May 2022

Molecular Evolution and Biogeography of the New World Eptesicus Bats

Xueling Yi
University of Wisconsin-Milwaukee

MOLECULAR EVOLUTION AND BIOGEOGRAPHY OF THE NEW WORLD *EPTESICUS* BATS

by

Xueling Yi

A Dissertation Submitted in

Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

in Biological Sciences

at

The University of Wisconsin-Milwaukee

May 2022

ABSTRACT

MOLECULAR EVOLUTION AND BIOGEOGRAPHY OF THE NEW WORLD *EPTESICUS* BATS

by

Xueling Yi

The University of Wisconsin-Milwaukee, 2022
Under the Supervision of Professor Emily K. Latch

Molecular evolution refers to a broad field of studies ranging from microevolution (e.g., population genetics) to macroevolution (e.g., phylogeny), including the bridging field of phylogeography. In natural populations, molecular studies are also combined with biogeography that links biological diversity with geographic distributions to provide a comprehensive understanding of evolutionary processes. The field of molecular evolution has been largely advanced from early exploratory descriptions to statistical tests on biological hypotheses and integrative analyses using sophisticated modeling. However, studies of molecular evolution still face some unresolved questions and challenges, especially in non-model systems. For example, the application of new technology has largely lagged behind in non-model systems, leaving plenty of knowledge gaps that also constrain developments of evolutionary theories. In addition, non-model organisms comprise the majority of global biodiversity and are urgently in need of conservation management, which requires better understanding of their evolution, biogeographic history, and population genetic structure. One of the non-model systems urgently in need of research is bats (order Chiroptera), the second largest mammalian order with > 1,400 globally distributed species. Bats have mysterious evolutionary histories and unique adaptations such as echolocation, powered flight, morphological convergence, adaptation in diverse ecological niches, and tolerance to viruses. The recent spillovers of bat-carrying viruses additionally call for

research on bat ecology, distribution, and evolution, to help predict and control future virus spillovers as well as for better conservation management of bats. In my doctoral dissertation, I studied the molecular evolution and biogeography of the cosmopolitan bat genus *Eptesicus* (family Vespertilionidae) with a focus on the New World species. First, I analyzed the phylogenetic relationships among New World *Eptesicus* species, including the morphological genus *Histiotus* which is endemic to South America and has been found closely related to New World *Eptesicus*. Second, I studied the range-wide nuclear phylogeography of the widespread *Eptesicus* species in North America, the big brown bat (*Eptesicus fuscus*). Third, I estimated the effects of nonrandom missing data on the population genetic structure inferred by the Principal Component Analysis (PCA). I found that the Old World *Eptesicus* bats most likely colonized the New World via the trans-Atlantic route from North Africa to the northern Neotropics in early to mid-Miocene. Cryptic diversity was indicated in the Neotropics, and the *Histiotus* species were found more closely related to *Eptesicus fuscus*. I found that distribution shifts of the North American *E. fuscus* during the Pleistocene glaciation cycles might have initiated the phylogeographic divergence shown by mitochondrial and nuclear DNA as well as morphological subspecies. On the other hand, strong secondary gene flow might have been merging the once diverged western phylogeographic lineages. In addition, I found that the population structure illustrated by PCA could be misinterpreted when using mean imputation of large amounts of nonrandom missing data, which could be common in non-model systems. I found that individuals biased with high amounts of missing data would be dragged towards the PCA origin and could be indistinguishable from truly admixed individuals. Accordingly, my dissertation research on *Eptesicus* bats covered a broad spatial-temporal scale to study their evolutionary history, biogeographic divergence, and inform conservation management. I showcase how the

application of genomics and integrative analyses in non-model systems can shed new light on our empirical as well as theoretical understanding of evolution and biodiversity.

TABLE OF CONTENTS

PAGE

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

First and foremost, I thank my advisor, Dr. Emily K. Latch, for offering me this opportunity to study abroad and do research on bats. Dr. Latch has been extremely accommodating and supportive throughout the years of my graduate education. I am grateful for her giving me full space to explore my own interest and encouraging me to build my career as a scientist. Dr. Latch has generously shared with me her knowledge about scientific research as well as experiences working in academia, which I am highly appreciated. Dr. Latch has also taught me the importance of being confident and optimistic, especially in scientific research where criticism has been the most common feedback. Lastly, I am so fortunate to have Dr. Latch as my advisor during the global pandemic, when her kindness, tolerance, and understanding have greatly supported me to survive and perhaps thrive as much as possible.

I thank my committee members who have nurtured me together. I thank Dr. Filipe Alberto for sharing his knowledge in species distribution modeling, R programming, and ecological research. I thank Dr. Peter Dunn for guiding me in research and asking great questions that always lead me to a step forward. I thank Dr. Rafael Rodríguez for sharing with me his experiences on scientific writing and conceptualization of biological hypotheses. I thank Dr. Gerlinde Höbel for her inspiring comments on my projects and data visualization. Thanks to all my committee members (including my advisor) for giving insightful and constructive comments on my drafts, and for generously spending their time on long conversations with me when I got confused in research.

Massive thanks to a long list of people and museums who have made my dissertation projects possible by generously sharing bat samples with me. First of all, thanks to Dr. Maarten Vonhof from Western Michigan University, who has generously shared a large collection of big

brown bat samples in support of my research. I thank the following researchers and institutions for their sharing of samples: Carly Malavé, USGS National Wildlife Health Center; Dale Paulson, Arkansas Department of Health Rabies Lab; Deahn Donner, Northern Research Station of the USDA Forest Service; Devaughn Fraser, California Department of Fish and Wildlife; Jeffrey Lorch, USGS National Wildlife Health Center; Julie Weckworth, USDA Forest Service National Genomic Center for Wildlife and Fish Conservation; Mammoth Cave National Park (MACA, Permit# MACA-2019-SCI-0015); MacKenzie Hall, New Jersey Division of Fish and Wildlife; Rolan Davis, Rabies Laboratory, Kansas State University; Sharon Messenger, California Department of Public Health. Thanks to the following museums for their generous loan of bat samples in support of my research: American Museum of Natural History (AMNH); Angelo State Natural History Collections (ASNHC); Denver Museum of Nature and Science (DMNS); Florida Museum of Natural History (FLMNH); Field Museum of Natural History (FMNH); Louisiana State University Museum of Natural Science (LSUMZ); University of New Mexico, Museum of Southwestern Biology (MSB); University of California-Berkeley Museum of Vertebrate Zoology (MVZ); Museum of Texas Tech University (NSRL); Royal Ontario Museum (ROM); the University of Arizona, Museum of Natural History (UAZ); Natural History Museum of Utah (UMNH); University of Montana Philip L. Wright Zoological Museum (UMZM); Smithsonian Institution National Museum of Natural History (USNM). In addition, I thank the following museum curators and coordinators for their accommodation and assistance in my sampling process: Adam Ferguson and Bruce Patterson from FMNH, Burton Lim from ROM, Ingrid Rochon from USNM, Melanie Bucci from UAZ, and many other people who have kindly helped me in the sampling process.

I thank Dr. Wes Larson and Dr. Kristen Gruenthal for training me on the library preparation of restriction site-associated DNA sequencing. Thanks to Dr. Anderson Feijó for his help in the morphological identification of museum specimens of *Histiotus* and his insights on the systematics.

Thanks to UWM staff and professors who are not on my committee but have shared with me their knowledge of scientific research and helped me on my projects. I thank Dr. Linda Whittingham for her wise advice and comments on my research projects. Thanks to Dr. Erica Young for teaching me scientific illustration and communication. I thank Dr. Jeffrey Karron for his inspirations on broad-scope evolutionary studies. Thanks to Jason Bacon, Darin Peetz, and Daniel Siercks for their help with the use of UWM High Performance Computing clusters. Special thanks to our graduate program coordinator, Rhianna Miles, for being a super helpful resource to keep track of progresses and paperwork in graduate school.

I thank members of Latch Lab for their massive help both in research and in my life in Milwaukee: Anaïs Tallon, Andrea Howells, Bennett Hardy, Brielle Shortreed, Chandika Rani Ganesh Babu, Genelle Uhrig, Madeline Opie, Margaret Haines, Peter Euclide, Rachel Cook, Rachael Giglio, Samantha Hauser. Special thanks to Dr. Rachael Giglio for her valuable comments and inspiration on my projects, her generous sharing of study and research experiences, and her encouragement to me throughout graduate school. I also thank friends from other labs in graduate school for their help and support, especially: Ambi Henschen, Bretta Speck, Ignacio Escalante Meza, Kane Stratman, Nicholas Sly, Olivia Feagles, and Wendy Semski. Many other friends have also helped and supported me, and I am grateful for having all of you as a warm and accommodating community during graduate school.

Many thanks to the funding sources that have supported me during graduate school. I thank the following student grants for supporting my dissertation research: the Grants-in-Aid of Research of the American Society of Mammalogists, the Theodore Roosevelt Memorial Fund of the American Museum of Natural History, and the Rosemary Grant Advanced Awards of the Graduate Research Excellence Grants of the Society for the Study of Evolution. Thanks to the Northwestern Mutual Data Science Institute for offering me a scholarship on data science research. Thanks to the UWM Graduate School for offering me the Distinguished Dissertation Fellowship and the Graduate Student Excellence Fellowship awards. I thank the UWM Department of Biological Sciences for offering me the following scholarships and awards: the Chancellor's Graduate Student Award, the Clifford H. Mortimer Scholarship, the Ruth I. Walker Memorial Scholarship, and the Joseph G Baier Memorial Scholarship. I also want to thank the generous donors of the above awards whose kindness has been a great support to my graduate studies and research.

Lastly, I thank my family members for their understanding and support to my graduate education. Thanks to my cousin Jin Wang for always being a sweet company and listener. I thank my best friend, Sophia Ma, and friends from undergraduate school for their constant supports. Thanks to my dear mom who has trained me, inspired me, and supported me to become who I am.

Chapter I. Introduction

Evolutionary biology has been revolutionized by the advent of molecular technologies and powerful computational analyses. Molecular data, such as genomic sequencing, provide unprecedented power to estimate biological diversity even on the individual level. Computational biology, such as bioinformatics and modeling, provides additional strength to process big datasets and carry out integrative analyses to shed light on the evolutionary processes underlying the generation and maintenance of diversity. Molecular evolution thus refers to a broad range of studies from microevolution (e.g., population genetics) to macroevolution (e.g., phylogeny), including the bridging field of phylogeography. Molecular studies in natural populations are also increasingly combined with biogeography that links biological diversity with ecology and geographic distributions to provide a comprehensive understanding of evolutionary processes.

The field of molecular evolution is also evolving with the development of technology, analytical approaches, and theoretical frameworks. Early studies tend to be more exploratory and descriptive (Avise 2000), and their characterized molecular patterns provided important foundations for following studies to conceptualize and test biological hypotheses statistically, such as using sophisticated modeling (Knowles 2009). The field of molecular evolution and biogeography also becomes additionally integrative by incorporating complementary analyses and multidisciplinary data such as genomics, ecological traits, landscape, and climatic variables. Such integrative studies can span broad spatial and temporal scales to better understand the underlying evolutionary mechanisms and processes that generate the observed biodiversity.

Despite the exciting revolution and advances, the field of molecular evolution still faces some unresolved questions and challenges, especially in non-model systems. For example, new technology has been mostly developed in model organisms, such as humans, while their

applications in non-model organisms have lagged behind, leaving plenty of knowledge gaps and evolutionary stories that await to be told. This biased knowledge of biological systems further constrains our consideration of alternative evolutionary processes. In other words, the conventional theories, null assumptions, and handy biological hypotheses may not be fully representative across taxonomic groups, and their extrapolation needs to be empirically tested. Importantly, non-model organisms comprise the majority of global biodiversity, making their biological studies urgently needed for better conservation management in the face of human-induced climate change and the debated sixth mass extinction in the Anthropocene (Ceballos et al. 2015). In my thesis, I approached some of the above problems by: 1) testing alternative biological hypotheses about historical biogeography and evolution, 2) integrating complementary data and analyses to shed light on eco-evolutionary dynamics under climate change, and 3) illustrating challenges of studies on natural populations to improve the application of molecular evolution in non-model systems.

Bats (order Chiroptera) are ideal for studying molecular evolution and biogeography for a number of reasons. First, bats have a global distribution (excluding Antarctica) and a high level of diversity as the second largest mammalian order with >1,400 recognized species (Fenton & Simmons 2015). Second, bats have a mysterious evolutionary history and unique adaptations such as echolocation, powered flight, and tolerance to viruses. Bats also occupy extremely diverse ecological niches and thus play important ecological roles across global communities, such as pest control, seed dispersal, and pollination (Kunz & Fenton 2005). Third, bats are difficult to study using traditional morphology and field work because of their nocturnal characteristics, cryptic behavior, and morphological convergence. Therefore, molecular evolution and biogeography have greatly improved our understanding of the evolution and

diversity of bats (Burland & Wilmer 2001; Teeling et al. 2005; Peixoto et al. 2018). However, plenty of unresolved questions and knowledge gaps remain in bats, including their evolutionary origin, global speciation processes, ecological characteristics, and contemporary diversity. The recent spillovers of bat-carrying viruses additionally call for research on bats, not only about their immunity and physiology, but also their ecology, distribution, and evolution, all of which are important for the prediction and control of future virus spillovers as well as better conservation management of bats (Letko et al. 2020; MacFarlane & Rocha 2020). In my dissertation thesis, I studied the molecular evolution and biogeography of the cosmopolitan bat genus *Eptesicus*, a group of insectivorous microbats in the most speciose bat family Vespertilionidae. My thesis focused on the New World species and includes projects of systematics, range-wide phylogeography, and population genetics, which covers a broad geographic range (nearly global) and an evolutionary time scale (from Eocene to the near future 2070). I showcase how the application of genomics and integrative analyses in non-model systems can shed new light on our empirical as well as theoretical understanding of evolution and biodiversity.

Chapter two examines the molecular phylogeny of *Eptesicus*, including the morphological genus *Histiotus* that is endemic in South America and has been found closely related to the New World *Eptesicus* but in unresolved phylogenetic relationships (Hoofer & van den Bussche 2003; Roehrs et al. 2010). I studied the colonization route between Old World and New World and test the on-land versus trans-marine dispersal hypotheses in these volant terrestrial mammals. Using extensive taxonomic and geographic sampling and high-power genomic data from thousands of ultra-conserved elements (UCEs), I identified four major clades in the New World and a novel topology of *Histiotus* and *E. fuscus* being sister clades that together diverged from two sister

clades of Neotropical *Eptesicus*. Historical biogeographic reconstruction identified a Neotropical origin of the New World clades and thus supported the trans-Atlantic colonization route, most likely from North Africa to the northern Neotropics. Divergence time estimations further indicated that the Miocene climatic events and hurricanes might have facilitated the long-distance dispersal of these bats, and possibly other taxonomic groups that diverged during the same time. The updated phylogenies also highlighted the Neotropical cryptic diversity that calls for taxonomic re-evaluation in future research. This study provides an empirical example of trans-marine dispersal promoting global colonization and diversification in terrestrial animals, in addition to the default assumption of on-land dispersal in biogeographic histories.

Chapter three examines the range-wide phylogeography of the only *Eptesicus* species identified north of Mexico, the big brown bat (*E. fuscus*). Big brown bats are widely distributed from southern Canada to northern South America, including most of the Caribbean Islands, with up to 11 identified morphological subspecies (Kurta & Baker 1990). Previous studies found mitochondrial divergence consistent with subspecies distributions but a lack of nuclear population structure, which was explained by mammalian male-biased gene flow homogenizing nuclear genomes (Turmelle et al. 2011). However, the lack of nuclear divergence could also result from limited analytical power. I hypothesized that population divergence in big brown bats had been shaped by historical isolation during climate change, such as the Pleistocene glaciation, and that such divergence signals could be detected in nuclear genomes despite secondary gene flow. Using genome-wide markers from the restriction site-associated DNA sequencing (RADseq), I characterized the fine-scale nuclear phylogeographic pattern that was overall consistent with mitochondrial and morphological divergence. However, discordances were found among nuclear trees and between mitochondrial and nuclear topologies, indicating strong effects

of (but not sex-biased) gene flow on the shallow within-species phylogeography. Integrative analyses of population genetics, networks, and demographic modeling demonstrated a complex evolutionary history of Pleistocene/Holocene isolation followed by secondary gene flow that has been merging once diverged lineages. Distribution modeling under future climate change predicted further northward range expansion and habitat loss especially on Caribbean Islands, indicating the importance of conservation management of this widespread species (Agosta 2002). Both climate change and gene flow will continue influencing the Anthropocene biodiversity. My work provides an empirical example of the within-species dynamics that could shed light on the speciation process.

Chapter four focuses on missing data from next-generation sequencing of non-model systems, a common problem that can bias result interpretation and even mislead downstream analyses or conservation management practices. Here I studied the effects of missing data on the Principal Component Analysis (PCA), a type of multidimensional analysis that has been widely used to characterize and visualize genetic relationships among individuals or populations. Because PCA does not tolerate missing data, it is conventional in population genetics to impute missing data with mean values (e.g., default in the R package adegenet). However, it is unclear how the mean imputed missing data might affect PCA plots and the interpreted genetic relationships, which has been overlooked in many empirical studies of non-model systems where missing data tend to be unavoidable due to variable sample quality and quantity. I simulated genetic datasets under various biological scenarios and incorporated different types (random, individual-biased, population-biased) and amounts of missing data for PCA estimation. I showed that when missing data are nonrandom across individuals, mean imputation would drag the individuals biased with high missingness towards the origin of the PCA plot, making them indistinguishable from the

true admixed individuals and potentially resulting in misinterpreted genetic relationships. Such effects of nonrandom missing data were also demonstrated using my empirical RADseq datasets from big brown bats (*Eptesicus fuscus*), where low-quality samples with high amounts of missing data were dragged away from their true population clusters towards the PCA origin. Based on these results, I suggested ways to detect missing data effects on PCA (e.g., color coding individuals by missing values) and better interpret population genetic structure (e.g., using various filtering strategies and complementary approaches) in non-model systems.

In summary, my research spans a wide spatial-temporal scale to demonstrate evolutionary processes that generate the observed biodiversity, using bats as empirical examples. My studies improved our understanding of the evolutionary history of *Eptesicus* bats and informed their conservation management, which provides important foundations for future research in these taxa. I showcased how the application of genomics and integrative analyses in non-model systems could help test biological hypotheses empirically and shed new light on evolutionary theories. Lastly, beyond the pursuit of scientific research, these projects reflect a personal passion for studying natural populations to embrace biodiversity and elucidate the beauty of nature.

References

Agosta, S. J. (2002). Habitat use, diet and roost selection by the Big Brown Bat (*Eptesicus fuscus*) in North America: a case for conserving an abundant species. *Mammal Review*, *32*(3), 179–198.

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

Burland, T. M., & Wilmer, J. W. (2001). Seeing in the dark: molecular approaches to the study of bat populations. *Biological Reviews*, *76*(3), 389–409.

Ceballos, G., Ehrlich, P. R., Barnosky, A. D., García, A., Pringle, R. M., & Palmer, T. M. (2015). Accelerated modern human-induced species losses: Entering the sixth mass extinction. *Science Advances*, *1*(5).

Fenton, M. B., & Simmons, N. B. (2015). *Bats: a world of science and mystery*. The University of Chicago Press.

Hoofer, S. R., & van den Bussche, R. A. (2003). Molecular Phylogenetics of the Chiropteran Family Vespertilionidae. *Acta Chiropterologica*, *5*(suppl), 1–63.

Knowles, L. L. (2009). Statistical Phylogeography. *Annual Review of Ecology, Evolution, and Systematics*, *40*(1), 593–612.

Kurta, A., & Baker, R. H. (1990). *Eptesicus fuscus*. *Mammalian Species*, *356*(356), 1–10.

Letko, M., Seifert, S. N., Olival, K. J., Plowright, R. K., & Munster, V. J. (2020). Bat-borne virus diversity, spillover and emergence. *Nature Reviews Microbiology 2020 18:8*, *18*(8), 461–471.

MacFarlane, D., & Rocha, R. (2020). Guidelines for communicating about bats to prevent persecution in the time of COVID-19. *Biological Conservation*, *248*, 108650.

Peixoto, F. P., Braga, P. H. P., & Mendes, P. (2018). A synthesis of ecological and evolutionary determinants of bat diversity across spatial scales. *BMC Ecology 2018 18:1*, *18*(1), 1–14.

Roehrs, Z. P., Lack, J. B., & van den Bussche, R. A. (2010). Tribal phylogenetic relationships within Vespertilioninae (Chiroptera: Vespertilionidae) based on mitochondrial and nuclear sequence data. *Journal of Mammalogy*, *91*(5), 1073–1092.

Teeling, E. C., Springer, M. S., Madsen, O., Bates, P., O'Brien, S. J., & Murphy, W. J. (2005). A molecular phylogeny for bats illuminates biogeography and the fossil record. *Science*, *307*(5709), 580–584.

Turmelle, A. S., Kunz, T. H., & Sorenson, M. D. (2011). A tale of two genomes: contrasting patterns of phylogeographic structure in a widely distributed bat. *Molecular Ecology*, *20*(2), 357–375.

Chapter II. Systematics of the New World bats *Eptesicus* and *Histiotus* indicate trans-marine dispersal followed by Neotropical cryptic diversification

Xueling Yi[1] and Emily K. Latch[1]

[1] Department of Biological Sciences, University of Wisconsin-Milwaukee, Milwaukee, Wisconsin, 53211, USA

Abstract

Biodiversity can be boosted by colonization of new habitats, such as different continents and remote islands. Molecular studies have suggested that recently evolved organisms probably colonized already separated continents by dispersal, either via land bridge connections or crossing the ocean. Here we test the on-land and trans-marine dispersal hypotheses by evaluating possibilities of colonization routes over Bering land bridge and across the Atlantic Ocean in the cosmopolitan bat genus *Eptesicus* (Chiroptera, Vespertilionidae). Previous molecular studies have found New World *Eptesicus* more closely related to *Histiotus*, a Neotropical endemic lineage with enlarged ears, than to Old World *Eptesicus*. However, phylogenetic relationships within the New World group remained unresolved and their evolutionary history was unclear. Here we studied the systematics of New World *Eptesicus* and *Histiotus* using extensive taxonomic and geographic sampling, and genomic data from thousands of ultra-conserved elements (UCEs). We estimated phylogenetic trees using concatenation and multispecies coalescent. All analyses supported four major New World clades and a novel topology where *E. fuscus* and *Histiotus* are sister clades that together diverged from two sister clades of Neotropical *Eptesicus*. Intra-clade divergence indicated cryptic diversity that has been concealed by morphological features, especially in the Neotropics where taxonomic re-evaluations are warranted. Molecular dating estimated that Old World and New World clades diverged around 17 million years ago followed by radiation of major New World clades in the mid-Miocene, when climatic changes might have facilitated global dispersal and radiation events. Biogeographic ancestral reconstruction supported the Neotropical origin of the New World clades, indicating a trans-Atlantic colonization route from North Africa to the northern

Neotropics. We highlight that trans-marine dispersal may be more prevalent than currently acknowledged and may be an important first step to global biodiversification.

Introduction

Diversification can be boosted when organisms colonize novel habitats that have been previously unreachable, such as remote islands and separated continents. Novel habitats provide the new arrivals with open niches that are both opportunities and challenges, possibly resulting in adaptation and rapid speciation. How organisms colonized novel habitats in the first place is thus an important question for understanding their diversification and evolutionary history. The classic vicariance theory explains that cosmopolitan taxa reached their global distribution during continental breakups. However, molecular estimates of diversification times that are after continental breakups indicated that more recently evolved organisms, such as the majority of terrestrial mammals, probably reached their global distribution by dispersal (Upchurch 2008). Two mechanisms have been hypothesized for the long-distance dispersal between separated lands. The on-land dispersal hypothesis proposes that organisms moved between continents via land connections, such as the Bering land bridge connecting Eurasia and North America (Jiang et al. 2019), and the Isthmus of Panama connecting North and South Americas (Bacon et al. 2016). Alternatively, the trans-marine dispersal hypothesis proposes that organisms colonized remote islands or continents by crossing the ocean via mechanisms such as seed drifting, rafting, swimming, and flying (de Queiroz 2005). For example, successful trans-marine dispersal has been reported in the colonization of Caribbean islands by lizards (Censky et al. 1998), the dispersal among Southeast Asian islands by fruit bats (Tsang et al. 2020), and trans-Atlantic dispersal by angiosperms (Renner 2004), land snails (Uit de Weerd & Gittenberger 2013), insects (Lovejoy et al. 2005; Murray & Heraty 2016), birds (Batista et al. 2020), primates (Bond et al. 2015), and rodents (Rowe et al. 2010). Such long-distance dispersal over water also happens at present, such as the remarkable journey between the African wintering grounds and

the North American Arctic breeding grounds of northern wheatears (*Oenanthe oenanthe*, Bairlein et al. 2012). However, despite accumulating evidence of trans-marine dispersal, on-land dispersal routes remain the default assumption in historical biogeography, possibly biased by human nature as terrestrial animals, and thus the prevalence of trans-marine dispersal in the evolutionary history is possibly under-estimated.

Successful trans-marine dispersal has been associated with certain ecological and/or physiological features, such as the wind-dispersal mechanism in plants (Munoz et al. 2004; Renner 2004), the tolerance of saltwater in lizards (Hsu et al. 2021), and the ability of long-distance flight in birds (Hosner et al. 2017) and bats. Bats (Order Chiroptera) are the only mammalian group that has powered flight which probably has facilitated their global distribution and speciation. Divergence of extant bat lineages was estimated around 65 million years ago (Teeling et al. 2005), well after continental breakups, indicating a global distribution achieved by dispersal rather than vicariance. However, although trans-marine dispersal may be feasible for bats due to their abilities of powered flight, bat dispersal might still be constraint on land by their ecological needs such as reliance on terrestrial roosting sites, food availability, and access to fresh water (Kunz & Fenton 2005). Therefore, bats are a good model system for testing the above dispersal hypotheses, both of which have been proposed in previous studies of bat global colonization. For example, studies in the genus *Myotis* suggested that the Old World common ancestor colonized the New World via the Bering land bridge (Stadelmann et al. 2007; Ruedi et al. 2013), while a study using parsimonious ancestral state reconstruction suggested that quite a few bat lineages colonized the New World via trans-Atlantic dispersal (Lim 2009). Additionally, with more than 1,400 species distributed in various environments worldwide (Fenton &

Simmons 2015), bats provide a great research opportunity to study whether and how long-distance dispersal and colonization of novel habitats might boost global diversification.

Here we focused on the genus *Eptesicus* (family Vespertilionidae), a cosmopolitan group of short-eared insectivorous bats. Previous phylogenetic studies of the Old World *Eptesicus* have found extensive cryptic diversity (Goodman et al. 2012; Juste et al. 2013; Koubínová et al. 2013; Amador et al. 2018) while relationships among the New World species remain ambiguous. Interestingly, molecular phylogenies showed that New World species were more closely related to the morphological genus *Histiotus* than to the Old World species, leading to the taxonomic suggestion of three subgenera *Cnephaeus* (Old World), *Eptesicus* (New World), and *Histiotus* (Hoofer & van den Bussche 2003; Roehrs et al. 2010). However, the obviously enlarged ears in the South American endemic *Histiotus* bats (Thomas 1916; Nowak & Walker 1994) made most subsequent studies continue to treat *Histiotus* as a unique genus (e.g., Handley & Gardner 2008; Feijó et al. 2015; Díaz et al 2019; Rodríguez-Posada et al. 2021; Velazco et al. 2021). Regardless of the taxonomy, unresolved phylogenies of the New World *Eptesicus* and *Histiotus* also indicated different colonization routes and diversification processes. On-land dispersal from Eurasia to Nearctic via Bering land was supported by mitochondrial phylogenies where Nearctic *E. fuscus* was the basal lineage of the New World clade (Hoofer & van den Bussche 2003; Roehrs et al. 2010; Amador et al. 2018). On the other hand, trans-Atlantic dispersal from Old World to the Neotropics was supported by the nuclear phylogeny where South American *Histiotus* was the basal lineage (Roehrs et al. 2010), or the possible topology where Neotropical *Eptesicus* was the basal lineage (e.g., Lim 2009). The lack of node resolution and weak statistic supports in previous phylogenetic studies may result from an evolutionary history of rapid early diversification in the family Vespertilionidae combined with limited sampling of taxonomic,

geographic, and genomic diversity (Hoofer & van den Bussche 2003; Lack & van den Bussche 2010). Therefore, extensive sampling and higher analytical power are required to estimate phylogenetics of the New World *Eptesicus* and *Histiotus* and to understand their evolutionary history.

Accordingly, we tested the on-land versus trans-marine dispersal hypotheses in the cosmopolitan bat genus *Eptesicus* and studied their New World diversification using a systematic approach. We collected as many nominated species as possible and incorporated extensive geographic sampling to help detect cryptic diversity. We used ultra-conserved elements (UCEs) as the genomic marker to gain higher analytical power. UCEs are variable sequences flanking thousands of orthologous nuclear regions that are conserved across taxonomic groups (Faircloth et al. 2012). UCEs have proven to be powerful for phylogenomic studies across diverse taxa such as insects (Blaimer et al. 2015), fish (Alda et al. 2019), snakes (Blair et al. 2019), frogs (Guillory et al. 2020), birds (Smith et al. 2014a), and mammals (Esselstyn et al. 2017) including the bat genus *Myotis* (Platt et al. 2018; Morales et al. 2019). The multi-locus data generated from hundreds to thousands of UCEs provide reliable estimates of phylogenies and work better to capture the species evolutionary history than single-locus data such as mitochondrial genomes (Platt et al. 2018). We constructed phylogenetic trees using concatenation as well as multispecies coalescent methods, and further incorporated divergence time estimation and ancestral state reconstruction to understand the evolutionary history of *Eptesicus* bats. Our study also sheds light on the broader mechanisms of long-distance dispersal promoting global biodiversification.

Materials and Methods

Taxonomic sampling

We obtained samples of 95 individual bats from researchers and museums (Appendix A Table S1, Fig S1). Tissue samples (frozen or preserved in ethanol) were requested if possible; otherwise, we collected one piece (about 2x3 mm) of wing skin per individual from dry museum specimens. We collected as many nominated *Eptesicus* species as possible and multiple individuals per species representing different subspecies or geographic populations to help detect cryptic diversity (Fig 1.1). Species identification was provided by source collections (i.e., museums or researchers). Based on the taxonomy in Bat Species of the World (Simmons & Cirranello 2020, https://batnames.org, accessed August 2021), our samples included approximately 8 of 10 nominated species from the New World subgenus *Eptesicus*, 6 of 8 species from *Histiotus*, and 6 of 16 species from the Old World subgenus *Cnephaeus*. Genus *Eptesicus* also includes a subgenus *Rhinopterus* with only one nominated species in the Old World (*E. floweri*; Simmons & Cirranello 2020) but we were not able to include it in the current study. We also collected 14 individual samples representing 12 outgroup species including two species that have been re-assigned from *Eptesicus* to genera *Laephotis* and *Rhyneptesicus*; a species from the most closely related genus *Scotomanes*; species from other genera (*Lasionycteris*, *Nycticeius*, *Glauconycteris*, *Rhogeessa*, *Antrozous*) in the subfamily Vespertilioninae; a *Myotis* species and a *Kerivoula* species from closely related subfamilies within family Vespertilionidae; and a *Miniopterus* species from the closely related family Miniopteridae (Hoofer & van den Bussche 2003; Miller-Butterworth et al. 2007; Lack & van den Bussche 2010; Roehrs *et al*. 2010; Amador et al. 2018).

UCE enrichment and bioinformatics

Genomic DNA was extracted from contemporary samples using the Qiagen DNeasy Blood & Tissue Kit, and from museum samples using an optimized aDNA protocol of phenol-chloroform extraction (Alminas et al. 2021). The extracted DNA was quantified using a Qubit fluorometer, concentrated into 1ug DNA per sample (all DNA concentrated if <1ug), and sent to Arbor Biosciences (Ann Arbor, MI, USA) for target sequencing following their myBaits Manual v4 protocol. Each sample was prepared into a standard double-stranded DNA library (ds library, if about 1ug DNA) or a single-stranded DNA library (ss library if <500 ng DNA; Appendix A Table S1). The ss library preparation was applied to low-quality or low-quantity DNA samples to improve efficiency for high-throughput sequencing (Gansauge et al. 2017). We replicated an individual using a ds library-prepared sample (about 1ug DNA) and an ss library-prepared sample (about 580 ng) to reassure comparable performance of the two library types. Accordingly, our experiments included 96 libraries that were dual-barcoded and pooled in sets of eight (sets of four if extractions from skin samples). Prepared libraries were then enriched for UCEs using the Tetrapods 5Kv1 probe set (5,472 baits targeting 5,060 UCEs, Faircloth et al. 2012) and sequenced at equal depths (about one Gb per library) for 150 bp paired-end reads using Illumina Novaseq 6000.

Demultiplexed sequencing data were obtained from Arbor Biosciences and processed in illumiprocessor 2.0.9 (Faircloth 2013) and Trimmomatic (Bolger et al. 2014) to trim adaptor sequences and low-quality bases (default threshold phred 33) into reads in minimum 40 bp. Trimmed reads were assembled into sample-specific contigs using SPAdes (Prjibelski et al. 2020) implemented in PHYLUCE v1.6.8 (Faircloth 2016). The following processes were done in PHYLUCE v1.7.0 with default settings unless stated otherwise.

The orthologous nature of UCEs allowed us to incorporate data from previous studies that used the same enrichment baits. Here we added four individuals sequenced by Platt et al. (2018) including two *E. fuscus* (eptesicus-fuscus-DAR4; eptfus1-genome-GCA000308155-1) and two *Myotis* (myoluc2-genome-GCA000147115-1; myotis-horsefeldi-1926039) whose contig assemblies were downloaded from the Dryad archive (https://doi.org/10.5061/dryad.5g205). The assembled contigs of total 100 samples (96 new plus 4 previous) were mapped to the tetrapods-UCE-5Kv1 probes (Faircloth et al. 2012) to identify UCE loci and remove duplications. The identified UCEs were extracted using the PHYLUCE command *phyluce_assembly_get_match_counts*, and the number of UCEs per sample was estimated using the command *phyluce_assembly_get_fastas_from_match_counts*. Individuals identified with <500 UCEs (less than 10% of the targeted loci) were removed as failed enrichments, probably due to low DNA qualities and quantities. The replicated sample generated comparable results using ds and ss library preparations (supplementary text), and thus we also removed the replicate that had less input DNA. Sequences of the remaining samples were aligned using MAFFT and internally trimmed using GBLOCKS as implemented in PHYLUCE. UCEs present in at least 75% individuals were output for downstream analyses.

Individual phylogenetic relationships

The output UCEs were concatenated in PHYLUCE. We used the single partition scheme in concatenation analyses due to computational constraints (run time and memory) and previous findings that partitioning does not impact tree topology or biological inference (Platt et al. 2018; Alda et al. 2019; Blair et al. 2019; but see Tagliacollo & Lanfear 2018). In addition, it remains unclear whether partition of UCE data in vertebrates is appropriate or beneficial (Guillory et al.

2020; Portik & Wiens 2021). The maximum likelihood (ML) phylogeny was conducted in RAxML v8.2.12 (Stamatakis 2014). A single complete analysis (command -f a) was run to conduct 100 rapid bootstraps and a thorough ML search using the GTRGAMMA model, and trees were rooted (constraint command -o) by the outgroup individual of *Miniopterus natalensis*. The best-scored ML tree was visualized in R 4.0 (R Core Team 2021) using the package ggtree (Yu 2020). The Bayesian inference (BI) analyses were conducted in ExaBayes v1.5.1 (Aberer et al. 2014) in two independent runs each having two coupled chains executed in parallel. All chains were run for 2 million generations with 25% burnin, parsimonySPR of 8, likeSpr of 4, branchMulti of 8, and blDistGamma of 6. The average standard deviation of split frequencies (SDSF) among all chains were estimated using the sdsf command, and the summary statistics of parameters were estimated using the postProcParam command in ExaBayes. Convergence of runs was checked by average SDSF <5% (average 0.23% in our data) and the effective sampling size (ESS) >200 for all parameters (ESS >500 in our data). Run results were built into consensus trees using the extended majority rule.

In addition to concatenation methods, we also used coalescent to account for the gene tree heterogeneity caused by incomplete lineage sorting. We estimated individual phylogenies using the quartet-based coalescent method SVDquartets (Chifman & Kubatko 2014) implemented in PAUP v4 (Swofford 2003). All individuals were treated as independent tips (i.e., no taxonomic assignments) and UCEs were combined in the input data. We did an exhaustive evaluation of all quartets (1,929,501 total) using the multispecies coalescent tree model, the multilocus analysis, and 100 bootstrap replicates. Results were built into consensus trees using the 50% Majority Rule. The BI and quartet-based trees were rooted by *Miniopterus natalensis* and visualized in FigTree (http://tree.bio.ed.ac.uk/software/figtree).

"Species trees" of operational taxonomic units

Individual phylogenetic trees indicated species misidentification of three samples which were excluded from the following analyses for clarity. The PHYLUCE command *phyluce_align_extract_taxa_from_alignments* was used to remove individuals from the 75% complete UCE alignment, and the command *phyluce_align_get_informative_sites* was used to estimate the number of informative sites per UCE in the retained data set. The remaining individuals were assigned into putative operational taxonomic units (OTUs) based on their species identities from source collections, geographic sampling sites, and positions in the individual phylogenetic trees generated above. We then estimated "species trees" on the assigned OTUs using multispecies coalescent in three complementary approaches.

First, the quartet-based method SVDquartets was conducted with individuals (excluding the misidentified ones) assigned into OTUs. All quartets (1,425,790 total) were evaluated and the other commands were the same as described above. Second, we estimated species trees using the summary method ASTRAL-III v5.7.7 (Zhang et al. 2018) which maximizes the number of quartets shared among input gene trees of UCEs. To construct gene trees, we did maximum likelihood analyses on each UCE in RAxML in the same way described above but without the outgroup constraint. We then compiled newick files of best-scored ML trees from all UCEs and the 500, 1000, and 2000 most informative UCEs based on their percent informative sites. The low-supported branches (bootstraps < 10%) in each compile were collapsed using Newick Utilities1.6 (Junier & Zdobnov 2010) to minimize gene tree errors and improve summary accuracy (Zhang et al. 2018). Individuals in gene trees were assigned into OTUs in ASTRAL-III and the species trees were estimated with local posterior probabilities (PP) based on quartet

supports (Sayyari & Mirarab 2016). Third, we used the full search method Bayesian Phylogenetics and Phylogeography (BPP) v4.4 (Flouri et al. 2018) to estimate species trees based on fixed taxonomic assignments (analysis A01; Yang 2015). Due to constraints on computational time and memory, we only used the alignment of the 500 most informative UCEs in BPP analyses and we set the guide tree as the ASTRAL-III topology generated from 500 UCEs. Two independent runs were conducted, each with 10,000 burnin, 2 sample frequency, 50,000 number of samples, the theta inverse-gamma prior (3, 0.004), the tau inverse-gamma prior (3, 0.09), and the other parameters as default. The estimated best trees from the two runs were compared to check consistency. All species trees were visualized in FigTree and rooted by *Miniopterus natalensis*.

Molecular timing

To estimate divergence times, we generated a pruned data set by including only one individual per OTU. When multiple individuals were assigned to the same OTU, we selected the individual that had higher data quality. Alignments of the selected individuals were extracted and concatenated in PHYLUCE using the 500 most informative UCEs or 500 randomly chosen UCEs for comparison. Then, these pruned data sets were used to estimate divergence time in the Bayesian Markov chain Monte Carlo (MCMC) program MCMCTree (Yang & Rannala 2006) implemented in PAML v4.9j (Yang 2007). We fixed the input tree as the ASTRAL-III topology of 500 UCEs and incorporated two soft-bound calibrations based on previous molecular phylogenies. We calibrated the root node as 43-54 million years, representing the most recent common ancestor (MRCA) of families Vespertilionidae and Miniopteridae, and we calibrated the MRCA of subfamilies Myotinae and Vespertilioninae as 20-36 million years (Eick et al. 2005;

Teeling et al. 2005; Miller-Butterworth et al. 2007; Lack & van den Bussche 2010; Agnarsson et al. 2011; Amador et al. 2018). The MCMCTree was run with the clock model of independent rates, the substitution model HKYG5, 1 million burnin, 5000 sample frequency, 20000 number of samples, and default settings for other parameters. To speed up analyses, we used the approximate likelihood calculation to first estimate branch lengths by maximum likelihood and then estimate divergence times using MCMC to approximate likelihoods (Reis & Yang 2011). For each data set (i.e., 500 most informative UCEs or 500 randomly chosen UCEs), we ran two independent MCMC runs and checked convergence based on ESS values using Tracer v1.7.2 (Rambaut et al. 2018) and convergence plots between runs. The timed trees were visualized in FigTree.

Historical biogeography and ancestral reconstruction

Ancestral geographic distribution was reconstructed using the R package BioGeoBEARS (Matzke 2013; Matzke 2014). We included ten biogeographic areas according to the proposed zoogeographic regions and realms (Holt et al. 2013) and our sample distribution (Appendix A Fig S1). We followed the terminology in Holt et al. (2013) where the Neotropical realm was divided into the Amazonian and South American regions, and it should be noted that the Amazonian area includes both the rainforest and the surrounding northern Andes. In addition, we labeled the Caribbean as an independent biogeographic area considering its insular nature. We coded the distribution of OTUs into biogeographic areas based on our sampling localities and the species distribution from the International Union for Conservation of Nature Red List (IUCN, https://www.iucnredlist.org/, accessed October 2021). For instance, the OTU of Continent West *E. fuscus* included samples from the western Nearctic, Panama, and the northern Neotropics, and

thus this OTU was coded by three biogeographic areas. The species *E. bottae* has a Saharo-Arabian distribution on the IUCN Red List but the only sample in our analyses was collected from Kyrgyzstan in the Palearctic area, and thus this OTU was coded by both biogeographic areas. Similarly, our sample putatively representing the species *E. pachyomus* was collected from Pakistan, but this species is also distributed in the Indian subcontinent (e.g., Juste et al. 2013), and thus this OTU was coded as both Saharo-Arabian and Palearctic areas.

BioGeoBEARS was conducted using default parameters (Matzke 2013) and the maximum range size of four. The timed phylogenetic tree from MCMCTree using the 500 most informative UCEs was used as the input tree. In addition, we also used the BPP tree topology, which was different from the other phylogenies, as the input to see if our interpretation differed. The Generalized Simulated Annealing (GenSA) was used for parameter optimization. Ancestral states were reconstructed using all six available biogeographic models, including Dispersal–Extinction–Cladogenesis (DEC), Dispersal-Vicariance Analysis (DIVALIKE), and BAYAREALIKE, and their counterparts with founder events (DEC+J, DIVALIKE+J, BAYAREALIKE+J). The null models estimate dispersal (the d parameter) and extinction (the e parameter) based on continuous geographic range expansion and contraction, which is more similar to expectations of the on-land dispersal hypothesis. On the other hand, the "+J" models additionally incorporate "jump dispersal" (the j parameter) between discontinuous biogeographic areas, which is especially suitable for island biogeography (Matzke 2014) and would represent the trans-marine dispersal hypothesis. Models were compared based on log likelihoods (LnL), the Akaike Information Criterion (AIC) values, and the AIC values with number of parameters controlled (AICc). The model with the lowest AIC and AICc values (or highest LnL) was identified as the better fit.

Results

UCE enrichment and individual phylogenetic relationships

Raw sequences of our 96 samples were trimmed to an average 5,229,362 (range 2,276 – 32,068,016) paired reads per sample. Assemblies generated an average of 138,474 (range 38 – 640,425) contigs per sample with an average length of 243 bp (87-384 bp). A total of 16 samples were removed, including one replicate and 15 failed enrichments (<500 UCE loci; Appendix A Fig S2), and 4 samples were incorporated from Platt et al. (2018), resulting in a data set of 84 individuals. A total of 4,972 UCEs were identified in the 84 samples. Additional filtering by 75% completeness retained 3,611 UCEs (present in >= 63 individuals), with an average alignment length of 307 bp (201 – 733 bp) and 31 (1-116) informative sites per UCE.

The concatenation methods ML and BI generated well supported and highly consistent phylogenies of the 84 individuals (Fig 1.2; Appendix A Fig S3). The only difference between ML and BI trees occurred in the subclade of the Continent East *E. fuscus* (Appendix A Fig S3) and probably reflects strong intra-population gene flow. The coalescent method SVDquartets generated an overall similar tree topology but showed lower statistical supports and limited resolution of recent nodes (Appendix A Fig S4), similar to patterns found in the previous studies comparing concatenation and SVDquartets (e.g., Blair et al. 2019). Multifurcating nodes in the quartet-based tree also indicated gene flow among closely related individuals such as those of *E. fuscus* (Appendix A Fig S4). Despite the minor discordances, all individual phylogenies consistently supported four monophyletic clades corresponding to *E. fuscus*, *Histiotus*, and two cryptic sister clades of Neotropical *Eptesicus* (Fig 1.2). In addition, we found *Histiotus* more closely related to *E. fuscus* than to their sympatric Neotropical *Eptesicus*, and we found two paraphyletic clades of the Old World *Eptesicus,* roughly representing divergence between

Eurasia and South Africa (Fig 1.2; Appendix A Fig S3, Fig S4). Outgroup relationships in our results support that genera *Rhogeessa* and *Antrozous* are more closely related to *Eptesicus* (e.g., Hoofer & van den Bussche 2003; Roehrs et al. 2010) rather than *Myotis* (Miller-Butterworth et al. 2007), consistent with current taxonomy. In addition, all individual phylogenies repeatedly showed misidentification of three samples whose morphological species identities did not match their positions in the molecular phylogeny (individuals in red; Fig 1.2; Appendix A Fig S3, Fig S4). Unfortunately, we were not able to provide morphological examination of these individuals as they were all collected as tissue samples, and additional evidence (e.g., morphology or other genetic markers) is required to further evaluate their species identity. Therefore, we suggested misidentification and putative taxonomy based on our molecular data (Appendix A Table S1) but excluded these individuals from downstream analyses for clarity.

Our results also supported the previous taxonomic updates of elevating *E. serotinus pachyomus* (the sample from Pakistan) and *E. serotinus isabellinus* (the sample from Tunisia) into full species *E. pachyomus* and *E. isabellinus* (Juste et al. 2013), renaming *E. nasutus* into a different genus *Rhyneptesicus* (Juste et al. 2013), and renaming *E. matroka* into a different genus *Laephotis* (following Simmons & Cirranello 2020; but Goodman et al. 2012 suggested genus *Neoromicia*). In addition, our individual phylogenies indicated cryptic diversity that has not been recognized in current taxonomy. Phylogenetic relationships of closely related individuals seemed more consistent with their geographic localities than morphological species identities. Therefore, we assigned the remaining 81 individuals into 54 operational taxonomic units (OTUs) based on species identities, geographic localities, and their positions in the individual phylogeny (Fig 1.2).

Species trees of putative taxonomic units

Species trees of the 54 OTUs varied slightly among methods and data sets but all supported the four major New World clades and the same topology of *E. fuscus* and *Histiotus* being sister clades that together diverged from two sister clades of the Neotropical *Eptesicus* (Fig 1.3). Although we used the conventional term "species tree", it should be noted that the OTUs do not necessarily represent valid species but could be subspecies or geographic populations (such as the continental lineages of *E. fuscus*). Therefore, putative gene flow among closely related OTUs might have resulted in the lower intra-clade statistical supports and the discordances among species trees (Fig 1.3). However, we decided to use the slightly over-split OTUs to fully represent the putative cryptic diversity that might have been concealed by unclear species identities based on morphology.

The species tree generated by the summary method ASTRAL-III using the 500 most informative UCEs (Fig 1.3A) is most consistent with individual phylogenies, species trees generated by other methods, and our expectations, and thus we relied on this tree in the following analyses and discussion. The ASTRAL-III analyses on four input data sets generated highly similar results with slight differences (Appendix A Fig S5). Analyses using 1000 and 2000 most informative UCEs showed different positions of *E. fuscus hispaniolae* (Dominican Republic), *E. serotinus* (Iran), and *E. bottae*; analyses using all UCEs showed additional discordances in the positions of *E. fuscus dutertreus* (Bahamas), *E. fuscus* (Jamaica), *H. velatus* (Peru), and *E. furinalis* (Belize, Nicaragua). Although ASTRAL-III analyses using more UCEs (i.e., more input gene trees) had relatively higher supports for recent nodes, the summary tree using the 500 UCEs had the highest supports for more ancient nodes (Appendix A Fig S5). These minor discrepancies may result from the competition between a gain of information and an increase of noise/error when more gene trees are included into the summary method (Zhang et al. 2018). The species tree from

SVDquartets included two multifurcating nodes and differed from the summary tree mainly in positions of closely related OTUs (Fig 1.3B). The two independent runs of BPP generated highly similar best trees and thus we only present the result that had higher probabilities ($p = 0.386$ versus $p = 0.166$). The BPP species tree is almost identical with the summary tree (which was used as the guide tree) but indicated monophyly of the Old World clades (Fig 1.3C), which differed from all other analyses in our study. Our estimated topology of *E. fuscus* and *Histiotus* being sister clades and Old World *Eptesicus* being paraphyletic is different from previous studies (Hoofer & van den Bussche 2003; Roehrs et al. 2010; Juste et al 2013; Amador et al. 2018), possibly because our more extensive taxonomic and geographic sampling helped balance the New World phylogeny and break up long branches (such as the *E. fuscus* lineage in previous phylogenies). We also used more powerful data from thousands of UCEs distributed across the nuclear genome to provide higher resolution in phylogenetic analyses. However, it should be noted that the node leading to *E. fuscus* and *Histiotus* was relatively less supported in all analyses (except BPP), indicating that this divergence may not be fully resolved.

Divergence time and the historical biogeography

For each of the pruned data sets, the two independent runs of MCMCTree showed convergence (ESS > 200 for all parameters) and highly consistent estimates of posterior mean divergence times. Thus we only presented results of the run that had relatively higher ESS values using the 500 most informative UCEs (Fig 1.4) and the 500 randomly chosen UCEs (Appendix A Fig S6). The overall results were similar using the two data sets. However, the 500 most informative UCEs appeared to generate better estimations because the mean posterior divergence times were slightly shorter at the more recent nodes, and the 95% highest posterior density (HPD) was

narrower, compared to corresponding estimations using the 500 randomly chosen UCEs (Appendix A Fig S7). Therefore, in the following analyses and interpretation, we focused on the results from the 500 most informative UCEs.

Divergence between the New World and Old World *Eptesicus* was estimated around 17.38 million years ago (mya) with the 95% HPD 11.19 – 23.88 mya, consistent with the previous estimates of mean 16.97 mya divergence time (Amador et al. 2018). Clades of *E. fuscus* and *Histiotus* were estimated to diverge from Neotropical *Eptesicus* around 14.82 mya (95% HPD 8.99 - 20.66 mya), relatively older than the previous estimates of mean 11 mya based on a different underlying topology (Amador et al. 2018). The sister clades *E. fuscus* and *Histiotus* diverged around 13.8 mya (95% HPD 8.29 - 19.50 mya), and the sister clades of Neotropical *Eptesicus* diverged around 12.47 mya (95% HPD 7.47 - 18.20 mya). These key divergence events were timed around early to mid-Miocene, followed by divergence in late Miocene to the Pleistocene (Fig 1.4). However, mean estimates of divergence time among closely related OTUs were overall much older than expected, such as the 8 mya divergence between continental lineages of *E. fuscus* compared to the previously estimated Holocene divergence using RADseq data (Yi & Latch 2022). This difference in time estimation may be due to the violated assumptions of no gene flow among OTUs. In addition, it has been found that concatenation methods (e.g., MCMCTree used here) with deep calibrations generally overestimate divergence times of closely related taxa (Tiley et al. 2020). Therefore, we suggest that the molecular divergence time between closely related OTUs may be overestimated in our study and thus should be interpreted with caution. However, the ancestral divergence, such as among major *Eptesicus* clades and outgroups, are expected to be more accurate.

The BioGeoBEARS results supported models with founder events (+J) over their corresponding

null models, indicating discontinuous range expansion via "jump dispersal" among geographic regions, such as via the trans-Atlantic route. When using the MCMCTree results (Fig 1.4) as the input, we found the DEC+J model as the best fit (Table 1.1) and the MRCA of the New World *Eptesicus* mostly likely in the Amazonian area, and the MRCA of Old World and New World *Eptesicus* most likely in the Saharo-Arabian area (Fig 1.5). When using the BPP topology (Fig 1.3C) as the input, again we found the DEC+J model as the best fit (Appendix A Table S2) and the MRCA of the New World Eptesicus in the Amazonian area, but the MRCA of Old World and New World Eptesicus was most likely in the Oriental area (Appendix A Fig S8).

Discussion

How organisms achieved their global distribution and diversification is an important question for understanding their evolution. Two key steps are required in the process of biodiversification: the successful colonization of new habitats, and the subsequent radiation in unoccupied ecological niches. It has been assumed that recently evolved organisms (such as the majority of terrestrial mammals) colonized separated continents via on-land dispersal, but it has been argued that the alternative trans-marine dispersal should be considered equally likely (de Queiroz 2005). Our phylogenetic study of the cosmopolitan bat genus *Eptesicus* favored the trans-marine dispersal hypothesis over on-land dispersal hypothesis, and indicated a colonization route from North Africa to the northern Neotropics. Using high-resolution genomic markers and extensive taxonomic as well as geographic sampling, we identified four well-supported New World clades and a novel topology where *Histiotus* is more closely related to *E. fuscus*, and we detected cryptic diversity in the Neotropics where current morphological identifications do not fully reflect molecular phylogenetic relationships. Our study in bats also supported previous findings

in birds (Smith et al. 2014b) that dispersal among heterogeneous environments drove the high level of Neotropical diversity, which could be a common mechanism of diversification in dispersive taxa. We highlight that trans-marine dispersal should be considered as an important hypothesis to be tested in terrestrial organisms, which could have an evolutionary significance in boosting global biodiversity.

Trans-Atlantic dispersal

Our results favored the trans-Atlantic dispersal over the on-land dispersal hypothesis in the global colonization of *Eptesicus* bats. The identified New World origin in the (northern) Neotropics is expected to be the landing area of the Old World common ancestor that first colonized the New World. Our ancestral reconstruction using the ASTRAL tree topology showed that this Old World common ancestor most likely originated from the Saharo-Arabian area (Fig 1.5). These ancestral reconstructions thus indicated a trans-Atlantic colonization route from North Africa to the northern Neotropics, which is also approximately the shortest geographic distance across the Atlantic. On the other hand, ancestral reconstructions using the BPP tree topology showed an Oriental origin of the Old World common ancestor (Appendix A Fig S8), indicating a trans-Pacific route that would be less parsimonious and thus less likely. It should be noted that the biogeographic areas were divided on a broad scale and the boundaries are not absolute in some regions, such as between the Saharo-Arabian and Palearctic areas and between the Amazonian and South American areas where some of our samples were located (Fig 1.1; Appendix A Fig S1), making these pairs of adjacent areas comparatively likely. In addition, caution is warranted that our limited sampling in the Old World, such as in the Oriental and Palearctic, might have impacted phylogenetic analyses (such as the paraphyly of Old World

clades) and precluded detection of more detailed biogeographic patterns. As a result, one might suggest that the Old World common ancestor could have originated from the eastern Palearctic where we lacked sampling. However, we find this scenario less likely than the Saharo-Arabian origin for the following reasons. First, the eastern Palearctic has a much lower species richness compared to the Saharo-Arabian area (Juste et al. 2013), and speciation is predicted to originate from areas of higher species richness due to the longer evolutionary time. Second, colonization from eastern Palearctic directly to northern Neotropics would also indicate a trans-marine dispersal over a much greater distance across the Pacific, a route that is less parsimonious than the shorter trans-Atlantic route.

Third, if a common ancestor in the eastern Palearctic first arrived in North America by an on-land dispersal route across Beringia, then one would expect to find initial radiations resulting in higher species diversity in North America than in South America. Such patterns of higher species richness in North America have been observed in the genus *Myotis* (e.g., Morales et al. 2019) which was suggested to have colonized the New World from Old World via Beringia (Stadelmann et al. 2007; Ruedi et al. 2013) and where the northern Nearctic species (*Myotis lucifugus*) is even distributed in Alaska. The reverse is true in *Eptesicus* which has 17 nominated species (including the *Histiotus* clade) in South America but only one species to the north of Mexico, and the extant New World species have a Neotropical origin (Fig 1.5; Appendix A Fig S8). Therefore, an on-land colonization explanation for *Eptesicus* evolution would require a mass extinction of all the ancestral lineages that initially colonized and diverged in North America. Instead, the discovered Nearctic *Eptesicus* fossils are rare, only dated back to late-Miocene to Pliocene, and likely represented the extant species *Eptesicus fuscus* (Czaplewski 2017). The available fossil record thus does not indicate a scenario of mass extinction of all ancestral

Nearctic *Eptesicus* lineages. Accordingly, we suggest that the trans-Atlantic colonization route is more likely than an on-land route in this bat lineage. Additional studies using more complete Old World sampling are needed to further test hypotheses about the origin of *Eptesicus* bats and their global colonization routes.

Trans-Atlantic dispersal between Africa and the Neotropics has also been suggested in other bat lineages (Lim 2009) and various taxonomic groups of plants and animals (e.g., reviewed in de Queiroz 2005). Bats most likely flew across the Atlantic, but we could not fully exclude the possibility of rafting such as proposed in other mammals (Rowe et al. 2010; Bond et al. 2015) and in the poorly flighted birds (Mayr et al. 2011) that made the trans-Atlantic dispersal. Successful colonization also requires a big enough population size that made the dispersal. A study in locust suggested that the swarming behavior might have facilitated their trans-Atlantic flight in high velocity winds and successful colonization of the New World (Lovejoy et al. 2005). Similarly, seasonal swarming in large groups may also have helped *Eptesicus* bats colonize the New World by flight, and the ability of females to store sperm for months (such as in *E. fuscus,* Kurta & Baker 1990) can further increase the chances of successful colonization. More data are needed to test these hypotheses. In addition, strong winds and hurricanes have been proposed to facilitate trans-marine dispersal (e.g., Renner 2004; Munoz et al. 2004), and bats can be good at following the wind (O'Mara et al. 2021). During the early to middle Miocene (around 20 mya) when Old World and New World *Eptesicus* diverged (Fig 1.4), the Atlantic was characterized by strong currents and hurricanes tracking from Africa to the northern Neotropics and Central America (Omta & Dijkstra 2003; Baarli et al. 2017), a geographic pattern consistent with the proposed colonization route. Divergence times among the major New World clades also roughly coincide with the Mid-Miocene Climatic Optimum (15-17 mya, Zachos et al. 2001),

indicating that warmer climates might have facilitated the long-distance dispersal and rapid radiation in novel niches, an evolutionary history possibly also shared by other taxonomic groups that diverged during this time.

Bat systematics using museum collections and UCEs

Molecular studies (e.g., Teeling et al. 2005; Miller-Butterworth et al. 2007) have shown that bat morphology can be unreliable and even misleading for indicating their phylogeny, possibly due to a combination of repeated convergence/parallel evolution and a lack of morphological diversity in cryptic lineages. On the other hand, bat phylogenies tend to show strong geographic signals, such as indicated in our results and previous molecular studies (Hoofer & van den Bussche 2003; Juste et al. 2013; Ruedi et al. 2013; Amador et al. 2018; Platt et al. 2018). Accordingly, geographic and intra-species sampling is especially important for bat phylogeny where cryptic diversity could be high and species misidentification is not rare. Such broad-scale sampling in a legitimate time frame is obviously difficult, making the use of museum collections highly beneficial for bat systematics. Our results demonstrated a successful application of genomic methods to museum samples, including ethanol-preserved tissues and wing pieces from dry skin specimens. Unsurprisingly, all samples (n=15) that failed UCE enrichment were dry skins, but the majority of dry skin samples (42 out of 57 total, 74%, including the replicate) generated sufficient data to be retained in our phylogenetic analyses, including collections as old as 1890s (Appendix A Table S1). However, it should be noted that the retained dry skin samples (n=41) had fewer identified UCEs (mean 3018 loci) and shorter sequences per UCE (mean 426 bp) than the retained tissue samples (n=39, mean 4015 UCE loci, mean 1178 bp per UCE). As a result, the skin samples tend to have higher proportions of missing data which might have

skewed their terminal branch lengths in the ML tree (Fig 1.2). To test this potential missing data bias, we plotted the terminal branch lengths of the ML tree in Figure 1.2 against the per-sample percentages of missing data and found a significant positive Pearson correlation (r=0.69, *p*=2.6e-13; Appendix A Fig S9A). Consistently, previous studies have shown that higher proportions of missing data tend to generate longer terminal branches although the phylogenetic relationships are not biased (e.g., Kimball et al. 2021). Therefore, the tree topology and taxonomic relationships indicated in Figure 1.2 are expected to be accurate, despite the biases in terminal branch lengths. The missing data bias might also contribute to the overestimated divergence times between closely related taxa. However, we found that the biasing effects were largely mitigated in the pruned data set (54 samples, the 500 most informative UCEs) based on the much weaker correlation (r=0.28, *p*=0.041; Appendix A Fig S9B) between per-sample missing data and the terminal branch lengths in the timed tree (Fig 1.4).

Our success with museum specimens was also attributed to the target sequencing of UCE enrichment, which is much more efficient on degraded samples than other next-generation sequencing methods (e.g., reduced representative sequencing, or RADseq). In addition, UCEs are especially suitable for phylogeny because they are orthologous loci present across distantly related lineages (e.g., tetrapods, Faircloth et al. 2012), and sequencing data from different studies can be easily combined as long as the same capture baits were used. For example, the incorporation of an Old World *Myotis* species from Platt et al. (2018) allowed us to suggest that the individual misidentified as *Eptesicus brasiliensis andinus* belongs to the genus *Myotis*. Future studies that collect UCEs from other bats can also easily incorporate our results to generate a more complete phylogeny; in particular, UCEs from Old World *Eptesicus* would allow additional testing of alternative colonization hypotheses and deeper understanding of the

evolutionary history of the genus. Accordingly, the good success rate and repeatability of UCEs make them a highly promising marker for systematic studies especially when using museum collections. We look forward to broader applications of UCEs in museum specimens, especially bats, to re-evaluate taxonomy and gain new insights into the species diversity and evolutionary history.

Eptesicus phylogeny and New World cryptic diversity

Our results confirmed *Histiotus* as a sub-group of *Eptesicus* but showed that the previous taxonomic suggestion of two subgenera in the New World (Hoofer & van den Bussche 2003) may be inappropriate. If *Histiotus* is an independent subgenus and *E. fuscus* is the type species for the subgenus *Eptesicus*, then our results clearly support at least one more subgenus for the Neotropical *Eptesicus* clades. Similarly, the Old World paraphyly indicates that the proposed subgenus *Cnephaeus* should be further divided. Future work combining extensive sampling from both New World and Old World is needed to provide a more complete *Eptesicus* phylogeny and suggest taxonomy. Despite strong supports of major New World clades, intra-clade divergence was less resolved and inconsistency was found between molecular phylogenies and morphological identities. We did not apply species delimitation in this study because it has been shown that contemporary species delimitation methods (such as BPP) could not distinguish between population divergence and speciation (Sukumaran & Knowles 2017) and tend to over-split geographically widespread species (Chambers & Hillis 2020). Reliable delimitation of biological species would require a combination of multiple approaches (Carstens et al. 2013) and integrative data including genomics, morphology, behavior, and ecological traits. However, our results clearly indicate species misidentification and cryptic diversity, highlighting that thorough

taxonomic re-evaluation is warranted. Based on our molecular phylogenies, here we provide taxonomic suggestions on the analyzed samples (Appendix A Table S1) and discuss below divergence within each of the major New World clades.

- The *E. fuscus* clade

Our sampling covered 9 of the 11 *E. fuscus* subspecies (Kurta & Baker 1990) based on museum taxonomy and geographic locations, excluding *E. f. osceola* in Florida and *E. f. petersoni* in Isla de la Juventud in the Caribbean. Our results supported the continent-island divergence and the continental west-east divergence found in previous phylogeographic studies (Turmelle et al. 2011; Yi & Latch 2022). The Continent West lineage includes multiple subspecies (*E. f. miradorensis*, *E. f. peninsulae*, *E. f. pallidus*, and *E. f. bernardinus*) that are weakly diverged based on UCEs, while previous studies using mitochondrial (Turmelle et al. 2011) and nuclear RADseq data (Yi & Latch 2022) both indicated genetic lineages roughly representing different western subspecies or geographic populations. This difference indicates the limited power of UCEs to detect fine-scale variation on the population level. On the other hand, divergence of the Caribbean subspecies was well supported in all studies. For example, the same individual from Jamaica was found as the distinct basal lineage in the Caribbean clade both using UCEs (this study) and using RADseq (Yi & Latch 2022), indicating a putative cryptic subspecies. Interestingly, the individual identified as *E. guadeloupensis* from Guadeloupe was well clustered within the Caribbean clade of *E. fuscus*, indicating that either *E. guadeloupensis* is a subspecies of *E. fuscus* or that the Caribbean lineages represent different species. We prefer the former interpretation of subspecies status considering the occasional gene flow in the Caribbean indicated by our phylogenetic discordances (e.g., Fig 1.3) and the previous network analyses (Yi & Latch 2022). However, the overall strong oceanic isolation and the unique insular habitats

with small population sizes might promote divergence on the Caribbean islands towards complete speciation.

- Neotropical *Eptesicus* A: the *diminutus_furinalis* group

The clade A of Neotropical *Eptesicus* is mainly composed of individuals identified as *E. diminutus* and *E. furinalis*. Although we labeled the clade as Neotropical, it should be noted that these species are also distributed in Central America such as Mexico (Davis 1966; Mies et al. 1996). This clade also included three samples identified as *E. brasiliensis* (Fig 1.2), a puzzling species that warrants taxonomic re-evaluation based on our molecular phylogenies (see 4.3.3). Therefore, we suggest that clade A is a genetic complex of *E. diminutus* and *E. furinalis*. These species are also hard to distinguish based on morphology (e.g., Mies et al. 1996; Ramírez-Chaves et al. 2021), further indicating hybridization and/or cryptic diversity. This clade also shows a strong geographic pattern of divergence, such as the Central American (Mexico, Belize, Nicaragua) lineage where the single *E. diminutus* individual may be a misidentified *E. furinalis* (Fig 1.2). In another case, the Venezuela lineage of one individual from each species may instead represent the new species *E. orinocensis* which was recently elevated from *E. diminutus* populations in Colombia and Venezuela (Ramírez-Chaves et al. 2021). It is thus likely that other geographic lineages may also represent cryptic species, such as the lineage in Ecuador and the lineage in Bolivia and Paraguay. Accordingly, our molecular phylogenies together with ongoing research in the Neotropics suggest that this clade of *E. diminutus* and *E. furinalis* may harbor high levels of cryptic diversity that has been so far concealed by morphological similarities but may be indicated by geographic distributions of these bats.

- Neotropical *Eptesicus* B: the *brasiliensis_chiriquinus* group

The clade B included all the other identified Neotropical *Eptesicus* species. Our phylogenies supported the divergence of *E. innoxius* but showed unclear relationships among individuals identified as *E. brasiliensis*, *E. chiriquinus*, and *E. andinus*, whose morphological similarities were also reported in previous studies (Davis 1965; Davis 1966). Some geographic pattern was found, such as the clustering of three individuals from Peru including both *E. brasiliensis* and *E. chiriquinus*, but the overall geographic signal seems weaker than that in clade A. The unresolved phylogenetic relationships could reflect gene flow in a species complex or cryptic diversity concealed by morphological similarities, the alternative scenarios that need to be tested in future studies using extensive fine-scale geographic sampling and more variant genomic markers (e.g., RADseq).

Our phylogenies indicated that *E. brasiliensis* is a puzzling species that warrants taxonomic re-evaluation. The sampled individuals identified as *E. brasiliensis* were not genetically related to each other but instead were sorted into various phylogenetic clades including both clades of Neotropical *Eptesicus*, the *Histiotus* clade, and even the outgroup clade of genus *Myotis* (Fig 1.2). These results showed a high rate of misidentification and indicated that *E. brasiliensis* might have served as a "basket name" in taxonomy to accommodate individuals that did not fit with morphological descriptions of the other identified species. Previous studies have suggested *E. brasiliensis* as a putative species complex (Davis 1965; Davis 1966) but our results showed that individuals identified in this group could be more diverged (e.g., since Miocene) than what is expected within a species complex. Accordingly, we suggest that the contemporary usage of *E. brasiliensis* has probably disguised a high level of cryptic diversity that awaits more rigorous taxonomic evaluations.

- The *Histiotus* clade

The *Histiotus* clade also indicated cryptic diversity that may not be fully represented by morphological identifications. Previous studies of *Histiotus* taxonomy have suggested various numbers of species from only four (Nowak & Walker 1994) to ten recognized species (Rodríguez-Posada et al. 2021). Our results favored the higher species number by supporting the elevation of *H. magellanicus* (Díaz et al. 2019; Giménez et al. 2019) from a previous subspecies in *H. montanus* (Handley & Gardner 2008), and the elevation of *H. laephotis* from a previous subspecies of *H. macrotus* or *H. montanus* (Barquez & Díaz 2001). The other *H. montanus* individuals also split into two diverged lineages representing *H. montanus montanus* and *H. montanus inambarus*, the latter being either a cryptic species or the synonym of *H. laephotis*. Previous studies have found *H. montanus* and *H. macrotus* difficult to distinguish using morphology and limited genes, especially when the samples were sympatric (Giménez et al. 2019; Rodríguez-Posada et al. 2021), indicating possible hybridization. Our results of geographically separated samples, on the other hand, showed clearer divergence. Therefore, it is likely that additional sampling would find individuals of these two species cluster by geographic localities rather than morphological identities, indicating unrecognized cryptic diversity. This hypothesis is partially supported by the recent elevation of *H. colombiae* as a full species from the previous subspecies of *H. montanus* (Rodríguez-Posada et al. 2021).

Our phylogenies also indicated cryptic diversity in the currently recognized *H. velatus*. Previous studies have identified *H. velatus* populations in northeastern and southwestern Brazil as a new species *H. diaphanopterus* which has transparent wings (Feijó et al. 2015; Semedo & Feijó 2017). Our *H. velatus* sample from southeastern Brazil was collected in different habitats and the museum specimen does not clearly demonstrate key morphological features of *H. diaphanopterus* (Appendix A Fig S10), making us hesitant to rename its species identity.

However, our two *H. velatus* samples from Brazil and Peru clearly represent different lineages both based on molecular phylogenies and their distinct morphological features (Appendix A Fig S10). Interestingly, the closely related *Histiotus* individual from Peru (AMNH:M−278524) was recently identified as a new species *H. mochica* (Velazco et al. 2021). If true, this identification combined with our phylogenies would further indicate species status of *H. velatus* in Brazil, *H. velatus* in Peru, and *H. macrotus* in Paraguay. Further taxonomic evaluations of *Histiotus* species are much needed not only using genomics and morphology but also ecological data such as behavior, geographic distribution, and habitat niche, which remains poorly known in *Histiotus* (Carvalho et al. 2013; Giménez et al. 2015; Díaz et al. 2019).

Diversification in the Neotropical hotspot

The current taxonomy (Simmons & Cirranello 2020) shows highly unbalanced speciation of the New World *Eptesicus* with only one species, *E. fuscus*, identified in the Nearctic north of Mexico, while 9 *Eptesicus* and 8 *Histiotus* species were identified in Central America and the Neotropics. This unbalance may be even more extreme considering the high level of Neotropical cryptic diversity indicated in our results and the increasing reports of new species in the Neotropics (e.g., Feijó et al. 2015; Sánchez et al. 2019; Acosta et al. 2021; Ramírez-Chaves et al. 2021; Rodríguez-Posada et al. 2021; Velazco et al. 2021). Elevated Neotropical species richness is in line with broader patterns such as the latitudinal biodiversity gradient and the Neotropical biodiversity hotspot. These patterns have been explained by multiple hypotheses and mechanisms, including longer evolutionary time in the tropics (Weir & Schluter 2007), variation of speciation rates (Schluter & Pennell 2017), and Neotropical geographic heterogeneity (Hoorn et al. 2010). These mechanisms are not mutually exclusive and all could have played a role in the

diversification of *Eptesicus* and *Histiotus*. For example, the Neotropical origin of the New World clades supported the hypothesis of longer evolutionary time in the Neotropics. The higher turnover rates in the Nearctic and/or lower extinction rates in the Neotropics might be indicated by the more ancient (up to Late-Miocene) Nearctic *Eptesicus* fossils compared to the younger (Pleistocene and Holocene) Neotropical *Eptesicus* fossils (Czaplewski 2017; Lim 2009), although this pattern may be biased by the rarity of bat fossils (Teeling et al. 2005). Neotropical environments were also suggested to promote speciation in these bats, such as the effects of Andean Mountains on the divergence of *H. magellanicus* (Giménez et al. 2019) and the Amazon basin on the divergence of *E. orinocensis* (Ramírez-Chaves et al. 2021).

In addition to the above mechanisms, here we emphasize the roles of dispersal and adaptation in promoting Neotropical biodiversity. A previous study in birds showed that Neotropical speciation was largely driven by dispersal across heterogeneous landscapes (Smith et al. 2014b), which may also hold true in the diversification of Neotropical bats. As the only mammalian group with powered flight, bats are generally more dispersive than other terrestrial mammals and are less impacted by geographic barriers (Bacon et al. 2016; López-Aguirre et al. 2018). In addition, bats also have high adaptive abilities that allow them to occupy various ecological niches. For example, *Myotis* species around the world have repeatedly evolved the three ecomorphs that are well adapted to different ecological niches (Ruedi & Mayer 2001; Platt et al. 2018; Morales et al. 2019). Similarly, it is likely that the long-eared *Histiotus* and short-eared *Eptesicus* also represent ecomorphs adaptive to different ecological niches such as different insect prey or feeding strategies (e.g., gleaner versus aerial netters in *Myotis*). Additionally, the enlarged ears in *Histiotus* might have allowed these species to avoid competition and co-distribute with the Neotropical *Eptesicus*. These hypotheses need to be tested with additional studies of *Histiotus*

especially their physiology and ecological traits. Taken together, we suggest that high dispersal abilities and potentials of rapid adaption are two key biological features that have promoted speciation of these Neotropical bats and other widespread speciose taxa.

Conclusions

Our phylogenetic study showed a trans-Atlantic colonization route from North Africa to the northern Neotropics followed by rapid Neotropical radiation, possibly triggered by the Miocene climatic events. High cryptic diversity was indicated in the Neotropics, which underscores the need for additional taxonomic evaluations of *Eptesicus* and *Histiotus* using genomics, morphology, and also ecological data such as behavior and life history traits. Our findings add to the accumulating evidence of trans-marine dispersal that might be more ubiquitous in terrestrial organisms than currently acknowledged. We suggest that such long-distance dispersal could be an important evolutionary mechanism for organisms to reach new habitats, a key first step prerequisite radiation in novel niches and the eventual successful global diversification.

Figure 1.1 Distribution of the *Eptesicus* individuals analyzed in this study. Each dot represents one individual (total n=65) and black dots are the ones included in the pruned dataset (n=31). Shapes represent the phylogenetic clade corresponding to the indicated UCE topology. Arrows represent colonization routes corresponding to the on-land and trans-marine dispersal hypotheses, and the latter in a solid arrow was favored in this study. Photos are the two bat specimens sampled and analyzed in this study, representing the short-eared *Eptesicus fuscus* (top, specimen AMNH_99048) and the long-eared *Histiotus montanus* (bottom, specimen AMNH_183876).

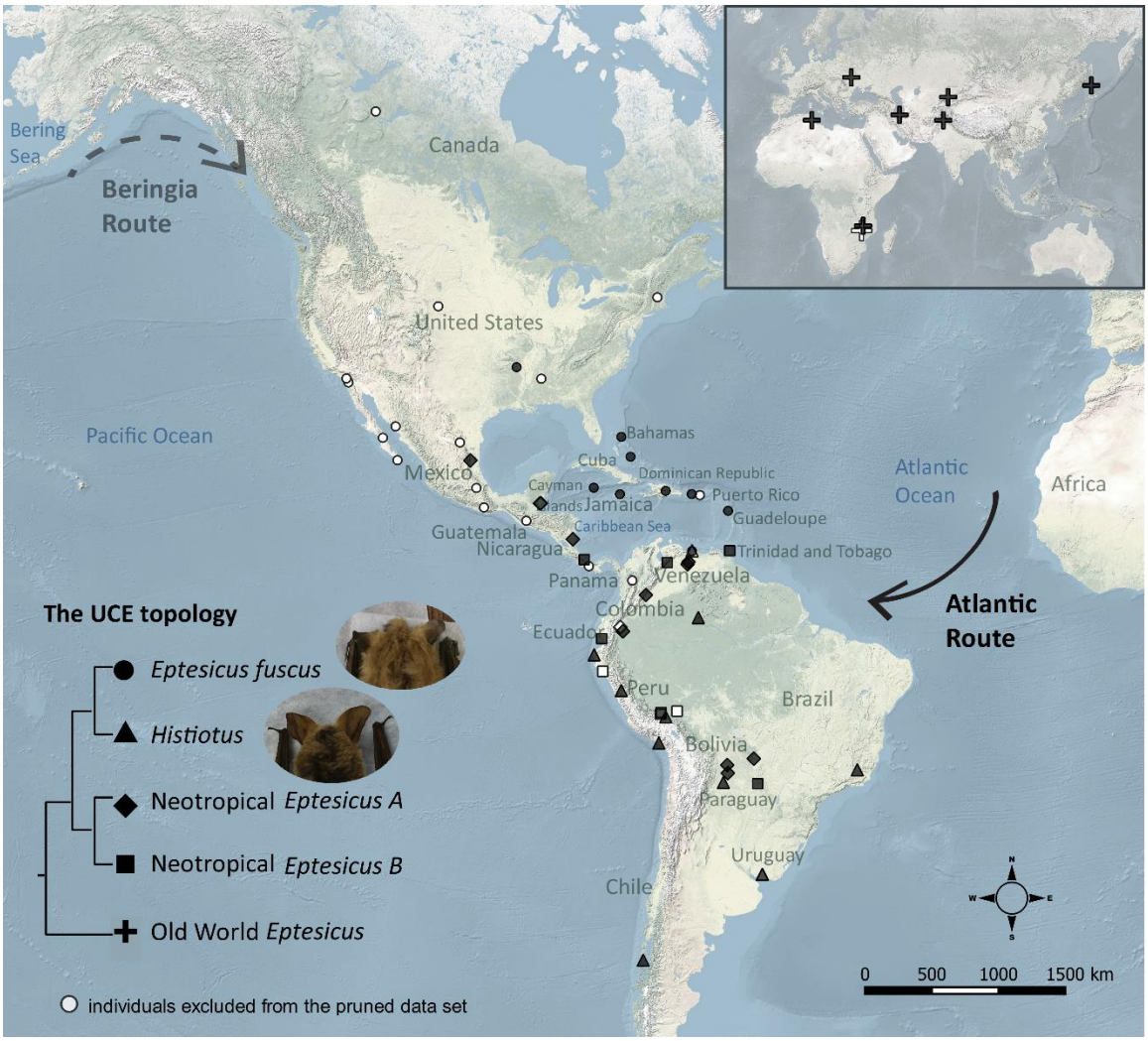


Figure 1.2 The maximum likelihood individual phylogeny generated in RAxML. The best-scored ML tree is visualized with branch lengths and bootstraps >75. Branches around the root node were truncated. Each tip represents one individual and is named by source species identity followed with sample source (e.g., museum catalogue) and the country of sample location. The three individuals in red text were misidentified species and were excluded from downstream analyses. The solid tips are the 54 individuals used in the pruned dataset. The assigned operational taxonomic units (OTUs) are given on the right and linked to corresponding individuals. The four major clades of the New World *Eptesicus* are highlighted.

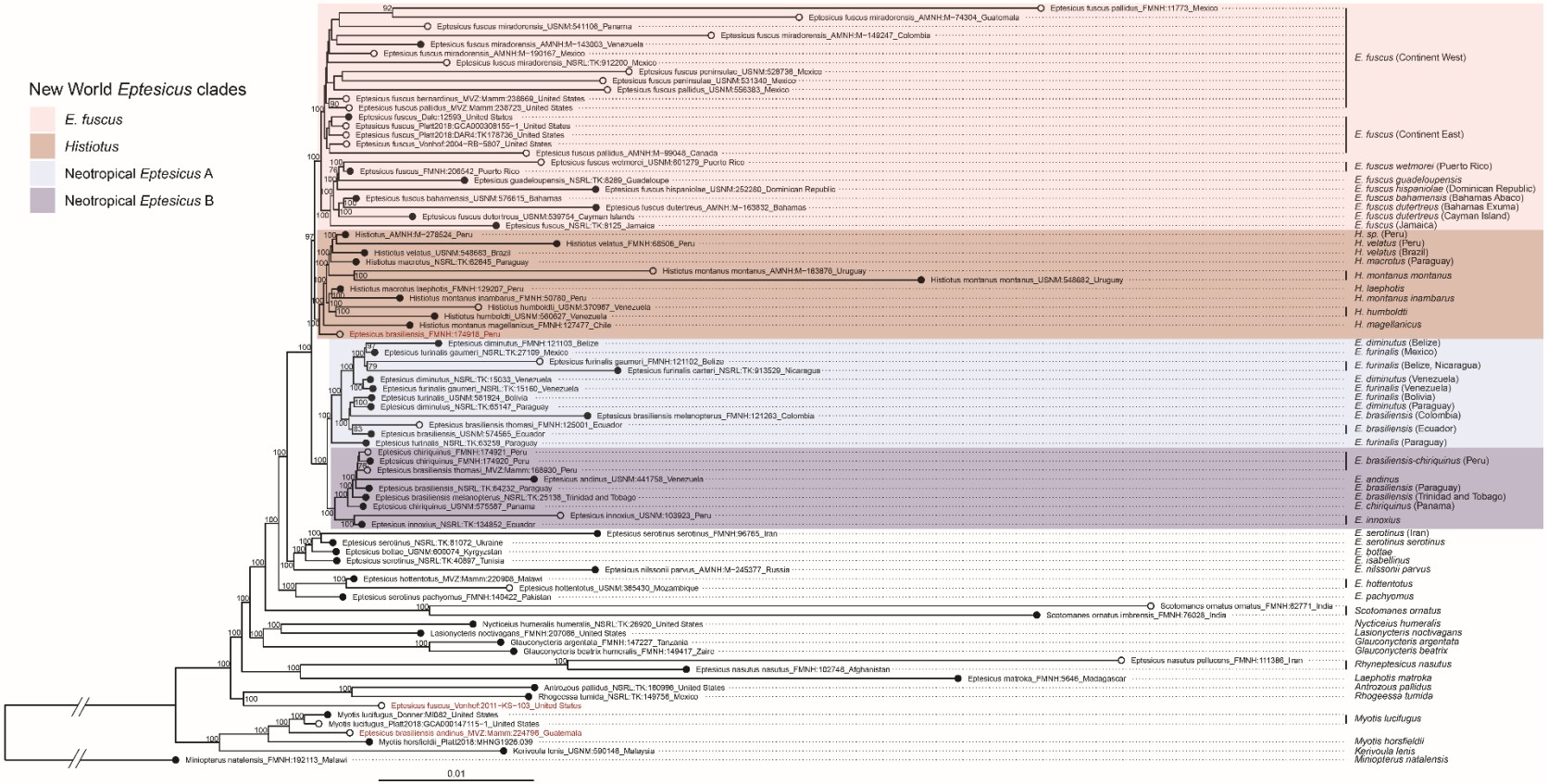


Figure 1.3 The species trees of 54 OTUs. A) The species tree summarized by ASTRAL-III using gene trees from the 500 most informative UCEs. Branches with <10% bootstraps were collapsed. Numbers on nodes represent posterior probabilities. Tip names are the assigned OTUs. **B)** The quartet-based tree generated by SVDquartets using all UCEs and OTU assignments. Numbers on nodes represent bootstrap supports. **C)** The species tree estimated by the Bayesian method BPP using the 500 most informative UCEs. The best tree is shown with posterior probabilities labeled on nodes. In all three trees, each tip represents one OTU that may correspond to more than one individual (see Fig 1.2). Red branches in B and C highlight differences from the ASTRAL tree in A and are connected with the corresponding tips by red dashed lines.

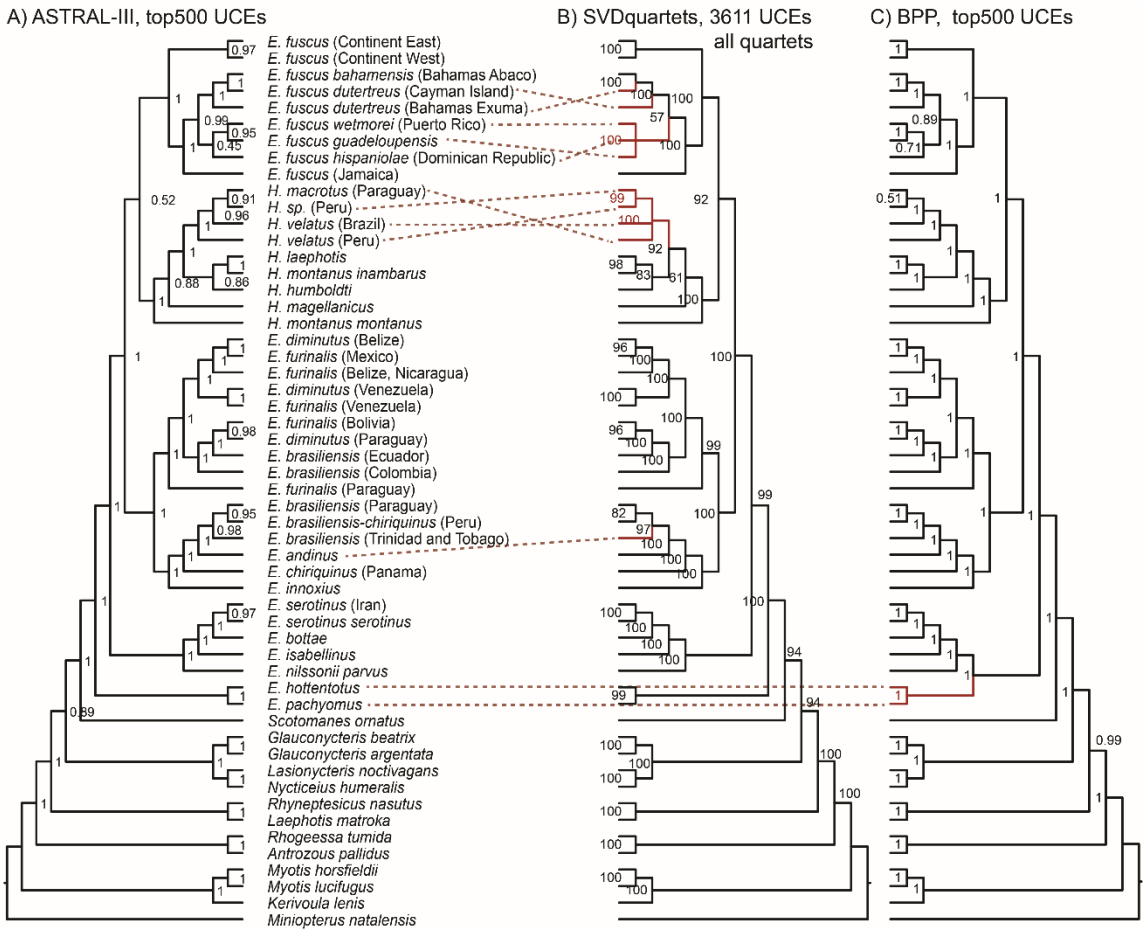


Figure 1.4 Divergence times estimated by MCMCTree using the 500 most informative UCEs. The inserted plot shows high consistency of posterior mean divergence times estimated in two independent runs. Each tip corresponds to one OTU represented by one individual in the pruned data set (black dots in Fig 1.1 and solid tips in Fig 1.2). The tree topology was fixed by results from ASTRAL-III using 500 UCEs (Fig 1.3A). Asterisks indicate the two soft-bound calibrations. Numbers on nodes are posterior mean divergence times and bars represent the 95% HPD. Key divergence times between major clades are enlarged and bolded. Corresponding geographic epochs are given at the bottom (Plio. for Pliocene; Pleist. for Pleistocene).

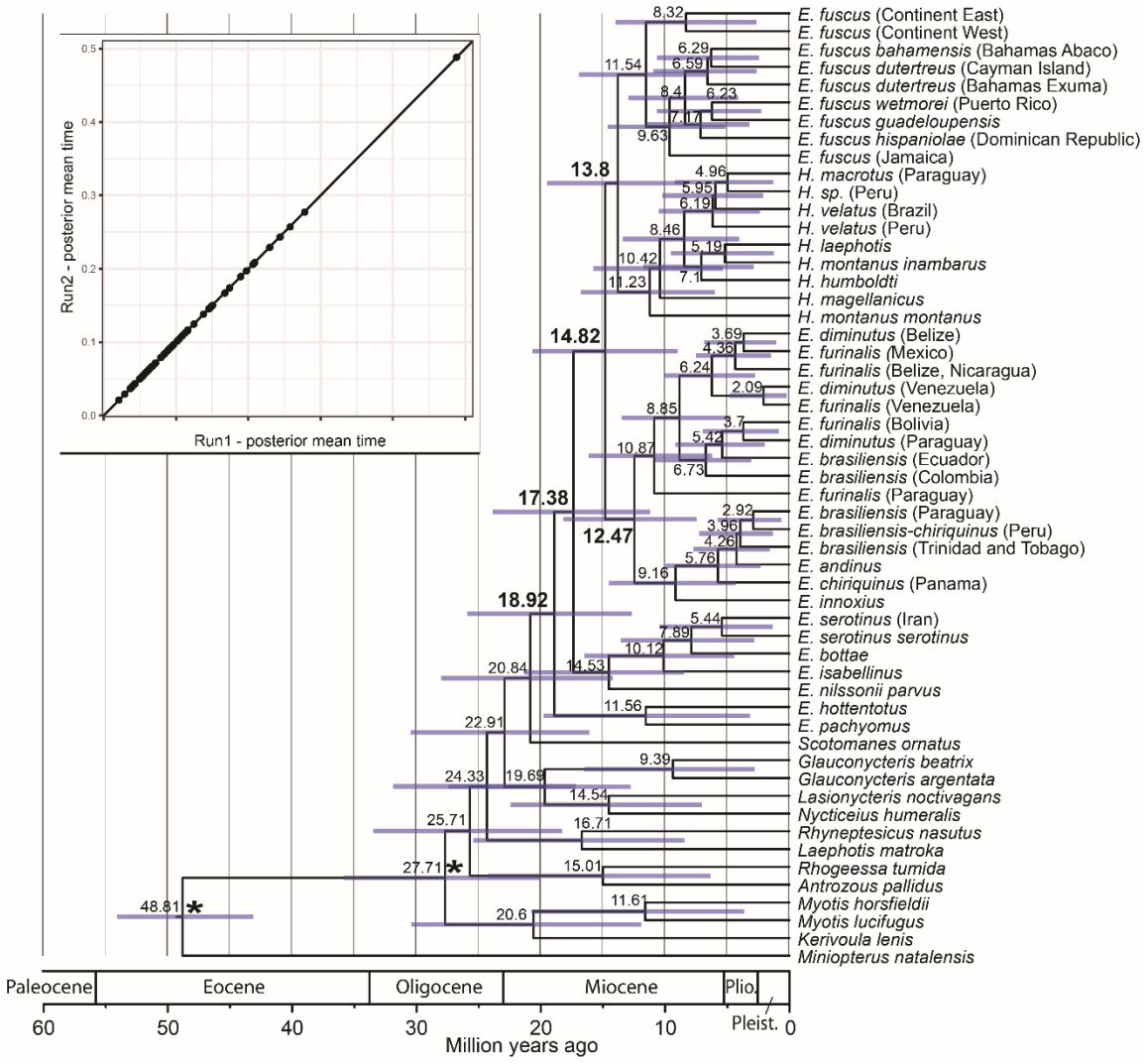


Figure 1.5 The reconstructed ancestral distribution in BioGeoBEARS using the DEC+J model. Tip letters represent the assigned biogeographic area(s) of each OTU (also in Appendix A Table S1) corresponding to the legends. Pie charts on nodes represent the estimated probabilities of ancestral geographic distribution. The most likely distributions are labeled for the key nodes representing the MRCA of *Histiotus* and *E. fuscus*, the MRCA of all New World clades, and the MRCA of the New World and Old World *Eptesicus*.

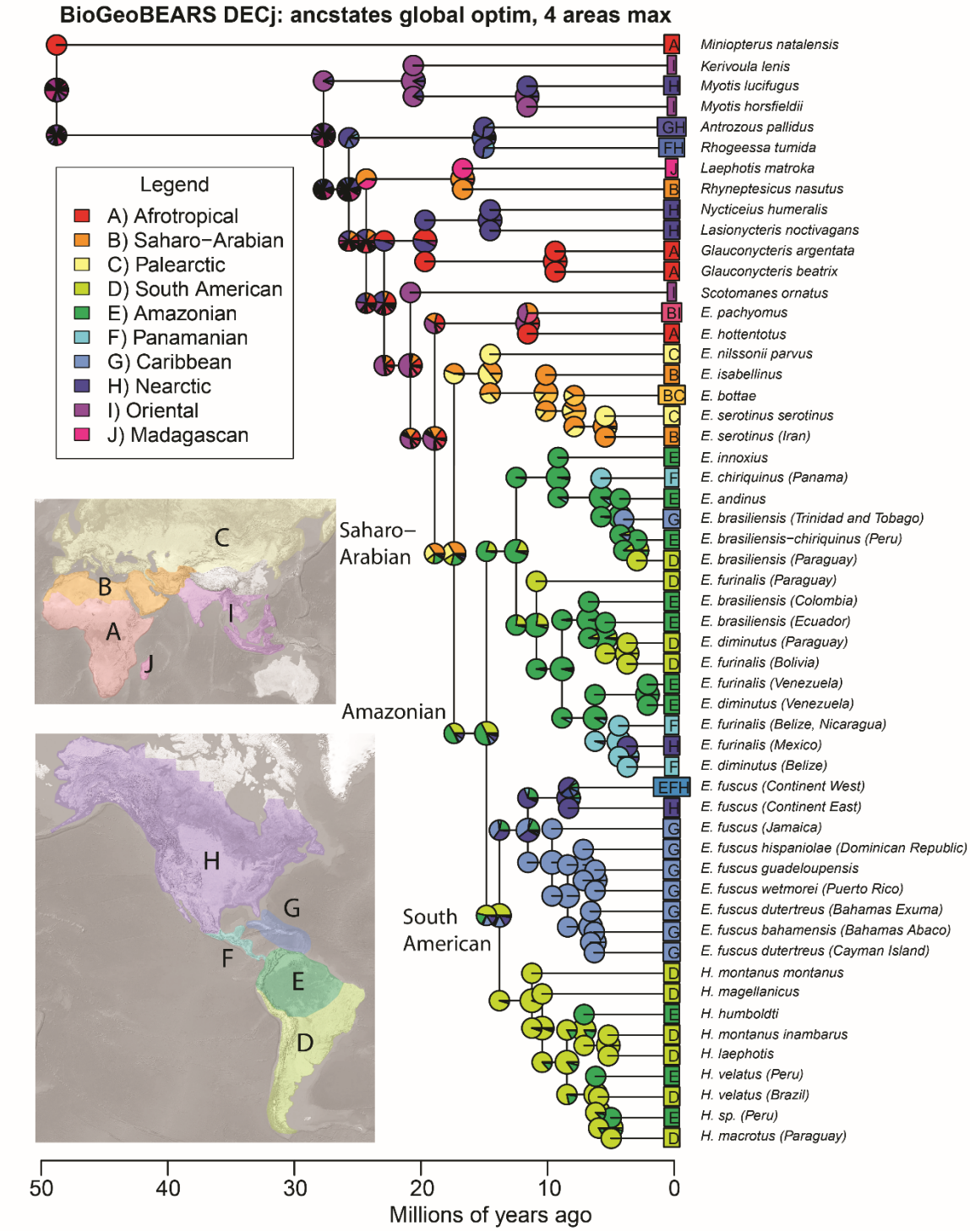


Table 1.1 Comparison of the six biogeographic models in BioGeoBEARS. Models are ordered based on AICc values. The input is the timed tree from MCMCTree analyses.

	LnL	d	e	j	AICc	AICc_wt
DEC+J	-125.9	0.10	1.00E-12	0.050	258.2	0.73
DIVALIKE+J	-126.9	0.11	1.00E-12	0.050	260.2	0.27
BAYAREALIKE+J	-130.9	0.09	1.00E-08	0.056	268.4	4.60E-03
DEC	-160.7	0.36	0.35	0	325.6	1.70E-15
DIVALIKE	-161.3	0.49	0.28	0	326.8	9.40E-16
BAYAREALIKE	-171.5	0.32	5.00	0	347.3	3.40E-20

References

Aberer, A. J., Kobert, K., & Stamatakis, A. (2014). ExaBayes: Massively Parallel Bayesian Tree Inference for the Whole-Genome Era. *Molecular Biology and Evolution*, *31*(10), 2553–2556.

Acosta, L. H., Poma-Urey, J. L., Ossa-López, P. A., Rivera-Páez, F. A., & Ramírez-Chaves, H. E. (2021). A new species of Eptesicus (Mammalia: Chiroptera: Vespertilionidae), from the sub-Andean Forest of Santa Cruz, Bolivia. *THERYA*, *12*(3), 391.

Agnarsson, I., Zambrana-Torrelio, C. M., Flores-Saldana, N. P., & May-Collado, L. J. (2011). A time-calibrated species-level phylogeny of bats (Chiroptera, Mammalia). *PLoS Currents*, *3*.

Alda, F., Tagliacollo, V. A., Bernt, M. J., Waltz, B. T., Ludt, W. B., Faircloth, B. C., Alfaro, M. E., Albert, J. S., & Chakrabarty, P. (2019). Resolving Deep Nodes in an Ancient Radiation of Neotropical Fishes in the Presence of Conflicting Signals from Incomplete Lineage Sorting. *Systematic Biology*, *68*(4), 573–593.

Alminas, O. S. V., Heffelfinger, J. R., Statham, M. J., & Latch, E. K. (2021). Phylogeography of Cedros and Tiburón Island Mule Deer in North America's Desert Southwest. *Journal of Heredity*, *112*(3), 260–275.

Amador, L., Moyers Arévalo, L., Cunha Almeida, F., & Catalano, S. A. (2018). Bat Systematics in the Light of Unconstrained Analyses of a Comprehensive Molecular Supermatrix Article. *Journal of Mammalian Evolution*.

Artyushin, I. v., Kruskop, S. v., Lebedev, V. S., & Bannikova, A. A. (2018). Molecular Phylogeny of Serotines (Mammalia, Chiroptera, *Eptesicus*): Evolutionary and Taxonomical Aspects of the E. serotinus Species Group. *Biology Bulletin*, 45(5), 469–477.

Baarli, B., Malay, M. C. (Machel) D., Santos, A., Johnson, M. E., Silva, C. M., Meco, J., Cachão, M., & Mayoral, E. J. (2017). Miocene to Pleistocene transatlantic dispersal of Ceratoconcha coral-dwelling barnacles and North Atlantic island biogeography. *Palaeogeography, Palaeoclimatology, Palaeoecology*, *468*, 520–528.

Bacon, C. D., Molnar, P., Antonelli, A., Crawford, A. J., Montes, C., & Vallejo-Pareja, M. C. (2016). Quaternary glaciation and the Great American Biotic Interchange. *Geology*, *44*(5), 375–378.

Bairlein, F., Norris, D. R., Nagel, R., Bulte, M., Voigt, C. C., Fox, J. W., Hussell, D. J. T., & Schmaljohann, H. (2012). Cross-hemisphere migration of a 25 g songbird. *Biology Letters*, *8*(4), 505–507.

Barquez, R. M., & Díaz, M. M. (2001). Bats of the Argentine Yungas: a systematic and distributional analysis. *Acta zoológica mexicana*, (82), 29-81.

Batista, R., Olsson, U., Andermann, T., Aleixo, A., Ribas, C. C., & Antonelli, A. (2020). Phylogenomics and biogeography of the world's thrushes (Aves, Turdus): new evidence for a more parsimonious evolutionary history. *Proceedings of the Royal Society B*,

287(1919).

Blaimer, B. B., Brady, S. G., Schultz, T. R., Lloyd, M. W., Fisher, B. L., & Ward, P. S. (2015). Phylogenomic methods outperform traditional multi-locus approaches in resolving deep evolutionary history: A case study of formicine ants. *BMC Evolutionary Biology*, *15*(1), 1–14.

Blair, C., Bryson, R. W., Linkem, C. W., Lazcano, D., Klicka, J., & McCormack, J. E. (2019). Cryptic diversity in the Mexican highlands: Thousands of UCE loci help illuminate phylogenetic relationships, species limits and divergence times of montane rattlesnakes (Viperidae: Crotalus). *Molecular Ecology Resources*, *19*(2), 349–365.

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120.

Bond, M., Tejedor, M. F., Campbell, K. E., Chornogubsky, L., Novo, N., & Goin, F. (2015). Eocene primates of South America and the African origins of New World monkeys. *Nature 2015 520:7548*, *520*(7548), 538–541.

Carstens, B. C., Pelletier, T. A., Reid, N. M., & Satler, J. D. (2013). How to fail at species delimitation. *Molecular Ecology*, *22*(17), 4369–4383.

Carvalho, W. D., Martins, M. A., Dias, D., & Esbérard, C. E. L. (2013). Extension of geographic range, notes on taxonomy and roosting of *Histiotus montanus* (Chiroptera: Vespertilionidae) in southeastern Brazil. *Mammalia*, *77*(3), 341-346.

Censky, E. J., Hodge, K., & Dudley, J. (1998). Over-water dispersal of lizards due to hurricanes. *Nature 1998 395:6702*, *395*(6702), 556–556.

Chambers, E. A., & Hillis, D. M. (2020). The Multispecies Coalescent Over-Splits Species in the Case of Geographically Widespread Taxa. *Systematic Biology*, *69*(1), 184–193.

Chifman, J., & Kubatko, L. (2014). Quartet Inference from SNP Data Under the Coalescent Model. *Bioinformatics*, *30*(23), 3317–3324.

Czaplewski, N. J. (2017). First report of bats (Mammalia: Chiroptera) from the Gray Fossil Site (late Miocene or early Pliocene), Tennessee, USA. *PeerJ*, *2017*(4), e3263.

Davis, W. B. (1965). Review of the *Eptesicus brasiliensis* complex in Middle America with the description of a new subspecies from Costa Rica. *Journal of Mammalogy*, *46*(2), 229-240.

Davis, W. B. (1966). Review of South American Bats of the Genus *Eptesicus*. *The Southwestern Naturalist*, *11*(2), 245.

de Queiroz, A. (2005). The resurrection of oceanic dispersal in historical biogeography. *Trends in Ecology & Evolution*, *20*(2), 68–73.

Díaz, M. M., Ossa, G., & Barquez, R. M. (2019). *Histiotus magellanicus* (Chiroptera: Vespertilionidae). *Mammalian Species*, *51*(973), 18–25.

Eick, G. N., Jacobs, D. S., & Matthee, C. A. (2005). A Nuclear DNA Phylogenetic Perspective on the Evolution of Echolocation and Historical Biogeography of Extant Bats (Chiroptera). *Molecular Biology and Evolution*, *22*(9), 1869–1886.

Esselstyn, J. A., Oliveros, C. H., Swanson, M. T., & Faircloth, B. C. (2017). Investigating Difficult Nodes in the Placental Mammal Tree with Expanded Taxon Sampling and Thousands of Ultraconserved Elements. *Genome Biology and Evolution*, *9*(9), 2308–2321.

Faircloth, B. C., McCormack, J. E., Crawford, N. G., Harvey, M. G., Brumfield, R. T., & Glenn, T. C. (2012). Ultraconserved Elements Anchor Thousands of Genetic Markers Spanning Multiple Evolutionary Timescales. *Systematic Biology*, *61*(5), 717–726.

Faircloth, B. C. (2013). illumiprocessor: a trimmomatic wrapper for parallel adapter and quality trimming.

Faircloth, B. C. (2016). PHYLUCE is a software package for the analysis of conserved genomic loci. *Bioinformatics*, *32*(5), 786–788.

Feijo, A., Da Rocha, P. A., & Althoff, S. L. (2015). New species of Histiotus (Chiroptera: Vespertilionidae) from northeastern Brazil. *Zootaxa*, *4048*(3), 412-427.

Fenton, M. B., & Simmons, N. B. (2015). Bats: a world of science and mystery. The University of Chicago Press.

Flouri, T., Jiao, X., Rannala, B., & Yang, Z. (2018). Species Tree Inference with BPP Using Genomic Sequences and the Multispecies Coalescent. *Molecular Biology and Evolution*, *35*(10), 2585–2593.

Gansauge, M. T., Gerber, T., Glocke, I., Korlević, P., Lippik, L., Nagel, S., Riehl, L. M., Schmidt, A., & Meyer, M. (2017). Single-stranded DNA library preparation from highly degraded DNA using T4 DNA ligase. *Nucleic Acids Research*, *45*(10), e79.

Giménez, A. L., Giannini, N. P., Schiaffini, M. I., & Martin, G. M. (2015). Geographic and potential distribution of a poorly known South American bat, *Histiotus macrotus* (Chiroptera: Vespertilionidae). *Acta Chiropterologica*, *17*(1), 143-158.

Giménez, A. L., Giannini, N. P., & Almeida, F. C. (2019). Mitochondrial genetic differentiation and phylogenetic relationships of three *Eptesicus* (*Histiotus*) species in a contact zone in patagonia. *Mastozool Neotrop*, *26*(2), 349-358.

Goodman, S. M., Taylor, P. J., Ratrimomanarivo, F., & Hoofer, S. (2012). The genus *Neoromicia* (Family Vespertilionidae) in Madagascar, with the description of a new species. *Zootaxa*, *3250*(1), 1-25.

Guillory, W. X., French, C. M., Twomey, E. M., Chávez, G., Prates, I., von May, R., de la Riva, I., Lötters, S., Reichle, S., Serrano-Rojas, S. J., Whitworth, A., & Brown, J. L. (2020). Phylogenetic relationships and systematics of the Amazonian poison frog genus *Ameerega* using ultraconserved genomic elements. *Molecular Phylogenetics and Evolution*, *142*, 106638.

Handley, C. O., Jr., and A. L. Gardner. 2008. Genus Histiotus P. Gervais, 1856. Pp. 450–457 in Mammals of South America. Volume 1, marsupials, xenarthrans, shrews, and bats (A. L. Gardner, ed.). The University of Chicago Press, Chicago, Illinois.

Holt, B. G., Lessard, J. P., Borregaard, M. K., Fritz, S. A., Araújo, M. B., Dimitrov, D., Fabre, P. H., Graham, C. H., Graves, G. R., Jønsson, K. A., Nogués-Bravo, D., Wang, Z., Whittaker, R. J., Fjeldså, J., & Rahbek, C. (2013). An update of Wallace's zoogeographic regions of the world. *Science*, *339*(6115), 74–78.

Hoofer, S. R., & van den Bussche, R. A. (2003). Molecular Phylogenetics of the Chiropteran Family Vespertilionidae. Acta Chiropterologica, 5(suppl), 1-63.

Hoorn, C., Wesselingh, F. P., ter Steege, H., Bermudez, M. A., Mora, A., Sevink, J., Sanmartín, I., Sanchez-Meseguer, A., Anderson, C. L., Figueiredo, J. P., Jaramillo, C., Riff, D., Negri, F. R., Hooghiemstra, H., Lundberg, J., Stadler, T., Särkinen, T., & Antonelli, A. (2010). Amazonia through time: Andean uplift, climate change, landscape evolution, and biodiversity. *Science*, *330*(6006), 927–931.

Hosner, P. A., Tobias, J. A., Braun, E. L., & Kimball, R. T. (2017). How do seemingly non-vagile clades accomplish trans-marine dispersal? Trait and dispersal evolution in the landfowl (Aves: Galliformes). *Proceedings of the Royal Society B: Biological Sciences*, *284*(1854).

Hsu, M. H., Lin, J. W., Liao, C. P., Hsu, J. Y., & Huang, W. S. (2021). Trans-marine dispersal inferred from the saltwater tolerance of lizards from Taiwan. *PLOS ONE*, *16*(2), e0247009.

Jiang, D., Klaus, S., Zhang, Y.-P., Hillis, D. M., & Li, J.-T. (2019). Asymmetric biotic interchange across the Bering land bridge between Eurasia and North America. *National Science Review*, *6*, 739–745.

Junier, T., & Zdobnov, E. M. (2010). The Newick utilities: high-throughput phylogenetic tree processing in the Unix shell. *Bioinformatics*, *26*(13), 1669–1670.

Juste, J., Benda, P., Garcia-Mudarra, J. L., & Ibáñez, C. (2013). Phylogeny and systematics of Old World serotine bats (genus *Eptesicus*, Vespertilionidae, Chiroptera): An integrative approach. *Zoologica Scripta*, *42*(5), 441–457.

Kimball, R. T., Hosner, P. A., & Braun, E. L. (2021). A phylogenomic supermatrix of Galliformes (Landfowl) reveals biased branch lengths. *Molecular Phylogenetics and Evolution*, *158*, 107091.

Koubínová, D., Irwin, N., Hulva, P., Koubek, P., & Zima, J. (2013). Hidden diversity in Senegalese bats and associated findings in the systematics of the family Vespertilionidae. *Frontiers in Zoology*, *10*(1), 1–16.

Kunz, Thomas H., and M. Brock Fenton, eds. *Bat ecology*. University of Chicago Press, 2005.

Kurta, A., & Baker, R. H. (1990). *Eptesicus fuscus*. *Mammalian Species*, *356*(356), 1–10.

Lack, J. B., & van den Bussche, R. A. (2010). Identifying the confounding factors in resolving phylogenetic relationships in Vespertilionidae. *Journal of Mammalogy*, *91*(6), 1435–1448.

Lim, B. K. (2009). Review of the origins and biogeography of bats in South America. *Chiroptera Neotropical*, *15*(1), 391-410.

López-Aguirre, C., Hand, S. J., Laffan, S. W., & Archer, M. (2018). Phylogenetic diversity, types of endemism and the evolutionary history of New World bats. *Ecography*, *41*(12), 1955–1966.

Lovejoy, N. R., Mullen, S. P., Sword, G. A., Chapman, R. F., & Harrison, R. G. (2005). Ancient trans-Atlantic flight explains locust biogeography: molecular phylogenetics of Schistocerca. *Proceedings of the Royal Society B: Biological Sciences*, *273*(1588), 767–774.

Matzke, N. J. (2013). *Probabilistic historical biogeography: new models for founder-event speciation, imperfect detection, and fossils allow improved accuracy and model-testing*. University of California, Berkeley.

Matzke, N. J. (2014). Model Selection in Historical Biogeography Reveals that Founder-Event Speciation Is a Crucial Process in Island Clades. *Systematic Biology*, *63*(6), 951–970.

Mayr, G., Alvarenga, H., & Mourer-Chauviré, C. (2011). Out of Africa: Fossils shed light on the origin of the hoatzin, an iconic Neotropic bird. *Naturwissenschaften*, *98*(11), 961–966.

Mies, R., Kurta, A., & King, D. G. (1996). *Eptesicus furinalis*. *Mammalian Species*, *526*, 1–7.

Miller-Butterworth, C. M., Murphy, W. J., O'Brien, S. J., Jacobs, D. S., Springer, M. S., & Teeling, E. C. (2007). A Family Matter: Conclusive Resolution of the Taxonomic Position of the Long-Fingered Bats, *Miniopterus*. *Molecular Biology and Evolution*, *24*(7), 1553–1561.

Morales, A. E., Ruedi, M., Field, K., & Carstens, B. C. (2019). Diversification rates have no effect on the convergent evolution of foraging strategies in the most speciose genus of bats, *Myotis**. *Evolution*, *73*(11), 2263–2280.

Muñoz, J., Felicísimo, Á. M., Cabezas, F., Burgaz, A. R., & Martínez, I. (2004). Wind as a long-distance dispersal vehicle in the Southern Hemisphere. *Science*, *304*(5674), 1144–1147.

Murray, E. A., & Heraty, J. M. (2016). Invading Africa: a novel transoceanic dispersal by a New World ant parasitoid. *Journal of Biogeography*, *43*(9), 1750–1761.

Nowak, R. M., & Walker, E. P. (1994). *Walker's bats of the world*. JHU Press.

O'Mara, M. T., Amorim, F., Scacco, M., McCracken, G. F., Safi, K., Mata, V., Tomé, R., Swartz, S., Wikelski, M., Beja, P., Rebelo, H., & Dechmann, D. K. N. (2021). Bats use topography and nocturnal updrafts to fly high and fast. *Current Biology*, *31*(6), 1311-1316.e4.

Omta, A. W., & Dijkstra, H. A. (2003). A physical mechanism for the Atlantic–Pacific flow reversal in the early Miocene. *Global and planetary Change*, *36*(4), 265-276.

Platt, R. N., Faircloth, B. C., Sullivan, K. A. M., Kieran, T. J., Glenn, T. C., Vandewege, M. W., Lee, T. E., Baker, R. J., Stevens, R. D., & Ray, D. A. (2018). Conflicting Evolutionary Histories of the Mitochondrial and Nuclear Genomes in New World *Myotis* Bats. *Systematic Biology*, *67*(2), 236–249.

Portik, D. M., & Wiens, J. J. (2021). Do Alignment and Trimming Methods Matter for Phylogenomic (UCE) Analyses? *Systematic Biology*, *70*(3), 440–462.

Prjibelski, A., Antipov, D., Meleshko, D., Lapidus, A., & Korobeynikov, A. (2020). Using SPAdes De Novo Assembler. *Current Protocols in Bioinformatics*, *70*(1), e102.

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Rambaut, A., Drummond, A. J., Xie, D., Baele, G., & Suchard, M. A. (2018). Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7. *Systematic Biology*, *67*(5), 901–904.

Ramírez-Chaves, H. E., Morales-Martínez, D. M., Pérez, W. A., Velásquez-Guarín, D., Mejía-Fontecha, I. Y., Ortiz-Giraldo, M., Ossalópez, P. A., & Rivera Páez, F. A. (2021). A new species of small *Eptesicus* Rafinesque (Chiroptera: Vespertilionidae) from northern South America. *Zootaxa*, *5020*(3), 489–520.

Reis, M. dos, & Yang, Z. (2011). Approximate Likelihood Calculation on a Phylogeny for Bayesian Estimation of Divergence Times. *Molecular Biology and Evolution*, *28*(7), 2161–2172.

Renner, S. (2004). Plant dispersal across the tropical Atlantic by wind and sea currents. *International Journal of Plant Sciences*, *165*(S4), S23-S33.

Rodríguez-Posada, M. E., Morales-Martínez, D. M., Ramírez-Chaves, H. E., Martínez-Medina, D., & Calderón-Acevedo, C. A. (2021). A new species of Long-eared Brown Bat of the genus *Histiotus* (Chiroptera) and the revalidation of *Histiotus colombiae*. *Caldasia*, *43*(2), 221–234.

Roehrs, Z. P., Lack, J. B., & van den Bussche, R. A. (2010). Tribal phylogenetic relationships within Vespertilioninae (Chiroptera: Vespertilionidae) based on mitochondrial and nuclear sequence data. *Journal of Mammalogy*, *91*(5), 1073–1092.

Rowe, D. L., Dunn, K. A., Adkins, R. M., & Honeycutt, R. L. (2010). Molecular clocks keep dispersal hypotheses afloat: evidence for trans-Atlantic rafting by rodents. *Journal of Biogeography*, *37*(2), 305–324.

Ruedi, M., & Mayer, F. (2001). Molecular Systematics of Bats of the Genus *Myotis* (Vespertilionidae) Suggests Deterministic Ecomorphological Convergences. *Molecular Phylogenetics and Evolution*, *21*(3), 436–448.

Ruedi, M., Stadelmann, B., Gager, Y., Douzery, E. J. P., Francis, C. M., Lin, L. K., Guillén-Servent, A., & Cibois, A. (2013). Molecular phylogenetic reconstructions identify East Asia as the cradle for the evolution of the cosmopolitan genus *Myotis* (Mammalia, Chiroptera). *Molecular Phylogenetics and Evolution*, *69*(3), 437–449.

Sánchez, R. T., Montani, M. E., Tomasco, I. H., Díaz, M. M., & Barquez, R. M. (2019). A new species of *Eptesicus* (Chiroptera, Vespertilionidae) from Argentina. *Journal of Mammalogy*, *100*(1), 118–129.

Sayyari, E., & Mirarab, S. (2016). Fast Coalescent-Based Computation of Local Branch Support from Quartet Frequencies. *Molecular Biology and Evolution*, *33*(7), 1654–1668.

Semedo, T. B. F., & Feijó, A. (2017). Filling the gap: first record of the transparent-winged big-eared bat *Histiotus diaphanopterus* (Chiroptera: Vespertilionidae) in southwestern Brazil. *Mammalia*, *81*(3), 323-327.

Schluter, D., & Pennell, M. W. (2017). Speciation gradients and the distribution of biodiversity. *Nature 2017 546:7656*, *546*(7656), 48–55.

Simmons, N.B. and A.L. Cirranello. 2020. *Bat Species of the World: A taxonomic and geographic database*. Accessed 2021. https://batnames.org

Smith, B. T., Harvey, M. G., Faircloth, B. C., Glenn, T. C., & Brumfield, R. T. (2014a). Target Capture and Massively Parallel Sequencing of Ultraconserved Elements for Comparative Studies at Shallow Evolutionary Time Scales. *Systematic Biology*, *63*(1), 83–95.

Smith, B. T., McCormack, J. E., Cuervo, A. M., Hickerson, M. J., Aleixo, A., Cadena, C. D., Pérez-Emán, J., Burney, C. W., Xie, X., Harvey, M. G., Faircloth, B. C., Glenn, T. C., Derryberry, E. P., Prejean, J., Fields, S., & Brumfield, R. T. (2014b). The drivers of tropical speciation. *Nature 2014 515:7527*, *515*(7527), 406–409.

Stadelmann, B., Lin, L. K., Kunz, T. H., & Ruedi, M. (2007). Molecular phylogeny of New World Myotis (Chiroptera, Vespertilionidae) inferred from mitochondrial and nuclear DNA genes. *Molecular Phylogenetics and Evolution*, *43*(1), 32–48.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, *30*(9), 1312–1313.

Sukumaran, J., & Knowles, L. L. (2017). Multispecies coalescent delimits structure, not species. *Proceedings of the National Academy of Sciences*, *114*(7), 1607–1612.

Swofford, D.L. and Sullivan, J., 2003. Phylogeny inference based on parsimony and other methods using PAUP*. The Phylogenetic Handbook: A Practical Approach to DNA and Protein Phylogeny, cáp, 7, pp.160-206.

Tagliacollo, V. A., & Lanfear, R. (2018). Estimating Improved Partitioning Schemes for Ultraconserved Elements. *Molecular Biology and Evolution*, *35*(7), 1798–1811.

Teeling, E. C., Springer, M. S., Madsen, O., Bates, P., O'Brien, S. J., & Murphy, W. J. (2005). A molecular phylogeny for bats illuminates biogeography and the fossil record. *Science*, *307*(5709), 580–584.

Tiley, G. P., Poelstra, J. W., dos Reis, M., Yang, Z., & Yoder, A. D. (2020). Molecular Clocks without Rocks: New Solutions for Old Problems. *Trends in Genetics*, *36*(11), 845–856.

Thomas, O. (1916). XXIX.—Notes on bats of the genus *Histiotus*. *Annals and Magazine of Natural History*, *17*(99), 272-276.

Tsang, S. M., Wiantoro, S., Veluz, M. J., Sugita, N., Nguyen, Y. L., Simmons, N. B., & Lohman, D. J. (2020). Dispersal out of Wallacea spurs diversification of *Pteropus* flying foxes, the world's largest bats (Mammalia: Chiroptera). *Journal of Biogeography*, *47*(2), 527–537.

Turmelle, A. S., Kunz, T. H., & Sorenson, M. D. (2011). A tale of two genomes: contrasting patterns of phylogeographic structure in a widely distributed bat. *Molecular Ecology*, *20*(2), 357–375.

Uit de Weerd, D. R., & Gittenberger, E. (2013). Phylogeny of the land snail family Clausiliidae (Gastropoda: Pulmonata). *Molecular Phylogenetics and Evolution*, *67*(1), 201–216.

Upchurch, P. (2008). Gondwanan break-up: legacies of a lost world? *Trends in Ecology & Evolution*, *23*(4), 229–236.

Velazco, P. M., Almeida, F. C., Cláudio, V. C., Giménez, A. L., & Giannini, N. P. (2021). A New Species of *Histiotus* Gervais, 1856 (Chiroptera, Vespertilionidae), from the Pacific Coast of Northern Peru. *American Museum Novitates*, *2021*(3979), 1-30.

Weir, J. T., & Schluter, D. (2007). The latitudinal gradient in recent speciation and extinction rates of birds and mammals. *Science*, *315*(5818), 1574–1576.

Yang, Z., & Rannala, B. (2006). Bayesian Estimation of Species Divergence Times Under a Molecular Clock Using Multiple Fossil Calibrations with Soft Bounds. *Molecular Biology and Evolution*, *23*(1), 212–226.

Yang, Z. (2007). PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Molecular Biology and Evolution*, *24*(8), 1586–1591.

Yang, Z. (2015). The BPP program for species tree estimation and species delimitation. *Current Zoology*, *61*(5), 854–865.

Yi, X., & Latch, E. K. (2022) Nuclear phylogeography reveals strong impacts of gene flow in big brown bats. *Journal of Biogeography*. (accepted)

Yu, G. (2020). Using ggtree to Visualize Data on Tree-Like Structures. *Current Protocols in Bioinformatics*, *69*(1), e96.

Zachos, J., Pagani, H., Sloan, L., Thomas, E., & Billups, K. (2001). Trends, rhythms, and aberrations in global climate 65 Ma to present. *Science*, *292*(5517), 686–693.

Zhang, C., Rabiee, M., Sayyari, E., & Mirarab, S. (2018). ASTRAL-III: Polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics*, *19*(6), 15–30.

Chapter III. Nuclear phylogeography reveals strong impacts of gene flow in big brown bats

Xueling Yi[1] and Emily K. Latch[1]

[1] Department of Biological Sciences, University of Wisconsin-Milwaukee, Milwaukee, Wisconsin, 53211, USA

ABSTRACT

Understanding speciation mechanisms requires disentangling processes that promote and erode population-level divergence. Three hypotheses are raised that contemporary population structure is mainly shaped by refugial isolation, gene flow, or both. Testing these hypotheses requires range-wide phylogeography and integrative analyses across scales. Here we aimed to 1) re-estimate the previously unresolved nuclear divergence within a widespread bat; 2) test the above three phylogeographic hypotheses; and 3) inform conservation management under climate change. We used the widespread big brown bat (*Eptesicus fuscus*) as the model system. We collected range-wide samples and genome-wide markers using restriction site-associated DNA sequencing. Population structure was analyzed by clustering methods and spatial estimations. Nuclear phylogeographic divergence was estimated using tree methods (concatenation and coalescence) and network analyses (TreeMix). Phylogeographic hypotheses were tested by comparing alternative evolutionary scenarios using demographic modeling. Species distribution modeling was used to help identify Pleistocene refugia and predict future range shifts under climate change. We identified three populations in the Caribbean, Eastern, and Western North America. The west population further split into three phylogeographic clades in Pacific, Southwestern North America, and Mexico. Discordances among mitochondrial and nuclear topologies reflected strong impacts of gene flow without sex bias. Demographic modeling supported scenarios of historical isolation followed by secondary gene flow and estimated Holocene divergence times. Species distribution was essentially continuous during glaciation with possible regional isolation, and northward range shifts were predicted under future climate change. Our results supported the hypothesis that the combined effects of historical isolation and secondary gene flow shaped the contemporary population divergence of big brown bats. We

showed that climate change probably triggered the initial divergence and that gene flow had a strong impact on the observed nuclear phylogeographic divergence. Our empirical study demonstrated that dynamic within-species processes had generated population divergences without speciation.

Introduction

Speciation begins with population divergence that can be driven by multiple factors such as geographic isolation and climate change (Rundle & Nosil 2005). However, the factors that promote divergence can change over time, and populations may not complete the process of speciation if divergence is eroded by high levels of gene flow. Repeated interactions between divergence and gene flow can generate complex patterns of genetic structure, the delineation of which remains challenging but important for understanding evolutionary processes and guiding conservation management.

Quaternary climate changes, especially the Pleistocene glaciation, have had an important influence on contemporary genetic divergences (Hewitt 1996; Hewitt 2004; Weir & Schluter 2004). Extensive ice sheets during glaciation constrained populations in suitable habitats and drove divergences between refugia, while warmer climates during inter- and post-glacial periods allowed range expansion and re-established gene flow that potentially merged refugial populations. Pleistocene glaciation cycles thus caused repeated interactions between refugial isolation and gene flow, and their competing effects on contemporary genetic structure can be summarized in three hypotheses. First, the refugial hypothesis proposes that contemporary population structure mainly originated from historical isolation, especially during the Last Glacial Maximum (LGM) when populations might have been restricted within refugia (Brunsfeld et al. 2001; Gómez & Lunt 2007; Waltari et al. 2007; Puckett et al. 2015). Second, the gene flow hypothesis proposes weak population structure which mainly arose from spatial variation in gene flow, such as due to geographic barriers and heterogeneous range expansion (e.g., Miller et al. 2021). The third hypothesis proposes no dominant process and signals of both historical isolation and secondary gene flow, potentially generating more complex patterns such

as genetic diversity hotspots (Petit et al. 2003; Dufresnes et al. 2016) and cytonuclear discordance (Dufresnes et al. 2020). These three phylogeographic hypotheses cover various divergence levels in the spectrum between a single genetic unit and complete speciation. Delineating genetic complexities in this grey area requires range-wide phylogeography and integrative analyses of evolutionary processes across spatial and temporal scales.

Bats are a good system to study diversification and test the above hypotheses because their evolution has been largely triggered by climate change (Teeling et al. 2005; Ruedi et al. 2013), while their ability to fly may allow high levels of gene flow. Some studies have proposed that refugial isolation shaped contemporary population structure in bats (Dixon 2011; Razgour et al. 2013; Boston et al. 2015), while others stressed the effects of gene flow on divergence patterns (Morales & Carstens 2018). Here we tested the above three hypotheses in the big brown bat (*Eptesicus fuscus*), a non-migratory microbat widely distributed from southern Canada to northern South America and on most of the Caribbean islands (Kurta & Baker 1990; IUCN 2018; Fig 2.1). This species is the only representative of the genus *Eptesicus* in northern North America but it co-distributes with a few other *Eptesicus* species in Mexico and South America (Kurta & Baker 1990). Up to 11 subspecies have been recognized based on morphology and life history traits (Burnett 1983; Kurta & Baker 1990; Fig 2.1). Both subspecies distribution and previously characterized mitochondrial phylogeography (Turmelle et al. 2011) showed clear divergence between western and eastern North America, suggesting Pleistocene isolation between two glacial refugia (Neubaum et al. 2007). In contrast, a lack of range-wide nuclear structure supported the second hypothesis that post-glacial gene flow (or even male-biased gene flow; Turmelle et al. 2011) has largely merged continental populations. However, some nuclear divergence was detected in regional studies between New York and Arizona (Neubaum et al.

2007) and between western and eastern Canada (Nadin-Davis et al. 2010), indicating that the lack of range-wide nuclear structure shown by Turmelle et al. (2011) may be due to limited analytical power using two genes with inadequate variation. Therefore, genome-wide markers are needed to better characterize nuclear divergence and test the above phylogeographic hypotheses in big brown bats. The refugial hypothesis predicts similar patterns in nuclear and mitochondrial phylogeographic divergence corresponding to the historical isolation between putative glacial refugia. The gene flow hypothesis predicts weak population structure and minimal geographic divergence. The third hypothesis predicts signals of both historical isolation and secondary gene flow in nuclear phylogeography.

Accordingly, in this study we had three major goals. First, we aimed to delineate nuclear population structure and phylogeographic divergence of big brown bats using range-wide sampling and genome-wide single nucleotide polymorphisms (SNPs) obtained from restriction site-associated DNA (RAD) sequencing. Second, we aimed to test the three hypotheses about whether historical isolation, post-glacial gene flow, or combined effects of both best explained the observed nuclear divergence. Third, we aimed to identify evolutionarily significant units and inform conservation management of this widespread species that serves ecological systems over a large geographic range (Agosta 2002; Barbosa et al. 2018). Our study shed light on the dynamic population-level processes that played important roles in the early divergence that potentially might lead to speciation.

Materials and Methods

Sampling and next-generation sequencing

Samples of big brown bats and closely related bat species were obtained from researchers and museums (Appendix B Table S1). DNA was extracted from contemporary samples using Qiagen DNeasy Blood & Tissue Kit, and from museum specimens using an optimized phenol-chloroform protocol (Alminas et al. 2021). Next-generation sequencing libraries were prepared using a total of 310 samples (275 big brown bats and 35 other bat species) following the bestRAD protocol (Ali et al. 2016). Briefly, DNA was normalized into 200 ng (250 ng if from dry skin specimens) in 10 ul of solution per individual, digested with the restriction enzyme *SbfI-HF*, ligated with barcodes and bestRAD adapters (Ali et al. 2016), and pooled into 480 ul per library. Samples from tissue and DNA (n=268) were pooled into three libraries (maximum 96 samples per library) and sheared in a Qsonica Q500 sonicator with 4 cycles of 30 sec on and 59 sec off. Samples from dry skin (n=42) were more degraded and fragmented, and thus were pooled into a separate library without shearing. All libraries were purified using AMPure XP beads (size-selection for 300-350 bp) and Dynabeads, and prepared using the NEBNext Ultra DNA library prep kit for Illumina sequencing on NovaSeq 6000 (one S4 lane, 150 bp paired-end) by Novogene Corporation.

Bioinformatic filtering

Bioinformatic analyses were processed using the University of Wisconsin-Milwaukee High Performance Computing cluster. Raw sequencing data were processed using STACKS v2.2 (Rochette et al. 2019). The function *process_radtags* was used to demultiplex data, remove reads with missing RAD sites or low qualities (default average Q < 10), and trim reads into 140 bp. Demultiplexed sequences of the individuals analyzed in this study are available on NCBI (accession SAMN23585471- SAMN23585657). The big brown bat reference genome

(GCF_000308155.1_EptFus1.0) was downloaded from NCBI and indexed using the Burrows-Wheeler Alignment tool (*bwa*, Li & Durbin 2009). Demultiplexed samples were aligned to the indexed reference genome using *bwa mem* with default settings. Alignments were saved as bam files using SAMtools (Li et al. 2009) and processed in *gstacks* to remove unmapped reads (8.9%) or PCR duplicates (84.4%) and identify RAD loci. Below we filtered a total of five datasets for analyses in this study (summarized in Table S2).

First, we filtered a range-wide dataset of big brown bats for population genetic analyses. All 275 big brown bat samples were processed in *populations* to filter for > 50% genotyping rates and 3 minimum minor allele counts (i.e., $\geq$ 3x depth per locus for presence in at least two diploid individuals). RAD loci were ordered and only the first SNP of each locus was output to minimize linkage disequilibrium (same below). The output was further filtered iteratively (Table S3) in VCFtools 0.1.16 (Danecek et al. 2011) with gradually increased stringency to retain as many individuals and SNPs as possible (O'Leary et al. 2018). The final filtering for > 90% genotyping rates and < 30% missing data per individual retained 182 big brown bats and 2,928 SNPs in the range-wide dataset. The second and third datasets were created by separating the retained big brown bats into continent and island, each filtered independently in *populations* for 3 minimum minor allele counts, 0.05 minimum minor allele frequency, and above 80% (continent) or 100% (island) genotyping rates. We filtered for minor allele frequency only in these regional datasets because of higher genetic similarity among individuals, meaning that alleles with extremely low minor frequency would likely be errors. We retained 24,957 SNPs using the 174 continental individuals, and 2,549 SNPs using the 8 insular individuals.

Fourth, we generated a phylogeography dataset by filtering both the retained 182 big brown bats and five high-quality outgroup samples of the Old World *Eptesicus serotinus* and the Neotropical

E. furinalis and *E. chiriquinus*. These 187 individuals were filtered together in *populations* for loci present in at least two species, > 50% genotyping rates per species, and 3 minimum minor allele counts. The output was transformed into phylip and nexus formats using vcf2phylip (Ortiz 2019) with the default filtering for at least 4 samples per SNP, retaining 149,766 SNPs in the phylogeography dataset. We were able to retain a large number of SNPs here by using genetically distinct individuals and a relatively loose filtering for the benefit of more loci in phylogenetic analyses (e.g., Huang & Knowles 2016; Eaton et al. 2017).

Finally, we generated a reduced dataset with a smaller sample size for some downstream analyses that did not tolerate missing data. We selected big brown bat individuals with the least missing data from each population and phylogeographic clade (all eight from the Caribbean, six from East, four from each western clade, Fig 2.1), and one outgroup individual from *Eptesicus furinalis*. The selected 27 individuals were filtered in *populations* for 3 minimum minor allele counts, 100% genotyping rate, and output in treemix and vcf formats. The filtered vcf file retained 4,079 SNPs and was reformatted in vcf2phylip.

Population genetics and spatial estimations

Population structure was characterized using the range-wide and continental datasets. For model-based analyses, datasets were reformatted in PGDspider 2.1.1 (Lischer & Excoffier 2011) and PLINK v1.9 (Chang et al. 2015). STRUCTURE 2.3.4 (Pritchard et al. 2000) was run with 10 iterations for each K (number of populations) from one to six, with 500,000 burn-in runs, 500,000 MCMC repetitions, admixture ancestry, and correlated allele frequencies (Falush et al. 2003). We did not run higher K in STRUCTURE due to computational constraints. The optimal K

was determined using the Evanno method (Evanno et al. 2005) in StructureSelector (Li & Liu 2018). ADMIXTURE 1.3.0 (Alexander et al. 2009) was run with 10 replicates for each K from one to ten, and the 10-fold cross-validation to select the optimal K. Results were compressed in CLUMPAK (Kopelman et al. 2015) and plotted in R 4.0 (R Core Team 2020). Individuals were assigned to the population where their estimated ancestry was > 50%. Multivariate analyses were carried out in R with datasets transformed into the genlight format using vcfR 1.1 and adegenet 2.1.3 (Jombart & Ahmed 2011). Individuals were assigned to population clusters using *find.clusters* with all PCs retained, and the optimal K was selected by the Bayesian Information Criterion (BIC). The Discriminant Analysis of Principal Components (DAPC) was carried out using *dapc* with 22 PCs retained in the range-wide dataset and 10 PCs in the continental dataset, based on *a-score* estimations (Jombart et al. 2010; Jombart &Collins 2015). The island dataset contained too few individuals for population clustering and thus was only analyzed in a Principal Component Analysis (PCA) using *glPca* with all PCs retained. This dataset contained no missing data to bias PCA analyses (Yi & Latch 2022).

To better illustrate spatial genetic patterns, we used the estimated effective migration surfaces (EEMS) to map genetic heterogeneity under the null model of isolation by distance and visualize spatial structure unidentified in clustering methods (Petkova et al. 2016). A habitat polygon was created using the Google Maps API v3 Tool (http://www.birdtheme.org/useful/v3tool.html). The habitat was equally divided into deme grids (density parameter nDemes=800, based on trial runs), and individuals were grouped into their closest deme grid by EEMS. The EEMS function *bed2diffs_v2* was used to calculate genetic dissimilarities between or within demes using the range-wide dataset. EEMS was run with three independent chains, each using 5,000,000 burn-in

runs and 5,000,000 MCMC iterations, and convergence was checked using the posterior trace plot. Results from all three chains were combined and plotted using *eems.plots*.

Characterizing nuclear phylogeographic divergence

To characterize the nuclear phylogeographic divergence, we first built a concatenated maximum likelihood (ML) tree in RAxML v8 (Stamatakis 2014) using the phylogeography dataset, the GTRGAMMA model, and the rapid bootstrap analysis (-f a) of 100 rapid bootstraps (-#) followed by searches for the best-scoring ML tree. Trees were rooted (-o) by the Old World outgroup *Eptesicus serotinus*, and the best-scoring tree was visualized in R using ggtree (Yu 2020). Individuals were assigned to phylogeographic clades based on the best ML tree. Because concatenation methods could suffer from incomplete lineage sorting (ILS) and might generate inaccurate topologies (e.g., Degnan & Rosenberg 2009), we also constructed a multispecies coalescent tree using the quartet-based method SVDquartets (Chifman & Kubatko 2014) implemented in PAUP4 (Swofford 2003). Big brown bats in the phylogeography dataset were assigned to operational taxonomic units representing nuclear populations or the ML phylogeographic clades. We evaluated all quartets with 100 standard bootstraps and constructed a 50% majority-rule consensus tree which was then rooted by *Eptesicus serotinus* and visualized in FigTree (http://tree.bio.ed.ac.uk/software/figtree).

Unsurprisingly, forcing tree methods to estimate genetic distances among conspecific individuals in the phylogeography dataset generated short tree branches, paraphyletic patterns, and discordant topologies. To clarify the tree topology and to test if discordances arose from sampling design, we divided the phylogeography dataset (using VCFTools and vcf2phylip) in

two ways and re-ran the tree analyses. First, we generated a subset (n=128) without Southwest individuals. Second, we generated a "pure" subset (n=35, five individuals per clade) by removing the individuals with Q-scores < 0.95 in ADMIXTURE and STRUCTURE (results of optimal K using the range-wide dataset), and removing the western individuals with relatively shorter branches in the complete ML tree. We ran RAxML and SVDquartets using these two subsets as described before. In addition, we ran SVDquartets without taxonomic assignments so that each tip represented one individual in the subsets.

Because gene flow among phylogeographic clades would violate tree assumptions, we also estimated nuclear phylogeographic divergence using the network method TreeMix1.13 (Pickrell & Pritchard 2012). TreeMix estimates population splits (based on maximum likelihood) with migration edges added to account for the tree-model unexplained covariance (Pickrell & Pritchard 2012). However, the TreeMix assumption of instantaneous gene flow suits scenarios of island colonization but does not apply to recurrent continental gene flow. Therefore, we mainly used TreeMix to estimate the Caribbean clade which had different positions in the ML and quartet-based trees (see Results), using the reduced dataset with continental individuals assigned to phylogeographic clades and insular individuals assigned to islands. We tested the number of migration edges (m) from 0 to 9, with 10 replicates for each m, the bootstrap command, no sample size correction, global rearrangements, and rooted by *Eptesicus furinalis*. The optimal m was chosen by the Evanno method in OptM (Fitak 2021) and results were visualized using the TreeMix function *plot_tree*. For comparison, we also ran RAxML and SVDquartets on the reduced dataset as described above.

Testing phylogeographic hypotheses in demographic modeling

We tested phylogeographic hypotheses by comparing the data fit of alternative evolutionary scenarios modeled in fastsimcoal 2.6 (Excoffier & Foll 2011; Excoffier et al. 2013). Because all samples were mapped to the big brown bat reference genome, we repolarized the reduced dataset using a custom R script vcf-repo (https://github.com/xuelingyi/repolarize-vcf) to filter for the loci that were homozygous in the outgroup and label outgroup alleles as the ancestral state. The repolarized dataset (26 big brown bats, 3,947 SNPs) was processed in vcf2sfs (Marques et al. 2019) to generate multidimensional derived site frequency spectrum (DSFS). We tested models of the three nuclear populations, the western phylogeographic clades, and all phylogeographic clades. The refugial hypothesis was modeled as strict isolation between clades; the gene flow hypothesis was modeled as continuous gene flow since divergence; and the third hypothesis was modeled as historical isolation followed by secondary gene flow. For simplicity and practicability, we only modeled symmetrical gene flow between adjacent continental clades, with constant population sizes based on a previous study (Chattopadhyay et al. 2019). We also tested the monophyly of western clades and the fit of tree topologies by comparing demographic models.

Each model had 50 independent runs, and each run 200,000 simulations (fsc26 command -n), 40 optimization (ECM) cycles (-L), and the maximum likelihood estimation (-M). Only the SFS counts > 10 were used (-C10) and monomorphic sites were ignored (-0). The ancestral population size was fixed as 500,000 diploid individuals (Chattopadhyay et al. 2019) for parameter estimation (Excoffier et al. 2013). Among the 50 replicates of each model, the best (i.e., with the highest maximum likelihood) was used to calculate the Akaike information criterion (AIC) using the script calculateAIC (https://github.com/speciationgenomics/scripts/blob/master/calculateAIC.sh). The model with

the lowest AIC was regarded as the better fit among tested scenarios, and the better-fit model including all phylogeographic clades was used to estimate demographic parameters.

Species distribution under climate change

Species distribution modeling (SDM) was used to identify putative Pleistocene refugia and predict range shifts under future climate change to inform conservation management. Occurrences of big brown bats with specimen records and coordinates were downloaded from the Global Biodiversity Information Facility (GBIF, https://doi.org/10.15468/dl.lpstz5), cleaned by removing spatial duplicates and records outside the current range (IUCN 2018), and filtered within years 1960-1990 for temporal consistency with bioclimatic data. To remove sampling bias (Elith et al. 2011), the retained occurrences were spatially thinned in spThin (Aiello-Lammes et al. 2015) by keeping records at least 100 km apart. Bioclimatic data for 19 variables were downloaded from WorldClim v1 (Hijmans et al. 2005) at the spatial resolution of 2.5 arc minutes and were trimmed around the current range in QGIS v3. We downloaded climatic data for the current time (1960-1990), mid-Holocene (~6,000 years ago), LGM (~22,000 years ago), and future (years 2050, 2070). The same Global Climate Models (CCSM4, MIROC-ESM, MPI-ESM-P) were used for the LGM and mid-Holocene. Future climates (only CCSM4 and MIROC-ESM available) were downloaded under two extreme climatic scenarios representing minimal (RCP2.6) and maximal (RCP8.5) greenhouse gas emissions.

SDM was constructed in MaxEnt 3.4 (Phillips et al. 2006; Phillips et al. 2017) using default settings unless otherwise stated. An initial model was built using all 19 bioclimatic variables for the current time, 65% occurrences for training and the rest for testing. We calculated pairwise

Pearson correlation coefficients for the 19 bioclimatic variables in ENMTools 1.3 (Warren et al. 2010), and selected between each highly correlated pair ($|r| > 0.8$) the variable that had a higher percentage contribution in the initial model (Warren et al. 2014). The variable bio18 was excluded as it made no contribution to the initial model. We then optimized MaxEnt parameters using the selected bioclimatic variables, the same training and testing occurrences, and combinations of 10 regularization multiplier (β) values (0.5 to 5 with an increment of 0.5) and 8 feature classes (L, H, LQ, LQP, LQH, LQPT, LQPH, LQPTH, Phillips et al. 2006; Elith et al. 2011). The 80 generated models were compared by AICc scores calculated in ENMTools. The optimal model (i.e., with the lowest AICc, and deltaAICc > 2) was run with 10-fold cross-validation to estimate the areas under the ROC curve (AUC) and response curves of present-day bioclimatic variables. The optimal model was then projected on to the other time periods using the corresponding bioclimatic data. QGIS was used to assemble results by averaging presence probabilities from different Global Climate Models at the same time period, and to calculate delta presence probabilities between future and present models.

Results

Nuclear population structure and phylogeographic patterns

We identified three population clusters in the Caribbean islands, Eastern North America, and Western North America using the range-wide dataset (Fig 2.1, Fig 2.2; Appendix B Fig S1, Fig S2). The continental dataset showed an optimal K of 2, corresponding with west and east populations, but ADMIXTURE results at K of 3 only had slightly higher cross-validation errors and implied a third continental cluster on the Pacific coast (Appendix B Fig S1, Fig S3).

Consistently, EEMS showed reduced effective migration around the Caribbean and between Western and Eastern North America, supporting the division of three major genetic populations, and lower effective migration among putative regional clades (Fig 2.3a; Appendix B Fig S4a). Higher effective genetic diversity was mapped along the Pacific coast and in southwestern and eastern North America (Fig 2.3b; Appendix B Fig S4b), indicating either long-persisting refugial populations or diversity generated by secondary contact (Petit et al. 2003).

The best-scoring ML tree of the phylogeography dataset supported three population clusters and indicated additional phylogeographic divergence within the west population (Fig 2.2; Appendix B Fig S5), roughly consistent with subspecies distributions (Fig 2.1). Specifically, the ML tree showed four monophyletic clades in the Caribbean (multiple insular subspecies), Eastern North America (*E. f. fuscus*), Pacific (*E. f. bernardinus*), and Mexico (*E. f. miradorensis*), and a paraphyletic clade in Southwestern North America (*E. f. pallidus*). All but four individuals were consistently assigned to the same population cluster across model-based methods, multivariate analyses, and the ML tree (Fig 2.1). The quartet-based consensus trees showed different topologies when using different taxonomic assignments and all topologies differed from the ML tree (Fig 2.4a), indicating effects of both ILS and gene flow. Subsets of the phylogeography dataset generated the same results and tree discordances (Appendix B Fig S6, Fig S7), rejecting potential bias from sampling design. In addition, quartet-based trees without taxonomic assignments showed a new topology that supported the monophyly of all western individuals (Appendix B Fig S6, Fig S7). Interestingly, none of the nuclear topologies agreed with the mitochondrial topology (Fig 2.4), although the two genomes showed highly similar phylogeographic clades. The reduced dataset generated similar trees but at lower resolution (Appendix B Fig S8), probably because of a limited number of loci.

Divergence and colonization in the Caribbean

Both the Caribbean clade in the ML tree and PCA of the island dataset showed evidence of geographic divergence among individuals grouped by islands (Fig 2.5a; Appendix B Fig S1), indicating strong isolation and genetic drift in the Caribbean. On the other hand, tree discordances indicated gene flow effects. All quartet-based nuclear trees showed an early divergence of the Caribbean clade from the continental clades, whereas the ML nuclear tree and mitochondrial phylogeography showed a more recent divergence of the Caribbean from the East after the west-east split (Fig 2.2a, Fig 2.4; Appendix B Fig S6, Fig S7). TreeMix analyses supported an early divergence of the Caribbean clade and added a migration edge from the East to the Caribbean and among Caribbean islands (Fig 2.5b), indicating that recolonization and gene flow may lead to the reconstruction of a sister relationship between these clades. We identified optimal four migration edges (Appendix B Fig S9) but found different positions of these edges among replicate runs. Because we identified loci using the ingroup reference genome, the ingroup-outgroup genetic distance was possibly underestimated and such migration edges were less reliable. Thus we presented the result without ingroup-outgroup edges and with the highest likelihood. Strong migration edges among western clades (Fig 2.5b) supported our result of one western population cluster (Fig 2.2b), but it should be noted that TreeMix assumes instantaneous gene flow (Pickrell & Pritchard 2012) which may not suit continental clades.

Historical isolation followed by secondary gene flow

We used demographic modeling to test the three hypotheses that contemporary divergence resulted from only historical isolation, continuous gene flow patterns, or historical isolation combined with secondary gene flow. Models of three nuclear populations supported a simultaneous divergence with historical isolation followed by secondary gene flow (Appendix B Fig S10a). Models of the three western clades supported the same scenario but showed the best fit (i.e., lowest AIC) when the Southwest clade diverged first (Appendix B Fig S10b). This result contradicted the geographical distribution of those clades and probably reflected the unmodeled gene flow between Southwest and East. In further support of this interpretation, models of all five phylogeographic clades showed that the Pacific clade diverged first, before the split between the Southwest and Mexico, and that all western clades form a monophyletic group (Appendix B Fig S10c, parameter estimates in the legend). Among the five-clade models of tree topology, the best fit was found in the quartet-based nuclear tree with tips representing individuals, which also supported western monophyly (Appendix B Fig S11). The tree-topology models were greatly improved when gene flow was added (Appendix B Fig S11), but still none was comparable to the alternative scenarios we tested (Appendix B Fig S10).

Species distribution under climate change

The SDM models were constructed using 328 spatially thinned occurrence records, seven selected bioclimatic variables (Appendix B Table S4, Fig S12), and the optimal MaxEnt parameters (β=2.5, feature class LQPT). The current model (1960-1990) had AUC scores of 0.856 ± 0.015 (standard deviation) and indicated good model fit. The Mean Temperature of Coldest Quarter (variable bio11) showed the largest impact on species distribution models with a bell-shaped quadratic response curve (Appendix B Fig S12). The LGM projection showed a

roughly continuous refugium in southern North America, possibly with multiple local refugia, such as in the Pacific southwest, Mexico, and the Florida Gulf Coast (Fig 6a). The mid-Holocene projection showed post-glacial range expansion into a widespread distribution similar to the current range (Fig 6b, c). Future (years 2050 and 2070) projections predicted an increase of suitable habitats in Canada and Alaska and potential habitat loss in South America, Mexico, and the Caribbean (Fig 6d; Appendix B Fig S13). The predicted northward range shifts were most pronounced in 2070 under the climatic scenario of the maximum greenhouse gas emission (RCP8.5, Fig 6d).

Discussion

We characterized fine scale nuclear phylogeographic patterns that supported the regional nuclear studies (Neubaum et al. 2007; Nadin-Davis et al. 2010), range-wide mitochondrial divergence (Turmelle et al. 2011), and morphological subspecies (Kurta & Baker 1990) of big brown bats. Using genome-wide SNPs and phylogeographic methods, we detected nuclear divergence that was masked by admixture in population genetic methods. Our analyses suffered from both assumption violations (e.g., gene flow in tree methods) and oversimplifications (e.g., demographic models). However, we highlight that these caveats were not simply methodological limitations, but in fact accurately reflected the nature of our study system which underwent dynamic processes of within-species divergence. By integrating alternative analyses with considerations of their limits, we showed that our results complemented (if not confirmed) each other and could be combined to shed light on the full picture of evolutionary processes. Our study provided an empirical example in which strong gene flow merged some phylogeographic

clades while weak gene flow maintained the population structure for others, giving a glimpse of the nuances of within-species divergence.

Strong effects of gene flow on nuclear phylogeography

Demographic modeling estimated the highest level of gene flow between the Pacific and Southwest, which might have generated an evolutionary melting pot (Petit et al. 2003; Dufresnes et al. 2016) and merged western phylogeographic clades into one population cluster. The West-East gene flow was also detected and corresponded with reported hybridization between *E. f. pallidus* and *E. f. fuscus*, such as in Kansas (Kunz 1974) and Nebraska (Hoffman & Genoways 2008). However, the estimated migration rate between Southwest and East was an order of magnitude lower than that between Southwest and Pacific (Appendix B Fig S10), supporting the Great Plains as a semi-permeable barrier to gene flow (Neubaum et al. 2007). Similarly, the EEMS-mapped areas of low effective migration overlapped with known geographical barriers such as the Great Plains, the Rocky Mountains, the Great Lakes, and the Cascade-Sierra Nevada Mountains (Engels 1936; Neubaum et al. 2007), which might have helped maintain the population structure. Nevertheless, EEMS results could be confounded by other factors such as historical isolation and local adaptation (Petkova et al. 2016). Putative adaptation may be important in enhancing population divergence and shaping subspecies morphology in big brown bats (Burnett 1983; Hoffman & Genoways 2008; Turmelle et al. 2011).

Our study provided an empirical example of how gene flow could affect tree-based analyses. Tree discordances have been normally attributed to ILS which is accounted for in multispecies coalescence (e.g., SVDquartets) but not concatenation (e.g., RAxML) methods, making the

former a preference for estimating species evolutionary history (e.g., Degnan & Rosenberg 2009; Jiang et al. 2020). However, tree discordances could also arise from gene flow which affects both concatenation and coalescent methods (Leaché et al. 2014; Solís-Lemus et al. 2016), which is commonly observed in the shallow phylogeography of recent and rapid radiations (e.g., Giarla & Esselstyn 2015) and population-level divergence (e.g., Rincon-Sandoval et al. 2019; this study). For example, SVDquartets assumes that individuals within the same taxonomic unit are more closely related to each other (i.e., with higher levels of gene flow) than individuals from different taxonomic units. As a result, grouping paraphyletic individuals into monophyletic taxa may lead to a loss of information and poor or even biased tree reconstruction, as indicated in our results. Our study thus highlighted the need to consider gene flow effects in the phylogeography of recently diverged genetic complexes.

Cytonuclear discordance caused by unbiased gene flow

Although we used a different set of samples from those of Turmelle et al. (2011), both studies analyzed the range-wide phylogeography of big brown bats and the identified genetic patterns are comparable. We showed that the cytonuclear discordance in big brown bats was not "extreme" but minor and comparable to the discordances among nuclear trees (Fig 2.4), indicating the effects of ILS and gene flow but no additional sex-biased processes. We thus rejected the suggestion by Turmelle et al. (2011) that cytonuclear discordance in big brown bats was due to male-biased gene flow homogenizing the nuclear genome while female philopatry drove mitochondrial divergence. Field studies of big brown bats also showed no evidence of extreme male-biased dispersal, but found that the species did not typically travel great distances (maximum 98 km, Beer 1955) and that gene flow occurred mainly by promiscuous mating at

swarming sites rather than long-distance individual dispersal (Veith et al. 2004; Vonhof et al. 2006). In addition, female dispersal and gene flow have been observed in big brown bats (Kurta & Baker 1990; Vonhof et al. 2008), rejecting strict female philopatry. We thus warn against invoking the assumed mammalian male-biased gene flow to explain cytonuclear discordance, which may mask the effects of other evolutionary processes (Lawson Handley & Perrin 2007; Zink & Barrowclough 2008; Toews & Brelsford 2012).

Historical divergence triggered by climate change

Demographic modeling supported the third hypothesis that contemporary population divergence is shaped by historical isolation combined with secondary gene flow (Appendix B Fig S10). The initial divergence was estimated in the Holocene (point estimation 6020 years ago, Appendix B Fig S10), indicating historical isolation during Holocene range expansion. However, this estimation might involve biases from sample size, genetic markers, prior parameter settings, and the underlying demographic model. Importantly, we used demographic modeling to test our three phylogeographic hypotheses rather than to estimate parameters, and the supported models are merely the better-fitting ones (based on AIC values) among our tests, whereas true demographic history would be much more complicated (i.e., with changes in population size and directional gene flow at different rates). Considering these limitations, it is possible that the actual divergence time was earlier than our estimation, dating back to the Pleistocene, which would indicate historical isolation in glacial refugia. Similarly, the LGM distribution showed an overall continuous range but indicated possible local refugia and isolation. Regardless of the exact divergence time, both the end Pleistocene and the early to mid-Holocene correspond with periods of climate change (Wanner et al. 2008; Clark et al. 2009) that had dramatic effects on species

distributions. Therefore, our data indicated that climate change might have triggered the initial phylogeographic divergence in big brown bats.

How population divergence may lead to speciation is an open question and possibly involves multiple mechanisms such as natural selection. We showed that divergent phylogeographic clades (e.g., Pacific and Southwest) could merge back to near panmixia (e.g., the west population cluster) under strong gene flow, potentially resulting in high genetic diversity but low species richness. Therefore, climate change alone may be insufficient to complete speciation, especially in highly dispersive and generalist taxa such as microbats (e.g., Morales & Carstens 2018). In such cases, complete speciation would require further reinforcement of the initial divergence through potentially different mechanisms such as natural and sexual selection and co-evolution of biomes.

Divergence in the Caribbean

Divergence in the Caribbean supported multiple insular subspecies (Fig 2.1; Fig 2.5a) but showed some differences from the distribution described by Kurta and Baker (1990). First, two subspecies were identified in the Bahamas islands, *E. f. bahamensis* and *E. f. dutertreus*, the latter also being present in Cuba, but our genetic data showed that all Bahamas individuals were more closely related to each other than to the Cuban individual. Second, the subspecies *E. f. hispaniolae* was identified both in Jamaica and the Dominican Republic, and an endemic subspecies (*E. f. wetmorei*) was identified in Puerto Rico. However, our genetic data showed that individuals from the Dominican Republic and Puerto Rico were more closely related to each other, while the Jamaican individual formed a distinct lineage in the Caribbean clade. However,

our results could be biased by limited samples in the Caribbean, and additional fine-scale sampling across the Caribbean islands would greatly help our understanding of the genetic and morphological diversity of these insular populations. Our results also indicated high levels of endemism and genetic drift in the Caribbean and supported open water as a strong barrier to gene flow in big brown bats (Burnett 1983), which was also suggested for several other Caribbean bats (e.g., Muscarella et al. 2011; Speer et al. 2017; Loureiro et al. 2019).

Phylogeographic analyses indicated an early divergence of the Caribbean clade followed by recolonization from eastern North America. Demographic modeling favored a simultaneous divergence of Caribbean and continental clades, but these data require caution as we did not model recolonization scenarios. Constrained species distributions in the LGM indicated that the common ancestor of big brown bats might have originated from southern North America, from where the species initially colonized the Caribbean and generated the early island-continent divergence. A similar colonization route has been reported in other widespread Caribbean bats (e.g., Loureiro et al. 2019). Holocene range expansion might have caused secondary island colonization from eastern North America. Occasional over-water migration of Caribbean bats may be facilitated by seasonal hurricanes in the Gulf of Mexico (Willig et al. 2010; Pedersen et al. 2013), helping maintain the viability of Caribbean populations.

Conservation management under climate change

Following the guidelines for defining conservation units (Barbosa et al. 2018), we suggest the five phylogeographic clades (Caribbean, East, Pacific, Mexico, Southwest) as distinct evolutionary units based on their historical isolation and detected divergence in nuclear and

mitochondrial DNA, and even morphology (Fig 2.1). We suggest the three populations (Caribbean, East, West) form distinct conservation units as they better represent current connectivity (Barbosa et al. 2018). Conservation should be prioritized in the Caribbean considering its genetic uniqueness, morphological diversity, small population sizes, and the predicted imminent and severe habitat loss under climate change. SDM supported strong effects of winter temperatures on species distributions (Table S4; Burnett 1983; Hoffman & Genoways 2008; Whitaker & Gummer 1992), indicating that range shifts might be the primary way for big brown bats to keep up with rapid climate change. However, caution is warranted, as we only modeled climatic variables whereas the realized distribution is also impacted by other factors such as insect abundance (Wagner et al. 2021), vegetation (Lenoir 2020), species ecological traits (Lyons et al. 2010), and human activities (Whitaker & Gummer 2000). In addition, adaptation to warmer climates (Kurta & Baker 1990) could make Caribbean populations less vulnerable or sensitive to climate change than predicted (Razgour et al. 2019), but the longevity of big brown bats (up to 19 years, Kurta & Baker 1990) might slow the process of adaptation to changing environments. Future studies with additional sampling on the Caribbean islands are needed to test the hypothesized local adaptation and to inform conservation management of these diverse insular populations.

Conclusion

In conclusion, we detected fine scale nuclear divergence with strong impacts of gene flow in big brown bat phylogeography. Our study tackled genetic complexities in early divergence governed by short evolutionary time, varying divergence mechanisms, and recurrent gene flow. The results

gave a snapshot of the highly dynamic within-species evolutionary processes and population-level divergence fluctuations in the absence of speciation.

Figure 2.1 Geographic distribution and sampling of big brown bats. The Mercator projection map shows the IUCN species distribution and the 11 subspecies based on Kurta and Baker (1990). Samples retained in this study (n=182) cover the range of all but three subspecies (in parenthesis). Each dot represents one individual and triangles represent those included in the reduced dataset (n=26, see Methods). Fill colors represent phylogeographic clades (labeled on the map) and outline colors represent the assigned population clusters in the Caribbean (orange), Eastern North America (red), and Western North America (blue). Four samples have inconsistent cluster assignments across methods and are shown with black outlines.

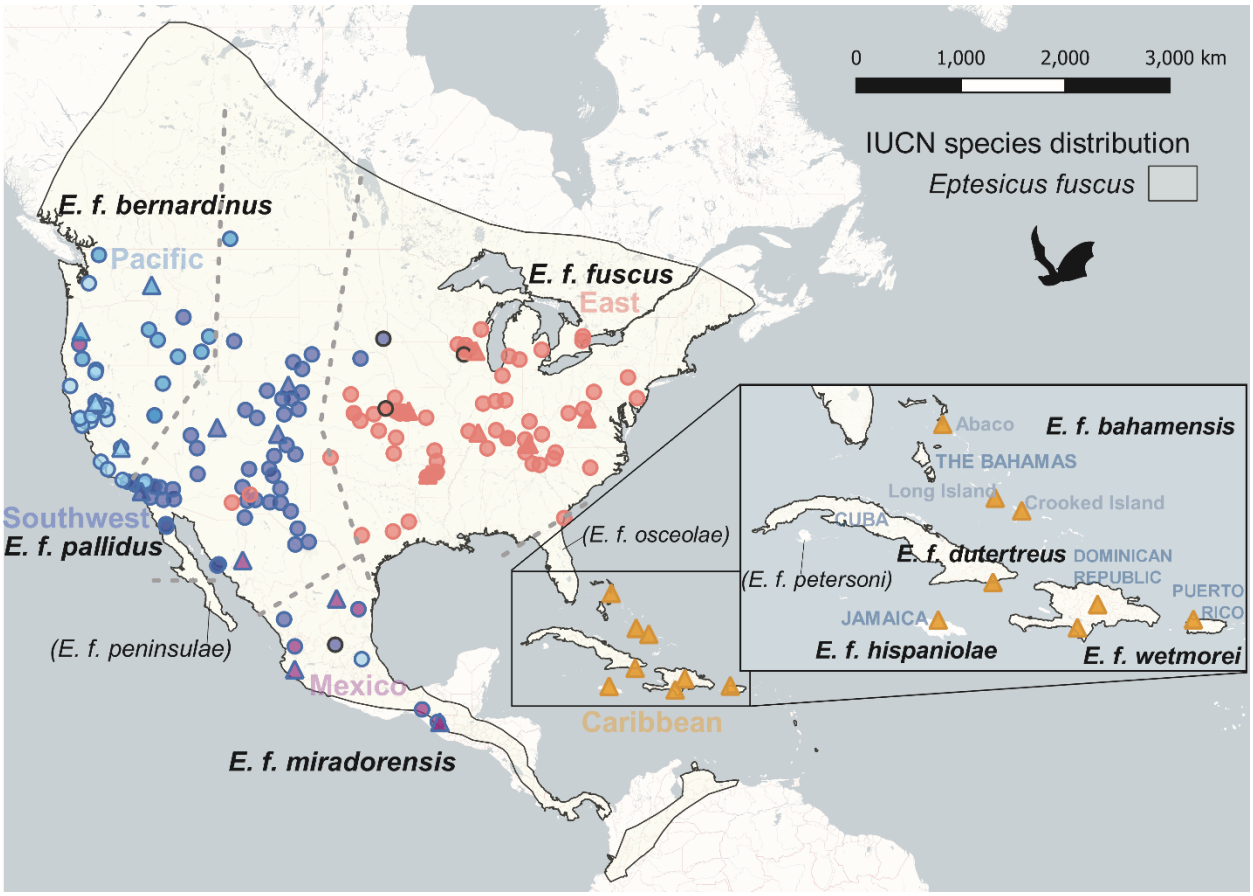


Figure 2.2 Maximum likelihood phylogeography and range-wide population clusters. a) The best-scoring maximum likelihood tree from RAxML with bootstraps > 75 labeled at the nodes. Each tip represents one individual (182 *Eptesicus fuscus*, the big brown bat, and 5 outgroups in the genus *Eptesicus*) and asterisks indicate those selected in the reduced dataset (n=27). Colors represent the assigned phylogeographic clades. The branch length scale bar is shown at the bottom and the outgroup branch is truncated. **b)** Bar plots of percent ancestry estimated in ADMIXTURE using the range-wide dataset (182 big brown bats) at K of three (optimal). Each bar represents one individual and is in the same order corresponding to the ML tree tips. Colors indicate population clusters: Eastern North America in red, the Caribbean in orange, and Western North America in blue.

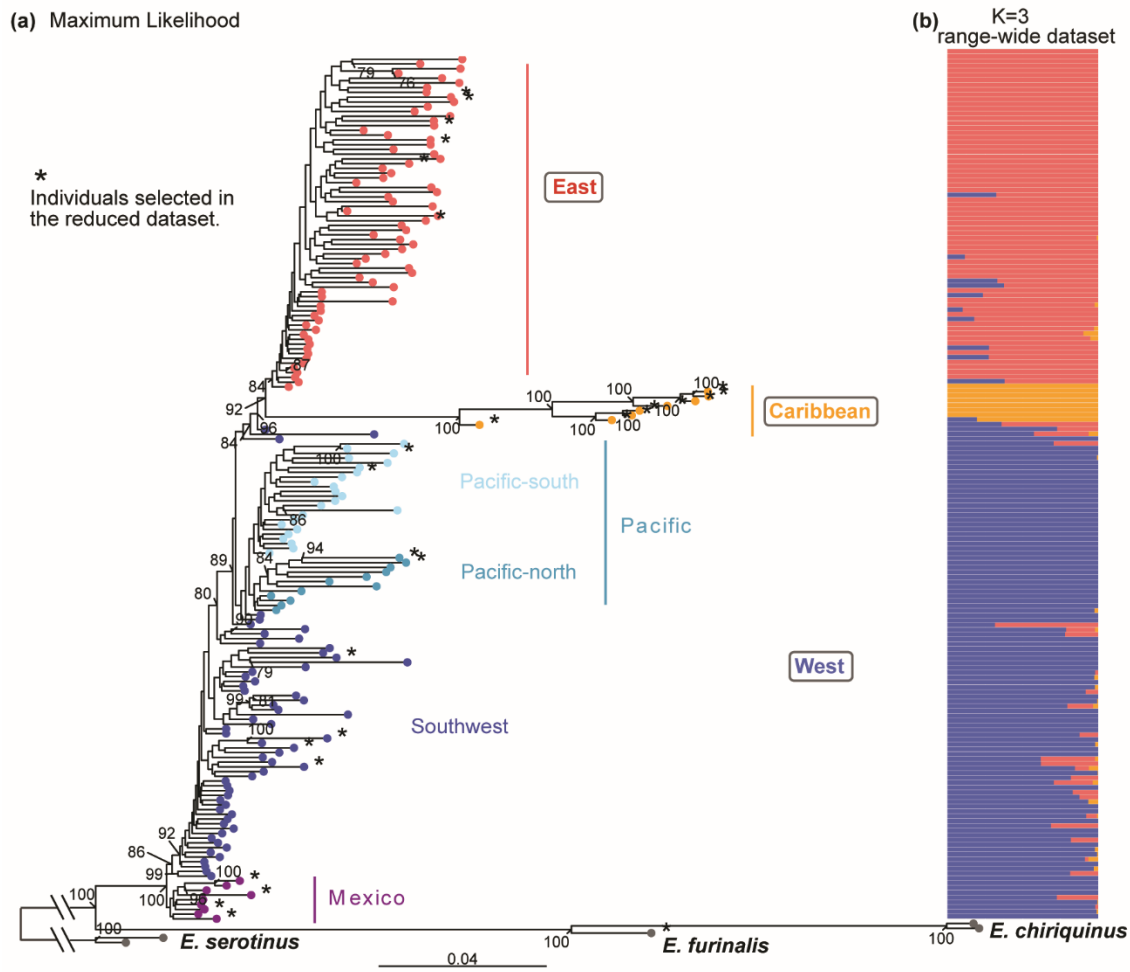


Figure 2.3 Estimated effective migration surfaces (EEMS). Mercator projection maps show black dots representing the deme grids with big brown bat samples and bigger dots indicating larger sample sizes. The division of all deme grids is shown in the supplementary Appendix B Fig S4. **a)** Effective migration rates estimated as genetic dissimilarities between demes. **b)** Effective diversity rates estimated as genetic dissimilarities between individuals within the same deme.

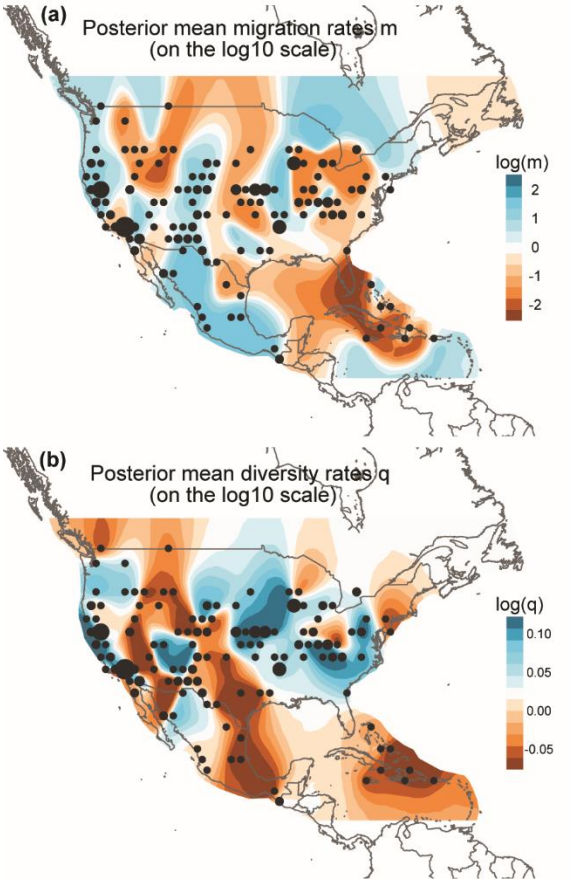


Figure 2.4 Discordances among quartet-based nuclear topologies and the mitochondrial topology of big brown bats. a) Nuclear trees estimated by the coalescent method SVDquartets using the phylogeography dataset. Western individuals grouped in different ways from one population to fully split phylogeographic clades. Numbers represent bootstrap supports. **b)** The mitochondrial phylogeography modified from Turmelle et al. (2011).

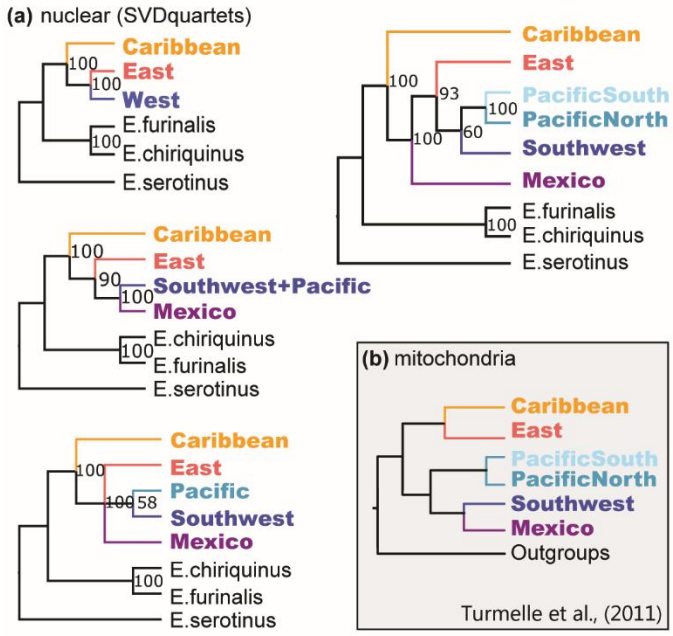


Figure 2.5 Divergence of big brown bats in the Caribbean and potential recolonization. a) PCA of the island dataset (n=8) and the Caribbean clade of the ML tree. PC1 is plotted on the y-axis and PC2 on the x-axis. Individuals are labeled by sampling sites. **b)** The TreeMix result showing optimal four migration edges among clades.

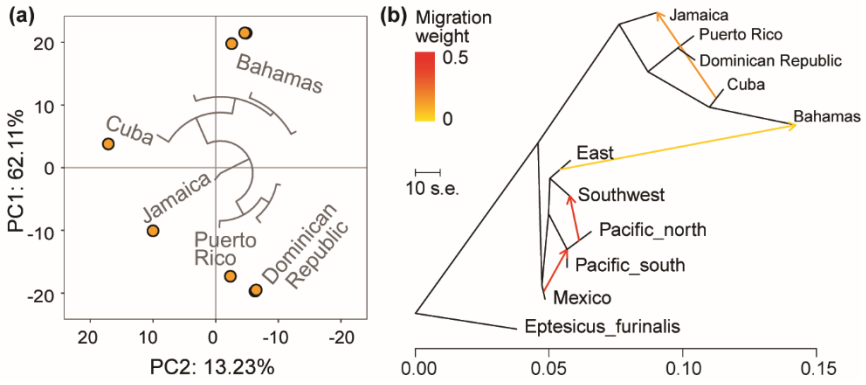
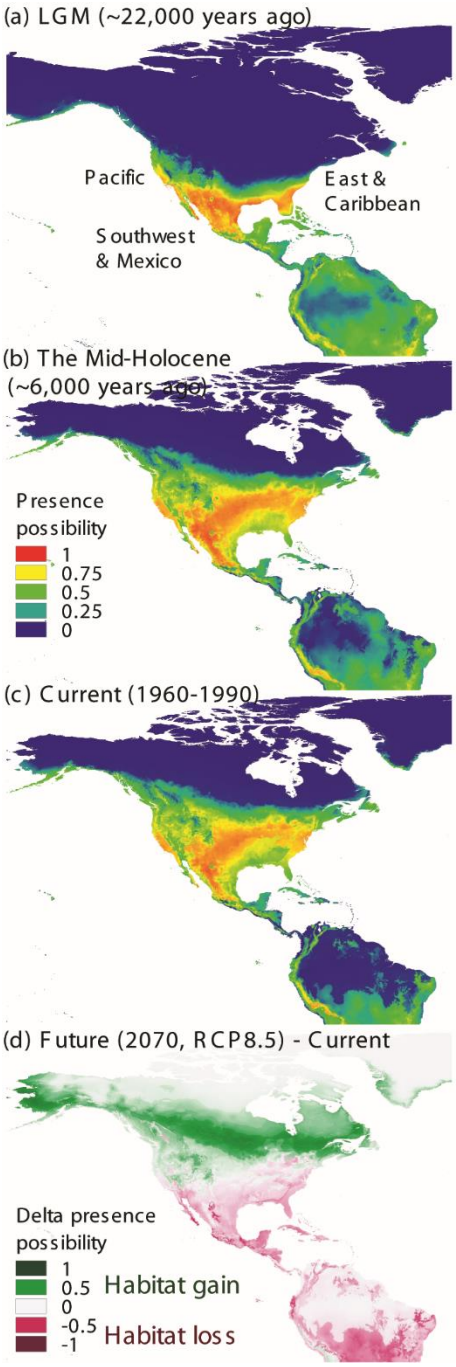


Figure 2.6 MaxEnt models of big brown bat distributions across evolutionary times. a) The Last Glacial Maximum**, b)** The mid-Holocene. **c)** The current time. Warmer colors indicate higher presence probabilities in the Mercator projection maps. **d)** The delta presence probability between the current model and the future model of maximal greenhouse gas emissions.



References

Agosta, S. J. (2002). Habitat use, diet and roost selection by the big brown bat (*Eptesicus fuscus*) in North America: a case for conserving an abundant species. *Mammal Review*, *32*(3), 179-198.

Aiello-Lammens, M. E., Boria, R. A., Radosavljevic, A., Vilela, B., & Anderson, R. P. (2015). spThin: an R package for spatial thinning of species occurrence records for use in ecological niche models. *Ecography*, *38*(5), 541-545.

Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome research*, *19*(9), 1655-1664.

Ali, O. A., O'Rourke, S. M., Amish, S. J., Meek, M. H., Luikart, G., Jeffres, C., & Miller, M. R. (2016). RAD capture (Rapture): flexible and efficient sequence-based genotyping. *Genetics*, *202*(2), 389-400.

Alminas, O. S., Heffelfinger, J. R., Statham, M. J., & Latch, E. K. (2021). Phylogeography of Cedros and Tiburón Island mule deer in North America's desert southwest. *Journal of Heredity*, *112*(3), 260-275.

Barbosa, S., Mestre, F., White, T. A., Paupério, J., Alves, P. C., & Searle, J. B. (2018). Integrative approaches to guide conservation decisions: using genomics to define conservation units and functional corridors. *Molecular Ecology*, *27*(17), 3452-3465.

Beer, J. R. (1955). Survival and movements of banded big brown bats. *Journal of Mammalogy*, *36*(2), 242-248.

Boston, E. S., Ian Montgomery, W., Hynes, R., & Prodöhl, P. A. (2015). New insights on postglacial colonization in western Europe: the phylogeography of the Leisler's bat (*Nyctalus leisleri*). *Proceedings of the Royal Society B: Biological Sciences*, *282*(1804), 20142605.

Brunsfeld, S. J., Sullivan, J., Soltis, D. E., & Soltis, P. S. (2001). Comparative phylogeography of northwestern North America: a synthesis. *Special Publication-British Ecological Society*, *14*, 319-340.

Burnett, C. D. (1983). Geographic and climatic correlates of morphological variation in *Eptesicus fuscus*. *Journal of Mammalogy*, *64*(3), 437-444.

Chattopadhyay, B., Garg, K. M., Ray, R., & Rheindt, F. E. (2019). Fluctuating fortunes: genomes and habitat reconstructions reveal global climate-mediated changes in bats' genetic diversity. *Proceedings of the Royal Society B*, *286*(1911), 20190304.

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience*, *4*(1), s13742-015.

Chifman, J., & Kubatko, L. (2014). Quartet inference from SNP data under the coalescent model. *Bioinformatics*, *30*(23), 3317-3324.

Clark, P. U., Dyke, A. S., Shakun, J. D., Carlson, A. E., Clark, J., Wohlfarth, B., ... & McCabe, A. M. (2009). The last glacial maximum. *Science*, *325*(5941), 710-714.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., ... & 1000 Genomes Project Analysis Group. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*(15), 2156-2158.

Dixon, M. D. (2011). Post-Pleistocene range expansion of the recently imperiled eastern little brown bat (*Myotis lucifugus lucifugus*) from a single southern refugium. *Ecology and evolution*, *1*(2), 191-200.

Dufresnes, C., Litvinchuk, S. N., Leuenberger, J., Ghali, K., Zinenko, O., Stöck, M., & Perrin, N. (2016). Evolutionary melting pots: a biodiversity hotspot shaped by ring diversifications around the Black Sea in the Eastern tree frog (*Hyla orientalis*). *Molecular Ecology*, *25*(17), 4285-4300.

Dufresnes, C., Nicieza, A. G., Litvinchuk, S. N., Rodrigues, N., Jeffries, D. L., Vences, M., ... & Martínez-Solano, Í. (2020). Are glacial refugia hotspots of speciation and cytonuclear discordances? Answers from the genomic phylogeography of Spanish common frogs. *Molecular Ecology*, *29*(5), 986-1000.

Eaton, D. A., & Ree, R. H. (2013). Inferring phylogeny and introgression using RADseq data: an example from flowering plants (Pedicularis: Orobanchaceae). *Systematic Biology*, *62*(5), 689-706.

Eaton, D. A., Spriggs, E. L., Park, B., & Donoghue, M. J. (2017). Misconceptions on missing data in RAD-seq phylogenetics with a deep-scale example from flowering plants. *Systematic Biology*, *66*(3), 399-412.

Elith, J., Phillips, S. J., Hastie, T., Dudík, M., Chee, Y. E., & Yates, C. J. (2011). A statistical explanation of MaxEnt for ecologists. *Diversity and Distributions*, *17*(1), 43-57.

Engels, W. L. (1936). Distribution of races of the brown bat (*Eptesicus*) in western North America. *American Midland Naturalist*, *17*(3), 653-660.

Evanno, G., Regnaut, S., & Goudet, J. (2005). Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Molecular ecology*, *14*(8), 2611-2620.

Excoffier, L., & Foll, M. (2011). Fastsimcoal: a continuous-time coalescent simulator of genomic diversity under arbitrarily complex evolutionary scenarios. *Bioinformatics*, *27*(9), 1332-1334.

Excoffier, L., Dupanloup, I., Huerta-Sánchez, E., Sousa, V. C., & Foll, M. (2013). Robust demographic inference from genomic and SNP data. *PLoS Genetics*, *9*(10), e1003905.

Falush, D., Stephens, M., & Pritchard, J. K. (2003). Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, *164*(4), 1567-1587.

Fitak, R. R. (2021). OptM: estimating the optimal number of migration edges on population trees using Treemix. *Biology Methods and Protocols*, *6*, bpab017.

Giarla, T. C., & Esselstyn, J. A. (2015). The challenges of resolving a rapid, recent radiation: empirical and simulated phylogenomics of Philippine shrews. *Systematic Biology*, *64*(5), 727-740.

Gómez, A., & Lunt, D. H. (2007). Refugia within refugia: patterns of phylogeographic concordance in the Iberian Peninsula. In S. Weiss & N. Ferrand (Eds.), *Phylogeography of southern European refugia* (pp. 155-188). Springer, Dordrecht.

Hewitt, G. M. (1996). Some genetic consequences of ice ages, and their role in divergence and speciation. *Biological Journal of the Linnean Society*, *58*(3), 247-276.

Hewitt, G. M. (2004). Genetic consequences of climatic oscillations in the Quaternary. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *359*(1442), 183-195.

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*(15), 1965-1978.

Hoffman, J. D., & Genoways, H. H. (2008). Characterization of a contact zone between two subspecies of the big brown bat (*Eptesicus fuscus*) in Nebraska. *Western North American Naturalist*, *68*(1), 36-45.

Huang, H., & Knowles, L. L. (2016). Unforeseen consequences of excluding missing data from next-generation sequences: simulation study of RAD sequences. *Systematic Biology*, *65*(3), 357-365.

IUCN. NatureServe and IUCN (International Union for Conservation of Nature) 2018. *Eptesicus fuscus. The IUCN Red List of Threatened Species. Version 2018.*1 http://oldredlist.iucnredlist.org/ Downloaded on 07 September 2018

Jiang, X., Edwards, S. V., & Liu, L. (2020). The multispecies coalescent model outperforms concatenation across diverse phylogenomic data sets. *Systematic Biology*, *69*(4), 795-812.

Jombart, T., Devillard, S., & Balloux, F. (2010). Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC genetics*, *11*(1), 1-15.

Jombart, T., & Ahmed, I. (2011). adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. *Bioinformatics*, *27*(21), 3070-3071.

Jombart, T., & Collins, C. (2015). A tutorial for discriminant analysis of principal components (DAPC) using adegenet 2.0. 0. *London: Imperial College London, MRC Centre for Outbreak Analysis and Modelling*.

Kopelman, N. M., Mayzel, J., Jakobsson, M., Rosenberg, N. A., & Mayrose, I. (2015). Clumpak: a program for identifying clustering modes and packaging population structure inferences across K. *Molecular Ecology Resources*, *15*(5), 1179-1191.

Kunz, T. H. (1974). Reproduction, growth, and mortality of the vespertilionid bat, *Eptesicus fuscus*, in Kansas. *Journal of Mammalogy*, *55*(1), 1-13.

Kurta, A., & Baker, R. H. (1990). *Eptesicus fuscus*. *Mammalian species*, (356), 1-10.

Lawson Handley, L. J., & Perrin, N. (2007). Advances in our understanding of mammalian sex-biased dispersal. *Molecular Ecology*, *16*(8), 1559-1578.

Leaché, A. D., Harris, R. B., Rannala, B., & Yang, Z. (2014). The influence of gene flow on species tree estimation: a simulation study. *Systematic Biology*, *63*(1), 17-30.

Lenoir, J. (2020). Rethinking climate context dependencies in biological terms. *Proceedings of the National Academy of Sciences, USA*, *117*(38), 23208-23210.

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, *25*(14), 1754-1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., ... & Durbin, R. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, *25*(16), 2078-2079.

Li, Y. L., & Liu, J. X. (2018). StructureSelector: a web-based software to select and visualize the optimal number of clusters using multiple methods. *Molecular Ecology Resources*, *18*(1), 176-177.

Lischer, H. E., & Excoffier, L. (2012). PGDSpider: an automated data conversion tool for connecting population genetics and genomics programs. *Bioinformatics*, *28*(2), 298-299.

Loureiro, L. O., Engstrom, M. D., & Lim, B. K. (2020). Comparative phylogeography of mainland and insular species of Neotropical molossid bats (*Molossus*). *Ecology and Evolution*, *10*(1), 389-409.

Lyons, S. K., Wagner, P. J., & Dzikiewicz, K. (2010). Ecological correlates of range shifts of Late Pleistocene mammals. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *365*(1558), 3681-3693.

Marques, D. A., Lucek, K., Sousa, V. C., Excoffier, L., & Seehausen, O. (2019). Admixture between old lineages facilitated contemporary ecological speciation in Lake Constance stickleback. *Nature Communications*, *10*(1), 1-14.

Miller, E. F., Leonardi, M., Beyer, R., Krapp, M., Somveille, M., Somma, G. L., ... & Manica, A. (2021). Post-glacial expansion dynamics, not refugial isolation, shaped the genetic structure of a migratory bird, the yellow warbler (*Setophaga petechia*). *bioRxiv*.

Morales, A. E., & Carstens, B. C. (2018). Evidence that *Myotis lucifugus* "subspecies" are five nonsister species, despite gene flow. *Systematic Biology*, *67*(5), 756-769.

Muscarella, R. A., Murray, K. L., Ortt, D., Russell, A. L., & Fleming, T. H. (2011). Exploring demographic, physical, and historical explanations for the genetic structure of two lineages of Greater Antillean bats. *PLoS One*, *6*(3), e17704.

Nadin-Davis, S. A., Feng, Y., Mousse, D., Wandeler, A. I., & Aris-Brosou, S. (2010). Spatial and temporal dynamics of rabies virus variants in big brown bat populations across Canada: footprints of an emerging zoonosis. *Molecular Ecology*, *19*(10), 2120-2136.

Neubaum, M. A., Douglas, M. R., Douglas, M. E., & O'Shea, T. J. (2007). Molecular ecology of the big brown bat (*Eptesicus fuscus*): genetic and natural history variation in a hybrid zone. *Journal of Mammalogy*, *88*(5), 1230-1238.

O'Leary, S. J., Puritz, J. B., Willis, S. C., Hollenbeck, C. M., & Portnoy, D. S. (2018). These aren't the loci you're looking for: principles of effective SNP filtering for molecular ecologists. *Molecular Ecology*, *27*, 3193–3206.

Ortiz, E.M. (2019). vcf2phylip v2.0: convert a VCF matrix into several matrix formats for phylogenetic analysis. doi:10.5281/zenodo.2540861

Pedersen, S. C., Genoways, H. H., Kwiecinski, G. G., Larsen, P. A., & Larsen, R. J. (2013). Biodiversity, biogeography, and conservation of bats in the Lesser Antilles.

Petkova, D., Novembre, J., & Stephens, M. (2016). Visualizing spatial population structure with estimated effective migration surfaces. *Nature genetics*, *48*(1), 94-100. http://dx.doi.org/10.1038/ng.3464

Petit, R. J., Aguinagalde, I., de Beaulieu, J. L., Bittkau, C., Brewer, S., Cheddadi, R., ... & Vendramin, G. G. (2003). Glacial refugia: hotspots but not melting pots of genetic diversity. *Science*, *300*(5625), 1563-1565.

Petit, R. J., & Excoffier, L. (2009). Gene flow and species delimitation. *Trends in Ecology & Evolution*, *24*(7), 386-393.

Phillips, S. J., Anderson, R. P., & Schapire, R. E. (2006). Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, *190*(3-4), 231-259.

Phillips, S. J., Anderson, R. P., Dudík, M., Schapire, R. E., & Blair, M. E. (2017). Opening the black box: an open-source release of Maxent. *Ecography*, *40*(7), 887-893.

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, *155*(2), 945-959.

Pickrell, J., & Pritchard, J. (2012). Inference of population splits and mixtures from genome-wide allele frequency data. *Nature Precedings*, 1-1.

Puckett, E. E., Etter, P. D., Johnson, E. A., & Eggert, L. S. (2015). Phylogeographic analyses of American black bears (*Ursus americanus*) suggest four glacial refugia and complex patterns of postglacial admixture. *Molecular Biology and Evolution*, *32*(9), 2338-2350.

R Core Team (2020). R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Razgour, O., Juste, J., Ibáñez, C., Kiefer, A., Rebelo, H., Puechmaille, S. J., ... & Jones, G. (2013). The shaping of genetic variation in edge-of-range populations under past and future climate change. *Ecology Letters*, *16*(10), 1258-1266.

Razgour, O., Forester, B., Taggart, J. B., Bekaert, M., Juste, J., Ibáñez, C., ... & Manel, S. (2019). Considering adaptive genetic variation in climate change vulnerability assessment reduces species range loss projections. *Proceedings of the National Academy of Sciences, USA*, *116*(21), 10418-10423.

Rincon-Sandoval, M., Betancur-R, R., & Maldonado-Ocampo, J. A. (2019). Comparative phylogeography of trans-Andean freshwater fishes based on genome-wide nuclear and mitochondrial markers. *Molecular Ecology*, *28*(5), 1096-1115.

Rochette, N. C., Rivera-Colón, A. G., & Catchen, J. M. (2019). Stacks 2: analytical methods for paired-end sequencing improve RADseq-based population genomics. *Molecular Ecology*, *28*(21), 4737-4754.

Ruedi, M., Stadelmann, B., Gager, Y., Douzery, E. J., Francis, C. M., Lin, L. K., ... & Cibois, A. (2013). Molecular phylogenetic reconstructions identify East Asia as the cradle for the evolution of the cosmopolitan genus *Myotis* (Mammalia, Chiroptera). *Molecular Phylogenetics and Evolution*, *69*(3), 437-449.

Rundle, H. D., & Nosil, P. (2005). Ecological speciation. *Ecology Letters*, *8*(3), 336-352.

Solís-Lemus, C., Yang, M., & Ané, C. (2016). Inconsistency of species tree methods under gene flow. *Systematic Biology*, *65*(5), 843-851.

Speer, K. A., Petronio, B. J., Simmons, N. B., Richey, R., Magrini, K., Soto-Centeno, J. A., & Reed, D. L. (2017). Population structure of a widespread bat (*Tadarida brasiliensis*) in an island system. *Ecology and Evolution*, *7*(19), 7585-7598.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, *30*(9), 1312-1313.

Swofford, D. (2003). PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4. Sinauer Associates.

Teeling, E. C., Springer, M. S., Madsen, O., Bates, P., O'Brien, S. J., & Murphy, W. J. (2005). A molecular phylogeny for bats illuminates biogeography and the fossil record. *Science*, *307*(5709), 580-584.

Toews, D. P., & Brelsford, A. (2012). The biogeography of mitochondrial and nuclear discordance in animals. *Molecular Ecology*, *21*(16), 3907-3930.

Turmelle, A. S., Kunz, T. H., & Sorenson, M. D. (2011). A tale of two genomes: contrasting patterns of phylogeographic structure in a widely distributed bat. *Molecular Ecology*, *20*(2), 357-375.

Veith, M., Beer, N., Kiefer, A., Johannesen, J., & Seitz, A. (2004). The role of swarming sites for maintaining gene flow in the brown long-eared bat (*Plecotus auritus*). *Heredity*, *93*(4), 342-349.

Vonhof, M. J., Barber, D., Fenton, M. B., & Strobeck, C. (2006). A tale of two siblings: multiple paternity in big brown bats (*Eptesicus fuscus*) demonstrated using microsatellite markers. *Molecular Ecology*, *15*(1), 241-247.

Vonhof, M. J., Strobeck, C., & Fenton, M. B. (2008). Genetic variation and population structure in big brown bats (*Eptesicus fuscus*): is female dispersal important? *Journal of Mammalogy*, *89*(6), 1411-1420.

Waltari, E., Hijmans, R. J., Peterson, A. T., Nyári, A. S., Perkins, S. L., & Guralnick, R. P. (2007). Locating Pleistocene refugia: comparing phylogeographic and ecological niche model predictions. *PLoS One*, *2*(7), e563.

Warren, D. L., Glor, R. E., & Turelli, M. (2010). ENMTools: a toolbox for comparative studies of environmental niche models. *Ecography*, *33*(3), 607-611.

Warren, D. L., Wright, A. N., Seifert, S. N., & Shaffer, H. B. (2014). Incorporating model complexity and spatial sampling bias into ecological niche models of climate change risks faced by 90 California vertebrate species of concern. *Diversity and Distributions*, *20*(3), 334-343.

Wagner, D. L., Grames, E. M., Forister, M. L., Berenbaum, M. R., & Stopak, D. (2021). Insect decline in the Anthropocene: Death by a thousand cuts. *Proceedings of the National Academy of Sciences*, *118*(2).

Wanner, H., Beer, J., Bütikofer, J., Crowley, T. J., Cubasch, U., Flückiger, J., ... & Widmann, M. (2008). Mid-to Late Holocene climate change: an overview. *Quaternary Science Reviews*, *27*(19-20), 1791-1828.

Weir, J. T., & Schluter, D. (2004). Ice sheets promote speciation in boreal birds. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *271*(1551), 1881-1887.

Willig, M. R., Presley, S. J., Bloch, C. P., & Genoways, H. H. (2010). 8. Macroecology of Caribbean Bats: Effects of Area, Elevation, Latitude, and Hurricane-Induced Disturbance. In T. Fleming & P. Racey (Eds.), *Island bats: Evolution, Ecology, and Conservation* (pp. 216-264). University of Chicago Press, Chicago.

Whitaker Jr, J. O., & Gummer, S. L. (1992). Hibernation of the big brown bat, *Eptesicus fuscus*, in buildings. *Journal of Mammalogy*, *73*(2), 312-316.

Whitaker Jr, J. O., & Gummer, S. L. (2000). Population structure and dynamics of big brown bats (*Eptesicus fuscus*) hibernating in buildings in Indiana. *The American Midland Naturalist*, *143*(2), 389-396.

Yi, X., & Latch, E. K. (2022) Nonrandom missing data can bias Principal Component Analysis inference of population genetic structure. *Molecular Ecology Resources*, *22*(2), 602-611.

Yu, G. (2020). Using ggtree to visualize data on tree-like structures. *Current Protocols in Bioinformatics*, *69*(1), e96.

Zink, R. M., & Barrowclough, G. F. (2008). Mitochondrial DNA under siege in avian phylogeography. *Molecular Ecology*, *17*(9), 2107-2121.

Chapter IV. Nonrandom missing data can bias PCA inference of population genetic structure

Xueling Yi[1], Emily K. Latch[1]

[1] Behavioral and Molecular Ecology Research Group, Department of Biological Sciences, University of Wisconsin-Milwaukee, Milwaukee, Wisconsin, USA

Abstract

Population genetic studies in non-model systems increasingly use next-generation sequencing to obtain more loci, but such methods also generate more missing data that may affect downstream analyses. Here we focus on the Principal Component Analysis (PCA) which has been widely used to explore and visualize population structure with mean-imputed missing data. We simulated data of different population models with various total missingness (1%, 10%, 20%) introduced either randomly or biased among individuals or populations. We found that individuals biased with missing data would be dragged away from their real population clusters to the origin of PCA plots, making them indistinguishable from true admixed individuals and potentially leading to misinterpreted population structure. We also generated empirical data of the big brown bat (*Eptesicus fuscus*) using restriction site-associated DNA sequencing (RADseq). We filtered three data sets with 19.12%, 9.87%, and 1.35% total missingness, all showing nonrandom missing data with biased individuals dragged towards the PCA origin, consistent with results from simulations. We highlight the importance of considering missing data effects on PCA in non-model systems where nonrandom missing data are common due to varying sample quality. To help detect missing data effects, we suggest to 1) plot PCA with a color gradient showing per sample missingness, 2) interpret samples close to the PCA origin with extra caution, 3) explore filtering parameters with and without the missingness-biased samples, and 4) use complementary analyses (e.g., model-based methods) to cross-validate PCA results and help interpret population structure.

Introduction

Studies of population structure provide a critical foundation for understanding ecology, evolution, and conservation management. Population structure indicates genetic divergence due to evolutionary and ecological processes such as migration, isolation, and adaptation. Characterizing contemporary population divergence also helps to delineate management units and conservation priorities, especially in non-model systems that are under-studied. Population delineation has been largely facilitated by the reduced-representation next-generation sequencing methods, such as restriction site-associated DNA sequencing (RADseq, Andrews et al. 2016), because these methods do not require prior knowledge of the genome and can sequence hundreds of thousands of loci with limited budgets. However, next-generation sequencing methods also intrinsically generate large amounts of missing data. Random missing data can arise from stochastic processes such as uneven sequencing depths. Nonrandom missing data can arise from biological processes, such as allele dropout due to mutation at the enzyme cutting site, or artificial processes such as variable sample quality and bioinformatic filtering (Arnold et al. 2013; Huang & Knowles 2016; Nunziata & Weisrock 2017). A recent study found that in non-model systems, RADseq libraries enriched from poor-quality DNA templates are likely to generate more (nonrandom) missing data (up to 60%) than what would be expected from allele dropout due to natural mutation (Rivera-Colon et al. 2021). Both the high amount and the nonrandom distribution of missing data are common in genetic studies of non-model systems. Therefore, it is important to know the potential effects of missing data on the delineation of population structure in non-model systems.

Several studies have tested missing data effects on population delineation but provided contradictory results and suggestions. Some found that missing data had little influence on

identifying population structure and suggested less stringent filtering to get more genetic loci (Huang & Knowles 2016; Eaton et al. 2017; Shafer et al. 2017; Hodel et al. 2017). By contrast, others found that missing data impact the identified population structure, and that optimal filtering parameters should be tuned in each specific study (Arnold et al. 2013; Wright et al. 2019). Despite the contradiction, these studies show that missing data can bias population delineation, but such effects may depend on specific data sets and analytical approaches.

In this study, we focus on the principal component analysis (PCA), a multivariate approach that has been widely used in genetic and genomic studies to explore divergence patterns, summarize data variation, and visualize overall population structure (Novembre et al. 2008; Novembre & Stephens 2008; Jombart et al. 2009). Briefly, PCA transforms the raw data matrix of multiple variables into a lower-dimension matrix of principal components (PC) ordered by their retained variation (for mathematic details refer to Wold et al. 1987; Queen et al. 2002). In the context of population genetics, raw variables of multiple genetic loci (e.g., microsatellites, single nucleotide polymorphisms or SNPs) are usually centered and scaled into a standardized matrix in PCA before being converted into fewer PC dimensions (Jombart et al. 2009). The dominant population structure is captured by the first two or three PCs that retain the highest variation, and thus these PCs are usually plotted to visualize the overall genetic pattern. However, PCA and other multivariate analyses do not allow missing data in the input. Some developed PCA programs can tolerate and skip missing data (e.g., Dray & Josse 2015) but those algorithms are adapted to specific goals (e.g., detecting adaptive loci in pcadapt, Luu et al. 2017), while PCA used for population delineation requires either removal or imputation of missing data. Complete elimination of missing data is difficult because samples unavoidably have variable quantities and qualities, and too stringent bioinformatic filtering could reduce power by removing the most

variable loci and/or samples (Huang & Knowles 2016). Imputation of missing data can be done by sophisticated methods such as in genomic association studies (e.g., human disease, animal and plant breeding) that have access to high-quality reference genomes, linked haplotypes, and strong genotype-phenotype correlations (Yu & Schaid 2007; Dray & Josse 2015; Xavier et al. 2016). However, such data are usually unavailable in population genetic studies of non-model systems which instead use the mean imputation of missing data, such as defaults of the most widely used R packages adegenet (Jombart & Ahmed 2011) and SNPRelate (Zheng et al. 2012). The mean imputation strategy replaces missing data by mean allele frequencies of the corresponding loci estimated across the entire data set (i.e., individuals from all populations considered altogether). Center standardization of the mean-imputed matrix would recode the original missing data as zeros and thus un-informative in PCA. The mean imputation strategy makes it easy to run PCA on any genetic data sets and seems to work well in many cases. For example, using the package SNPRelate on empirical RADseq datasets, a study found that despite more missing data, less stringent filtering worked better to identify population structure in PCA (Hodel et al. 2017). However, in theory, mean imputation of high amounts of missing data would result in lots of zeros in the standardized matrix, leaving less overall and sample-specific variation to be analyzed in PCA. For example, a study using PCoA, a similar multivariate approach, found that RADseq data sets with an intermediate amount of missing data (i.e., 10% missingness compared to 5% and 20%) generated the best results for population structure (Wright et al. 2019). Therefore, missing data effects remain unclear and under-considered in PCA inference of population genetics.

In this study we ask three major questions.1) Do the mean-imputed missing data affect PCA interpretation? If so, how? 2) Are such effects related to the type of missing data (random or

biased) and the type of population structure (highly diverged or admixed)? 3) How to detect and mitigate the effects of missing data on PCA interpretation of population structure? We predict that high amounts of missing data will reduce the detected genetic variation, and that nonrandom missing data will bias PCA-based inference of population structure (Dray & Josse 2015). To test these predictions, we simulated genetic data sets under different population divergence patterns and introduced various amounts of missing data either randomly or nonrandomly (i.e., biased among individuals and/or populations). We also tested missing data effects using empirical RADseq data sets of the big brown bat (*Eptesicus fuscus*), a widespread species distributed from southern Canada to northern South America (Kurta & Baker 1990). We collected samples from eastern North America, western North America, and Caribbean Islands, the three putative populations identified based on morphology (Kurta & Baker 1990; Hoffman & Genoways 2008) and genetics (Neubaum et al. 2007; Turmelle et al. 2011). All PCA runs were conducted with the default mean imputation in the R package adegenet (Jombart & Ahmed 2011), a widely used method in population genetic studies. We demonstrate how missing data can affect PCA results and bias interpretation of population structure, especially when missing data are nonrandom among individuals. We suggest an easy check of missing data effects by color-coding PCA plots based on missing values, and we call for caution in the interpretation of samples or populations located around the origin of PCA plots.

Materials and Methods

Simulated data sets

We simulated genetic data of biological populations with different divergence patterns (Fig 3.1). We used the R package Coala (Staab and Metzler, 2016) to create evolutionary models and to call the coalescent simulator *ms* (Hudson 2002) to generate genetic data sets. For each model, we simulated three populations each with 25 diploid individuals (sample_size=25, ploidy=2) and 5,000 biallelic SNP loci (loci_number=5000, loci_length=1, mutation rate=1, fixed number). All three populations diverged from each other simultaneously at 0.9 time units backwards. We chose the three-population scenario for clarity and a good representation of common biological patterns simulated by different per-generation migration rates (m). Specifically, we simulated highly diverged populations that have no gene flow (m=0, model p3); weakly diverged populations with high gene flow (m=0.5, model p3_mig); population cline (i.e., stepping-stone or isolation by distance) where gene flow (m=0.5) only occurs between adjacent populations (model cline); and a continent-island system where gene flow between continent and island (m=0.05) is a magnitude lower than the within-continent gene flow (m=0.5, model island; Fig 3.1). Each model was simulated five times and consistency among replicates was confirmed based on their PCA plots (see below). Subsequent analyses were processed using one replicate simulation per model.

The simulated data were output as haploid segregation site matrices and were transformed into diploid SNP matrices where individual genotypes are coded as 0 (homozygotes of the ancestral allele), 1 (heterozygotes), or 2 (homozygotes of the alternative allele). These SNP matrices were the raw complete data, and we then introduced three levels of total missingness by replacing 1%, 10%, and 20% of the raw genotypes with "NA" (i.e., the code for missing data). These levels were selected to cover the range of missing values that are likely to be used in empirical studies (e.g., Wright et al. 2019). In addition, for each level of total missingness, we introduced missing

data either randomly or nonrandomly with bias among individuals or populations. Random missing data were introduced across the entire raw SNP matrix. Individual-biased missing data were introduced in all populations by having 80% missingness per population condensed in five individuals (20% of the population), while the other 20% missingness were introduced randomly among the remaining 80% individuals from that population. Population-biased missing data were introduced by having 80% of the total missingness condensed within one population (random among individuals within that population), while the other missing data were introduced randomly among and within the other two populations. We did not introduce locus-biased missing data because the genetic loci biased with a high percent of missing data (i.e., low genotyping rates) would have been filtered out through standard bioinformatics processing.

The raw SNP matrices and the matrices introduced with missing data were all transformed into the genlight format and analyzed using the function *glPca* of the R package adegenet 2.1.3 (Jombart & Ahmed 2011). We applied the default mean imputation on missing data, and the default standardization using center and scale of the number of alleles. We also ran PCA without centering (center=F in *glPca*) on a few representative data sets to see if the centering process could explain the detected missing data effects. But it should also be noted that no centering is *not* recommended for studies of population structure (Jombart et al. 2009). All PC axes were retained to calculate their percent variation, and results were visualized by plotting scores of the first two PCs using the package ggplot2 (Wickham 2016). All the simulation, incorporation of missing data, and PCA analyses were done in R 4.0.1 (R Core Team 2020), and the custom R scripts are available on GitHub (https://github.com/xuelingyi/missing_data_PCA).

Empirical data sets

We collected 96 samples of big brown bats from researchers and museums (Smithsonian National Museum of Natural History, USNM; American Museum of Natural History, AMNH; museum catalogs in Appendix C Table S1). Genomic DNA was extracted from contemporary tissue samples using Qiagen DNeasy Blood & Tissue Kit, and from museum samples using an optimized phenol-chloroform protocol (Alminas et al. 2021). Extracted DNA was Qubit quantified, normalized into 200ng in 10ul volume per sample, and built into a RADseq library following the bestRAD protocol in Ali et al. (2016). In brief, normalized DNA was digested using the restriction enzyme *SbfI-HF*, ligated with bestRAD adapters and unique 8bp barcodes, and then pooled by 5ul per sample and sheared in a Qsonica Q500 sonicator with 4 cycles of 30 sec on and 59 sec off. The sheared DNA fragments were visually checked on an E-gel agarose gel, size-selected using AMPure XP beads, purified using Dynabeads, and prepared with the NEBNext Ultra DNA library prep kit for Illumina. The library was sequenced using Illumina Novaseq 6000 (paired-end 150bp) by Novogene Corporation.

The raw sequencing data were processed using the University of Wisconsin-Milwaukee's research computational cluster. We used the program *process_radtags* in STACKS v2.2 (Rochette et al. 2019) to demultiplex raw data, trim reads into 140bp, and remove the reads that had missing RAD cut sites or low qualities (average Q < 10 for the window slide 15% of read length, default). The big brown bat reference genome (GCF_000308155.1_EptFus1.0) was downloaded from the National Center for Biotechnology Information and indexed using the Burrows-Wheeler Alignment tool (*bwa*, Li & Durbin 2009). The demultiplexed reads were aligned to the indexed reference genome using *bwa mem* in SAMtools (Li et al. 2009). The program *gstacks* was used to remove unmapped reads and PCR duplicates from the alignments and identify RAD loci. The program *populations* was used to filter for a minor allele frequency >

0.02 and to output only the first SNP of each RAD locus to minimize linkage disequilibrium. Data from *populations* were ordered and export into a vcf file which was further filtered in VCFtools 0.1.16 (Danecek et al. 2011) for a minimum depth of 3 and > 50% genotyping rates. Samples with more than 90% missing data across the remaining SNPs were removed to generate the first empirical data set. This data set was further filtered into two additional data sets using 85% and 98% minimum genotyping rates. The amount of missing data per individual in all three data sets was calculated using VCFtools. It is important to note that the above filtering processes were *not* to obtain optimal data for population genetic analyses, but only to generate empirical data sets with variable total missingness to demonstrate missing data effects, which is the goal of this study.

The filtered data sets were transformed from vcf into the genlight format using the package vcfR 1.11.0 (Knaus & Grünwald 2017). The total amount of missingness in each filtered data set was obtained from the summary of the genlight files. PCA was conducted in the same manner above using *glPca* with default mean imputation and standardization and all PC axes were retained. To help visualize missing data effects, we labeled the individuals from Vermont and North Carolina (the East population), Oregon (the West population), and Nebraska (potential hybrid zone) on the PCA plots.

Results

Simulated data sets

All simulated population structure was accurately identified using PCA on the raw SNP matrices without missing data (Fig 3.1), and the five replicates of each model showed almost identical

PCA plots (Appendix C Fig S1) which confirmed the consistency of our simulation. High migration rates reduced inter-population divergence and increased intra-population variation, resulting in bigger and looser population clusters in PCA plots. It is worth noting that no individual fell around the PCA origin in the p3_mig model (Fig 3.1b; Appendix C Fig S1), whereas a few individuals from the admixed population (pop2) in the cline model were always located around the origin (Fig 3.1c; Appendix C Fig S1). These results indicate that admixed or hybrid individuals would be located close to the origin on PCA plots if populations diverge under the stepping-stone model, or isolation by distance. In addition, the difference between the variance retained on PC1 and PC2 is always bigger in the cline model than in the p3_mig model, which is consistent with the lack of migration between end populations in the cline model. PCA of the island model also accurately reflected the true population structure by showing the highest divergence on PC1 between island (pop2) and continental populations, followed by the much lower divergence on PC2 between continental populations (Fig 3.1d).

The full PCA plots on the matrices introduced with missing data are in the Supplemental Information (Appendix C Fig S2-S5). We found that introducing a total of 1% missing data did not affect PCA results regardless of missing data types or evolutionary models. These results suggest that the mean imputation strategy should work well for PCA when missing data are rare. Similarly, random missing data did not have obvious impacts on PCA results across models. However, high amounts of random missing data reduced the explained genetic variation (e.g., lower variation on PC1) and resulted in looser population clusters with unclear boundaries on PCA plots. For example, with total 20% random missing data, PCA showed highly ambiguous boundaries among weakly diverged populations in the model p3_mig (Appendix C Fig S3), between adjacent populations in the cline model (Appendix C Fig S4), and between continental

populations in the island model (Appendix C Fig S5). Accordingly, our results supported the prediction that high amounts of random missing data will reduce the amount of genetic variation that would be detected if the data set contained little or no missing data. In other words, PCA does not have much power with mean-imputed high amounts of missing data, and this lack of analytical power may be misinterpreted as a lack of population structure based on PCA plots.

We found that nonrandom missing data can bias PCA results, and the biasing effects are influenced by the total amounts of missingness and the underlying population structure. Overall, no obvious bias was detected when the total missingness is low (1%), whereas higher missingness always exaggerated the biasing effects. When nonrandom missing data are condensed within a few individuals in each population (individual-biased), those individuals tend to be dragged away from their real population clusters and towards the origin of PCA plots, making them indistinguishable from any true admixed individuals that originated from high migration rates (Fig 3.2). On the other hand, if migration is low and population structure is strong (e.g., in the p3 and island models), the unbiased individuals would still tightly cluster into populations while the individuals biased with missing data would seem to form an additional "cluster" around the PCA origin (e.g., Fig 3.3a), which might be misinterpreted as those individuals being more similar to each other and as cryptic structure. When nonrandom missing data are condensed within a population (population-biased), the biased population would show reduced intra-population variation and would be dragged towards the origin of PCA plots, resulting in under-estimated divergence between the biased and unbiased populations. For example, in the island model, high amounts of missing data condensed in the island population resulted in a much lower level of island-continent divergence as detected by the PC1 variation (Fig 3.3b). In addition, effects of population-biased missing data depend on the biased

population. For example, almost identical PCA plots were obtained when high amounts of missing data are biased in one population of the p3_mig model (Appendix C Fig S3) and in the admixed population (pop2) of the cline model (Fig 3.4a). Both results indicate the biased population as an admixture, an incorrect interpretation for the p3_mig model. On the other hand, when high amounts of missing data are condensed in an end-population (pop3) of the cline model, mean imputation using average allele frequencies not only drag this population to the PCA origin, but also make it overlap with the true admixed pop2 and seem to "disappear" on the PCA plot (Fig 3.4b). In summary, our results showed that nonrandom missing data will drag the biased individuals (or populations) towards the PCA origin and make them indistinguishable from true admixed individuals (or populations).

PCA without centering did not work well to visualize population structure, as shown in the plots of the raw SNP matrices of p3_mig and cline models (Appendix C Fig S6). PCA without centering did help distinguish the missing-data biased individuals (or populations) from the true admixed individuals (or populations) because the latter were not around the origin when data were not centered (Appendix C Fig S6), but the same information is also expressed in the plots of centered PCA when points are color coded by their level of missing data. Accordingly, our data confirmed that PCA without centering is not appropriate for population genetic studies.

Empirical data sets

We sequenced 96 big brown bat individuals and obtained a total of 847,133,692 reads. Demultiplexing retained 675,346,692 (79.7%) reads in total and 7,725 to 33,773,190 reads per individual (mean 7,034,861, SD 7,381,381), indicating large variation and potential bias of

sequencing data among individuals. All individuals were aligned to the reference genome and 58.9% reads were removed as PCR duplicates. The VCFtools filtering retained 72 individuals (Fig 3.5; Appendix C Table S1) and 76,809 SNPs with total 19.12 % missing data (herein the bat20 data set). Further filtering by genotyping rates generated 25,674 SNPs with total 9.87 % missingness in the bat10 data set, and 100 SNPs with total 1.35 % missing data in the bat1 data set.

PCA on all empirical data sets showed the highest divergence on PC1 between the West and East populations, followed by the continent-island divergence on PC2 in bat10 and bat20 data sets (Fig 3.5). This pattern is consistent with the range-wide genetic study of big brown bats, which found closer evolutionary relationships between East and Caribbean populations than those between East and West populations (Turmelle et al. 2011). The data set bat1 grouped Caribbean individuals within the East population probably due to the limited power from only 100 SNPs (Fig 3.5a). On the other hand, data sets bat10 and bat20 both detected strong continent-island divergence (Fig 3.5b, c), indicating high genetic drift in the island population, although the small sample size of two Caribbean individuals could also affect the PCA indicated divergence. Plots of continental individuals using higher ordered PCs indicated substructure within the Western population, including a roughly latitudinal divergence on PC3 and roughly longitudinal divergence on PC4 (Appendix C Fig S7).

All three empirical data sets had individual-biased nonrandom missing data, as shown by the gradient coloring based on per individual missingness (Fig 3.5), and we detected similar effects of missing data as those in the simulated data sets. One individual from North Carolina had the highest missingness and was always placed around the PCA origin, while the other NC individuals with low missingness were well clustered within the East population. Similarly, a

few Oregon individuals biased with missing data were dragged from deep inside the West population towards the PCA origin, a process that is visible as the higher total missingness resulted in increasingly biased missing data. On the other hand, individuals from Nebraska (labeled with "x" and "+" in Fig 3.5) had low missingness in all data sets but were also located close to the PCA origin, suggesting genetic admixture of these individuals which is consistent with the other findings of Nebraska as a potential hybrid zone (Hoffman & Genoways 2008). It is also worth noting that the individual labeled with "x" is not only geographically closest to the East population, but also was always closer to the PCA origin than the other two Nebraska individuals that are labeled with "+" (Fig 3.5). The one individual from Vermont was identified as an outlier and was found more closely related to the West population in all three data sets, contradictory to its sampling location (Fig 3.5; Appendix C Fig S7). This individual was also biased with high missingness in the bat10 and bat20 data sets, making it unclear whether the detected pattern arose from a real genetic signal, the missing data bias, or potential errors (e.g., contamination). Plots using higher ordered PCs were also impacted by the missing data effects with high-missingness individuals located around the origin (Appendix C Fig S7).

Discussion

Results from simulated and empirical data sets support our predictions that both the total amounts and the distribution of missing data affect PCA results. The mean imputation strategy only worked well when missing data are rare (total 1%), or when missing data are of medium amounts (total 10%) and randomly distributed. When there are high amounts of random missing data, the mean imputation can reduce data variation and analytical power, although no bias was identified in our results. Accordingly, the previous study found minimal effects of missing data

on PCA (Hodel et al. 2017) possibly because missing data were overall random in their data sets. However, when missing data are nonrandom, mean imputation using average allele frequencies can bias PCA plots by artificially making the missingness-biased individuals more similar to the global "average" of the full input data set and closer to the origin of PCA plots. Such effects are exaggerated with higher amounts or more extremely biased missing data, and may further lead to misinterpretation of population structure based on PCA plots. In summary, high amounts of random missing data reduce analytical power and can lead to under-estimated population divergence. Nonrandom missing data drag missingness-biased individuals (or populations) toward the PCA origin and make them indistinguishable from true admixed individuals (or populations), potentially resulting in misidentified admixture and population structure. It should be noted that the missing data effects or biases do *not* suggest flaws of the PCA algorithm but arise from the imputation of input missing data. In other words, PCA characterizes patterns of the imputed data sets rather than the originally input incomplete data set, and thus results should be interpreted with consideration of the imputed missing data.

The missing data effects are also detected in plots using higher ordered PCs, which increases the difficulty to interpret those PC axes. In population genetic studies, PCA results are normally plotted only with the first two or three PCs for visualization purposes, and because the dominant pattern should be captured in PC1 and PC2 where the highest variation is retained (Wold et al. 1987). Although higher ordered PCs may also capture biologically interesting patterns, it is challenging to visualize plots of more than two dimensions, and variation on higher ordered PCs would be hard to interpret without the context of dominant structure (captured in PC1 and PC2) and with missing data effects. The higher ordered PCs may be more suitable in data exploration

and may help formulate hypotheses for downstream analyses, such as to analyze regional divergence using hierarchical STRUCTURE.

The effects of missing data shown in this study are more likely to be observed in non-model systems, where variable sample quantities and qualities tend to generate high amounts of nonrandom missing data in the next-generation sequencing (Arnold et al. 2013; Rivera-Colon et al. 2021). Overall, our data supported the previous suggestions that intermediate amounts of missing data (e.g., 10% total missingness) are more likely to generate optimal results (Arnold et al. 2013; Wright et al. 2019), and that high amounts of missingness could be tolerated (Shafer et al. 2017; Hodel et al. 2017) if missing data are random and the population structure is strong. Clearly, there is no universal "rule" for bioinformatic filtering on missing values (i.e., genotyping rates), and data sets need to be optimized for specific systems, analyses, and aims of the study. Such optimization also applies to other filtering parameters, such as minor allele frequencies and linkage disequilibrium, which also play important roles in empirical data sets and analyses (O'Leary et al. 2018; Linck & Battey 2019; Wright et al. 2019). These parameters do not specifically filter for missing values but can indirectly affect the amount and distribution of missing data. For example, a filtering for lower minor allele frequencies would retain more loci and higher genetic variation, but probably also higher amounts of nonrandom missing data such as due to allele dropout. Our empirical results showed that for PCA analyses, it is difficult (if not impossible) to eliminate missing data effects without sacrificing analytical power (i.e., number of genetic loci or number of samples). Therefore, potential missing data effects need to be considered for accurate interpretation of PCA results, especially when the population structure is weak or unknown. We suggest researchers 1) plot PCA scores with a color gradient showing per sample (either individual or population, depending on the input) missing data to detect

potential biasing effects; 2) when samples biased with missing data are found around the PCA origin, interpret those samples with high uncertainty; 3) if possible, remove the low-quality samples with highest missingness and refilter the remaining samples (e.g., Cerca et al. 2021) to get potentially more loci and more randomly distributed missing data; 4) use complementary analyses (e.g., model-based methods) to cross-validate PCA results and help delineate population structure. Because each method has its own power and caveats, no single method would be the best under all conditions, but consistent signals across methods (e.g., multivariate analyses and model-based methods) would increase confidence in the interpretation of population structure.

Here we tested effects of missing data on PCA using the mean imputation strategy, which is contemporarily the most common approach in population genetic studies of non-model systems (such as defaults of the R packages adegenet and SNPRelate). Although more sophisticated imputation and adapted PCA algorithms that can tolerate missing data are available (Yu & Schaid 2007; Dray & Josse 2015; Xavier et al. 2016; Luu et al. 2017), these approaches were not developed to study population structure, and they remain largely impractical in non-model systems where genomic structure, divergence patterns, and phenotype-genotype associations are largely unknown. In addition, the existing sophisticated imputation methods were found to generate different results with high amounts of missing data (20% in Xavier et al. 2016), which may also bias PCA in different ways. We expect to see further advances in PCA-like approaches for population genetic analyses, but it may take some time for such advances to be thoroughly vetted and widely applied in non-model systems. One potentially convenient alternative to consider would be to use within-group mean imputation by imputing missing data using mean allele frequencies of the population of the imputed individual. Although within-group mean imputation would not drag the missingness-biased individuals to the PCA origin, this method

would instead drag those individuals to the population used to generate imputation values. Therefore, within-group mean imputation can also generate biased results where the missingness-biased individuals are indicated to be more closely related to their *a priori* assigned population, or even placed within the wrong population cluster if the *a priori* population assignment was wrong. Given these limitations, we do not recommend a within-group mean imputation strategy which largely depends on the *a priori* population designation and can easily bring artificial biases that are much more difficult to detect or justify than the straightforward bias shown here. In addition, population genetic studies use PCA to visualize dominant structure, explore data, assign individuals into groups, and identify outlier individuals. It would be a logical fallacy to impute missing data based on the *a priori* population assignment, and then use PCA of that imputed data to confirm or support the same *a priori* population assignment. Accordingly, we recommend the default strategy of grand mean imputation over the possible within-group mean imputation for assessing population genetic structure, especially when the population divergence is weak or not fully understood. Regardless of the approach used, it is important to always consider potential missing data biases when interpreting PCA-described population structure, which is critical for further studies of evolution and conservation management in non-model systems.

Figure 3.1 The simulated models and their original PCA without missing data. Each of the three populations was simulated with 25 diploid individuals and 5,000 SNP loci. The four evolutionary models were simulated with different migration parameters: a) p3 model, no migration, b) p3_mig model, migration rate 0.5 per generation between all population pairs, c) cline model, migration rate 0.5 per generation between adjacent populations, and d) island model, migration rate 0.5 per generation between continental populations, while 0.05 per generation between the island and continental populations. PCA plots of the four replicate simulations of each model are highly consistent (Appendix C Fig S1).

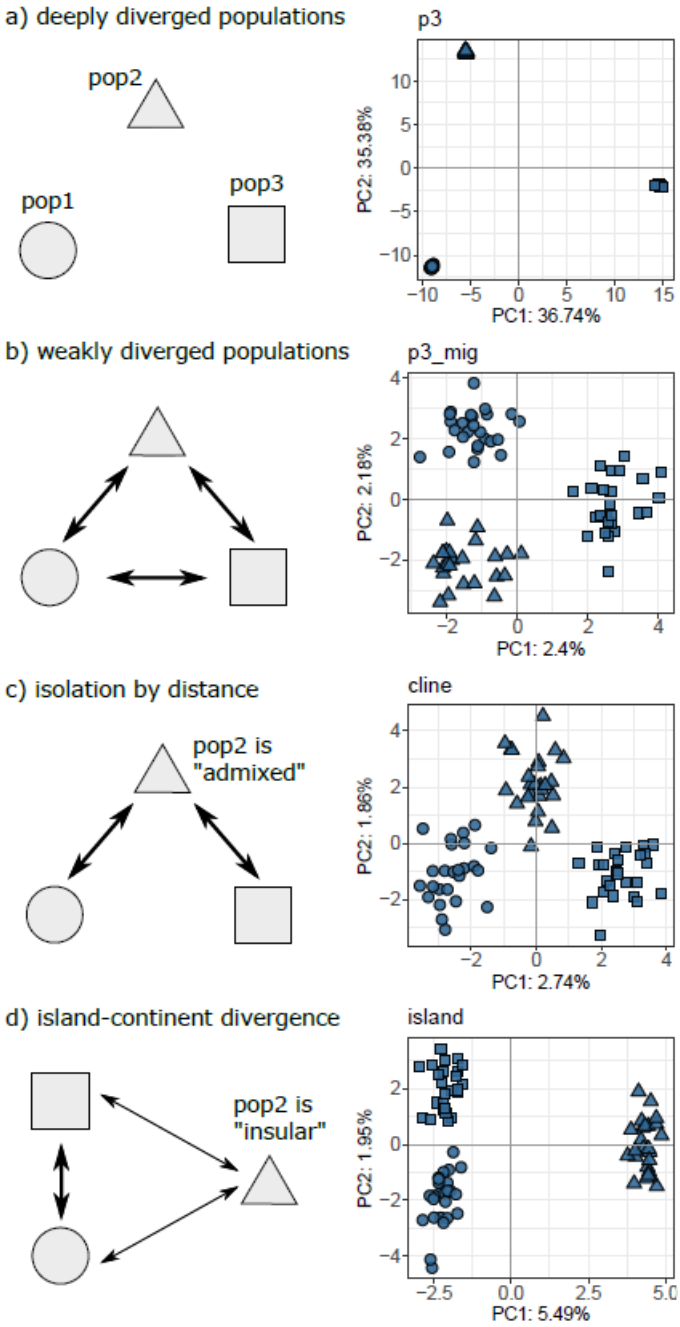


Figure 3.2 PCA on the individual-biased missing data introduced to a) p3_mig and b) cline models. Individual shapes represent their population and are consistent with labels in Fig 3.1. Individual colors represent their amounts of missing data with relatively higher missingness shown in lighter blue in each plot. The following Figs are plotted in the same way.

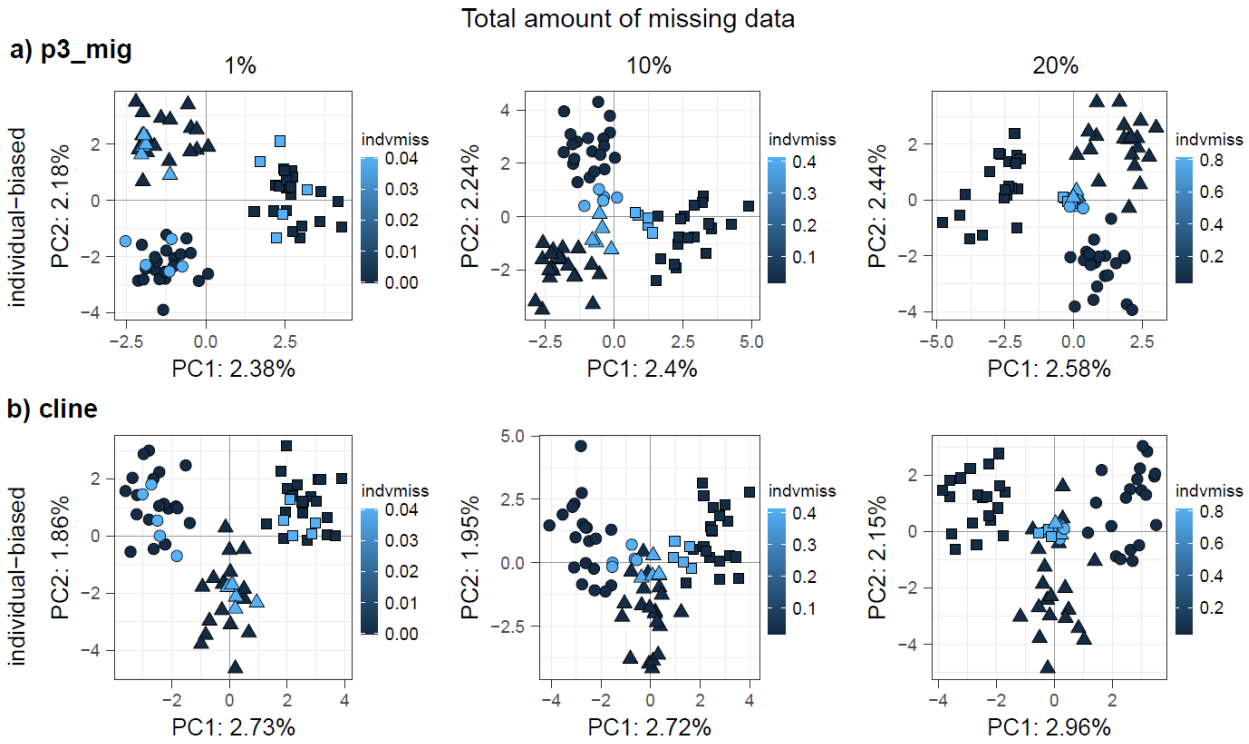


Figure 3.3 PCA on the island model with a) individual-biased and b) population-biased missing data (the island population is biased). Only the extreme levels (1% and 20%) of total missing data are shown here and results of the intermediate missingness are available in the supplementary information (Appendix C Fig S5).

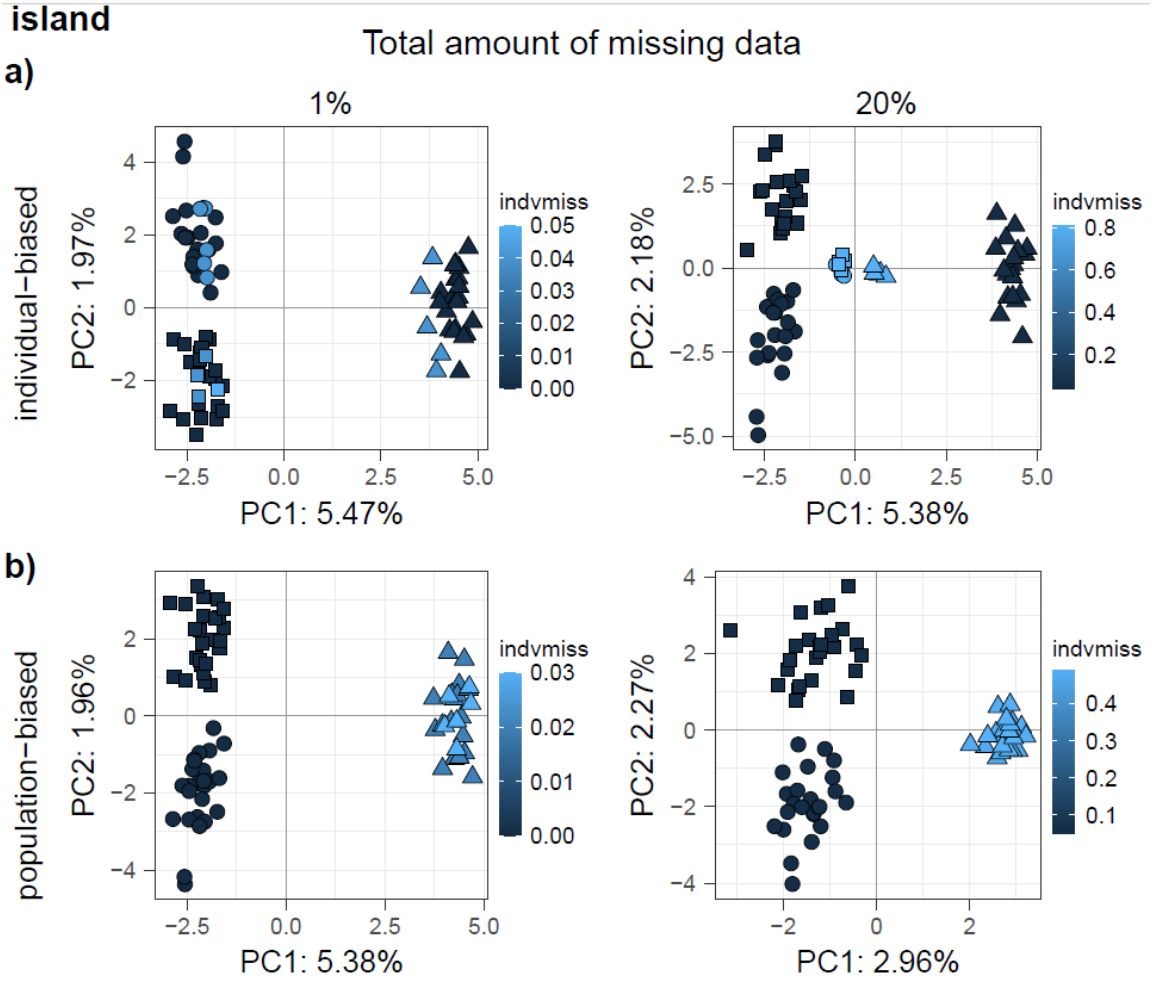


Figure 3.4 PCA on the cline model with missing data condensed in the a) admixed population and b) one end population. The admixed population (pop2) has high migration with both end populations (pop1 and pop3) while no migration occurs between the end populations (see Fig 3.1).

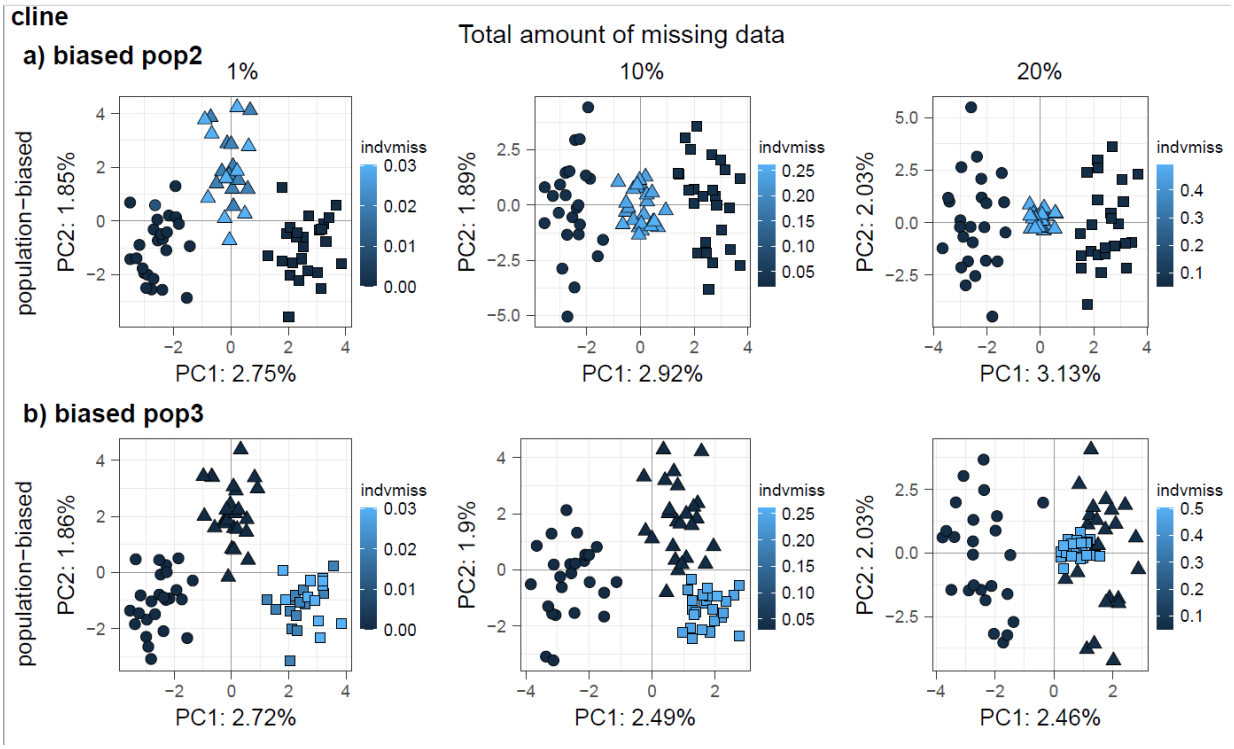
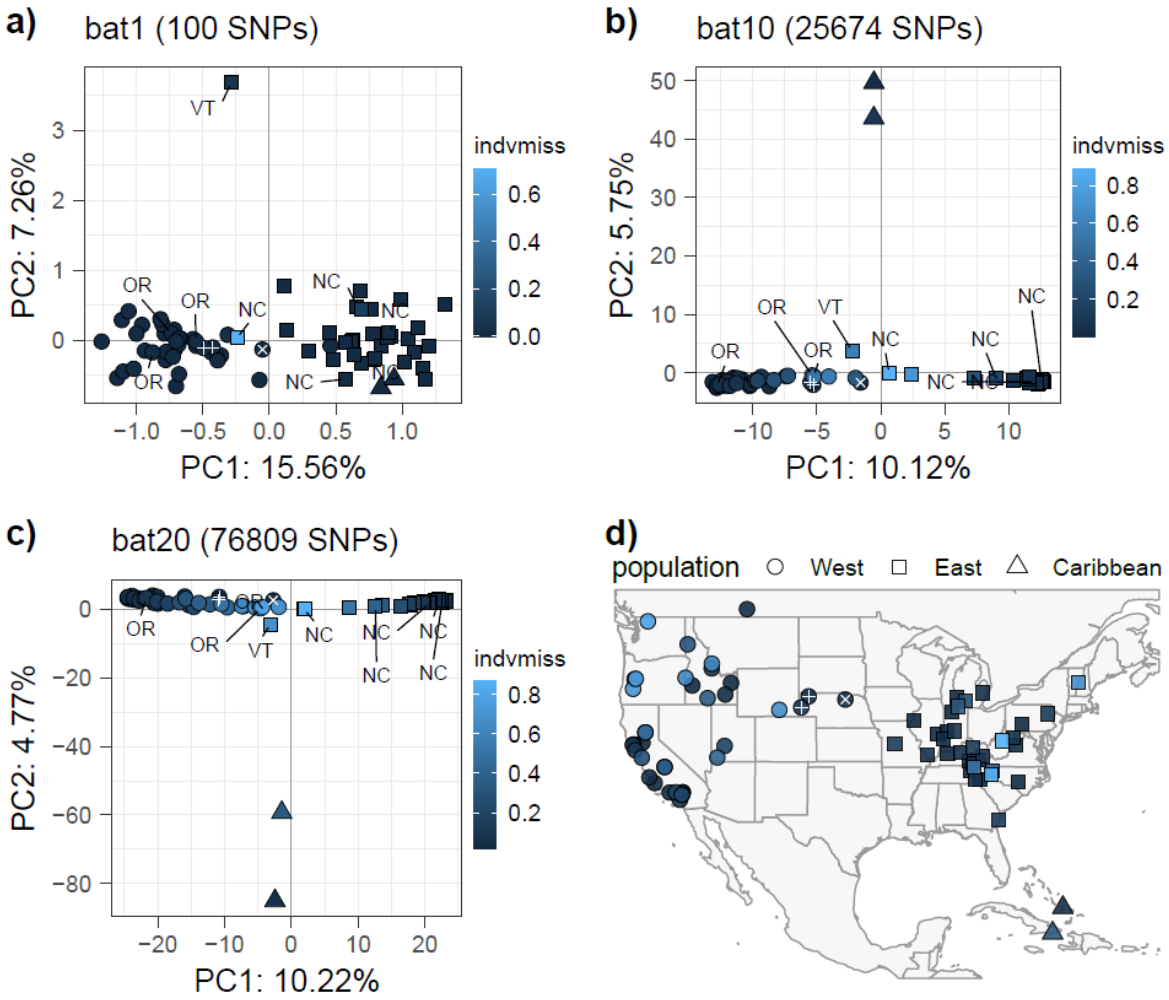


Figure 3.5 Empirical data sets of the big brown bat. Plots **a-c** correspond to the filtered empirical data sets with roughly 1%, 10%, and 20% total missing data. The number of retained SNPs in each data set is given in the title of PCA plots. Individuals are shaped by population corresponding to **d)** the distribution map of the retained 72 samples. PCA plots are colored by the per individual missingness in each data set, and the map is colored by the per individual missingness in the dataset bat20. Individuals from Oregon (the West population) and Vermont and North Carolina (the East population) are labeled by state abbreviations. The three individuals from Nebraska (potential hybrid zone) are labeled with "x" and "+" on top of the corresponding dots, where the "x" individual was both geographically closest to the East population and closer to PCA origins than the "+" individuals.



References

Ali, O. A., O'Rourke, S. M., Amish, S. J., Meek, M. H., Luikart, G., Jeffres, C., & Miller, M. R. (2016). Rad capture (Rapture): Flexible and efficient sequence-based genotyping. *Genetics*, *202*(2).

Alminas, O. S. V., Heffelfinger, J. R., Statham, M. J., & Latch, E. K. (2021). Phylogeography of cedros and tiburón island mule deer in North America's desert southwest. *Journal of Heredity*, *112*(3).

Andrews, K. R., Good, J. M., Miller, M. R., Luikart, G., & Hohenlohe, P. A. (2016). Harnessing the power of RADseq for ecological and evolutionary genomics. *Nature Reviews Genetics 2016 17:2*, *17*(2), 81–92.

Arnold, B., Corbett-Detig, R. B., Hartl, D., & Bomblies, K. (2013). RADseq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling. *Molecular Ecology*, *22*(11), 3179–3190.

Cerca, J., Maurstad, M. F., Rochette, N. C., Rivera-Colón, A. G., Rayamajhi, N., Catchen, J. M., & Struck, T. H. (2021). Removing the bad apples: A simple bioinformatic method to improve loci-recovery in de novo RADseq data for non-model organisms. *Methods in Ecology and Evolution*, *12*(5), 805–817.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., & Durbin, R. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*(15).

Dray, S., & Josse, J. (2015). Principal component analysis with missing values: a comparative survey of methods. *Plant Ecology*, *216*(5), 657–667.

Eaton, D. A. R., Spriggs, E. L., Park, B., & Donoghue, M. J. (2017). Misconceptions on missing data in RAD-seq phylogenetics with a deep-scale example from flowering plants. *Systematic Biology*, *66*(3).

Hodel, R. G. J., Chen, S., Payton, A. C., McDaniel, S. F., Soltis, P., & Soltis, D. E. (2017). Adding loci improves phylogeographic resolution in red mangroves despite increased missing data: comparing microsatellites and RAD-Seq and investigating loci filtering. *Scientific Reports 2017 7:1*, *7*(1), 1–14.

Hoffman, J. D., & Genoways, H. H. (2008). Characterization of a contact zone between two subspecies of the big brown bat (Eptesicus fuscus) in Nebraska. *Western North American Naturalist*, *68*(1).

Huang, H., & Lacey Knowles, L. (2016). Unforeseen Consequences of Excluding Missing Data from Next-Generation Sequences: Simulation Study of RAD Sequences. *Systematic Biology*, *65*(3), 357–365.

Hudson, R. R. (2002). Generating samples under a Wright–Fisher neutral model of genetic variation. *Bioinformatics*, *18*(2), 337–338.

Jombart, T., Pontier, D., & Dufour, A. B. (2009). Genetic markers in the playground of multivariate analysis. *Heredity 2009 102:4*, *102*(4), 330–341.

Jombart, T., & Ahmed, I. (2011). adegenet 1.3-1: New tools for the analysis of genome-wide SNP data. *Bioinformatics*, *27*(21).

Knaus, B. J., & Grünwald, N. J. (2017). vcfr: a package to manipulate and visualize variant call format data in R. *Molecular Ecology Resources*, *17*(1), 44–53.

Kurta, A., & Baker, R. H. (1990). *Eptesicus fuscus*. *Mammalian species*, (356), 1-10.

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, *25*(14), 1754-1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., ... & Durbin, R. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, *25*(16), 2078-2079.

Linck, E., & Battey, C. J. (2019). Minor allele frequency thresholds strongly affect population structure inference with genomic data sets. *Molecular Ecology Resources*, *19*(3), 639–647.

Luu, K., Bazin, E., & Blum, M. G. B. (2017). pcadapt: an R package to perform genome scans for selection based on principal component analysis. *Molecular Ecology Resources*, *17*(1), 67–77.

Neubaum, M. A., Douglas, M. R., Douglas, M. E., & O'Shea, T. J. (2007). Molecular ecology of the big brown bat (*Eptesicus fuscus*): genetic and natural history variation in a hybrid zone. *Journal of Mammalogy*, *88*(5), 1230-1238.

Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A. R., Auton, A., Indap, A., King, K. S., Bergmann, S., Nelson, M. R., Stephens, M., & Bustamante, C. D. (2008). Genes mirror geography within Europe. *Nature 2008 456:7219*, *456*(7219), 274–274.

Novembre, J., & Stephens, M. (2008). Interpreting principal component analyses of spatial population genetic variation. *Nature Genetics 2008 40:5*, *40*(5), 646–649.

Nunziata, S. O., & Weisrock, D. W. (2017). Estimation of contemporary effective population size and population declines using RAD sequence data. *Heredity 2018 120:3*, *120*(3), 196–207.

O'Leary, S. J., Puritz, J. B., Willis, S. C., Hollenbeck, C. M., & Portnoy, D. S. (2018). These aren't the loci you're looking for: principles of effective SNP filtering for molecular ecologists. *Molecular Ecology*, *27*, 3193–3206.

Queen, J.P., Quinn, G.P. and Keough, M.J., 2002. Experimental design and data analysis for biologists. Cambridge university press.

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rivera-Colón, A. G., Rochette, N. C., & Catchen, J. M. (2021). Simulation with RADinitio improves RADseq experimental design and sheds light on sources of missing data. *Molecular Ecology Resources*, *21*(2), 363–378.

Rochette, N. C., Rivera-Colón, A. G., & Catchen, J. M. (2019). Stacks 2: analytical methods for paired-end sequencing improve RADseq-based population genomics. *Molecular Ecology*, *28*(21), 4737-4754.

Shafer, A. B. A., Peart, C. R., Tusso, S., Maayan, I., Brelsford, A., Wheat, C. W., & Wolf, J. B. W. (2017). Bioinformatic processing of RAD-seq data dramatically impacts downstream population genetic inference. *Methods in Ecology and Evolution*, *8*(8), 907–917.

Staab, P. R., & Metzler, D. (2016). Coala: an R framework for coalescent simulation. *Bioinformatics*, *32*(12), 1903–1904.

Turmelle, A. S., Kunz, T. H., & Sorenson, M. D. (2011). A tale of two genomes: contrasting patterns of phylogeographic structure in a widely distributed bat. *Molecular Ecology*, *20*(2), 357-375.

Wold, S., Esbensen, K., & Geladi, P. (1987). Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, *2*(1–3), 37–52. https://doi.org/10.1016/0169-7439(87)80084-9

Wickham H (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. ISBN 978-3-319-24277-4, https://ggplot2.tidyverse.org.

Wright, B. R., Grueber, C. E., Lott, M. J., Belov, K., Johnson, R. N., & Hogg, C. J. (2019). Impact of reduced-representation sequencing protocols on detecting population structure in a threatened marsupial. *Molecular Biology Reports 2019 46:5*, *46*(5), 5575–5580.

Xavier, A., Muir, W. M., & Rainey, K. M. (2016). Impact of imputation methods on the amount of genetic variation captured by a single-nucleotide polymorphism panel in soybeans. *BMC Bioinformatics*, *17*(1), 1–9.

Yu, Z., & Schaid, D. J. (2007). Methods to impute missing genotypes for population data. *Human Genetics*, *122*(5), 495–504.

Zheng, X., Levine, D., Shen, J., Gogarten, S. M., Laurie, C., & Weir, B. S. (2012). A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics*, *28*(24), 3326–3328.

Appendix A

Supplementary materials

Systematics of the New World bats *Eptesicus* and *Histiotus* indicate trans-marine dispersal followed by Neotropical cryptic diversification

Xueling Yi[1] and Emily K. Latch[1]
[1] Department of Biological Sciences, University of Wisconsin-Milwaukee, Milwaukee, Wisconsin, 53211, USA

Data of the replicate samples from the same individual

The individual AMNH_M-190167 (*Eptesicus fuscus miradorensis* from Mexico) was a dry skin specimen and the extracted DNA was split into two replicate samples (EF112_1, EF112_2) in sequencing. The replicate samples had different amounts of input DNA and were prepared using either double-stranded (ds) or single-stranded (ss) libraries. The sample prepared using the ds library was included in analyses. Data of the replicate samples are summarized below.

Comparable numbers of UCE loci were identified from the two replicates, but the ss library output fewer reads and shorter contigs probably due to the less input DNA. Interestingly, the ss library generated more UCEs than the ds library, although the mean lengths of UCE loci were much shorter. We think this is because the ss library can ligate short DNA fragments to allow sequencing on the degraded molecules that would be filtered out in the ds library, resulting in slightly higher number of total UCEs but those loci were also much shorter. On the other hand, the ss library generates fewer contigs than the ds library, indicating that the ss library preparation effectively maximized the output of targeted loci rather than ligating and sequencing errors or random reads (the majority of contigs). Accordingly, our results showed that the ss library preparation performed well on degraded DNA and generated comparable results as the ds library preparation.

ID	lib	DNA /ng	#Trimmed reads	Trimmed reads mean length	#contigs	Contig mean length	#UCE	UCE mean length
EF112_1 (included)	ds	1006.5	22049978	127.6	543019	267.2	3861	1042.4
EF112_2	ss	585.6	10851526	117.9	102224	270.3	4005	445.9

Table S1. The sampling information. Both the 96 samples sequenced in this study and the four samples from Platt et al. (2018) are included. Museum catalogues of our collected samples are provided.

Note: Table S1 is an independent Supplemental File.

Table S2. Comparison of the six biogeographic models in BioGeoBEARS. Models are ordered based on AICc values. The input is the BPP tree topology.

	LnL	d	e	j	AICc	AICc_wt
DEC+J	-127.5	0.003	1.00E-12	0.049	261.5	0.930
DIVALIKE+J	-130.1	0.004	1.00E-12	0.047	266.8	0.066
BAYAREALIKE+J	-135.2	0.003	1.00E-08	0.059	276.8	4.00E-04
DIVALIKE	-150.3	0.014	8.30E-03	0	304.8	3.60E-10
DEC	-152.6	0.009	1.00E-12	0	309.5	3.50E-11
BAYAREALIKE	-169.5	0.012	0.2	0	343.3	1.60E-18

Figure S1. The ten biogeographic areas used in BioGeoBEARS analyses. All 95 individuals collected for this study are mapped and colored by their biogeographic codes (Table S1). Solid lines depict the 11 zoogeographic realms, and dashed lines depict the 20 zoogeographic regions identified in Holt et al. (2013). Note that the Caribbean area was coded separately from the Panamanian area in our analyses.

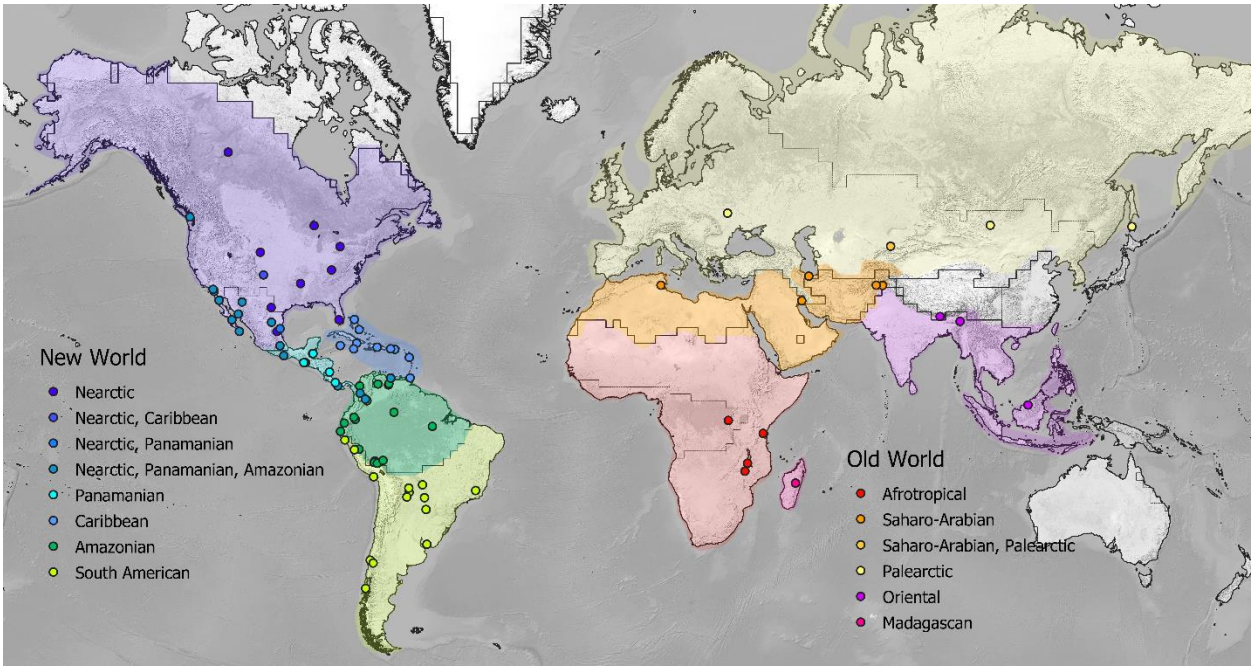


Figure S2. The number of UCEs per sample. Each row represents one sample including the 96 samples sequenced in this study and the 4 samples selected from Platt et al. (2018). Row names correspond with sample IDs in Table S1. The dashed red line shows the threshold of 500 UCEs, and red bars are the 16 samples excluded from analyses (one replicate and 15 samples with <500 UCEs).

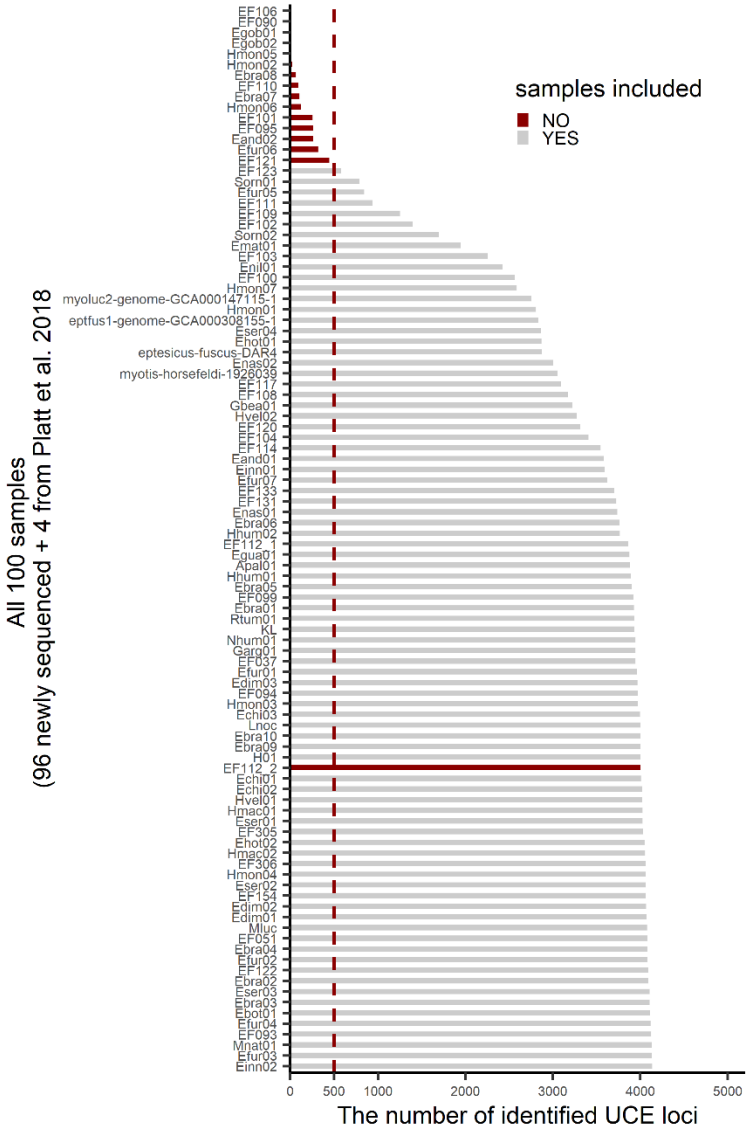


Figure S3. The Bayesian individual phylogeny generated in ExaBayes. Numbers on the nodes are posterior probabilities. Each tip represents one individual and is named by species identity and sample source. The three individuals in red had misidentified species identities and were excluded from downstream analyses. Branches in red highlight differences from the ML individual phylogeny in Figure 2.

Figure S4. The quartet-based individual phylogeny generated in SVDquartets. Numbers on the nodes are bootstrap supports. Each tip represents one individual and is named by species identity and sample source. The three individuals in red had misidentified species identities and were excluded from downstream analyses. Branches in red highlight differences from the ML individual phylogeny in Figure 2.

Figure S5. Species trees generated by the summary method ASTRAL-III. Results were obtained from gene trees of 500, 1000, and 2000 most informative UCEs (A-C) or all 3611 UCEs (D). Low-supported branches (bootstraps < 10) in the gene trees were collapsed. Numbers on the nodes represent posterior probabilities. Red tips and branches in B-D highlight differences from the species tree using 500 UCEs in A (also Fig3A) which was fixed in downstream analyses and used in discussion.

Note: figures S3-S5 are independent Supplemental Files.

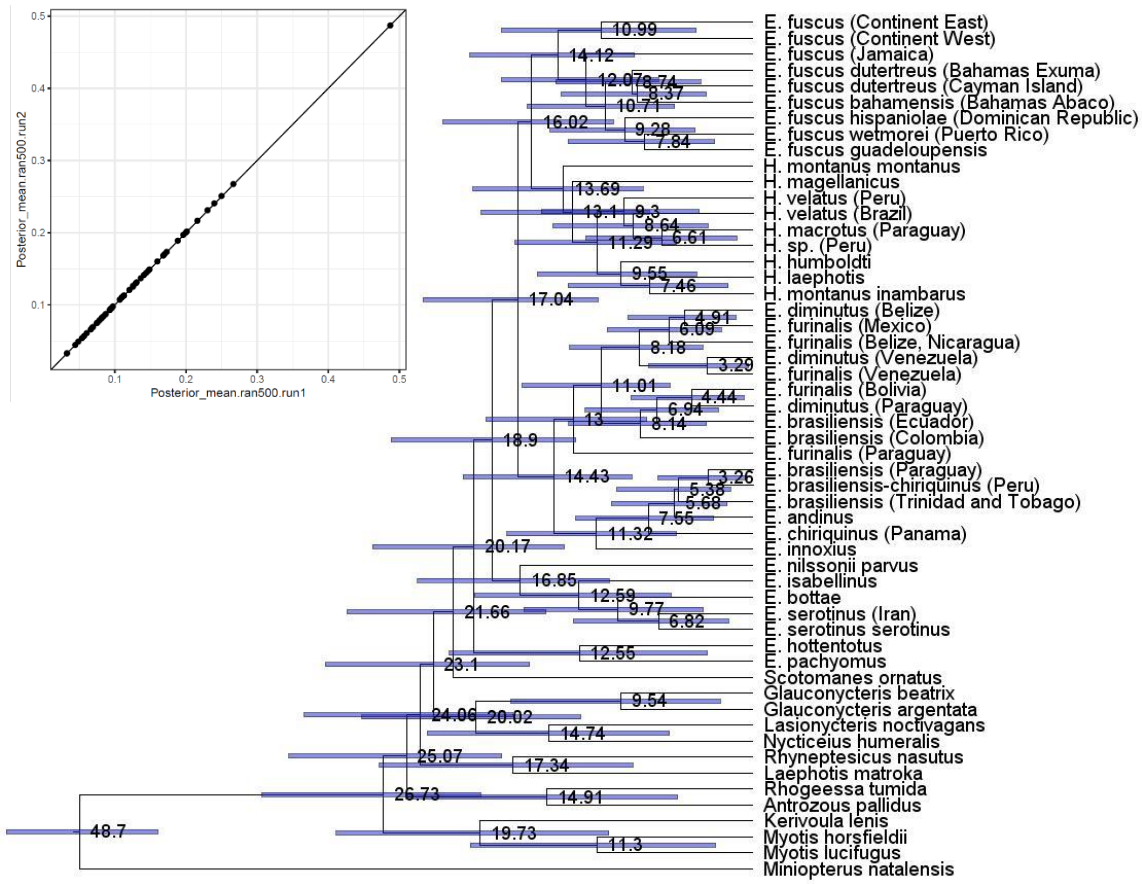


Figure S6. Divergence times estimated by MCMCTree using the 500 randomly chosen UCEs. The inserted plot shows high consistency of posterior mean divergence times estimated in two independent runs. Each tip corresponds to one OTU represented by one individual in the pruned data set (black dots in Fig 1.1 and solid tips in Fig 1.2). The tree topology was fixed by results from ASTRAL-III using 500 UCEs (Fig 1.3A). Asterisks indicate the two soft-bound calibrations. Numbers on nodes are posterior mean divergence times (millions of years ago) and bars represent the 95% HPD.

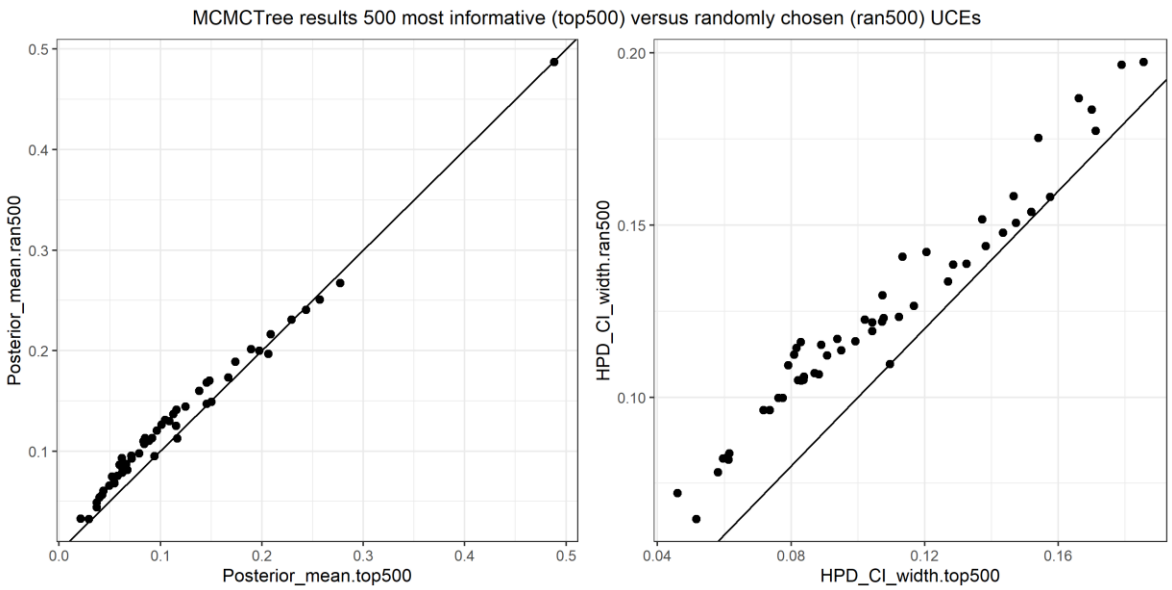


Figure S7. Comparisons of MCMCTree estimations using the 500 most informative UCEs (x axis) or 500 randomly chosen UCEs (y axis). The left panel compares estimates of the mean divergence times, and the right panel compares the 95% HPD at the same nodes. Solid lines indicate the expectation of identical estimates (i.e., y=x) from the two analyses.

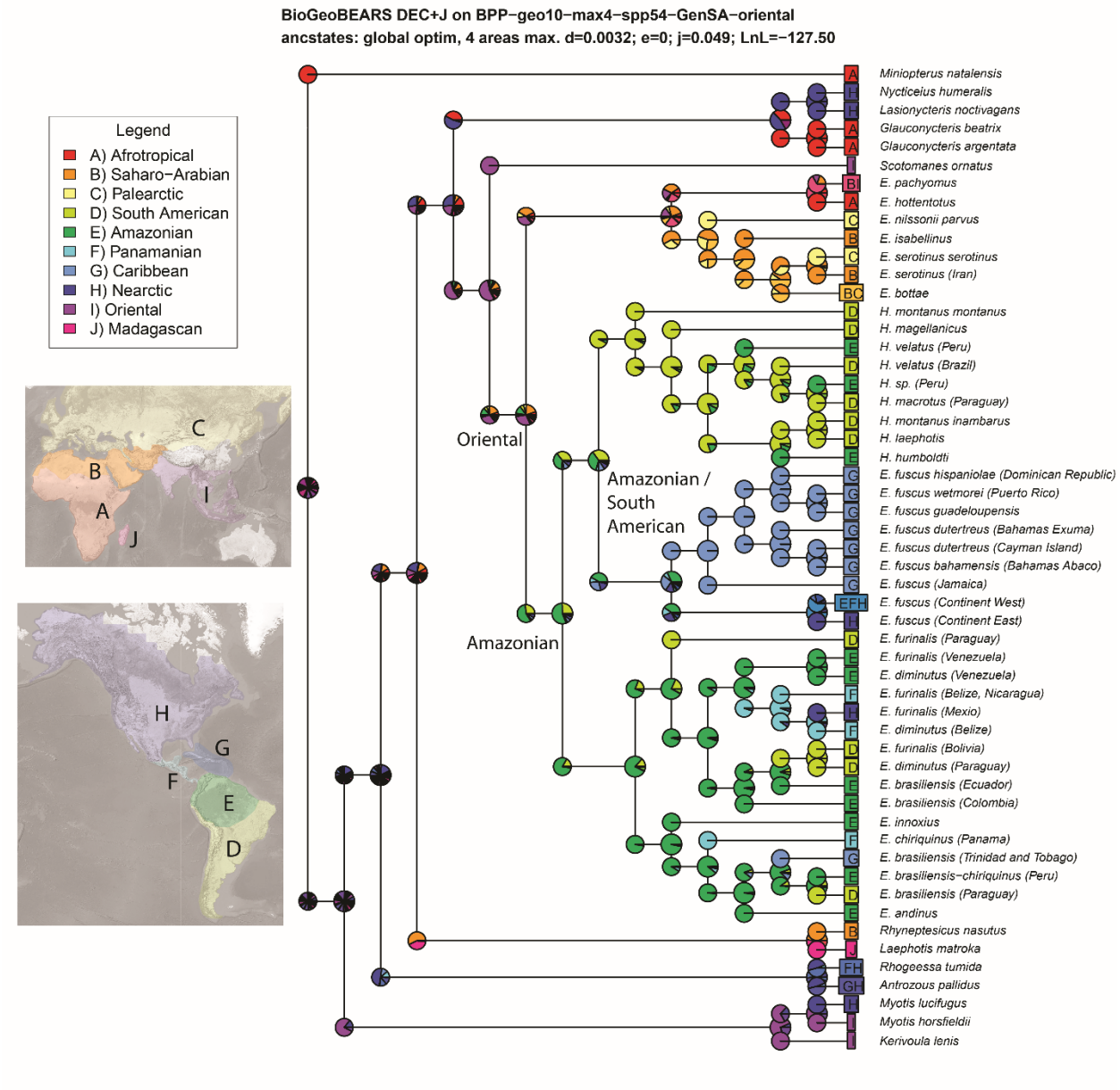


Figure S8. The BioGeoBEARS reconstructed ancestral distribution using the DEC+J model and the BPP topology. Tip letters represent the assigned biogeographic area(s) of each OTU (also in Appendix A Table S1) corresponding to the legends. Pie charts on nodes represent the estimated probabilities of ancestral geographic distribution. The most likely distributions are labeled for the key nodes representing the MRCA of *Histiotus* and *E. fuscus*, the MRCA of all New World clades, and the MRCA of the New World and Old World *Eptesicus*.

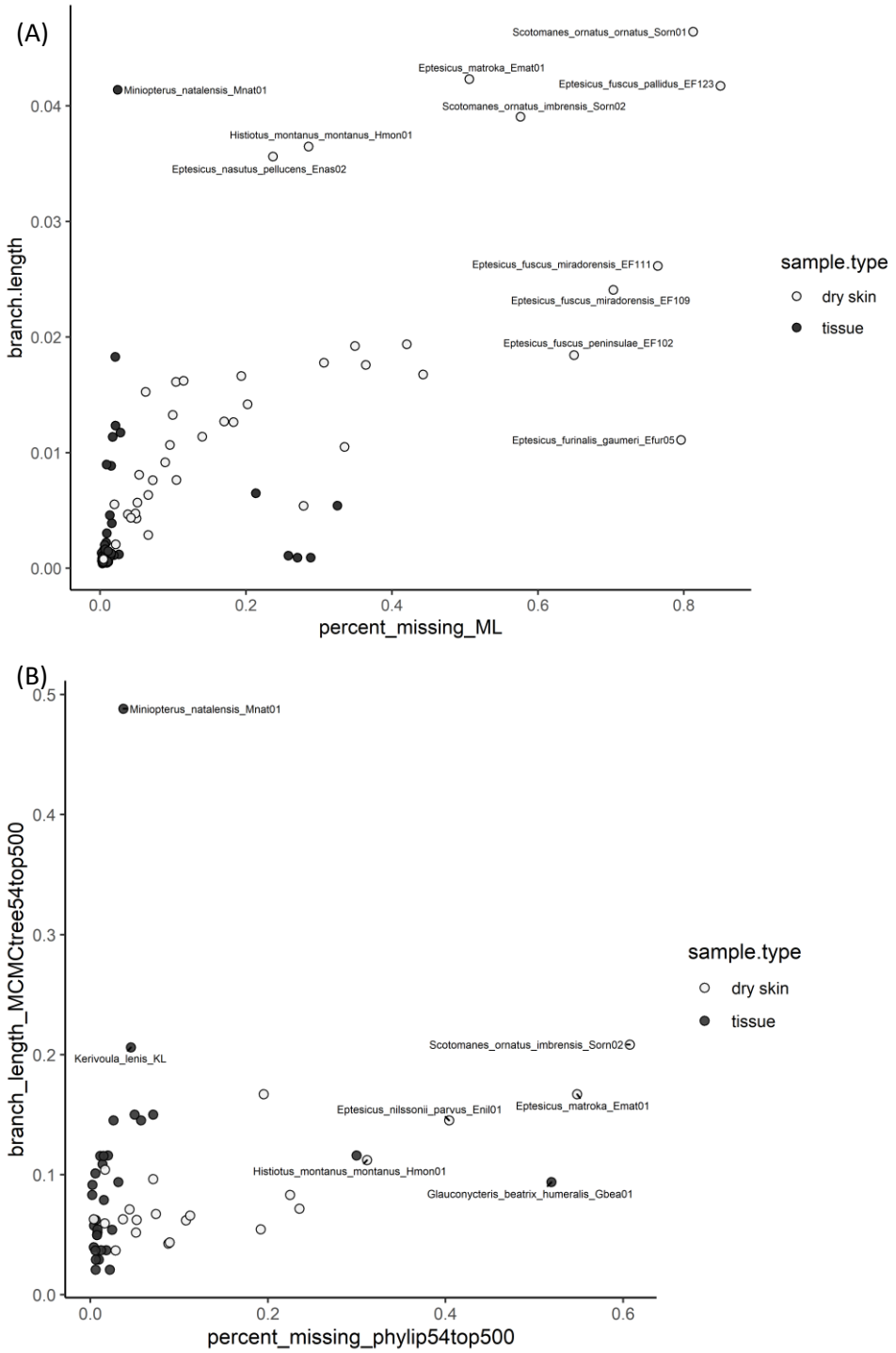


Figure S9. The relationship between missing data and terminal branch lengths. The percent missingness was calculated as the number of missing data (coded as "?") divided by the total length of the alignment in per sample. Individuals from Platt et al. (2018) are included as tissue samples. **A)** The full data set using 84 samples and 3611 UCEs. Terminal branch lengths are from the ML tree in Figure 1.2. **B)** The pruned data set using 54 samples and the 500 most informative UCEs. Terminal branch lengths are from the timed tree in Figure 1.4.

Figure S10. Photos of the two analyzed museum specimens identified as *Histiotus velatus*. Photos were taken by Xueling Yi during sampling. **Left**: USNM_548683 collected from Brazil. The yellowish fur color, paler skin color, and rounder ear tips make it look similar to the new species *H. diaphanopterus* described in Feijó et al. (2015). However, the key feature, transparent wings, was not identifiable in this museum specimen and it did not show a pale uropatagium found in *H. diaphanopterus* (Feijó et al. 2015). Accordingly, we hesitated to rename it as a different species with the data in hand. **Right**: FMNH_68506 collected from Peru. This specimen has more typical morphological features described in *H. velatus*, including the dark brown fur color, dark skin color, and triangular ears (Feijó et al. 2015).

Appendix B

Supplementary materials

Nuclear phylogeography reveals strong impacts of gene flow in big brown bats

Xueling Yi[1] and Emily K. Latch[1]

[1] Department of Biological Sciences, University of Wisconsin-Milwaukee, Milwaukee, Wisconsin, 53211, USA

Table S1. A separate excel file listing the sample information. Catalogues of the museum specimens retained in the analyses are included in the table.

Note: Table S1 is an independent Supplemental File.

The following museum specimens of big brown bats were loaned but yielded insufficient data for analyses:

USNM_296712, USNM_103810, USNM_300532, USNM_576615, USNM_539754, USNM_296709, USNM_296706, USNM_565065, USNM_329579, USNM_541106, USNM_556383, USNM_507124, USNM_528738, USNM_531340, USNM_601279, USNM_121906, USNM_129052, USNM_103809, USNM_252280, AMNH_M-149247, AMNH_M-68688, AMNH_M-74304, AMNH_M-190167, AMNH_M-205167, AMNH_M-163832, AMNH_M-21447, AMNH_M-188550, AMNH_M-143003, AMNH_M-141817, AMNH_M-38491, AMNH_M-99048, FMNH_49154, FMNH_11773, NSRL_TK_8125, NSRL_TK_912200, NSRL_TK_972145, ASNHC_16872, ASNHC_16961, ASNHC_16997, ASNHC_17839, ASNHC_18125, ASNHC_18806, ASNHC_18827, ASNHC_18829, ASNHC_19096, ASNHC_19097, ASNHC_19099, ASNHC_19100, ASNHC_19104, ASNHC_19105, ASNHC_19297, UMNH_32395, UMNH_39930, LSUMZ_8081, LSUMZ_11932, LSUMZ_12976, LSUMZ_12980, LSUMZ_17088, LSUMZ_M-10718, LSUMZ_M-12, LSUMZ_M-1825, LSUMZ_M-559, LSUMZ_M-8, UMZM:Mamm:20593, DMNS:Mamm:14138, DMNS:Mamm:17767, DMNS:Mamm:19461, DMNS:Mamm:14021, DMNS:Mamm:14283, DMNS:Mamm:19671, UAZ_04104, UAZ_17019, UAZ_07240, UAZ_25964, UAZ_23999, UAZ_11901, UAZ_16552, UAZ_10918, UAZ_15031, UAZ_10840, UAZ_25972, UAZ_03626, UAZ_09937, UAZ_19334, UAZ_23946, UAZ_13444, UAZ_13600, MSB:Mamm:125614, MSB:Mamm:156913, MSB:Mamm:126643, MSB:Mamm:126644, FLMNH_33826, FLMNH_32624, MVZ:Mamm:238711, MVZ:Mamm:238669, MVZ:Mamm:238649, MVZ:Mamm:238673, MVZ:Mamm:228337, MVZ:Mamm:225373, MVZ:Mamm:225371, MVZ:Mamm:225363, MVZ:Mamm:225360

Table S2. Summary of the five datasets used in this study.

Dataset	# Individuals	# SNPs	Analyses
Range-wide	182	2,928	ADMIXTURE, Structure, DAPC, EEMS, IBD
Continent	174	24,957	Admixture, Structure, DAPC
Island	8	2,549	PCA
Phylogeography	187	149,766	RAxML, SVDquartets
Reduced (repolarized)	27 (26)	4,079 (3,947)	RAxML, SVDquartets, TreeMix, fastsimcoal2

Table S3. Summary of the iterative filtering on the range-wide dataset.

Software	Filtering command	# Individuals retained	# SNPs retained
STACKS v2.2-populations	-r 0.5, --min_mac 3	275	111,977
VCFtools v0.1.16 *Note:* *The --max-missing value equals to the minimum genotyping rate.* *The imiss value sets the maximum percentage of missing data per individual.*	--imiss<0.8	233	111,977
	--max-missing 0.7	233	42,707
	--imiss<0.6	207	42,707
	--max-missing 0.8	207	19,883
	--imiss<0.5	195	19,883
	--max-missing 0.85	195	10,021
	--imiss<0.4	184	10,021
	--max-missing 0.9	184	2,928
	--imiss<0.3	182	2,928

Table S4. The seven selected bioclimatic variables out of all 19 variables from the WorldClim. The initial species distribution model was constructed with all variables and their percent contributions are shown below. The shaded 7 variables were selected for subsequent distribution modeling.

ID	Bioclimatic variable	% contribution in the initial model
BIO1	Annual Mean Temperature	0.9
BIO2	Mean Diurnal Range (Mean of monthly (max temp - min temp))	0.5
BIO3	Isothermality (BIO2/BIO7) (* 100)	0.3
BIO4	Temperature Seasonality (standard deviation *100)	7.4
BIO5	Max Temperature of Warmest Month	14.3
BIO6	Min Temperature of Coldest Month	6.2
BIO7	Temperature Annual Range (BIO5-BIO6)	2.7
BIO8	Mean Temperature of Wettest Quarter	1.1
BIO9	Mean Temperature of Driest Quarter	1.3
BIO10	Mean Temperature of Warmest Quarter	0.1
BIO11	Mean Temperature of Coldest Quarter	59.9
BIO12	Annual Precipitation	0.3
BIO13	Precipitation of Wettest Month	1.6
BIO14	Precipitation of Driest Month	1
BIO15	Precipitation Seasonality (Coefficient of Variation)	0.4
BIO16	Precipitation of Wettest Quarter	0
BIO17	Precipitation of Driest Quarter	0.2
BIO18	Precipitation of Warmest Quarter	0
BIO19	Precipitation of Coldest Quarter	1.8

Figure S1. Model-based and multivariate analyses on the population structure of the big brown bat. A) Bar plots of percent ancestry estimated in STRUCTURE using the range-wide dataset (optimal K=3) and the continent dataset (optimal K=2). Each bar represents one individual and is ordered in the same way by the assigned cluster then by longitude. **B)** Identification of the optimal number of clusters in STRUCTURE (the Evanno method), ADMIXTURE (the cross-validation error), and DAPC - *find.clusters* (the BIC values). The range-wide dataset showed the lowest BIC at K=2, corresponding to the continent-island divergence, but K=3 had similar support (deltaBIC = 1.1) and was thus determined optimal. **C)** DAPC plots of the continental (2 clusters) and range-wide (3 clusters) datasets. **D)** PCA of the island dataset with PC2 on the x-axis and PC3 on the y-axis. Samples were plotted separately in two groups that have already diverged on PC1 (see Fig 2.5a).

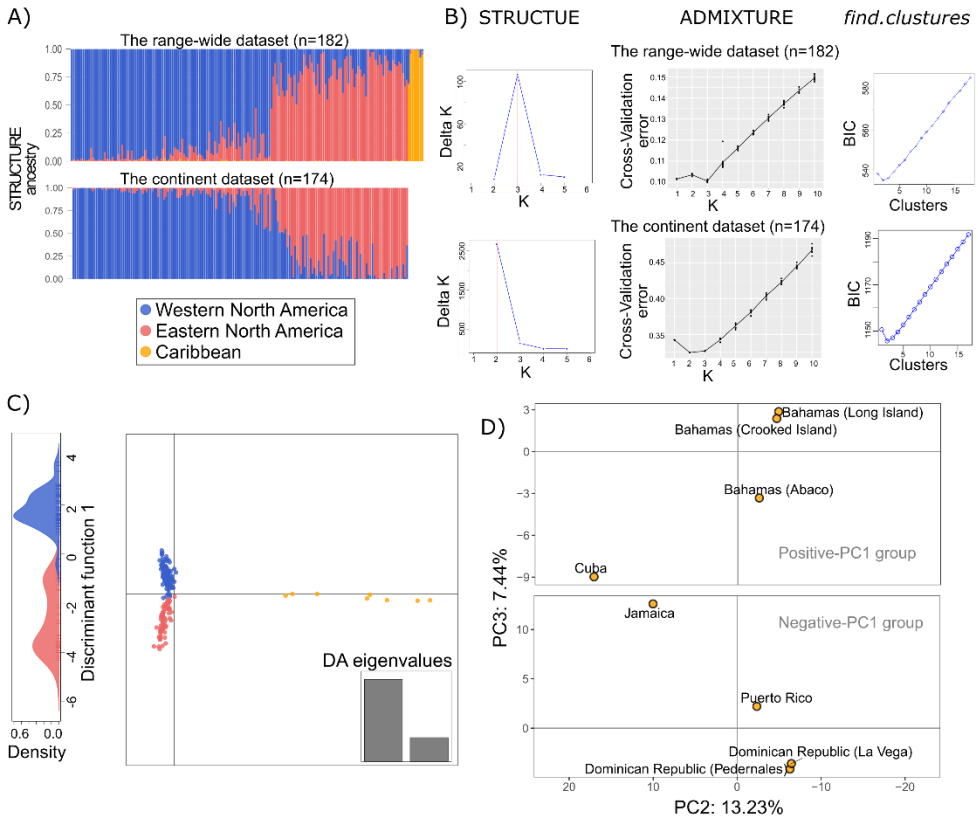


Figure S2. ADMIXTURE results of the range-wide dataset at K of two and three (optimal). **Left**: each bar represents one individual. Individuals are ordered according to the tips of the ML tree (Fig 2.2; Fig S4) and labeled with the State/Province of their sampling sites. **Right**: Mercator projection maps showing distributions of all retained big brown bat samples (n=182). Each pie chart represents one individual and those selected in the reduced dataset (n=26) are in a black outline.

Figure S3. ADMIXTURE results of the continent dataset at K of two (optimal) and three. **Left**: each bar represents one individual. Individuals are ordered according to the tips of the ML tree (Fig 2.2; Fig S4) and labeled with the State/Province of their sampling sites. **Right**: Mercator projection maps showing distributions of the retained big brown bat samples in the continental dataset (n=174). Each pie chart represents one individual and those selected in the reduced dataset (n=18) are in a black outline.

Note: figures S2, S3 are independent Supplemental Files.

Figure S4. The estimated effective migration surfaces (EEMS) with deme grids. The input habitat is equally divided into triangle deme grids (density parameter nDemes=800) in grey lines on the Mercator projection maps. Black dots represent the demes with samples. **A)** Effective migration rates estimated as genetic dissimilarities between demes. **B)** Effective diversity rates estimated as genetic dissimilarities between individuals within the same deme.

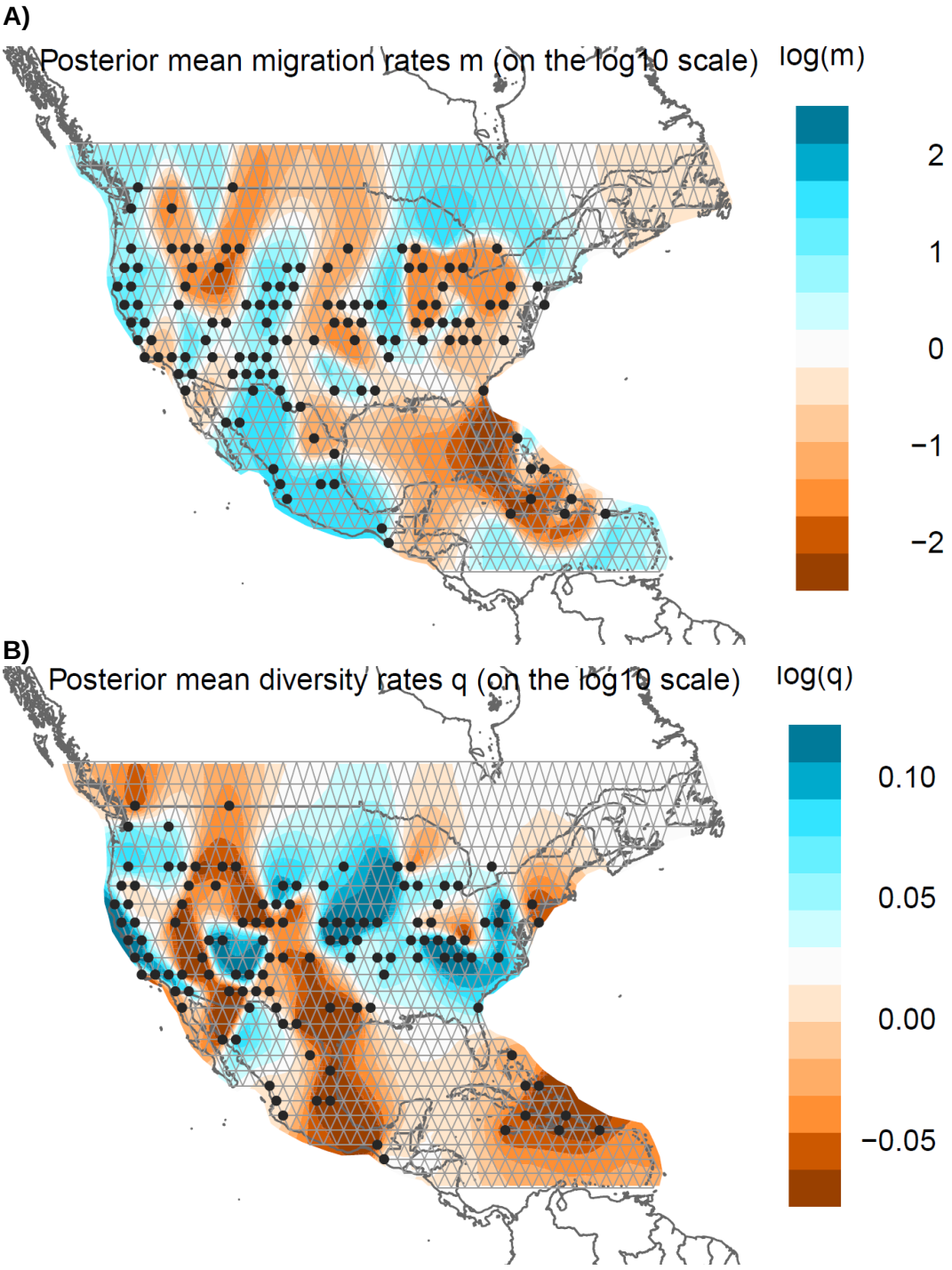


Figure S5. The maximum likelihood tree of the phylogeography dataset. Tips are shown in the same order as in Fig 2.2 and are labeled by species names (5 outgroups) or Country_State/Province of sampling sites (182 ingroups). Tip labels are coloured based on the assigned phylogeographic clades. Tip nodes in black represent the subset of 35 putatively "pure" individuals used to clarify the topology. Numbers at the nodes represent bootstrap supports (only values > 75 are shown).

Figure S6. Tree analyses of the subset phylogeography dataset (n=128, excluding the Southwest individuals). Numbers at the nodes represent bootstrap supports. Both the RAxML and the SVDquartets-individual trees show tips representing individuals that are labeled by species names (5 outgroups) or Country_State/Province of sampling sites (123 ingroups). Tip labels are coloured based on the assigned phylogeographic clades. The SVDquartets-group trees show tips representing phylogeographic clades with the two Pacific subclades combined or separated in the analyses.

Figure S7. Tree analyses of the subset phylogeography dataset using potentially "pure" individuals (n=35). A) The Mercator projection map showing geographic distributions of the selected big brown bats (n=30, coloured by phylogeographic clades) and the other individuals (n=157 including the 5 outgroups, in grey) in the phylogeography dataset. **B)** The ML tree of the subset dataset. **C)** The quartet-based tree with each tip representing one individual. Individual tips in B and C are labeled by species names (5 outgroups) or Country_State/Province of sampling sites (30 ingroups). **D)** The quartet-based trees with tips representing phylogeographic clades.

Figure S8. Tree analyses of the reduced dataset (n=27) using A) RAxML, B) SVDquartets with tips representing phylogeographic clades, and C) SVDquartets with tips representing individuals. Individual tips in A and C are labeled by species names (1 outgroups) or Country_State/Province of sampling sites (26 ingroups).

Note: Figures S5-S8 are independent Supplemental Files.

Figure S9. The summary of TreeMix results with different migration edges in optM. The explained variance shows model fitness, and the Evanno method is used to select the optimal number of migration edges (m=4).

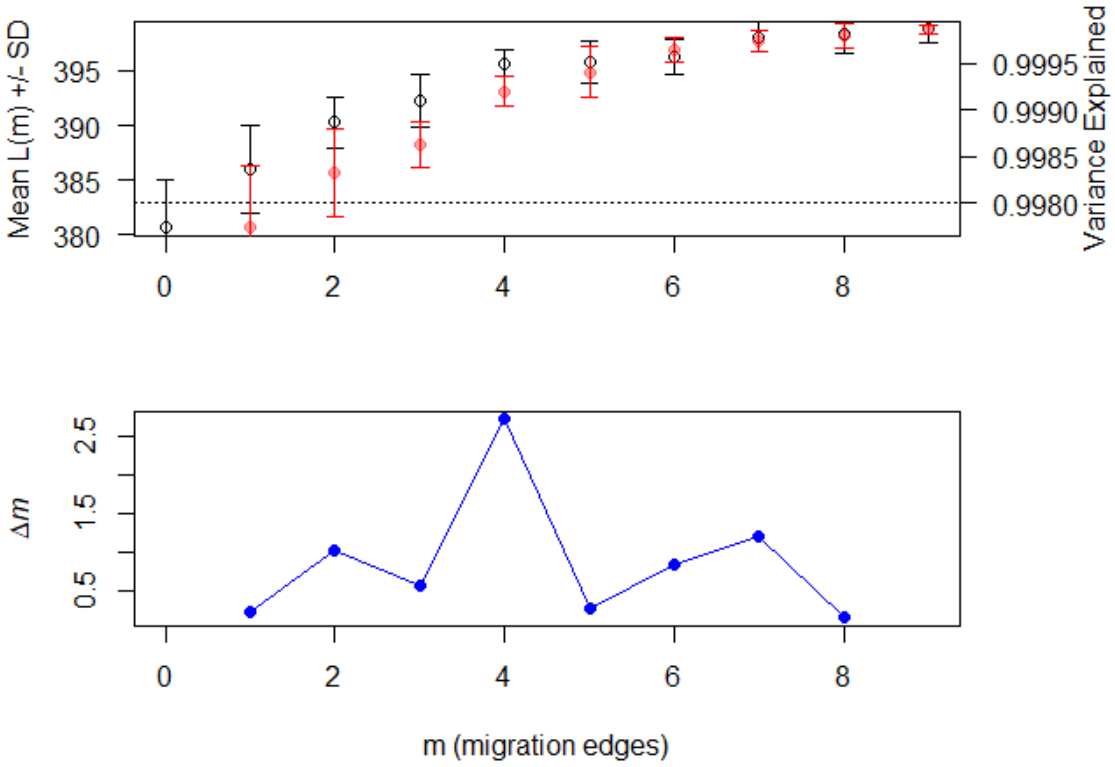
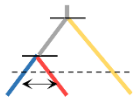
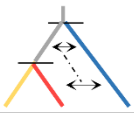
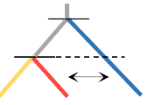
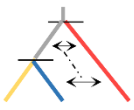
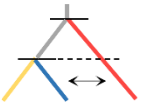
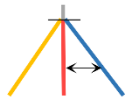
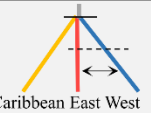
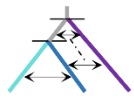
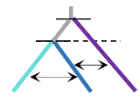
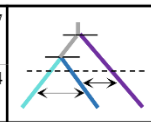
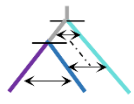
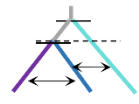
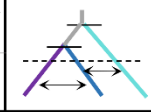
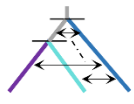
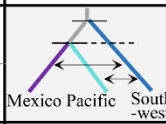
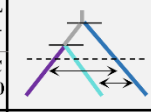
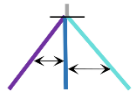
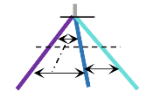
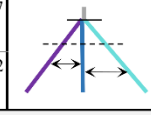


Figure S10. Hypothesis testing using demographic models. Arrows represent bidirectional gene flow, solid lines represent divergence events, and dashed lines represent historical isolation (i.e., a change of migration matrix in fastsimcoal2). Numbers on the right of each model are the delta likelihoods (top, estimated-observed likelihoods) and delta AIC values (bottom, deviation from the lowest AIC using the same dataset). Note that the delta likelihoods are high and indicate that these models are not the best estimations of our data, but comparisons among models are still informative and useful for our hypotheses testing. **A) Models of three nuclear populations for testing phylogeographic hypotheses.** The west population includes individuals from Pacific, Southwest, and Mexico. **B) Models of the three western clades for testing hypotheses of western refugia and divergence patterns. C) Models of all five phylogeographic clades for testing western monophyly and divergence patterns.** The better-fit model (i.e., the lowest AIC) of all five phylogeographic clades was used for parameter estimation. Point estimates were from the best run using the observed DSFS, and confidence intervals (CIs) were from non-parametric bootstrapping on 100 pseudoreplicate DSFS generated by random replacement using vcf2sfs. Each pseudoreplicate DSFS had 50 independent runs under the best model, with initial values from the best run of the observed DSFS (--initValues), 100,000 simulations, 20 ECM cycles, and the other settings same as those in the main text. Parameter estimates from the 100 best runs of pseudoreplicate DSFS were used to calculate 95% percentile CIs in the R function *quantile* and results are given below. Divergence time (assuming 2 years per generation): T1=2110 (95% CI: 2067, 2634) years, T2=3794 (3734, 6536) years, T3=6020 (5935, 8739) years. Migration rate per generation: Southwest-East=4.4E-04 (3.5E-04, 4.6E-04), Southwest-Pacific=1.1E-03 (8.0E-04, 1.4E-03), Southwest-Mexico=7.3E-04 (3.2E-04, 9.9E-04). Effective population size (# diploid individuals): NeCaribbean=1425 (1411, 2254), NeEast=5893 (5592, 8866), NeSouthwest=5909 (5864, 8061), NePacific=5356 (5220, 8066), NeMexico=5679 (5419, 18083).

A) Three nuclear populations: west -east isolation + secondary gene flow

	Strict isolation	Continuous gene flow	Historical isolation + secondary gene flow
Caribbean first	 2389.91 / 77.77	 2384.29 / 53.90	 2379.22 / 32.53
West first	 2488.27 / 530.71	 2381.74 / 42.12	 2380.06 / 34.42
East first	 2487.42 / 526.82	 2381.03 / 38.88	 2381.19 / 39.62
Simultaneous divergence	 2481.37 / 496.97	 2374.19 / 5.38	 **deltaL** 2372.59 / **deltaAIC** 0 Caribbean East West

B) Three western clades: historical isolation

	Continuous gene flow	Two isolated clades	Three isolated clades
Mexico first	 1340.95 / 724.40	 1341.47 / 726.84	 1323.37 / 645.47
Pacific first	 1341.67 / 727.74	 1342.32 / 730.71	 1342.90 / 735.42
Southwest first	 1340.69 / 723.21	 **deltaL** 1183.64 / **deltaAIC** 0 Mexico Pacific South -west	 1183.59 / **1.77**
Simultaneous divergence	 1341.31 / 724.10	 1341.47 / 726.82	 1341.38 / 726.40

C) Five phylogeographic clades: monophyletic western population with Pacific diverged first

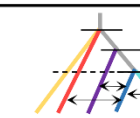
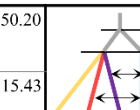
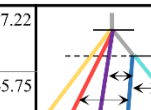
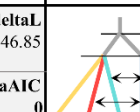
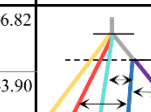
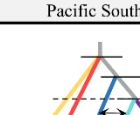
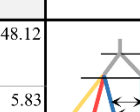
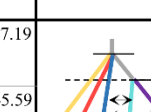
	West monophyly	West paraphyly	Simultaneous divergence
Mexico first	 2550.20 / 15.43	 2557.22 / 45.75	 2549.20 / 8.81
Pacific first	T3 T2 T1 Caribbean East Mexico Pacific Southwest **deltaL** 2546.85 / **deltaAIC** 0	 2556.82 / 43.90	 2549.92 / 12.14
Southwest first	 2548.12 / 5.83	 2557.19 / 45.59	 2550.84 / 16.38

Figure S11. Testing the fit of tree topologies using demographic models. Arrows represent bidirectional gene flow, solid lines represent divergence events, and dashed lines represent historical isolation (i.e., a change of migration matrix in fastsimcoal2). Numbers on the right of each model are the delta likelihoods (top, estimated-observed likelihoods) and delta AIC values (bottom, deviation from the lowest AIC using the same dataset). The highlighted models have the lowest AIC among the displayed scenarios.

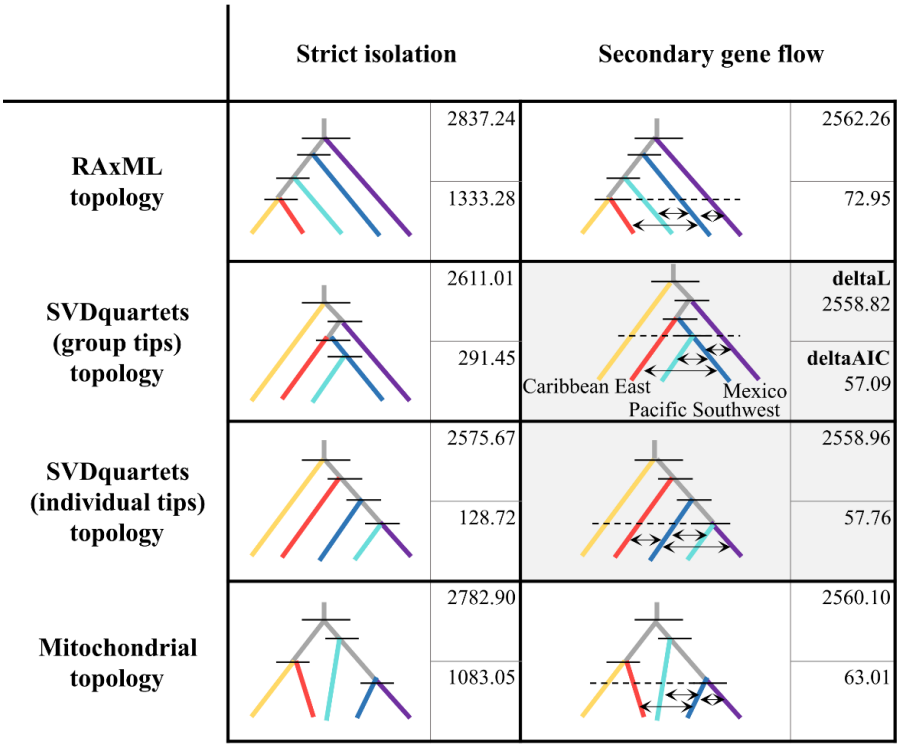


Figure S12. Response curves of the seven selected bioclimatic variables estimated using the current species distribution model (1960-1990). See Table S4 for variable descriptions. The x-axis shows the range of the variable, and the y-axis shows the estimated probability of presence. The curves are shown with standard deviation error bars generated from the 10-fold cross-validation. Note that MaxEnt estimates each response curve using a model built with only the corresponding bioclimatic variable.

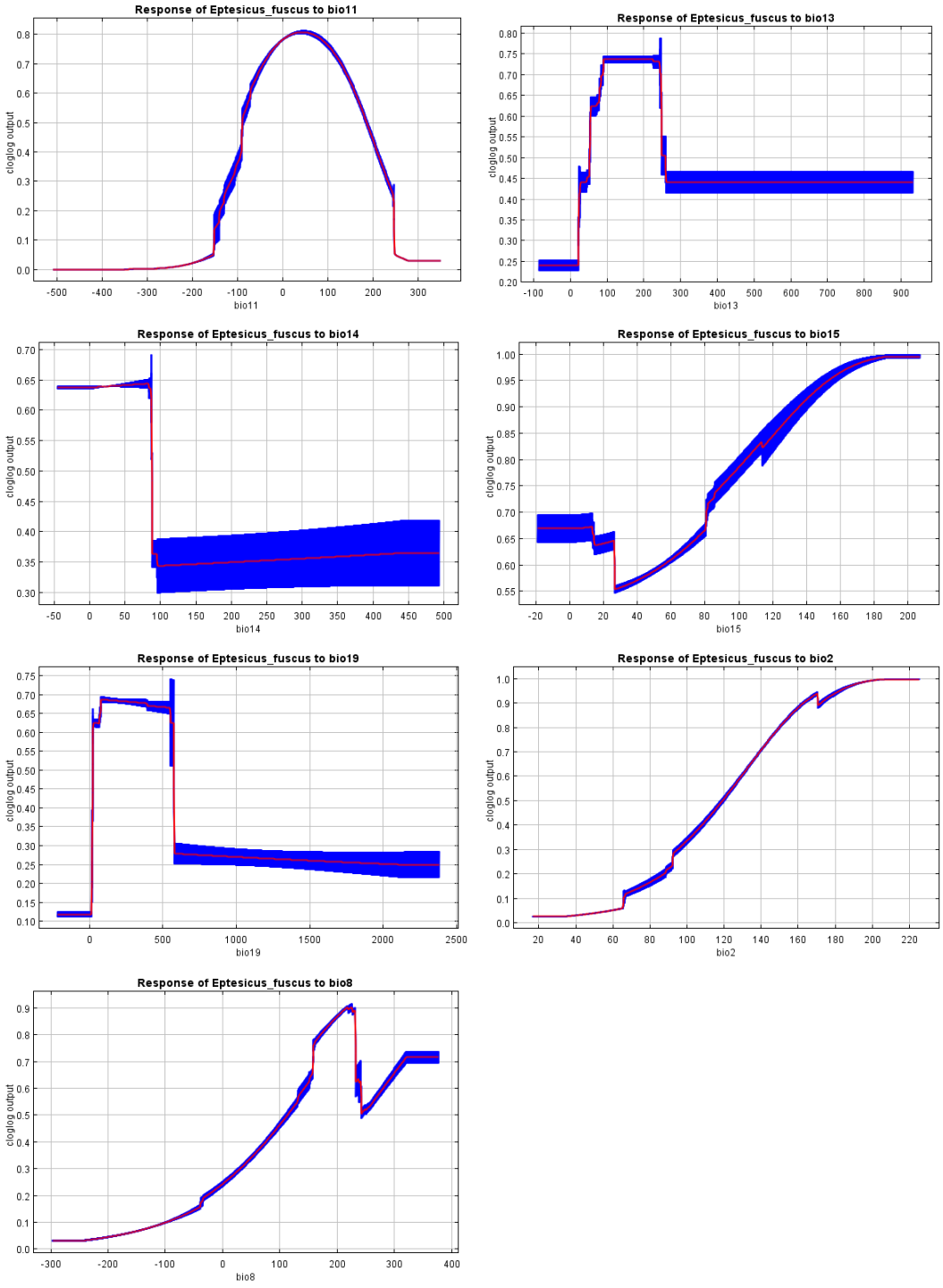
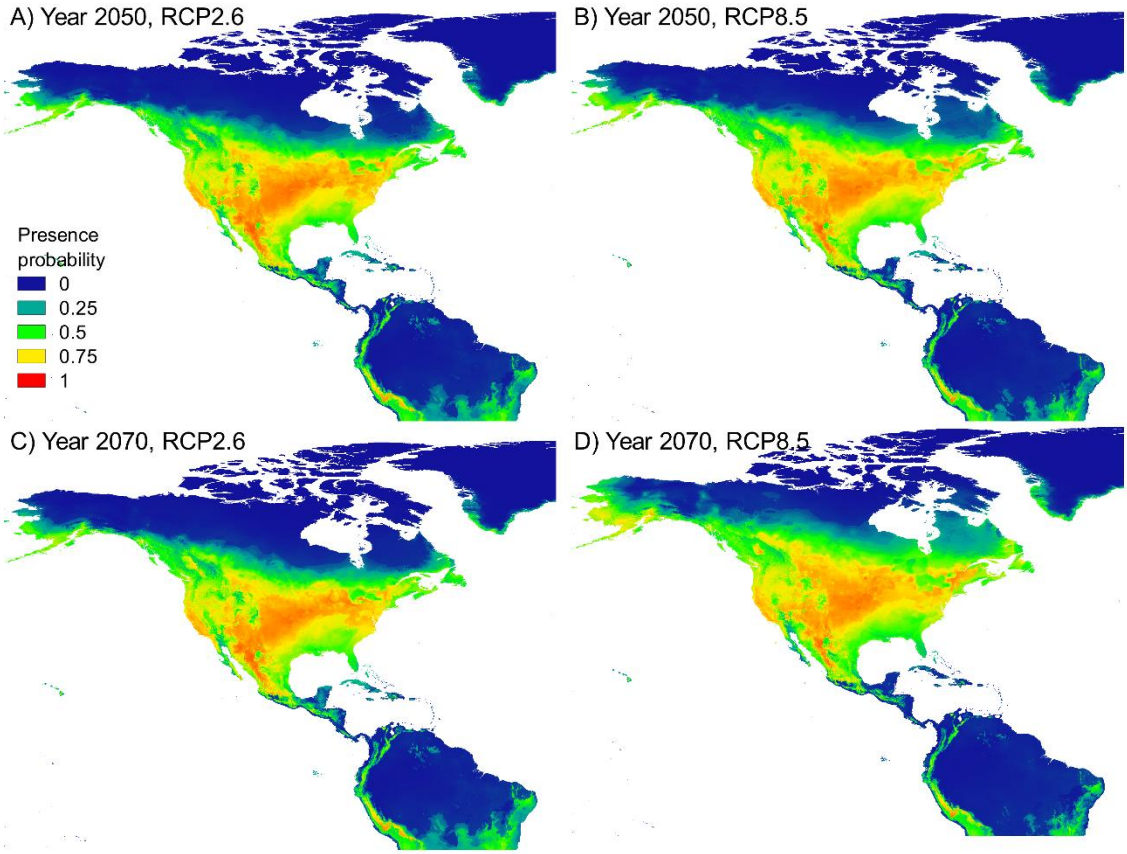


Figure S13. Predicted future species distribution in 2050 (A & B) and 2070 (C & D) using the Mercator projection. Two climate change scenarios were used representing the minimum (RCP 2.6, A & C) and maximum (RCP 8.5, B & D) greenhouse gas emissions.



Appendix C

Supplementary materials

Nonrandom missing data can bias PCA inference of population genetic structure

Xueling Yi[1], Emily K. Latch[1]
[1] Behavioral and Molecular Ecology Research Group, Department of Biological Sciences, University of Wisconsin-Milwaukee, Milwaukee, Wisconsin, USA

Table S1. Sample information of the 72 big brown bats retained in the empirical data sets. Museum specimens are listed with the corresponding catalogues. Individual missing values in the three empirical data sets are included.

Note: Table S1 is an independent Supplemental File.

Figure S1. PCA on the raw SNP matrices from replicated simulations. Each model was simulated additional four times. PCA plots of the replicates simulated using the same model were highly consistent.

Figure S2. PCA on the missingness-introduced matrices in the p3 model.
Figure S3. PCA on the missingness-introduced matrices in the p3_mig model.
Figure S4. PCA on the missingness-introduced matrices in the cline model.
Figure S5. PCA on the missingness-introduced matrices in the island model.

** Note: figures S1-S5 are independent Supplemental Files.*

Figure S6. Examples of PCA without centering standardization. The analyzed data sets are indicated in the plot titles. Population labels correspond to the diagram in Fig 3.1. All incomplete data sets have the total 20% missingness. PCA was conducted using the function glPca without centering (center=F) and all PCs were retained.

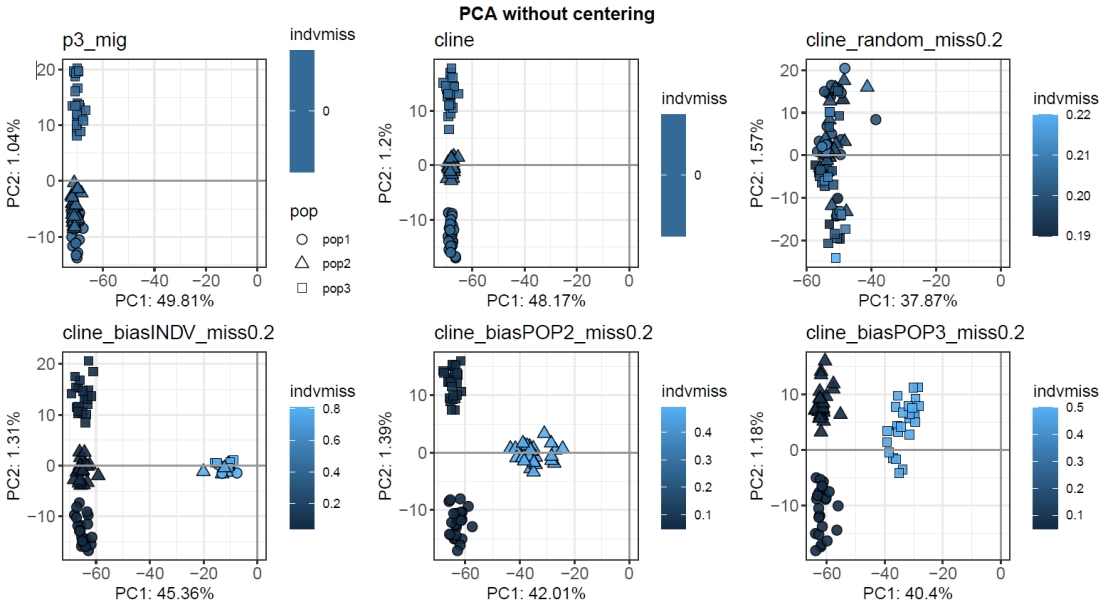


Figure S7. Continental divergence in the empirical bat data indicated by PC3 and PC4. Individuals from the East population are plotted in squares, and individuals from the West population are plotted by the states / provinces where they were collected. PC2 captured the continent-island divergence (see in Fig 3.5) and thus was not included here. The outlier individual in the data set bat10 was collected from Vermont. The data set bat20 indicates substructure within the West population with roughly latitudinal divergence on PC3 and roughly longitudinal divergence on PC4. Importantly, these plots of higher ordered PCs also indicate missing data effects with high-missingness individuals being dragged to the PCA origin.