Received: 6 December 2017 | Revised: 30 April 2018 | Accepted: 6 May 2018

DOI: 10.1111/jbi.13382

RESEARCH PAPER

WILEY Journal of Biogeography

A global analysis of bats using automated comparative phylogeography uncovers a surprising impact of Pleistocene glaciation

Bryan C. Carstens | Ariadna E. Morales | Kathryn Field | Tara A. Pelletier

Department of Evolution, Ecology, and Organismal Biology, The Ohio State University, Columbus, Ohio

Correspondence
Bryan Carstens, Department of Evolution, Ecology, and Organismal Biology, 318 W. 12th Ave., Columbus, OH 43210-1293.
Email: carstens.12@osu.edu

Funding information
Division of Environmental Biology, Grant/Award Number: 60046038; Ohio Supercomputer Center, Grant/Award Number: PAS1184; Consejo Nacional de Ciencia y Tecnologia, Grant/Award Number: Reg. 217900 CVU 324588

Editor: Camila Ribas

Abstract

Aim: Our work seeks to understand the global demographical response of bat species to the climate change that occurred at the Last Glacial Maximum (LGM).

Location: All continents except Antarctica.

Methods: Mitochondrial DNA sequences were sampled from bat species throughout the planet where we could associate a georeferenced sample with a given DNA sequence. Our investigation estimates the historical demographical response using over 12,000 samples from >300 nominal species of bats. Custom Python and R scripts were written to aggregate sequence data from GenBank, locality information from GBIF, and to associate these records to individual samples. We conducted approximate Bayesian computation to calculate the posterior probability of demographical bottleneck and expansion responses to the end of the Pleistocene, and then collected organismal trait data to identify traits that were associated with either demographical response. We also used R to estimate current and end-Pleistocene species distribution models (SDM) for species where >10 georeferenced samples were available.

Results: Analysis of the genetic data indicate that some temperate insectivores responded to the end of the Pleistocene by undergoing a demographical expansion. However, the neotropical family Phyllostomidae experienced the most dramatic response, with many of its species undergoing demographical bottlenecks. Larger bats, and those with shorter forewings, were more likely to undergo a demographical bottleneck. In contrast with the results of the genetic data analysis, the automated SDMs all predicted range expansion since the LGM.

Main conclusions: Historical populations of Neotropical bats that rely on Angiosperms for resources (i.e., pollen, nectar, fruit) were negatively influenced by the climate change that occurred at the end of the Pleistocene. Our work highlights the utility of incorporating exploratory trait-based analyses in phylogeography. It serves as an example of automated big data phylogeography, and suggests that repurposed data can lead to new insights about global biodiversity.

KEYWORDS

automated phylogeography, big data, Chiroptera, Holocene, Phylogeography, Pleistocene, species distribution modelling

1 | INTRODUCTION

> *An appreciation of the scale and frequency of climatic oscillations in the past few million years is modifying our views on how evolution proceeds. Such major events caused extinction and repeated changes in the ranges of those taxa that survived. Their spatial effects depend on latitude and topography, with extensive extinction and recolonization in higher latitudes and altitudinal shifts and complex refugia nearer the tropics. The associated population dynamics varied with life history and geography, and the present genetic constitution of the populations and species carry attenuated signals of these past dynamics.* (Hewitt, 2004).

One key research goal of phylogeography is to understand how broad scale climatic fluctuations have influenced the evolution of species. In particular, the global change initiated by the retreat of the glaciers that followed the Last Glacial Maximum (LGM) has been shown to influence a wide range of species in tropical (e.g., Thomé et al., 2010), temperate (e.g., Garrick et al., 2004), and high latitude (e.g., Hope et al., 2010) habitats. One key finding is that climate change over the last twenty thousand years has led to substantial changes in the geographical range of many species (Hewitt, 2004). Some temperate species have expanded their range northward following the retreat of glaciers, often increasing their range to the north by hundreds of kilometers (e.g., Emerson et al., 2010). In many cases, species response was dictated by newly available habitat left behind in the wake of glacial retreat (e.g., Carstens, Stevenson, Degenhardt, & Sullivan, 2004). Other species occupy habitat patches that oscillate between broad and contiguous to small and fragmented. For example, species restricted to high elevation habitat in the present climate likely moved down in elevation, and thus into more contiguous ranges, during cooler glacial periods (e.g., Knowles, 2001). Hundreds of phylogeographical investigations have invoked some role for Pleistocene climate change in influencing the evolution of species (e.g., Avise, 2000).

Two predominant genetic consequences of quaternary climate change have been identified (Hewitt, 2004). Species that dramatically expanded their geographical range often exhibit genetic diversity characteristic of population expansion (e.g., Rogers & Harpending, 1992). For example, in Europe, intraspecific genetic diversity is highest in the southern peninsulas that functioned as glacial refugia, and lowest in the northern regions that were recently colonized (Hewitt, 2004). Species that underwent a population bottleneck during the Holocene, presumably due to the fragmentation of a geographical range (e.g., Heller, Lorenzen, Okello, Masembe, & Siegismund, 2008), typically have fewer alleles with deeper coalescences among these alleles (e.g., Knowles, Carstens, & Keat, 2007). Entire communities of species can demonstrate a concerted response to climate change (e.g., Chan, Schanzenback, & Hickerson, 2014; Riddle, Hafner, Alexander, & Jaeger, 2000), and typically have high genetic diversity in stable regions with less in ephemeral habitat (e.g., Carnaval, Hickerson, Haddad, Rodrigues, & Moritz, 2009; Hugall, Moritz, Moussalli, & Stanisic, 2002).

Meta-analyses, where the results of many investigations are synthesized to elucidate broader inferences, have been proposed as an important tool for comparative phylogeography (Dawson, 2014). However, they are difficult to apply to this question due to differences among species and geographical regions. For example, despite the tenuous correlation between species dispersal abilities and geographical range size (Lester, Ruttenberg, Gaines, & Kinlan, 2007), long-distance dispersal appears to be a key component of population expansion following glaciation (Hewitt, 2004), which suggests that organismal traits related to dispersal at least partially influence the capacity of a particular species to respond to newly-opened habitat (e.g., Morales, Villalobos, Velazco, Simmons, & Piñero, 2016; Pelletier & Carstens, 2016). Similarly, the response of a given species to broad-scale climate change is influenced by both the organismal life history and the particular habitat that it occupies. Traits largely dictate how a particular species acquires resources and interacts with other species, and reflect adaptation to its habitat, which has a unique temperature, precipitation and albedo profile. Furthermore, differences in study design can render comparisons across results problematic.

Rather than attempt a meta-analysis while accounting for these factors, we proceed by repurposing existing genetic data from bats, but do not limit our investigation to a single geographical region. Our global comparison of phylogeographical patterns attempts to identify organismal traits that predict the historical demographical response of species to climate change. Our sample of species includes disparate range attributes (i.e., size, location, connectivity) and incorporates species distribution modelling to quantify how the ranges of hundreds of bat species may have changed since the end of the Pleistocene. Our analysis consists of two primary data types: georeferenced gene sequences and climate data that enable us to predict species range for each species. The former are used to estimate whether the species experienced population demographical expansion or a bottleneck coincident with the end of the Pleistocene, and the latter are used to estimate the current and historical (i.e., at LGM) species range. We assess the association of traits with a particular response, and use machine learning in an attempt to develop a model that predicts species response to the LGM climate change.

2 | MATERIALS AND METHODS

2.1 | Repurposed data

Custom PYTHON scripts were written to aggregate data as follows, and are available at: https://carstenslab.osu.edu/software.html. We downloaded all Global Biodiversity Information Facility (GBIF) accessions that matched the search term "*chiroptera*", identified records that included GenBank accessions, and retrieved these data from

GenBank by downloading all sequences using the Batch Entrez nucleotide search feature (accessed February, 2016). All data downloaded from GBIF were filtered to only include those with GPS points with "no known coordinate issues". Once the georeferenced specimens were identified, we used MUSCLE (Edgar, 2004) to align the sequence data from these samples by species. Unless stated otherwise, all statistical analyses were done using R v3.2.3 (R Development Core Team 2016).

2.2 | Species distribution modelling

Species distributions of 115 bat species were modelled independently for current climatic conditions. The present-day species distribution models (SDMs) were projected onto a model of the historical climate at the LGM (21,000 yr BP) in order to quantify how the species range may have changed. Analyses were limited to species with 10 or more samples in an attempt to balance broad representation across chiropteran families and zoogeographical regions while producing a relatively accurate estimate of the species distribution. Briefly, we used environmental variables from WorldClim (Hijmans, Cameron, Parra, Jones, & Jarvis, 2005) and removed variables that were correlated on a species-by-species basis. We used MAXENT (Phillips, Anderson, & Schapire, 2006) implemented in *biomod2* (Thuiller, Lafourcade, Engler, & Araujo, 2009) with 1,000 iterations of pseudoabsence data. Models were evaluated using measurements of ROC and TSS. Full details of this process are available (Supporting Information 1).

2.3 | Approximate Bayesian computation

Approximate Bayesian computation (ABC, Csilléry, Blum, Gaggiotti, & François, 2010; Pritchard, Seielstad, Perez-Lezaun, & Feldman, 1999) was used to calculate the posterior probability of two models of historical demography given the empirical data from each of the species of bats included in the analysis, using custom PYTHON scripts (Supporting Information). Prior distributions from two models for each species were generated using **ms** (Hudson, 2002) and dimensions that match the empirical data from the particular species. One (hereafter referred to as the *bottleneck* model) modelled population contraction coincident with the LGM, while the second (hereafter referred to as the *expansion* model) modelled exponential population growth that was initiated at the LGM (Figure 1). In order to mitigate the single locus data and (in some cases) smaller than ideal sample sizes, we only estimated two parameters from each data set: One associated with the timing of population size change ($\tau_{\text{size change}}$) and a second associated with the magnitude of this event. The former is scaled by 4N* generations, where *N* is the effective population size and generations is the generation length, and was drawn from a uniform prior with a range of 0.001–0.4. Priors on the magnitude of the population size change were defined using uniform distributions ranging from 0.01–1.0 for the *bottleneck* model, and 1.0–10.0 for the *expansion* model. Such a design allows both the bottleneck and expansion models to simulate data that vary from effectively no size change (i.e., when prior values are near 1.0) to extreme events (i.e., a substantial bottleneck or expansion) and was implemented to make the recognition of *bottleneck* and *expansion* species conservative such that we are more likely to fail to detect a species as belonging to one of these categories than to mistakenly classify a species with modest historical demographical change.

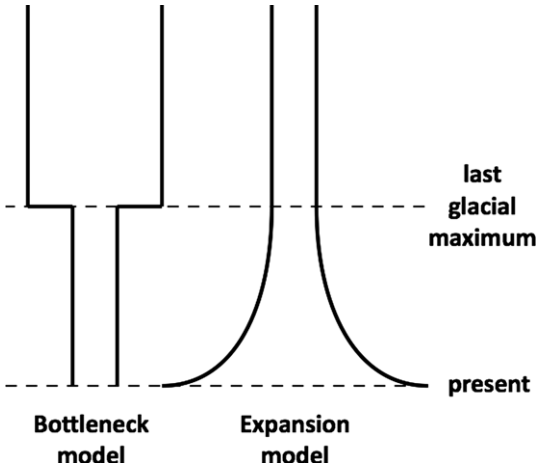

FIGURE 1 Description of models used in the ABC analysis. Demographic size change was modelled to occur at the Last Glacial Maximum, roughly 21,000 years before present (Pielou 1991). Bottlenecks were modelled as an instantaneous contraction of population size and expansion was modelled as exponential population growth

Our experimental design is intended to be a simplification of what are likely to be more complex patterns of demographical change in any particular bat species. For example, a species may have exhibited a bottleneck during the Pleistocene followed by an expansion during the Holocene, and if this is the case our models will not capture the complexity of the demographical history. There are dozens of combinations of demographical response that may be appropriate to any given species in our analysis, and it is nearly impossible in a comparative analysis with hundreds of species to tailor models to any particular species. Given challenges associated with choosing among multiple models (Pelletier & Carstens, 2014), the limitation of our analysis to two models with opposite responses is a conservative approach that should enable us to identify species where demography left a clear signal on their genetic variation.

One advantage of limiting our investigation to Chiroptera is that the included species should vary less in the life history traits that may influence the simulations. For example, the analysis assumes that generation lengths are between 3–5 years and effective population sizes are between 10,000 and 100,000 individuals, values consistent with previous investigations (e.g. Carstens & Dewey, 2010). Such generalizations would be much more difficult to justify in an analysis of all vertebrates, for example, than they are here. We simulated our data to match the observed number of segregating sites rather than attempting to estimate $\theta = 2N_{ef}\mu$ for each species, because such estimates would likely be poor given our single locus data (Felsenstein, 2005). Nucleotide diversity (π) and Tajima's *D* (*D*), both calculated using **sample_stats** (Hudson, 2002), were used to

summarize both the empirical and simulated data. The prior distribution for each model consisted of 100,000 simulated datasets, and closest 100 (0.001%) of these were retained via a rejection algorithm as implemented in **msReject** (Hickerson, Stahl, & Takebayashi, 2007). The proportion of datasets in the posterior distribution that were contributed by the *expansion* and *bottleneck* models represent the posterior probabilities of each of these models given the data.

The presence of unaccounted population genetic structure can lead to inference errors when estimating parameters related to population size change, including the expansion and bottleneck models utilized here (Gehara et al., 2016). To prevent our results from being biased in this manner, we identified species in our dataset that contained multiple entities as described in a recent analysis by Pelletier and Carstens (in review). This analysis examined >20,000 barcoding sequences collected from bats and applied the General Mixed Yule Coalescent (GMYC) model (Pons et al. 2006) to identify genetically structured intraspecific entities. While the GMYC has been criticized as being prone to the oversplitting of lineages (Esselstyn, Evans, Sedlock, Khan, & Heaney, 2012), this bias (if present) should identify the meaningful population genetic clusters within the species included here. Once these entities were identified, we examined the distribution of samples that composed these entities on a map in order to assess whether the GMYC entities are allopatric or sympatric. We then separated the GMYC entities within species that contained two or more such entities when these entities were allopatric under the rationale that the genetic structure was likely to result from geographical isolation and be indicative of separate gene pools for these entities. The GMYC entities were then analysed in the same manner as that reported to each species (below), provided that four or more samples were available per GMYC entity. We did not analyse entities with fewer than four samples because preliminary simulations suggested that we lacked power to differentiate the genetic expansion and bottleneck models with these low sample sizes. There were between two and four usable GMYC entities within approximately 30 species, which increased the number of the number of "species" in our analysis from 302 to 354. To evaluate the accuracy of the data for differentiating the bottleneck and expansion models, we conducted a cross-validation test for 10 simulations for each model. The average posterior probability for each known model was used to guide us in identifying a threshold for categorizing species as *expansion* or *bottleneck*.

2.4 | Trait differences in species response

Parametric tests were used to explore whether organismal traits, attributes of the geographical range, or spatial variables could explain species demographical change since the LGM (trait database available as Supplemental Materials). Species organismal traits were represented by several variables collected from various sources, including the panTHERIA trait database (Jones et al., 2009) and via descriptions of bat species and biology from a number of sources (Altringham, 1999; Emmons & Feer, 1997; Li et al., 2007; Marinello & Bernard, 2014; Moosman, Thomas, & Veilleux, 2012; Neuweiler, 2000; Norberg & Rayner, 1987; Nowak, 1994). Trait variables included: average adult body mass, average adult forearm length, species dietary niche, breeding habit, and roosting location. The sizes of the species ranges were quantified from the SDMs in km^2 at multiple thresholds for habitat suitability (0.5, 0.7, 0.9), including the relative difference between the estimated current and LGM ranges. The maximum latitude of the range, mean precipitation, and mean temperature from panTHERIA were also included.

We first classified bat species as either *expansion* or *bottleneck* using a threshold for the posterior probability (0.9) such that a species was required to exceed this threshold for being included. We then used a *t*-test to ask if there was a significant difference between bottleneck and expansion species for several variables: Adult body mass, forearm length, maximum range latitude, mean precipitation, mean temperature, the predicted area of range in the present and at the LGM, and the predicted proportion of range size change. A χ^2 test was used to explore differences between classifications for: dietary niche; breeding and roosting habits; Family; and zoogeographical province.

2.5 | Machine learning analysis

Random forest analysis (Liaw & Wiener, 2002) was used to identify variables that were important predictors of demographical expansion or bottlenecks in bat species. Random forest is a machine learning ensemble approach that utilizes multiple decision trees (a forest) to predict the response based on many potential predictor variables. In this case, we designed the algorithm to predict whether or not a species exhibits genetic expansion or bottleneck. In order to avoid bias induced by correlation among predictor variables, each individual decision tree consists of a subset of the data and a random ordering of variables at the nodes. Individually each tree is a weak predictor, but when many trees are combined the resulting consensus prediction is very strong (Biau, 2012; Breiman, 2001). Random forest samples the data with replacement for each decision tree in the forest and uses the unsampled datasets to test the model. This information is used to build a confusion matrix for the prediction and calculate the out of bag error rates (OOB). The importance of each predictor variable is determined by measuring the mean decrease in accuracy (MDA) of the prediction after the removal of each variable in the predictive function. We included taxonomic information about species in the random forest in an attempt to identify results that were being driven by phylogeny.

Bat species were classified as expansion or bottleneck as above, and at two lower thresholds for exploratory purposes (pp > 0.7, 0.8). We conducted the machine learning analysis for each combination of probability threshold using several sets of predictor variables from our trait database. Because the error rates were extremely unbalanced (see Table 2), we reran the random forest analysis using a downsampling scheme to even the response variables. Here, the majority class (bottleneck) was randomly subsampled to match the minority class (expansion). This was done 100 times and error rates were average across analyses. For the downsampling schemes, MDA was averaged across all 100 analyses. Full data tables for the machine learning analyses are available as Supplemental Materials.

3 | RESULTS

3.1 | Repurposed data

We obtained data from 302 nominal bat species, a number representing approximately a quarter of described bat species diversity (Wilson & Reeder, 2005). The average number of sequences was 48.2, the alignment length (in base pairs) was 675, and the average number of segregating sites was 24. Of these, 115 species were represented by ≥10 sampling localities, with the average number of unique GPS localities = 27.9. Most of the datasets consisted of data from DNA barcoding genes such as *Cytochrome oxidase I* and *Cytochrome b*. The samples were drawn from most families of bats in rough proportion to their species richness (Figure 2), although there are likely taxonomic and geographical biases.

3.2 | Species distribution modelling

At all threshold values of minimum probability of occurrence (0.5, 0.7, 0.9) all species were modelled to have expanded their range since the LGM (Supporting Information Table S1). With a threshold of 0.5, the proportion of range expansion goes from minimal (an increase in 0.88% in *Myotis nattereri*) to substantial (an increase in 24.45% in *Rhinolophus stheno*). Similar results are observed using thresholds of 0.7 (increases ranging from 0.37% in *Artibeus concolor* to almost 24.45% in *R. stheno*) and 0.9 (increases ranging from 0.20% in *Cormura brevirostris* to almost 44.53% in *R. stheno*). Note that regardless of the threshold of probability of occurrence, *R. stheno*, a species from the family Rhinolophidae found in Myanmar, Vietnam, Thailand, Lao PDR, Peninsular Malaysia, Sumatra and Java (Indonesia), is the species with the largest predicted range expansion.

3.3 | Approximate Bayesian computation

ABC was used to calculate the posterior probability of the expansion and bottleneck models in 302 nominal species of bats. Approximately 30 of the species contained more than one GMYC entity (Supporting Information Table S2), so we conducted in total 354 ABC analyses collectively using >70 million draws from prior distributions (Figure 3). While some temperate insectivorous bat species exhibited a strong signal of expansion coincident with the LGM (e.g., *Myotis mystacinus*), there were a relatively small number of species overall with such a pattern. Posterior probabilities >0.9 for the expansion model were only observed in 17 species, and these included as many species from tropical as temperate regions. In contrast, a far greater number (90) of species exhibited a signal of strong population bottleneck at the LGM, and this group also included both temperate and tropical species. Finally, the majority of species in our dataset (245) did not exhibit a strong signal (pp < 0.9) for either model, which may be a sign that they did not have a strong demographical response to the end of the Pleistocene, or may result from the conservative design of the ABC analysis and/or sampling artefacts.

Cross-validation testing indicates that we have some power to differentiate models for species so long as there is sufficient sampling and genetic variation (i.e., more than 10 samples or segregating sites). Both values are positively correlated with the accuracy of the analysis (Figure 4), and the number of segregating sites ($R^2_{\text{bottleneck}} = 0.313$, $p < 2.2 \times 10^{-16}$; $R^2_{\text{expansion}} = 0.2437$, $p < 2.2 \times 10^{-16}$) explain more of the variance in the accuracy than does the sample size ($R^2_{\text{bottleneck}} = 0.1441$, $p = 1.509 \times 10^{-13}$; $R^2_{\text{expansion}} = 0.187$, $p < 2.2 \times 10^{-16}$). In addition, the magnitude of population demographical size change also influences the accuracy (Supporting Information Figure S1).

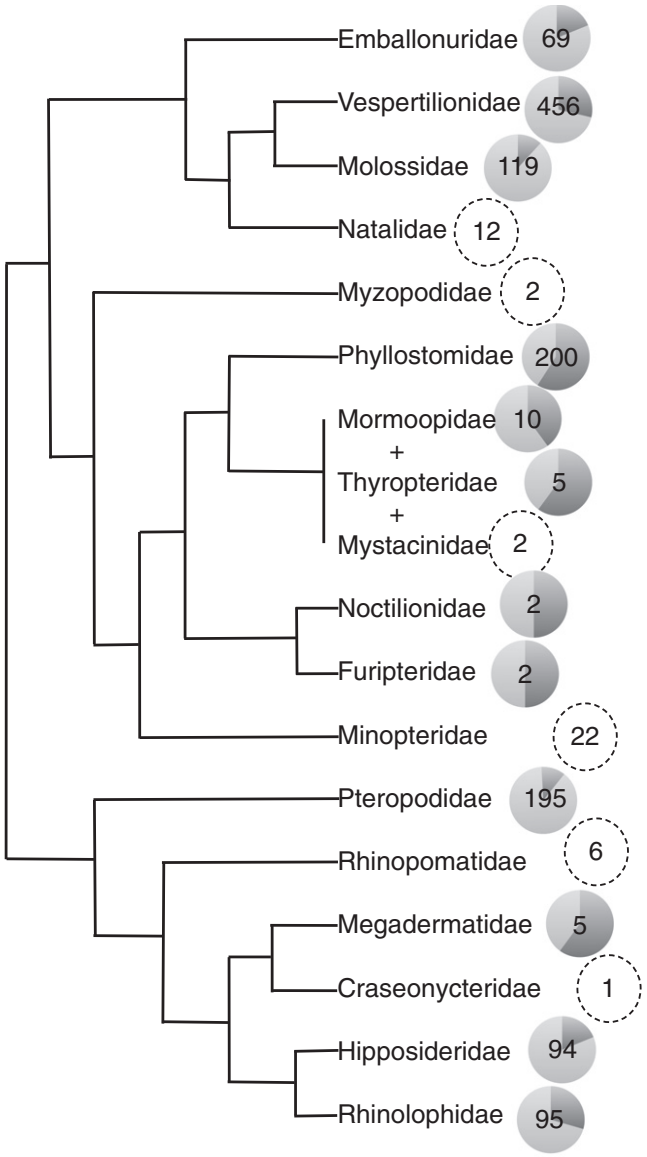

FIGURE 2 Sampling included in this investigation. A phylogeny, redrawn from fig. 3 of Agnarsson, Zabrana-Torrelio, Fores-Saldana, and May-Collado (2011), is shown with chiropteran families as the operational taxonomic units (OTU). A pie chart is shown next to each OTU, to represent the number of nominal species sampled from each clade for our analysis (darker pie piece). Each pie chart also includes the number of species in each family to illustrate differences in species richness across the phylogeny. Families that lack representation in our analysis are depicted using a circle with a dotted line

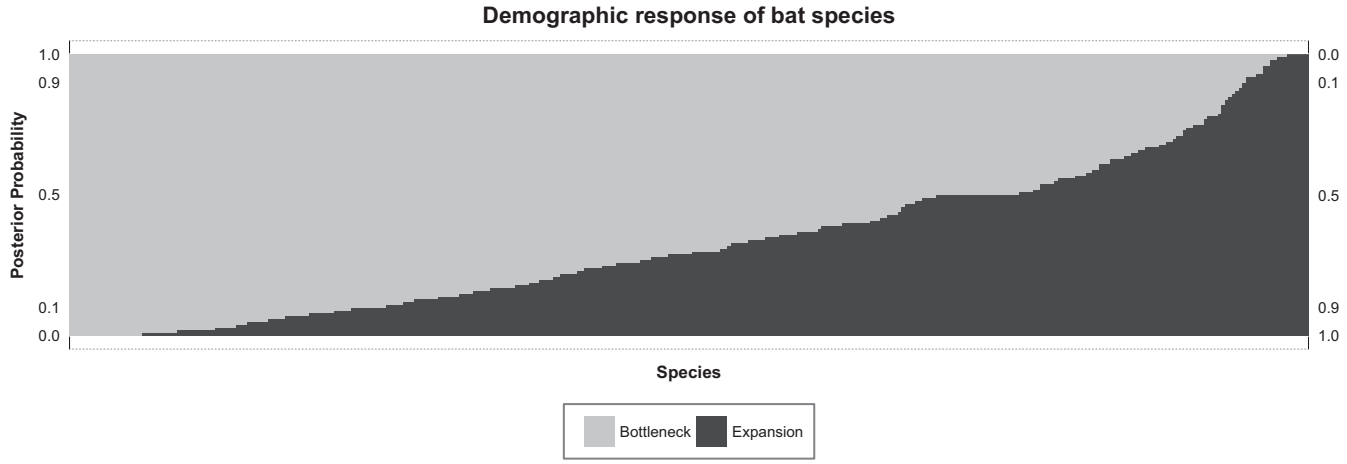

FIGURE 3 The posterior probabilities of the bottleneck and expansion models are shown for each of 354 species or GMYC entities in the analysis. Light grey is used to depict bottleneck species, with black used for expansion species. Results labelled by species are included in Supporting Information Table S2

3.4 | Trait differences in species response

There were no significant differences between species classified as bottleneck or expansion among the geographical and climatic variables or those derived from species distribution modelling (Table 2). Several organismal traits did significantly differ between these groups, including two related to size (adult body mass, $p = 0.0163$; adult forearm length, $p < 0.001$) and several categorical variables (dietary niche, $p < 0.003$; Family, $p < 0.00001$; zoogeographical province, $p < 0.00000001$). Categorical variables describing roosting or breeding habits did not differ in a significant manner (Table 1).

3.5 | Machine learning analyses

Regardless of probability threshold used to classify a species as either expansion or contraction, the random forest had difficulty correctly assigning species (Table 2 for pp > 0.9). This is likely due to the small number of species that experienced genetic expansion since the Pleistocene (13%–19%). Downsampling evened out the error rates across classes, but reduced the overall accuracy to ~50%, again probably due to low sample sizes in the data table. Variable importance, as measured by MDA, are included in the supplemental material (Supporting Information Table S3) but are not discussed further as these do a poor job of predicting expansion or bottleneck using the random forest model.

4 | DISCUSSION

4.1 | Global response of bats to Pleistocene climate change

The global change at the LGM had a substantial impact on the demography of bat species. Slightly under 1/3 of the analysed entities (107/352) exhibited a strong response in genetic variation (i.e., pp > 0.9 for either the bottleneck or expansion models). Far more species experienced a genetic bottleneck than an expansion at the LGM and this response was largely dictated by organismal rather than geographical traits (Table 1). For example, bottleneck species are nearly 10 g larger than expansion species, with correspondingly longer wings. Bottleneck species were less likely to be insectivorous, and more likely to be neotropical and from the family Phyllostomidae. In fact, many of the significant categories are characteristics of this family and it is likely that phyllostomid bats are the primary influence on the signal detected here.

The diversity of feeding niches in the Phyllostomidae is perhaps unmatched in any other mammalian family; it includes species that specialize on fruit, nectar, pollen, blood, vertebrates and insects. It may be that that many large fruigivorous species experienced a genetic bottleneck at the LGM because of changes in precipitation patterns (e.g., Bush & Silman, 2004) that impacted angiosperms that are used as a resource by these species. In other words, dietary niche explains the genetic patterns, rather than habitat availability in a broad sense. While many of the palaeotropical bottleneck species are larger frugivores, there are also a number of insectivorous species from these regions. Sampling biases may prevent recognition of similar signal from palaeotropical frugivores, or the differences between these communities (i.e., neo-vs. palaeo-tropical frugivorous bats) may result from other factors.

The SDMs predict that habitat suitability expanded in all species since the LGM. While habitat suitability does not directly translate to species population size (either census or effective), the uniformity of the result in the SDMs is surprising. Several explanations, including error, sampling artefacts, or unmet assumptions could account for the discrepancy observed here. We initially suspected that these results were driven by low sample sizes, but found no correlation between the magnitude of the difference between predicted historical and current range and the sample size (Supporting Information Figure S2). We suspect that habitat suitability predictions at the LGM are not good estimates of the historical range of these species. Furthermore, SDMs do not reflect biotic interactions, and as such

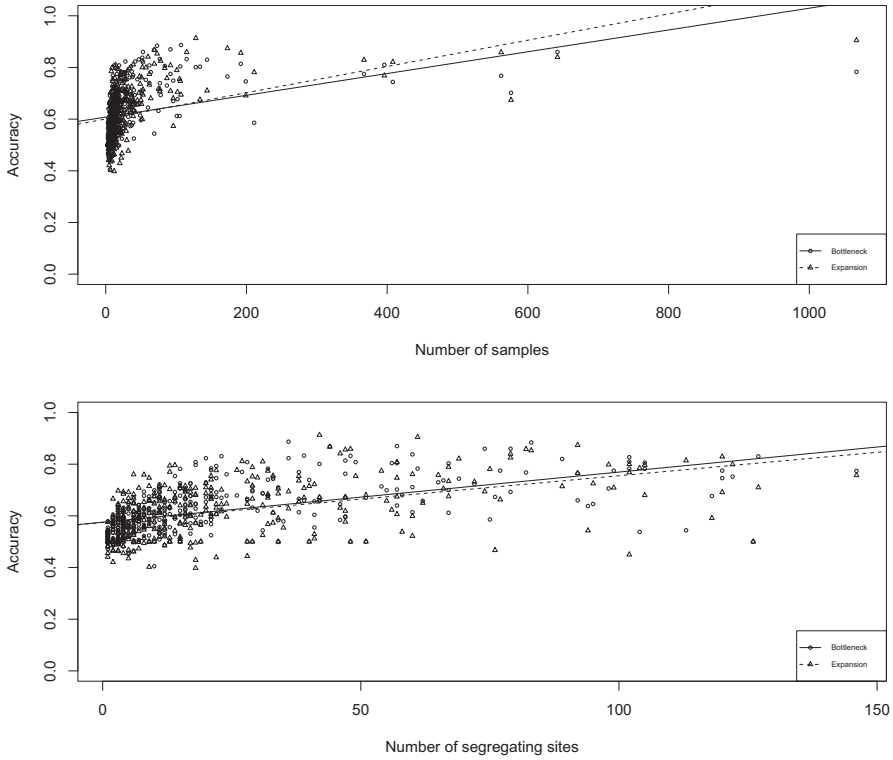

FIGURE 4 Results of simulation testing. For each simulation test, the posterior probability of the true model was calculated, and these values were averaged across replicates. Simulations were binned according to the number of samples and number of segregating sites (S) that they contained

oversimplify the factors that influence species demography. In general, they are difficult to verify against other sources of data (Gavin et al., 2014). These uncertainties, combined with the lack of a recent fossil record for most bat species, limits our power to infer historical species ranges.

4.2 | Trait-based comparative phylogeography

One outstanding goal of phylogeographical research is to more explicitly incorporate organismal phenotype into investigations (Zamudio, Bell, & Mason, 2016), and researchers have adopted a variety of approaches to integrating organismal traits into comparative phylogeography. For example, Paz, Ibanez, Lip, and Crawford (2015) used generalized linear modelling to demonstrate that body size and reproductive mode predicted genetic distance in Central American Frogs. Their approach incorporated species range attributes and organismal trait data in an attempt to explain patterns inherent to the genetic data, and such an approach could be conducted in an automated fashion. Alternatively, Papadopoulou and Knowles (2016) argued that organismal traits should be used to predict species-specific responses that can then be evaluated using a hypothesis-testing framework. Both approaches share a key feature with most previous comparative phylogeographical investigations in that they investigate systems consisting of closely related species that share broad similarities in their geographical distributions. However, they offer a marked contrast in how organismal trait data are incorporated into the analysis pipeline. Paz et al. (2015) seek to explore *how* organismal traits explain key differences in phylogeographical structure. They chose traits that are suspected of having an influence, but conduct statistical modelling to determine how and in what combination these traits exert an influence on the response variable. Papadopoulou and Knowles (2016) advocate choosing a trait on an a priori basis to evaluate via hypothesis testing, and make the implicit assumption that such traits can be recognized. Their approach is likely more difficult to implement on an automated basis. Our investigation, like that of Paz et al. (2015), is explorative in the sense that it seeks to identify organismal traits that are associated with a particular pattern in the data. While our analysis is not comparative phylogeography in the traditional sense because it investigates species that do not all share a geographical distribution, the inclusion of species from various biomes across the globe has clear benefits for the incorporation of organismal trait data.

4.3 | Automated big data phylogeography

Our findings are made possible by repurposing existing georeferenced genetic data (e.g., Sidlauskas et al., 2010), Such data contain immense potential for insight (Dawson, 2014; Soltis & Soltis, 2016) because thousands of studies investigating the geographical distribution of genetic variation have been published to date, each collecting data from hundreds of samples (Garrick et al., 2015). Unfortunately, only 7% of GenBank accessions of barcoding genes, such as COI, include latitude and longitude, and only 18% list museum catalogue information that can be used to link the sequence to a particular specimen (Marques, Maronna, & Collins, 2013). The disassociation of genetic and geographical accessions limits the utility of open source databases and must be addressed if biodiversity scientists are to leverage the information contained within existing data to meet the

TABLE 1 Comparison of trait differences between expansion and bottleneck species. For several categories of species traits, the mean value of species categorized as expansion and bottleneck by the ABC analysis are reported, along with the *p*-value and the test used

	Mean expansion	Mean bottleneck	*p*-value	Test
Geography and climate				
Maximum latitude of range (°)	26.5	23.7	0.4342	*t*
Mean precipitation (cm)	136.0	145.7	0.3782	*T*
Mean temperature (°C)	21.5	22.0	0.6182	*T*
Predicted size of range (km^2)				
At present, 50% suitability	3,657,870	3,321,545	0.662	*T*
At LGM, 50% suitability	262,582.5	169,225.3	0.1711	*T*
Proportion increase, 50% suitability	18.6	28.9	0.07301	*T*
At present, 90% suitability	536,075.6	408,056.7	0.1846	*T*
At LGM, 90% suitability	49,528.7	33,544.0	0.2787	*T*
Proportion increase, 90% suitability	19.2	37.4	0.1682	*T*
Organismal traits				
Adult body mass (gm)	13.3	23.0	0.01613	*T*
Adult forearm length (mm)	41.8	46.4	0.00861	*T*
Dietary niche			0.002481	χ^2
Breeding habit			0.7853	χ^2
Roosting habit			0.9975	χ^2
Taxonomic Family			2.59E-06	χ^2
Zoogeographic province			1.87E-09	χ^2

TABLE 2 Results of machine learning analysis. Shown for the random forest analysis are the out of bag error rates, expansion and bottleneck error rates, the number of variables in the analysis (*n*), and the number of species in the expansion and bottleneck categories. Results of two analyses (full, and with down sampling) are shown

Full analysis	**OOB error**	**Expansion error**	**Bottleneck error**	***n***	***n* expansion**	***n* bottleneck**
Organismal traits only	0.2198	0.8823	0.0675	91	17	74
Organismal traits with geography & climate	0.1418	1	0.0212	54	7	47
Down sampling	**OOB error**	**Expansion error**	**Bottleneck error**	***n***	***n* expansion**	***n* bottleneck**
Organismal traits only	0.4951	0.5263	0.4638	34	17	17
Organismal traits with geography & climate	0.6229	0.5935	0.6082	14	7	7

challenges associated with conservation of species and understanding patterns in evolution on a global scale (Pope, Liggins, Keyse, Carvalho, & Riginos, 2015). A concerted effort by researchers, journal editors and program officers is needed to increase the availability of georeferenced genetic data in published databases.

Our investigation relied on the development and implementation of automated phylogeographical analyses (e.g., Gratton et al., 2016). Using Python and R, we analysed data from hundreds of species in batch form, and used a combination of basic statistical comparisons, machine learning, and niche modelling analyses to extract signal from these data. The former returned several clear findings, including results such as the over-representation of the "*neotropical*" and "*Phyllostomid*" categories among bottleneck species. However, we were unable to build an accurate classification function using the Random Forest machine learning approach (Table 2). Several factors may contribute to this result. First, while general trends can result in significant differences between groups (as seen in the trait differences in species response), there are species that represent exceptions to these general trends (e.g., there are Phyllostomid species classified in the expansion set), and such interactions among variables may lead the random forest analysis to perform poorly with the small sample sizes used here. Second, it is likely that inadequate or biased sampling within some species limits our ability to strongly support a model in the ABC analysis, which would have obvious ramifications to the machine learning analysis. Finally, by relying on automated niche modelling, we likely have suboptimal estimates of the species geographical range. This is a potential problem with automated analyses, but one that can only be avoided by severely limiting the number of species included in a study so that all could be analysed manually. Regardless of these challenges, automated phylogeographical analyses have considerable potential.

Automated big data phylogeography represents an important direction for the field. While single species studies, particularly those that utilize multilocus data (e.g., Hotaling et al., 2017), can compare

a broad range of detailed models, and comparative studies into a single system enable a more nuanced investigation of how changes in the landscape have influenced evolutionary history (e.g., Mather, Hanson, Pope, & Riginos, 2017), neither has the potential to uncover generalized patterns on continental or global scales. Automated studies can incorporate organismal trait data in a manner that is not possible in more limited studies, which offers an important connection between evolutionary and ecological processes (Rissler, 2016). They complement traditional approaches to phylogeography and, by repurposing existing data, represent an efficient research program accessibly to most researchers.

ACKNOWLEDGEMENTS

Funding for TAP and KEF was provided by the National Science Foundation (DEB-60046038). Support for AEM was provided by a graduate fellowship at The Ohio State University funded by Consejo Nacional de Ciencia y Tecnologia (Reg.217900 CVU324588). The Ohio Supercomputer Center allocated resources to support part of this study (PAS1184 and PAS 1201). We thank members of the Carstens Lab and three reviewers for providing useful comments on drafts of this manuscript.

REFERENCES

Agnarsson, I., Zabrana-Torrelio, C. M., Fores-Saldana, N. P., & May-Collado, L. J. (2011). A time-calibrated species-level phylogeny of bats (Chiroptera, Mammalia). *PLoS Currents*, *4*, RRN1212.

Altringham, J. D. (1999). *Bats: Biology and behavior*. New York, NY: Oxford University Press.

Avise, J. C. (2000). *Phylogeography: The history and formation of species*. Cambridge, MA: Harvard University Press.

Biau, G. (2012). Analysis of a random forests model. *Journal of Machine Learning Research*, *13*, 1063–1095.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32. https://doi.org/10.1023/A:1010933404324

Bush, M. B., & Silman, M. R. (2004). Observations on Late Pleistocene cooling and precipitation in the lowland Neotropics. *Journal of Quaternary Science*, *19*, 677–684. https://doi.org/10.1002/(ISSN)1099-1417

Carnaval, A. C., Hickerson, M. J., Haddad, C. F. B., Rodrigues, M. T., & Moritz, C. (2009). Stability predicts genetic diversity in the Brazilian Atlantic Forest hotspot. *Science*, *323*, 785–789. https://doi.org/10.1126/science.1166955

Carstens, B. C., & Dewey, T. A. (2010). Species delimitation using a combined coalescent and information theoretic approach: An example from North American *Myotis* bats. *Systematic Biology*, *59*, 400–414. https://doi.org/10.1093/sysbio/syq024

Carstens, B. C., Stevenson, A. L., Degenhardt, J. D., & Sullivan, J. (2004). Testing nested phylogenetic and phylogeographic hypotheses in the *Plethodon vandykei* species group. *Systematic Biology*, *53*(5), 781–792. https://doi.org/10.1080/10635150490522296

Chan, Y. L., Schanzenback, D., & Hickerson, M. J. (2014). Detecting concerted demographic response across community assemblages using hierarchical approximate Bayesian computation. *Molecular Biology and Evolution*, *13*, 2501–2515. https://doi.org/10.1093/molbev/msu187

Csilléry, K., Blum, M. G., Gaggiotti, O. E., & François, O. (2010). Approximate Bayesian computation (ABC) in practice. *Trends in Ecology & Evolution*, *25*(7), 410–418. https://doi.org/10.1016/j.tree.2010.04.001

Dawson, M. N. (2014). Natural experiments and meta-analyses in comparative phylogeography. *Journal of Biogeography*, *41*(1), 52–65. https://doi.org/10.1111/jbi.12190

Edgar, R. C. (2004). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, *32*(5), 1792–1797. https://doi.org/10.1093/nar/gkh340

Emerson, K. J., Merz, C. R., Catchen, J. M., Hohenlohe, P. A., Cresko, W. A., Bradshaw, W. E., & Holzapfel, C. M. (2010). Resolving postglacial phylogeography using high-throughput sequencing. *Proceedings of the National Academy of Sciences*, *107*(37), 16196–16200. https://doi.org/10.1073/pnas.1006538107

Emmons, L. H., & Feer, F. (1997). *Neotropical rainforest mammals: A field guide*. Chicago, IL: University of Chicago Press.

Esselstyn, J. A., Evans, B. J., Sedlock, J. L., Khan, F. A. A., & Heaney, L. R. (2012). Single-locus species delimitation: A test of the mixed Yule–coalescent model, with an empirical application to Philippine round-leaf bats. *Proceedings of the Royal Society of London B: Biological Sciences*, *297*, 3678–3686. https://doi.org/10.1098/rspb.2012.0705

Felsenstein, J. (2005). Accuracy of coalescent likelihood estimates: Do we need more sites, more sequences, or more loci? *Molecular Biology and Evolution*, *23*(3), 691–700.

Garrick, R. C., Bonatelli, I. A., Hyseni, C., Morales, A., Pelletier, T. A., Perez, M. F., … Carstens, B. C. (2015). The evolution of phylogeographic data sets. *Molecular Ecology*, *24*(6), 1164–1171. https://doi.org/10.1111/mec.13108

Garrick, R. C., Sands, C. J., Rowell, D. M., Tait, N. N., Greenslade, P., & Sunnucks, P. (2004). Phylogeography recapitulates topography: Very fine-scale local endemism of a saproxylic 'giant' springtail at Tallaganda in the Great Dividing Range of south–east Australia. *Molecular Ecology*, *13*(11), 3329–3344. https://doi.org/10.1111/j.1365-294X.2004.02340.x

Gavin, D. G., Fitzpatric, M. C., Gugger, P. F., Heath, K. S., Rodriquez-Sanches, F., Dobrowski, S. Z., … Williams, J. W. (2014). Climate refugia: Using fossils, genetics, and spatial modeling to explain the past and project the future of biodiversity. *New Phytologist*, *204*, 37–54. https://doi.org/10.1111/nph.12929

Gehara, M., Garda, A. A., Werneck, F. P., Oliveira, E. F., da Fonseca, E. M., Camurugi, F., … Burbrink, F. T. (2016). Estimating synchronous demographic changes across populations using hABC and its application for a herpetological community from northeastern Brazil. *Molecular Ecology*, *26*, 4756–4771.

Gratton, P., Marta, S., Bocksberger, G., Winter, M., Trucchi, E., & Kühl, H. (2016). A world of sequences: Can we use georeferenced nucleotide databases for a robust automated phylogeography? *Journal of Biogeography*, *44*, 475–486.

Heller, R., Lorenzen, E. D., Okello, J. B., Masembe, C., & Siegismund, H. R. (2008). Mid-Holocene decline in African buffalos inferred from Bayesian coalescent-based analyses of microsatellites and mitochondrial DNA. *Molecular Ecology*, *17*(22), 4845–4858. https://doi.org/10.1111/j.1365-294X.2008.03961.x

Hewitt, G. M. (2004). Genetic consequences of climatic oscillations in the Quaternary. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *359*(1442), 183–195. https://doi.org/10.1098/rstb.2003.1388

Hickerson, M. J., Stahl, E., & Takebayashi, N. (2007). msBayes: Pipeline for testing comparative phylogeographic histories using hierarchical approximate Bayesian computation. *BMC Bioinformatics*, *8*(1), 268. https://doi.org/10.1186/1471-2105-8-268

Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G., & Jarvis, A. (2005). Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, *25*, 1965–1978. https://doi.org/10.1002/(ISSN)1097-0088

Hope, A. G., Waltari, E., Dokuchaev, N. E., Abramov, S., Dupal, T., Tsvetkova, A., … Cook, J. A. (2010). High-latitude diversification within Eurasian least shrews and Alaska tiny shrews (Soricidae). *Journal of*

Mammalogy, *91*(5), 1041–1057. https://doi.org/10.1644/09-MAMM-A-402.1

Hotaling, S., Muhfeld, C. C., Giersch, J. J., Ali, O. A., Jordan, S., Miller, M. R., … Weisrock, D. W. (2017). Demographic modelling reveals a history of divergence with gene flow for a glacially tied stonefly in a changing post-Pleistocene landscape. *Journal of Biogeography*, *45*, 304–317. https://doi.org/10.1111/jbi.13125

Hudson, R. R. (2002). Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*, *18*(2), 337–338. https://doi.org/10.1093/bioinformatics/18.2.337

Hugall, A., Moritz, C., Moussalli, A., & Stanisic, J. (2002). Reconciling paleodistribution models and comparative phylogeography in the Wet Tropics rainforest land snail *Gnarosophia bellendenkerensis* (Brazier 1875). *Proceedings of the National Academy of Sciences*, *99*(9), 6112–6117. https://doi.org/10.1073/pnas.092538699

Jones, K. E., Bielby, J., Cardillo, M., Fritz, S. A., O'Dell, J., Orme, C. D., … Connolly, C. (2009). PanTHERIA: A species-level database of life history, ecology, and geography of extant and recently extinct mammals. *Ecology*, *90*(9), 2648. https://doi.org/10.1890/08-1494.1

Knowles, L. L. (2001). Did the Pleistocene glaciations promote divergence? Tests of explicit refugial models in montane grasshopprers. *Molecular Ecology*, *10*(3), 691–701.

Knowles, L. L., Carstens, B. C., & Keat, M. L. (2007). Coupling genetic and ecological-niche models to examine how past population distributions contribute to divergence. *Current Biology*, *17*(11), 940–946. https://doi.org/10.1016/j.cub.2007.04.033

Lester, S. E., Ruttenberg, B. I., Gaines, S. D., & Kinlan, B. P. (2007). The relationship between dispersal ability and geographic range size. *Ecology Letters*, *10*(8), 745–758. https://doi.org/10.1111/j.1461-0248.2007.01070.x

Li, G., Liang, B., Wang, Y., Zhao, H., Helgen, K. M., Lin, L., … Zhang, S. (2007). Echolocation calls, diet, and phylogenetic relationships of Stoliczka's trident bat, *Aselliscus stoliczkanus* (Hipposideridae). *Journal of Mammalogy*, *88*(3), 736–744. https://doi.org/10.1644/06-MAMM-A-273R.1

Liaw, A., & Wiener, M. (2002). Classification and regression by random Forest. *R News*, *2*(3), 18–22.

Marinello, M. M., & Bernard, E. (2014). Wing morphology of Neotropical bats: A quantitative and qualitative analysis with implications for habitat use. *Canadian Journal of Zoology*, *92*(2), 141–147. https://doi.org/10.1139/cjz-2013-0127

Marques, A. C., Maronna, M. M., & Collins, A. G. (2013). Putting GenBank data on the map. *Science*, *341*, 1341. https://doi.org/10.1126/science.341.6152.1341-a

Mather, A. T., Hanson, J. O., Pope, L. C., & Riginos, C. (2017). Comparative phylogeography of two co-distributed but ecologically distinct rainbowfishes of far-northern Australia. *Journal of Biogeography*, *45*, 127–141. https://doi.org/10.1111/jbi.13117

Moosman, P. R. Jr, Thomas, H. H., & Veilleux, J. P. (2012). Diet of the widespread insectivorous bats *Eptesicus fuscus* and *Myotis lucifugus* relative to climate and richness of bat communities. *Journal of Mammalogy*, *93*(2), 491–496. https://doi.org/10.1644/11-MAMM-A-274.1

Morales, A., Villalobos, F., Velazco, P. M., Simmons, N. B., & Piñero, D. (2016). Environmental niche drives genetic and morphometric structure in a widespread bat. *Journal of Biogeography*, *43*(5), 1057–1068.

Neuweiler, G. (2000). *The biology of bats*. New York, NY: Oxford University Press.

Norberg, U. M., & Rayner, J. M. (1987). Ecological morphology and flight in bats (Mammalia; Chiroptera): Wing adaptations, flight performance, foraging strategy and echolocation. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *316*(1179), 335–427. https://doi.org/10.1098/rstb.1987.0030

Nowak, R. M. (1994). *Walker's bats of the world*. Baltimore, MD: Johns Hopkins University Press.

Papadopoulou, A., & Knowles, L. L. (2016). Towards a paradigm shift in comparative phylogeography driven by trait-based hypotheses. *Proceedings of the National Academy of Sciences*, *13*(29), 8018–8024. https://doi.org/10.1073/pnas.1601069113

Paz, A., Ibanez, R., Lip, K. R., & Crawford, A. J. (2015). Testing the role of ecology and life history in structuring genetic variation across a landscape: A trait-based phylogeographic approach. *Molecular Ecology*, *24*, 3723–3737. https://doi.org/10.1111/mec.13275

Pelletier, T. A., & Carstens, B. C. (2014). Model choice in phylogeography using a large set of models. *Molecular Ecology*, *23*, 3028–3043. https://doi.org/10.1111/mec.12722

Pelletier, T. A., & Carstens, B. C. (2016). Comparing range evolution in two western *Plethodon* salamanders: Glacial refugia, competition, ecological niches, and spatial sorting. *Journal of Biogeography*, *43*, 2237–2249. https://doi.org/10.1111/jbi.12833

Pelletier, T. A., & Carstens, B. C. (In review) Barcoding genes reveal high numbers of cryptic species in bats and demonstrate the need for basic taxonomic research.

Phillips, S. J., Anderson, R. P., & Schapire, R. E. (2006). Maximum entropy modelling of species geographic distributions. *Ecological Modelling*, *190*, 231–259. https://doi.org/10.1016/j.ecolmodel.2005.03.026

Pielou, E. C. (1991). *After the Ice Age: The return of life to glaciated North America*, pp. 376. University of Chicago Press, Chicago, IL.

Pons, J., Barraclough, T. G., Gomez-Zurita, J., Cardoso, A., Duran, D. P., Hazell, S., & … Vogler, A. P. (2006). Sequence-based species delimitation for the DNA taxonomy of undescribed insects. *Systematic Biology*, *55*(4), 595–609.

Pope, L. C., Liggins, L., Keyse, J., Carvalho, S. B., & Riginos, C. (2015). Not the time or the place: The missing spatio-temporal link in publically available genetic data. *Molecular Ecology*, *24*(15), 3802–3809. https://doi.org/10.1111/mec.13254

Pritchard, J. K., Seielstad, M. T., Perez-Lezaun, A., & Feldman, M. W. (1999). Population growth of human Y chromosomes: A study of Y chromosome microsatellites. *Molecular Biology and Evolution*, *16*(12), 1791–1798. https://doi.org/10.1093/oxfordjournals.molbev.a026091

R Development Core Team. (2016). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Riddle, B. R., Hafner, D. J., Alexander, L. F., & Jaeger, J. R. (2000). Cryptic vicariance in the historical assembly of a Baja California Peninsular Desert biota. *Proceedings of the National Academy of Sciences*, *97*(26), 14438–14443. https://doi.org/10.1073/pnas.250413397

Rissler, L. J. (2016). Union of phylogeography and landscape genetics. *Proceedings of the National Academy of Sciences*, *113*(29), 8079–8086. https://doi.org/10.1073/pnas.1601073113

Rogers, A. R., & Harpending, H. (1992). Population growth makes waves in the distribution of pairwise genetic differences. *Molecular Biology and Evolution*, *9*(3), 552–569.

Sidlauskas, B., Ganapathy, G., Hazkani-Covo, E., Jenkins, K. P., Lapp, H., McCall, L. W., … Kidd, D. M. (2010). Linking big: The continuing promise of evolutionary synthesis. *Evolution*, *64*(4), 871–880.

Soltis, D. E., & Soltis, P. S. (2016). Mobilizing and integrating big data in studies of spatial and phylogenetic patterns of biodiversity. *Plant Diversity*, *38*, 264–270. https://doi.org/10.1016/j.pld.2016.12.001

Thomé, M. T., Zamudio, K. R., Giovanelli, J. G., Haddad, C. F., Baldissera, F. A., & Alexandrino, J. (2010). Phylogeography of endemic toads and post-Pliocene persistence of the Brazilian. *Atlantic Forest*. *Molecular Phylogenetics and Evolution*, *55*(3), 1018–1031. https://doi.org/10.1016/j.ympev.2010.02.003

Thuiller, W., Lafourcade, B., Engler, R., & Araujo, M. B. (2009). BIOMOD—a platform for ensemble forecasting of speciesdistributions. *Ecography*, *32*, 369–373. https://doi.org/10.1111/j.1600-0587.2008.05742.x

Wilson, D. E., & Reeder, D. M. (2005). *Mammal species of the world: A taxonomic and geographic reference*. Baltimore, MD: Johns Hopkins University Press.

Zamudio, K. R., Bell, R. C., & Mason, N. A. (2016). Phenotypes in phylogeography: Species' traits, environmental variation, and vertebrate distribution. *Proceedings of the National Academy of Sciences*, *113*(29), 8041–8048. https://doi.org/10.1073/pnas.1602237113

BIOSKETCH

B.C.C. is an Associate Professor who is interested in developing novel approaches to the collection and analysis of phylogeographic data. Additional information about research in the Carstens Lab is available at https://www.carstenslab.osu.edu. T.A.P. is a postdoctoral researcher interested in combining genetic, geographic, environmental and morphological data to investigate the eco-evolutionary processes shaping current biodiversity patterns. A.E.M. studies evolution of bats by integrating genomic, environmental and phenotypic data. K.E.F. is planning to apply to graduate school in the upcoming year.

Author contributions: B.C.C. and T.A.P. designed the project; B.C.C., T.A.P. and A.E.M. conducted the analyses; K.E.F. and B.C.C. collected and organized the data. All authors assisted in preparing the manuscript.

DATA ARCHIVING

Files used in the acquisition and analysis of these data will be made available upon publication via the Dryad archive. Files are also available at https://carstenslab.osu.edu/software.html.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Carstens BC, Morales AE, Field K, Pelletier TA. A global analysis of bats using automated comparative phylogeography uncovers a surprising impact of Pleistocene glaciation. *J Biogeogr*. 2018;45:1795–1805. https://doi.org/10.1111/jbi.13382