UNIVERSITY OF CALIFORNIA, MERCED

A tale of two rodents: Understanding past distribution of genetic diversity, population structure, and areas of refugia

A dissertation submitted in partial satisfaction of the requirements
for the degree Doctor of Philosophy

in

Quantitative Systems Biology

by

Robert A. Boria

Committee in charge:

Professor Michael N Dawson, Chair
Professor Jessica L. Blois
Professor Emily Jane McTavish
Professor Jason Sexton

2021

The Dissertation of Robert A. Boria is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____
Jessica L. Blois, Advisor

_____
Michael N Dawson, Chair

_____
Jason P. Sexton

_____
Emily Jane McTavish

Table of Contents

List of Tables

List of Figures

Acknowledgements

This dissertation was made possible by numerous people. First, I'd like to thank my wife, Sasha Acosta, and our children for allowing me to follow my dreams and complete this degree. Their love and support over the last six years has kept me going through all ups and downs. Secondly, I would like to thank my parents, Arcadio and Josephine Boria, and two sisters for providing the structural support over the many years.

I would like to thank my wonderful advisor, Dr. Jessica Blois, for her time, guidance, and feedback at every stage of my doctoral degree. Dr. Blois displayed an enormous amount of patience while I gained wet lab experience and she never once questioned my competency. I would also like to thank Dr. Emily Jane McTavish for her time and expertise, she never once turned me away from her office. I would also like to thank my other two committee members, Dr. Michael Dawson and Jason Sexton, for their guidance and help along this journey. Finally, I would like to thank all the members of the Blois lab over the last six years: Nate Fox, Eric Williams, Sarah Brown, Danaan DeNeve Weeks, Mairin Balisi, Paulo Ricardo de Oliveira Roth, Kinsey Brock, Laura Van Vranken, Giovanni Rapacciuolo, Corey Shaver, and Kaitlin Maguire. I would like to specifically acknowledge Dr. Sarah K. Brown for training me in the wet lab. She never wavered in her support and answered every question I had including after she left the lab. Lastly, I would like to thank everyone in the Blois-McTavish lab meeting, they have all helped in numerous ways and I am forever grateful.

I would like to thank collaborators and researchers from other Universities that have provided guidance and encouragement throughout my Ph.D. Dr. Marjorie Matocq was an amazing collaborator on my second chapter. Dr. Diego Alvarado-Serrano answered every question I had during my Ph.D. and was always available to help troubleshoot scripts used throughout my dissertation. Dr. Robert Anderson had provided guidance and support throughout my scientific career. Dr. Joshua Reece was my NSF AGEP mentor during my time at Fresno State. He taught me new techniques for teaching and mentoring his students.

My dissertation would not be possible without the help of Museum collections, which are extremely valuable resources. I gratefully acknowledge the Burke Museum, Field Museum, Museum of Southwestern Biology, Museum of Vertebrate Zoology, and University of Alaska Museum of the North. Dr. Dean Pearson also provided tissues used in this project from his lab collections.

This project was funded in part by Ford Foundation and National Science Foundation grants HRD-1820875, 1820876, 1820886, and 1820895. These resources allowed me to focus almost exclusively on my research throughout my

graduate career. The lab costs for this project were also funded in part by National Science Foundation, EAR-1750597 and Larry D. Agenbroad Legacy grant program at the Mammoth Site.

Chapter 1 is a full reprint of the material as it appears in Ecological Modelling: https://www.sciencedirect.com/science/article/pii/S0304380018302825?via%3Dihub. © 2018 Ecological Modelling. The dissertation author, Robert A. Boria, designed the study, led the acquisition, analysis, and interpretation of data, and wrote the article. Jessica L. Blois directed and supervised this research, which forms the basis for the dissertation, and revised the manuscript.

Chapter 2 is a full reprint of the material as it appears in Heredity: https://www.nature.com/articles/s41437-020-00393-7. © 2021 Heredity Open Access CC BY 4.0. Robert A. Boria designed the study, led analysis, and interpretation of data, and wrote the article. Sarah K. Brown helped acquire data and revised the manuscript. Marjorie D. Matocq helped interpret the results and revised the manuscript. Jessica L. Blois directed and supervised this research, which forms the basis for the dissertation, provided analytical support, and revised the manuscript.

Curriculum vitae

Education:

University of California at Merced, Merced, CA, USA
Ph.D. *in* Quantitative and Systems Biology (2015−present)
Research advisor: Dr. Jessica L. Blois

The City College of New York, New York, NY
Master of Science (Biology), 2014

The City College of New York, New York, NY
Bachelor of Science (*cum laude)*, 2012
Major: Biology; Minor: Education

Fellowships:

Graduate Dean's Dissertation Fellowship, University of California at Merced
$11,000 one-semester stipend, Spring 20202

National Science Foundation AGEP Fellowship (Through UC Merced)
$12,000 one-semester stipend, Fall 2019

Ford Foundation Predoctoral Fellowship
$27,000 per year stipend, August 2017 – August 2019; August 2020 – August 2021

Quantitative and Systems Biology Graduate Group Recruitment Fellowship, University of California at Merced
$5,000, August 2015

Bridge to Doctorate Fellow, *Luis Stokes Alliance for Minority Participation in Science, Math, Engineering, and Technology Research Experience*
$30,000 per year stipend, February 2012 – February 2014

Peer-reviewed Publications:

Boria, RA, SK Brown, MD Matocq, and JL Blois. Understanding the population structure and historical demography of the *Neotoma fuscipes* in California*. Heredity*, 126: 521–536

Boria, RA, and JL Blois. 2018. The effect of large sample sizes on ecological niche models: Analysis using a North American rodent, *Peromyscus maniculatus*. *Ecological Modelling*, 386: 83–88.

Boria, RA., LE Olson, SM Goodman, and RP Anderson. 2017. A single-algorithm approach to creating an ensemble of ecological niche models and generating maps of uncertainty. *Diversity and Distributions*, 23: 196–208.

Aiello-Lammens, ME, **RA Boria**, A Radosavljevic, and RP Anderson. 2015. spThin: An R function for spatial thinning of species occurrence records for use in ecological niche models. *Ecography*, 38: 541–545.

Muscarella, R, PJ Galante, M Soley-Guardia, **RA Boria**, JM Kass, M Uriarte, and RP Anderson. 2014. ENMeval: An R package for tuning MAXENT

ecological niche models with spatially independent evaluations. *Methods in Ecology and Evolution*, 5: 1198–1205.

Boria, RA, LE Olson, SM. Goodman, and RP Anderson. 2014. Spatial filtering to reduce sampling bias can improve the performance of ecological niche models. *Ecological Modelling*, 275: 73–77.

Gutiérrez, EE, **RA Boria**, and R. P. Anderson. 2014 Can biotic interactions cause allopatry? Niche models, competition, and distributions of South American mouse opossums. *Ecography*, 37: 741–753. (featured on journal cover)

Manuscripts in Review:

Kass, JM, R Muscarella, PJ Galante, CL. Bohl, GE Pinilla-Buitrago, M Soley-Guardia, **RA Boria**, and RP Anderson. ENMeval 2.0: redesigned for customizable and reproducible evaluations for models of species' niches and distributions (Submitted to Methods in Ecology and Evolution)

Boria, RA, JA Aguilar, S Bever, M Estrada, CN Graves, R Kaur, TR Langdon, BK Mulvey, TR Nicolo, CO O'Keefe, KD Peters, J Salazar, RT Seward, CN Thomas, E Torres, MM Vidal, MP Woodbury, D Xiong, E Meyer, and J Reece. Impacts of the Rough Fire on bat biodiversity in Sequoia and Kings Canyon National (Submitted to *PLoS One*)

Manuscripts in Preparation:

Boria, RA and JL Blois. Phylogeography of two lineages of *Peromyscus maniculatus* (deer mouse): Understanding past distribution of genetic diversity and areas of refugia. (To be submitted to *Molecular Ecology*)

Research Support:

2018. Larry D. Agenbroad Legacy grant program, the Mammoth site. Funding to perform aDNA research on deer mice in northern California. $1,000.

2014. Funding to Attend Scientific Conference, National Science Foundation: DEB-1119915 to RP Anderson. $426.50.

2014. The City College Student Development Fund Award. Travel award. $244.

2014. Funding to Attend Scientific Conference, National Science Foundation: DEB-1119915 to RP Anderson. $985.75.

2014. Funding for *Research assistant* on grant (DEB-1119915) "Collaborative Research: Testing species limits, phylogeographic concordance, and niche evolution in Madagascar's endemic small mammals" to RP Anderson. $3,750.

2013. International Biogeography Society Registration and Workshop Waiver Award. $280.

2011. Luis Stokes Alliance for Minority Participation in Science, Math, Engineering, and Technology Research Experience. $5,000.

2011. Luis Stokes Alliance for Minority Participation in Science, Math, Engineering, and Technology Research Experience. $1,500.

2010. Associated Medical Schools of New York. $4,500.

2010. Luis Stokes Advancement for Minority Participation in Science, Math, Engineering, and Technology Research Experience Summer program. $4,000.

Professional Presentations:

Departmental Seminars:

A tail of two rodents: Understanding past distribution of genetic diversity, population structure, and areas of refugia. California State University, Fresno, 06 September 2019, Fresno, CA. **R. A. Boria**

Oral presentations at scientific meetings:

Understanding the population structure and historical demography of the *Neotoma fuscipes* in California. The American Society of Mammalogists, 25–29 June 2018, Manhattan, KS. **R. A. Boria**, S. K. Brown, and J. L. Blois

Generating maps of uncertainty for ecological niche models using a single-algorithm ensemble approach. Ecological Society of America 2014 Meeting, 10–15 August 2014, Sacramento, CA. **R. A. Boria** and R. P. Anderson.

A single-algorithm approach to creating an ensemble of ecological niche models and generating maps of uncertainty. The American Society of Mammalogists, 6–10 June 2014, Oklahoma City, OK. **R. A. Boria** and R. P. Anderson.

Assessing niche evolution and population connectivity over glacial cycles for two Malagasy small mammals. The American Society of Mammalogists, 14–18 June 2013, Philadelphia, PA. **R. A. Boria**, M. Shcheglovitova, A. Radosavljevic, S. A. Jansa, L. E. Olson, and R. P. Anderson.

Comparing species-specific tuning versus AIC-c to select optimally complex ecological niche models. The American Society of Mammalogists, 14–18 June 2013, Philadelphia, PA. P. J. Galante, **R. A. Boria,** R. P. Anderson. (Presented by P.J. Galante)

Using ENMs to study the effects of climatic change on species distributions: analyses with the Malagasy small mammals *Eliurus majori* and *Microgale longicaudata*. The American Society of Mammalogists, 22–26 June 2012, Reno, NV. **R. A. Boria**, M. Shcheglovitova, A. Radosavljevic, S. A. Jansa, L. E. Olson, and R. P. Anderson.

Poster Presentations:

scientific meetings:

Genetic structure and diversity through time for northern California *Peromyscus maniculatus*. Society of the Study Evolution 2017 Meeting, 23–27 June 2017, Portland, OR. **R. A. Boria**, S. K. Brown, and J. L. Blois.

The effect of large sample sizes on Ecological Niche Models: analysis using a North American rodent, *Peromyscus maniculatus.* International Society of

Biogeography 2017 Meeting, 9-13 January 2017, Tuscon, AZ. **R.A. Boria** and J. L. Blois.

Assessing niche evolution and population connectivity over glacial cycles for two Malagasy small mammals. Ecological Society of America 2013 Meeting, 4–9 August 2013, Minneapolis, MN. **R. A. Boria**, M. Shcheglovitova, A. Radosavljevic, S. A. Jansa, L. E. Olson, and R. P. Anderson.

Comparing species-specific tuning versus AICc to select optimally complex ecological niche models. Ecological Society of America 2013 Meeting, 4–9 August 2013, Minneapolis, MN. P. J. Galante, **R. A. Boria,** R. P. Anderson. (Presented by P.J. Galante)

The effect of sampling bias on Ecological Niche Models: analysis using the Malagasy mammal *Microgale cowani*. International Society of Biogeography 2013 Meeting, 9-13 January 2013, Miami, FL. **R.A. Boria** and R.P. Anderson.

Using ENMs to study the effects of climatic change on species distributions: analyses with the Malagasy small mammals *Eliurus majori* and *Microgale longicaudata*. Society of the Study Evolution 2012 Meeting, 6–10 July 2012, Ottawa, Ontario, Canada. **R. A. Boria**, M. Shcheglovitova, A. Radosavljevic, S. A. Jansa, L. E. Olson, and R. P. Anderson.

Employment History:

Teaching Assistant, **University of California, Merced**, Merced, CA

Ess 050, Ecosystems of California (discussion section), University of California, Merced. Spring, 2016

Teaching Assistant, **University of California, Merced**, Merced, CA

Bio 001, Contemporary Biology (laboratory section), University of California, Merced. Fall, 2015

Adjunct Instructor, **Baruch College**, New York, NY

ENV 103L, Fundamentals of Ecology (recitation section), City University of New York (CUNY). Spring, 2015

Adjunct Instructor, **Baruch College**, New York, NY

Bio 1015, 1016, Human Biology (recitation/laboratory section), City University of New York (CUNY). Spring, 2015

Adjunct Instructor, **The City College of New York**, New York, NY

Bio 100, Human Biology (laboratory section), City University of New York (CUNY). Spring, 2015

Adjunct Instructor, **The City College of New York**, New York, NY

Bio 228, Ecology and Evolution (laboratory section), City University of New York (CUNY). Fall, 2014

Graduate Teaching Assistant, **The City College of New York**, New York, NY

Bio 102, Foundations of Biology II (laboratory sections), City University of New York (CUNY). Spring, 2014

Research Intern, **The City College of New York**, New York, NY

Assist graduate students and lab-mentor on their research projects. My tasks included: proofing spreadsheets; creating graphs; organizing and proofing bibliographies; lab up-keep. Spring, 2011

Manuscript Review:

Biogeosciences, Biological Journal of the Linnean Society, Diversity and Distributions, Ecography, Journal of Animal Ecology, Journal of Biogeography, Journal of Mammalogy, Journal of Wildlife Management and Wildlife Monographs, Landscape Ecology, Nature Communications

Abstract of the Dissertation

A tale of two rodents: Understanding past distribution of genetic diversity, population structure, and areas of refugia

by

Robert A. Boria

Doctor of Philosophy, Quantitative Systems Biology
University of California, Merced 2021
Dr. Jessica L. Blois, Graduate Advisor
Dr. Michael N. Dawson, Chair

Forecasting how biodiversity will change in the future due to natural and anthropogenic impacts is a primary focus of both ecology and evolutionary biology. By understanding the historical processes influencing the current geographic distribution of biodiversity, we can determine the relative importance of different factors shaping biodiversity, now and in the future. **The main question that drove this dissertation: What historical processes led to the current distribution of diversity across the landscape?** One omnipresent influence on the geographic distribution of diversity at multiple levels– e.g., genes and species– is climate. Understanding the spatial distributions of intraspecific genetic diversity and the role of climate and climate refugia in evolutionary and ecological processes is important because it shapes species potential for persistence in the face of future climate change. My dissertation focuses on how populations have responded to past climate change, and how the historical distributions and past areas of climate refugia will influence future climate change responses, using mammals as the study system.

Studying population histories through time, we can uncover how different populations with similar genetic reservoirs respond differently to the same environmental stressor (climate). Determining how the distribution of intraspecific diversity of North American taxa was directly influenced by climate and landscape changes may illuminate broad-scale patterns of species' responses to other climatic events, or more generally, to barriers impeding or constraining gene flow. My dissertation research utilized an interdisciplinary approach– next generation sequencing; GIS data; ecological modeling; bioinformatics– to understand how historical events have shaped the current distribution of genetic diversity within mammals. My aim was to study how mammal populations respond to climate change through time by determining the range dynamics and potential areas of refugia (Chapter 2 and 3).

Understanding the spatial distributions of intraspecific genetic diversity and the role of climate refugia in the evolutionary and ecological processes of populations is important because it may determine their potential for persistence in the face of future climate change. My dissertation examined how two small mammals responded to the glacial cycles of the Pleistocene in North America. First, I determined *Neotoma fuscipes* has three historical populations in California—two northern and one southern population (Chapter 2). The major split between the northern and southern populations is older than 1.7 million years and occurred in the San Francisco Bay-Delta region, a historically significant region with high lineage diversification in mammals, amphibians, and reptiles. I detected multiple refugia within the species, including several origins of expansions and contractions (particularly with the northern populations). Second, I examined two western lineages within *Peromyscus maniculatus* and identified three main populations: 1) southern California; 2) a small population nested within the broader Pacific Northwest; 3) the Pacific Northwest through central California and across the Rocky Mountains (Chapter 3). These populations diverged within the last 160,000 years with very little migration, and evidence of recent population expansion. I found evidence for multiple areas of refugia including southeastern Alaska, a known refugia for several mammal species.

In order to connect patterns and processes observed in the past with projections of change for the future, I pair my focus on generating empirical genetic & genomic data with different modeling types. Specifically, ecological niche models (ENM) are powerful tools for approximating the abiotically suitable area of a species by comparing environmental conditions at localities where the species occurs with the overall conditions available in the study region. Many ENM studies struggle with small sample sizes, but modeling widely distributed and well-sampled cosmopolitan species raises computational issues as well (Chapter 1). I thus determined the number of localities needed to model the distribution of a cosmopolitan species (*Peromyscus maniculatus*) with many occurrence records. I discovered when modeling species with a large number of occurrence records, it may not be necessary to use all localities for ENMs and could potentially affect model performance negatively.

Chapter 1: The effect of large sample sizes on Ecological Niche Models: analysis using a North American rodent, *Peromyscus maniculatus*

1.1 Abstract

Correlative ecological niche models (ENMs) aim to approximate the environmentally suitable areas for a species. Recently, studies have explored the minimum number of occurrence records needed to implement ENMs; however, cosmopolitan species with many occurrence records have their own challenges and the effects of larger sample sizes on ENM performance have yet to be determined. To address this issue, we focused on a New World rodent, *Peromyscus maniculatus*. We obtained locality data from GBIF (13,334 unique records), and spatially filtered the localities. We then modeled suitable area for the species using Maxent, two different environmental datasets (at different spatial resolutions) and different numbers of occurrence records (with 25 replicates per sample size). We evaluated the models with k-fold cross-validation, AUC, and two omission rates. Further, we calculated the variability among predictions within and between datasets to indicate variation in geography. Generally, the AUC and omission rate both decreased as sample size increased. Lastly, as sample size increased, similarities in geography increased within and between datasets. For *P. maniculatus*, we get similar performing models, both in terms of geographic predictions and evaluation statistics, with as few as 10% – 20% of the maximum number of localities for each environmental dataset. Using a large number of occurrence records may not be necessary for ENMs, and in fact may hinder model performance.

1.2 Introduction

Ecological niche models (ENMs; oft termed Species Distribution Models) are correlative models that approximate the environmentally suitable areas for a species. They do so by comparing the overall environmental conditions available to the species with the conditions at localities where the species occurs (Peterson et al., 2011). Because ENMs can be projected across time and space they are frequently used in a wide range of ecological and evolutionary studies. However, despite their relevance to many fields, several methodological issues (such as, sample bias; correlation of variables; model evaluation) hinder their implementation in many systems (Anderson, 2012; Merow et al., 2013). Here, we study the effects of large numbers of occurrence records on ENMs.

Several studies have determined the minimum number of occurrence records needed to implement ENMs for different algorithms

(Stockwell and Peterson, 2002; Papeş and Gaubert, 2007; Wisz et al., 2008; Proosdij et al., 2016), revealing that as few as five localities can produce biologically meaningful models (Pearson et al., 2007; Shcheglovitova and Anderson, 2013; Galante et al., 2018). However, cosmopolitan species with many occurrence records present their own challenges and the effects of larger sample sizes on models have yet to be determined. These species typically have large geographic extents encompassing heterogeneous environments, which could potentially pose problems when trying to estimate the environmental suitability of areas. Using too few localities may not capture the species' entire niche; however, using too many occurrence records may not increase model accuracy and could potentially hinder model performance (i.e., by limiting the amount of available background localities and reducing discriminatory ability; (Stockwell and Peterson, 2002; VanDerWal et al., 2009). Further, as sample size increases so do associated issues of sampling bias and georeferencing errors, which could potentially lead to overfit models (Boria et al., 2014; Bloom et al., 2018).

In this study, we aim to determine how large sample sizes affect ENM evaluations and predictions in geography. Specifically, using a widely distributed species (*Peromyscus maniculatus*), we calibrated and evaluated our models using a spatial partition method, two different environmental datasets (at different spatial resolutions) – worldclim (Hijmans et al., 2005a) and Community Climate Simulation Model 3 (CCSM3; (Lorenz et al., 2016) – and an increasing number of localities. Further, we evaluated the predictions in geographic space to determine differences between the different datasets when they are projected across space. We explored if using smaller datasets would produce similar models to larger datasets, without noticeable improvements in model performance and in predictions. The results suggest that as sample size increases there isn't necessarily a noticeable increase in model performance. When modelling species with a large number of occurrence records, it may not be necessary to use all localities for ENMs and could potentially affect model performance negatively.

1.3 Materials and Methods

1.3.1 Input data

We conducted analyses with a New World rodent, *Peromyscus maniculatus* (North American deer mouse), which is indigenous to and distributed widely across North America (Shorter et al., 2012). This species can be found in every terrestrial ecosystem, and is the most diverse and widespread species of deer mouse (Dewey and Dawson, 2001; Bedford and Hoekstra, 2015). This species provides a good system because there are currently more than 100,000 occurrence records for *P. maniculatus* in the Global Biodiversity Information

Facility (GBIF). We downloaded all GBIF records that contained coordinates, a preserved specimen, and didn't have any geospatial errors, using the R (v. 3.4.1, R Development Core Team, 2016) package *rgbif* (Chamberlain, 2017). We then removed all duplicate records, and all localities that were placed in the ocean and outside the known geographic range (based on the NatureServe range estimate; (Patterson et al., 2007). These steps led to a final set of 13,334 unique *P. maniculatus* localities. Despite efforts to improve data quality, we acknowledge errors associated with GBIF data (e.g., Beck et al., 2014); however, the main goal of this study is the effects of large number of occurrence records on ENMs and not determine *P. maniculatus* environmental suitability.

Generally, easily accessible roads are sampled more frequently, generating issues with sampling bias (Hijmans et al., 2000; Reddy and Dávalos, 2003; Kadmon et al., 2004). Additionally, spatial biases within GBIF datasets have been shown to negatively affect ENM performance (Beck et al., 2014). To reduce sampling bias, we spatially filtered the occurrence dataset to ensure that no two localities were within a pre-determined distance of one another (while retaining the most localities possible) using the *spThin* package in R (Aiello-Lammens et al., 2015). This method effectively reduces artificially induced spatial auto-correlation (Kramer-schadt et al., 2013; Fourcade et al., 2014; Boria et al., 2014). Because of the difference in resolution among our climate predictor datasets (*see below*), we used a different distance for each climate dataset, based on their grid cell size: 25km for the worldclim dataset and 75km for CCSM3. The worldclim dataset was reduced to 3,334 localities after applying the spatial filter, and the CCSM3 dataset was condensed to 990 occurrence records. From these 'full' datasets, we then randomly sampled different numbers of occurrence records, with 25 replicates for each sample size. The worldclim sample sizes were: 25; 50; 100; 500; 1,000; 1,500; 2,000; 2,500; 3,000. For CCSM3 we used: 25; 50; 75; 100; 150; 200; 300; 400; 500; 600; 700; 800. These were the occurrence datasets used in the modeling exercises described below.

We used two different climate simulations that include climate inferences for the present day: worldclim (Hijmans et al., 2005a) and Community Climate Simulation Model 3 (Liu et al., 2009; Lorenz et al., 2016). These two simulations emphasize different tradeoffs: worldclim variables are downscaled to 1km x 1km grid cells for the entire world for present day, as well as 7,000 and 21,000 years ago, and thus have better spatial resolution. CCSM3 variables are downscaled to 0.5 x 0.5 degrees (~50km x 50km) grid cells for North America from 21,000 years ago to the present day at 500-year intervals, and thus have better temporal resolution. Although we do not consider paleoclimates in this paper (all ENMs are based only on contemporary climate inferences), our ultimate research goal is to infer *P. maniculatus* climate suitability and hindcast the ENMs to the past, which motivated our inclusion of these two climate models. The worldclim dataset has 19 bioclimatic variables that reflect aspects of temperature and precipitation (Hijmans et al., 2005b)). The worldclim variables are known to be correlated;

however, MaxEnt is a machine-learning algorithm that uses regularization to reduce complexity (especially regarding correlated variables), and thus not all variables are necessarily included in the final model (Phillips and Dudík, 2008; Elith et al., 2011). The CCSM3 environmental variables consist of 27 variables that also reflect features of temperature and precipitation (Lorenz et al., 2016). However, for the CCSM3 variables we followed the procedure of (Maguire et al., 2016) and only used the six least correlated variables over the last 21,000 years (minimum precipitation of the driest quarter, maximum temperature of the warmest quarter, mean yearly potential evapotranspiration, maximum precipitation of the wettest quarter, mean yearly water deficit index and mean yearly actual evapotranspiration).

To approximate modeling assumptions regarding dispersal and biotic interactions more closely, we delimited a custom study region for the full dataset (13,199 localities), specifically by drawing a minimum convex polygon around the localities and adding a 0.5° buffer (Anderson and Raza, 2010; Barve et al., 2011). Background localities for calibration (default number of 10,000) were taken from within the delimited study region only.

1.3.2 Ecological Niche Modeling

We used a machine learning approach, Maxent (v 3.3.3k; (Phillips et al., 2006; Phillips and Dudík, 2008), to generate the ENMs. This method is a presence-background technique and has performed well in comparison with other techniques (Elith et al., 2006; Wisz et al., 2008); but see (Royle et al., 2012; Fitzpatrick et al., 2013). To simplify the current work, we employed the default settings (feature class and regularization) in Maxent for each sample size.

We calibrated and evaluated the ENMs using a spatial partition approach using the R package *ENMeval* (Muscarella et al., 2014). The geographic range was divided into four quadrants k=4, with approximately equal number of occurrence records. We then built models using a jackknife approach (k – 1) and evaluated models within the unused partition (Boria et al., 2014; Radosavljevic and Anderson, 2014). Maxent sampled background data for the environmental variables from only the regions corresponding to the quadrants used during calibration (Phillips, 2008). We did this a total of four times for each dataset so that each partition was used for evaluation once; we calculated evaluation statistics (see below) for each quadrant and averaged across the four iterations.

We evaluated the models using threshold-independent and -dependent methods. The threshold-independent measure, Area Under the Curve (AUC) of the Receiver Operating Characteristic plot, is a rank-based measure of discriminatory ability of the model. We calculated these two ways: 1) using all localities for each dataset (according to the chosen sample size) projected

across the entire study region (Full model AUC), 2) calculating AUC for each evaluation quadrant and averaging the four iterations (Mean AUC). The threshold-dependent measures were based on different omission rate thresholding rules: 1) Minimum Training Presence (MTP) and 2) 10% calibration omission rate. Omission rate is the proportion of evaluation localities that are not correctly predicted as present, and measures model overfitting. Overfit models have omission rates higher than theoretical expectations. The MTP sets the threshold at the smallest value of the prediction for any grid cell that contains a calibration locality and has an expected omission rate of zero for evaluation localities and is a more conservative measure of model fitness. Similarly, the 10% calibration omission rate rule sets the threshold at a value that excludes the 10% of calibration localities with lowest prediction and has an expected omission rate of 0.10 (Pearson et al., 2007). We averaged the evaluation statistics across all 25 replicates for each sample size and calculated standard deviation. Finally, for each environmental dataset we ran models and calculated the evaluation statistics using all filtered localities available (the “full” datasets).

To determine variability in geography among models, we measured the pairwise similarity by calculating Schoener's D values using the R package *ENMeval* (Warren et al., 2008; Muscarella et al., 2014). Schoener's D is a pixel-by-pixel comparison between two predictions across the entire study region with scores ranging from 0 (no overlap) to 1 (identical models) (Warren et al., 2008). We calculated pairwise Schoener's D between all 25 replicate models within each sample size to determine variability with each dataset size. Additionally, we calculated the pairwise D values between each sample size and the maximum sample size. This allowed us to determine how similar the models from each sample size were compared with the datasets that contained the most information (the largest sample of occurrence records). We averaged the pairwise D values within and between samples sizes and calculated standard deviations.

We evaluated the distribution of means for each sample size using two variables for each environmental dataset. We extracted the climate data for every occurrence record from the two variables that provided the largest percent contribution for the full dataset maxent model. Then we calculated the mean value for each dataset per sample size and plotted the density of the distribution of means for each sample size. Lastly, we plotted the mean value for the full dataset to determine how the distribution changes as we increase sample size.

1.4 Results

1.4.1 Model Evaluation

For the CCSM3 datasets, as sample size increased, AUC and omission rate decreased (Fig. 1.1). When using the full dataset, the AUC was lower and the omission rate was higher than samples sizes of similar magnitude (Fig. 1.1). Additionally, as sample size increased, similarities in geography increased within datasets and between each dataset and the maximum set (800 occurrence records; Fig. 1.3). Lastly, as we increased the sample size the distribution of means for each sample size narrowed around the mean value of the full dataset (Appendix Fig. 1.8.1 A & B).

For the worldclim datasets, we recovered similar patterns for model performance and geographic predictions as CCSM3. As we increased sample size, Full model AUC decreased slightly; however, Mean AUC increased slightly initially (Fig. 1.2). Again, both measures of omission rates decreased as we increased sample sizes (Fig. 1.2). When using the full dataset, AUC values for both measures were about the same and the omission rates were slightly higher than samples sizes within a similar range (Fig. 1.2). Lastly, as sample size increased, similarities in geography increased within datasets and between each dataset and the maximum set (3,000 occurrence records; Fig. 1.3). Similarly, to the CCSM3 variables, as we increased the sample size the distribution of means for each dataset narrowed around the mean value of the full dataset (Appendix Fig. 1.8.1 C & D).

1.5 Discussion

1.5.1 Comparisons among datasets

Several studies have documented the minimum number of occurrence records needed for ENMs (Stockwell and Peterson, 2002; Papeş and Gaubert, 2007; Wisz et al., 2008; Proosdij et al., 2016), with (Pearson et al., 2007) developing a test of significance when modeling with a small number of localities. However, there are almost no recommendations or guidelines for modeling species with large distributions and many occurrence records. When using many localities, users could be limiting the availability of background localities leading to a hindrance in model performance and predictive accuracy (Stockwell and Peterson, 2002). Additionally, issues with the input occurrence data (sampling bias and georeferencing errors) may be amplified as sample size increases, negatively affecting ENM results (Beck et al., 2014; Boria et al., 2014; Bloom et

al., 2018). Here we show that for a species with many occurrence records, using a smaller dataset can produce very similar models (Fig. 1.1 & 1.2).

For *Peromyscus maniculatus*, we get similar performing models both in geography and with evaluation statistics once a threshold number of localities has been exceeded; while the exact number of optimal localities varies, in general as few as 10% – 20% of the maximum number of localities for each climate dataset seems sufficient (Fig. 1.1–1.3). Further, when using the full spatially filtered dataset for both climate sets, model performance does not increase and in most instances performs worse than similarly sized samples. Interestingly, we generally see a slight decrease in both measures of AUC (model discriminatory ability) as sample size increases (Fig. 1.1 & 1.2), which is most likely due to the increased number of occurrence records making it more difficult for the models to distinguish between presence and background localities (Lobo et al., 2008; VanDerWal et al., 2009). For species with a large number of localities, researchers may want to use a smaller number of georeferenced and spatially-filtered occurrence records when fitting ENMs because, as shown here, performance and predictions robustness does not substantially increase. Finally, Maxent models with larger sample sizes are necessarily more complex, and could potentially be more difficult to transfer across space and time (Boria et al., 2017).

Whereas the maximum number of localities to employ generally goes unaccounted for in ecological niche modeling, the current results support using a smaller number of georeferenced localities (assuming the minimum standards are met). Indeed, our results suggest that subsampling is ideal, since model performance is generally worse for the full dataset with a large number of localities. We recommend that researchers modeling broadly distributed species use multiple occurrence datasets to determine if similar patterns occur within their system. Researchers could also try experimenting with higher numbers of background localities (if possible) to determine the effects on evaluation statistics (VanDerWal et al., 2009). The ideal would be to use a minimum number of occurrence records that can be correctly (and reasonably) georeferenced, one of the often-overlooked aspects of ENMs (Bloom et al., 2018), while retaining the species-specific niche signal.

1.6 References

Aiello-Lammens ME, Boria RA, Radosavljevic A, Vilela B, Anderson RP (2015). spThin: An R package for spatial thinning of species occurrence records for use in ecological niche models. *Ecography* **38**: 541–545.

Anderson RP (2012). Harnessing the world's biodiversity data: promise and peril in ecological niche modeling of species distributions. *Ann N Y Acad Sci* **1260**: 66–80.

Anderson RP, Raza A (2010). The effect of the extent of the study region on GIS models of species geographic distributions and estimates of niche evolution: preliminary tests with montane rodents (genus Nephelomys) in Venezuela. *J Biogeogr* **37**: 1378–1393.

Barve N, Barve V, Jiménez-Valverde A, Lira-Noriega A, Maher SP, Peterson AT, et al. (2011). The crucial role of the accessible area in ecological niche modeling and species distribution modeling. *Ecol Model* **222**: 1810–1819.

Beck J, Böller M, Erhardt A, Schwanghart W (2014). Ecological Informatics Spatial bias in the GBIF database and its effect on modeling species ' geographic distributions. *Ecol Inform* **19**: 10–15.

Bedford NL, Hoekstra HE (2015). Peromyscus mice as a model for studying natural variation. *eLife* **4**: 1–13.

Bloom TDS, Flower A, DeChaine EG (2018). Why georeferencing matters: Introducing a practical protocol to prepare species occurrence records for spatial analysis. *Ecol Evol* **8**: 765–777.

Boria RA, Olson LE, Goodman SM, Anderson RP (2014). Spatial filtering to reduce sampling bias can improve the performance of ecological niche models. *Ecol Model* **275**: 73–77.

Boria RA, Olson LE, Goodman SM, Anderson RP (2017). A single-algorithm ensemble approach to estimating suitability and uncertainty: cross-time projections for four Malagasy tenrecs. *Divers Distrib* **23**: 196–208.

Chamberlain S (2017). rgbif: Interface to the global 'biodiversity'information facility API. R package version 0.9. 5. : https://CRAN.R-project.org/package=rgbif.

Dewey MJ, Dawson WD (2001). Deer Mice : "The Drosophila of North American Mammalogy". *Genesis* **29**: 105–109.
Elith J, Graham CH, Anderson RP, Dudík M, Ferrier S, Guisan A, et al. (2006). Novel methods improve prediction of species' distributions from occurrence data. *Ecography* **29**: 129–151.

Elith J, Phillips SJ, Hastie T, Dudík M, Chee YE, Yates CJ (2011). A statistical explanation of MaxEnt for ecologists. *Divers Distrib* **17**: 43–57.

Fitzpatrick MC, Gotelli NJ, Ellison AM (2013). MaxEnt versus MaxLike : empirical comparisons with ant species distributions. *Ecosphere* **4**: 1–15.

Fourcade Y, Engler JO, Rödder D, Secondi J (2014). Mapping Species Distributions with MAXENT Using a Geographically Biased Sample of Presence Data: A Performance Assessment of Methods for Correcting Sampling Bias. *PloS One* **9**: e97122.

Galante PJ, Alade B, Muscarella R, Jansa SA, Goodman SM, Anderson RP (2018). The challenge of modeling niches and distributions for data-poor species: a comprehensive approach to model complexity. *Ecography* **41**: 726–736.

Hijmans RJ, Cameron SE, Parra JL, Jones PG, Jarvis A (2005a). Very high resolution interpolated climate surfaces for global land areas. *Int J Climatol* **25**: 1965–1978.

Hijmans RJ, Cameron SE, Parra JL, Jones PG, Jarvis A (2005b). Very high resolution interpolated climate surfaces for global land areas. *Int J Climatol* **25**: 1965–1978.

Hijmans RJ, Garrett KA, Huaman Z, Zhang DP, Schreuder M, Bonierbale M (2000). Assessing the Geographic Representativeness of Genebank Collections: the Case of Bolivian Wild Potatoes. *Conserv Biol* **14**: 1755–1765.

Kadmon R, Farber O, Danin A (2004). Effect of roadside bias on the accuracy of predictive maps produced by bioclimatic models. *Ecol Appl* **14**: 401–413.

Kramer-schadt S, Lindenborn J, Reinfelder V, Stillfried M, Schr B, Heckmann I, et al. (2013). The importance of correcting for sampling bias in MaxEnt species distribution models. *Divers Distrib* **19**: 1366–1379.

Liu Z, Otto-Bliesner BL, He F, Brady EC, Tomas R, Clark PU, et al. (2009). Transient Simulation of Last Deglaciation with a New Mechanism for Bølling-Allerød Warming. *Science* **325**: 310–314.

Lobo JM, Jiménez-Valverde A, Real R (2008). AUC: a misleading measure of the performance of predictive distribution models. *Glob Ecol Biogeogr* **17**: 145–151.
Lorenz DJ, Nieto-lugilde D, Blois JL, Fitzpatrick MC, Williams JW (2016). Data Descriptor : Downscaled and debiased climate simulations for North America from 21, 000 years ago to 2100 AD. *Sci Data* **3**: 160048.

Maguire KC, Nieto-Lugilde D, Blois JL, Fitzpatrick MC, Williams JW, Ferrier S, et al. (2016). Controlled comparison of species- and community-level models across novel climates and communities. *Proc R Philos Soc B* **283**: 20152817.

Merow C, Smith MJ, Silander JA (2013). A practical guide to MaxEnt for modeling species' distributions: what it does, and why inputs and settings matter. *Ecography* **36**: 1058–1069.

Muscarella R, Galante PJ, Soley-Guardia M, Boria RA, Kass JM, Uriarte M, et al. (2014). ENMeval: An R package for conducting spatially independent evaluations and estimating optimal model complexity for Maxent ecological niche models. *Methods Ecol Evol* **5**: 1198–1205.

Papeş M, Gaubert P (2007). Modelling ecological niches from low numbers of occurrences : assessment of the conservation status of poorly known viverrids (Mammalia , Carnivora) across two continents. *Divers Distrib* **13**: 890–902.

Patterson BD, Ceballos G, Sechrest W, Tognelli MF, Brooks T, Luna L, et al. (2007). Digital distribution maps of the mammals of the Western Hemisphere, version 3.0. *CD-ROM Compil NatureServe*.

Pearson RG, Raxworthy CJ, Nakamura M, Peterson AT (2007). Predicting species distributions from small numbers of occurrence records: a test case using cryptic geckos in Madagascar. *J Biogeogr* **34**: 102–117.

Peterson TA, Soberon J, Pearson RG, Anderson RP, Martínez-Meyer E, Nakamura M, et al. (2011). *Ecological Niches and Geographic Distributions*. Princeton University Press.

Phillips SJ (2008). Transferability, sample selection bias and background data in presence-only modelling: A response to Peterson et al. (2007). *Ecography* **31**: 272–278.

Phillips SJ, Anderson RP, Schapire RE (2006). Maximum entropy modeling of species geographic distributions. *Ecol Model* **190**: 231–259.

Phillips SJ, Dudík M (2008). Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. *Ecography* **31**: 161–175.

Proosdij ASJ Van, Sosef MSM, Wieringa JJ, Raes N (2016). Minimum required number of specimen records to develop accurate species distribution models. *Ecography* **39**: 542–552.

Radosavljevic A, Anderson RP (2014). Making better Maxent models of species distributions: complexity, overfitting and evaluation (M Araújo, Ed.). *J Biogeogr* **41**: 629–643.

Reddy S, Dávalos LM (2003). Special Paper: Geographical Sampling Bias and Its Implications for Conservation Priorities in. *J Biogeogr* **30**: 1719–1727.

Royle JA, Chandler RB, Yackulic C, Nichols JD (2012). Likelihood analysis of species occurrence probability from presence-only data for modelling species distributions. *Methods Ecol Evol* **3**: 545–554.

Shcheglovitova M, Anderson RP (2013). Estimating optimal complexity for ecological niche models: A jackknife approach for species with small sample sizes. *Ecol Model* **269**: 9–17.

Shorter KR, Crossland JP, Webb D, Szalai G, Felder MR, Vrana PB (2012). Peromyscus as a Mammalian Epigenetic Model. *Genet Res Int* **2012**: 1–11.

Stockwell DRB, Peterson AT (2002). Effects of sample size on accuracy of species distribution models. *Ecol Model* **148**: 1–13.

VanDerWal J, Shoo LP, Graham C, Williams SE (2009). Selecting pseudo-absence data for presence-only distribution modeling: How far should you stray from what you know? *Ecol Model* **220**: 589–594.
Warren DL, Glor RE, Turelli M (2008). Environmental niche equivalency versus conservatism: Quantitative approaches to niche evolution. *Evolution* **62**: 2868–2883.

Wisz MS, Hijmans RJ, Li J, Peterson AT, Graham CH, Guisan A (2008). Effects of sample size on the performance of species distribution models. *Divers Distrib* **14**: 763–773.

1.7 Figures

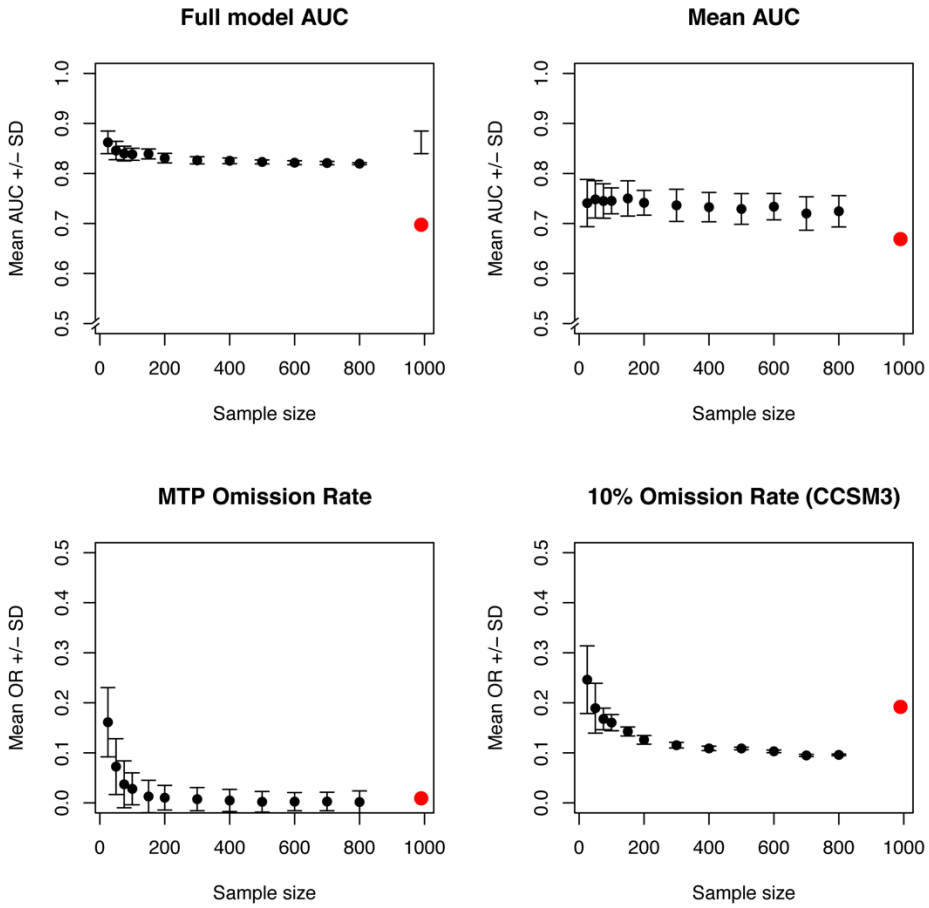


Figure 1.1 Evaluation statistics for *Peromyscus maniculatus* in North America using the CCSM3 climate variables for: (top left) Full model AUC; (top right) Mean AUC; (bottom left) Minimum training presence (MTP) omission rate; and, (bottom right) 10% omission rate. Values represent the mean value (± the standard deviation [SD]) across 25 replicates for each sample size. The red point is the full dataset (990 localities) evaluation statistics, and it had the lowest AUC values and an increased omission rate (relative to the most similar sample size, N=800).

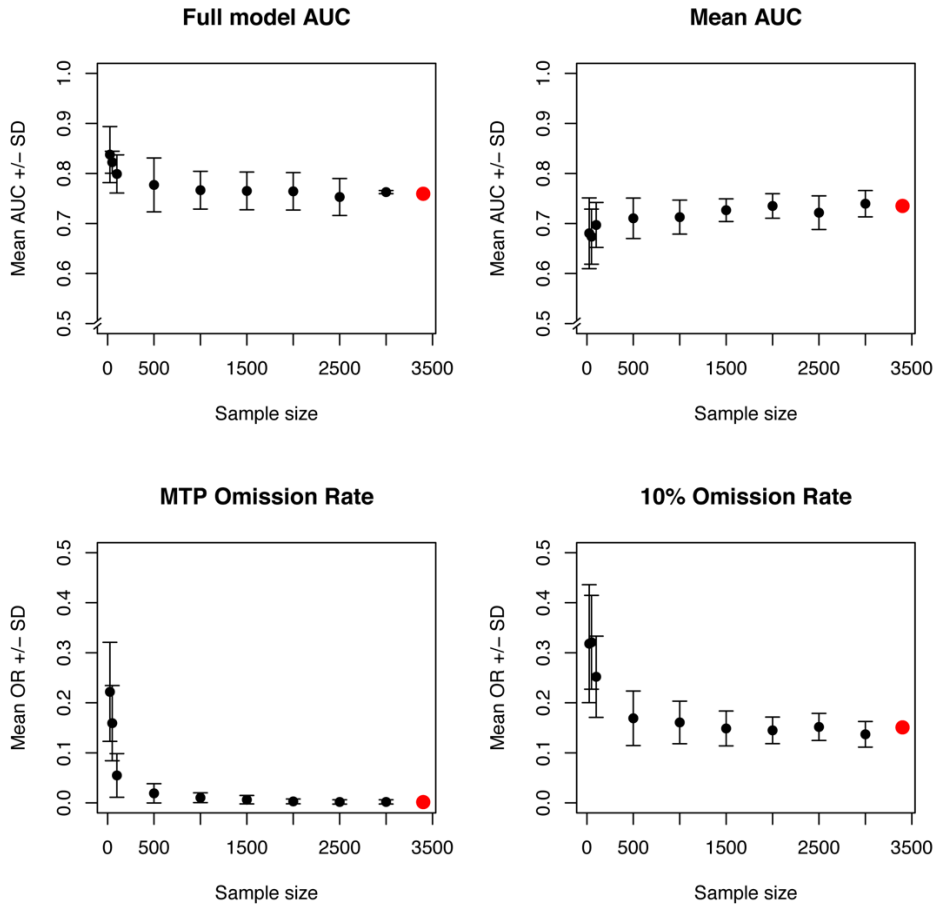


Figure 1.2 Evaluation statistics for *Peromyscus maniculatus* in North America using the worldclim climate variables for: (top left) Full model AUC; (top right) Mean AUC; (bottom left) Minimum training presence (MTP) omission rate; and, (bottom right) 10% omission rate. Values represent the mean value (± the standard deviation [SD]) across 25 replicates for each sample size. The red points represent the full dataset (3,334 localities) evaluation statistics, and this dataset had similar AUC values and a slightly increased omission rate (relative to the most similar sample size, N=3000).

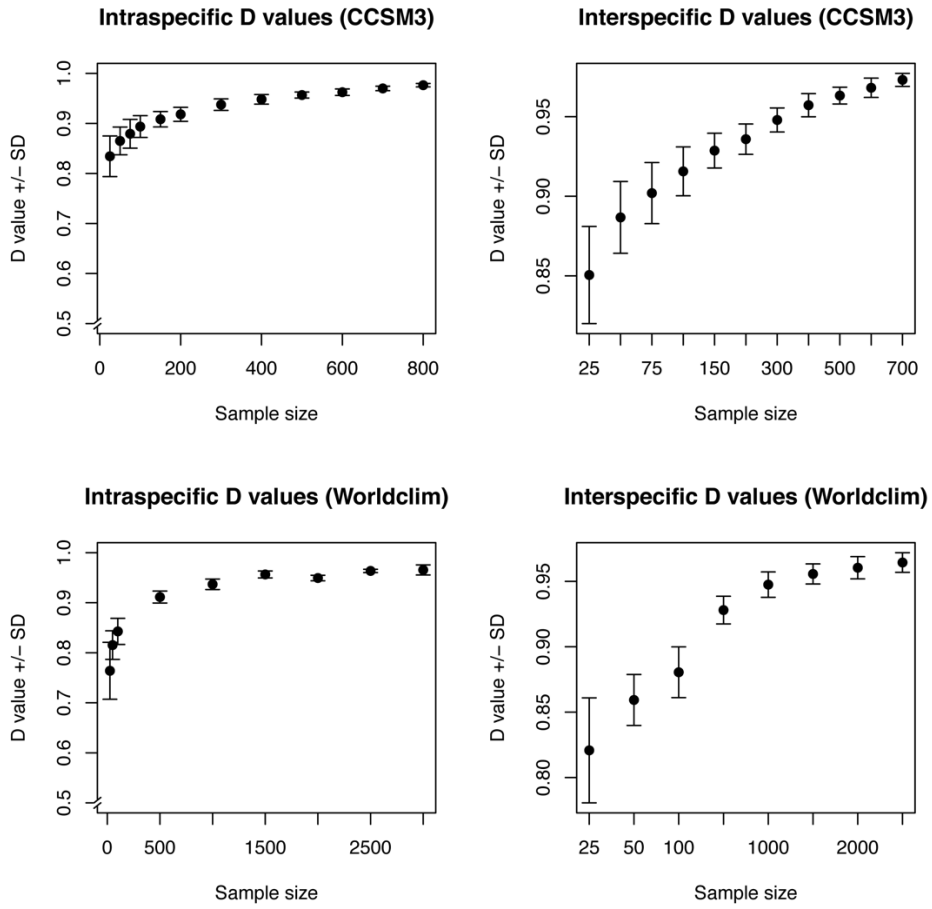
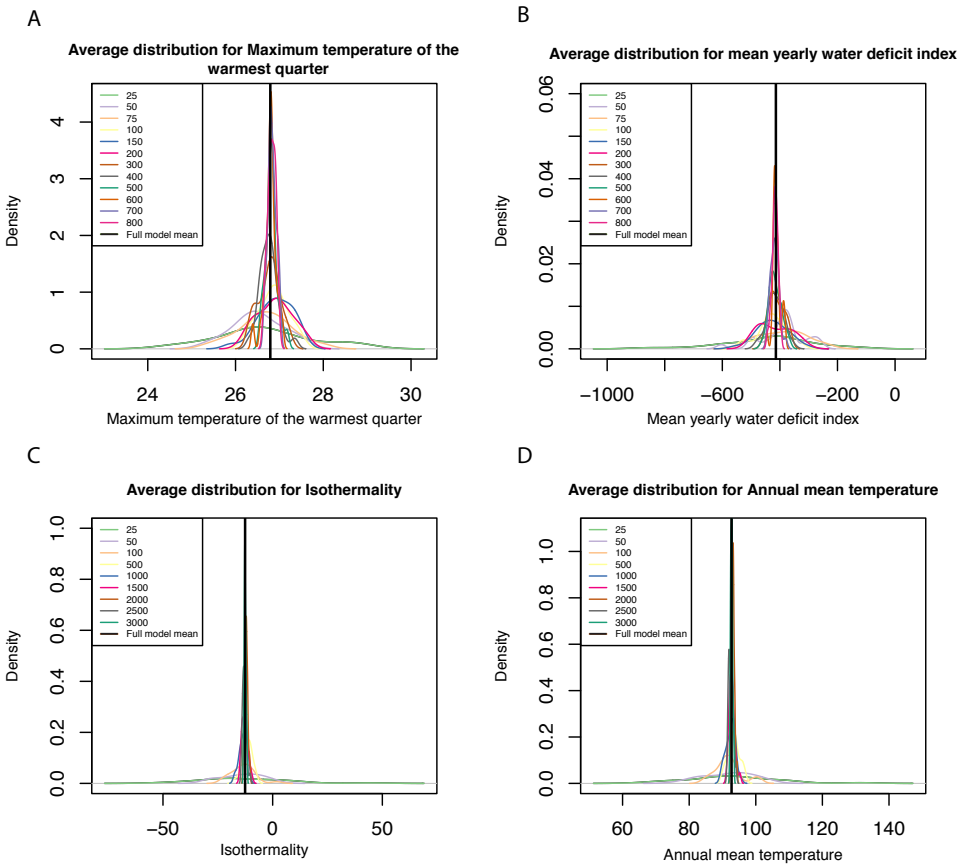


Figure 1.3 Schoener's D statistic for *Peromyscus maniculatus* in North America for both sets of climate variables. This similarity index ranges from 0 (completely different) – 1 (identical). We compared the predictions in geographic space, both within (intraspecific) and between each sample size and the largest sampled dataset (interspecific).

1.8 Appendix



Appendix Figure 1.8.1: Density distribution of means for each sample size of *Peromyscus maniculatus* in North America using two variables each from CCSM3 (A; Maximum temperature of the warmest quarter, B; Mean yearly water deficit index) and worldclim (C; Isothermality, D; Annual mean temperature). The mean value for the full dataset is shown with black line. Note that as we increased the sample size the distribution of means for each sample size narrowed around the mean value of the full dataset for all variables.

Chapter 2: Genome-wide genetic variation coupled with demographic and ecological niche modeling of the dusky-footed woodrat (*Neotoma fuscipes*) reveal patterns of deep divergence and widespread Holocene expansion across northern California

2.1 Abstract

Understanding how species have responded to past climate change may help refine projections of how species and biotic communities will respond to future change. Here, we integrate estimates of genome-wide genetic variation with demographic and niche modeling to investigate the historical biogeography of an important ecological engineer: the dusky-footed woodrat, *Neotoma fuscipes*. We use RADseq to generate a genome-wide dataset for 71 individuals from across the geographic distribution of the species in California. We estimate population structure using several model-based methods and infer the demographic history of regional populations using a site frequency spectrum-based approach. Additionally, we use ecological niche modeling to infer current and past (Last Glacial Maximum) environmental suitability across the species' distribution. Finally, we estimate the directionality and possible spatial origins of regional population expansions. Our analyses indicate this species is subdivided into three regionally distinct populations, with the deepest divergence occurring ~1.7 million years ago across the modern-day San Francisco-Bay Delta region; a common biogeographic barrier for the flora and fauna of California. Our models of environmental suitability through time coincide with our estimates of population expansion, with relative long-term stability in the southern portion of the range, and more recent expansion into the northern end of the range. Our study illustrates how the integration of genome-wide data with spatial and demographic modeling can reveal the timing and spatial extent of historic events that determine patterns of biotic diversity and may help predict biotic response to future change.

2.2 Introduction

The glacial-interglacial cycles of the Quaternary have long been considered primary drivers of contemporary patterns of biodiversity (Hewitt 2000; Peterson 2009). The dynamic climatic oscillations of this time period were characterized by cooler glacials followed by warmer interglacials, and very short transitions from cool to warm periods (Dansgaard et al. 1993). For example, in North America during the Last Glacial Maximum (LGM) 22,000 to 18,000 years before present, large ice sheets covered most of the northern portion of the continent, and climatic conditions were generally cooler and drier than at present (Hopkins et al. 2013), except in the southwest where it was wetter (Clark et al. 2012). These global climatic oscillations and associated landscape changes likely influenced the current distribution of biodiversity through repeated

reduction, isolation, and subsequent expansion of species across large geographic areas (Hewitt 2004; Svenning and Skov 2004).

Paleoecological, distribution modeling, and genetic evidence support the dynamic changes that species experienced with past climate change. During the Quaternary, many plant and animal species retreated into areas of glacial climate refugia (termed refugia hereafter)—local areas with the appropriate environmental conditions to allow species to persist—and expanded their ranges when environmental conditions were favorable once again (Jackson and Overpeck 2000; Hewitt 2004; Nogués-Bravo et al. 2008). This pattern of range shift is seen in North America during the most recent glacial-interglacial transition from the LGM to the present (Graham et al. 1996; Jackson and Overpeck 2000; Waltari et al. 2007). In addition to range shifts, many species in North America experienced a reduction in habitat during the glacial periods (e.g., Knowles and Massatti 2017; Reid et al. 2018) due to the spread of continental ice sheets during the LGM and the reduction or fragmentation of species preferred habitat (Hewitt 2004; Nogués-Bravo et al. 2008). These dramatic changes to species ranges can leave distinct patterns on the present-day genetic structure of a species (Waltari et al. 2007; Knowles and Massatti 2017; Reid et al. 2019). For example, populations that recolonized areas that were previously covered by glaciers or otherwise inhospitable should show signs of recent rapid genetic expansion (Lessa et al. 2003; Burbrink et al. 2016).

California offers a unique opportunity to study how climate oscillations influence patterns of diversity and differentiation in detail across a wide breadth of ecologically distinct taxa. The California Floristic Province is a recognized diversity hotspot due to the combination of high species endemism and significant conservation threat (Myers et al. 2000). The high topographic complexity of California leads to significant environmental gradients in temperature, precipitation, and many other variables, across latitudes, meridians, and elevations that contribute to the generation and maintenance of genetic and biotic diversity (Parisi 2003; Davis et al. 2008). High topographic complexity has magnified the ecological dynamics and climatic oscillations of the past 3-5 million years (Badgley et al. 2017) leading to repeated broad-scale contractions and expansions of species ranges resulting, in many cases, in concordant patterns of diversity and differentiation across multiple species and taxonomic groups (LaPointe and Rissler 2005; Rissler et al. 2006). In particular, concordant phylogeographic breaks exist in many plant and animal taxa across the San Francisco Bay-Delta, the Monterey Bay, between the northern Sierra Nevada and southern Cascade Mountains, along the western slope of the Sierra Nevada, and across the Tehachapi-Transverse Ranges (reviewed in Schierenbeck 2014). These concordant phylogeographic patterns within California further support the role of repeated climatic oscillations across the geographic landscape as drivers of differentiation and diversity across species (Lessa et al. 2003; Grivet et al. 2006).

Species that should be of particular interest for reconstructing phylogeographic history are those that have numerous ecological connections to their communities. Such ecological connections include predator-prey and host-pathogen relationships, as well as species that are environmental engineers that determine the presence of other community members (i.e. 'strong interactors'; Soulé et al. 2005). One such species in California is the dusky-footed woodrat (*Neotoma fuscipes*; (Goldman 1910; Hooper 1938; Matocq 2002a,b)). This taxon is well known for the large stick houses (commonly ~1 meter in height) that individuals build (Carraway and Verts 1991). These structures provide a stable microclimate in comparison to the outside environment (Brown 1968) not only for the woodrat that occupies the house, but also for the many other small vertebrate and invertebrate species that occupy these structures (Whitford and Steinberger 2010). In addition to their role as environmental engineers, woodrats are the most common prey item of the northern spotted owl (Sakai and Noon 1993; Thome et al. 1999), and one of the primary mammalian hosts of *Borrelia burdorferi*, the causative agent of Lyme disease (Lane and Brown 1991) and other tick-borne pathogens (Foley et al. 2016). Beyond their central role in the communities they occupy, species of the genus *Neotoma* are also well known for the rich paleorecord they have left behind. The contents of their complex, multi-chambered houses (e.g. plant fragments, scat, raptor pellets) often become fossilized and have provided a uniquely detailed view of ecological and environmental change through time in western North America (Jackson et al. 2005). These "middens" can persist for thousands of years and have been used to document change in body size (Smith et al. 1995) and community composition across tens of thousands of years (Blois et al. 2010). By understanding the evolutionary history and historical demography of woodrats, we can begin to shed light on species-level dynamics that may have had community-level consequences, and that can be uniquely tested given the existence of the high resolution paleorecord generated by these animals.

The range of *Neotoma fuscipes* (*sensu* Matocq 2002a) extends from western Oregon through the Coast Ranges of northern California, the San Francisco Bay Area and south through the Inner Coast Range, and throughout the northern Sierra Nevada foothills, Modoc Plateau and Cascade Range (Figure 1). Using mitochondrial data, Matocq (2002b) showed that *N. fuscipes* comprises two lineages separated by the San Francisco Bay, and suggested that *N. fuscipes* had recently expanded into northern California, given relatively low mtDNA diversity across that region. Here, we revisit many of these original collections and substantially augment them with newly available samples, especially from northeastern California (courtesy of the Museum of Vertebrate Zoology and the Grinnell Resurvey Project). In addition to augmented spatial coverage, we also augment our genomic sampling by using high-throughput sequencing across the nuclear genome. We integrate these genetic data with both niche and demographic modeling to provide a temporally comprehensive and spatially detailed view of the demographic and biogeographic history of this

species across its distribution. Specifically, we ask 1) Does *N. fuscipes* exhibit significant subdivision within its distribution that coincides with known phylogeographic breaks in other species? 2) Did *N. fuscipes* experience range contractions/expansions in the recent past despite occurring in largely unglaciated areas, and if so, were there multiple refugial areas? 3) What is the timing of major events in the history of this species including divergence, contraction, and expansion? By integrating genome-wide estimates of genetic variation with niche and demographic modeling, we provide a temporally and spatially detailed view of the history of this taxon across its range.

2.3 Materials and Methods

2.3.1 ddRAD library preparation and sequencing

DNA was extracted from 71 individual *Neotoma fuscipes* and 8 *N. macrotis* (to provide outgroup rooting; Figure 2.1; Appendix Table 2.9.1) using the Qiagen DNeasy Blood and Tissue kit (Qiagen, Inc.). All samples used in this study are housed in the Museum of Vertebrate Zoology, University of California, Berkeley. We generated multilocus SNP datasets following the double-digest RADSeq protocol (Peterson et al. 2012). Briefly, we used 500 ng of genomic DNA per sample and digested them overnight with the common-cutting restriction enzyme *EcoR1* and the rare-cutting enzyme *MSP1*. We cleaned the digested samples with Ampure (Invitrogen) beads and eluted the samples with water. These products were tagged with a unique five base-pair (bp) barcode adaptor. We pooled and purified the tagged samples (8-10 individuals per pool) and used a pippin prep to size select for fragments between 300-500 bp long. Each library was amplified via PCR with a second index primer to differentiate individuals. Finally, the libraries were sequenced at the U.C. Berkeley QB3 Vincent J. Coates Genomics Sequencing Laboratory on two lanes of an Illumina HiSeq 2500 to produce 100 bp single end reads.

The Illumina HiSeq raw reads were cleaned using the Stacks pipeline (v. 1.21; (Catchen et al. 2013). The *process_radtags* script (from Stacks) was used to filter out low quality reads and demultiplex the reads according to their unique barcodes. The demultiplexed reads were imported into *ipyrad* (v. 0.7.17; https://github.com/dereneaton/ipyrad/blob/master/docs/index.rst) pipeline to cluster first within and then among individuals and *de novo* assemble the reads. We ran several iterations of *ipyrad* because with next generation sequencing, there can be large amounts of variation between loci due to missing data (Huang and Knowles 2016). We ran *ipyrad* with all samples (both *N. fuscipes* and *N. macrotis*) to include the outgroup for phylogenetic analyses. We then re-clustered the *N. fuscipes* samples alone to be used in the population structure and demographic analyses focused solely on this taxon. We used two different

thresholds for the minimum number of individuals needed to retain a locus to determine the effects of missing data: 50% and 85%, referred to hereafter as the 50% and 85% datasets, respectively (see Appendix S1 for program inputs). This first filtering step controls for the amount of missing data across individuals. Following read clustering, we used the --missing-indv tool in vcftools (v. 0.1.17; Danecek et al. 2011) to calculate the amount of missing data within each individual and removed three samples that had more than 70% missing data. Once we explored the effect of missing data and determined it has little influence on inference of population structure, we ran the range expansion and genetic diversity analyses using a 67% threshold to strike a balance between missing data from natural sequencing variation and the different number of individuals per population. We subsequently re-ran *ipyrad* on each of the populations identified in the population structure analyses using a 67% threshold.

2.3.2 Phylogenetic analyses

We inferred phylogenetic relationships among sequenced individuals with *N. macrotis* as the outgroup using RAxML (v. 8.2.12; Stamatakis 2014) a software program used for large datasets that employs a maximum likelihood (ML) algorithm. Note, this analysis is not appropriate for many phylogenetic questions involving intraspecific SNP data; however, we used this method to understand how individuals grouped together, rather than to infer their specific phylogenetic relationships (following Harrington et al. 2017). To understand the effects of missing data, we ran the RAxML analysis on both the 50% and 85% datasets. Further, we used the full concatenated sequence (invariant and variant sites) to avoid acquisition bias (Leaché et al. 2015). We used rapid bootstrapping (with an automatic stop under the autoMRE criterion), and searched for the best tree under the GTRGAMMA model with final optimization of trees using GTR+Γ. We used Splitstree (V 4.14.8; Huson and Bryant 2006) with only the *N. fuscipes* individuals to generate a phylogenetic network. We calculated distances between all individuals using the Jukes Cantor model (Jukes and Cantor 1969) and used the neighbornet method to build the network. Finally, we used these pairwise distance estimates to compare genetic variation within and among clades.

2.3.3 Population structure and isolation by distance

To investigate population structure, we used two model-based and one non-model-based approaches. All methods attempt to determine the optimal number of populations present within the dataset. We used only one SNP per locus for each population structure analysis. We ran STRUCTURE, a Bayesian-clustering algorithm that identifies population structure (Pritchard et al. 2000), using 50,000 burn-in generations, 1,000,000 generations, K = 1-5 populations, using the admixture and allele frequencies correlated models, and for ten iterations. We used the Evanno method (Evanno et al. 2005) in STRUCTURE

HARVESTER (V. 0.6.94; Earl and vonHoldt 2012) to determine the optimal number of populations. Because STRUCTURE HARVESTER may not accurately infer the correct number of groups, typically inferring the upper most level of genetic structure (often *K* = 2; Janes et al. 2017), we used a hierarchical approach, and performed a second set of STRUCTURE analyses on each of the initial populations identified (Pritchard et al. 2000). We used 50,000 burn-in generations, 200,000 generations, *K* = 1-5 populations, using the admixture and allele frequencies correlated models, and for five iterations. We also used ADMIXTURE (v. 1.3), a maximum likelihood approach to determine the optimal number of populations (Alexander et al. 2009). We used cross-validation error values across different values of *K* to determine the optimal number of populations. Finally, we used the non-model based method DAPC (Discriminate analysis of principal components) in the adegenet package (Jombart 2008; Jombart and Ahmed 2011) in R (v. 3.5.1; R core team 2106) to infer structure. For the DAPC analysis, we used Bayesian information criterion to determine the optimal number of populations. Once optimal populations were identified from each method, we estimated population differentiation by calculating pairwise F_{ST} using vcftools (v. 0.1.17; Danecek et al. 2011). To understand the effects of missing data, we ran all of these analyses on the 50% and 85% datasets.

We assessed whether there was a significant signal of isolation by distance (IBD) within the entire species and each regional group identified by STRUCTURE. A pattern of IBD is expected to arise at a regional scale once populations reach an equilibrium between dispersal among populations and genetic drift within populations (as shown in Hutchison and Templeton, 1999 [Case1]; van Strien et al. 2015). This specific pattern of IBD, showing increased genetic differences with an increase in geographic distance across the whole range, is generally expected to arise in regions that have been stably occupied (Hutchison and Templeton 1999). We plotted the relationship between genetic (F_{ST}) and geographic distance among populations within regional genetic groups and determined the significance of these relationships using a Mantel test and 10,000 permutations in the adegenet package (Jombart 2008; Jombart and Ahmed 2011).

2.3.4 Demographic model testing

We used a coalescent-based approach to model the demographic history of *Neotoma fuscipes* in a temporal framework, specifically with respect to isolation and migration. We selected and parameterized the best-fit demographic model using fastsimcoal2 (FSC2; v2.6.0.3; Excoffier et al. 2013), which estimates demography from the site frequency spectrum (SFS). We followed guidance from Hotaling et al. (2018) in setting up several of our input files. FSC2 allows users to generate hypotheses with differing levels of complexity and uses simulations to estimate the likelihood of competing hypotheses with different parameters. Individuals were assigned using a 3-population model based on the population

structure analyses (see above), with admixed individuals assigned to the majority population. We generated the observed joint SFS using code developed by Isaac Overcast (https://github.com/isaacovercast/easySFS).

We tested 17 different demographic scenarios (Fig. 2). Briefly, they describe variations of a 3-population model that have two divergence events with every possible topology and vary in the amounts of migration and timing in divergence between populations, while also estimating the effective population size (N_E) of every regional grouping. Additionally, we modeled trifurcation scenarios with different levels of historical and recent migration events. For *Neotoma fuscipes*, we used a nuclear mutation rate of 2.5×10^{-8} mutations/site/generation based on estimates from human loci (Nachman and Crowell 2000) and used one-year estimated generation time. For each model we ran 75 replicate FSC2 analyses, each using 250,000 coalescent simulations. We selected the best fit model by calculating AIC and △AIC scores to account for number of parameters following the guideline of Excoffier et al. (2013). Once the best model was obtained, we re-estimated parameter values by simulating 100 SFS from the maxL.par file to obtain a mean parameter estimate and 95% confidence intervals. We accounted for *N. fuscipes* being diploid by dividing the FSC2 estimates in half. For migration rates, we multiplied the values by population size, and then divided by two to account for ploidy.

2.3.5 Ecological niche models

We generated ecological niche models (EMNs) to infer past areas of refugia. For ENMs, we limited the scope of the analyses to only individuals for which we had genetic data (Supplementary Table 1). Because of the deep divergence between the two northern and the southern populations we modeled each separately (combined north, south). To remove spatial biases, we spatially filtered the dataset to ensure no two localities were within 5 km of one another (Boria et al. 2014) using the R package spThin (Aiello-Lammens et al. 2015). For environmental data, we used Community Climate Simulation Model 3 (CCSM3; Liu et al. 2009). CCSM3 variables are downscaled to 50 x 50 km degree grid cells for North America from 21,000 years ago to present day at 500-year intervals (Lorenz et al. 2016). To approximate modeling assumptions regarding dispersal and biotic interactions more closely, we delimited a custom study region, specifically by drawing a minimum convex polygon around the localities and adding a 3.0° buffer (Anderson and Raza 2010; Barve et al. 2011). We used a machine learning algorithm, maxent (V3.4.1; Phillips et al. 2017; Phillips et al. 2006) to infer the ENMs. We calibrated and evaluated the models using a jackknife approach (Pearson et al. 2007) in the ENMeval package in R (Muscarella et al. 2014). To select species-specific model settings approximating optimal levels of complexity, we tuned model settings by varying different

combinations of feature class and regularization multiplier (RM; Shcheglovitova and Anderson 2013). To identify the optimal parameter settings, we evaluated model performance using sequential criteria, lowest average omission rate and secondarily on the highest average AUC value (minimizing overfitting and then maximizing discriminatory ability; Shcheglovitova and Anderson 2013; Muscarella et al. 2014). We used the optimal settings to project the ENM for each regional population into current climatic conditions and the LGM.

2.3.6 Range Expansion

We inferred origins of recent range expansion within the southern and combined northern populations using the rangeExpansion package in R (Peter and Slatkin 2013). Briefly, this method detects range expansion and gives an estimated location of the origin of expansion. It does so by calculating a directionality index (ψ), using the genetic data and spatial coordinates, based on allele frequency clines found between multiple populations (Peter and Slatkin 2013). Populations at the expanding edge will tend to have lower genetic diversity because of serial founder events and higher fixation rates due to small effective population sizes (Peter and Slatkin 2015). For each population, we only used loci that were present in at least two thirds of the individuals.

2.3.7 Genetic Diversity

We estimated the population genetic diversity parameter, θ, using ThetaMeta (Adams et al. 2018). This program uses an infinite sites likelihood model to estimate a posterior probability distribution of θ for a given dataset. To understand the population dynamics further, we used the population structure results and calculated θ for each of the three identified populations. For each population, we only used loci that were present in at least two thirds of the individuals. We used the ThetaMater.M1 simulation to generate a posterior probability distribution of θ with a burnin of 100,000 and a MCMC run of one million generations.

2. 4 Results

2.4.1 ddRAD data

We obtained 216,373,426 raw single end reads from the Illumina HiSeq run across all individuals. There were 140,817 loci present before ipyrad quality filtering. After filtering, there were 15,334 loci (mean across individuals = 13,576) for the 85% missing-data threshold and 57,860 loci (mean = 43,670) for the 50% missing-data threshold. For the final 50% and 85% datasets, there were 34,434 (85%) and 122,636 (50%) total SNPs, respectively.

2.4.2 Phylogenetic analyses

We obtained very similar tree topologies with RAxML with both the 50% and 85% datasets (Fig 2.1; Appendix Figure 2.9.1A). Essentially, we recovered one monophyletic group of individuals south of the San Francisco Bay-Delta region, and one large monophyletic group of individuals north of the Bay-Delta region (Fig. 2.1). The northern group is further split into two monophyletic groups, one that comprises the majority of the northern range (North A hereafter), and a second that is primarily restricted to the western slopes of the Sierra Nevada (North B hereafter). The Splitstree results were very similar, clearly separating the northern and southern populations (Appendix Figure 2.9.2), as well as separating the North A and North B populations.

Inter-individual genetic distances were higher among individuals in the southern clade than in either the North A or North B clades. On average, southern individuals were differentiated from one another by a genetic distance of 0.073 (min.-max.: 0.044 – 0.097), North A had average pairwise divergence of 0.04 (0.015 – 0.054) and North B had 0.033 (0.013 – 0.042; Appendix Table 2.9.2). These distinctions are visually evident (Fig. 2.1) in the deeper genetic structure of the southern clade in contrast to the shallow, minimal pairwise divergences that characterize both northern clades.

2.4.3 Population structure and isolation by distance

STRUCTURE initially gave a $K = 2$ model as the best fit for the data; however, when we ran STRUCTURE on the two subpopulations, the northern population was split further into two distinct populations (Fig. 2.1) and for the southern population the optimal number of subpopulations was $K = 4$, though the number of individuals sampled was low in each subpopulation (Appendix Figure 2.9.3). Although $K = 2$ was the optimal model in STRUCTURE, $K = 3$ inferred the same northern split of samples we found by running the northern region alone (Fig. 2.1). ADMIXTURE indicated $K = 3$ as the optimal number of populations based on cross-validation (Fig. 2.1; Appendix Table 2.9.3 & Appendix Table 2.9.4) regardless of the amount of missing data included. DAPC analysis identified $K = 3$ as the optimal number of populations for both datasets (Appendix Table 2.9.4, Appendix Figure 2.9.1C). Both ADMIXTURE and STRUCTURE identified some admixture between the two Northern populations. All methods assigned the same individuals to the same population across analyses (and datasets), and ADMIXTURE and STRUCTURE identified the same individuals in the northern group as being admixed. Pairwise F_{ST} showed strong differentiation between the southern and northern groups (ave. $F_{ST} = 0.36$) and moderate differentiation between the North A and North B groups ($F_{ST} = 0.08$; Appendix Figure 2.9.1D; Appendix Table 2.9.5). We adjusted downstream analyses to reflect the population structure results, generally relying on $K = 3$ except for the niche modeling and range expansion analyses.

The Mantel test showed significant isolation by distance for the overall dataset (r= 0.605, p<0.001). However, we did not recover a distinct cline, indicating lack of a clear IBD signature (Appendix Figure 2.9.5; Appendix Table 2.9.6). Each of the regional groups showed a significant relationship between genetic and geographic distance, all following a similar pattern of increasing genetic differentiation with increased geographic distance (Case 1 of Hutchison and Templeton (1999); Fig. 2.3). The southern region showed a pronounced pattern of IBD (r= 0.731, p < 0.001) and the North A group had a slight pattern of IBD (r= 0.537, p < 0.001). The North B region showed an IBD pattern (r= 0.544, p < 0.001) though the samples from this region occurred over a very small spatial scale.

2.4.4 Demographic modeling

Using FSC2, the model that best fit the data according to AIC was scenario 11 (Table 2.1; Appendix Table 2.9.7; Fig. 2.4). This consisted of the two northern populations coalescing most recently, with an older coalescent event between the north and south clusters and both historical and recent admixture (Fig. 2.4; Appendix Table 2.9.8; see Appendix Table 2.9.9 for confidence intervals). No other scenario had a reasonably good fit (Table 2.1; Appendix Table 2.9.7).

Scenario 11 estimated the two northern populations coalesced about 76,000 ka (73,063 – 79,644) and the ancestral northern and southern lineages coalesced approximately 1.72 million years ago (1.71 – 1.73; Fig. 4). The southern population has the largest effective population size (over 800,000 individuals), North A population has the second largest population size (over 300,000), and North B has the smallest effective population size (approximately 4,000). The migration rates between most of the populations (historical and recent) were all less than one individual per year (Fig. 2.4).

2.4.5 Ecological niche models

The optimal Maxent settings for the northern population was the Hinge feature with a RM = 1.0 (Appendix Table 2.9.10). This model had an AUC of 0.83 and an omission rate of 0.045. The optimal model for the southern population used the Linear and Quadratic features and a RM = 1.5 and had an AUC of 0.88 and an omission rate of 0.091. ENMs for the northern population inferred suitable environment present throughout the entire known range of *N. fuscipes* for current conditions, and a severe reduction of habitat at the LGM relative to the contemporary range (Fig. 2.5). For the southern population, ENMs inferred suitable conditions in the southern part of the *N. fuscipes* range for contemporary climatic conditions, matching the current distribution of the population. The

distribution of the southern population was inferred to occur even further south during the LGM (Fig. 2.5).

2.4.6 Range expansion

The range expansion analyses indicated that both population expansion origins were close to the middle of the contemporary *Neotoma fuscipes* range in California, near the San Francisco Bay-Delta area (Fig. 2.5; Appendix Figure 2.9.6). The north origin of expansion located at a slightly higher latitude and closer to the Sierra Nevada ($p < 0.001$), while the southern origin was located closer to coastal California (not significant; $p > 0.1$), just east of the San Francisco Bay.

2.4.7 Genetic Diversity

Results from ThetaMater showed the southern population had the highest median posterior θ, indicating it holds the highest levels of genetic diversity (Fig. 2.6), followed by North A and then North B.

2.5 Discussion

Quaternary climate oscillations across California's landscape had dramatic effects on patterns of biological and genetic diversity (Rissler et al. 2006). Here, we show that the dusky-footed woodrat responded to past climatic shifts through widespread contraction and expansion of its range, which led to lineage divergence and a spatial distribution of genetic variation that remains evident today. By integrating genomic data with demographic and ecological niche modeling, we were able to reconstruct the dynamic history of this species across its range.

2.5.1 Deep divergence and the distribution of suitable environmental conditions

Matocq (2002b) proposed that *Neotoma fuscipes*, and its sister species *Neotoma macrotis* diverged from one another approximately 2 million years ago in the foothills of the central Sierra Nevada, likely between Auburn and Placerville, the current distributional limits of each species (Matocq and Murphy 2007). Under this scenario, within a short window of time, early *N. fuscipes* would have expanded to the north and west around a partly inundated northern Central Valley and experienced a major divergence event leading to the northern and southern *N. fuscipes* groups (referred to as the northern and west central groups by Matocq 2002b). Our estimates of this early divergence within *N. fuscipes* using genome-wide SNPs dates to approximately 1.7 million years ago, largely consistent with the original mtDNA-based estimate of 1.8 mya (Matocq 2002b). This genetic divergence is mirrored by morphological differentiation within *N.*

fuscipes (Hooper 1940), supporting not only the depth of this divergence but also its potential functional significance.

Fossil evidence of the history of *N. fuscipes* is sparse but not inconsistent with the genetic data. *Neotoma* fossils are widespread throughout the more southerly parts of California and the west by the Pliocene (Paleobiology DB search for "Neotoma" on 17 April 2020, paleobiodb.org), perhaps mirroring the more stable habitat available in the southern parts of the state. The earliest *Neotoma* fossils in the central and northern parts of the state are not found until the Irvingtonian 1.8 - 0.3 Ma, around the same time as or after the initial divergence events among species and between clades within *N. fuscipes* inferred from the genetic data. These earliest fossils are found in the San Francisco Bay Area (Alameda County; Savage 1951) and the San Joaquin Valley (Fairmead Landfill, Madera County; Dundas et al. 1996), and thus within the overall region identified as key origins of expansion for both lineages within *N. fuscipes*. These early fossils are not identified to individual species within *Neotoma*, so additional collections focused on the potential regions of divergence and subsequent expansion, aided by new methods for species-level identifications of specimens, would be extremely useful.

The central portion of California was particularly dynamic in the mid to late Pleistocene (Bartow 1991), and many factors could have contributed to the generation and maintenance of early divergence within *N. fuscipes*. These processes include continued and repeated glaciations extending down the western slopes of the Sierra Nevada (Richmond and Fullerton 1986), large (Sacramento River) and shifting freshwater drainages (Lock et al. 2006), extensive volcanic activity including the emergence of the Sutter Buttes in the northern Central Valley at ~1.6 mya (Prothero 2017), the inundation of the Central Valley by Corcoran Lake from ~700-600 ka (Bartow 1991), and the sudden establishment of the outflow of Corcoran Lake through the Carquinez Strait and what would become the modern San Francisco Bay-Delta ~600 ka (Sarna-Wojcicki et al. 1985; Bartow 1991). This historic landscape dynamism coupled with the fact that the modern San Francisco Bay-Delta presents a strong contemporary barrier to north-south movement for terrestrial species has led to this region being a concordant phylogeographic break among many taxa (Rodríguez-Robles et al. 2001; Feldman and Spicer 2006; Martínez-Solano et al. 2007; Phuong et al. 2014; Lavin et al. 2018; reviewed by Rissler et al. 2006 and Gottscho, 2016).

Following initial divergence of the northern and southern lineages of *N. fuscipes*, our demographic modeling shows there was little admixture between the two groups, with one group largely isolated to the north of the modern day San Francisco Bay-Sacramento-San Joaquin Delta, and the other largely restricted to the south, but primarily in the Coast Ranges on the western side of the Central Valley. Our niche modeling of present and LGM distributions of

suitable environmental conditions for the northern and southern lineages can be viewed as rough proxies of conditions during repeated glacial (LGM) and interglacial (present) conditions of the mid to late Pleistocene (Millar and Woolfenden 2016). These models suggest that the southern lineage would have had fairly continuous access to suitable conditions in the South Coast Ranges through both glacial and interglacial times, while the northern lineage would likely have experienced much greater shifts in suitable conditions, likely leading to repeated episodes of range expansion and contraction. Our estimates of the location of lineage coalescence, or the source area for subsequent expansion, for the northern lineage is located at a central point in the Central Valley, while the source point of the southern lineage is at the northern end of the South Coast Ranges near the modern San Francisco Bay-Delta.

2.5.2 The northern range

The dynamic history of the northern range of *N. fuscipes* makes it possible that there were multiple episodes of range expansion and contraction. One or more of these episodes led to a significant divergence within the northern lineage approximately 76 ka where the northern Sierra Nevada meet the southern Cascade Range. This timing is potentially consistent with an early Wisconsin glaciation where there is glacial evidence (Tahoe till) in the east-central Sierra Nevada that is younger than 118 - 119 ka and a till from Crater Lake National Park, Oregon that is older than 67 and 72 ka (Richmond and Fullerton 1986). In addition to repeated glaciations of the northern Sierra Nevada and southern Cascades, the southern Cascades have been particularly volcanically active (Sarna-Wojcicki et al. 1985). In California, Mount Shasta and Mount Lassen along with lava flows across the Modoc Plateau have likely repeatedly influenced local to regional habitat suitability for woodrats throughout the Pleistocene and Holocene. Spatially concordant with the genetic subdivision we found at the Sierra Nevada-Cascades transition, Hooper (1940) found a morphological subdivision between the subspecies *N. f. fuscipes* and *N. f. streatori*, suggesting the potential ecological significance of this subdivision. Matocq's (2002b) limited sampling in the northern Sierra Nevada and mtDNA analysis led to a failure to detect the distinction of the North A and North B clades.

In addition to woodrats, many other species show genetic discontinuities between the northern Sierra Nevada and southern Cascades (e.g., Rodríguez-Robles et al. 2001; Barrowclough et al. 2005; Kuchta et al. 2009; Phuong et al. 2014; Lavin et al. 2018). As in other parts of California, these common phylogeographic breaks are characterized by different depths of divergence across taxa depending on which particular glacial or volcanic event impacted a taxon in that region (Matocq 2002b). Of particular note is that spotted owls, an important predator of woodrats, have a similar distribution to *N. fuscipes* in northern California and share a similar spatial genetic disjunction at the boundary

of the northern Sierra Nevada and southern Cascades (Barrowclough et al. 2005).

2.5.3 Range expansion and secondary contact

Our niche modeling suggests that the southern lineage has occupied a part of the range that potentially had suitable environmental conditions through both glacial and interglacial periods of the Quaternary. Our range expansion analysis suggests that sometime after the initial divergence with the northern lineage, the southern lineage expanded into the South Coast Range region from the modern San Francisco Bay region. The possible north to south colonization of the Inner Coast Range by this lineage of *N. fuscipes* is also reflected in mtDNA data, with more northern haplotypes of the Inner Coast Range ancestral to haplotypes at the southern end of the range (Matocq 2002b). Using our SNP data, our estimates of relatively large effective population size, high genetic diversity, deeper divergences among genotypes (Figure 1; Figure 4; Figure 6), and a relatively pronounced pattern of isolation by distance (Figure 3) suggest relative stability in this portion of the range in comparison to the northern lineage. This is consistent with mtDNA data that also showed greater diversity and clade structure in the southern (west central) range in comparison to the north (Matocq 2002b). Regions inferred to be climatically stable through time are typically characterized by the accumulation of higher levels of genetic diversity and deeper genetic structure (Lessa et al. 2003; Carnaval et al. 2009; Jezkova et al. 2015, 2016 ; but see Hewitt 2004; Excoffier and Ray 2008), as seen here in the southern portion of the *N. fuscipes* range.

In contrast to the relative stability of the southern range, the northern range was likely characterized by repeated expansions and contractions of *N. fuscipes*, based on the known history of the region and our ENM's for this portion of the range (Fig. 2.5). The relatively small effective population size (Fig. 2.4), low genetic diversity (Fig. 2.6), shallow-star-like clade structure (Fig. 2.1; Slatkin and Hudson 1991; Avise 2000) and weak pattern of isolation by distance (Fig. 2.3) suggest that modern patterns of genetic variation in northern *N. fuscipes* are the result of relatively recent expansion (re-expansion) into northern California. It is likely that clades North A and North B re-expanded separately into their current distributions, but our lack of thorough sampling of the North B range precludes robust modeling of this process. Additionally, these methods assume populations are in a continuous and homogenous landscape and incorporating fragmented landscapes should be explored in future studies. Nonetheless, we were able to identify an area of lineage admixture just west of the Mount Lassen area. Interestingly, here too, there is striking concordance with an important predator of *N. fuscipes*, the spotted owl. This is the region in which northern and California spotted owl meet, and these two lineages admix in the area near Mount Lassen, just like *N. fuscipes* (Barrowclough et al. 2005). These authors suggest that this area of admixture between spotted owl lineages is due to a density trough related

to unsuitable habitat. In particular, much of the area is characterized by open understory pine forest, whereas owls prefer habitat with a more well-developed understory (Gutiérrez and Barrowclough 2005). Likewise, woodrats typically only occupy forested areas that have a well-developed understory in which they can build their houses (Sakai and Noon 1993; Innes et al. 2007). Interestingly, then, both predator and prey are characterized by a shared regional phylogeographic break and a concordant area of subsequent secondary contact and admixture. Space use (Carey et al. 1992; Zabel et al. 1995) and fitness (Thome et al. 1999) of spotted owls are correlated to woodrat abundance, so while both species may be responding independently to habitat structure in the area of admixture, owls are also likely responding to woodrat availability. This intriguing pattern suggests that building community-wide genomic, ENM and demographic modeling datasets may shed light on the ecological interactions and evolutionary processes that underlie patterns of phylogeographic and population genetic concordance among taxa.

2.5.4 Future directions

One limitation of our work, as with most efforts to hindcast species distributions, is the use of only climatic variables in our estimations. Ideally, future modeling efforts would incorporate biotic interactions, including those with vegetation and predators, and for these woodrats in particular—their interactions with congeners. The range of *N. fuscipes* today and through time has been strongly influenced by the distribution of closely related congeners with which they often compete or even hybridize (Matocq 2002a; Matocq and Murphy 2007; Matocq et al. 2012; Coyner et al. 2015; Dochtermann and Matocq 2016; Hunter et al. 2017). Woodrat species often compete for resources such as optimal nest sites and food resources (Dial 1988), leading to minimal overlap of species ranges and fine-scale parapatry (Matocq and Murphy 2007; Coyner et al. 2015; Shurtliff et al. 2014). The notion that differentially-adapted woodrat species replace each other in response to changing climate conditions is evident in the paleorecord (Smith et al. 1995, 2009) and even on a multi-annual scale (Gillespie et al. 2008; Hunter et al. 2017). In northern California, we suspect the distributional response of *N. fuscipes* has been influenced in part by the changing distributions of its relatively cold-adapted congener, the bushy-tailed woodrat *N. cinerea* (Hornsby and Matocq 2012) and the relatively arid-adapted *N. lepida* (Gillespie et al. 2008). As stated previously, most specimens are not distinguished at the species level in the fossil record of northern California [though both *N. cinerea* and *N. fuscipes* have been identified from cave deposits in northern California (Furlong 1904; Sinclair 1907; Stock 1917)], so morphology-based work on the genus would be useful, and genetic distinction of these species and reconstruction of their changing occupancy through time could be reconstructed using ancient DNA methods (reviewed in Shapiro and Hofreiter 2014). Ecological niche modeling for *N. fuscipes*, like many species, will be improved with integration of ecologically and physiologically-relevant parameters.

Our demographic analyses would also be improved with greater sampling, especially in the range of the North B clade and in the southern portion of the distribution. Another limitation to our study is the relatively small number of demographic scenarios we explored, given that the history and demography of *N. fuscipes* are certainly much more complex. For example, it is likely that these taxa, like many others, experienced asymmetric gene flow in their history. We only explored four asymmetric gene flow scenarios as a subset of scenario 11 (Appendix Figure 2.9.7). Our results suggest that while asymmetric models may be better (Appendix Table 2.9.11), parameter estimates are consistent between symmetric and asymmetric models (Appendix Table 2.9.7). Finally, the assumptions of mutation rate and generation time used in this study directly affect parameter estimates. For example, having a longer generation time or a slower mutation rate would shift the estimates of population subdivision further back in time. Nonetheless, genome-wide datasets and the potential to exploit SFS patterns is a huge advance in the field of population genetics and these data and methods will continue to be refined and integrated with even more sophisticated simulation and spatial modeling approaches that promise improved insight into complex biogeographic histories.

2.6 References

Adams RH, Schield DR, Card DC, Corbin A, Castoe TA (2018). ThetaMater: Bayesian estimation of population size parameter θ from genomic data (Oliver, Ed.). *Bioinformatics* **34**: 1072–1073.

Aiello-Lammens ME, Boria RA, Radosavljevic A, Vilela B, Anderson RP (2015). spThin: An R package for spatial thinning of species occurrence records for use in ecological niche models. *Ecography* **38**: 541–545.

Alexander DH, Novembre J, Lange K (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* **19**: 1655–1664.

Anderson RP, Raza A (2010). The effect of the extent of the study region on GIS models of species geographic distributions and estimates of niche evolution: preliminary tests with montane rodents (genus Nephelomys) in Venezuela. *J Biogeogr* **37**: 1378–1393.

Avise JC (2000). *Phylogeography: The History and Formation of Species*. Harvard University Press.

Badgley C, Smiley TM, Terry R, Davis EB, DeSantis LRG, Fox DL, et al. (2017). Biodiversity and Topographic Complexity: Modern and Geohistorical Perspectives. *Trends Ecol Evol* **32**: 211–226.

Bartow JA (1991). *The Cenozoic evolution of the San Joaquin Valley, California*.

Barve N, Barve V, Jiménez-Valverde A, Lira-Noriega A, Maher SP, Peterson AT, et al. (2011). The crucial role of the accessible area in ecological niche modeling and species distribution modeling. *Ecol Model* **222**: 1810–1819.

Blois JL, McGuire JL, Hadly EA (2010). Small mammal diversity loss in response to late-Pleistocene climatic change. *Nature* **465**: 771–774.

Boria RA, Olson LE, Goodman SM, Anderson RP (2014). Spatial filtering to reduce sampling bias can improve the performance of ecological niche models. *Ecol Model* **275**.

Brown JH (1968). Adaptation to environmental temperature in two species of woodrats, Neotoma cinerea and N. albigula.

Burbrink FT, Chan YL, Myers EA, Ruane S, Smith BT, Hickerson MJ (2016). Asynchronous demographic responses to Pleistocene climate change in Eastern Nearctic vertebrates. *Ecol Lett* **19**: 1457–1467.

Carey AB, Horton SP, Biswell BL (1992). Northern Spotted Owls: Influence of Prey Base and Landscape Character. *Ecol Monogr* **62**: 223–250.

Carnaval AC, Hickerson MJ, Haddad CFB, Rodrigues MT, Moritz C (2009). Stability predicts genetic diversity in the Brazilian Atlantic forest hotspot. *Science* **323**: 785–789.

Carraway LN, Verts BJ (1991). Neotoma fuscipes. *Mamm Species*: 1–10.

Catchen J, Hohenlohe PA, Bassham S, Amores A, Cresko WA (2013). Stacks: an analysis tool set for population genomics. *Mol Ecol* **22**: 3124–3140.

Clark PU, Shakun JD, Baker PA, Bartlein PJ, Brewer S, Brook E, et al. (2012). Global climate evolution during the last deglaciation. *Proc Natl Acad Sci* **109**: E1134–E1142.

Coyner BS, Murphy PJ, Matocq MD (2015). Hybridization and asymmetric introgression across a narrow zone of contact between Neotoma fuscipes and N. macrotis (Rodentia: Cricetidae). *Biol J Linn Soc* **115**: 162–172.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. (2011). The variant call format and VCFtools. *Bioinformatics* **27**: 2156–2158.

Dansgaard W, Johnsen SJ, Clausen HB, Dahl-Jensen D, Gundestrup NS, Hammer CU, et al. (1993). Evidence for general instability of past climate from a 250-kyr ice-core record. *Nature* **364**: 218–220.

Davis EB, Koo MS, Conroy C, Patton JL, Moritz C (2008). The California Hotspots Project: identifying regions of rapid diversification of mammals. *Mol Ecol* **17**: 120–138.

Dial KP (1988). Three sympatric species of Neotoma: dietary specialization and coexistence. *Oecologia* **76**: 531–537.

Dochtermann NA, Matocq MD (2016). Speciation along a shared evolutionary trajectory. *Curr Zool* **62**: 507–511.

Dundas RG, Smith RB, Verosub KL (1996). The Fairmead Landfill Locality (Pleistocene, Irvingtonian), Madera County, California: preliminary report and significance. *PaleoBios* **17**: 50–58.

Dyke AS, Moore A, Robertson L (2003). Deglaciation of North America. *Geol Surv Can Open File 1574*.

Earl DA, vonHoldt BM (2012). STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conserv Genet Resour* **4**: 359–361.

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

Excoffier L, Dupanloup I, Huerta-Sanchez E, Sousa VC, FollSousa M (2013). Robust Demographic Inference from Genomic and SNP Data. *PLoS Genet* **9**: e1003905–e1003905.

Excoffier L, Ray N (2008). Surfing during population expansions promotes genetic revolutions and structuration. *Trends Ecol Evol* **23**: 24–28.

Feldman CR, Spicer GS (2006). Comparative phylogeography of woodland reptiles in California: repeated patterns of cladogenesis and population expansion. *Mol Ecol* **15**: 2201–2222.

Foley J, Rejmanek D, Foley C, Matocq M (2016). Fine-scale genetic structure of woodrat populations (Genus: Neotoma) and the spatial distribution of their tick-borne pathogens. *Ticks Tick-Borne Dis* **7**: 243–253.

Furlong EL (1904). An Account of the Preliminary Excavations in a Recently Explored Quaternary Cave in Shasta County, California. *Science* **20**: 53–55.

Gillespie SC, Vuren DHV, Kelt DA, Eadie JM, Anderson DW (2008). Dynamics of Rodent Populations in Semiarid Habitats in Lassen County, California. *West North Am Nat* **68**: 76–82.

Goldman EA (1910). *Revision of the Wood Rats of the Genus Neotoma*. U.S. Government Printing Office.

Gottscho AD (2016). Zoogeography of the San Andreas Fault system: Great Pacific Fracture Zones correspond with spatially concordant phylogeographic boundaries in western North America: Zoogeography of the San Andreas Fault system. *Biol Rev* **91**: 235–254.

Graham RW, Lundelius EL, Graham MA, Schroeder EK, Toomey RS, Anderson E, et al. (1996). Spatial Response of Mammals to Late Quaternary Environmental Fluctuations. *Science* **272**: 1601–1606.

Gutiérrez RJ, Barrowclough GF (2005). Redefining the Distributional Boundaries of the Northern and California Spotted Owls: Implications for Conservation. *The Condor* **107**: 182–187.

Harrington SM, Hollingsworth BD, Higham TE, Reeder TW (2017). Pleistocene climatic fluctuations drive isolation and secondary contact in the red diamond rattlesnake (Crotalus ruber) in Baja California. *J Biogeogr* **45**: 64–75.

Hewitt G (2000). The genetic legacy of the Quaternary ice ages. *Nature* **405**: 907–913.

Hewitt GM (2004). Genetic consequences of climatic oscillations in the Quaternary. *Philos Trans R Soc Lond B Biol Sci* **359**: 183–195.

Hooper ET (1940). Geographic variation in bushy-tailed wood rats. *Univ Calif Press* **42**: 407–424.

Hopkins DM, Matthews JV, Schweger CE (2013). *Paleoecology of Beringia*. Elsevier.

Hornsby AD, Matocq MD (2012). Differential regional response of the bushy-tailed woodrat (Neotoma cinerea) to late Quaternary climate change. *J Biogeogr* **39**: 289–305.

Hotaling S, Muhlfeld CC, Giersch JJ, Ali OA, Jordan S, Miller MR, et al. (2018). Demographic modelling reveals a history of divergence with gene flow for a glacially tied stonefly in a changing post- Pleistocene landscape. *J Biogeogr* **45**: 304–317.

Huang H, Knowles LL (2016). Unforeseen consequences of excluding missing data from next-generation sequences: Simulation study of rad sequences. *Syst Biol* **65**: 357–365.

Hunter EA, Matocq MD, Murphy PJ, Shoemaker KT (2017). Differential Effects of Climate on Survival Rates Drive Hybrid Zone Movement. *Curr Biol* **27**: 3898-3903.e4.

Huson DH, Bryant D (2006). Application of Phylogenetic Networks in Evolutionary Studies. *Mol Biol Evol* **23**: 254–267.

Hutchison DW, Templeton AR (1999). Correlation of Pairwise Genetic and Geographic Distance Measures: Inferring the Relative Influences of Gene Flow and Drift on the Distribution of Genetic Variability. *Evolution* **53**: 1898–1914.

Innes RJ, Van Vuren DH, Kelt DA, Johnson ML, Wilson JA, Stine PA (2007). Habitat Associations of Dusky-Footed Woodrats (Neotoma fuscipes) in Mixed-Conifer Forest of the Northern Sierra Nevada. *J Mammal* **88**: 1523–1531.

Jackson ST, Betancourt JL, Lyford ME, Gray ST, Rylander KA (2005). A 40,000-year woodrat-midden record of vegetational and biogeographical dynamics in north-eastern Utah , USA. *J Biogeogr* **32**: 1085–1106.

Jackson ST, Overpeck JT (2000). Responses of plant populations and communities to environmental changes of the late Quaternary. *Paleobiology* **26**: 194-220.

Janes JK, Miller JM, Dupuis JR, Malenfant RM, Gorrell JC, Cullingham CI, et al. (2017). The K = 2 conundrum. *Mol Ecol* **26**: 3594–3602.

Jezkova T, Jaeger JR, Oláh-Hemmings V, Jones KB, Lara-Resendiz RA, Mulcahy DG, et al. (2016). Range and niche shifts in response to past climate change in the desert horned lizard Phrynosoma platyrhinos. *Ecography* **39**: 437–448.

Jezkova T, Riddle BR, Card DC, Schield DR, Eckstut ME, Castoe TA (2015). Genetic consequences of postglacial range expansion in two codistributed rodents (genus Dipodomys) depend on ecology and genetic locus. *Mol Ecol* **24**: 83–97.

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

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

Jukes, TH, Cantor CR (1969). *Mammalian Protein Metabolism*. Elsevier.

Knowles LL, Massatti R (2017). Distributional shifts – not geographic isolation – as a probable driver of montane species divergence. *Ecography* **40**: 1475–1485.

Kuchta SR, Parks DS, Mueller RL, Wake DB (2009). Closing the ring: historical biogeography of the salamander ring species *Ensatina eschscholtzii*. *J Biogeogr* **36**: 982–995.

Lane RS, Brown RN (1991). Wood Rats and Kangaroo Rats: Potential Reservoirs of the Lyme Disease Spirochete in California. *J Med Entomol* **28**: 299–302.

Lavin BR, Wogan GOU, McGuire JA, Feldman CR (2018). Phylogeography of the Northern Alligator Lizard (Squamata, Anguidae): Hidden diversity in a western endemic. *Zool Scr* **47**: 462–476.

Leaché AD, Banbury BL, Felsenstein J, De Oca ANM, Stamatakis A (2015). Short tree, long tree, right tree, wrong tree: New acquisition bias corrections for inferring SNP phylogenies. *Syst Biol* **64**: 1032–1047.

Lessa EP, Cook JA, Patton JL (2003). Genetic footprints of demographic expansion in North America, but not Amazonia, during the Late Quaternary. *Proc Natl Acad Sci* **100**: 10331–10334.

Liu Z, Otto-Bliesner BL, He F, Brady EC, Tomas R, Clark PU, et al. (2009). Transient Simulation of Last Deglaciation with a New Mechanism for Bølling-Allerød Warming. *Science* **325**: 310–314.

Lock J, Kelsey H, Furlong K, Woolace A (2006). Late Neogene and Quaternary landscape evolution of the northern California Coast Ranges: Evidence for Mendocino triple junction tectonics. *GSA Bull* **118**: 1232–1246.

Lorenz DJ, Nieto-Lugilde D, Blois JL, Fitzpatrick MC, Williams JW (2016). Downscaled and debiased climate simulations for North America from 21,000 years ago to 2100AD. *Sci Data* **3**: 160048.

Martínez-Solano I, Jockusch EL, Wake DB (2007). Extreme population subdivision throughout a continuous range: phylogeography of *Batrachoseps attenuatus* (Caudata: Plethodontidae) in western North America. *Mol Ecol* **16**: 4335–4355.

Matocq MD (2002a). Morphological and Molecular Analysis of a Contact Zone in the Neotoma Fuscipes Species Complex. *J Mammal* **83**: 866–883.

Matocq MD (2002b). Phylogeographical structure and regional history of the dusky-footed woodrat , Neotoma fuscipes. *Mol Ecol* **11**: 229–242.

Matocq MD, Kelly PA, Phillips SE, Maldonado JE (2012). Reconstructing the evolutionary history of an endangered subspecies across the changing landscape of the Great Central Valley of California. *Mol Ecol* **21**: 5918–5933.

Matocq MD, Murphy PJ (2007). Fine-Scale Phenotypic Change Across a Species Transition Zone in the Genus Neotoma: Disentangling Independent Evolution from Phylogenetic History. *Evolution* **61**: 2544–2557.

Millar CI, Woolfenden WB (2016). Ecosystems Past: In: Mooney H, Zavaleta E, Chapin MC (eds) *Ecosystems of California*, University of California Press, pp 131–154.

Muscarella R, Galante PJ, Soley-Guardia M, Boria RA, Kass JM, Uriarte M, et al. (2014). ENMeval: An R package for conducting spatially independent evaluations and estimating optimal model complexity for Maxent ecological niche models. *Methods Ecol Evol* **5**: 1198–1205.

Myers N, Mittermeier RA, Mittermeier CG, Fonseca GAB da, Kent J (2000). Biodiversity hotspots for conservation priorities. *Nature* **403**: 853–858.

Nachman MW, Crowell SL (2000). Estimate of the Mutation Rate per Nucleotide in Humans. *Genetics* **156**: 297–304.

Nogués-Bravo D, Rodríguez J, Hortal J, Batra P, Araújo MB (2008). Climate Change, Humans, and the Extinction of the Woolly Mammoth. *PLOS Biol* **6**: e79.

Parisi M (ed.) (2003). Atlas of the Biodiversity of California. *Calif Dep Fish Game Sacram CA*.

Pearson RG, Raxworthy CJ, Nakamura M, Townsend PA (2007). Predicting species distributions from small numbers of occurrence records: a test case using cryptic geckos in Madagascar. *J Biogeogr* **34**: 102–117.

Peter BM, Slatkin M (2013). Detecting range expansion from genetic data. *Evolution* **67**: 3274–3289.

Peter BM, Slatkin M (2015). The effective founder effect in a spatially expanding population. *Evolution* **69**: 721–734.

Peterson AT (2009). Phylogeography is not enough: The need for multiple lines of evidence. *Front Biogeogr* **1**: 19–25.

Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE (2012). Double digest RADseq: An inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PLoS ONE* **7**.

Phillips SJ, Anderson RP, Dudík M, Schapire RE, Blair ME (2017). Opening the black box: an open-source release of Maxent. *Ecography* **40**: 887–893.

Phillips SJ, Anderson RP, Schapire RE (2006). Maximum entropy modeling of species geographic distributions. *Ecol Model* **190**: 231–259.

Phuong MA, Lim MCW, Wait DR, Rowe KC, Moritz C (2014). Delimiting species in the genus *Otospermophilus* (Rodentia: Sciuridae), using genetics, ecology, and morphology: Species delimitation in ground squirrels. *Biol J Linn Soc* **113**: 1136–1151.

Pritchard JK, Stephens M, Donnelly P (2000). Inference of Population Structure Using Multilocus Genotype Data. *Genetics* **155**: 945–959.

Prothero DR (2017). *California's Amazing Geology*. CRC Press.

Reid BN, Kass JM, Wollney S, Jensen EL, Russello MA, Viola EM, et al. (2019). Disentangling the genetic effects of refugial isolation and range expansion in a trans-continentally distributed species. *Heredity* **122**: 441–457.

Richmond GM, Fullerton DS (1986). Summation of quaternary glaciations in the United States of America. *Quat Sci Rev* **5**: 183–196.

Rissler LJ, Hijmans RJ, Graham CH, Moritz C, Wake DB (2006). Phylogeographic Lineages and Species Comparisons in Conservation Analyses : A Case Study of California Herpetofauna. *Am Nat* **167**: 655–666.

Rodríguez-Robles JA, Stewart GR, Papenfuss TJ (2001). Mitochondrial DNA-Based Phylogeography of North American Rubber Boas, Charina bottae (Serpentes: Boidae). *Mol Phylogenet Evol* **18**: 227–237.

Sakai HF, Noon BR (1993). Dusky-Footed Woodrat Abundance in Different-Aged Forests in Northwestern California. *J Wildl Manag* **57**: 373–382.

Sarna-Wojcicki AM, Meyer CE, Bowman HR, Hall NT, Russell PC, Woodward MJ, et al. (1985). Correlation of the Rockland Ash Bed, a 400,000-Year-Old Stratigraphic Marker in Northern California and Western Nevada, and Implications for Middle Pleistocene Paleogeography of Central California. *Quat Res* **23**: 236–257.

Savage D (1951). Late Cenozoic vertebrates of the San Francisco Bay region, California. *Publ Depart Geol Sci* **28**: 1–30.

Schierenbeck KA (2014). *Phylogeography of California: An Introduction*. Univ of California Press.

Shapiro B, Hofreiter M (2014). A Paleogenomic Perspective on Evolution and Gene Function: New Insights from Ancient DNA. *Science* **343**.

Shcheglovitova M, Anderson RP (2013). Estimating optimal complexity for ecological niche models: A jackknife approach for species with small sample sizes. *Ecol Model* **269**: 9–17.

Shurtliff QR, Murphy PJ, Matocq MD (2014). Ecological Segregation in a Small Mammal Hybrid Zone: Habitat-Specific Mating Opportunities and Selection Against Hybrids Restrict Gene Flow on a Fine Spatial Scale. *Evolution* **68**: 729–742.

Sinclair WJ (1907). *The Exploration of the Potter Creek Cave*. University of California Press.

Slatkin M, Hudson RR (1991). Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. *Genetics* **129**: 555–562.

Smith FA, Brown JH, Betancourt, JL (1995). Evolution of Body Size in the Woodrat over the Past 25,000 Years of Climate Change. *Science* **270**: 2012–2014.

Smith FA, Crawford DL, Harding LE, Lease HM, Murray IW, Raniszewski A, et al. (2009). A tale of two species: Extirpation and range expansion during the late Quaternary in an extreme environment. *Glob Planet Change* **65**: 122–133.

Soulé ME, Estes JA, Miller B, Honnold DL (2005). Strongly Interacting Species: Conservation Policy, Management, and Ethics. *BioScience* **55**: 168–176.

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

Stamatakis A, Ott M, Ludwig T (2005). RAxML-OMP: An Efficient Program for Phylogenetic Inference on SMPs. In: Malyshkin V (ed) *Parallel Computing Technologies*, Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp 288–302.

Stock C (1917). *The Pleistocene Fauna of Hawver Cave*. University of California Press.

van Strien MJ, Holderegger R, Van Heck HJ (2015). Isolation-by-distance in landscapes: considerations for landscape genetics. *Heredity* **114**: 27–37.

Svenning J-C, Skov F (2004). Limited filling of the potential range in European tree species: Limited range filling in European trees. *Ecol Lett* **7**: 565–573.

Thome DM, Zabel CJ, Diller LV (1999). Forest Stand Characteristics and Reproduction of Northern Spotted Owls in Managed North-Coastal California Forests. *J Wildl Manag* **63**: 44–59.

Waltari E, Hijmans RJ, Peterson AT, Nyári ÁS, Perkins SL, Guralnick RP (2007). Locating Pleistocene Refugia: Comparing Phylogeographic and Ecological Niche Model Predictions (J Chave, Ed.). *PLoS ONE* **2**: e563.

Whitford WG, Steinberger Y (2010). Pack rats (Neotoma spp.): Keystone ecological engineers? *J Arid Environ* **74**: 1450–1455.

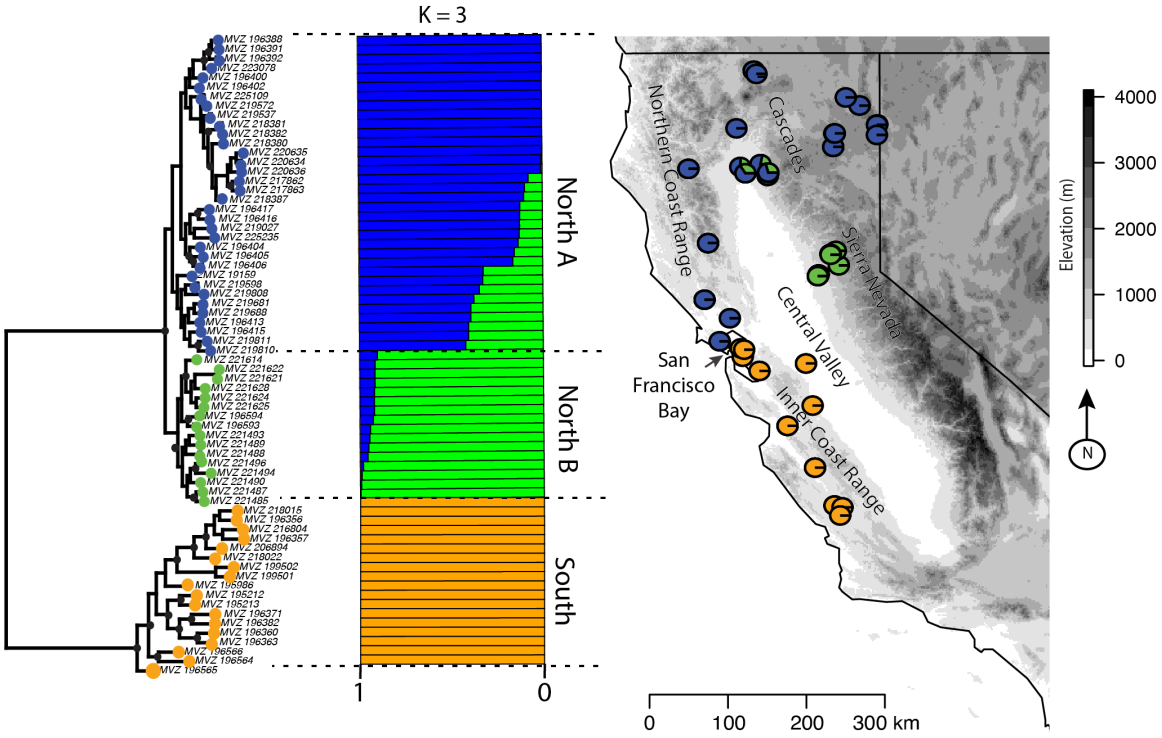
Zabel CJ, McKelvey K, Ward Jr. JP (1995). Influence of primary prey on home-range size and habitat-use patterns of northern spotted owls (Strix occidentalis caurina). *Can J Zool* **73**: 433–439.

2.7 Tables

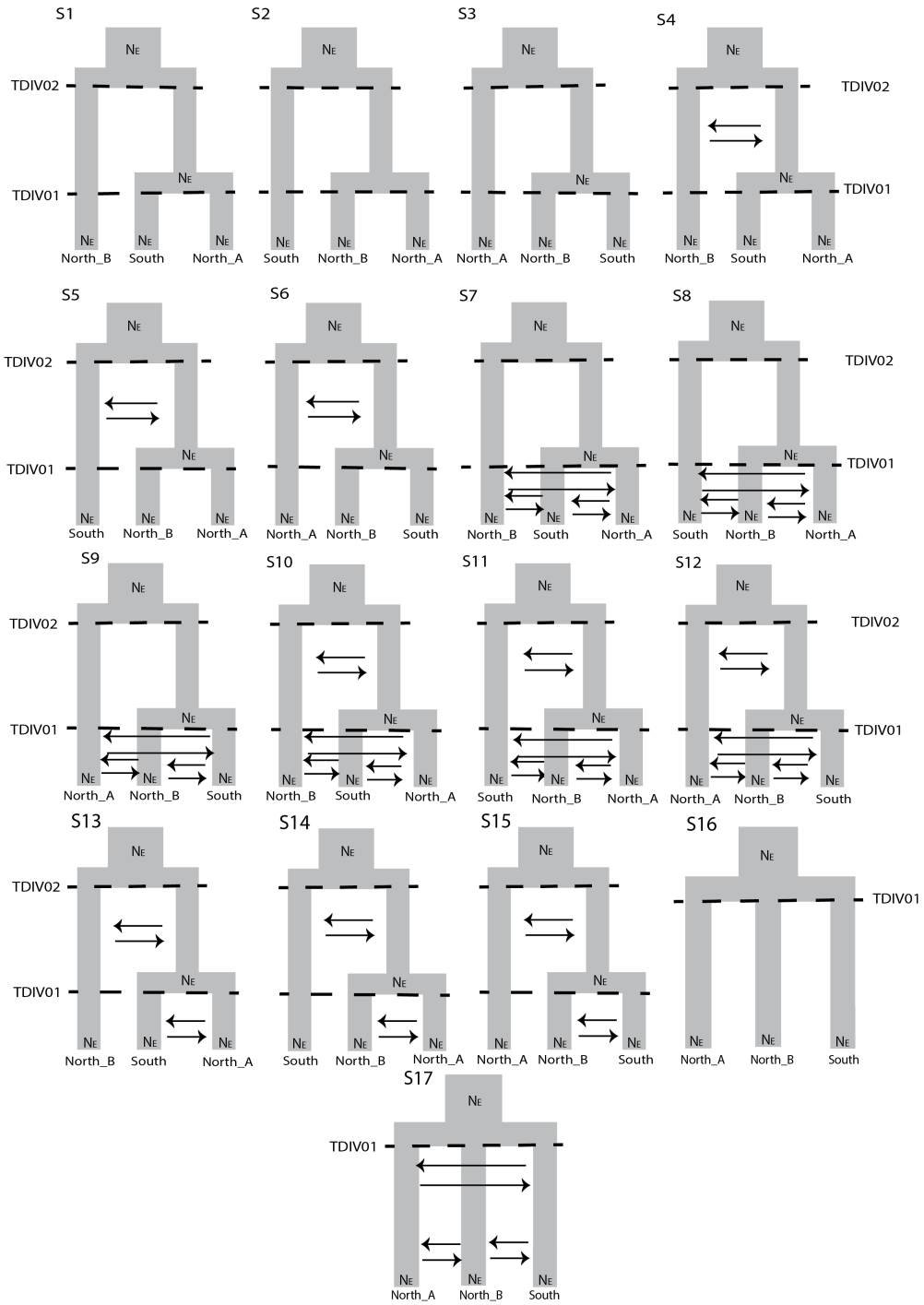
2.7.1: Top five demographic scenarios using fastsimcoal2 according to AIC scores for *Neotoma fuscipes* in California. Scenario 11 was indicated as the model that best fit the data.

Scenarios	Parameters	AIC	△AIC	AIC_W
S11	15	117943.029	0	.99
S14	11	117947.791	4.762	4.508×10^{-5}
S12	15	118037.638	94.609	1.393×10^{-24}
S8	15	118193.532	250.503	1.959×10^{-58}
S10	15	118237.327	294.298	6.053×10^{-68}

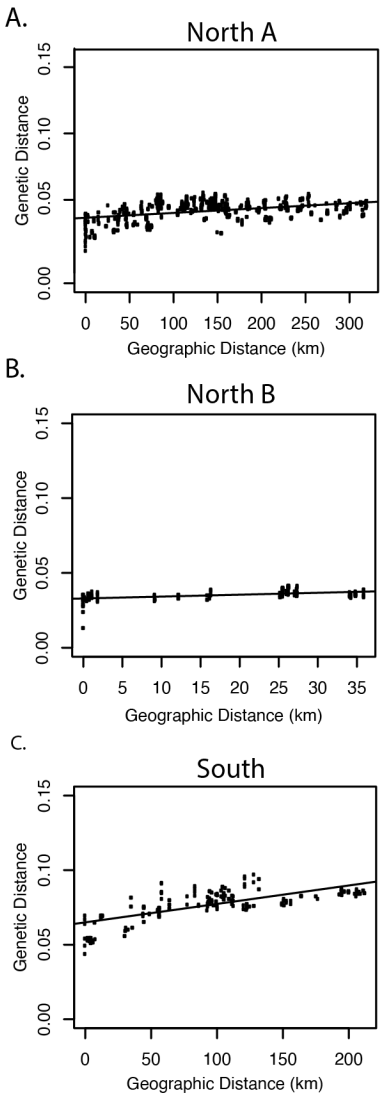
2.8 Figures



2.1 The geographic structure of *Neotoma fuscipes* across its range in California based on the 85% missing-data threshold ddRAD matrix. Maximum likelihood tree, ADMIXTURE, and STRUCTURE results showing three distinct clusters (*K* = 3): one southern and two northern populations. Colors on the branch tips correspond to the ADMIXTURE/ STRUCTURE population individuals were placed within, and the colored circles on map indicate the proportion of ancestry each individual has to a population. The black circles at nodes indicate bootstrap support > 70%.

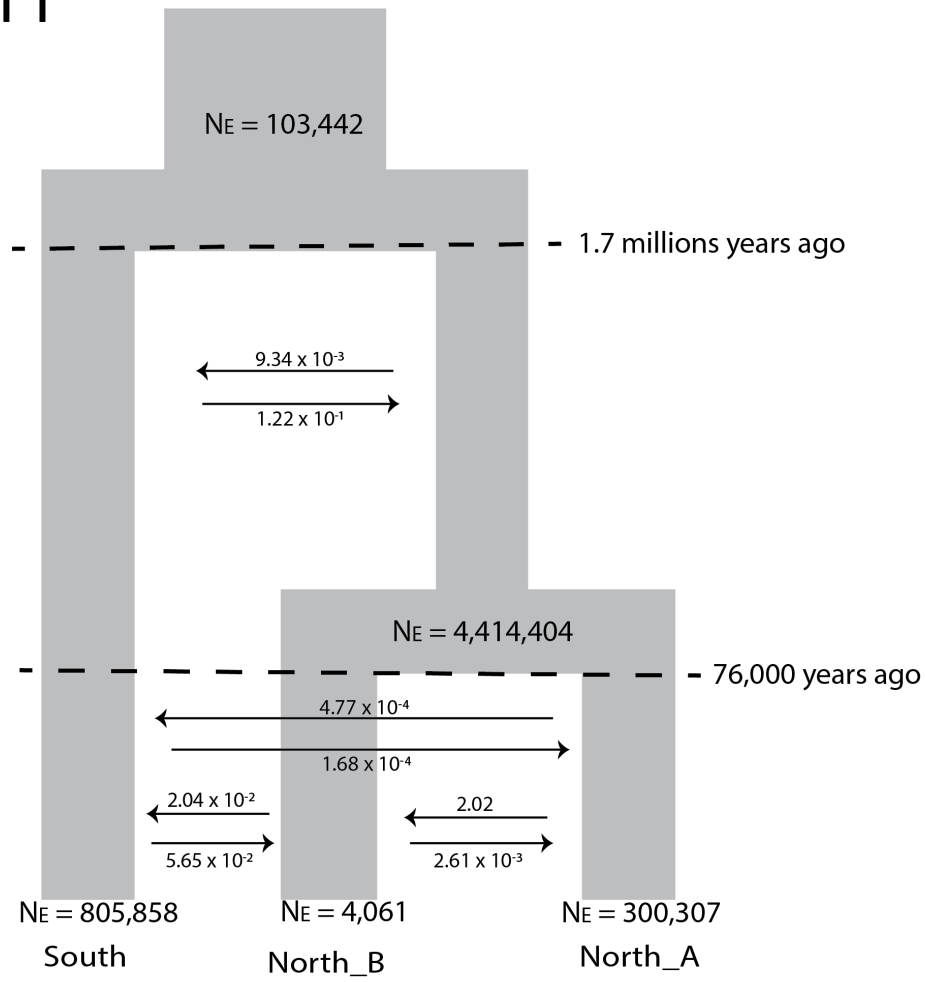


2.2 The 17 phylogeographic hypotheses used to simulate demographic history for *Neotoma fuscipes* in California. The parameters included contemporary population size for the three populations, their ancestral population sizes, migration rates between the populations, and the timing of divergence.

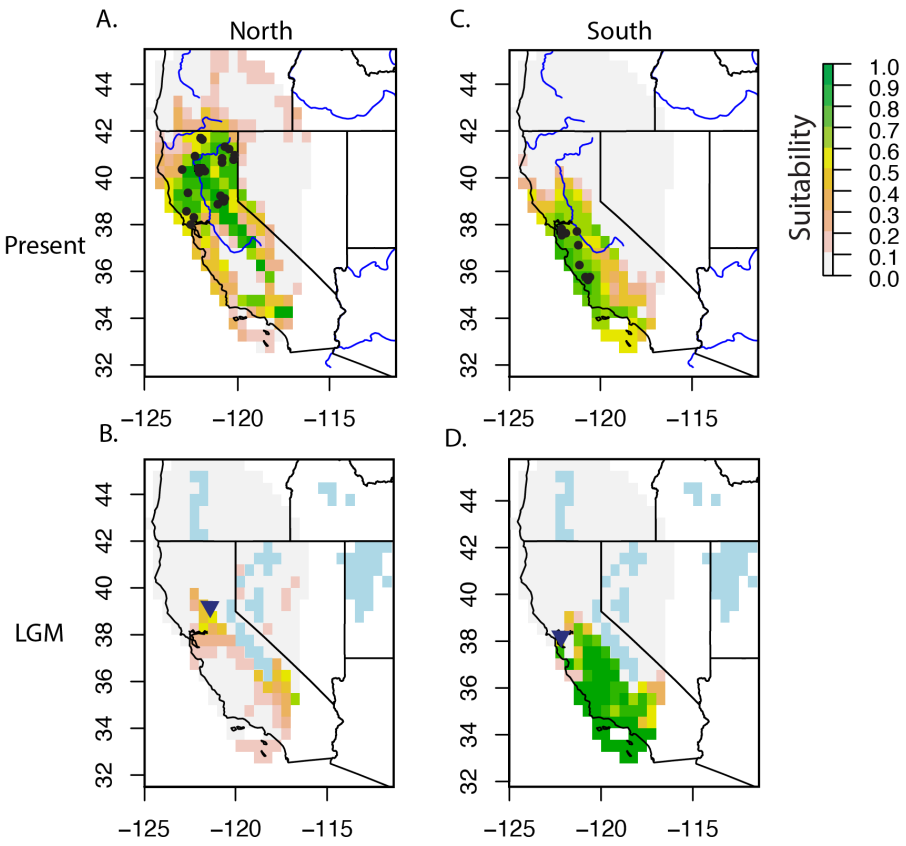


2.3 Plots of genetic distance by geographic distance for all populations of *Neotoma fuscipes* in California: A) North A ($r = 0.537$, $p < 0.001$); B) North B ($r = 0.544$, $p < 0.001$); C) South ($r = 0.731$, $p < 0.001$). Note that there is significant isolation by distance found within each population.

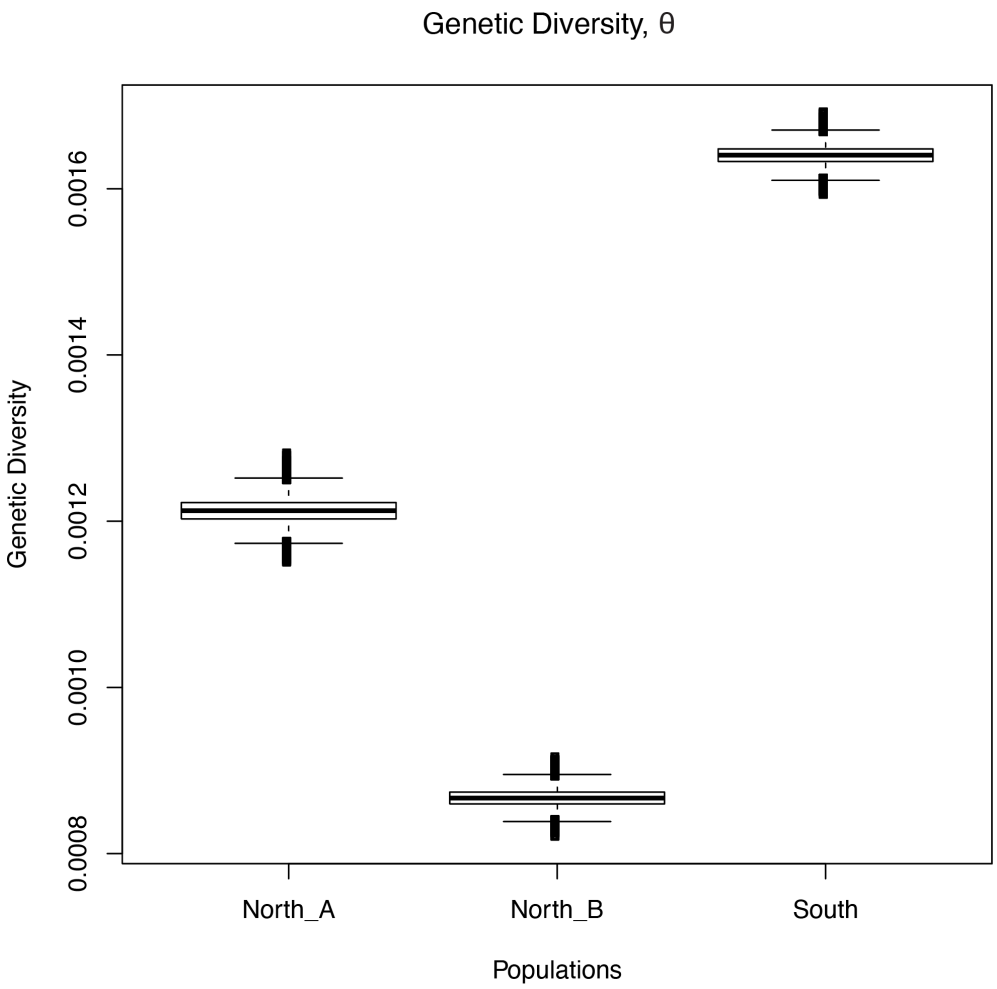
S11



2.4 The best fit model and parameter estimates of demographic history for *Neotoma fuscipes* in California.



2.5 Ecological niche models for the combined northern and southern populations of *Neotoma fuscipes* in California: A) combined north present day; B) combined north last glacial maximum; C) south present day; D) south last glacial maximum. Note the reduction of possible suitable areas in the past for both the northern and southern populations. Blue lines in A and C indicate rivers and lake centerlines greater than 50 M (made with Natural Earth. Free vector and raster map data @ naturalearthdata.com), and blue pixels in B and D indicate ice or paleolakes during the last glacial maximum (Dyke et al. 2003). The blue triangle in B and D indicate the origin of expansion for the combined north and south, respectively.



2.6 The posterior probability distribution of genetic diversity analyses from ThetaMater for the three populations, North A, North B, and South, of *Neotoma fuscipes* in California. Note the southern population had the highest diversity, followed by North A, and then North B.

2.9 Appendix

Appendix Table 2.9.1: Gazetteer of specimens corresponding to localities of *Neotoma fuscipes* employed in this study. Museum, catalog number, and geographic coordinates (in decimal degrees), collector/preparator, clade/population assignment if included in previous studies, and clade/population assignment in current study. All samples used in this study are held at the University of California Museum of Vertebrate Zoology.

ID	DEC_LAT	DEC_LONG	Collector	Matocq 2002a,b	Population
MVZ_195212	36.27	-121.06	JLP17570		South
MVZ_195213	36.27	-121.06	JLP17571		South
MVZ_195986	37.12	-121.10	JLP17862		South
MVZ_196356	37.87	-122.15	MDM432	West Central	South
MVZ_196357	37.86	-122.13	MDM662	West Central	South
MVZ_196360	35.75	-120.77	AEJ15	West Central	South
MVZ_196363	35.75	-120.77	JLP17489	West Central	South
MVZ_196371	35.73	-120.65	JLP17483	West Central	South
MVZ_196382	35.61	-120.69	AEJ32	West Central	South
MVZ_196388	41.71	-121.98	MDM375	Northern	North_A
MVZ_196391	41.71	-121.98	MDM380	Northern	North_A
MVZ_196392	41.66	-121.94	MDM382	Northern	North_A
MVZ_196400	40.36	-122.96	MDM413	Northern	North_A
MVZ_196402	40.36	-122.96	MDM415	Northern	North_A
MVZ_196404	39.35	-122.66	MDM420		North_A

MVZ_196405	39.35	-122.66	MDM421	Northern	North_A
MVZ_196406	39.35	-122.66	MDM422		North_A
MVZ_196413	40.27	-121.77	MDM358	Northern	North_A
MVZ_196415	40.27	-121.77	MDM360	Northern	North_A
MVZ_196416	38.57	-122.72	MDM368	Northern	North_A
MVZ_196417	38.57	-122.72	MDM369	Northern	North_A
MVZ_196564	36.84	-121.48	MDM433	West Central	South
MVZ_196565	36.84	-121.48	MDM434	West Central	South
MVZ_196566	36.84	-121.48	MDM435	West Central	South
MVZ_196593	39.04	-120.71	MDM352	Northern	North_B
MVZ_196594	39.04	-120.71	MDM353	Northern	North_B
MVZ_199501	37.70	-121.20	CJC703		South
MVZ_199502	37.70	-121.20	CJC704		South
MVZ_206894	37.79	-122.15	ALI7		South
MVZ_216804	37.90	-122.18	JLP22463		South
MVZ_217862	40.66	-120.79	JDP28		North_A
MVZ_217863	40.66	-120.79	CJC1888		North_A
MVZ_218015	37.89	-122.12	JTW1232		South
MVZ_218022	37.59	-121.90	KT59		South
MVZ_218380	40.97	-120.13	RD134		North_A
MVZ_218381	40.97	-120.14	JAC65		North_A
MVZ_218382	40.97	-120.14	JAC66		North_A
MVZ_218387	40.83	-120.14	JDP232		North_A
MVZ_219027	38.00	-122.49	RD183		North_A
MVZ_219159	40.43	-121.88	JDP432		North_A

MVZ_219537	41.23	-120.40	SAM72		North_A
MVZ_219572	41.23	-120.40	SAM82		North_A
MVZ_219598	40.39	-122.18	RD205		North_A
MVZ_219681	40.30	-122.11	ANL23		North_A
MVZ_219688	40.30	-122.11	ANL38		North_A
MVZ_219808	40.31	-121.76	CJC2250		North_A
MVZ_219810	40.31	-121.77	CJC2255		North_A
MVZ_219811	40.31	-121.77	CJC2270		North_A
MVZ_220634	40.85	-120.77	CJC2385		North_A
MVZ_220635	40.85	-120.77	CJC2386		North_A
MVZ_220636	40.85	-120.77	FPR66		North_A
MVZ_221485	38.90	-121.01	CJC2431		North_B
MVZ_221487	38.90	-121.01	CJC2471		North_B
MVZ_221488	38.90	-121.01	CJC2422		North_B
MVZ_221489	38.92	-121.01	CJC2465		North_B
MVZ_221490	38.92	-121.01	CJC2443		North_B
MVZ_221493	38.92	-121.01	CJC2463		North_B
MVZ_221494	38.89	-121.02	CJC2421		North_B
MVZ_221496	38.89	-121.02	CJC2430		North_B
MVZ_221614	39.24	-120.75	KRC315		North_B
MVZ_221621	39.20	-120.83	CJC2521		North_B
MVZ_221622	39.20	-120.83	CJC2522		North_B
MVZ_221624	39.19	-120.83	CJC2508		North_B
MVZ_221625	39.19	-120.83	CJC2528		North_B
MVZ_221628	39.19	-120.83	CJC2511		North_B
MVZ_223078	41.34	-120.60	JLP24444		North_A

MVZ_225109	40.92	-122.24	JLB70		North_A
MVZ_225235	38.32	-122.34	REJ446		North_A
MVZ_196387	41.83	-123.14	MDM374	Northern	Missing Data
MVZ_219021	38.01	-122.47	IL59		Missing Data
MVZ_219599	40.36	-122.17	RD194		Missing Data
MVZ_196422	36.38	-121.56	MDM546	*N. macrotis*	macrotis
MVZ_197373	35.09	-119.78	MCKP1		macrotis
MVZ_198571	35.34	-120.71	JLP16711	*N. macrotis*	macrotis
MVZ_198573	35.34	-120.71	JLP16713	*N. macrotis*	macrotis
MVZ_201919	37.72	-119.67	JLP19933		macrotis
MVZ_216081	34.47	-120.23	CJC1654		macrotis
MVZ_218708	37.38	-121.74	JTW1249		macrotis
MVZ_230700	35.33	-120.86	SP2760		macrotis

Appendix Table 2.9 2: Inter-individual genetic distances within populations of *Neotoma fuscipes* in California.

Populations	Inter-individual genetic distances
North A	0.040
North B	0.033
South	0.073

Appendix Table 2.9.3: The cross-validation results of admixture for *Neotoma fuscipes* in California indicating *K*=3 model as optimal. These results are for the 85% dataset.

Populations	CV error
K=1	0.55171
K=2	0.24569
K=3	0.23397
K=4	0.24831
K=5	0.24987

Appendix Table 2.9.4: The cross-validation results of admixture for *Neotoma fuscipes* in California indicating *K*=3 model as optimal. These results are for the 50% dataset.

Populations	CV error
K=1	0.58831
K=2	0.28679
K=3	0.27287
K=4	0.29439
K=5	0.28647

Appendix Table 2.9.5: Pairwise F_{ST} between populations of *Neotoma fuscipes* in California.

	North A	North B	South
North A	-	-	-
North B	0.077	-	-
South	0.342	0.377	-

Appendix Table 2.9.6: The slope of the isoloation-by-distance plots of *Neotoma fuscipes* in California.

Populations	Slope
All	0.07
North A	3.22×10^{-05}
North B	1.27×10^{-04}
South	1.24×10^{-04}

Appendix Table 2.9.7: The demographic scenarios using fastsimcoal2 according to △AIC for *Neotoma fuscipes* in California. Scenario 11 was indicated as the model that best fit the data. See Figure 2 for a full explanation of the different scenarios.

Scenarios	Parameters	AIC	△AIC	AIC_W
S1	7	120688.354	2760.577	$< 9.594 \times 10^{-279}$
S2	7	118941.815	1014.038	6.373×10^{-221}
S3	7	120107.402	2179.625	$< 9.594 \times 10^{-279}$
S4	9	119362.97	1435.193	$< 9.594 \times 10^{-279}$
S5	9	118435.42	507.6428	5.844×10^{-111}
S6	9	119913.703	1985.925	$< 9.594 \times 10^{-279}$
S7	15	118710.191	782.414	1.262×10^{-170}
S8	15	118193.532	265.755	1.959×10^{-58}
S9	15	119208.097	1280.320	9.594×10^{-279}
S10	15	118237.327	309.550	6.053×10^{-68}
S11	15	117927.777	0	0.999
S12	15	118037.638	109.860	1.393×10^{-24}
S13	11	119162.068	1234.290	9.488×10^{-269}
S14	11	117947.791	20.014	4.508×10^{-05}
S15	11	118826.692	898.915	6.356×10^{-196}
S16	5	120810.632	2882.855	$< 9.594 \times 10^{-279}$
S17	11	119378.842	1451.065	$< 9.594 \times 10^{-279}$

Appendix Table 2.9.8: Direct parameter estimates for best fit models using symmetrical and asymmetrical gene flow from Fastsimcoal2 for *Neotoma fuscipes* in California.

Parameters	Scenario 11 (symmetrical)	Scenario 11 (asymmetrical)
N South	1434802	1449199.5
N North A	585838	634456.5
N North B	3926	11320.5
N ancestral All	7083.5	11095
N ancestral North	3359048.5	3996724.5
T_{DIV} North	94456.5	137851.5
MIG South – North A	2.670E-04	-
MIG North A – South	2.124E-04	1.89E-04
MIG South – North B	8.521E-04	-
MIG North B - South	3.302E-02	1.30E-05
MIG North B – North A	3.346E-01	-
MIG North A – North B	3.015E-05	2.62E-01
MIG North – South	7.087E-04	1.960E-03
MIG South – North	5.700E-02	2.11E-01
T_{DIV} All	1711044	1803740.5

Appendix Table 2.9.9: Re-estimated parameter values from 100 simulated Site Frequency Spectrums (and 95% confidence intervals) for best fit model, scenario 11, from Fastsimcoal2 for *Neotoma fuscipes* in California. See Figure 2 for a graphical depiction of the scenario.

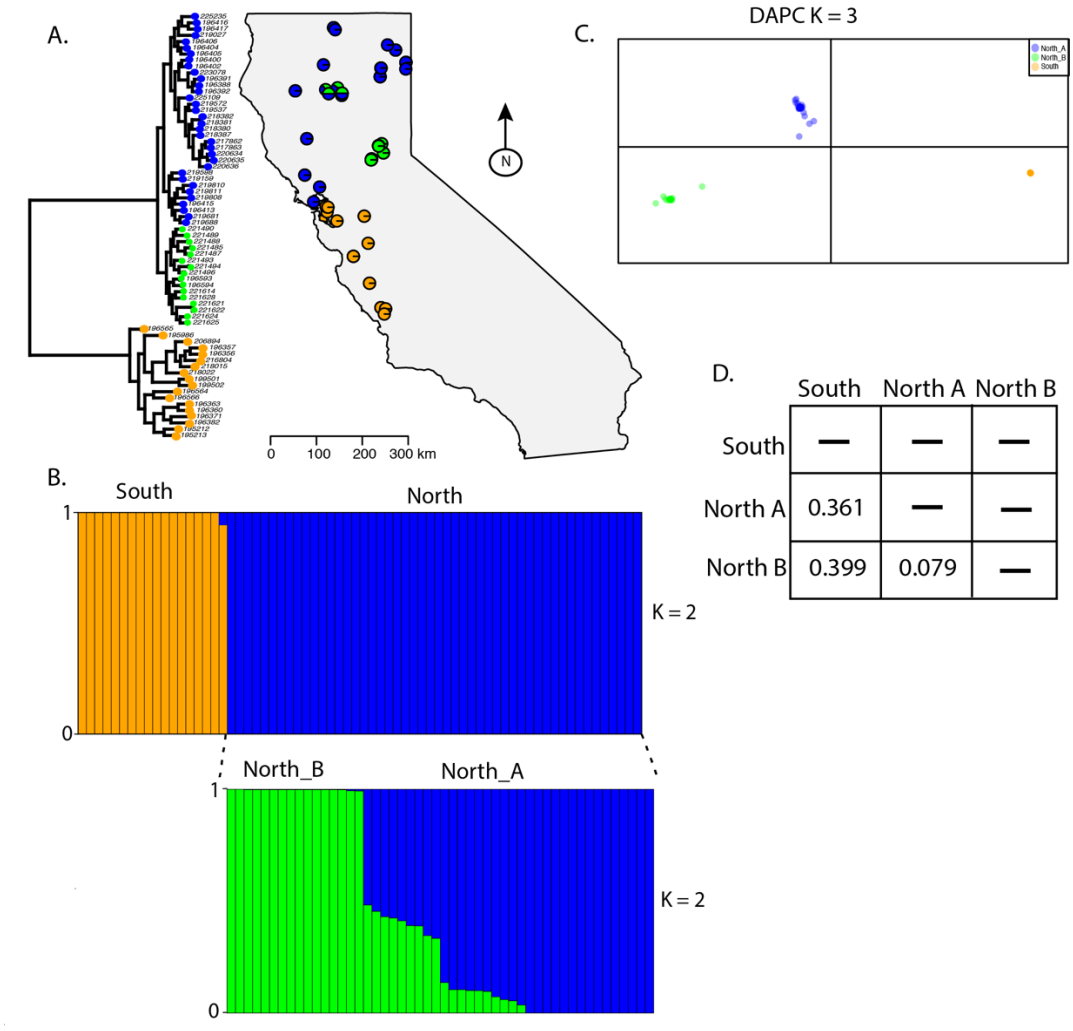
Parameters	Estimates
N South	805,858 (802,421 - 809,295)
N North A	300,307 (289,184 - 312,229)
N North B	4,061 (3,830 - 4,292)
N ancestral All	103,442 (94,944 – 111,939)
N ancestral North	4,414,404 (4,351,869 - 4,476,939)
T_{DIV} North	76,353 (73,063 - 79,644)
MIG South – North A	5.65×10^{-2} (5.47×10^{-2} - 5.85×10^{-2})
MIG North A – South	2.04×10^{-2} (1.74×10^{-2} - 2.34×10^{-2})
MIG South – North B	1.68×10^{-4} (8.47×10^{-5} - 2.51×10^{-4})
MIG North B - South	4.77×10^{-4} (1.54×10^{-4} - 8.03×10^{-4})
MIG North B – North A	2.61×10^{-3} (2.22×10^{-3} - 3.00×10^{-3})
MIG North A – North B	2.02 (4.16×10^{-1} - 3.63)
MIG North – South	9.34×10^{-3} (7.63×10^{-3} - 1.10×10^{-2})
MIG South – North	1.22×10^{-1} (1.21×10^{-1} - 1.23×10^{-1})
T_{DIV} All	1,724,549 (1,714,493 - 1,734,606)

Appendix Table 2.9.10: Summary statistics for the top five performing Maxent models for the combined northern and southern populations of *Neotoma fuscipes* in California. We used the top performing setting for each population to generate the ENM predictions.

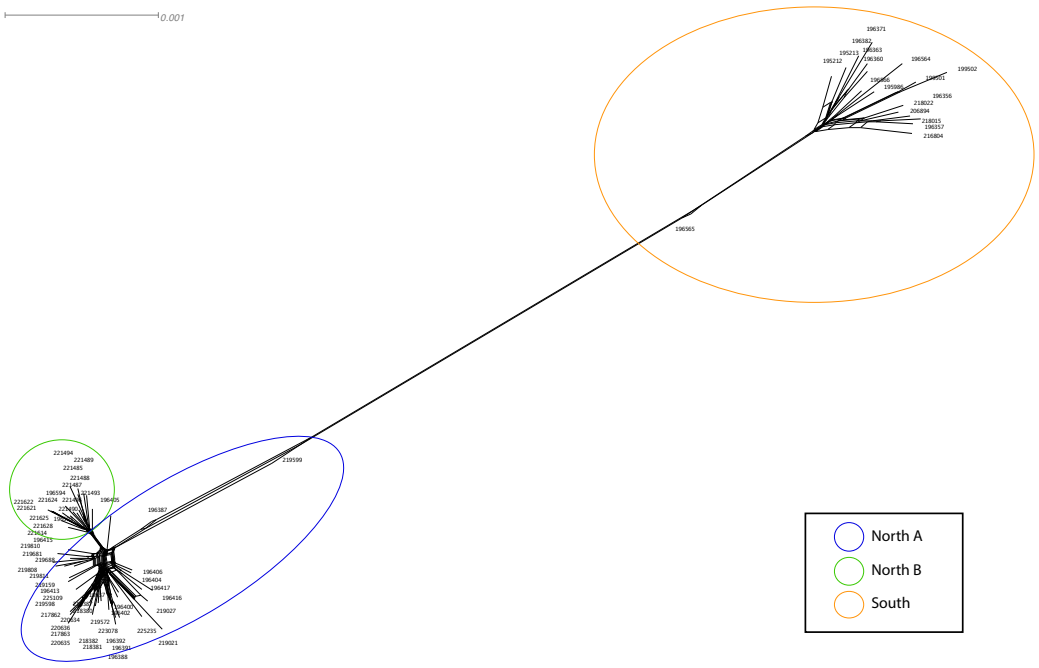
Combined North			South		
Settings	OR_{MTP}	AUC_{test}	Settings	OR_{MTP}	AUC_{test}
H_1	0.045	0.827	LQ_1.5	0.091	0.878
H_1.5	0.045	0.821	L_0.5	0.091	0.877
LQ_0.5	0.045	0.815	LQ_2	0.091	0.876
LQH_1	0.045	0.814	LQ_2.5	0.091	0.872
H_2	0.045	0.813	LQ_1	0.091	0.870

Appendix Table 2.9.11: The asymmetric demographic scenarios using fastsimcoal2 according to AIC for *Neotoma fuscipes* in California.

Scenarios	Parameters	AIC
S18	12	117911.342
S19	12	117887.045
S20	12	117936.246
S21	12	117928.768

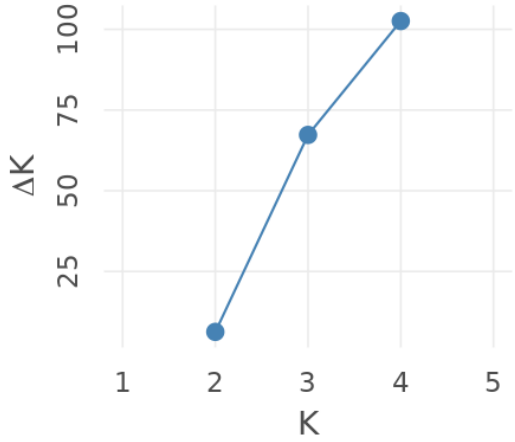


Appendix Figure 2.9.1: The geographic structure of *Neotoma fuscipes* across its range in California based on the 50% missing-data threshold ddRAD matrix. A. Maximum likelihood tree and ADMIXTURE results showing three distinct clusters (*K* = 3): one southern and two northern populations. Colors on the branch tips correspond to the ADMIXTURE population within which individuals were placed, and the colored circles on map indicate the proportion of ancestry each individual has to a population. B. The admixture results from STRUCTURE on the whole population indicating *K* = 2, and the results examining fine scale population structure using only the northern individuals (also indicating *K* = 2). Each line represents one individual depicting the amount of shared ancestry, colors correspond to the phylogenetic and ADMIXTURE results. C. The discriminate analysis of principal components (DAPC) results also showing *K* = 3. D. Pairwise Fst values between the three different clusters, which show much more differentiation between the northern and southern populations than between the two northern populations.

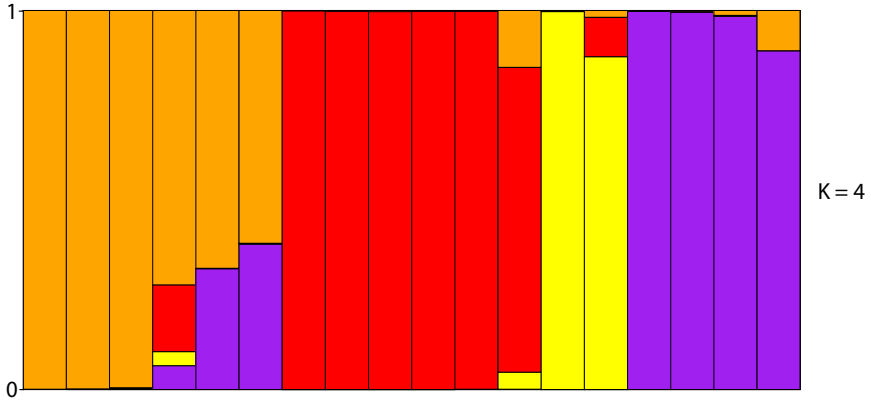


Appendix Figure 2.9.2: The Splitstree network for *Neotoma fuscipes* across its range in California. The northern and southern populations are clearly delineated, with the two northern populations clustered separately as well.

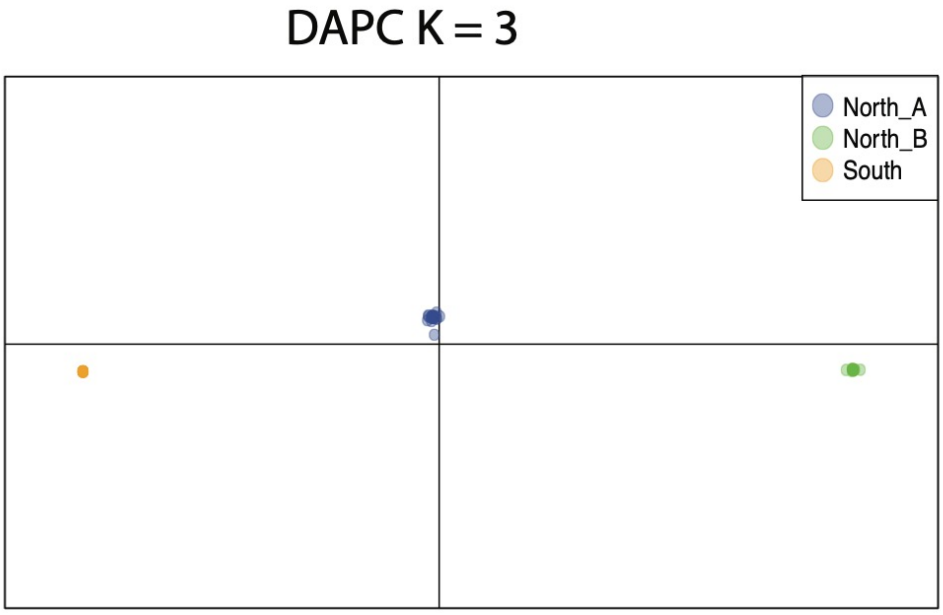
A.



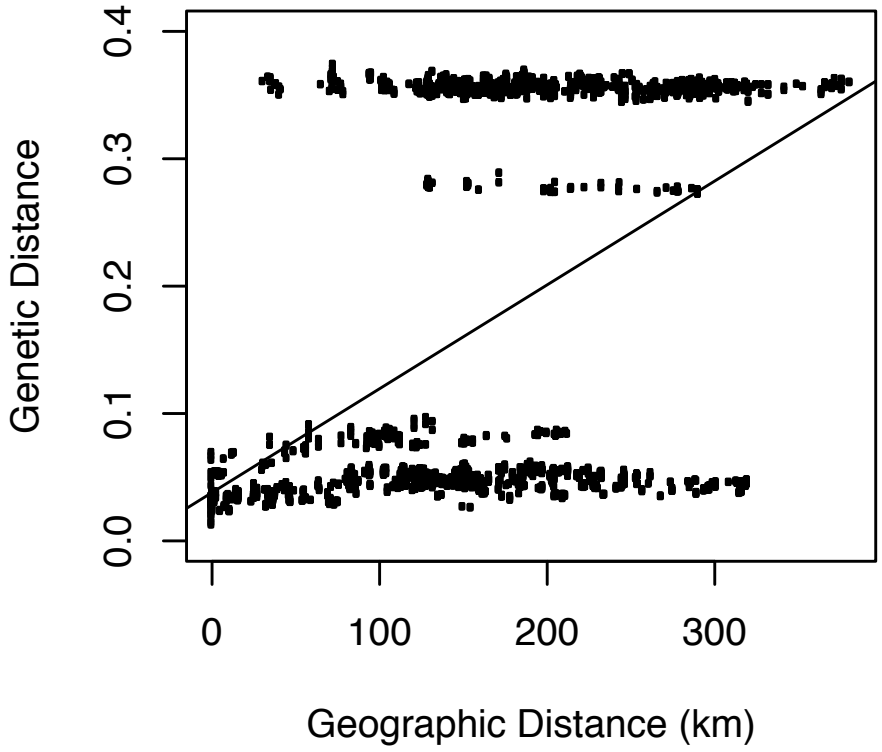
B.



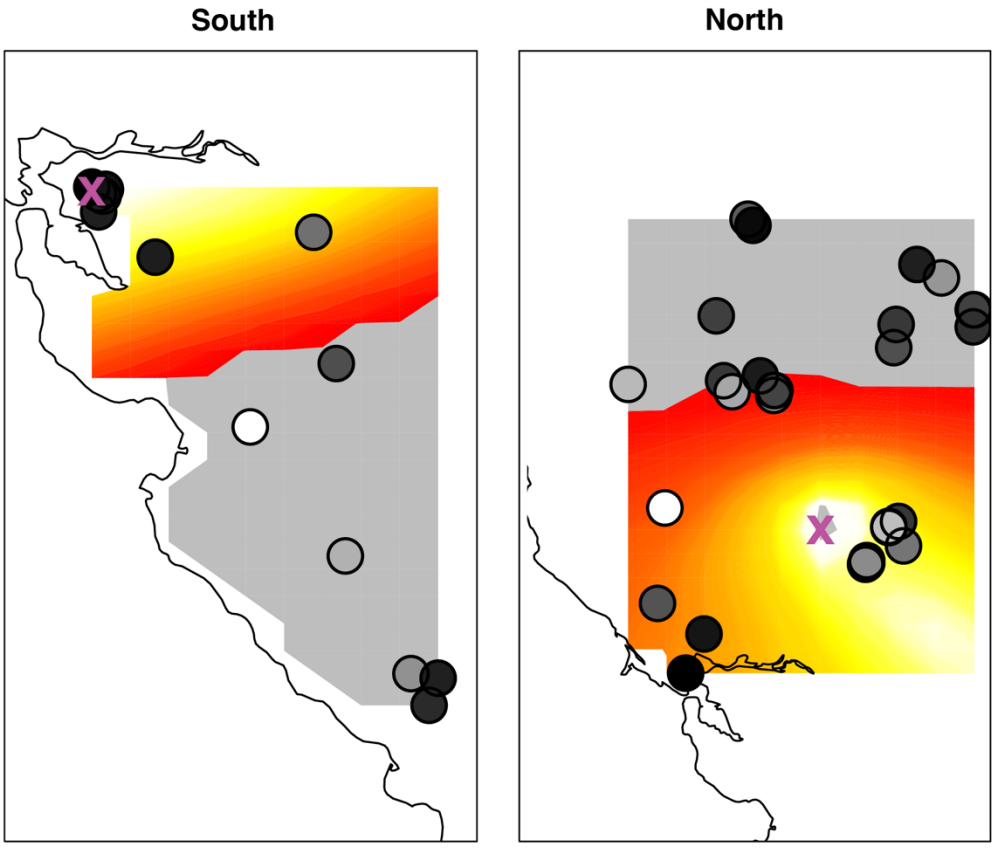
Appendix Figure 2.9.3: The results from STRUCTURE on the southern population of *Neotoma fuscipes*. A. The △K results used to determine the optimal *K* = 4. B. The STRUCTURE plot using *K* = 4. Each line represents one individual depicting the amount of shared ancestry.



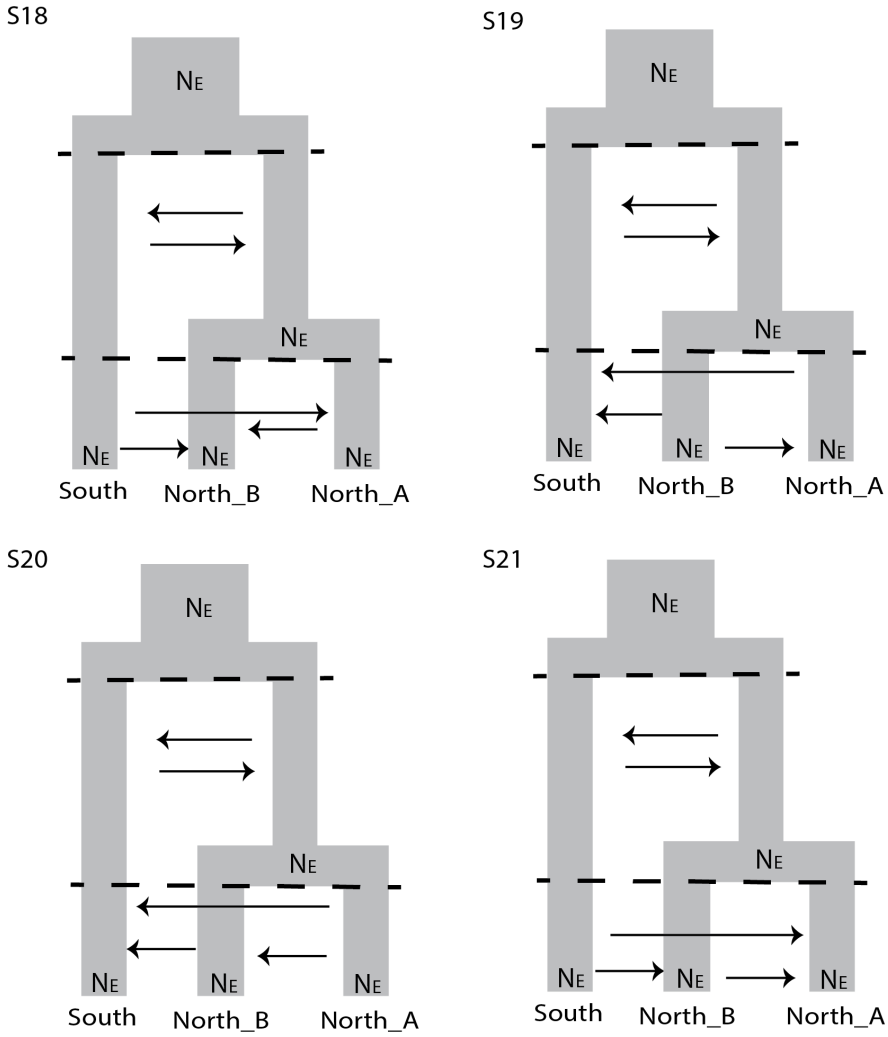
Appendix Figure 2.9.4: The discriminate analysis of principal components (DAPC) using the 85% dataset results showing *K* = 3 on the entire of population of *Neotoma fuscipes* in California.



Appendix Figure 2.9.5: Plot of genetic distance by geographic distance for all sampled individuals of *Neotoma fuscipes* in California. Note that although there is significant isolation by distance ($r = 0.605$, $p < 0.001$), we do not see the typical representative cline indicating historical population structure.



Appendix Figure 2.9.6: The range expansion analyses for the southern and northern populations of *Neotoma fuscipes* in California. The purple X indicates the origin of expansion, a heat map shows areas of likely expansion origin (yellow) to areas of low probability of expansion origin (red). Colors of localities are indicative of heterozygosity: light circles indicate high heterozygosity and dark circles show low heterozygosity.



Appendix Figure 2.9.7: The four asymmetric phylogeographic hypotheses used to simulate demographic history for *Neotoma fuscipes* in California. The parameters included contemporary population size for the three populations, their ancestral population sizes, migration rates between the populations, and the timing of divergence.

Chapter 3: Phylogeography of *Peromyscus maniculatus*: Understanding past distribution of genetic diversity and areas of refugia in western North America

3.1 Abstract

The effects of anthropogenic climate change on biodiversity have been recognized on every continent, ocean, and across different taxonomic groups. Here, we study the range dynamics and demography of a cosmopolitan species: the deer mouse, *Peromyscus maniculatus*. We generated a multilocus SNP dataset using the ddRADseq protocol for 215 individuals across the geographic range within two western North American lineages. We evaluated population structure using several methods and explored the correlation between geographic and genetic distances. We modeled the demographic history using a site frequency spectrum approach and used a machine learning algorithm to infer current and past (last glacial maximum) environmental suitability. Lastly, we explored the origin of population expansion for the identified populations. We identified three regionally distinct populations— 1) southern California; 2) the Pacific Northwest; 3) a broad "Northern" population spanning the Pacific Northwest through central California and across the Rocky Mountains into the Great Plains. Demographic analysis indicated the split between the Pacific Northwest population and the other two was about 160 thousand years ago. ENMs for these lineages predicted suitable environment present throughout the known range of *P. maniculatus* for current conditions, and a severe reduction of northern habitat in the past. The deer mouse has responded to past climate episodes by expanding its range during interglacial periods and contracting its range during glacial periods leading to population differentiation and multiple areas of refugia. By understanding the historical processes that led to the contemporary geographic distribution of biodiversity, we can determine the relative importance of different factors that shape biodiversity, now and in the future.

3.2 Introduction

The effects of anthropogenic climate change on biodiversity have been recognized on every continent, ocean, and across different taxonomic groups (see Parmesan, 2006 for a thorough review of this literature). Given the looming global biodiversity crisis due to climate change, understanding how species will respond is a central question in ecology and evolutionary biology (Dawson et al., 2011). Species will typically respond with one of three outcomes: adapt to changing conditions, track their climatic niche, or go extinct (Parmesan, 2006; Moritz and Potter, 2013). Because conditions are changing quickly (Moritz and Agudo, 2013) and many species are unable to keep pace with these rates of change through either adaptation or dispersal (Warren et al., 2013), the risk of

extinction is elevated (Sinervo et al., 2010). When a species goes extinct, not only do we lose the unique ecological roles played by a species but also all of the genetic information uniquely contained in that species. Broadly distributed, cosmopolitan species have a lower risk of climate change-related extinction risk (Schwartz et al., 2006), but even for these species, population extirpation could lead to a reduction of genetic diversity and constrain the adaptive potential of the species (Sexton, 2019). It is thus imperative to understand how populations have responded to past climate change.

Populations within a species are the units distributed across space and time that are subject to different constraints and opportunities (O'Neill et al., 2008). Determining how genetic diversity within a species is distributed across the landscape is crucial because the structure of this diversity is a fundamental to lineage divergence and speciation (Frankham et al., 2002; Yannic et al., 2014), adaptive evolution (Sexton, 2019), and to probability of extinction (Hughes et al., 1997; Urban, 2018). To fully understand the impacts of climate change on biodiversity we need to recognize how it will affect intraspecific variation. Climate change can influence the intraspecific distribution of genetic diversity through alterations to the ranges of historically significant lineages; alterations to phenotypic plasticity of individuals in different populations; and/or local adaptations to novel environmental conditions (Pauls et al., 2013). These changes in population structure usually reduce genetic diversity for many species (Bálint et al., 2011; Pironon et al., 2017). Furthermore, the responses of populations to recent, human-induced climate change are taking place immediately following the loss of intraspecific genetic diversity due to climate change over the last 21,000 years and longer (Hewitt, 2000; Magyari et al., 2011; Miraldo et al., 2016).

To truly understand how intraspecific genetic diversity has been influenced by climate and landscape, we need to integrate our understanding of dynamic climate change through time with spatial landscape processes (Knowles and Alvarado-Serrano, 2010; Alvarado-Serrano and Knowles, 2014). Regions that should be of particular interest are climate refugia (termed refugia hereafter), local areas with the appropriate environmental conditions to allow species to persist (Keppel et al., 2012). Identifying the geographic locations of refugia, and how species contracted into and expanded out of those refugia, is biologically important (Keppel et al., 2012; Gavin et al., 2014): these refugial areas could harbor greater genetic diversity and influence the ability of a species to adapt to new environmental changes because this capability is dependent on their historical population structure (Dynesius and Jansson, 2000). Thus, identifying refugial regions for single species, and more importantly, identifying refugial locations shared by multiple species, could be highly relevant to conservation efforts; refugia could indicate areas that are climatically stable areas throughout time and support higher genetic diversity even in unsampled species (Carnaval et al., 2009). By studying the dynamic climatic oscillations of the Pleistocene,

cooler glacial periods followed by warmer interglacial periods (Dansgaard et al., 1993), we can begin to understand the processes that determine genetic diversity across the landscape.

One species within the genus *Peromyscus*, *P. maniculatus* (the deer mouse), represents a good study system for addressing the impacts of climate-mediated historical population distribution and refugia on genetic diversity. *Peromyscus maniculatus* is indigenous to and distributed widely across North America (Shorter et al., 2012) and can be found in every terrestrial ecosystem (Dewey and Dawson, 2001), and thus is exposed to different environmental and climatic conditions across its geographic range. Based on mitochondrial DNA, *P. maniculatus* is likely a species complex, with six hypothesized monophyletic lineages currently: 1) Northeastern USA and Canada; 2) Northeastern USA and central Canada; 3) USA central plains; 4) Southeastern USA; 5) New Mexico, Arizona and Northern Mexico; 6) Canada and USA Pacific Northwest, California and as far east as Kansas, USA (Dragoo et al., 2006; Kalkvik et al., 2012). Additionally, there are several other recognized species that cluster within *P. maniculatus*: *P. keeni* (Rhoads, 1894; Zheng et al., 2003*)*, *P. polionotus* (Osgood, 1909; Hall, 1981), and *P*. sp (Lucid and Cook, 2007). *Peromyscus polionotus* occupies southeastern USA (the one area *P. maniculatus* is not found), and both *P. keeni* and *P.* sp occupy the Pacific Northwest in North America.

It is important to understand the evolutionary history and population dynamics of *Peromyscus maniculatus* because they are vectors for several diseases, such as hantavirus and Lyme disease (Dragoo et al., 2006), and potentially SARS-CoV-2 (Fagre et al., 2020). But most relevant here, different populations of this species have already exhibited different responses to recent climate change, demonstrating potential for population-level variation that may have long-term effects. A population of deer mice in Yosemite National Park has not shifted its distribution in response to recent environmental change and has maintained a similar range over the last 100 years (Moritz et al., 2008), though there may have been a cryptic geographic shift in genetic variation (Yang et al., 2011). In contrast, a population of *P. maniculatus* has shifted its distributional range in Michigan (moving northward; Myers et al., 2009), demonstrating that climate has likely also influenced this population in the recent past. Populations of *P. maniculatus* at different elevations exhibit different physiological performances and show different patterns of gene expression, which may be due to differences in local population structure and adaptation (Storz et al., 2019). Thus, changes to gene flow and population structure through time as a result of climate may have direct fitness consequences for different populations of deer mice.

Here, we focus on the ecological and evolutionary processes that have structured genetic diversity within two western North American lineages of

Peromyscus maniculatus (Fig. 1). We generated a multilocus SNP dataset on 215 individuals to answer several questions: 1) Do we recover the two mtDNA lineages found in Dragoo et al. (2006) and Kalkvik et al. (2012) using a nuclear genomic dataset? 2) What were the potential areas of refugia for *P. maniculatus,* given that its contemporary range encompasses both previously glaciated and unglaciated regions, throughout western North America? It has been suggested that *P. maniculatus* has only inhabited areas south of the continental ice sheets during the Last Glacial Maximum (Dragoo et al., 2006; Sawyer et al., 2017), and so we determine if the genomic data provides evidence for refugia in the Pacific Northwest coast or Beringia. We first detect the historical population structure found within these two lineages and then use several different analyses to detect possible areas of refugia throughout North America.

3.3 Materials and Methods

3.3.1 ddRAD library preparation and sequencing:

Genomic DNA was extracted from 215 individuals of *Peromyscus maniculatus (*and three individuals of *P. californicus* to serve as an outgroup) from 32 localities across the western North American range (Fig. 3.1B; Appendix Table 3.1) using a Qiagen extraction kit (Qiagen Inc.). Samples were obtained from fieldwork and several museum collections: Burke Museum (*n* = 30), Field Museum (*n* = 33), Museum of Southwestern Biology (*n* = 25), Museum of Vertebrate Zoology (*n* = 75), and University of Alaska Museum of the North ((*n* = 21). Additionally, we received tissues from 10 individuals from Dr. Dean Pearson. We generated genomic SNP datasets using the ddRAD protocol (Peterson et al., 2012) and as detailed in Boria et al. (2020). Briefly, we digested 500ng of genomic DNA with restriction enzymes *EcoR1* and the rare-cutting enzyme *MSP1*. We pooled and purified the tagged samples (12 individuals per pool) and used a Pippin Prep to size-select for fragments between 300-500 bp long. Each library was amplified via PCR with a second index primer to differentiate between all individuals. Finally, the libraries were sequenced at UC Berkeley on two lanes of an Illumina HiSeq 4000 to produce 50-bp single end reads.

We cleaned and filtered the Illumina HiSeq raw reads using the iPyrad pipeline (v 0.9.65; Eaton and Overcast, 2020). Raw reads were aligned to a draft genome of *Peromyscus maniculatus* (NCBI: GCA 000500345.1). We used a 75% threshold as the minimum number of individuals needed to retain a locus. Following read clustering, we excluded 25 individuals that had greater than 70% missing data. Additionally, four individuals clustered with outgroup specimens in several preliminary analyses and were excluded from any downstream analyses because they may have been misidentified as *P. maniculatus* upon original

collection. We retained 186 individuals for analyses: 183 *P. maniculatus* and 3 *P. californicus*. Finally, we further filtered the dataset to include only one biallelic SNP per locus.

3.3.2 Population structure:

We used two model-based and one non-model-based approaches to investigate population structure. All methods attempt to determine the optimal number of populations present within the dataset. One model-based approach used to determine the optimal number of populations was Admixture (v. 1.3; Alexander *et al.*, 2009), a maximum likelihood method. We used cross-validation error values across different values of K to determine the optimal number of populations. Another model-based method we ran was Structure (v 2.3.4; Pritchard *et al.*, 2000), a Bayesian-clustering algorithm that identifies population structure. We ran it for 50,000 burn-in generations, 250,000 generations, K = 1-5 populations, and for five iterations. We used the Evanno method (Evanno et al., 2005) in Structure Harvester (v. 0.6.94; Earl and vonHoldt, 2012) to determine the optimal number of populations. Because Structure Harvester may not accurately infer the correct number of groups, typically inferring the uppermost level of genetic structure (often K = 2; Janes et al., 2017), we used a hierarchical approach, and re-ran Structure on the largest initial population identified (Pritchard et al., 2000). Finally, we used the non-model based method DAPC (Discriminate analysis of principal components) in the Adegenet package (Jombart, 2008; Jombart and Ahmed, 2011) in R (v. 3.5.1; R core team, 2020) to infer structure. For the DAPC analysis, we used Bayesian information criterion to determine the optimal number of populations. Once we identified the optimal number of populations, we calculated pairwise Fst values between each population using the hierfstat R package(v. 0.1.17; Danecek *et al.*, 2011).

To determine if the patterns indicated by the population structure programs were influenced by historical population structure and not isolation by distance (IBD), we ran a Mantel test on the genetic and geographic distances between individuals. We examined the Mantel test significance with 10,000 permutations in the Adegenet package (Jombart, 2008; Jombart and Ahmed, 2011). Further, we plotted the genetic and geographic distances among all individuals to determine if there was a distinct cline indicative of IBD or if separate patches of individuals were representative of distinct populations. Lastly, we assessed significance of IBD within each of the populations identified by the structure analyses.

3.3.3 Phylogenetic analyses:

We inferred phylogenetic relationships between individuals using two methods: 1) SVDquartets (Chifman and Kubatko, 2015); 2) RAxML (Stamatakis, 2014). We implemented SVDquartets, using a multi-species coalescent model, in

PAUP* (v4.0a; Swoford, 2003) utilizing all quartets and 100 bootstrap replicates. We also used RAxML (v. 8.2.12;), a maximum likelihood (ML) algorithm for large datasets. We used rapid bootstrapping (with an automatic stop under the autoMRE criterion), and searched for the best tree under the GTRGAMMA model with final optimization of trees using GTR+Γ. Note, this analysis is not appropriate for many phylogenetic questions involving intraspecific SNP data; however, we used this method to narrowly understand how individuals were generally grouped together, rather than infer the specific relationships among individuals (following Harrington *et al*., 2017).

3.3.4 Demographic modeling:

We used a coalescent-based approach to model the demographic history of *Peromyscus maniculatus* using a temporal framework, specifically with respect to isolation, population growth, and migration. We selected and parameterized the best-fit demographic model using fastsimcoal2 (FSC2, v2.5.2; Excoffier *et al*., 2013), which estimates demography from the site frequency spectrum (SFS). Users specify hypotheses with differing levels of complexity and fastsimcoal2 uses simulations to estimate the likelihood of competing hypothesis with different parameters. Individuals were assigned to populations using a 3-population model based on the population structure analyses (see above), with admixed individuals assigned to the majority population. We generated the observed joint SFS using code developed by Isaac Overcast (https://github.com/isaacovercast/easySFS).

We tested 21 different demographic scenarios (Fig. 3.2). Briefly, they describe variations of a 3-population model that have two divergence events with every possible topology and vary in the amounts of migration and timing between populations. Additionally, we used a nuclear mutation rate of 2.5×10^{-8} mutations/site/generation based on estimates from human loci (Nachman and Crowell, 2000) and used a one-year estimated generation time. For each model we ran 75 replicate fastsimcoal2 analyses, each using 250,000 coalescent simulations. We selected the best fit model by calculating AIC and $\triangle$AIC scores to account for the differing number of parameters, following the guidelines of Excoffier et al. (2013). To ensure the simulations were not only sampling parameter estimates from a local maximum, we ran five replicates for each model starting with different initial conditions and chose the model with the highest likelihood score. Once the best model was obtained, we re-estimated parameter values by simulating 100 SFS from the maxL.par file and determining the mean parameter estimate and 95% confidence intervals. We accounted for *P. maniculatus* being diploid by dividing the fastsimcoal2 parameter estimates in half.

3.3.5 Ecological niche modeling:

We generated ecological niche models (EMNs) to predict areas of refugia. Occurrence localities were compiled from the fieldwork and museum specimens used in this study (see above, Appendix Table 3.X.1). Additionally, we supplemented these data with localities that had associated genetic data from Kalkvik et al. (2012) and Sawyer et al. (2017). To remove spatial biases, we spatially filtered the dataset to ensure no two localities were within 50 km of one another (Boria et al., 2014) using the R package spThin (Aiello-Lammens et al., 2015). For environmental data, we used Community Climate Simulation Model 3 (CCSM3; Liu et al., 2009). CCSM3 variables are downscaled to 50 x 50 km degree grid cells for North America from 21,000 years ago to the present day (Lorenz, Nieto-Lugilde, Blois, Fitzpatrick, & Williams, 2016). To approximate modeling assumptions regarding dispersal and biotic interactions more closely, we delimited a custom study region by drawing a minimum convex polygon around the localities and adding a 5.0° buffer (Anderson and Raza, 2010; Barve et al., 2011). We used a machine learning algorithm, maxent (V3.4.1; Phillips *et al*., 2017; Phillips *et al*., 2006) to infer the ENMs. We calibrated and evaluated the models using a geographically structured 5 *k-fold* approach (Radosavljevic and Anderson, 2014) in the ENMeval package in R (Muscarella et al., 2014). To select species-specific model settings approximating optimal levels of complexity, we tuned model settings by varying different combinations of feature class and regularization multiplier (RM; Shcheglovitova and Anderson, 2013). To identify the optimal parameter settings, we evaluated model performance using sequential criteria (minimizing overfitting and then maximizing discriminatory ability; Shcheglovitova and Anderson, 2013; Muscarella *et al*., 2014). We used the optimal settings to project the ENM into current climatic conditions and the LGM.

3.3.6 Range Expansion:

We inferred origins of recent range expansion within the three identified populations: 1) Pacific Northwest (PNW); 2) Northern; 3) southern California (see *Results*). For the Northern population, we calculated the origin of expansion two ways: 1) we included all individuals in the northern population; 2) we excluded the northernmost locality (in Canada) because it is one of the few highly supported groups within both phylogenetic analyses (Fig. 1; SFig. 5). We used the rangeExpansion package in R (Peter and Slatkin, 2013). Briefly, this method detects range expansion and gives an estimated location of the origin of expansion. It does so by calculating a directionality index (ψ), using SNP datasets and spatial coordinates, based on allele frequency clines found between multiple populations (Peter and Slatkin, 2013). Populations at the expanding edge will tend to have lower genetic diversity because of serial founder events and higher fixation rates (Peter and Slatkin, 2015).

3.3.7 Interpolating genetic diversity:

We calculated heterozygosity on a per-individual basis using vcftools (v. 0.1.17; Danecek et al. 2011), and then averaged all scores across all sampling sites. We interpolated heterozygosity across evenly spaced 50km grid cells to visualize genetic patterns across the landscape. We used the inverse distance weighting interpolation procedure in the R package phylin (v 2.0.2; Tarroso et al., 2015) and projected the heterozygosity values across the ENM study region.

3.4 Results:

3.4.1 Sequence data:

We obtained 482,988,240 raw reads from the Illumina HiSeq 4000 runs across all individuals, comprising 551,767 prefiltered loci. After iPyrad quality filtering, 9,640 loci were retained with 69,726 SNPs across loci. Following filtering for one biallelic SNP per loci, we retained 9,535 SNPs.

3.4.2 Population structure and isolation by distance:

Admixture indicated the optimal population model based on cross validation was *K* = 3: 1) southern California; 2) the Pacific Northwest (PNW hereafter); 3) a broad “Northern” population spanning the Pacific Northwest through central California and across the Rocky Mountains into the Great Plains (Northern hereafter; Appendix Figure 3.9.1A; Appendix Table 3.9.2). Structure initially showed a *K* = 2 (Northern + southern California and PNW) model best fit the data (Appendix Figure 3.9.2A), however, a *K* = 3 model provides the same population structure as Admixture and DAPC. Further, when we ran Structure on the large subpopulation it split into the Northern and southern California populations (Appendix Figure 3.9.2B, C). DAPC analysis identified *K* = 3 as the optimal number of populations (Appendix Figure 3.9.1B). All methods assigned the same individuals to the same population across analyses. Pairwise F_{ST} showed strong differentiation between the PNW and both the Northern (F_{ST} = 0.22) and southern California (F_{ST} = 0.28) groups and moderate differentiation between the Northern and southern California populations (F_{ST} = 0.06). We adjusted downstream analyses to reflect the population structure results, *K* = 3.

The Mantel test showed significant isolation by distance for the overall dataset (r= 0.614, p<0.001) with a distinct cline (Appendix Figure 3.9.3). We also discovered significant IBD for each subpopulation with distinct clines (Appendix Figure 3.9.4): Northern: r = 0.469 (p<0.001); southern California: r = 0.266 (p<0.001); PNW: r = 0.917 (p<0.001).

3.4.3 Phylogenetic analyses:

Both SVDquartets and RAxML produced weakly supported phylogenetic trees (Fig. 3.1; Appendix Figure 3.9.5), with different hierarchies of clustering among individuals and populations. All individuals are clustered within the Northern population for SVDquartets; conversely, all individuals are clustered within the southern California population, for RAxML. Although the trees were weakly supported, there were several well supported clusters of individuals (bootstrap support > 50) across both analyses. The PNW population was very well supported in both analyses and always clustered within the Northern population (Fig. 3.1; Appendix Figure 3.9.5). All individuals in the southern California population clustered together (and some relationships were well supported) in both analyses but the overall group was weakly supported. Another group of individuals that were very well supported in both analyses were 7 individuals from the same locality in Canada, which were a part of the Northern population (Fig. 3.1).

3.4.4 Demographic modeling:

Using FSC2, the model that best fit the data according to AIC was scenario 15 (Table 3.1; Appendix Table 3.9.3; Fig. 3.1). This involved the Northern and southern California populations coalescing most recently, with an older coalescent event between the Northern/ southern California cluster and the PNW population (Fig. 3.3; see Appendix Table 3.9.4 for confidence intervals). There were several other scenarios that had a reasonably good fit, however (scenarios 1, 10, 13, 14, and 19: delta AIC <5; Table 3.1). They all had the same sequence of coalescent events, but with recent migration events only between the northern and southern California populations, and different population expansion scenarios (see Appendix Table 3.9.5 for direct parameter estimates for these models).

Scenario 15 estimated the Northern and southern California populations coalescing about 78 ka (60,747– 95,799) and the ancestral Northern/southern California lineages coalescing with the PNW population about 158 kya (136,945 – 178,201; Fig. 3.3). All three populations had moderate to large effective population sizes (> 50 thousand individuals), with the PNW group having a recent population expansion event.

3.4.5 Ecological niche modeling:

The optimal Maxent settings for *Peromyscus maniculatus* was the Linear, Hinge, and Quadratic features with a RM = 6.0 (Appendix Table 3.9.6). This model had an AUC of 0.71 and an omission rate of 0.12. ENM for *P. maniculatus* inferred suitable environment present throughout the entire known range of these

lineages for current conditions, and a severe reduction of northern habitat south of the ice sheets during the LGM relative to the contemporary range (Fig. 3.4A, B).

3.4.6 Range expansion:

For the whole “Northern” population, the range expansion analysis indicated the origin of expansion was in northwest British Columbia, Canada (Fig. 3.4B, C) although the results were not significant ($p>0.05$). When we removed the Canadian locality, the range expansion analysis indicated the origin of expansion for the Northern population was in the southwest United States of America; however, these results were again not significant ($p>0.05$; Fig. 3.4B, C). The origin of expansion for the southern California population was the San Francisco Bay area ($p<0.001$), and for the PNW population it was Southeast Alaska, the Heceta Islands ($p<0.001$; Fig. 3.4B, C).

3.4.7 Interpolating genetic diversity:

The interpolation analysis indicated the region with the highest genetic diversity across all localities was in the Pacific Northwest (Fig. 3.4C). There is also a smaller area of high diversity in Southern California, and a region of moderate diversity on the coast of central California to southern Oregon. Areas with the least diversity were across the Rocky Mountains and the Great Plains (Fig. 3.4C).

3.5 Discussion:

The spatial distributions of intraspecific genetic diversity and the role of climate refugia shapes the evolutionary and ecological processes of populations and possibly determines the potential for population and species persistence in the face of future climate change (Pauls et al., 2013; Yannic et al., 2014; Brown et al., 2016). By using a multilocus SNP dataset combined with ENMs and demographic modeling, we were able to determine the dynamic history of several lineages within *Peromyscus maniculatus* across western North America. Here, we show how three populations of *P. maniculatus* have responded to past climatic shifts by contracting and expanding its range over the last 200,000 years, and the influence of those changes on overall genetic diversity across the species.

3.5.1 Impacts of glacial cycles on population structure:

The Pleistocene glacial-interglacial cycles have impacted the distribution of genetic diversity (Hewitt, 2004; Hope et al., 2015; Burbrink et al., 2016) and led to cycles of population fragmentation and merger in many North American species (Arbogast, 2007; Malaney et al., 2013; Jezkova et al., 2015, 2016;

Massatti and Knowles, 2016; Sim et al., 2016; Reid et al., 2019; Boria et al., 2020). All population-level analyses indicate there is significant population structure within *Peromyscus maniculatus.* These analyses support a three-population model—1) southern California; 2) a small population nested within the broader Pacific Northwest [PNW]; 3) the Pacific Northwest through central California, across the Rocky Mountains and through the Great Plains (Fig. 3.1; Appendix Figure 3.9.1; [Northern]). Although these lineages have been previously sequenced with only a few loci, overall population structure was based solely on phylogenetic trees (see future direction for further discussion on deer mouse phylogenetics; Dragoo et al., 2006; Kalkvik et al., 2012; Sawyer et al., 2017) and detailed population structure was unknown. However, we did recover the Northern and southern California populations similar to lineages One and Six (respectively) in Kalkvik et al. (2012) and a PNW population that is similar to a lineage described as "*P. manicualtus* west" in Sawyer et al. (2017).

The demographic analysis based on the SFS indicated the divergence between the PNW population and the lineage leading to the Northern/southern California populations dates back to about 160 thousand years ago (Fig. 3.3), which is very similar to previous estimates (Sawyer et al. 2017). This timing coincides with Marine Isotope Stage (MIS) 6 glacial period (Lisiecki and Raymo, 2005; Kawamura et al., 2007). *P. maniculatus* likely expanded from a refugium in the southern part of North America as temperatures warmed during the MIS 7 interglacial (approximately 243 kya; Kawamura et al., 2007). None of the demographic scenarios included any migration between the PNW and either of the other two populations (Table 3.1), and there is also very high population differentiation between the PNW and Northern and southern California populations. Together, this indicates that the three populations remained isolated enough through the proceeding glacial and interglacial periods to maintain significant population differentiation. Lastly, the PNW population has experienced a recent population expansion event (Fig. 3.3) coincident with genetic evidence in the pacific northwest (Sawyer et al. 2017) and fossil evidence from northern California (Blois et al., 2010) with this species, and also very similar to a pattern of recent expansion events seen in North American snakes (Burbrink et al., 2016).

The Northern and southern California populations diverged much more recently, about 78 kya during the MIS5 interglacial-glacial transition (Fig. 3.3; Lisiecki and Raymo, 2005; Kawamura et al., 2007). This timing estimate is slightly younger than previous estimates (Sawyer et al., 2017) and this time period corresponds to a highly variable, relatively warm interglacial substage of MIS5 (MIS5a). These two populations are much less differentiated and several demographic scenarios having a reasonably good fit included migration between these two groups (Table 3.1). Although the best fit model did not include recent population expansion, three of the top six models included recent expansion events for both populations (Table 3.1, Fig. 3.2).

3.5.2 Potential areas of refugia and expansion:

During the glacial phases of the Pleistocene, colder climate and the physical formation of large glaciers that covered most of northern North America (Dansgaard et al., 1993; Kleman et al., 2010) forced species to retreat to refugia, local areas where species were able to persist (Stewart et al., 2010). The exact nature of climate and landscape change was variable across regions, but within a region species were broadly exposed to similar environmental conditions. Yet, evidence suggests many species responded individualistically to these environmental changes (Jackson and Overpeck, 2000). Typically in western North America, species persisted in refugia located in: 1) Beringia; 2) Pacific Northwest coast; 3) one of the many potential regions south of the continental ice sheets (see Waltari et al., 2007; Pielou, 1991; Fleming and Cook, 2002; Eddingsaas et al., 2004; Sawyer et al., 2017), with almost no mammal species occupying all three regions (for a possible exception see Dawson et al., 2014). Here, we provide evidence that *Peromyscus maniculatus* responded to Pleistocene climate changes by expanding its range during interglacial periods and contracting its range during glacial periods (Fig. 3.4), occupying at least two of the main refugial areas in the western United States — 1) Pacific Northwest coast; 2) south of the continental ice sheets.

Previously, *Peromyscus maniculatus* was thought to have only occupied regions south of the continental ice sheets during glacial periods and expanded north only during the interglacial periods (Dragoo et al., 2006), with other *Peromyscus* species inhabiting the Pacific Northwest coast (*P. keeni*; Lucid and Cook, 2004) and Beringia (*Peromyscus* sp; Sawyer et al., 2017) glacial refugia. However, previous studies only used mitochondrial DNA (e.g., Dragoo et al., 2006), or mitochondrial DNA with four nuclear loci (Sawyer et al., 2017). Here, we have assembled the largest multilocus SNP dataset to date for these populations. We show that at least one population of *P. maniculatus*, PNW, had a glacial refugium located on the Pacific Northwest coast (Fig. 3.4). The ENMs indicate moderate suitability during the LGM in this region (Fig. 3.4B), and range expansion analysis that identifies the likely origin of expansion as the Heceta Islands, Alaska (Fig. 3.4C), an area that was possibly accessible during the LGM when sea levels were much lower in the region (Mann and Hamilton, 1995). The map of interpolated heterozygosity indicate the regions of highest diversity are in the Pacific Northwest, providing further evidence the PNW lineage has been inhabiting stable environmental conditions through the glacial-interglacial cycles and accumulating genetic diversity (Fig. 3.4C; Carnaval et al., 2009).

The much larger Northern population likely had multiple refugia during glacial periods. During the LGM, there was a severe reduction in the suitability of northern regions south of the continental ice sheets, with moderate suitability in the Pacific Northwest and high suitability from northern California to areas further

to the south (Fig. 3.4). The range expansion analysis for the entire Northern population reveals a possible origin of expansion in northwest British Columbia, Canada (near the southeast Alaska border) although the results were not significant. This location is similar to the “minor refugium” found in other species (e.g., Sim et al., 2016; reviewed by Shafer et al., 2010). The most northern locality within this group is in northwest British Columbia, Canada and it has the highest level of diversity (Fig. 3.1; Fig. 3.4), indicative of a refugium for the Northern population. If the Canadian individuals are removed, the origin of expansion for the Northern lineage is further south in Arizona, USA (Fig. 3.4C), though again the results were not significant. This region was predicted as highly suitable during the LGM, although it is a region with low genetic diversity (Fig. 3.4C), which is contrary to expectations for a refugium.

Much of the contemporary distribution of the southern California population was also highly suitable during the LGM (Fig. 3.4B). The origin of expansion for this lineage was near the San Francisco Bay-Delta region, a known lineage diversity hotspot in California for reptiles, amphibians (Rissler et al., 2006), and possibly mammals (Boria et al., 2020). This region has a moderate level of genetic diversity (Fig. 3.4C), possibly indicating that the San Francisco Bay-Delta region served as a refugium as well. There is also a hotspot of genetic diversity and high suitability during the LGM in Baja California, Mexico; however, greater sampling depth for the southern California population (specifically along the Baja peninsula) is needed to determine if this could have also been a refugium.

3.5.3 Future directions:

Peromyscus maniculatus likely represents a species complex with several described species within the clade. Highly supported phylogenetic trees show six monophyletic lineages within *P. maniculatus* (see Dragoo et al., 2006; Kalkvik et al., 2012); however, these studies only relied on mitochondrial DNA. Recently, Sawyer et al. (2017) used mitochondrial DNA and four nuclear loci with an emphasis on assessing the relationships of *Peromyscus* species in the Pacific Northwest. This study produced a mostly unsupported tree, similar to the results found here, likely due to incomplete lineage sorting (Sawyer et al., 2017). Much more data (genomic and morphological, across all geographic regions and not just western North America) is needed to fully resolve the lineage relationships within *P. maniculatus*. Furthermore, although we are confident the PNW population identified here is clustered firmly within *P. maniculatus*, it is possible that additional data would show the group we identified is *P. keeni*. More in-depth genomic sampling for this species is needed to understand the dynamic, population-level processes in the Pacific Northwest.

Site frequency spectrum simulations are an extremely valuable tool, albeit one with some limitations (Hickerson, 2014). Specifically, although we choose a

range of different models, we are only exploring a small sample of parameter space, and the demographic history of *P*. *maniculatus* is probably much more complex. For example, there were several models with a reasonably good fit, and they all had very similar point estimates for demographic parameters (Appendix Table 3.9. 5). Additionally, the mutation rate assumed in these models are based on data from humans and could impact the inferences presented here. Lastly, the demographic analyses would also be improved with greater sampling so that we could model the history for each individual population as well. Nonetheless, genome-wide datasets and the potential to exploit SFS patterns is a huge advance in the field of population genetics and these data and methods will continue to provide insight into complex biogeographic histories.

3.6 References:

Aiello-Lammens ME, Boria RA, Radosavljevic A, Vilela B, Anderson RP (2015). spThin: An R package for spatial thinning of species occurrence records for use in ecological niche models. *Ecography* **38**: 541–545.

Alexander DH, Novembre J, Lange K (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* **19**: 1655–1664.

Alvarado-Serrano DF, Knowles LL (2014). Ecological niche models in phylogeographic studies: applications, advances and precautions. *Mol Ecol Resour* **14**: 233–248.

Anderson RP, Raza A (2010). The effect of the extent of the study region on GIS models of species geographic distributions and estimates of niche evolution: preliminary tests with montane rodents (genus Nephelomys) in Venezuela. *J Biogeogr* **37**: 1378–1393.

Arbogast BS (2007). A Brief History of the New World Flying Squirrels: Phylogeny, Biogeography, and Conservation Genetics. *J Mammal* **88**: 840–849.

Bálint M, Domisch S, Engelhardt CHM, Haase P, Lehrian S, Sauer J, et al. (2011). Cryptic biodiversity loss linked to global climate change. *Nat Clim Change* **1**: 313–313.

Barve N, Barve V, Jiménez-Valverde A, Lira-Noriega A, Maher SP, Peterson AT, et al. (2011). The crucial role of the accessible area in ecological niche modeling and species distribution modeling. *Ecol Model* **222**: 1810–1819.

Blois JL, McGuire JL, Hadly EA (2010). Small mammal diversity loss in response to late-Pleistocene climatic change. *Nature* **465**: 771–774.

Boria RA, Brown SK, Matocq MD, Blois JL (2020). Genome-wide genetic variation coupled with demographic and ecological niche modeling of the dusky-footed woodrat (Neotoma fuscipes) reveal patterns of deep divergence and widespread Holocene expansion across northern California. *Heredity*: 1–16.

Boria RA, Olson LE, Goodman SM, Anderson RP (2014). Spatial filtering to reduce sampling bias can improve the performance of ecological niche models. *Ecol Model* **275**.

Brown JL, Weber JJ, Alvarado-Serrano DF, Hickerson MJ, Franks SJ, Carnaval AC (2016). Predicting the genetic consequences of future climate change: The power of coupling spatial demography, the coalescent, and historical landscape changes. *Am J Bot* **103**: 153–163.

Burbrink FT, Chan YL, Myers EA, Ruane S, Smith BT, Hickerson MJ (2016). Asynchronous demographic responses to Pleistocene climate change in Eastern Nearctic vertebrates. *Ecol Lett* **19**: 1457–1467.

Carnaval AC, Hickerson MJ, Haddad CFB, Rodrigues MT, Moritz C (2009). Stability predicts genetic diversity in the Brazilian Atlantic forest hotspot. *Science* **323**: 785–789.

Chifman J, Kubatko L (2015). Identifiability of the unrooted species tree topology under the coalescent model with time-reversible substitution processes, site-specific rate variation, and invariable sites. *J Theor Biol* **374**: 35–47.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. (2011). The variant call format and VCFtools. *Bioinformatics* **27**: 2156–2158.

Dansgaard W, Johnsen SJ, Clausen HB, Dahl-Jensen D, Gundestrup NS, Hammer CU, et al. (1993). Evidence for general instability of past climate from a 250-kyr ice-core record. *Nature* **364**: 218–220.

Dawson NG, Hope AG, Talbot SL, Cook JA (2014). A multilocus evaluation of ermine (Mustela erminea) across the Holarctic, testing hypotheses of Pleistocene diversification in response to climate change. *J Biogeogr* **41**: 464–475.

Dawson TP, Jackson ST, House JI, Prentice IC, Mace GM (2011). Beyond Predictions: Biodiversity Conservation in a Changing Climate. *Science* **332**: 53–58.

Dewey MJ, Dawson WD (2001). Deer mice: “The Drosophila of North American mammalogy”. *genesis* **29**: 105–109.

Dragoo JW, Lackey JA, Moore KE, Lessa EP, Cook JA, Yates TL (2006). Phylogeography of the deer mouse (Peromyscus maniculatus) provides a predictive framework for research on hantaviruses. *J Gen Virol* **87**: 1997–2003.

Dynesius M, Jansson R (2000). Evolutionary consequences of changes in species’ geographical distributions driven by Milankovitch climate oscillations. *Proc Natl Acad Sci* **97**: 9115–9120.

Earl DA, vonHoldt BM (2012). STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conserv Genet Resour* **4**: 359–361.

Eaton DAR, Overcast I (2020). ipyrad: Interactive assembly and analysis of RADseq datasets. *Bioinformatics* **36**: 2592–2594.

Eddingsaas AA, Jacobsen BK, Lessa EP, Cook JA (2004). Evolutionary History of the Arctic Ground Squirrel (Spermophilus parryii) in Nearctic Beringia. *J Mammal* **85**: 601–610.

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

Excoffier L, Dupanloup I, Huerta-Sanchez E, Sousa VC, FollSousa M (2013). Robust Demographic Inference from Genomic and SNP Data. *PLoS Genet* **9**: e1003905–e1003905.

Fagre A, Lewis J, Eckley M, Zhan S, Rocha SM, Sexton NR, et al. (2020). SARS-CoV-2 infection, neuropathogenesis and transmission among deer mice: Implications for reverse zoonosis to New World rodents. *bioRxiv*.

Fleming MA, Cook JA (2002). Phylogeography of endemic ermine (Mustela erminea) in southeast Alaska. *Mol Ecol* **11**: 795–807.

Frankham R, Ballou SEJD, Briscoe DA, Ballou JD (2002). *Introduction to Conservation Genetics*. Cambridge University Press.

Gavin DG, Fitzpatrick MC, Gugger PF, Heath KD, Rodríguez-Sánchez F, Dobrowski SZ, et al. (2014). Climate refugia: joint inference from fossil records, species distribution models and phylogeography. *New Phytol* **204**: 37–54.

Hall ER (1981). *The mammals of North America*. Wiley: New York.

Harrington SM, Hollingsworth BD, Higham TE, Reeder TW (2017). Pleistocene climatic fluctuations drive isolation and secondary contact in the red diamond rattlesnake (Crotalus ruber) in Baja California. *J Biogeogr* **45**: 64–75.

Hewitt G (2000). The genetic legacy of the Quaternary ice ages. *Nature* **405**: 907–913.

Hewitt GM (2004). Genetic consequences of climatic oscillations in the Quaternary. *Philos Trans R Soc Lond B Biol Sci* **359**: 183–195.

Hickerson MJ (2014). All models are wrong. *Mol Ecol* **23**: 2887–2889.

Hope AG, Waltari E, Malaney JL, Payer DC, Cook JA, Talbot SL (2015). Arctic biodiversity : increasing richness accompanies shrinking refugia for a cold-associated tundra fauna. *Ecosphere* **6**: 1–67.

Hughes JB, Daily GC, Ehrlich PR (1997). Population Diversity: Its Extent and Extinction. *Science* **278**: 689–692.

Jackson ST, Overpeck JT (2000). Responses of plant populations and communities to environmental changes of the late Quaternary. *Paleobiology* **26**: 194-220.

Janes JK, Miller JM, Dupuis JR, Malenfant RM, Gorrell JC, Cullingham CI, et al. (2017). The K = 2 conundrum. *Mol Ecol* **26**: 3594–3602.

Jezkova T, Jaeger JR, Oláh-Hemmings V, Jones KB, Lara-Resendiz RA, Mulcahy DG, et al. (2016). Range and niche shifts in response to past climate change in the desert horned lizard Phrynosoma platyrhinos. *Ecography* **39**: 437–448.

Jezkova T, Riddle BR, Card DC, Schield DR, Eckstut ME, Castoe TA (2015). Genetic consequences of postglacial range expansion in two codistributed rodents (genus Dipodomys) depend on ecology and genetic locus. *Mol Ecol* **24**: 83–97.

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

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

Kalkvik HM, Stout IJ, Doonan TJ, Parkinson CL (2012). Investigating niche and lineage diversification in widely distributed taxa: Phylogeography and ecological niche modeling of the Peromyscus maniculatus species group. *Ecography* **35**: 54–64.

Kawamura K, Parrenin F, Lisiecki L, Uemura R, Vimeux F, Severinghaus JP, et al. (2007). Northern Hemisphere forcing of climatic cycles in Antarctica over the past 360,000 years. *Nature* **448**: 912–916.

Keppel G, Niel KPV, Wardell-Johnson GW, Yates CJ, Byrne M, Mucina L, et al. (2012). Refugia: identifying and understanding safe havens for biodiversity under climate change. *Glob Ecol Biogeogr* **21**: 393–404.

Kleman J, Jansson K, De Angelis H, Stroeven AP, Hättestrand C, Alm G, et al. (2010). North American Ice Sheet build-up during the last glacial cycle, 115–21kyr. *Quat Sci Rev* **29**: 2036–2051.

Knowles LL, Alvarado-Serrano DF (2010). Exploring the population genetic consequences of the colonization process with spatio-temporally explicit models: Insights from coupled ecological, demographic and genetic models in montane grasshoppers. *Mol Ecol* **19**: 3727–3745.

Lisiecki LE, Raymo ME (2005). A Pliocene-Pleistocene stack of 57 globally distributed benthic δ18O records. *Paleoceanography* **20**.

Liu Z, Otto-Bliesner BL, He F, Brady EC, Tomas R, Clark PU, et al. (2009). Transient Simulation of Last Deglaciation with a New Mechanism for Bølling-Allerød Warming. *Science* **325**: 310–314.

Lorenz DJ, Nieto-Lugilde D, Blois JL, Fitzpatrick MC, Williams JW (2016). Downscaled and debiased climate simulations for North America from 21,000 years ago to 2100AD. *Sci Data* **3**: 160048.

Lucid MK, Cook JA (2004). Phylogeography of Keen's Mouse (Peromyscus keeni) in a Naturally Fragmented Landscape. *J Mammal* **85**: 1149–1159.

Lucid MK, Cook JA (2007). Cytochrome-b haplotypes suggest an undescribed Peromyscus species from the Yukon. *Can J Zool* **85**: 916–919.

Magyari EK, Major Á, Bálint M, Nédli J, Braun M, Rácz I, et al. (2011). Population dynamics and genetic changes of Picea abies in the South Carpathians revealed by pollen and ancient DNA analyses Population dynamics and genetic changes of Picea

abies in the South Carpathians revealed by pollen and ancient DNA analyses. *BMC Evol Biol* **11**: 66–66.

Malaney JL, Conroy CJ, Moffitt LA, Spoonhunter HD, Patton JL, Cook JA (2013). Phylogeography of the western jumping mouse (Zapus princeps) detects deep and persistent allopatry with expansion. *J Mammal* **94**: 1016–1029.

Mann DH, Hamilton TD (1995). Late Pleistocene and Holocene paleoenvironments of the North Pacific coast. *Quat Sci Rev* **14**: 449–471.

Massatti R, Knowles LL (2016). Contrasting support for alternative models of genomic variation based on microhabitat preference: species-specific effects of climate change in alpine sedges. *Mol Ecol* **25**: 3974–3986.

Miraldo A, Li S, Borregaard MK, Flórez-rodríguez A, Gopalakrishnan S, Rizvanovic M, et al. (2016). An Anthropocene map of genetic diversity. *Science* **353**: 1532–1535.

Moritz C, Agudo R (2013). The Future of Species Under Climate Change: Resilience or Decline? *Science* **341**: 504–508.

Moritz C, Patton JL, Conroy CJ, Parra JL, White GC, Beissinger SR (2008). Impact of a Century of Climate Change on Small-Mammal Communities in Yosemite National Park, USA. *Science* **322**: 261–265.

Moritz CC, Potter S (2013). The importance of an evolutionary perspective in conservation policy planning. *Mol Ecol* **22**: 5969–5971.

Muscarella R, Galante PJ, Soley-Guardia M, Boria RA, Kass JM, Uriarte M, et al. (2014). ENMeval: An R package for conducting spatially independent evaluations and estimating optimal model complexity for Maxent ecological niche models. *Methods Ecol Evol* **5**: 1198–1205.

Myers P, Lundrigan BL, Hoffman SMG, Haraminac AP, Seto SH (2009). Climate-induced changes in the small mammal communities of the Northern Great Lakes Region. *Glob Change Biol* **15**: 1434–1454.

Nachman MW, Crowell SL (2000). Estimate of the Mutation Rate per Nucleotide in Humans. *Genetics* **156**: 297–304.

O'Neill GA, Hamann A, Wang T (2008). Accounting for population variation improves estimates of the impact of climate change on species' growth and distribution. *J Appl Ecol* **45**: 1040–1049.

Osgood WH (1909). *Revision of the Mice of the American Genus Peromyscus*. U.S. Government Printing Office.

Parmesan C (2006). Ecological and Evolutionary Responses to Recent Climate Change. *Annu Ecol Evol Syst* **37**: 637–669.

Pauls SU, Nowak C, Bálint M, Pfenninger M (2013). The impact of global climate change on genetic diversity within populations and species. *Mol Ecol* **22**: 925–946.

Peter BM, Slatkin M (2013). Detecting range expansion from genetic data. *Evolution* **67**: 3274–3289.

Peter BM, Slatkin M (2015). The effective founder effect in a spatially expanding population. *Evolution* **69**: 721–734.

Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE (2012). Double digest RADseq: An inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PLoS ONE* **7**.

Phillips SJ, Anderson RP, Dudík M, Schapire RE, Blair ME (2017). Opening the black box: an open-source release of Maxent. *Ecography* **40**: 887–893.

Phillips SJ, Anderson RP, Schapire RE (2006). Maximum entropy modeling of species geographic distributions. *Ecol Model* **190**: 231–259.

Pielou EC (1991). *After the Ice Age: The Return of Life to Glaciated North America*. University of Chicago Press.

Pironon S, Papuga G, Angert AL, María B, Thompson JD (2017). Geographic variation in genetic and demographic performance : new insights from an old biogeographical paradigm. *Biol Rev Camb Philos Soc* **92**: 1877–1909.

Pritchard JK, Stephens M, Donnelly P (2000). Inference of Population Structure Using Multilocus Genotype Data. *Genetics* **155**: 945–959.

Radosavljevic A, Anderson RP (2014). Making better Maxent models of species distributions: complexity, overfitting and evaluation. *J Biogeogr* **41**: 629–643.

Reid BN, Kass JM, Wollney S, Jensen EL, Russello MA, Viola EM, et al. (2019). Disentangling the genetic effects of refugial isolation and range expansion in a trans-continentally distributed species. *Heredity* **122**: 441–457.

Rhoads SN (1894). Descriptions of Four New Species and Two Subspecies of White-Footed Mice from the United States and British Columbia. *Proc Acad Nat Sci Phila* **46**: 253–261.

Rissler LJ, Hijmans RJ, Graham CH, Moritz C, Wake DB (2006). Phylogeographic Lineages and Species Comparisons in Conservation Analyses : A Case Study of California Herpetofauna. *Am Nat* **167**: 655–666.

Sawyer E, Flamme MJ, Jung TS, Macdonald SO, Cook JA (2017). Diversification of deermice (Rodentia : genus Peromyscus) at their north- western range limit : genetic consequences of refugial and island isolation. *J Biogeogr* **44**: 1572–1585.

Schwartz MW, Iverson LR, Prasad AM, Matthews SN, O'Connor RJ (2006). Predicting Extinctions as a Result of Climate Change. *Ecology* **87**: 1611–1615.

Sexton JP (2019). The Adaptive Continuum and How Species Succeed and Fail. *Philos Theory Pract Biol* **11**.

Shafer ABA, Cullingham CI, Côté SD, Coltman DW (2010). Of glaciers and refugia: A decade of study sheds new light on the phylogeography of northwestern North America. *Mol Ecol* **19**: 4589–4621.

Shcheglovitova M, Anderson RP (2013). Estimating optimal complexity for ecological niche models: A jackknife approach for species with small sample sizes. *Ecol Model* **269**: 9–17.

Shorter KR, Crossland JP, Webb D, Szalai G, Felder MR, Vrana PB (2012). Peromyscus as a Mammalian Epigenetic Model. *Genet Res Int* **2012**: 1–11.

Sim Z, Hall JC, Jex B, Hegel TM, Coltman DW (2016). Genome-wide set of SNPs reveals evidence for two glacial refugia and admixture from postglacial recolonization in an alpine ungulate. *Mol Ecol* **25**: 3696–3705.

Sinervo B, Méndez-de-la-Cruz F, Miles DB, Heulin B, Bastiaans E, Cruz MVS, et al. (2010). Erosion of lizard diversity by climate change and altered thermal niches. *Science* **328**: 894–899.

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

Stamatakis A, Ott M, Ludwig T (2005). RAxML-OMP: An Efficient Program for Phylogenetic Inference on SMPs. In: Malyshkin V (ed) *Parallel Computing Technologies*, Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp 288–302.

Stewart JR, Lister AM, Barnes I, Dalen L (2010). Refugia revisited: individualistic responses of species in space and time. *Proc R Soc B Biol Sci* **277**: 661–671.

Storz JF, Cheviron ZA, McClelland GB, Scott GR (2019). Evolution of physiological performance capacities and environmental adaptation: insights from high-elevation deer mice (Peromyscus maniculatus). *J Mammal* **100**: 910–922.

Swoford DL (2003). *PAUP* Phylogenetic Analysis Using Parsimony (*and Other Methods).* Sinauer Associates: Sunderland, Massachusetts.

Tarroso P, Velo-Antón G, Carvalho SB (2015). phylin: an r package for phylogeographic interpolation. *Mol Ecol Resour* **15**: 349–357.

Urban MC (2018). Escalator to extinction. *Proc Natl Acad Sci* **115**: 11871–11873.

Warren R, VanDerWal J, Price J, Welbergen JA, Atkinson I, Ramirez-Villegas J, et al. (2013). Quantifying the benefit of early climate change mitigation in avoiding biodiversity loss. *Nat Clim Change* **3**: 678–682.

Yang DS, Conroy CJ, Moritz C (2011). Contrasting responses of Peromyscus mice of Yosemite National Park to recent climate change. *Glob Change Biol* **17**: 2559–2566.

Yannic G, Pellissier L, Ortego J, Lecomte N, Couturier S, Cuyler C, et al. (2014). Genetic diversity in caribou linked to past and future climate change. *Nat Clim Change* **4**: 132–137.

Zheng X, Arbogast BS, Kenagy GJ (2003). Historical demography and genetic structure of sister species: deermice (Peromyscus) in the North American temperate rain forest. *Mol Ecol* **12**: 711–724.

3.7 Tables

Table 3.1: Top six demographic scenarios using fastsimcoal2 according to AIC scores for *Peromyscus maniculatus* across its range in western North America. Scenario 15 was indicated as the model that best fit the data.

Scenarios	Parameters	AIC	△AIC	AIC_W
S15	8	751.068	0	0.43
S1	7	752.490	1.422	0.21
S10	9	753.496	2.428	0.13
S13	8	754.531	3.463	0.08
S14	8	754.784	3.716	0.07
S19	10	755.188	4.120	0.05

3.8 Figures

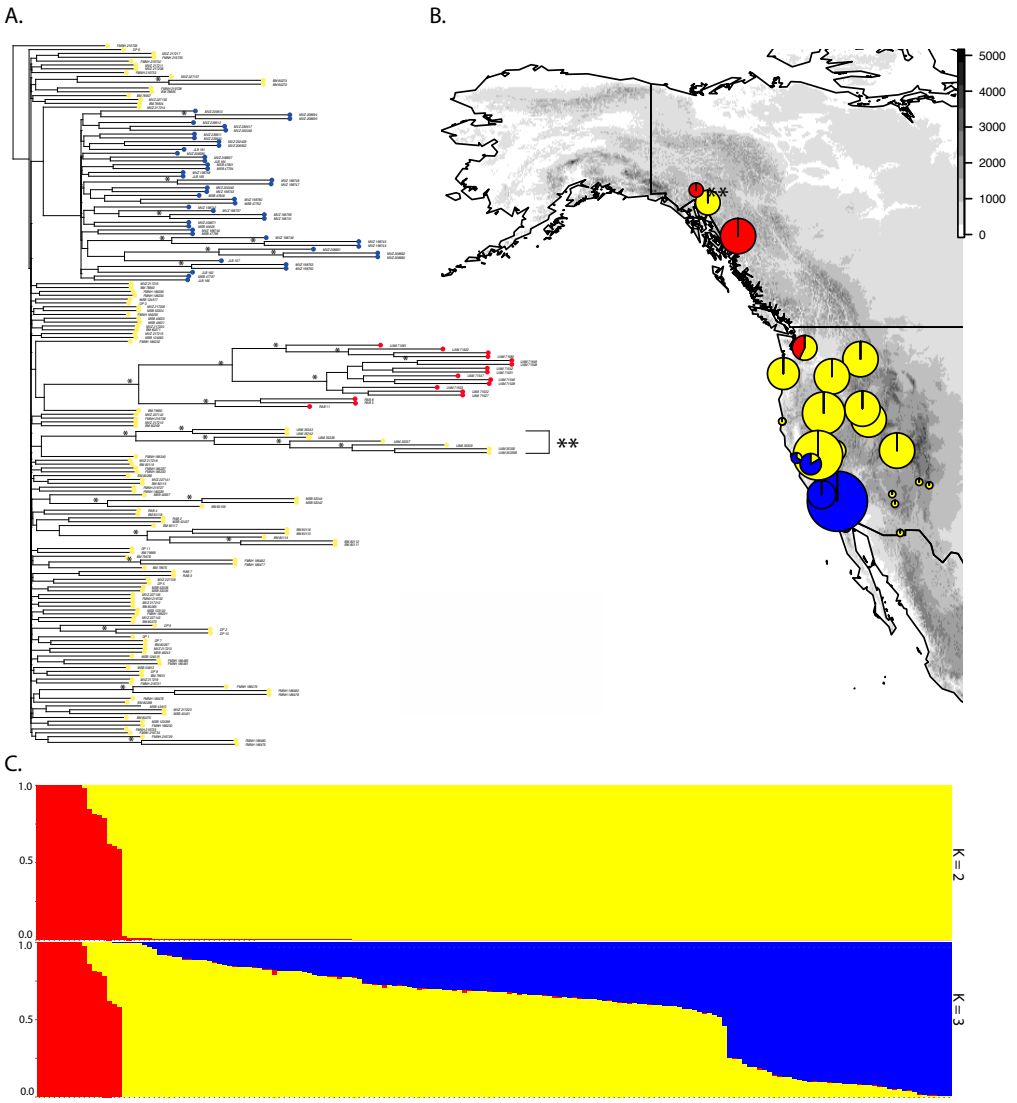
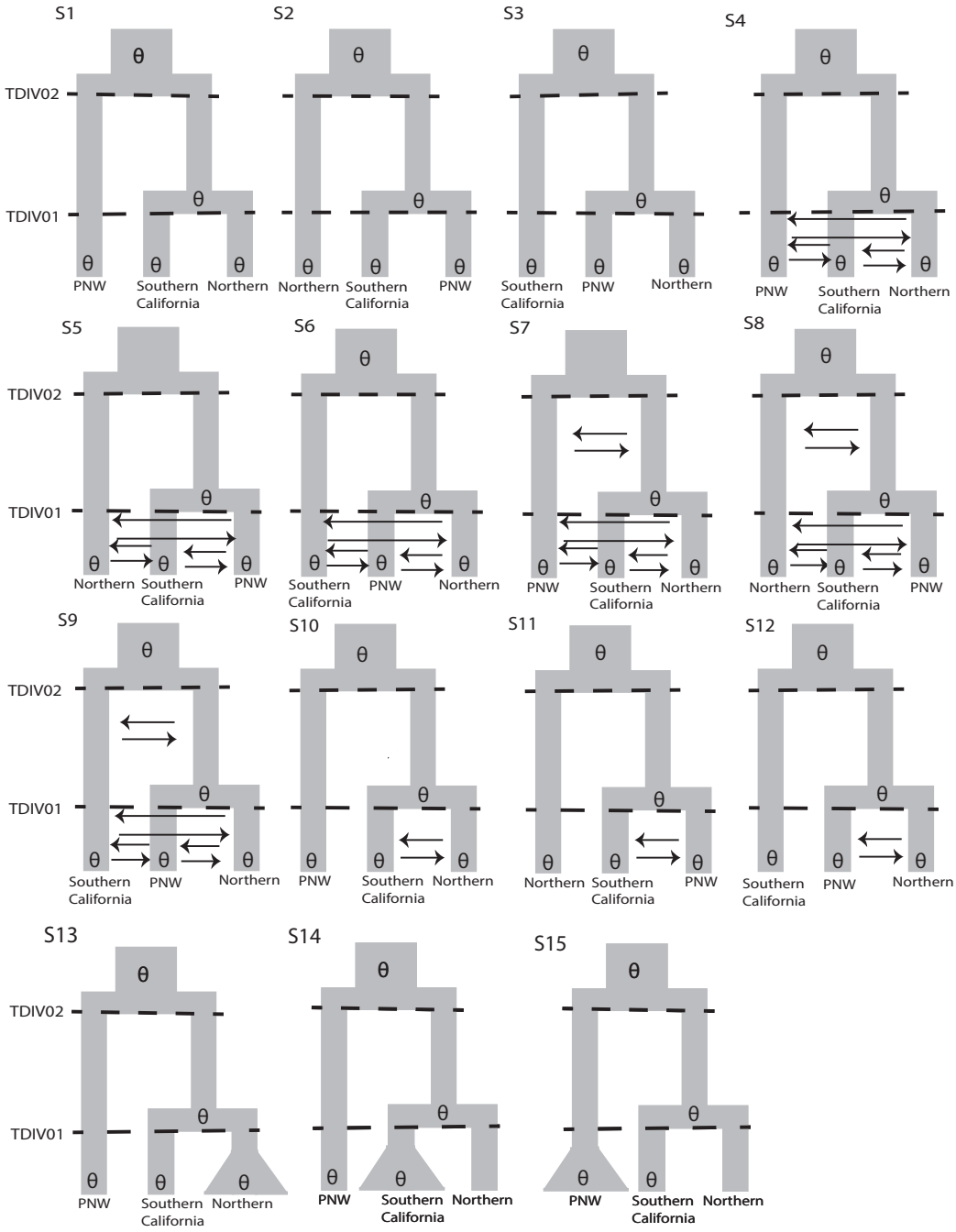


Figure 3.1: Figure 1. The geographic structure of *Peromyscus maniculatus* across its range in western North America. Species tree, geographic position of the sampled localities, and STRUCTURE results showing *K* = 2 and *K* = 3. Colors on the branch tips correspond to the population individuals were placed within STRUCTURE, and the size of the circles on the map indicate the number of individuals sampled per locality. The black circles at nodes indicate bootstrap support > 50%. ** is used to denote the Canadian locality found within the Northern population.


S1
S2
S3
S4
S5
S6
S7
S8
S9
S10
S11
S12
S13
S14
S15
θ
TDIV02
TDIV01
PNW
Southern California
Northern

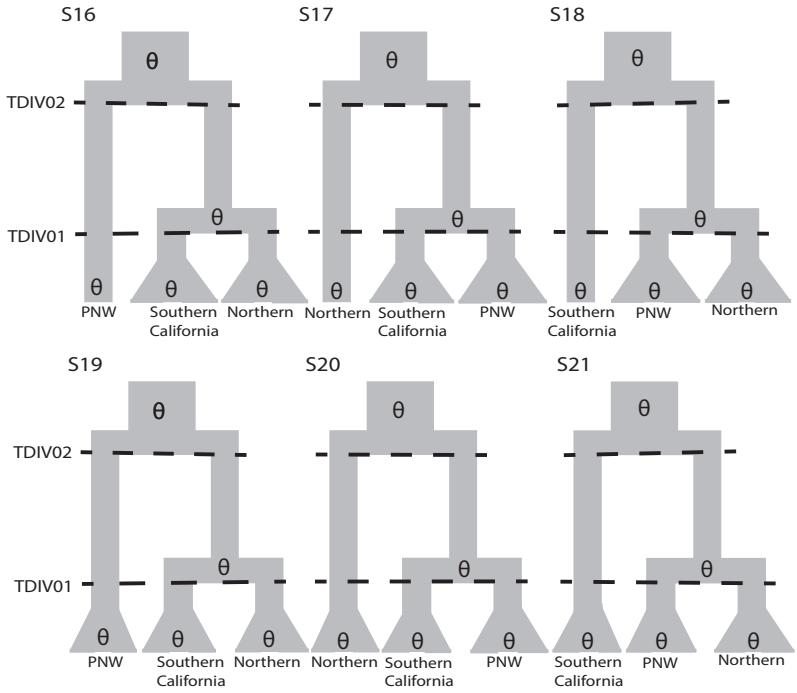


Figure 3.2: The 21 phylogeographic hypotheses used to simulate demographic history for *Peromyscus maniculatus* in western North America. The parameters included contemporary population size for the three populations, their ancestral population sizes, migration rates between the populations, recent population expansion events, and the timing of divergence.

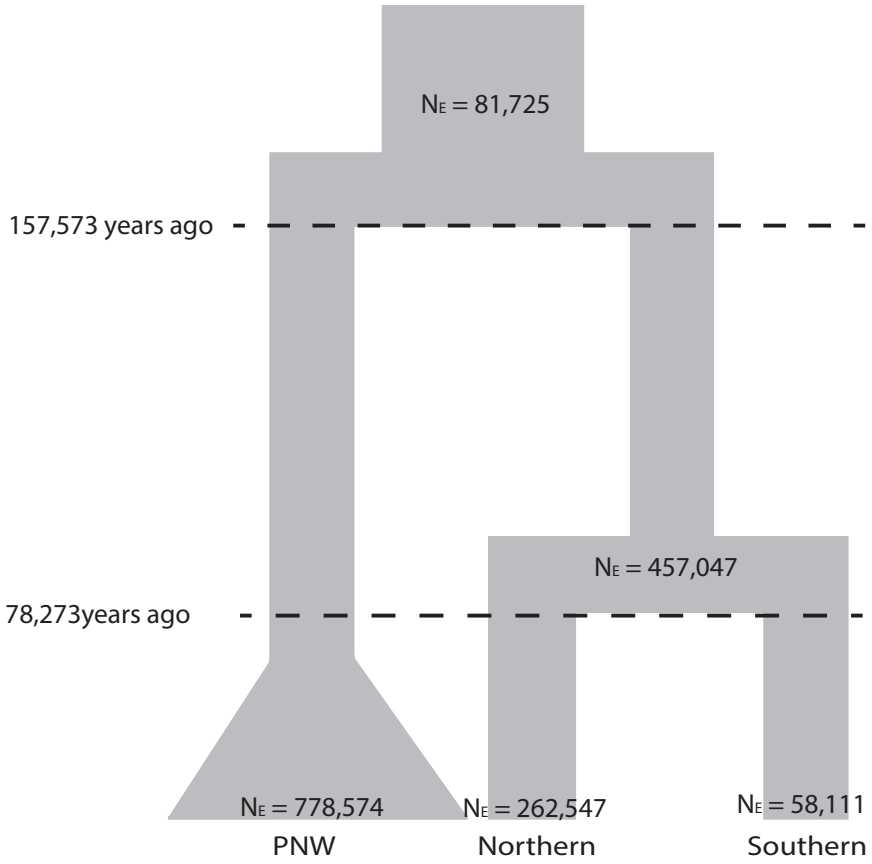


Figure 3.3: The best fit model and parameter estimates of demographic history for *Peromyscus maniculatus* in western North America.

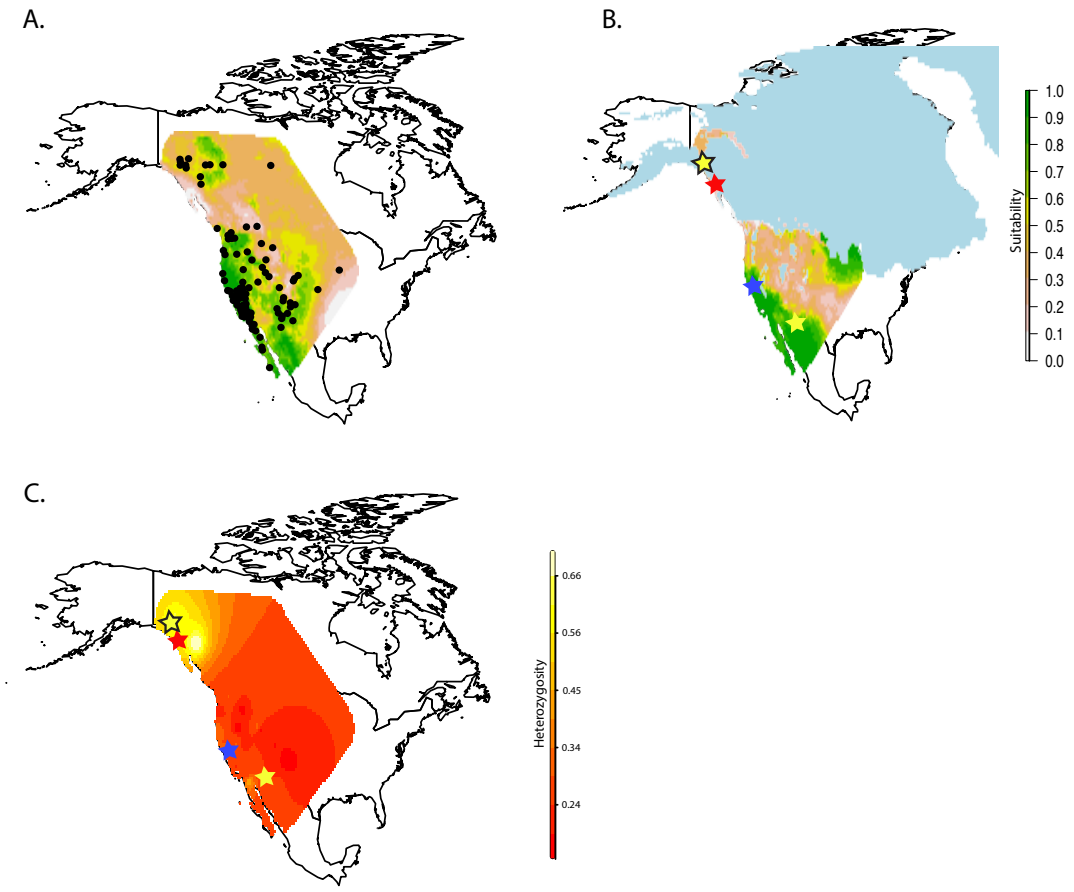


Figure 3.4: Ecological niche models and interpolation of genetic diversity for *Peromyscus maniculatus* in western North America: A) inferred habitat suitability at the present; B) inferred habitat suitability during the Last Glacial Maximum; C) interpolation of heterozygosity. Note the reduction of possible suitable areas in the past for *P. maniculatus* and the high amount of genetic diversity in the Pacific Northwest. The blue pixels in B indicate ice or paleolakes during the Last Glacial Maximum. The yellow, blue, and red stars in B and C indicate the origin of expansion for the Northern, southern California, and PNW populations, respectively. The yellow star with the black outline represents the origin of expansion for the Northern population with all individuals and the yellow star with no outline is the origin of expansion without the Canadian population.

3.9 Appendix

Appendix Table 3.9.1: Gazetteer of specimens corresponding to localities of *Peromyscus maniculatus* employed in this study. Species, museum catalog number, and geographic coordinates (in decimal degrees). Samples used in this study are held at the Burke Museum (BM), Field Museum, Museum of Southwestern Biology (MSB), Museum of Vertebrate Zoology (MVZ), and University of Alaska Museum of the North (UAM). Dr. Dean Pearson (DP), Dr. Jessica Blois, (JLB) and Robert Boria (RAB) also provided tissues used in this project.

species	Sample_Name	Latitude	Longitude
Peromyscus maniucalatus	BM_78895	42.08	-113.74
Peromyscus maniucalatus	BM_78900	42.08	-113.71
Peromyscus maniucalatus	BM_78904	42.09	-113.7
Peromyscus maniucalatus	BM_78907	42.09	-113.7
Peromyscus maniucalatus	BM_79650	42.08	-113.7
Peromyscus maniucalatus	BM_79651	42.08	-113.69
Peromyscus maniucalatus	BM_79669	42.07	-113.68
Peromyscus maniucalatus	BM_79673	42.09	-113.72
Peromyscus maniucalatus	BM_79675	42.08	-113.68
Peromyscus maniucalatus	BM_79676	42.08	-113.73
Peromyscus maniucalatus	BM_80111	45.05	-123.93
Peromyscus maniucalatus	BM_80112	45.05	-123.93
Peromyscus maniucalatus	BM_80113	45.05	-123.93
Peromyscus maniucalatus	BM_80114	45.05	-123.93
Peromyscus maniucalatus	BM_80115	45.05	-123.93
Peromyscus maniucalatus	BM_80116	45.05	-123.93
Peromyscus maniucalatus	BM_80117	45.05	-123.93
Peromyscus maniucalatus	BM_80118	45.05	-123.93

Peromyscus maniucalatus	BM_80119	45.05	-123.93
Peromyscus maniucalatus	BM_80159	45.05	-123.93
Peromyscus maniucalatus	BM_80265	44.81	-117.68
Peromyscus maniucalatus	BM_80266	44.81	-117.68
Peromyscus maniucalatus	BM_80267	44.81	-117.68
Peromyscus maniucalatus	BM_80268	44.81	-117.68
Peromyscus maniucalatus	BM_80269	44.81	-117.68
Peromyscus maniucalatus	BM_80270	44.81	-117.68
Peromyscus maniucalatus	BM_80271	44.81	-117.68
Peromyscus maniucalatus	BM_80272	44.81	-117.68
Peromyscus maniucalatus	BM_80273	44.81	-117.68
Peromyscus maniucalatus	BM_80274	44.81	-117.68
Peromyscus maniucalatus	DP_1	46.28	-113.99
Peromyscus maniucalatus	DP_10	46.28	-113.99
Peromyscus maniucalatus	DP_11	46.28	-113.99
Peromyscus maniucalatus	DP_2	46.28	-113.99
Peromyscus maniucalatus	DP_3	46.28	-113.99
Peromyscus maniucalatus	DP_5	46.28	-113.99
Peromyscus maniucalatus	DP_6	46.28	-113.99
Peromyscus maniucalatus	DP_7	46.28	-113.99
Peromyscus maniucalatus	DP_8	46.28	-113.99
Peromyscus maniucalatus	DP_9	46.28	-113.99
Peromyscus maniucalatus	FMNH_186230	38.52	-109.28
Peromyscus maniucalatus	FMNH_186231	38.52	-109.28
Peromyscus maniucalatus	FMNH_186232	38.52	-109.28
Peromyscus maniucalatus	FMNH_186233	38.52	-109.28

Peromyscus maniucalatus	FMNH_186234	38.52	-109.28
Peromyscus maniucalatus	FMNH_186235	38.52	-109.28
Peromyscus maniucalatus	FMNH_186236	38.53	-109.28
Peromyscus maniucalatus	FMNH_186237	38.53	-109.28
Peromyscus maniucalatus	FMNH_186239	38.53	-109.28
Peromyscus maniucalatus	FMNH_186240	38.53	-109.28
Peromyscus maniucalatus	FMNH_186475	41.16	-112.93
Peromyscus maniucalatus	FMNH_186476	41.16	-112.93
Peromyscus maniucalatus	FMNH_186477	41.16	-112.93
Peromyscus maniucalatus	FMNH_186478	41.16	-112.93
Peromyscus maniucalatus	FMNH_186479	41.16	-112.93
Peromyscus maniucalatus	FMNH_186480	41.16	-112.93
Peromyscus maniucalatus	FMNH_186481	41.16	-112.93
Peromyscus maniucalatus	FMNH_186482	41.16	-112.93
Peromyscus maniucalatus	FMNH_186483	41.16	-112.93
Peromyscus maniucalatus	FMNH_186484	41.16	-112.93
Peromyscus maniucalatus	FMNH_186485	41.16	-112.93
Peromyscus maniucalatus	FMNH_219725	41.69	-118.74
Peromyscus maniucalatus	FMNH_219726	41.69	-118.74
Peromyscus maniucalatus	FMNH_219727	41.69	-118.74
Peromyscus maniucalatus	FMNH_219728	41.69	-118.74
Peromyscus maniucalatus	FMNH_219729	41.69	-118.74
Peromyscus maniucalatus	FMNH_219730	41.69	-118.74
Peromyscus maniucalatus	FMNH_219731	41.69	-118.74
Peromyscus maniucalatus	FMNH_219732	41.69	-118.74
Peromyscus maniucalatus	FMNH_219733	41.69	-118.74

Peromyscus maniucalatus	FMNH_219734	41.69	-118.74
Peromyscus maniucalatus	FMNH_219735	41.69	-118.74
Peromyscus maniucalatus	FMNH_219736	41.69	-118.74
Peromyscus maniucalatus	JLB_149	37.37	-120.41
Peromyscus maniucalatus	JLB_152	37.37	-120.41
Peromyscus maniucalatus	JLB_157	37.37	-120.41
Peromyscus maniucalatus	JLB_161	37.37	-120.41
Peromyscus maniucalatus	JLB_162	37.37	-120.41
Peromyscus maniucalatus	JLB_163	37.37	-120.41
Peromyscus maniucalatus	JLB_164	37.37	-120.41
Peromyscus maniucalatus	JLB_165	37.37	-120.41
Peromyscus maniucalatus	JLB_166	37.37	-120.41
Peromyscus maniucalatus	JLB_192	37.37	-120.41
Peromyscus maniucalatus	MSB_123069	34.82	-109.9
Peromyscus maniucalatus	MSB_123100	34.87	-109.79
Peromyscus maniucalatus	MSB_124063	34.05	-109.54
Peromyscus maniucalatus	MSB_124516	35.86	-106.42
Peromyscus maniucalatus	MSB_124517	35.86	-106.42
Peromyscus maniucalatus	MSB_40606	37.09	-119.72
Peromyscus maniucalatus	MSB_40623	37.09	-119.72
Peromyscus maniucalatus	MSB_40697	40.99	-124.05
Peromyscus maniucalatus	MSB_42374	33.85	-100.79
Peromyscus maniucalatus	MSB_42379	33.85	-100.79
Peromyscus maniucalatus	MSB_43431	47.22	-120.99
Peromyscus maniucalatus	MSB_43437	47.22	-120.99
Peromyscus maniucalatus	MSB_43442	47.22	-120.99

Peromyscus maniucalatus	MSB_43631	41	-124.09
Peromyscus maniucalatus	MSB_43633	41	-124.09
Peromyscus maniucalatus	MSB_46243	31.53	-108.87
Peromyscus maniucalatus	MSB_46452	31.65	-108.76
Peromyscus maniucalatus	MSB_46453	31.65	-108.76
Peromyscus maniucalatus	MSB_47792	33.37	-116.83
Peromyscus maniucalatus	MSB_47794	33.37	-116.83
Peromyscus maniucalatus	MSB_47795	33.37	-116.83
Peromyscus maniucalatus	MSB_47797	32.05	-115.91
Peromyscus maniucalatus	MSB_47800	32.05	-115.91
Peromyscus maniucalatus	MSB_47801	32.05	-115.91
Peromyscus maniucalatus	MSB_53334	38.47	-119.26
Peromyscus maniucalatus	MSB_53335	38.47	-119.26
Peromyscus maniucalatus	MSB_53338	38.47	-119.26
Peromyscus maniucalatus	MSB_53342	44.33	-123.5
Peromyscus maniucalatus	MSB_53344	44.33	-123.5
Peromyscus maniucalatus	MSB_55813	35.52	-105.18
Peromyscus californicus	MVZ_198617	34.56	-118.56
Peromyscus californicus	MVZ_198618	34.71	-118.54
Peromyscus californicus	MVZ_198687	34.31	-117.54
Peromyscus maniucalatus	MVZ_198740	34.22	-116.97
Peromyscus maniucalatus	MVZ_198741	34.22	-116.97
Peromyscus maniucalatus	MVZ_198742	34.22	-116.97
Peromyscus maniucalatus	MVZ_198743	34.22	-116.97
Peromyscus maniucalatus	MVZ_198744	34.22	-116.97
Peromyscus maniucalatus	MVZ_198745	34.22	-116.97

Peromyscus maniucalatus	MVZ_198746	34.22	-116.97
Peromyscus maniucalatus	MVZ_198747	34.22	-116.97
Peromyscus maniucalatus	MVZ_198748	34.22	-116.97
Peromyscus maniucalatus	MVZ_198750	34.22	-116.94
Peromyscus maniucalatus	MVZ_198753	34.22	-116.94
Peromyscus maniucalatus	MVZ_198757	34.22	-116.98
Peromyscus maniucalatus	MVZ_198758	34.22	-116.98
Peromyscus maniucalatus	MVZ_198759	34.22	-116.98
Peromyscus maniucalatus	MVZ_198760	34.79	-119
Peromyscus maniucalatus	MVZ_200045	34.22	-116.98
Peromyscus maniucalatus	MVZ_200046	34.79	-119
Peromyscus maniucalatus	MVZ_202865	34.79	-119
Peromyscus maniucalatus	MVZ_202866	36.37	-121.54
Peromyscus maniucalatus	MVZ_206952	36.37	-121.54
Peromyscus maniucalatus	MVZ_208660	37.89	-122.23
Peromyscus maniucalatus	MVZ_208661	34.79	-119
Peromyscus maniucalatus	MVZ_208662	34.79	-119
Peromyscus maniucalatus	MVZ_208671	34.79	-119
Peromyscus maniucalatus	MVZ_208693	34.8	-119
Peromyscus maniucalatus	MVZ_208694	34.8	-119
Peromyscus maniucalatus	MVZ_208695	34.8	-119
Peromyscus maniucalatus	MVZ_208697	34.8	-119
Peromyscus maniucalatus	MVZ_217208	38.12	-119.48
Peromyscus maniucalatus	MVZ_217209	38.12	-119.48
Peromyscus maniucalatus	MVZ_217210	38.12	-119.48
Peromyscus maniucalatus	MVZ_217211	38.12	-119.48

Peromyscus maniucalatus	MVZ_217212	38.12	-119.48
Peromyscus maniucalatus	MVZ_217213	38.12	-119.48
Peromyscus maniucalatus	MVZ_217214	38.12	-119.48
Peromyscus maniucalatus	MVZ_217215	38.12	-119.48
Peromyscus maniucalatus	MVZ_217216	38.12	-119.48
Peromyscus maniucalatus	MVZ_217217	38.12	-119.48
Peromyscus maniucalatus	MVZ_217218	38.12	-119.48
Peromyscus maniucalatus	MVZ_217219	38.12	-119.48
Peromyscus maniucalatus	MVZ_217220	38.12	-119.48
Peromyscus maniucalatus	MVZ_217223	38.12	-119.48
Peromyscus maniucalatus	MVZ_219103	38.5	-117.25
Peromyscus maniucalatus	MVZ_227135	37.89	-122.24
Peromyscus maniucalatus	MVZ_227136	38.5	-117.25
Peromyscus maniucalatus	MVZ_227137	38.5	-117.25
Peromyscus maniucalatus	MVZ_227138	38.5	-117.25
Peromyscus maniucalatus	MVZ_227139	38.5	-117.25
Peromyscus maniucalatus	MVZ_227140	38.5	-117.25
Peromyscus maniucalatus	MVZ_227141	38.5	-117.25
Peromyscus maniucalatus	MVZ_227142	38.5	-117.25
Peromyscus maniucalatus	MVZ_227143	38.5	-117.25
Peromyscus maniucalatus	MVZ_227146	38.5	-117.25
Peromyscus maniucalatus	MVZ_227162	38.5	-117.25
Peromyscus maniucalatus	MVZ_227163	37.9	-122.18
Peromyscus maniucalatus	MVZ_227164	37.87	-122.18
Peromyscus maniucalatus	MVZ_227166	37.9	-122.18
Peromyscus maniucalatus	MVZ_228329	38.5	-117.25

Peromyscus maniucalatus	MVZ_228331	38.5	-117.25
Peromyscus maniucalatus	MVZ_228334	38.5	-117.25
Peromyscus maniucalatus	MVZ_228335	37.9	-122.18
Peromyscus maniucalatus	MVZ_228336	37.9	-122.18
Peromyscus maniucalatus	MVZ_230285	37.85	-122.2
Peromyscus maniucalatus	MVZ_232439	37.89	-122.25
Peromyscus maniucalatus	MVZ_239610	34.81	-118.89
Peromyscus maniucalatus	MVZ_239611	34.81	-118.89
Peromyscus maniucalatus	MVZ_239612	34.81	-118.89
Peromyscus maniucalatus	MVZ_239613	34.81	-118.89
Peromyscus maniucalatus	MVZ_239617	34.81	-118.89
Peromyscus maniucalatus	RAB_1	47.24	-121.19
Peromyscus maniucalatus	RAB_11	47.24	-121.19
Peromyscus maniucalatus	RAB_2	47.24	-121.19
Peromyscus maniucalatus	RAB_3	47.24	-121.19
Peromyscus maniucalatus	RAB_4	47.24	-121.19
Peromyscus maniucalatus	RAB_5	47.24	-121.19
Peromyscus maniucalatus	RAB_6	47.24	-121.19
Peromyscus maniucalatus	RAB_7	47.24	-121.19
Peromyscus maniucalatus	RAB_8	47.24	-121.19
Peromyscus maniucalatus	RAB_9	47.24	-121.19
Peromyscus maniucalatus	UAM_ 35337	59.56	-133.67
Peromyscus maniucalatus	UAM_ 35338	59.56	-133.67
Peromyscus maniucalatus	UAM_ 35339	59.56	-133.67
Peromyscus maniucalatus	UAM_ 35343	59.56	-133.67
Peromyscus maniucalatus	UAM_ 71532	56.72	-129.78

Peromyscus maniucalatus	UAM_ 71542	56.69	-129.77
Peromyscus maniucalatus	UAM_ 71548	56.69	-129.77
Peromyscus maniucalatus	UAM_ 71549	56.69	-129.77
Peromyscus maniucalatus	UAM_ 71580	60.63	-135.17
Peromyscus maniucalatus	UAM_ 71581	60.63	-135.17
Peromyscus maniucalatus	UAM_ 71622	60.71	-135.2
Peromyscus maniucalatus	UAM_35335B	59.56	-133.67
Peromyscus maniucalatus	UAM_35336	59.56	-133.67
Peromyscus maniucalatus	UAM_35336B	59.56	-133.67
Peromyscus maniucalatus	UAM_35342	59.56	-133.67
Peromyscus maniucalatus	UAM_71527	56.72	-129.78
Peromyscus maniucalatus	UAM_71528	56.72	-129.78
Peromyscus maniucalatus	UAM_71529B	56.72	-129.78
Peromyscus maniucalatus	UAM_71531	56.72	-129.78
Peromyscus maniucalatus	UAM_71531B	56.72	-129.78
Peromyscus maniucalatus	UAM_71533	56.72	-129.78
Peromyscus maniucalatus	UAM_71533B	56.72	-129.78
Peromyscus maniucalatus	UAM_71546	56.69	-129.77
Peromyscus maniucalatus	UAM_71546B	56.69	-129.77
Peromyscus maniucalatus	UAM_71547	56.69	-129.77
Peromyscus maniucalatus	UAM_71620	60.71	-135.2
Peromyscus maniucalatus	UAM_71620B	60.71	-135.2

Appendix Table 3.9.2: The cross-validation results of ADMIXTURE for *Peromyscus maniculatus* across its range in western North America indicating *K*=3 model as optimal.

Populations	CV error
K=1	0.17229
K=2	0.16606
K=3	0.16369
K=4	0.16705
K=5	0.17383

Appendix Table 3.9.3: The demographic scenarios using fastsimcoal2 according to △AIC for *Peromyscus maniculatus* across its range in western North America. Scenario 15 was indicated as the model that best fit the data. See Figure 2 for a full explanation of the different scenarios.

models	Parameters	AIC	delat AIC	weights
S1	7	752.489	1.42164145	0.21006496
S2	7	767.378211	16.3101567	0.00012284
S3	7	767.392027	16.3239722	0.000122
S4	13	761.464099	10.3960446	0.00236367
S5	13	761.47331	10.405255	0.00235281
S6	13	761.795672	10.7276169	0.00200257
S7	15	765.468705	14.4006498	0.00031915
S8	15	761.264184	10.1961294	0.00261215
S9	15	765.468705	14.4006498	0.00031915
S10	9	753.496336	2.42828083	0.12698858
S11	9	771.373606	20.3055515	1.6663E-05
S12	9	771.387422	20.319367	1.6549E-05
S13	8	754.531143	3.46308798	0.0756936
S14	8	754.784427	3.71637234	0.06668976
S15	8	751.068055	0	0.42762111
S16	9	756.517327	5.44927247	0.02803914
S17	9	771.373606	20.3055515	1.6663E-05
S18	9	768.122356	17.0543013	8.4677E-05
S19	10	755.187789	4.11973442	0.05450925
S20	10	771.692719	20.6246644	1.4206E-05
S21	10	770.163803	19.0957479	3.0512E-05

Appendix Table 3.9.4: Re-estimated parameter values from 100 simulated Site Frequency Spectrums (and 95% confidence intervals) for best fit model, scenario 15, from Fastsimcoal2 for *Peromyscus maniculatus* across its range in western North America. See Figure 2 for a graphical depiction of the scenario.

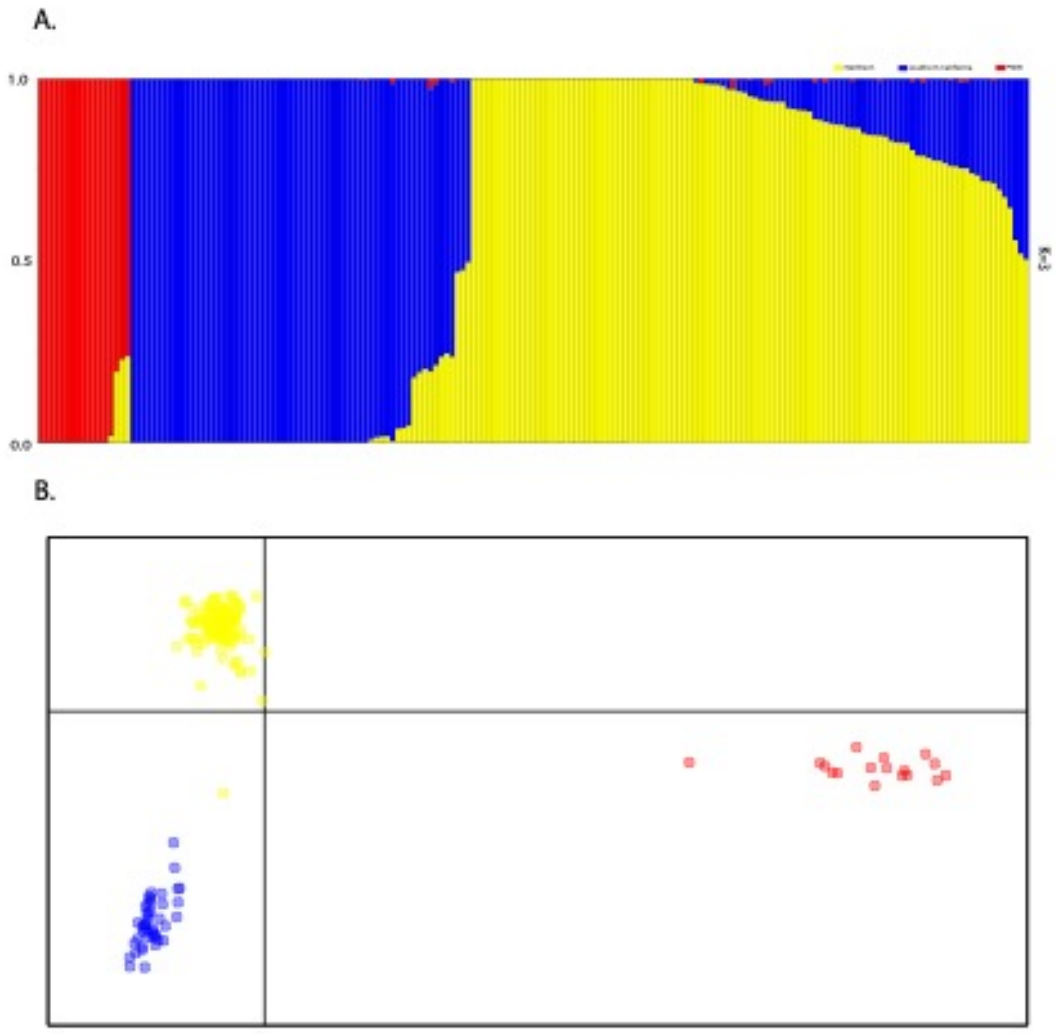
Parameters	Estimates
Northern	262,547 (136,740 – 388,354)
southern California	58,111 (11,035 – 105,187)
PNW	778,574 (548,678 – 1,008,471)
N_E ancestral All	81,725 (62,937 - 100,513)
N_E ancestral Northern + southern California	457,047 (343,020- 571,074)
TDIV Northern + southern California	78,272.95 (60,747.4 – 95,798.5)
R1	-2.21×10^{-05} (-2.79×10^{-05} - 1.63×10^{05})
TDIV All	157,572.75 (136,944.6 – 178,200.9)

Appendix Table 3.9.5: Direct parameter estimates for the six best fit models using Fastsimcoal2 for *Peromyscus maniculatus* across its range in western North America.

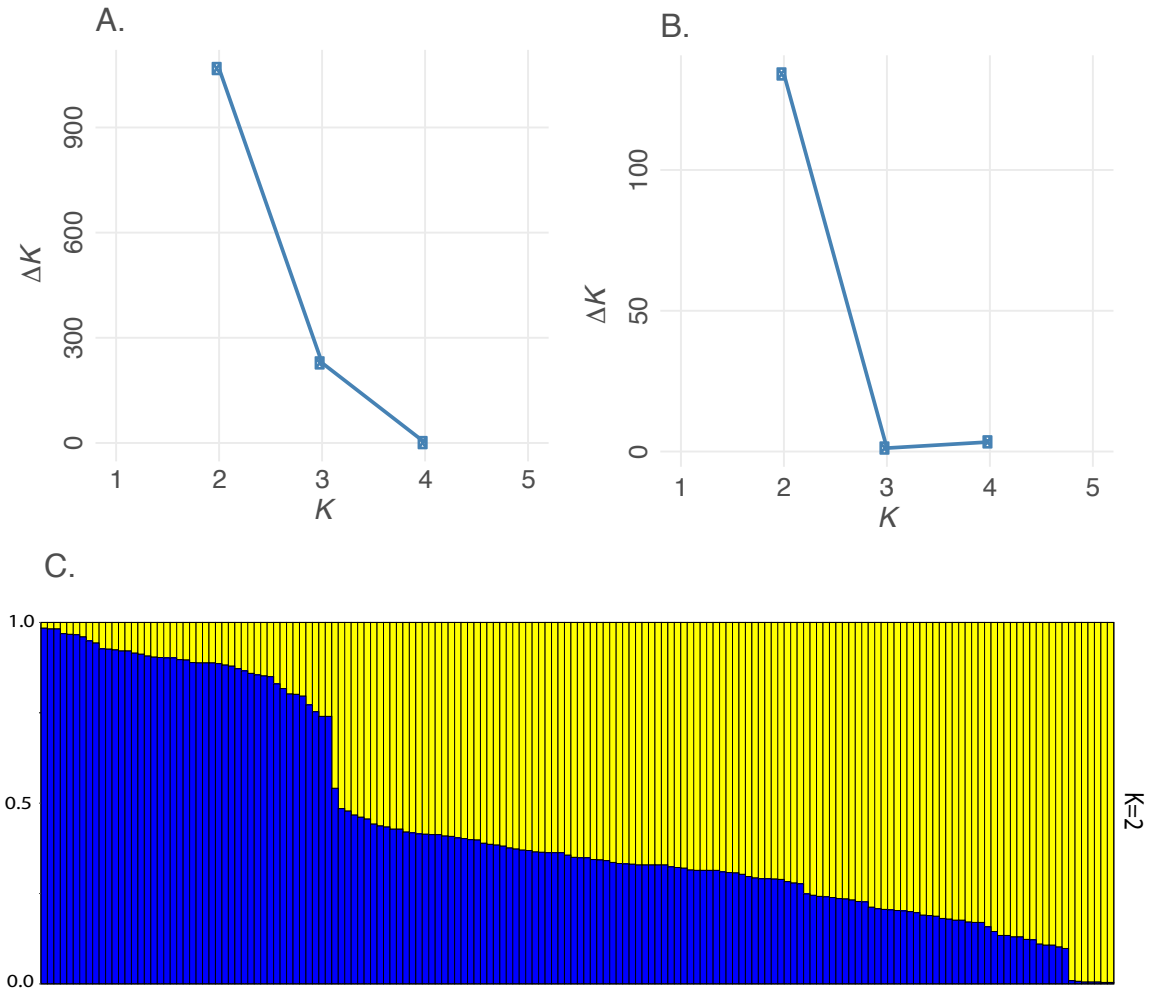
Parameters	S 1	S 10	S13	S 14	S 15	S 19
Northern	7,444,378	7,180,900	7,551,339	7,549,571	7,549,053	7,544,285
southern California	2,820,041	4,200,050	2,934,639	2932913.5	2,675,763	2,844,682
PNW	534,564	654,517	534,062	529893.5	4,354,998	4,793,944
N_E ancestral All	110	496.5	305	190	337.5	442.5
N_E ancestral Northern + southern California	7,537,017	311,817	7,480,539	7260701.5	7,556,107	7,405,540
TDIV Northern + southern California	104,782.5	348,328	102,555	109,891	94,090.5	101,778
R1	-	-	-1.166×10^{-9}	-	-	-4.095×10^{-9}
R2	-	-	-	-6.317×10^{-9}	-	-6.296×10^{-9}
R3	-	-	-	-	-9.434×10^{-6}	-9.669×10^{-6}
MIG	-	5.693×10^{-11}	-	-	-	-
MIG	-	9.376×10^{-11}	-	-	-	-
TDIV Northern + southern California	355,057	349,292	355,293	352,434	363,787	358,182

Appendix Table 3.9.6: Summary statistics for the top five performing Maxent models for *Peromyscus maniculatus* across its range in western North America. We used the top performing setting to generate the ENM predictions.

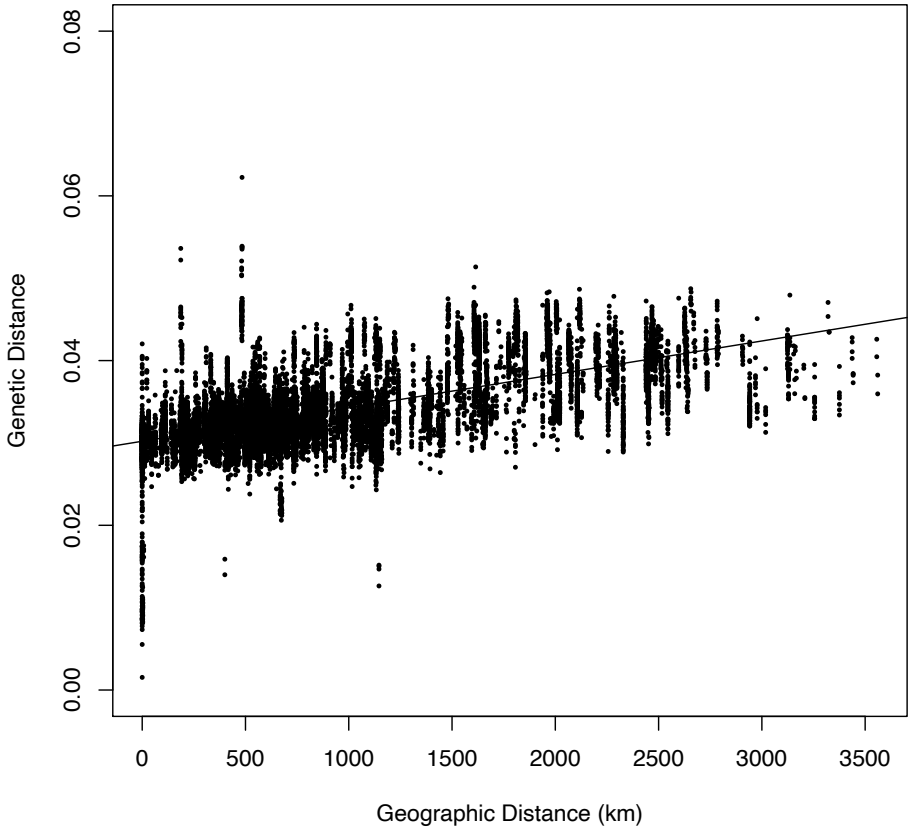
Summary statistics		
Settings	OR_{10}	AUC_{test}
LQH_6	0.12125	0.7055
LQHT_6	0.12125	0.7053
LQH_3	0.14125	0.7602
LQHT_ 3	0.14125	0.7581
LQH_ 4	0.15125	0.7489



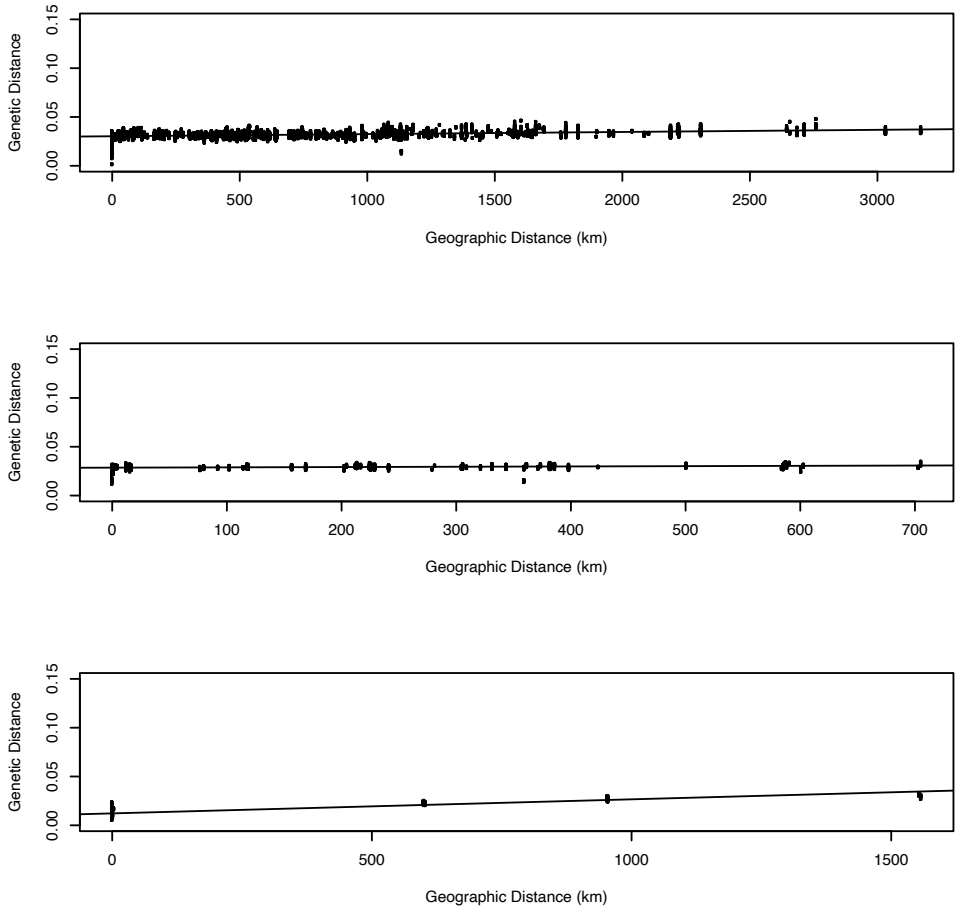
Appendix Figure 3.9.1: The population structure of *Peromyscus maniculatus* across its range in western North America. A. The Admixture results on all individuals indicating *K* = 3 B. The discriminate analysis of principal components (DAPC) results also showing *K* = 3.



Appendix Figure 3.9.2: The results from Structure on *Peromyscus maniculatus* across its range in western North America and for the northern and southern California populations. A. The △K results for all individuals used to determine the optimal *K* = 2. B. The △K results for the northern and southern California populations used to determine the optimal *K* = 2. C. The Structure plot for the northern and southern California populations using *K* = 2. Each line represents one individual depicting the amount of shared ancestry.



Appendix Figure 3.9.3: Plot of genetic distance by geographic distance for all sampled individuals of *Peromyscus maniculatus* across its range in western North America. Note there is significant isolation by distance ($r = 0.614$, $p < 0.001$).



Appendix Figure 3.9.4: Plots of genetic distance by geographic distance for all populations of *Peromyscus maniculatus* across its range in western North America: A) Northern ($r = 0.469$, $p<0.001$); B) southern California ($r = 0.266$, $p<0.001$); C) PNW: $r = 0.917$ ($p<0.001$). Note that there is significant isolation by distance found within each population.

Appendix Figure 3.9.5: The RAxML phylogenetic tree for *Peromyscus maniculatus* across its range in western North America. Note that all individuals were nested within the southern California population.