig IR. 2012 Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. *WIREs Data Mining Knowledge Discov.* **2**, 493–507. (doi:10.1002/widm.1072)
65. Knights D, Costello EK, Knight R. 2011 Supervised classification of human microbiota. *FEMS Microbiol. Rev.* **35**, 343–359. (doi:10.1111/j.1574-6976.2010.00251.x)
66. Pudlo P, Marin JM, Estoup A, Cornuet JM, Gautier M, Robert CP. 2016 Reliable ABC model choice via random forests. *Bioinformatics* **32**, 859–866. (doi:10.1093/bioinformatics/btv684)
67. Liaw A, Wiener M. 2002 Classification and regression by randomForest. *R News* **2**, 18–22.
68. Chen C, Liaw A, Breiman L. 2004 Using random forest to learn imbalanced data. Statistics technical reports, 1–12. University of California Berkeley.
69. Garrick RC *et al.* 2015 The evolution of phylogeographic data sets. *Mol. Ecol.* **24**, 1164–1171. (doi:10.1111/mec.13108)
70. Hickerson MJ, Meyer CP. 2008 Testing comparative phylogeographic models of marine vicariance and dispersal using a hierarchical Bayesian approach. *BMC Evol. Biol.* **8**, Artn 322. (doi:10.1186/1471-2148-8-322)
71. Wenger SJ, Olden JD. 2012 Assessing transferability of ecological models: an underappreciated aspect of statistical validation. *Methods Ecol. Evol.* **3**, 260–267. (doi:10.1111/j.2041-210X.2011.00170.x)
72. Patton JL, Huckaby DG, Álvarez-Castañeda ST. 2014 *The evolutionary history and a systematic revision of woodrats of the* Neotoma lepida *group*. London, UK: University of California Press.
73. Espíndola A, Ruffley M, Smith ML, Carstens BC, Tank DC, Sullivan J. 2016 Data from: Identifying cryptic diversity with predictive phylogeography. Dryad Digital Repository. (http://dx.doi.org/10.5061/dryad.ss7d6)