UNIVERSITY OF CALIFORNIA

Los Angeles

Towards Genomics-Informed Biodiversity Conservation: Case Studies on Environmental DNA, Fin Whales and Bobcats Genomics and the Distribution of Fitness Effects

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

by

Meixi Lin

2022

ABSTRACT OF THE DISSERTATION

Towards Genomics-Informed Biodiversity Conservation: Case Studies on Environmental DNA, Fin Whales and Bobcats Genomics and the Distribution of Fitness Effects

by

Meixi Lin

Doctor of Philosophy in Biology

University of California, Los Angeles, 2022

Professor Kirk Edward Lohmueller, Co-Chair

Professor Robert Wayne, Co-Chair

Global biodiversity is declining at an alarming rate. Genetics approaches have been invaluable for conservation practices across scales, from identifying impacts of inbreeding in endangered species to mapping global biodiversity patterns. The rapid development of next-generation sequencing technologies and computational advances further enabled genomics-informed biodiversity conservation in the 21st century. Compared with a handful of genetic markers that could only be acquired in relatively high-quality genetic materials, genomics approaches generate data with whole genome coverage and allow analyses on environmental DNA (eDNA). Here I demonstrate the applications of genomics in biodiversity conservation in four case studies that encompass a wide spectrum of genetic material types, from eDNA to whole genome sequencing, and a wide spectrum of topics, from population genetics to landscape ecology. In the first chapter, I analyzed the landscape biodiversity pattern derived from eDNA metabarcoding using surface soil samples collected across California. Combining eDNA with environmental predictors, including remote sensing data, have capacity to model biodiversity at landscape scales and to create new biodiversity baselines that span the tree of life. In the second chapter, I analyzed the population genomics pattern in a pair of fin whale (*Balaenoptera physalus*) populations with contrasting demographic trajectories and whaling pressures. I was able to detect the severity of whaling in the Eastern North Pacific population and found that even low levels of migration are crucial to the sustenance of the small and isolated Gulf of California population. In the third chapter, I extended the single-species population genomics to a comparative framework and evaluated the extent of the Distribution of Fitness Effects variation in eight animal species with diverse phylogenetic relationships. I found that the DFE is more similar in more closely related species. In the appendix chapter, I provided another example of how genomics could aid conservation by describing a *de novo* genome assembly of the bobcat (*Lynx rufus*), an indicator species for landscape connectivity.

Overall, this dissertation illustrates the promise and prospect of including genomics in conservation biology through four case studies.

The dissertation of Meixi Lin is approved.

Beth Shapiro

Thomas Welch Gillespie

Kirk Edward Lohmueller, Committee Co-Chair

Robert Wayne, Committee Co-Chair

University of California, Los Angeles

2022

In loving memory of my grandfather

Table of Contents

List of Figures

Chapter 1: Landscape Analyses Using eDNA Metabarcoding and Earth Observation Predict Community Biodiversity in California

Chapter 2: The Genomic Footprint of Whaling and Isolation in Fin Whale Populations

Chapter 3: Variation of the Distribution of Fitness Effects Across Animals

Appendix Chapter: A Reference Genome Assembly of the Bobcat, *Lynx rufus*

List of Tables

Chapter 1: Landscape Analyses Using eDNA Metabarcoding and Earth Observation Predict Community Biodiversity in California

Chapter 3: Variation of the Distribution of Fitness Effects Across Animals

Appendix Chapter: A Reference Genome Assembly of the Bobcat, *Lynx rufus*

Supplementary Materials

Chapter 2: The Genomic Footprint of Whaling and Isolation in Fin Whale Populations

Chapter2_Supplementary_Information.pdf (File Format .pdf).

Chapter 3: Variation of the Distribution of Fitness Effects Across Animals

Chapter3_Supplementary_Information.pdf (File Format .pdf)

Chapter3_Supplementary_Tables.xlsx (File Format .xlsx)

Appendix Chapter: A Reference Genome Assembly of the Bobcat, *Lynx rufus*

AppendixChapter_Supplementary_Tables.xlsx (File Format .xlsx)

Acknowledgements

It’s with my great gratitude to thank my mentors, friends and family in this journey.

First and foremost, I would like to thank my advisors Dr. Bob Wayne and Dr. Kirk Lohmueller. Thank you Bob for showing me the endless potential in research, and motivating me to be innovative and asking the bigger questions. Thank you Kirk for taking me in without much knowledge in Population Genetics, teaching me so much about research, and not being afraid of new things. Thank you both for cultivating me to become a young scientist and helping me navigate through grad school whenever I was in doubt. Their guidance and encouragement have enabled me to explore every aspect I was interested in and made this dissertation possible. I am grateful to Dr. Thomas Gillespie for the helpful discussions on GIS and remote sensing, and your encouragement along every step of my PhD. I thank Dr. Beth Shapiro for introducing me to the wonderful world of ancient DNA and the several fun and inspiring trips up to Santa Cruz.

I am grateful for mentors and collaborators in various projects that I was involved with. Dr. Rachel Meyer is essential in my first chapter and thank you for inspiring me that science is not just publishing and community engagement is equally, if not more, important. To Dr. Sergio Nigenda-Morales, thank you for introducing me into the field of whales and our weekly meetings have been so fun. Your rigorous attitudes in science have shaped my belief in research. To Dr. Michael Buchalski, thank you for believing in me early on. You’ve assured me that genetics can make a difference in real-life conservation management, and I am so honored to be a small part of it. To my undergrad mentor Dr. Meng Yao, thank you for introducing me to research, accommodating my hiccups and always supporting and inspiring me to do better. I could not have made it this far without all your support.

During my PhD, I am surrounded by outstanding colleagues and friends. To Ana, thank you for always opening the doors for me, for the coffees, hikes and fun conversations. To Gabi, thank you for sharing with me the struggles of an international student, and of course, friend's dogs are the best. To Miro, thank you for all the pandemic livestream trips that kept me sane, and not minding my persistent preference for burgers. To Emily, Teia, Maura, Sabrina, Ryan, Fabian, Zack, Gaurav, and Rachel T., thank you for the tea and cookies time, the GIS crash course and welcoming me as I started grad school. To Jacqueline, Annabel, Alice, Chris and Daniel, thank you for holding my hands through GATK pipelines, Hoffman schedulers and various software as I ventured into genomics. To Eduardo, Jon and Annabel, without the DFE chat group, my third chapter would not been possible. To Izabel and Jun, thank you for always answering my career questions. I thank Laurel, Devaughn and Seth for their unlimited knowledge on field biology. To Erin, Dan, Tara, Jackie and Cri, thank you for bearing with my wet lab questions. To Merly, thank you for being the genome assembly guru and always so patient with me. To Chris, Tina, Jesse and Emma, it's great pleasure to have entered grad school together with you and life has taken good turns for us all. To Joey and Stella, I am so excited for your adventures and the lunch breaks have made life so much more interesting. I also have fond memories for the class we taught together with Ana, Maura, Audra, Daniel, Chloe, Kim and Marcel. Thank you to Paulina for being the absolute undergrad rock star, to May and Kiu for trusting me in the first class I TAed, to Nitin and Ramal for your contributions as undergraduate researchers. It takes a village to complete a PhD, and I thank everyone in the Wayne Lab and Lohmueller Lab, and the EEB Department for their friendship and guidance. I also want to thank the staff at the EEB Department, Tessa and Gabby are indispensable in my PhD journey. Alberto and Carlos are always friendly and taught me how to fill a Dewar.

My friends have been a great treasure and I am so lucky to have met you. To Yao, Xiuling and Kiki, you all are amazing, and I am grateful to share the room and life with you. To Fan, Zelan, Molly, James, Riddle, Wei, Jing, etc., thank you for the chats during those harder moments. To Jia, Zhu and Lily, thank you for the fun times we spent in climbing trips, which reminded me what is more important. To Qiao, thank you for believing in me and sharing this journey.

In the end, I would like to thank my family. To my grandfathers, I miss those young sunny afternoons. You supported me whole-heartedly and taught me to strive to become someone that helps make the world a better place, even just a little bit. I miss you so much. To my grandmothers, thank you for being there with me all the time and I hope I can go home and see you soon. To my parents, there aren't enough words I can express my gratitude. You have raised me right. When I chose a different career path, you simply stood by me unconditionally and you are always proud of me even when I don't believe myself. I love you.

To the mountains and rivers that I find peace in, thank you.

Chapter 1 is a version of a manuscript that was originally published in *Ecological Applications* ©2021 by the Ecological Society of America:

Lin, Meixi, Ariel Levi Simons, Ryan J. Harrigan, Emily E. Curd, Fabian D. Schneider, Dannise V. Ruiz-Ramos, Zack Gold, Melisa G. Osborne, Sabrina Shirazi, Teia M. Schweizer, Tiara N. Moore, Emma A. Fox, Rachel Turba, Ana E. Garcia-Vedrenne, Sarah K. Helman, Kelsi Rutledge, Maura Palacios Mejia, Onny Marwayana, Miroslava N. Munguia Ramos, Regina Wetzer, N. Dean Pentcheff, Emily Jane McTavish, Michael N. Dawson, Beth Shapiro, Robert K. Wayne, and Rachel S. Meyer. 2021. "Landscape Analyses Using eDNA Metabarcoding and Earth Observation Predict Community Biodiversity in California." *Ecological Applications* 31

(6). https://doi.org/10.1002/eap.2379. Reprinted with permission from Ecological Society of America.

Funding for the CALeDNA sample processing, infrastructure, and personnel was provided by the University of California Research Initiatives (UCRI) Catalyst grant CA-16-376437 and Howard Hughes Medical Institute (HHMI) Professors Grant GT10483. Additional funding for personnel and computational infrastructure was provided by the National Science Foundation (NSF) 1759756. The research carried out at the Jet Propulsion Laboratory, California Institute of Technology, was under a contract with the National Aeronautics and Space Administration (80NM0018D0004). Government sponsorship is acknowledged. AEGV is a postdoctoral fellow supported by UPLIFT: UCLA Postdocs' Longitudinal Investment in Faculty (Award # K12 GM106996). This is contribution number 7 of the Natural History Museum of Los Angeles County's Diversity Initiative for the Southern California Ocean (DISCO). Graduate student support was additionally provided by the National Council for Scientific and Technological Development of Brazil [Grant No. 209261/2014-5] and the University of California, Los Angeles Department of Ecology and Evolutionary Biology (EEB) Summer Research Fellowship. We thank the UC Natural Reserves System managers, other natural areas managers, and the hundreds of volunteers for collections. We thank A. Mahinan, N. Stavros, W.-Y Kwan, for assisting with spatial environmental data and to A. DeVries and L. Bulbenko for preparing extractions. We thank T. Gillespie, C.M. Mueller and Z. Kurtz for help optimizing analyses.

RSM, RKW, BAS, EEC designed the study. ML, RSM, RJH, ALS, MO, EEC and ZG selected analyses. RSM and EEC coordinated public sampling. ML, EF, TMS, and RSM made DNA libraries. ML, MPM, FDS, AEGV, DRR, RSM and EEC curated environmental metadata.

ML, FDS, and RJH generated statewide data layers. ML led biodiversity and gradient forest analyses, and ML, ALS, and RSM generated plots. EJM generated the synthetic phylogeny. All authors performed analyses and interpretation. ML, RSM, and RKW wrote the manuscript with input from all authors.

Chapter 2 is currently a version of a manuscript in preparation for submission to *Nature Communications*:

Nigenda-Morales, Sergio F.*, Meixi Lin*, Paulina G. Nuñez-Valencia, Christopher C. Kyriazis, Annabel C. Beichman, Jacqueline A. Robinson, Aaron P. Ragsdale, Jorge Urbán R, Frederick I. Archer, Lorena Viloria-Gómora, María José Pérez-Álvarez, Elie Poulin, Kirk E. Lohmueller, Andrés Moreno-Estrada, and Robert K. Wayne. 2022. "The Genomic Footprint of Whaling and Isolation in Fin Whale Populations." In Prep. (*: contributed equally)

Acknowledgments. We would like to thank the members of the research program of marine mammals from Universidad Autónoma de Baja California Sur for their help collecting the samples in the Gulf of California. Sergio Flores-Ramírez for his initial support in the Gulf of California project. Cei Abreu-Goodger for his support and providing laboratory space during the initial analysis of the data, Phil Morin for reviewing an early draft of the manuscript. Unpublished genome assemblies and sequencing data for the Bryde's whale and fin whale are used with permission from the DNA Zoo Consortium (dnazoo.org). For the fin whale DNAzoo assemblies, the sample for the assembly was collected by The Marine Mammal Center under the Marine Mammal Health and Stranding Program (MMHSPR) Permit No. 18786-04 issued by the National Marine Fisheries Service (NMFS) in accordance with the Marine Mammal Protection Act (MMPA) and Endangered Species Act (ESA). The work at DNA Zoo was performed under Marine Mammal Health and Stranding Response Program (MMHSRP) Permit No. 18786-03.

This work used computational and storage services associated with the Hoffman2 Shared Cluster provided by UCLA Institute for Digital Research and Education's Research Technology Group.

Funding sources. This work was supported by the Mexican National Council for Science and Technology (CONACYT) grant FONCICYT/50/2016, National Science Foundation (DEB Small Grant #1556705), UCMEXUS-CONACYT collaborative grant 2006. SNM was supported by CONACYT Postdoctoral Fellowship 724094 and the Mexican Secretariat of Agriculture and Rural Development Postdoctoral Fellowship. ML was supported by the University of California, Los Angeles Department of Ecology and Evolutionary Biology (EEB) Summer Research Fellowship. KEL and CK were supported by NIH grant R35GM11985 to KEL. ACB was supported by the Biological Mechanisms of Healthy Aging Training Program NIH T32AG066574. MJPA was supported by ANID under Grant Program FONDECYT Iniciación 11170182.

Author contributions. SN-M, AM-E and RKW conceived the study. AM-E and RKW contributed reagents, materials, and analysis tools. JUR, LV-G and FIA collected and contributed the samples and sample information. SN-M carried out laboratory work. SN-M, ML, PN-V and CK performed the analysis of the data. ACB, JAR and AR provided scripts for some analyses. AM-E, ACB, JAR, AR, MJP-A, EP and KEL provided guidance and advised the project. AM-E, KEL and RKW performed funding acquisition. SN-M, ML, PN-V, CCK and RKW wrote the manuscript with input from all the authors.

Data availability. The scripts used to perform the data processing and analysis are available in this github repository: https://github.com/snigenda/FinWhale_PopGenomics_2021. The raw sequence data will be available from NCBI's Sequence Read Archive database website. Additional supportive information will be available from a data repository like Dryad or Zenodo.

Declaration of interests. The authors declared that they do not have any conflict of interest.

Chapter 3 is a version of a manuscript in preparation for submission.

Lin, Meixi, Eduardo Guerra Amorim, Sergio F. Nigenda-Morales, Annabel C. Beichman, Paulina G. Nuñez-Valencia, Jonathan Mah, Jacqueline A. Robinson, Christopher C. Kyriazis, Christian Huber, Andrew E. Webb, Sarah D. Kocher, Frederick I. Archer, Andrés Moreno-Estrada, Robert K. Wayne, Kirk E. Lohmueller. 2022. "Variation of the Distribution of Fitness Effects Across Animals." In Prep.

Acknowledgments. For the mosquito dataset, we are grateful for the MalariaGen team with providing open access data from Ag1000G Phase 1 release. For the vaquita dataset, we thank Dr. Morin for providing early access to the SFS data. We thank the Southwest Fisheries Science Center's Marine Mammal and Sea Turtle Research Collection for use of archival vaquita tissue samples. All samples were imported to the US under appropriate CITES and US Marine Mammal Protection Act permits. We thank F. Gulland, The Marine Mammal Center, and NOAA Fisheries for funding genome resequencing. For the fin whale dataset, we would like to thank Cei Abreu-Goodger for his support and laboratory space during the initial analysis of the data. Samples were collected by the Southwest Fisheries Science Center (California, USA) in accordance with national guidelines and regulations. We are grateful for Y. Zhen, B. Kim, M. I. A. Cavassim, J. Garcia, and E. Wade for their helpful discussions. This work is supported by National Institutes of Health grant R35GM119856 to KEL.

Author contributions. ML and KEL conceived the study. ML, EGA, SN-M, ACB, PN-V and JAR generated the data. ML, EGA, ACB and JM performed the analysis of the data. CCK and CH provided scripts for some analyses. FIA collected and contributed the samples and sample information. KEL performed funding acquisition. ML wrote the manuscript. ML, KEL and RKW revised the manuscript with input from all the authors.

Appendix Chapter is a version of a manuscript submitted to the *Journal of Heredity*:

Lin, Meixi, Merly Escalona, Ruta Sahasrabudhe, Oanh Nguyen, Eric Beraut, Michael R. Buchalski* and Robert K. Wayne*. 2022. "A Reference Genome Assembly of the Bobcat, *Lynx rufus*" Submitted to the *Journal of Heredity*. (*: contributed equally)

Funding. This work was supported by the California Conservation Genomics Project, with funding provided to the University of California by the State of California, State Budget Act of 2019 [UC Award ID RSI-19-690224]. Sample collection was supported by funding from the U.S. Fish and Wildlife Service, Wildlife Restoration Act (Grant No. P1580009).

Acknowledgements. PacBio Sequel II library prep and sequencing was carried out at the DNA Technologies and Expression Analysis Cores at the UC Davis Genome Center, supported by NIH Shared Instrumentation Grant 1S10OD010786-01. Deep sequencing of Omni-C libraries used the NovaSeq S4 sequencing platforms at the Vincent J. Coates Genomics Sequencing Laboratory at UC Berkeley, supported by NIH S10 OD018174 Instrumentation Grant. We thank the staff at the UC Davis DNA Technologies and Expression Analysis Cores and the UC Santa Cruz Paleogenomics Laboratory for their diligence and dedication to generating high quality sequence data. We thank Dr. Devaughn Fraser, Dr. Laurel Serieys, Dr. Kirk Lohmueller, Dr. Brad Shaffer and Dr. Erin Toffelmier for inputs on study design; Mr. Daniel R. Oliveira, Dr. Courtney Miller and Ms.Tara Luckau for coordination on sample submissions; Mr. Barry Rowan and Dr. Laurel Serieys for sharing bobcat photos in Figure A.1. We thank the staff at Sonoma Wildlife Care and California's wildlife rehabilitation facilities for their generous help with providing the samples. This work used computational and storage services associated with the

Hoffman2 Shared Cluster provided by UCLA Institute for Digital Research and Education's Research Technology Group.

Author contributions. ML, MRB and RKW conceived the study. RS, ON, and EB performed the lab work. ME performed the genome assembly. ML and ME performed the analysis of the data. MRB collected and contributed the samples and sample information. MRB and RKW provided guidance and advised the project. MRB and RKW performed funding acquisition. ML, ME, MRB and RKW wrote the manuscript with input from all the authors.

Data Availability. Data generated for this study are available under NCBI BioProject PRJNA777191. Raw sequencing data for sample CCGP_SWC_20201006 (NCBI BioSample SAMN23391104) are deposited in the NCBI Short Read Archive (SRA) under SRR17978068 for PacBio HiFi sequencing data and SRR17978066-67 for Omni-C Illumina Short read sequencing data. GenBank accessions for both primary and alternate assemblies are GCA_022079265.1 and GCA_022079275.1; and for genome sequences JAJSDN000000000 and JAJSDO000000000. The NCBI RefSeq accession corresponding to the primary assembly GCA_022079265.1 is GCF_022079265.1. The GenBank organelle genome assembly for the mitochondrial genome is CM039064.1. Assembly scripts and other data for the analyses presented can be found at the following GitHub repository: https://www.github.com/ccgproject/ccgp_assembly. The scripts for genome assembly comparisons can be found at the following GitHub repository: https://github.com/meixilin/ccgp_bobcat_joh.

VITA

Meixi Lin

EDUCATION:

Bachelor of Science (*Magna cum laude*), Biological Science, Peking University	June 2017
Bachelor of Science (*Double Major*), Economics, Peking University	June 2017

TEACHING APPOINTMENTS:

Teaching Assistant:

1. EEB 108 “Biodiversity in the Age of Humans” Spring 2019, 2020, Winter 2021, UCLA
2. EEB 87 “California's DNA, a field course” Fall 2018, 2020, UCLA
3. EEB 136 “Restoration Ecology” Spring 2021, UCLA
4. LS 30A “Mathematics for Life Scientists” Winter 2020, UCLA
5. EEB 116 “Conservation Biology” Winter 2019, UCLA

PEER REVIEWED PUBLICATIONS:

1. **Lin, Meixi**, Ariel Levi Simons, Ryan J. Harrigan, Emily E. Curd, Fabian D. Schneider, Dannise V. Ruiz-Ramos, Zack Gold, et al. 2021. “Landscape Analyses Using eDNA Metabarcoding and Earth Observation Predict Community Biodiversity in California.” *Ecological Applications* 31 (6).
2. Meyer, Rachel S., Miroslava Munguia Ramos, Meixi Lin, Teia M. Schweizer, Zack Gold, et al. 2021. “The CALeDNA program: Citizen scientists and researchers inventory California's biodiversity.” *California Agriculture* 75 (1): 20–32.
3. Rocchini, Duccio, Matteo Marcantonio, Daniele Da Re, Giovanni Bacaro, Enrico Feoli, Giles M. Foody, Reinhard Furrer, et al. 2021. “From Zero to Infinity: Minimum to Maximum Diversity of the Planet by Spatio-Parametric Rao’s Quadratic Entropy.” Global Ecology and Biogeography 30 (5): 1153–62.
4. **Lin, Meixi**, Shan Zhang, and Meng Yao. 2019. “Effective Detection of Environmental DNA from the Invasive American Bullfrog.” *Biological Invasions* 21 (7): 2255–68.
5. Curd, Emily E., Zack Gold, Gaurav S. Kandlikar, Jesse Gomer, Max Ogden, Taylor O’Connell, Lenore Pipes, et al. 2019. “Anacapa Toolkit: An Environmental DNA Toolkit for Processing Multilocus Metabarcode Datasets.” *Methods in Ecology and Evolution* 10 (9): 1469–75.

SELECTED PRESENTATIONS:

1. AGA 2021 Symposium - Conservation Genomics, Oct 2021, Poster presentation. Title: Inference of the distribution of fitness effects of new mutations in the fin whale, a large-bodied mammal.
2. CABW/Cal-SFS Meeting, Oct 2020, invited 15 min oral presentation. Title: A Biodiversity Composition Map of California Derived from Environmental DNA Metabarcoding and Earth Observation.

3. GIBI, CAPES-UTFORSK and BRC Workshop, July 2020, invited 15 min oral presentation. Title: A Biodiversity Composition Map of California Derived from Environmental DNA Metabarcoding and Earth Observation.
4. 8th International Barcode of Life Conference, June 2019, 15 min oral presentation. Title: Mapping California biodiversity using remote sensing and community science eDNA surveys.
5. 8th International Barcode of Life Conference, June 2019, Poster presentation. Title: Evaluation of species-specific primers and extraction methods for eDNA-based detection of the American bullfrog.
6. 8th International Barcode of Life Conference, June 2019, Poster presentation. Title: Comparison of capture array, metabarcoding, and shotgun sequencing in recovering mammalian eDNA from contemporary soil samples.

SELECTED AWARDS AND HONORS:

University of California, Los Angeles

1. The Josephine Reich Quarter Fellowship, UCLA ($7500) Fall 2021
2. Departmental Research Travel Grant, UCLA ($3000) July 2021
3. Departmental Quarter Fellowship, UCLA ($7500) Fall 2019, Summer 2019,2020
4. Departmental Research Travel Grant, UCLA ($1700) July 2019
5. La Kretz Center & Stunt Ranch Reserve Research Grant (joint-award), UCLA ($2600) 2018
6. Departmental Research Travel Grant, UCLA ($1100) July 2018

SERVICE:

- Reviewer, Molecular Ecology Resources 2020
- Undergraduate Mentor, CALeDNA Summer Research Internship 2019, 2021
- Professional development officer, EEB Graduate Committee, UCLA 2020
- Host, eDNA Tea and Cookies weekly seminars, 2019-2020

Chapter 1: Landscape Analyses Using eDNA Metabarcoding and Earth Observation Predict Community Biodiversity in California

Originally published in *Ecological Applications*

Supplementary materials available online from *Ecological Applications*:

https://onlinelibrary.wiley.com/doi/10.1002/eap.2379

Abstract

Ecosystems globally are under threat from ongoing anthropogenic environmental change. Effective conservation management requires more thorough biodiversity surveys that can reveal system-level patterns and that can be applied rapidly across space and time. Using modern ecological models and community science, we integrate environmental DNA and Earth observations to produce a time snapshot of regional biodiversity patterns and provide multi-scalar community-level characterization. We collected 278 samples in spring 2017 from coastal, shrub, and lowland forest sites in California, a complex ecosystem and biodiversity hotspot. We recovered 16,118 taxonomic entries from eDNA analyses and compiled associated traditional observations and environmental data to assess how well they predicted alpha, beta, and zeta diversity. We found that local habitat classification was diagnostic of community composition and distinct communities and organisms in different kingdoms are predicted by different environmental variables. Nonetheless, gradient forest models of 915 families recovered by eDNA analysis and using BIOCLIM variables, Sentinel-2 satellite data, human impact, and topographical features as predictors, explained 35% of the variance in community turnover. Elevation, sand percentage, and photosynthetic activities (NDVI32) were the top predictors. In addition to this signal of environmental filtering, we found a positive relationship between environmentally predicted families and their numbers of biotic interactions, suggesting environmental change could have a disproportionate effect on community networks. Together, these analyses show that coupling eDNA with environmental predictors including remote sensing data has capacity to test proposed Essential Biodiversity Variables and create new landscape biodiversity baselines that span the tree of life.

Key words: environmental DNA; citizen science; ecological modeling; gradient forest; remote sensing; community ecology; beta diversity; zeta diversity; biomonitoring

Introduction

Species are being rapidly lost worldwide (Pimm et al. 2014, Ceballos et al. 2015, Díaz et al. 2019) with many key habitats that harbor high biodiversity (Myers et al. 2000) threatened by climate change and environmental degradation. The scientific community needs rapid bioinventory tools to provide critical baseline biodiversity data with minimal cost and effort that can be applied globally (Bush et al. 2017). Essential Biodiversity Variables (EBVs; Pereira et al. 2013) are a minimal set of measurements needed to support multi-purpose, long-term planning at various scales. Example EBVs include community composition, genetic composition, and ecosystem structure, which can be extrapolated from in situ and remote sensing observations. Scaling up from in situ biological measures to enable system-wide projections remains challenging (Pereira et al. 2013). Bioinventories remain often taxonomically or spatiotemporally restricted because technical feasibility limits large scale monitoring (Cristescu 2014), and thus, very few studies attempt to assess the complex composition of the total biotic environment (Karimi et al. 2018, George et al. 2019) that could provide unbiased EBVs needed to aid systems-level biodiversity conservation.

Technology-assisted citizen and community science (CCS) is a growing means to obtain in situ biodiversity observations to complement those made by taxonomic experts, and CCS observations from photographs and sounds have already eclipsed other biomonitoring data records such as physical collections (Theobald et al. 2015, Kobori et al. 2016). However, most CCS observations favor diurnal macroscopic species and often omit cryptic and microbial taxa (Theobald et al. 2015). In response, our program, CALeDNA (by the University of California Conservation Genomics Consortium; CALeDNA 2021), and several other fledging programs, have focused on giving community scientists the capacity to sample environmental DNA (eDNA) from their surroundings (Biggs et al. 2015, Miralles et al. 2016, Meyer et al. 2021), which can be probed for nearly any taxonomic group using multi-locus metabarcoding methods (Bohmann et al. 2014, Deiner et al. 2016, Thompson et al. 2017, Franklin et al. 2019).

Multi-locus metabarcoding of eDNA from surface soil and sediment retains a record of taxa recently present in the local area, including bacteria and archaea, often-overlooked meiofauna, protozoans, non-vascular plants, algae, and fungi in addition to the vertebrate and vascular plant communities that are easier to observe directly. These methods are increasing in accuracy as reference DNA sequence databases grow and informatic tools improve, and are decreasing in cost as library preparation and sequencing technology become less expensive. Community-powered eDNA surveys can be coupled with remote sensing measures of ecosystem properties to model community composition, generate EBVs and advance ecological theories about how community diversity is regulated by biotic and abiotic traits (Yamasaki et al. 2017). On the ground and space-based technologies yield increasingly copious and accessible abiotic data (Pettorelli et al. 2014, Schimel et al. 2019) on land cover, topography, soil property (Hengl et al. 2017), bioclimate (Fick and Hijmans 2017), human impact (WCS and CIESIN 2005), and vegetation (e.g., Sentinel-2; European Space Agency), which can be used to model eDNA biodiversity changes across landscapes (Crowther et al. 2019, van den Hoogen et al. 2019). Biotic-abiotic interactions among soil properties (e.g., pH and nutrient availabilities), climate, plant coverage, and habitat type have been shown to affect soil alpha and beta diversity in different taxonomic groups (Fierer and Jackson 2006, Ranjard et al. 2013, George et al. 2019, White et al. 2020) from tropical mountains to temperate ecosystems (Thompson et al. 2017, Karimi et al. 2018, Montagna et al. 2018, Peters et al. 2019). However, these studies have largely focused on a single habitat, region, or phylogenetic clade with few exceptions, notably, a national-scale soil eDNA survey in England showed that animal and microbial richness responded to different environment factors but beta-diversity trends were shared across taxonomic groups (George et al. 2019).

Our study attempts to use multi-locus metabarcoding from CCS-collected eDNA in a biodiversity-ecological response model that spans kingdoms and habitats of California. Similar to other biodiversity hotspots, we expect discontinuous environmental clines and high endemism (Myers et al. 2000, Thompson et al. 2017) to be apparent in eDNA community patterns. Our objectives are threefold. First, we identify the taxonomic occurrence patterns recovered in eDNA surveys and assess their reliability and concordance with traditional observations. Second, we assess the relationship of eDNA alpha, beta, and zeta diversity to environmental measures to determine how the environment filters species richness and community composition. Third, we apply joint-species gradient forest and ecological co-occurrence network modeling to generate a community turnover map of the entire state of California and characterize the taxonomic families that are found to be most sensitive to environmental filtering. These analyses reveal the abiotic and biotic variables that are the most predictive of community composition patterns and provide a framework for using CCS-generated eDNA with remote sensing to refine static maps of ecological delineations and provide effective EBVs.

Methods

Sampling design

Volunteers for CALeDNA sampled biodiversity from a wide variety of habitats, including coast, shrub, and lowland forest sites across the state of California using target sampling and eDNA metabarcoding. Sample location metadata were collected by a smartphone webform made in Kobo Toolbox and included a photograph (http://kobotoolbox.org/). Surface samples were collected by filling three 2-mL tubes with substrate from <2 cm depth, each 30 cm apart. Samples were frozen at −80°C immediately upon their return to CALeDNA headquarters at UC Los Angeles.

To minimize the potential effect of seasonal variations in eDNA profiles, we selected samples from March 2017 to June 2017, with two-thirds of samples collected in April. We classified the predominant biome using photographs and a variety of geolocation data. We selected 100 samples from each of three transect types, coast, shrub/scrub (abbreviated as "shrub"), and forest, that covered the broadest latitudinal range possible. Samples with ambiguous metadata were removed, resulting in a total of 278 samples (98 coast, 89 shrub, and 91 forest) used in subsequent analyses (Table 1.1; Data S1.1).

Compilation of environmental variables

We assembled environmental variables across six main categories: location, habitat, bioclimate, soil properties, topography and vegetation (including surface reflectance properties) variables (Supplemental Methods; Figures S1.1, S1.2; Data S1.1). Uncertainty layers were downloaded if available as well (Figure S1.3). All raster layers were aligned and projected to a unified 100 x 100 m grid from Google Earth Engine (Coordinate Reference System for this project: ESPG 4326, WGS84). Layers were stacked and clipped to California's extent, and used for point extraction. For coastal sites outside of the raster's geographical coverage, values were extracted by the closest point available in 0.5 km radius or assigned an "NA" value if not available. All computation and analyses were performed in R version 3.5.3 (R Core Team 2019). Raster operations were performed using R package *raster* (v. 2.8-19; Hijmans 2019).

Considering that many environmental variables are correlated, we evaluated the Pearson's correlation coefficient of the 56 numerical environmental variables and hierarchically clustered the variables according to the coefficients into variable groups using R functions *cor*, *hclust* and *cutree*. To reduce collinearity and improve interpretability in community modeling, we created a "reduced" set of 33 numerical environmental variables that had an $R^2 < 0.8$ (Table 1.1; Figures S1.1, S1.2) for downstream analysis.

DNA extraction, amplification and sequencing

DNA extraction, amplification and sequencing followed Curd et al. 2019. Briefly, three 250 mg biological replicate soil samples from each site were fully homogenized and pooled per site. DNA was extracted using QIAGEN DNeasy PowerSoil Kit (Qiagen, Valencia, CA, USA) according to the manufacturer's instructions. Negative controls were included in every batch of 12-18 extractions. DNA was amplified by polymerase chain reaction (PCR), using primers for five barcode regions: *16S* (515F and 806R; Caporaso et al. 2012), *18S* (Euk_1391f and EukBr; Amaral-Zettler et al. 2009), *CO1* (mlCOIintF and Fol-degen-rev; Yu et al. 2012, Leray et al. 2013), fungal *ITS1* ("*FITS*"; ITS5 and 5.8S; White et al. 1990, Epp et al. 2012), and plant *ITS2* ("*PITS*"; ITS-S2F and ITS-S3R; Gu et al. 2013). Primer sequences and thermocycling profiles can be found in Tables S1.1 and S1.2. All PCR amplifications were performed in triplicate and with additional PCR negative controls. Triplicate positive amplifications confirmed by gel electrophoresis, were pooled by sample and barcode to equimolar levels, indexed and sequenced on an Illumina MiSeq v6 platform for 2x300 bp reads (QB3-Berkeley FGL) with a target sequencing depth of 50,000 reads/sample/metabarcode (Supplemental Methods). Five of the 278 sites were processed as biological replicates by different technicians to inspect taxonomic variation in independent DNA extraction and technical processing.

Bioinformatics and data processing

We used default settings in the Anacapa Toolkit (Curd et al. 2019) for multi-locus sequence data processing and taxonomy assignment. In brief, quality control of raw sequences was performed using Cutadapt (Martin 2011) and FastX-Toolkit (Gordon et al. 2010), and inference of Amplicon Sequence Variants (ASVs) was made with DADA2 (Callahan et al. 2016). Taxonomy assignment was made on each ASV using Bowtie2 (Langmead and Salzberg 2012) and the Bayesian Lowest Common Ancestor algorithm (BLCA; Gao et al. 2017) on custom metabarcode-specific reference databases that were created using Creating Reference libraries Using eXisting tools (CRUX; Curd et al. 2019). Taxonomy assignments with a bootstrap confidence cutoff score over 0.6 were kept for each ASV. ASVs with the exact same inferred LCA passing confidence filter were summed into one "taxonomic entry" as the species/phylotype/MOTU equivalent in this study (Supplemental Methods).

To informatically control for contamination, we further removed all singleton or doubleton taxa, and removed taxa that occurred in more than two reads in all blank samples, from subsequent analyses. To prepare data for alpha and beta diversity analyses requiring rarefaction, we performed rarefaction in 10 replicates and took the mean using the custom_rarefaction function in the R package ranacapa (v. 0.1.0; Text S1.1; Table S1.3; Kandlikar et al. 2018). Reads with no assignment were not removed before rarefaction. We also estimated concordance between biological replicates (Text S1.2).

Comparison of eDNA taxonomic output with traditional surveys

To compare the eDNA taxonomic results to traditional surveys, we compared eDNA results to the curated species inventory of the University of California Natural Reserve System (UCNRS), which records Chordata, Arthropoda, and Streptophyta. We counted how many taxon records were shared or unique to eDNA results or traditional records at classification levels of order, family and genus combining all reserves and within each reserve.

We developed a metric of traditional observation score (TOS) in eDNA taxonomic assignment. TOS uses all observation and collection records in the Global Biodiversity Information Facility (GBIF) database from a broad region centered on California to score whether the taxon assignment of an eDNA ASV has been observed. A TOS > 0 suggests there is support for the assignment of an ASV based on its presence in the TOS region (Supplemental Methods).

Community alpha, beta and zeta diversity relationships with environmental variables

We used the rarefied dataset for alpha and beta diversity analyses to control for variations in read depth. Alpha diversity was calculated using Observed and Shannon's Diversity Index in the R package vegan (v. 2.5-2; Oksanen et al. 2018). These two measures weigh relative sequence abundance differently. Shannon's Index penalizes rare sequences compared to the Observed Index (Calderón-Sanou et al. 2020). We evaluated relationships of alpha diversity measures using the Kruskal-Wallis Test for categorical environmental variables, and individual linear models and partial least squares models for numerical variables (Supplemental Methods, Text S1.3).

Beta diversity was visualized by plotting sample relative abundance of the top ten phyla for metabarcodes 16S, 18S, and CO1, and top ten classes for PITS and FITS. Composition profiles were analyzed using unconstrained ordination to reveal turnover across sites. We calculated the binary Jaccard dissimilarity distance to only consider presence-absence patterns given eDNA relative abundance can be influenced by stochastic processes of DNA shedding, deposition, and decay. We performed principal coordinate analysis (PCoA; function ordinate), permutational multivariate ANOVA analysis (PERMANOVA; function adonis), and tested for the assumption of homogeneity of dispersion (function betadisp) in the R packages phyloseq (v. 1.24.2; McMurdie and Holmes 2013) and vegan. We also partitioned the data by the four categories in the majorhab variable (aquatic; herbaceous; shrub and tree dominated habitats) and performed PCoA and PERMANOVA analyses within each major habitat. Additionally, we tested for the effects on community turnover of coastal sites and spatial correlation (Text S1.4). Post hoc explanation of the ordination axes was performed by fitting the reduced set of numerical variables (Table 1.1) onto the PCoA result using functions envfit and ordisurf in the R package vegan (Supplemental Methods).

Zeta diversity was used to measure the fraction of unique categories of organisms held in common among nearby sets of communities, which unlike beta diversity, considers the composition of metacommunities composed of more than two sites. We set cluster size to four nearby sites, calculated and scaled zeta four diversity (ζ_4) using the R package ZETADIV (v. 1.1.1; Latombe et al. 2018). We tested the likelihood of two model forms of the relationship between zeta diversity and sample numbers (zeta decline). Based on prior analyses (Hui et al. 2014), declines which follow a power-law of the form $\zeta_N = \zeta_1 N^{-b}$, or an exponential of the form $\zeta_N = \zeta_1 e^{b(N-1)}$, were associated with a niche differentiation or stochastic process of community assembly, respectively (Supplemental Methods). Scaled ζ_4 diversity values were then plotted on a map of California using the R package Leaflet (v. 2.0.2; Cheng et al. 2018). Environmental factor groups were made by binning environmental variables according to their categories (Table 1.1). We used generalized linear models (GLM) to determine the variation in ζ_4 diversity attributed to either geographic distance or an environmental factor group.

Gradient forest modeling and ecological network analysis to predict and interpret community turnover across California

We used the gradient forest classification model in the R package gradientForest (v. 0.1-17; Ellis et al. 2012) to test which environmental variables best explained eDNA-detected community turnover patterns across California using all 272 sites without any missing metadata collected from three transects (six out of 278 sites excluded due to missing metadata). We chose to perform predictive modeling on beta diversity because it is less affected by molecular artefacts, such as PCR errors or tag-jumps, or variations in bioinformatics pipelines, and more likely to reflect ecologically meaningful community composition patterns compared to alpha diversity, which is more sensitive to eDNA processing strategies (Calderón-Sanou et al. 2020, Shirazi et al. 2020) and does not require the clustering of sites that zeta diversity does. Due to large variation in the coastal sites, we also performed additional analyses excluding all coastal sites using the same methods described below. The gradient forest model was built with the reduced set of 33 numerical environmental variables (Table 1.1). We fit a classification-tree based gradient forest model using default settings to the eDNA-derived biological matrix, but increased the number of trees to 2000 per family to increase the stability of the model (Breiman 2001). To assess model robustness, we repeated the gradient forest model 20 times. To assess model power and reliability, we randomized the predictor matrix 100 times and ran the model with the same settings (Bay et al. 2018; Supplemental Methods).

To visualize the community turnover gradient forest model over space, we used the input of all 33 environmental variables from 100 m x 100 m grids in the extent of California without extrapolation (Pitcher et al. 2011). We used the top three principal components from the transformed environmental variables and visualized them by red, green and blue (RGB) bands

(Ellis et al. 2012). To differentiate model performance from the high-dimensional nature of the environmental variable matrix and provide prediction uncertainty estimates, we scaled the environmental variables and performed the same PCA and visualization procedure without using the model ("uninformed map") and performed a mantel test and a monotonic regression between the biological matrix and either the uninformed map or gradient forest informed map. We also estimated which area contained more uncertainty by mapping the sites in gradient forest informed map to the biological matrix using a Procrustes rotation and evaluated the residuals (Ellis et al. 2012; Supplemental Methods).

To explore the biotic interactions underlying the gradient forest patterns, results for each metabarcode were summarized by family, filtered on read depth and frequency and used in ecological co-occurrence network analysis using the R package SpiecEasi (v. 0.1.4; Kurtz et al. 2015) for cross domain analysis that incorporates all five metabarcodes into one complex network (see Tipton et al. 2018). Topological parameters were determined in Cytoscape (v. 3.6.1; Shannon et al. 2003) using the NetworkAnalyzer tool. To observe the relationship between network degrees and the prediction R^2 of each family from gradient forest, an ordinary least squares (OLS) linear regression model was made using the lm function in R and interactions were visualized with R package Interactions (v. 1.1.1; Long 2019). To evaluate the co-occurrence and gradient forest predictor patterns in a phylogenetic framework, the 915 families used in the gradient forest modeling were mapped onto the Open Tree of Life (tree.opentreeoflife.org) and a synthetic tree was generated using synthesis release v12.3. Datasets were mapped next to the phylogeny tips using the Interactive Tree of Life (https://itol.embl.de/).

Results

eDNA metabarcoding recovered taxonomic entries across 86 phyla

The 278 selected samples from coast, shrub, and forest areas across California (Figure 1.1A) were sequenced with five metabarcodes. Each metabarcode recovered their target groups as expected (Figure 1.1C; Table S1.1), with 16S amplifying Bacteria and Archaea, 18S and CO1 broadly amplifying eukaryotes including Animalia, Chromista, Fungi, Protozoa and some Plantae, ITS1 amplifying Fungi ('FITS') from Ascomycota, Basidiomycota and other phyla, and the ITS2 region amplifying plants ('PITS') across both Chlorophyta and Streptophyta.

Sequencing the 278 samples, five repeated "biological replicate" samples, and 23 negative controls as PCR blanks or extraction blanks amounted to 75,830,796 reads for the five metabarcoding loci and averaged 54,554 reads per sample per metabarcode. After several steps of quality control, taxonomic assignment and sequence decontamination, a total of 16,157,425 reads were assigned to 16,118 unique taxonomic entries, i.e. best taxonomic hypotheses (Data S1.2). The median assigned read depth was 7,717 (Figure S1.4) and mean taxa identified was 778 per sample. Assignments spanned 86 phyla with most reads and taxonomic entries being assigned to Proteobacteria, Ascomycota and Basidiomycota (Figure 1.1B, 1.1C). Despite fairly deep sequencing, stringent sample filtration and validation on eDNA result concordance were necessary to meet sufficiency metrics practiced by the metabarcoding community (Goldberg et al. 2016, Taberlet et al. 2018; Text S1.2; Figure S1.5; Data S1.3). Sequence rarefaction for diversity analyses that require even read depth across samples was able to be set near the taxon accumulation curve asymptote, suggesting we did not undersample during sequencing, although we did have to remove a small number of sample sites to meet the depth requirement (Text S1.1; Figures S1.6, S1.7; Table S1.3).

Comparison with traditional surveys: eDNA results partially overlap with traditional observations

Our first objective to assess the concordance between eDNA surveys and traditional observations initially utilized the UC Natural Reserve System curated species list of Streptophyta, Arthropoda and Chordata made by traditional surveys. Forty-four Streptophyta families were only found in eDNA, 77 were only in traditional observations, 65 were recovered from both methods. We found that 110 Arthropoda families were only recovered from eDNA, 139 were only in traditional observations, and 16 were recovered from both methods. No Chordata families were jointly recovered from both methods, since our metabarcoding markers did not specifically target Chordata. Evaluating concordance at order, family, and genus levels, we determined that family was the classification level that could be best validated by traditional observation at our UCNRS sample sites (Data S1.4).

To further evaluate eDNA and traditional observation concordance without relying on restricted local surveys, we assigned a Traditional Observation Score (TOS) for eDNA taxon entries using the GBIF records from Western North America and the Eastern Pacific which represent hypotheses of correct matches if eDNA entries overlap with the region specific GBIF records. Only taxonomic entries resolved to at least the level of order were assigned a TOS, hence 1700 eDNA entries were omitted. Results showed only 5.6% of eDNA entries had an adjusted TOS of 0 (no GBIF support for assignment), and 50.0% of entries had an adjusted TOS of 1 (strong GBIF support for assignment; Data S1.5). Partial concordance was found in the remaining entries. No relationship was found between TOS and the frequency at which a taxon was found in eDNA samples (Pearson's R^2=0.004; P< 1e-5), suggesting the TOS is not heavily biased toward common or ubiquitous taxa. As with the UCNRS comparison, the TOS was highest at the family level, so we selected family level classification for downstream gradient forest and network analyses.

Beta and zeta diversity are structured by minor habitat and vegetation variables

We examined relationships of alpha, beta and zeta diversity to environmental measures as our second objective. Alpha diversity varies at the local scale and across the terrestrial-marine interface (Figure S1.8), with high spatial stratification among loc (reported location names) and minorhab (minor habitat) variables for all metabarcodes besides CO1 (Figure S1.9), and stratification for the clust variable (neighboring cluster of sites within a radius of 0.5 km derived from GPS record) in Shannon Index for 16S and FITS (Data S1.6), indicating bacterial and fungal alpha diversity are locally constrained in California. Post-hoc Dunn tests of categorical groups (Figures S1.10 – S1.13; Data S1.6), as well as individual linear regressions (Data S1.7) and partial least squares models (Data S1.8) of observed richness and Shannon's diversity indices with numerical environmental observations showed alpha diversity is predicted by many environmental variables and is most strongly predicted in fungi (FITS; Text S1.3; Figure S1.14; Data S1.6 – S1.8).

Similarly, beta diversity patterns exhibited variations by habitat characteristics and were structured by environmental filtering. We found visually apparent differences in dominant taxa by habitat grouping (Figure S1.15). In community dissimilarity analyses, beta diversity was significantly different across major habitat groups despite many overlapping sites in the ordination plots (PERMANOVA; Figures 1.2A, 1.2B, S1.16 – S1.19; Data S1.9). In particular, samples from aquatic environments were more dispersed in the ordination (Figure 1.2A, 1.2B). Beta dispersion also showed significant heterogeneity of multivariate dispersion (variance) within groups for all metabarcode and category combinations except loc, majorhab, transect, and clust for the PITS metabarcode (Data S1.9).

Further investigation into beta diversity patterns revealed that minor habitat (minorhab) composition within each of the four major habitats contributed strongly to dissimilarity in all markers (PERMANOVA, adjusted $P < 0.01$; Figures 1.2C, S1.20; Data S1.10). Jaccard dissimilarity PCoA revealed finer-scale habitat partitions for some, but not all, minor habitat categories, suggesting eDNA may be useful to complement minor habitat classifications as distinct management units (McKnight et al. 2007). For example, within aquatic major habitat, many of the marine nearshore categories overlapped, while marine and freshwater (lacustrine and riverine) sites separated (Figures 1.2C, S1.20). Patterns of environmental filtering remained after exclusion of coastal sites and spatial correlation effects (Text S1.4; Figures S1.21, S1.22; Data S1.11, S1.12). For numerical variables, post hoc explanation of the ordination axes showed that photosynthetic activities (NDVI32 and greenness) were most highly correlated with 16S, 18S and FITS (Table 1.2; Figure S1.23; Data S1.13). Soil organic carbon content (orcdrc) was most highly correlated with CO1, and Isothermality (bio3) was most highly correlated with PITS (Table 1.2).

Zeta diversity describes the degree of overlap in the number of unique categories of organisms held in common between N sites or communities (ζ_N) (Figure S1.24A), which as N increases captures more variation due to turnover. This framework allows for an assessment in trends in regional scale turnover of relatively common organisms which are less biased towards the presence of rare, or spuriously detected taxa (Hui et al. 2018). Environmental factor groups explained 1 to 32% of the observed variation in ζ diversity (Table 1.3). Vegetation variables were among the top predictors for 18S, CO1, FITS and PITS datasets, with the highest variance explained at 32% for the FITS dataset. Variables related to small-scale location describe minimal variation (< 1%) in ζ_4 diversity for communities (Table 1.3). To better understand the likeliest processes associated with the spatial assembly of communities, two models of zeta diversity decline were tested using the power law model and the exponential model. The power law model was found to be a better fit for more than 83% communities described in all but the PITS metabarcode results, 31% of which followed the exponential model, suggesting lower spatial autocorrelation in plant and algal communities (Figure S1.24; Data S1.14).

Gradient forest models map high resolution biodiversity turnover in California

Our third objective used gradient forest and ecological co-occurrence network modeling to map and characterize the taxonomic families that are predicted by the environment. Our gradient forest model included 272 sites x 915 eDNA-derived families as a response variable matrix and 272 sites x 33 environmental variables as a predictor matrix (Data S1.15). The gradient forest model explained 35% of variation in the biotic matrix, and all 915 families were able to be effectively modeled (i.e. had an $R^2 > 0$) with high stability across 20 replicated runs (Average $R^2 = 0.349 \pm 0.0004$; Average families effectively modeled = 915 ± 0; Data S1.16). Using a permutation approach, we confirmed the mean overall R^2 and number of families with positive R^2 for true observations were significantly higher than all the permuted runs (Figure S1.25). Many of the most responsive families were from marine aquatic sites, and some of these were low in observation frequency (Figures 1.3B, S1.26).

Gradient forest provides information on the rate of community turnover along environmental gradients (Ellis et al. 2012). We plotted the relative density of splits and cumulative importance for environmental variables. Within the top three environmental variables, we found nonlinear community changes. For elevation, rapid community turnover (high splits density) occurred at 0 m and above 1,000 m (Figure 1.3C, 1.3D). For sand percentage, important splits were mainly distributed at 23%, 43% and 74% sand (local maxima with the highest density; Figure 1.3C, 1.3D), which have similarity to the soil texture triangle in the USDA system (Groenendyk et al. 2015). For photosynthetic activities (NDVI32), important splits were mainly distributed along -0.16, 0.05, and 0.28 (scale: -1 to 1; Figure 1.3C, 1.3D).

Our map of California biodiversity resembled EPA North America Level II and California Level III Ecoregion maps (U.S. Environmental Protection Agency 2010, 2012), which were created with different input data and methods (Figure 1.4C – 1.4E). For example, in the gradient forest map (Figure 1.4A), the majority of central and southwestern CA community type (red) corresponded to Mediterranean California (Figure 1.4C, pale green, Level II 11.1.), characterized by medium photosynthetic activities (NDVI32), lower elevation (elev), and higher precipitation seasonality (bio15).

We assessed the model prediction robustness and prediction uncertainties by regenerating our community turnover map of California without using any information obtained from eDNA surveys (Figure 1.4B), and the resulting map neither resembled California published maps such as the EPA North America Level II Ecoregion map (U.S. Environmental Protection Agency 2010, Omernik and Griffith 2014; Figure 1.4C) nor did it separate regions as sharply as the eDNA-informed map (Figure 1.4A). This purely physical approach of community turnover mapping showed adding eDNA improves gradient forest informed mapping by a 1.4% reduction in stress performance statistics and a 5.6% increase in Mantel correlation R^2 (Figure S1.27). We quantified the prediction uncertainties at each site by Procrustes rotation errors and found that predictions for coastal sites harbor more deviation from real eDNA communities (Dunn test, $P < 0.001$; Figure S1.28). We also were curious how robust our map was when coastal sites were removed, since several of the most predicted families were marine, and found that we could still explain 30% of the variation in the biotic matrix (Text S1.5; Figure S1.29).

Biotic co-occurrence has a weak positive relationship with gradient forest predictability

To characterize the biotic relationships of families across the spectrum of their predictability in the gradient forest models, which indicates environmental filtering (Horner-Devine et al. 2007), we modeled the relationship between each family's ecological co-occurrence network degrees and their predictor R^2 using an OLS linear model. Co-occurrence patterns reflect biotic niche processes that maintain biodiversity patterns which theoretically hold no expected relationship with abiotic environmental filtering. A family-level co-occurrence network produced 916 edges connecting 290 nodes (families) out of the total 304 families that met minimum frequency thresholds for analysis (Figure 1.5A; Data S1.17). In the OLS linear model, interaction effects of site frequency were also considered. Model results showed a modest positive relationship (Adj R^2= 0.22) between the number of edges and R^2 for families, indicating the families determined by gradient forest to be under more environmental filtering (higher R^2) were also the families most integrated in ecological networks based on their numbers of degrees. However, the interaction between frequency in sites and network degrees was also significant ($P<0.02$; Figure 1.5B). In a phylogenetic analysis of these patterns, we observed that families with high network degrees and high gradient forest predictor values were widely distributed across clades and kingdoms, but most frequent in the clades containing the class Flavobacteriia and the SAR supergroup (Stramenopiles, Alveolates and Rhizaria; Figure 1.5C), suggesting ecological networks containing these families might have the lowest resilience under abiotic change.

Discussion

Species observations by the public will continue to outpace both field collections and on-the-ground observations made by scientists (Theobald et al. 2015). With eDNA as a CCS tool (Biggs et al. 2015, Miralles et al. 2016, Larson et al. 2020), broader taxonomic inventories and assessments from minimally invasive environmental collections can be accomplished. Soils and sediments used in this study, collected by CCS volunteers, had an average of 778 taxonomic lineages identified in each DNA sample, and were easily obtained from a broad area within a seasonal snapshot. Co-analysis of eDNA from these collections and readily available environmental data provides predictor values for hundreds of families that evade traditional observations.

Our first objective concerning the concordance between eDNA results and traditional observations revealed relatively low overlap with UCNRS surveys, despite high support by GBIF traditional observation score, which suggests eDNA CCS surveys complement but do not replace traditional surveys. Ongoing efforts to sequence species and build a global taxonomic biodiversity reference database in the next decade (e.g., the Earth BioGenome Project, Lewin et al. 2018; the Centre for Biodiversity Genomics, Hobern 2020) are positioned to ameliorate shortcomings of current DNA reference sequences. Emerging alternatives to metabarcoding may additionally help mitigate detection bias currently in favor of small body size in eDNA studies (Figure 1.1; Data S1.4, S1.5). For example, DNA capture approaches to target larger organisms (Seeber et al. 2019) may improve detection of large-bodied species, but these are not yet as cost-effective for CCS as multi-locus metabarcoding. Another challenge is that different DNA extractions from the same soil or sediment sample exhibit heterogeneity (Text S1.2; Data S1.3). We are examining stability and stochasticity of taxonomic profiles under varied sample processing (Castro et al. in review) and DNA library preparation steps (Shirazi et al. 2020) in response to calls for research about these potential biases (Prosser 2010, Goldberg et al. 2016). In this study we used several standard approaches for reducing these biases.

Our second aim to test predictors of alpha, beta, and zeta diversity revealed that most environmental categories can significantly partition samples according to taxonomic composition (Figures 1.2, S1.15 – S1.20; Data S1.6 – S1.13), suggesting that surface communities are largely filtered by ecological rather than neutral processes (Bahram et al. 2018). These patterns remained significant after exclusion of coastal sites and location effects (Figures S1.21, S1.22; Data S1.11, S1.12). However, we found substantial overlap in community composition ordinations, as has been shown in the global Earth Microbiome Project (Thompson et al. 2017) and regional soil biodiversity ordination plots (George et al. 2019; Figure 1.2A, 1.2B; Data S1.9). In our ordinations, groups separated from each other when fine-scale categories are used, such as minor habitat within partitioned major habitat, suggesting a large amount of community partitioning is harbored within major habitats categories (Figure S1.20). We found prokaryotic diversity was particularly diagnostic of minor habitats in ordinations (Figures 1.2C, S1.20). We propose eDNA-based composition could be EBVs for planning management units such as minor habitat delineations and for detecting ecotones (Jetz et al. 2019).

Environmental variables (Tables 1.2 and 1.3) can have power to predict general biotic patterns and can illuminate possible drivers of community turnover (Figure S1.23) because they can readily be compared across studies (Omernik and Griffith 2014). For example, photosynthetic activities (NDVI32/greenness) had the highest correlation with the observed fungal alpha diversity pattern and beta diversity structure in bacteria (16S), eukaryotes (18S) and fungi (FITS) in the envfit analyses (Table 1.2; Figure S1.23). We note indices of photosynthetic activity have not been included as part of most microbiome studies (Karimi et al. 2018, Bahram et al. 2018, George et al. 2019) so their importance is still being discovered. For the subset of studies we found that had included NDVI as a predictor, it was observed to be important in modulating soil fungal and herbivore nematodes communities (Timling et al. 2014, Delgado-Baquerizo et al. 2016, Yang et al. 2017, van den Hoogen et al. 2019). Isothermality (bio3) has strong positive associations with PITS beta diversity turnover, suggesting inland arid California regions with low isothermality display nestedness in the biodiversity encompassed by these markers, as has been shown with plants in Australia (Gibson et al. 2012) and in South American seasonally dry forests (Silva and Souza 2018). Organic carbon (orcdrc) was strongly associated with CO1 community turnover, which mirrors associations reported in soil meiofaunal communities, particularly nematodes (Jackson et al. 2019). Overall, zeta diversity largely supports the envfit results, although zeta diversity had poorer explanatory power for 16S patterns, which can be attributed to its greater sensitivity to common groups (Table 1.3; Simons et al. 2019) such as the nearly ubiquitous taxa in Proteobacteria.

Previous efforts have successfully integrated abiotic environmental data and models with traditional observational records such as herbarium specimens (Baldwin et al. 2017) to produce maps used to conserve threatened species (Jenkins et al. 2015), assess deforestation (Zarnetske et al. 2019) and evaluate species richness and endemism (Baldwin et al. 2017). However, remotely sensed variables such as from the Sentinel-2 instrument and local-scale eDNA observations of taxonomy biodiversity enable community mapping at a grid size finer than 5 km (Jenkins et al. 2013, 2015, Pimm et al. 2014, Baldwin et al. 2017, Zarnetske et al. 2019), which aligns better with in situ biodiversity (e.g., Wang et al. 2018). Our objective to project community composition across California's landscape achieved a higher resolution than currently available statewide maps (Figure 1.4). Elevation (elev), sand percentage (sndppt), photosynthetic activities (NDVI32) and mean temperature in the wettest quarter (bio8) were the among the most important predictors (Figure 1.3A) and all of these variables had been proposed to be prominent drivers in community structures worldwide. For example, sand percentage, an inverse of clay percentage, is known to explain differences in plant community guilds (Cornelius et al. 1991), correlates with presence of halophytes (Lee et al. 2016, Moreno et al. 2018) and influences microbial community structures (Sessitsch et al. 2001, Ehrlich et al. 2015).

Space, flight, tower and drone-based remote sensing information are becoming increasingly available and accessible (Pettorelli et al. 2014). By providing more direct, spatially continuous measures of plant functional diversity and ecosystem functioning at regional (Schneider et al. 2017, Durán et al. 2019) to global scales (Schimel et al. 2019, Schneider et al. 2020), we expect that future analyses will uncover new important environmental predictors and develop prediction maps on species richness (alpha diversity) or community turnover at higher dimensions (zeta diversity), expanding on the beta diversity map presented here. eDNA composition could potentially be better predicted with more remote sensing and in situ bioinventory data from different spatial and temporal scales with improved gradient forest R^2 from what we achieved at $R^2 = 0.35$ and decreased prediction uncertainties. Bayesian hierarchical modeling and artificial neural networks are also receiving increasing attention for community modeling with more application potentials for improved spatial-temporal biodiversity predictions with associated uncertainty estimates (Hefley and Hooten 2016, Nieto-Lugilde et al. 2018, Pollock et al. 2020). We are looking forward to applying Bayesian hierarchical models in future CALeDNA meta-analyses.

Finally, we suggest eDNA ecological network analyses should be leveraged so that the biotic interaction dependence can be contrasted with dependence or sensitivity to the abiotic environment. Our work shows a weak but positive relationship between the number of degrees a family has and its propensity for environmental filtering based on gradient forest predictability. This positive relationship persists across phylogenetic groups (Figure 1.5). Other studies focused on a single kingdom have obtained similar conclusions, such as in microbial variation in an altitudinal gradient in the Atacama Desert, Chile (Mandakovic et al. 2018).

Conclusion

In conclusion, we demonstrate the emerging potential of coupling CCS observations and eDNA data from samples that CCS volunteers collect in combination with remote sensing and ecological modeling to assess community-environment interactions and ultimately map community turnover. We provide one of the most comprehensive surveys of terrestrial biodiversity across three domains of life over a large, environmentally diverse state. We show the predictive and explanatory power of environmental variables on alpha, beta, and zeta diversity across highly diverse regions and at local geographic scales. The beta diversity map for California, as a continuous surface of community turnover, shares many similar boundaries to the standard US Ecoregion maps, but with nuanced detail. Computationally intensive and artificial intelligence driven models are producing maps for mitigating the challenges of global change (Harfouche et al. 2019, Pollock et al. 2020). Our approach contributes to the development of strategies to model living systems which could be directly used as Essential Biodiversity Variables for tracking biodiversity change, advancing ecological understanding, and managing ecosystems.

Tables

Table 1.1

Table 1.1. List of the categorical and a reduced set of numerical variables used in the diversity analysis and gradient forest modeling. For a complete list of variables, detailed description and data URL, refer to Data S1.1.

Variable	Category	Description and definition
Categorical variables		
loc	Location	Name of places visited reported by volunteers
clust	Location	Neighboring cluster of sites within a radius of 0.5 km derived from GPS record
ecoregion	Habitat	EPA Level III Ecoregions of California (Conterminous United States)
majorhab	Habitat	Major habitat type classified according to California Wildlife Habitat Relationships System
minorhab	Habitat	Minor habitat type classified according to California Wildlife Habitat Relationships System
transect	Habitat	Original classification of the predominant biome type (coast/coastal, shrub/ShrubScrub, and forest)
NLCD	Habitat	USGS national land cover classification 2011
SoS	Soil Properties	Volunteers' classification of substrate type (Sediment, Soil, Sand)
taxousda	Soil Properties	Predicted most probable class in USDA soil taxonomy
Reduced set of numerical variables		
Longitude	Location	Longitude of sample sites
hfp	Habitat	Global human footprint index
bio1	BioClim	Annual Mean Temperature
bio2	BioClim	Mean Diurnal Range (Mean of monthly (max temp - min temp))
bio3	BioClim	Isothermality (BIO2/BIO7) (* 100)
bio4	BioClim	Temperature Seasonality (standard deviation *100)
bio5	BioClim	Max Temperature of Warmest Month
bio6	BioClim	Min Temperature of Coldest Month
bio8	BioClim	Mean Temperature of Wettest Quarter
bio14	BioClim	Precipitation of Driest Month
bio15	BioClim	Precipitation Seasonality (Coefficient of Variation)
phihox	Soil Properties	Soil pH x 10 in H_2O at depth 0.00 m
orcdrc	Soil Properties	Soil organic carbon content (fine earth fraction) in g / kg at depth 0.00 m
cecsol	Soil Properties	Cation exchange capacity of soil in cmolc / kg at depth 0.00 m

sndppt	Soil Properties	Sand content (50 to 2000 μm) mass fraction in % at depth 0.00 m
bldfie	Soil Properties	Bulk density (fine earth) in kg / m^3 at depth 0.00 m
ntot	Soil Properties	Weight percentage of total nitrogen at depth 0.00 m
elev	Topography	Elevation of sample sites
Slope	Topography	The rate of change of elevation for each digital elevation model (DEM) cell
aspect	Topography	The direction of the maximum rate of change in the z-value from each cell in a raster surface
CTI	Topography	Compound Topographic Index
DAH	Topography	Diurnal Anisotropic Heating
B1	Vegetation	Sentinel-2 spectral band 1 (Wavelength: 443.9nm (S2A) / 442.3nm (S2B); Description: Aerosols)
B4	Vegetation	Sentinel-2 spectral band 4 (Wavelength: 664.5nm (S2A) / 665nm (S2B); Description: Red)
B6	Vegetation	Sentinel-2 spectral band 6 (Wavelength: 740.2nm (S2A) / 739.1nm (S2B); Description: Red Edge 2)
B9	Vegetation	Sentinel-2 spectral band 9 (Wavelength: 945nm (S2A) / 943.2nm (S2B); Description: Water vapor)
B10	Vegetation	Sentinel-2 spectral band 10 (Wavelength: 1373.5nm (S2A) / 1376.9nm (S2B); Description: Cirrus)
B11	Vegetation	Sentinel-2 spectral band 11 (Wavelength: 1613.7nm (S2A) / 1610.4nm (S2B); Description: SWIR 1)
NDVI32	Vegetation	Normalized Difference Vegetation Index in 32 days period
NBRT	Vegetation	Normalized Burn Ratio Thermal index in 32 days period
greenness	Vegetation	Annual Greenest Pixel in the year of 2017
imprv	Habitat	Percent of the pixel covered by developed impervious surface
ptrcv	Habitat	Percent of the pixel covered by tree canopy

Table 1.2

Table 1.2. *Post hoc* fitting of environmental variables on PCoA ordination (*Envfit*) for each metabarcode. Here, we present the three significant ($P < 0.001$) environmental variables with the highest correlation coefficient. The significance of the correlation was tested by 1999 permutations. For a complete result of all variables, please refer to Data S1.13. The direction of changes is included in Figure S1.23.

Metabarcode	**1st variable**	**R^2**	**2nd variable**	**R^2**	**3rd variable**	**R^2**
16S	NDVI32	0.49	greenness	0.47	B1	0.42
18S	NDVI32	0.51	greenness	0.49	B1	0.43
CO1	orcdrc	0.41	ptrcv	0.36	NBRT	0.33
FITS	greenness	0.52	B1	0.5	orcdrc	0.46
PITS	bio3	0.21	sndppt	0.2	B11	0.13

Table 1.3

Table 1.3. Variation in ζ_4 (zeta) diversity attributed to geographic separation distance between site clusters (*VarDistance*) versus variation in an environmental factor group between the same site clusters (*VarFactor*). Within each metabarcode, factor groups were ordered from lowest to highest contributions to variations in zeta diversity. Communities were defined at family levels.

Metabarcode	FactorGroup	NumSamples	VarFactor	VarDistance	VarUnknown
16S	Location	184	0.00%	0.29%	99.70%
16S	Topography	184	0.94%	0.17%	98.90%
16S	Habitat	156	1.33%	0.00%	98.70%
16S	Vegetation	169	5.92%	0.00%	94.10%
16S	BioClim	184	7.17%	0.00%	92.20%
16S	Soil Properties	180	9.21%	0.00%	90.70%
18S	Location	184	0.14%	0.00%	99.90%
18S	Habitat	156	5.49%	0.00%	94.50%
18S	Topography	184	7.15%	0.00%	92.80%
18S	BioClim	184	7.30%	0.00%	92.70%
18S	Soil Properties	180	15.30%	0.00%	84.70%
18S	Vegetation	169	18.50%	0.00%	81.50%
CO1	Location	184	0.11%	0.22%	99.60%
CO1	Habitat	156	1.86%	0.00%	98.10%
CO1	Topography	184	3.30%	0.46%	96.20%
CO1	BioClim	184	12.00%	0.00%	88.00%
CO1	Vegetation	169	18.20%	0.31%	81.10%
CO1	Soil Properties	180	18.60%	0.00%	81.30%
FITS	Topography	184	0.69%	0.55%	98.70%
FITS	Location	184	0.93%	0.38%	98.20%
FITS	Habitat	156	2.24%	0.37%	97.10%
FITS	BioClim	184	18.50%	0.00%	80.40%
FITS	Soil Properties	180	22.40%	0.00%	77.50%
FITS	Vegetation	169	32.40%	1.05%	66.40%
PITS	Location	184	0.03%	0.00%	100.00%
PITS	BioClim	184	1.30%	0.00%	98.70%
PITS	Habitat	156	2.16%	0.00%	97.80%
PITS	Topography	184	2.98%	0.03%	96.90%
PITS	Soil Properties	180	4.23%	0.00%	95.70%
PITS	Vegetation	169	9.00%	0.00%	91.00%

Figures

Figure 1.1

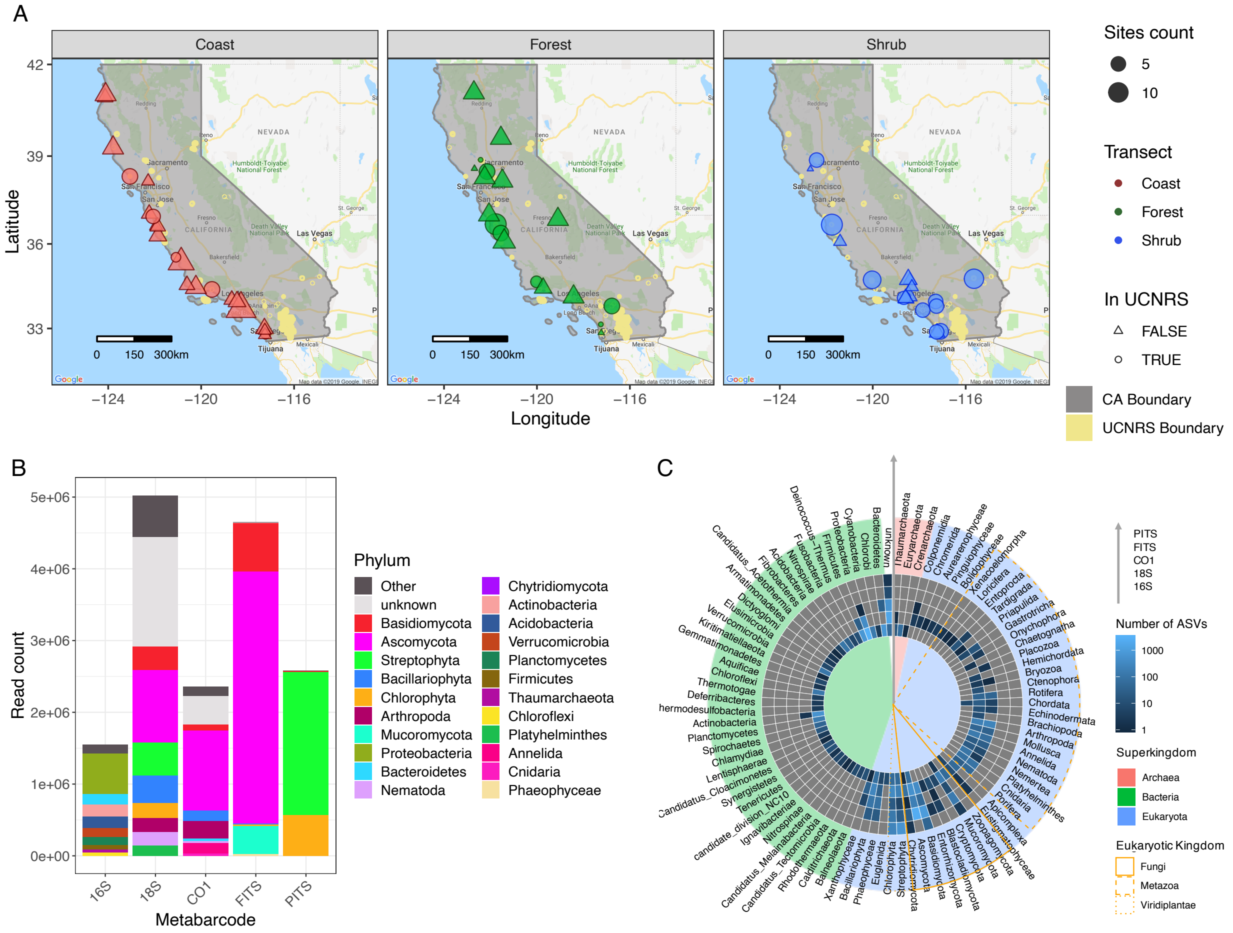

Figure 1.1. Map of 278 sites included in this study and illustration of taxonomic entries recovered with five metabarcodes. **(A)** Study area (gray shade) is defined within the State of California, United States. Sample sites are colored by three transect designations: coast (red), forest (green) and shrub (blue). Size of the points corresponds to the number of samples taken in the same area. Shape of the points represents areas within (circles) and outside (triangles) of the University of California's Natural Reserve System (UCNRS, yellow shade, area size not to scale for visibility). **(B)** Read abundance is grouped by the phylum they belong to after taxonomy assignment and decontamination for five metabarcodes targeting Bacteria and Archaea (*16S*),

Eukaryota (*18S*), Metazoa (*CO1*), Fungi (*FITS*) and Viridiplantae (*PITS*). Only the most abundant 10 phyla are plotted for each metabarcode. All other phyla are summarized in the “Other” category. **(C)** Heatmap shows each metabarcode’s taxonomic specificity. The results from each metabarcode (*16S, 18S, CO1, FITS, PITS*) are represented from inner to outer rings (gray arrow). Lighter blue in one cell represents more taxonomic entries were recovered by that metabarcode for that phylum, gray color represents no entries. Phyla are indicated on the periphery. Background color of each pie wedge denotes the superkingdom (Red: Archaea, Blue: Eukaryota, Green: Bacteria, No background: Unknown) to which the phyla belonged at the time of taxonomy assignment (taxonomy file downloaded from NCBI on January 19, 2018). For eukaryotic phyla, kingdoms are marked by different line types in an orange outline: Fungi (solid), Metazoa (dashed) and Viridiplantae (dotted).

Figure 1.2

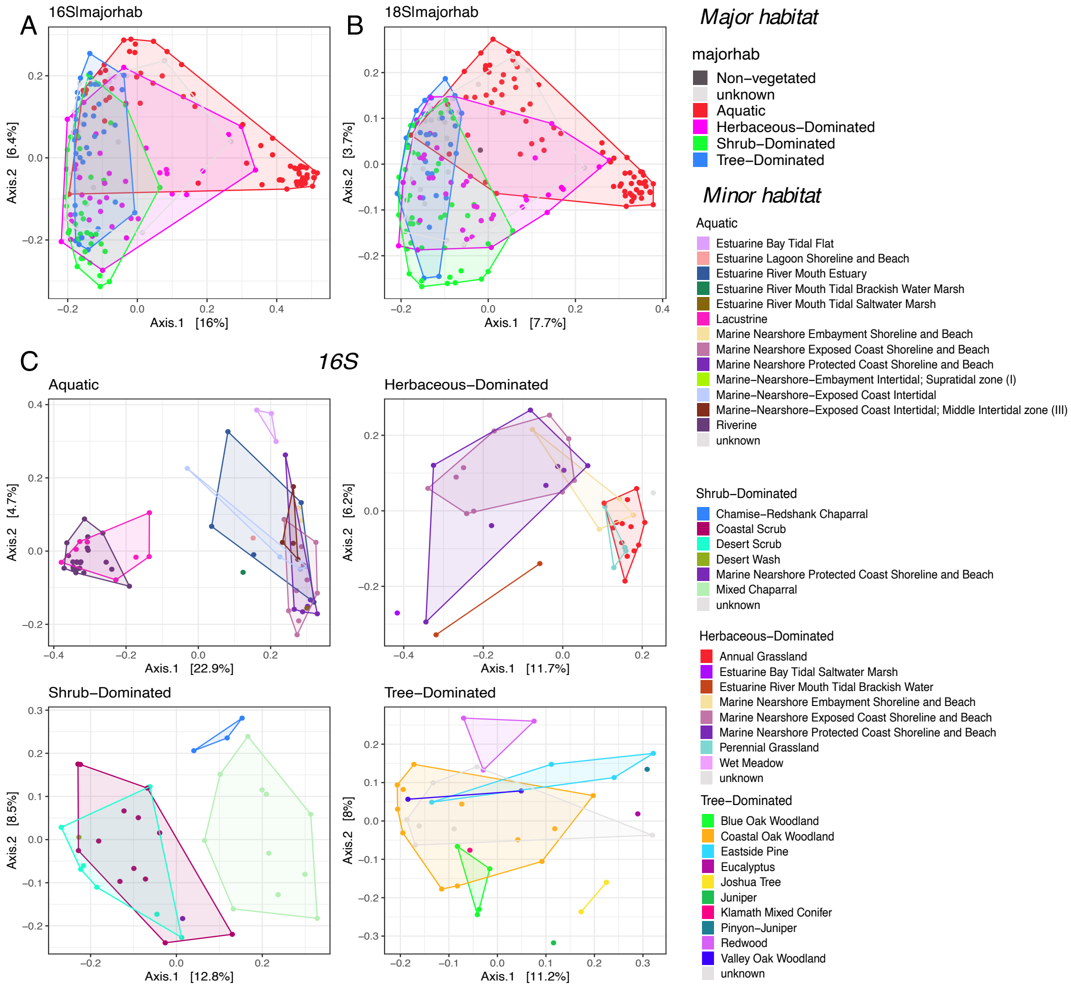

Figure 1.2. Beta diversity plots based on Jaccard dissimilarity. The first two principal coordinates are plotted with percentage of variance explained included in axis label. We show selected Principal Coordinate Analysis (PCoA) plots from **(A)** *16S* and **(B)** *18S* for major habitat. Each point stands for a sample site. **(C)** Example PCoA plots based on Jaccard dissimilarity with samples grouped by minor habitat and plotted within *aquatic* major habitat for *16S* metabarcode. Some minor habitat groups separate while others overlap, and patterns of compositional similarity (overlap) are different for different metabarcodes (Figure S1.20).

Figure 1.3

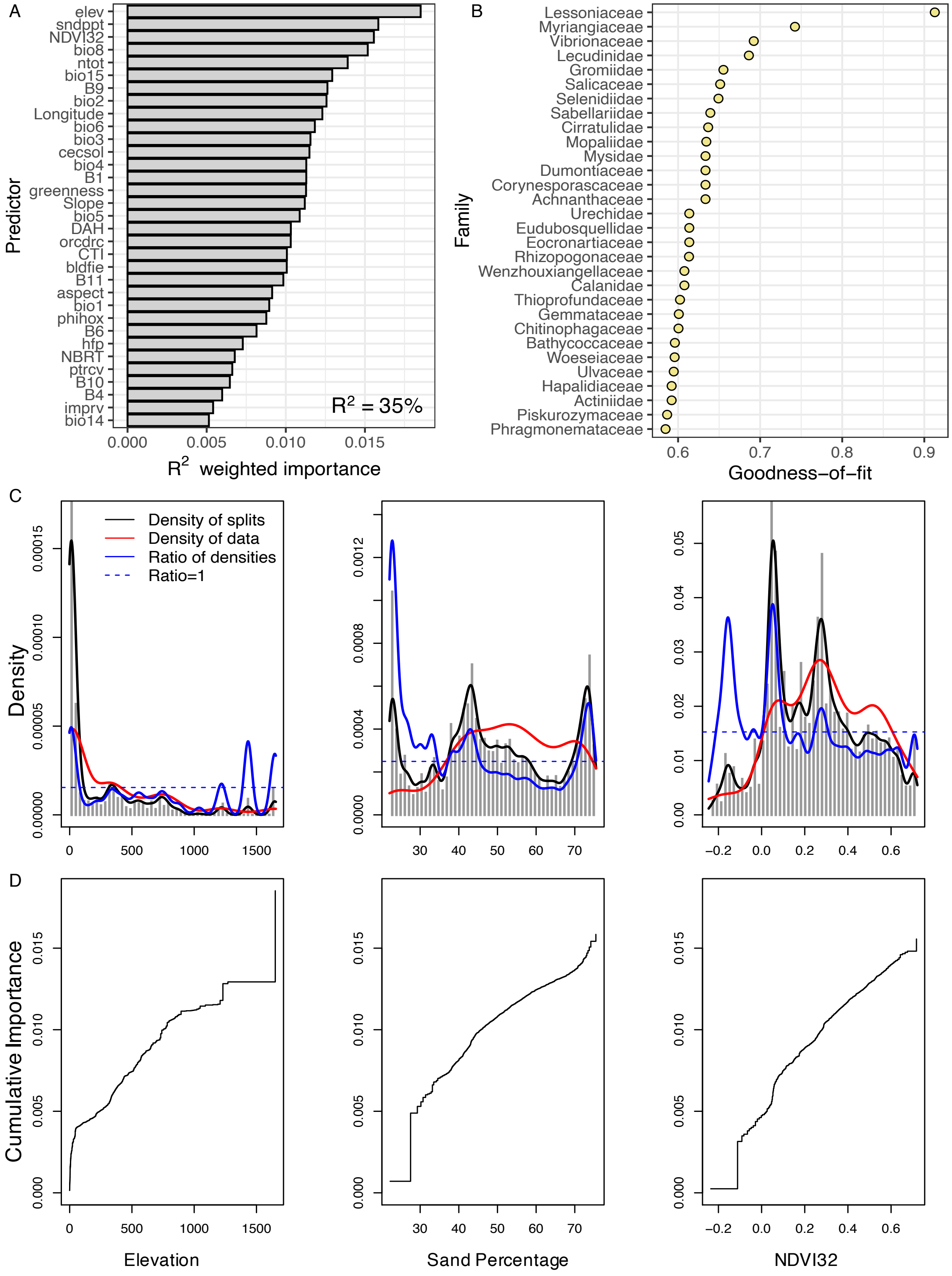
A
Predictor
elev
sndppt
NDVI32
bio8
ntot
bio15
B9
bio2
Longitude
bio6
bio3
cecsol
bio4
B1
greenness
Slope
bio5
DAH
orcdrc
CTI
bldfie
B11
aspect
bio1
phihox
B6
hfp
NBRT
ptrcv
B10
B4
imprv
bio14
$R^2 = 35\%$
0.000
0.005
0.010
0.015
R^2 weighted importance
B
Family
Lessoniaceae
Myriangiaceae
Vibrionaceae
Lecudinidae
Gromiidae
Salicaceae
Selenidiidae
Sabellariidae
Cirratulidae
Mopaliidae
Mysidae
Dumontiaceae
Corynesporascaceae
Achnanthaceae
Urechidae
Eudubosquellidae
Eocronartiaceae
Rhizopogonaceae
Wenzhouxiangellaceae
Calanidae
Thioprofundaceae
Gemmataceae
Chitinophagaceae
Bathycoccaceae
Woeseiaceae
Ulvaceae
Hapalidiaceae
Actiniidae
Piskurozymaceae
Phragmonemataceae
0.6
0.7
0.8
0.9
Goodness−of−fit
C
Density
Density of splits
Density of data
Ratio of densities
Ratio=1
0.00000
0.00005
0.00010
0.00015
0
500
1000
1500
0.0000
0.0004
0.0008
0.0012
30
40
50
60
70
0.00
0.01
0.02
0.03
0.04
0.05
−0.2
0.0
0.2
0.4
0.6
D
Cumulative Importance
0.000
0.005
0.010
0.015
0
500
1000
1500
Elevation
0.000
0.005
0.010
0.015
30
40
50
60
70
Sand Percentage
0.000
0.005
0.010
0.015
−0.2
0.0
0.2
0.4
0.6
NDVI32

Figure 1.3. Gradient forest result for filtered CALeDNA dataset. **(A)** Ranked overall importance for 33 environmental predictors. **(B)** Ranked goodness-of-fit (1 - Relative Error rates) for the top 30 families (response variables). **(C)** and **(D)** show the community turnover along the three most important environmental gradients: elevation, sand percentage and photosynthetic activity proxy (*NDVI32*). **(C)** The gray histogram shows binned split importance at each gradient. Kernel density of splits (black lines), of observed predictor values (red lines) and of splits standardized by observation density (blue lines) are overlaid. The horizontal dashed line indicates where the ratio is 1. Each curve integrates to the importance of the predictor. **(D)** The line shows cumulative importance distributions of splits improvement scaled by R^2 weighted importance and standardized by density of observations, averaged over all families.

Figure 1.4

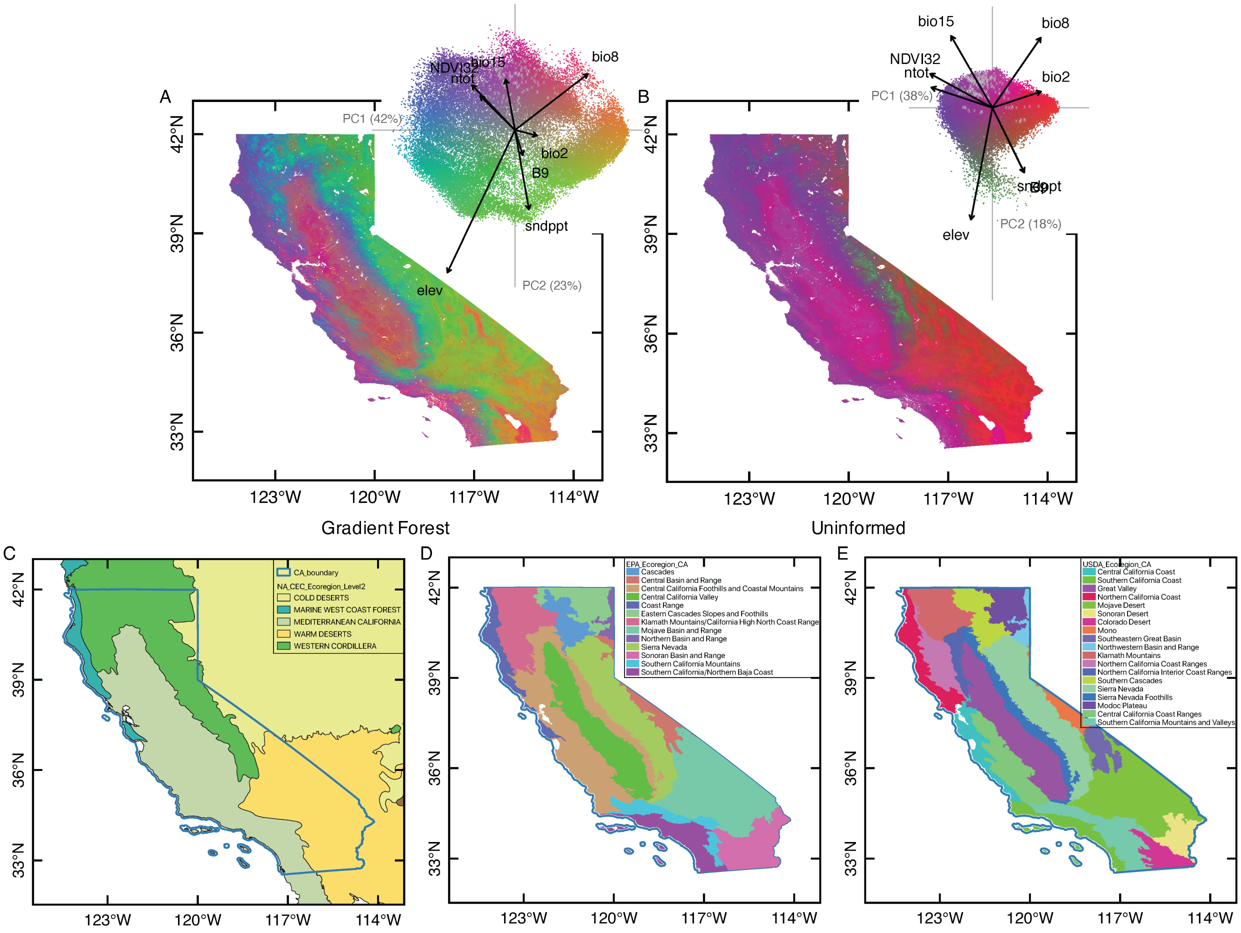

Figure 1.4. Gradient forest predicted community turnover map in California. **(A)** Map of transformed environmental variables following gradient forest predictions of biodiversity turnover from eDNA results compared with **(B)** uninformed, standardized environmental variables and **(C-E)** current major ecoregion maps in California. The map shows the first three principal dimensions of **(A)** biologically predicted or **(B)** uninformed community compositions with an RGB color palette with 100 m resolution. The biplot of the first two PCs of the transformed environment space with (inset A) or without (inset B) biological information provides a color key for the compositional variation (n = 50,000). Similar colors approximate similar community in the transformed environmental space. The gray crosses denote the input

eDNA sites (n = 272). Vectors denote the direction and magnitude of the eight most important environmental correlates. **(C-E)** Selected major ecoregions maps are provided for comparisons with **(A)** the gradient forest map. **(C)** EPA Level II Ecoregions of North America (U.S. Environmental Protection Agency 2010). **(D)** EPA Level III Ecoregions of California (U.S. Environmental Protection Agency 2012). **(E)** USDA Ecoregion Sections in California (USDA Forest Service 2007).

Figure 1.5

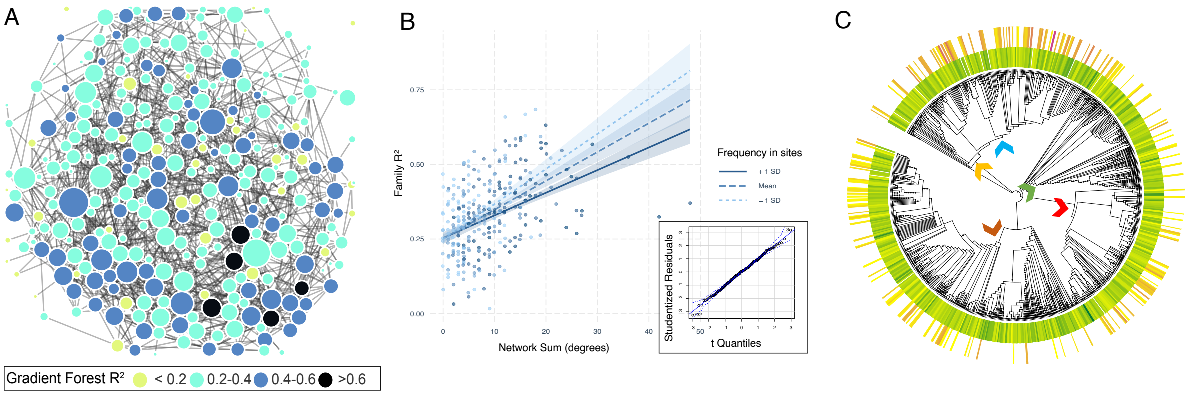

Figure 1.5. eDNA-based ecological co-occurrence network and relationship with gradient forest model goodness-of-fit R^2. **(A)** 369 families (as nodes) are included in the network and 290 of those have at least one edge connecting them to another node. Dark blue and black nodes represent families with R^2 predictor values >0.4. The size of the node is scaled to the number of network degrees. **(B)** OLS linear regression and quantile-quantile plot showing the interaction between network sum of degrees and frequency of taxa in sample sites with the dependent variable of gradient forest Family goodness-of-fit R^2. There were 304 families included as joint observations in gradient forest and network results. The adjusted $R^2 = 0.22$, network sum estimate = 0.01 (t-value=5.44; p = 0.00), frequency in sites estimate = 0.00 (t-value = 0.18; p = 0.86), and interaction between network sum and frequency in sites = 0.00 (t-value = -2.38; p = 0.02). **(C)** Phylogenetic tree made with the Open Tree of Life targeting input families as tips. Heatmap labels correspond to the range of gradient forest R^2 (0.078-0.913) from yellow to dark green (inner circle), and to the range of network degrees (0-48) from yellow to purple (outer circle). Families too rare to be included in the network analysis (in fewer than 28 sites) are not colored in heatmaps. Arrows indicate the following clades: brown = fungi, mustard = Enterobacteriaceae, blue = Flavobacteriia, green = Streptophyta, red = SAR supergroup.

References

Amaral-Zettler, L. A., E. A. McCliment, H. W. Ducklow, and S. M. Huse. 2009. A Method for Studying Protistan Diversity Using Massively Parallel Sequencing of V9 Hypervariable Regions of Small-Subunit Ribosomal RNA Genes. PLOS ONE 4:e6372.

Bahram, M., F. Hildebrand, S. K. Forslund, J. L. Anderson, N. A. Soudzilovskaia, et al. 2018. Structure and function of the global topsoil microbiome. Nature 560:233–237.

Baldwin, B. G., A. H. Thornhill, W. A. Freyman, D. D. Ackerly, M. M. Kling, et al. 2017. Species richness and endemism in the native flora of California. American Journal of Botany 104:487–501.

Bay, R. A., R. J. Harrigan, V. L. Underwood, H. L. Gibbs, T. B. Smith, and K. Ruegg. 2018. Genomic signals of selection predict climate-driven population declines in a migratory bird. Science 359:83–86.

Biggs, J., N. Ewald, A. Valentini, C. Gaboriaud, T. Dejean, et al. 2015. Using eDNA to develop a national citizen science-based monitoring programme for the great crested newt (Triturus cristatus). Biological Conservation 183:19–28.

Bohmann, K., A. Evans, M. T. P. Gilbert, G. R. Carvalho, S. Creer, et al. 2014. Environmental DNA for wildlife biology and biodiversity monitoring. Trends in Ecology & Evolution 29:358–367.

Breiman, L. 2001. Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). Statistical Science 16:199–231.

Bush, A., R. Sollmann, A. Wilting, K. Bohmann, B. Cole, et al. 2017. Connecting Earth observation to high-throughput biodiversity data. Nature Ecology & Evolution 1:0176.

Calderón-Sanou, I., T. Münkemüller, F. Boyer, L. Zinger, and W. Thuiller. 2020. From environmental DNA sequences to ecological conclusions: How strong is the influence of methodological choices? Journal of Biogeography 47:193–206.

Callahan, B. J., P. J. McMurdie, M. J. Rosen, A. W. Han, A. J. A. Johnson, and S. P. Holmes. 2016. DADA2: High-resolution sample inference from Illumina amplicon data. Nature Methods 13:581–583.

Caporaso, J. G., C. L. Lauber, W. A. Walters, D. Berg-Lyons, J. Huntley, et al. 2012. Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. The ISME Journal 6:1621–1624.

Castro, L. R., A. M. Lagos, B. Shapiro, R. S. Meyer, S. Shirazi, et al. in review. Metabarcoding meiofauna biodiversity assessment in four beaches of Northern Colombia: Effects of sampling protocols and primer choice. in review.

Ceballos, G., P. R. Ehrlich, A. D. Barnosky, A. García, R. M. Pringle, and T. M. Palmer. 2015. Accelerated modern human–induced species losses: Entering the sixth mass extinction. Science advances 1:e1400253.

Cheng, J., B. Karambelkar, and Y. Xie. 2018. leaflet: Create Interactive Web Maps with the JavaScript “Leaflet” Library. https://CRAN.R-project.org/package=leaflet

Cornelius, J. M., P. R. Kemp, J. A. Ludwig, and G. L. Cunningham. 1991. The distribution of vascular plant species and guilds in space and time along a desert gradient. Journal of Vegetation Science 2:59–72.

Cristescu, M. E. 2014. From barcoding single individuals to metabarcoding biological communities: towards an integrative approach to the study of global biodiversity. Trends in Ecology & Evolution 29:566–571.

Crowther, T. W., J. van den Hoogen, J. Wan, M. A. Mayes, A. D. Keiser, et al. 2019. The global soil community and its influence on biogeochemistry. Science 365:eaav0550.

Curd, E. E., Z. Gold, G. S. Kandlikar, J. Gomer, M. Ogden, et al. 2019. *Anacapa Toolkit* : an environmental DNA toolkit for processing multilocus metabarcode datasets. Methods in Ecology and Evolution:2041–210X.13214.

Deiner, K., E. A. Fronhofer, E. Mächler, J.-C. Walser, and F. Altermatt. 2016. Environmental DNA reveals that rivers are conveyer belts of biodiversity information. Nature Communications 7:12544.

Delgado-Baquerizo, M., F. T. Maestre, P. B. Reich, T. C. Jeffries, J. J. Gaitan, et al. 2016. Microbial diversity drives multifunctionality in terrestrial ecosystems. Nature Communications 7:1–8.

Díaz, S., J. Settele, E. S. Brondízio, H. T. Ngo, J. Agard, et al. 2019. Pervasive human-driven decline of life on Earth points to the need for transformative change. Science 366:eaax3100.

Durán, S. M., R. E. Martin, S. Díaz, B. S. Maitner, Y. Malhi, et al. 2019. Informing trait-based ecology by assessing remotely sensed functional diversity across a broad tropical temperature gradient. Science Advances 5:eaaw8114.

Ehrlich, R., S. Schulz, M. Schloter, and Y. Steinberger. 2015. Effect of slope orientation on microbial community composition in different particle size fractions from soils obtained from desert ecosystems. Biology and Fertility of Soils 51:507–510.

Ellis, N., S. J. Smith, and C. R. Pitcher. 2012. Gradient forests: calculating importance gradients on physical predictors. Ecology 93:156–168.

Epp, L. S., S. Boessenkool, E. P. Bellemain, J. Haile, A. Esposito, et al. 2012. New environmental metabarcodes for analysing soil DNA: potential for studying past and present ecosystems. Molecular Ecology 21:1821–1833.

Fick, S. E., and R. J. Hijmans. 2017. WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. International Journal of Climatology 37:4302–4315.

Fierer, N., and R. B. Jackson. 2006. The diversity and biogeography of soil bacterial communities. Proceedings of the National Academy of Sciences 103:626–631.

Franklin, T. W., K. S. McKelvey, J. D. Golding, D. H. Mason, J. C. Dysthe, et al. 2019. Using environmental DNA methods to improve winter surveys for rare carnivores: DNA from snow and improved noninvasive techniques. Biological Conservation 229:50–58.

Gao, X., H. Lin, K. Revanna, and Q. Dong. 2017. A Bayesian taxonomic classification method for 16S rRNA gene sequences with improved species-level accuracy. BMC Bioinformatics 18:247.

George, P. B. L., D. Lallias, S. Creer, F. M. Seaton, J. G. Kenny, et al. 2019. Divergent national-scale trends of microbial and animal biodiversity revealed across diverse temperate soil ecosystems. Nature Communications 10:1107.

Gibson, N., R. Meissner, A. S. Markey, and W. A. Thompson. 2012. Patterns of plant diversity in ironstone ranges in arid south western Australia. Journal of Arid Environments 77:25–31.

Goldberg, C. S., C. R. Turner, K. Deiner, K. E. Klymus, P. F. Thomsen, et al. 2016. Critical considerations for the application of environmental DNA methods to detect aquatic species. Methods in Ecology and Evolution:1299–1307.

Gordon, A., G. Hannon, and others. 2010. Fastx-toolkit. FASTQ/A short-reads preprocessing tools (unpublished) http://hannonlab. cshl. edu/fastx_toolkit 5.

Groenendyk, D. G., T. P. A. Ferré, K. R. Thorp, and A. K. Rice. 2015. Hydrologic-Process-Based Soil Texture Classifications for Improved Visualization of Landscape Function. PLOS ONE 10:e0131299.

Gu, W., J. Song, Y. Cao, Q. Sun, H. Yao, et al. 2013. Application of the ITS2 Region for Barcoding Medicinal Plants of Selaginellaceae in Pteridophyta. PLOS ONE 8:e67818.

Harfouche, A. L., D. A. Jacobson, D. Kainer, J. C. Romero, A. H. Harfouche, et al. 2019. Accelerating Climate Resilient Plant Breeding by Applying Next-Generation Artificial Intelligence. Trends in Biotechnology 37:1217–1235.

Hefley, T. J., and M. B. Hooten. 2016. Hierarchical Species Distribution Models. Current Landscape Ecology Reports 1:87–97.

Hengl, T., J. M. de Jesus, G. B. M. Heuvelink, M. R. Gonzalez, M. Kilibarda, et al. 2017. SoilGrids250m: Global gridded soil information based on machine learning. PLOS ONE 12:e0169748.

Hijmans, R. J. 2019. raster: Geographic Data Analysis and Modeling. https://CRAN.R-project.org/package=raster

Hobern, D. G. 2020. BIOSCAN: DNA Barcoding to accelerate taxonomy and biogeography for conservation and sustainability. Genome.

van den Hoogen, J., S. Geisen, D. Routh, H. Ferris, W. Traunspurger, et al. 2019. Soil nematode abundance and functional group composition at a global scale. Nature 572:194–198.

Horner-Devine, M. C., J. M. Silver, M. A. Leibold, B. J. M. Bohannan, R. K. Colwell, et al. 2007. A Comparison of Taxon Co-Occurrence Patterns for Macro- and Microorganisms. Ecology 88:1345–1353.

Hui, C., M. A. McGeoch, A. E. S. Harrison, and E. J. L. Bronstein. 2014. Zeta Diversity as a Concept and Metric That Unifies Incidence-Based Biodiversity Patterns. The American Naturalist 184:684–694.

Hui, C., W. Vermeulen, and G. Durrheim. 2018. Quantifying multiple-site compositional turnover in an Afrotemperate forest, using zeta diversity. Forest Ecosystems 5:15.

Jackson, L. E., T. M. Bowles, H. Ferris, A. J. Margenot, A. Hollander, et al. 2019. Plant and soil microfaunal biodiversity across the borders between arable and forest ecosystems in a Mediterranean landscape. Applied Soil Ecology 136:122–138.

Jenkins, C. N., S. L. Pimm, and L. N. Joppa. 2013. Global patterns of terrestrial vertebrate diversity and conservation. Proceedings of the National Academy of Sciences 110:E2602–E2610.

Jenkins, C. N., K. S. Van Houtan, S. L. Pimm, and J. O. Sexton. 2015. US protected lands mismatch biodiversity priorities. Proceedings of the National Academy of Sciences 112:5081–5086.

Jetz, W., M. A. McGeoch, R. Guralnick, S. Ferrier, J. Beck, et al. 2019. Essential biodiversity variables for mapping and monitoring species populations. Nature Ecology & Evolution 3:539–551.

Kandlikar, G. S., Z. J. Gold, M. C. Cowen, R. S. Meyer, A. C. Freise, et al. 2018. ranacapa: An R package and Shiny web app to explore environmental DNA data with exploratory statistics and interactive visualizations. F1000Research 7:1734.

Karimi, B., S. Terrat, S. Dequiedt, N. P. A. Saby, W. Horrigue, et al. 2018. Biogeography of soil bacteria and archaea across France. Science Advances 4:eaat1808.

Kobori, H., J. L. Dickinson, I. Washitani, R. Sakurai, T. Amano, et al. 2016. Citizen science: a new approach to advance ecology, education, and conservation. Ecological Research 31:1–19.

Kurtz, Z. D., C. L. Müller, E. R. Miraldi, D. R. Littman, M. J. Blaser, and R. A. Bonneau. 2015. Sparse and Compositionally Robust Inference of Microbial Ecological Networks. PLOS Computational Biology 11:e1004226.

Langmead, B., and S. L. Salzberg. 2012. Fast gapped-read alignment with Bowtie 2. Nature Methods 9:357–359.

Larson, E. R., B. M. Graham, R. Achury, J. J. Coon, M. K. Daniels, et al. 2020. From eDNA to citizen science: emerging tools for the early detection of invasive species. Frontiers in Ecology and the Environment 18:194–202.

Latombe, G., M. A. McGeoch, D. A. Nipperess, and C. Hui. 2018. zetadiv: functions to compute compositional turnover using ζ diversity. https://cran.r-project.org/package=zetadiv

Lee, J.-S., J.-W. Kim, S. H. Lee, H.-H. Myeong, J.-Y. Lee, and J. S. Cho. 2016. Zonation and soil factors of salt marsh halophyte communities. Journal of Ecology and Environment 40:4.

Leray, M., J. Y. Yang, C. P. Meyer, S. C. Mills, N. Agudelo, et al. 2013. A new versatile primer set targeting a short fragment of the mitochondrial COI region for metabarcoding metazoan diversity: application for characterizing coral reef fish gut contents. Frontiers in Zoology 10:34.

Lewin, H. A., G. E. Robinson, W. J. Kress, W. J. Baker, J. Coddington, et al. 2018. Earth BioGenome Project: Sequencing life for the future of life. Proceedings of the National Academy of Sciences:201720115.

Long, J. A. 2019. interactions: Comprehensive, User-Friendly Toolkit for Probing Interactions. https://cran.r-project.org/package=interactions

Mandakovic, D., C. Rojas, J. Maldonado, M. Latorre, D. Travisany, et al. 2018. Structure and co-occurrence patterns in microbial communities under acute environmental stress reveal ecological factors fostering resilience. Scientific Reports 8:1–12.

Martin, M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet. journal 17:10–12.

McMurdie, P. J., and S. Holmes. 2013. phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. PLOS ONE 8:e61217.

Meyer, R. S., M. Munguia Ramos, M. Lin, T.M. Schweizer, Z. Gold, et al. 2021. The CALeDNA program: Citizen scientists and researchers inventory California's biodiversity. California Agriculture 75 (1): 20–32.

Miralles, L., E. Dopico, F. Devlo-Delva, and E. Garcia-Vazquez. 2016. Controlling populations of invasive pygmy mussel (Xenostrobus securis) through citizen science and environmental DNA. Marine Pollution Bulletin 110:127–132.

Montagna, M., A. Berruti, V. Bianciotto, P. Cremonesi, R. Giannico, et al. 2018. Differential biodiversity responses between kingdoms (plants, fungi, bacteria and metazoa) along an Alpine succession gradient. Molecular Ecology 27:3671–3685.

Moreno, J., A. Terrones, A. Juan, and M. Á. Alonso. 2018. Halophytic plant community patterns in Mediterranean saltmarshes: shedding light on the connection between abiotic factors and the distribution of halophytes. Plant and Soil 430:185–204.

Myers, N., R. A. Mittermeier, C. G. Mittermeier, G. A. B. da Fonseca, and J. Kent. 2000. Biodiversity hotspots for conservation priorities. Nature 403:853–858.

Nieto-Lugilde, D., K. C. Maguire, J. L. Blois, J. W. Williams, and M. C. Fitzpatrick. 2018. Multiresponse algorithms for community-level modelling: Review of theory, applications, and comparison to species distribution models. Methods in Ecology and Evolution 9:834–848.

Oksanen, J., F. G. Blanchet, M. Friendly, R. Kindt, P. Legendre, et al. 2018. vegan: Community Ecology Package. https://CRAN.R-project.org/package=vegan

Omernik, J. M., and G. E. Griffith. 2014. Ecoregions of the Conterminous United States: Evolution of a Hierarchical Spatial Framework. Environmental Management 54:1249–1266.

Pereira, H. M., S. Ferrier, M. Walters, G. N. Geller, R. H. G. Jongman, et al. 2013. Essential Biodiversity Variables. Science 339:277–278.

Peters, M. K., A. Hemp, T. Appelhans, J. N. Becker, C. Behler, et al. 2019. Climate–land-use interactions shape tropical mountain biodiversity and ecosystem functions. Nature 568:88–92.

Pettorelli, N., K. Safi, and W. Turner. 2014. Satellite remote sensing, biodiversity research and conservation of the future. Phil. Trans. R. Soc. B 369:20130190.

Pimm, S. L., C. N. Jenkins, R. Abell, T. M. Brooks, J. L. Gittleman, et al. 2014. The biodiversity of species and their rates of extinction, distribution, and protection. Science 344:1246752.

Pitcher, C. R., N. Ellis, and S. J. Smith. 2011. Example analysis of biodiversity survey data with R package gradientForest:16. http://gradientforest.r-forge.r-project.org/biodiversity-survey.pdf

Pollock, L. J., L. M. J. O'Connor, K. Mokany, D. F. Rosauer, M. V. Talluto, and W. Thuiller. 2020. Protecting Biodiversity (in All Its Complexity): New Models and Methods. Trends in Ecology & Evolution:S0169534720302305.

Prosser, J. I. 2010. Replicate or lie. Environmental Microbiology 12:1806–1810.

R Core Team. 2019. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Ranjard, L., S. Dequiedt, N. Chemidlin Prévost-Bouré, J. Thioulouse, N. P. A. Saby, et al. 2013. Turnover of soil bacterial diversity driven by wide-scale environmental heterogeneity. Nature Communications 4:1–10.

Schimel, D., F. D. Schneider, and JPL Carbon and Ecosystem Participants. 2019. Flux towers in the sky: global ecology from space. New Phytologist 224:570–584.

Schneider, F. D., A. A. Ferraz, S. Hancock, L. I. Duncanson, R. O. Dubayah, et al. 2020. Towards mapping the diversity of canopy structure from space with GEDI. Environmental Research Letters.

Schneider, F. D., F. Morsdorf, B. Schmid, O. L. Petchey, A. Hueni, et al. 2017. Mapping functional diversity from remotely sensed morphological and physiological forest traits. Nature Communications 8:1441.

Seeber, P. A., G. K. McEwen, U. Löber, D. W. Förster, M. L. East, et al. 2019. Terrestrial mammal surveillance using hybridization capture of environmental DNA from African waterholes. Molecular Ecology Resources 19:1486–1496.

Sessitsch, A., A. Weilharter, M. H. Gerzabek, H. Kirchmann, and E. Kandeler. 2001. Microbial Population Structures in Soil Particle Size Fractions of a Long-Term Fertilizer Field Experiment. Applied and Environmental Microbiology 67:4215–4224.

Shannon, P., A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, et al. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome research 13:2498–2504.

Shirazi, S., R. Meyer, and B. Shapiro. 2020. PCR replication in environmental DNA metabarcoding. preprint, Authorea.

Silva, A. C., and A. F. Souza. 2018. Aridity drives plant biogeographical sub regions in the Caatinga, the largest tropical dry forest and woodland block in South America. PLOS ONE 13:e0196130.

Simons, A. L., R. Mazor, E. D. Stein, and S. Nuzhdin. 2019. Using alpha, beta, and zeta diversity in describing the health of stream-based benthic macroinvertebrate communities. Ecological Applications 29:e01896.

Taberlet, P., A. Bonin, L. Zinger, and E. Coissac. 2018. Environmental DNA: For Biodiversity Research and Monitoring. Oxford University Press.

Theobald, E. J., A. K. Ettinger, H. K. Burgess, L. B. DeBey, N. R. Schmidt, et al. 2015. Global change and local solutions: Tapping the unrealized potential of citizen science for biodiversity research. Biological Conservation 181:236–244.

Thompson, L. R., J. G. Sanders, D. McDonald, A. Amir, J. Ladau, et al. 2017. A communal catalogue reveals Earth’s multiscale microbial diversity. Nature 551:457.

Timling, I., D. A. Walker, C. Nusbaum, N. J. Lennon, and D. L. Taylor. 2014. Rich and cold: diversity, distribution and drivers of fungal communities in patterned-ground ecosystems of the North American Arctic. Molecular Ecology 23:3258–3272.

Tipton, L., C. L. Müller, Z. D. Kurtz, L. Huang, E. Kleerup, et al. 2018. Fungi stabilize connectivity in the lung and skin microbial ecosystems. Microbiome 6:12.

U.S. Environmental Protection Agency. 2010. NA_CEC_Eco_Level2. U.S. EPA Office of Research and Development (ORD) - National Health and Environmental Effects Research Laboratory (NHEERL), Corvallis, OR. ftp://ftp.epa.gov/wed/ecoregions/cec_na/NA_CEC_Eco_Level2.zip

U.S. Environmental Protection Agency. 2012. Level III Ecoregions of California. U.S. EPA Office of Research and Development (ORD) - National Health and Environmental Effects Research Laboratory (NHEERL), Corvallis, OR. ftp://newftp.epa.gov/EPADataCommons/ORD/Ecoregions/ca/ca_eco_l3.zip

USDA Forest Service. 2007. USDA Ecoregion Sections, California. USDA Forest Service - Pacific Southwest Region - Remote Sensing Lab. https://databasin.org/datasets/81a3a809a2ae4c099f2e495c0b2ecc91

Wang, R., J. A. Gamon, J. Cavender-Bares, P. A. Townsend, and A. I. Zygielbaum. 2018. The spatial sensitivity of the spectral diversity–biodiversity relationship: an experimental test in a prairie grassland. Ecological Applications 28:541–556.

Wildlife Conservation Society and Center for International Earth Science Information Network, Columbia University. 2005. Last of the Wild Project, Version 2, 2005 (LWP-2): Global Human Footprint Dataset (Geographic). Palisades, NY: NASA Socioeconomic Data and Applications Center (SEDAC). https://doi.org/10.7927/H4M61H5F.

White, H. J., L. León-Sánchez, V. J. Burton, E. K. Cameron, T. Caruso, et al. 2020. Methods and approaches to advance soil macroecology. Global Ecology and Biogeography 29:1674–1690.

White, T. J., T. Bruns, S. Lee, J. Taylor, and others. 1990. Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. PCR protocols: a guide to methods and applications 18:315–322.

Yamasaki, E., F. Altermatt, J. Cavender-Bares, M. C. Schuman, D. Zuppinger-Dingley, et al. 2017. Genomics meets remote sensing in global change studies: monitoring and predicting phenology, evolution and biodiversity. Current Opinion in Environmental Sustainability 29:177–186.

Yang, T., J. M. Adams, Y. Shi, J. He, X. Jing, et al. 2017. Soil fungal diversity in natural grasslands of the Tibetan Plateau: associations with plant diversity and productivity. New Phytologist 215:756–765.

Yu, D. W., Y. Ji, B. C. Emerson, X. Wang, C. Ye, et al. 2012. Biodiversity soup: metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring: Biodiversity soup. Methods in Ecology and Evolution 3:613–623.

Zarnetske, P. L., Q. D. Read, S. Record, K. D. Gaddis, S. Pau, et al. 2019. Towards connecting biodiversity and geodiversity across scales with satellite remote sensing. Global Ecology and Biogeography 28:548–556.

Chapter 2: The Genomic Footprint of Whaling and Isolation in Fin Whale Populations

In preparation for submission to *Nature Communications*

Supplementary materials available online as this dissertation's Supplementary Materials: Chapter2_Supplementary_Information.pdf

Abstract

Twentieth century industrial whaling pushed several species to the brink of extinction. Fin whales (*Balaenoptera physalus*) were the most impacted with over 75,000 individuals harvested in the North Pacific. However, a small, resident population in the Gulf of California was not targeted by whaling. Here we analyzed 50 whole-genome sequences from the Eastern North Pacific and Gulf of California fin whale populations to investigate their demographic history and the genomic effects of natural and human-induced bottlenecks. We show that the two populations diverged between 16,000 and 25,000 years ago, after which the Eastern North Pacific population expanded and then suffered a dramatic 99% reduction in effective population size during the recent whaling period. In contrast, the Gulf of California population remained small and highly isolated throughout this period, receiving less than one migrant per generation. However, this extremely low level of migration has been crucial for maintaining its viability, despite an increased genetic load caused by isolation. Our genomic analysis demonstrates that the magnitude and time of recent anthropogenic population bottlenecks can be assessed using contemporary samples, exposes the severity of whaling, emphasizes the importance of migration, and demonstrates the use of genome-based analyses and simulations to inform conservation strategies.

Key words: Population genetics, demographic history, deleterious variation, conservation, gene flow, inbreeding, demographic simulations, North Pacific, Gulf of California

Introduction

Due to increasing recent human impacts, many vertebrate species have experienced drastic population declines and now persist as small and fragmented populations[1–3]. Small populations are at higher risk of population declines due to stochastic environmental and genetic factors[4–6]. Both anthropogenic and naturally occurring population declines reduce genetic diversity, increase inbreeding and genetic load due to the stronger action of genetic drift which diminish the long-term survival and adaptive potential of populations[7,8]. However, the impact of these processes depends on the often unknown population-specific demographic histories and life history traits. For example, gene flow as low as one effective migrant per generation may counteract genetic drift and reduce the frequency of deleterious variation[9–11], but might also reduce metapopulation genetic variation[12], or introduce strongly deleterious alleles[13]. Therefore, uncovering population history and determining how detrimental genetic patterns arise in declining populations are challenging questions, but the answers are critical to developing effective conservation strategies[14].

Industrial whaling during the 20th century is arguably one of the most disruptive ecological events caused by humans[15], which decimated all great whale species and drove many of them to the brink of extinction[16,17]. Estimating the decline of whale populations is crucial to evaluate the full impact of whaling and design appropriate recovery policies, not only on whale abundance but on entire ecosystems[15,17,18]. However, quantifying the magnitude of known recent population declines in endangered vertebrate species from contemporary samples has proven difficult because the estimates based on genetic diversity capture long-term effective sizes rather than recent demographic events[19,20]. Additionally, the long life-span and generation time of whales complicate the inference of recent population size changes[21] because less generation turnover occurs in a given amount of time. Due to these challenges, previous genetic studies using contemporary samples have only indirectly estimated the impact of whaling, determining that historical abundance from whaling records and recent ecological studies are orders of magnitude lower than those based on the diversity of a few mitochondrial or nuclear markers [17–24], suggesting a slower recovery of whale populations after the end of whaling. Therefore, a direct estimation of the time and magnitude of the whaling bottleneck is still lacking for most whale populations. The analysis of whole-genome data can provide results with more power and resolution to detect recent demographic changes[25].

The fin whale (*Balaenoptera physalus*) is the second largest whale and the one most impacted by industrial whaling world-wide. In the North Pacific alone, more than 75,500 fin whales were harvested[26]. However, fin whales in the Gulf of California, Mexico, were not targeted by whalers. Nevertheless, their population has been small with limited gene flow for thousands of years[27,28]. In contrast, the Eastern North Pacific population was large, interconnected, and overexploited[29], though the population along the U.S. west coast has shown evidence of growth at 3% per year since the 1990's[30]. Here we provide one of the first direct large-scale genome-wide demographic reconstruction of whaling in a previously large population, in comparison to a never-whaled but small and isolated population. We analyze and model the whole-genome diversity of these fin whale populations having contrasting demographic histories to identify the genetic and evolutionary impacts of population reductions in large, long-lived marine mammals. Understanding the complex interaction between demographic and evolutionary factors shaping the genetic diversity in whale populations is key to improving their conservation, especially given current and future whaling threats from some countries and the challenges of climate change and human inputs to marine ecosystems[17].

Evaluating the genomic consequences of contrasting population reductions in threatened populations make our results relevant for the conservation of other small and endangered populations.

Results

Sampling, population structure and differentiation

To assess the genome-wide impact of human-induced and natural bottlenecks on fin whale populations, we generated high coverage (mean average 27x) whole genome resequencing data from 50 samples of free-ranging individuals collected between 1995 and 2017 (Figure 2.1A; Table S2.1). Thirty individuals are from regions that survived intensive whaling pressure in the Eastern North Pacific (ENP), along the coasts of California (CA; N = 9), Oregon (OR; N = 4), Washington (WA; N = 2), British Columbia (BC; N = 3) and Alaska (AK; N = 12). Additionally, we included 20 individuals from a naturally small population in the Gulf of California, Mexico (GOC), that has maintained a low population size between 300-600 individuals for thousands of years and avoided the impacts of whaling[28,29,31].

The sequences were aligned, genotyped, annotated and filtered using the minke whale genome as a reference (BalAcu1.0). Principal component analysis (PCA) separated the ENP and GOC individuals on PC1 with tight clustering of the GOC samples (Figure 2.1B). A wider dispersion pattern is observed for the ENP samples, with the Alaska samples remaining relatively clustered, suggesting some degree of differentiation of this northern population from those to the south (Figure S2.1). Admixture analysis of all the samples supports a $K = 2$ partition of ENP and GOC samples (Figures 2.1C, S2.2). We identified one ~50% admixed individual from each population (ENPCA09 and GOC010) and a small admixture fraction from GOC in the ENP population (Figure 2.1B, 2.1C). Additional admixture analysis of only ENP samples supports a K

= 1 partition of this population (Figure S2.3). F_{ST} values are higher between the GOC and ENP (F_{ST} = 0.073, P = 0.001) than between all locations within the ENP (F_{ST} = 0-0.008; Table S2.2). Assuming the highest F_{ST} of 0.008 observed within ENP, this substructure would at most inflate effective population size (*Ne*) estimates by 0.8% (Alter et al. 2007). Also, a phylogenetic analysis separated both populations into different clades, with the nodes within the ENP clade showing no bootstrap support. The two admixed individuals clustered with ENP but showed early divergence (Figure S2.4). These results indicate there are two main populations in our sample, one off the Pacific coast and the other in the Gulf of California, consistent with previous microsatellite and mitochondrial data[28,31]. In addition, our findings confirm the strong isolation of the geographically distinct Gulf population[28,32], whereas weak population substructure was observed in the eastern North Pacific.

Genome-wide patterns of variation and runs of homozygosity

We explored the genome-wide diversity patterns of fin whale populations by calculating average genome-wide heterozygosity and per-site heterozygosity in nonoverlapping 1-Mb windows. In GOC individuals we found patterns of reduced variation, with an average 1.13 heterozygotes per kb (het/kb) and an increased proportion of genomic regions with low heterozygosity (46% of windows contain < 1 het/kb). In contrast, the ENP population had much higher diversity (1.76 het/kb; two-tailed Mann-Whitney U [MWU] test P = 1.15E-10; Figure 2.2A) and few regions of low heterozygosity (12% of windows with < 1 het/kb; Figures 2.2B, S2.5, S2.6). These genome-wide results imply contrasting demographic histories of long-term small and large population size in the Gulf and North Pacific, respectively[28]. Compared with other marine mammals that have experienced different levels of population contractions, such as the diminutive vaquita in the Gulf of California[33,34] (0.1 het/kb), abundant minke whale[35] (0.6

het/kb) and endangered blue whale[36] (2.1 het/kb), the GOC fin whales have maintained moderate genome-wide patterns of variation (Figure 2.2A), suggesting that evolutionary mechanisms such as migration have maintained genetic diversity. However, the GOC population has an enriched number of 1-Mb windows with null or very low heterozygosity (0-0.1 het/kb) compared with more endangered mysticete species such as the North Atlantic right whale and blue whale (Figure S2.7), indicating that populations of these endangered species were historically larger than the Gulf of California fin whale population and a reassessment towards a more threatened status of the GOC population is needed.

To characterize the history of inbreeding events, we identified runs of homozygosity (ROH), which are genomic stretches within an individual that are assumed to be identical by descent, using two model-based methods[37,38] (Figure S2.8). Long ROH (≥ 5 Mb) typically result from recent close inbreeding whereas shorter ROH indicate either older inbreeding or older reductions in population size[39]. Overall, GOC individuals contained considerably more ROH segments than ENP individuals (two-tailed MWU test P = 9.42E-08), but most of the ROH were of short (0.1 – 1 Mb) or intermediate (1 – 5 Mb) length (Figure 2.2A). Long ROH were present in all GOC individuals, except the admixed sample GOC010, and only in three ENP individuals. Nevertheless, they comprise a small fraction of total ROH length in both populations ($F_{ROH \geq 5M} = 0.4 - 3.1\%$). To further explore the timing of inbreeding, we estimated the average time at which two homologous haplotypes could coalesce within our ROH categories for each population, assuming a recombination rate of 1 cM/Mb[40]. For short ROH, haplotypes coalesced on average approximately 145 and 250 generations ago in GOC and ENP, respectively, whereas for intermediate ROH the average haplotype coalescent time was 28 and 30 generations ago. These findings suggest a lack of recent inbreeding in both populations (Figures 2.2A, S2.9).

However, the higher number and longer ROH observed in the GOC fin whales (Figures 2.2A, S2.8, S2.9), together with the high proportion of their genome contained in ROH larger than 1Mb ($F_{ROH\geq 1M(\mathrm{GOC})} = 17.5 - 23.4\%$; Table S2.3), indicate that genomic segments in this population share a more recent common ancestor than they do in the Pacific population. Finally, we determined the relatedness between individuals in both populations and found significantly higher average kinship coefficient among GOC individuals (0.054) than in the ENP population (0.0032; two-tailed MWU test P < 2.2E-16), indicating greater identity-by-descent in the GOC, which further demonstrate higher inbreeding levels in this population (Figure S2.10A). We divided the ENP into location groups to account for larger geographical coverage and continued to observe significantly higher kinship in the GOC (Figure S2.10B, S2.10C). In summary, these results reflect the greater historical isolation and small population size of the GOC[27] and a lack of recent inbreeding in both populations.

Demographic reconstruction of whaling, divergence and gene flow

We reconstructed the demographic history of fin whale populations using the site frequency spectrum (SFS) to assess the impact of whaling in the Eastern North Pacific population and to determine the demographic events that have shaped the genomic diversity of the Gulf of California population. First, using the SFS from each population, we tested different single-population effective size (N_e) change models, employing coalescent[41] (fastsimcoal2) and diffusion approximation[42] (∂a∂i) methods. We assumed a generation time of 25.9 years[43] and a mutation rate of 2.77E-08 mutation/bp/generation[35], and tested several nested models with increasing numbers of size-change epochs (Figure S2.11). Both inference methods provided concordant findings and ∂a∂i results are shown throughout the text, except when noted (see Tables S2.4 – S2.6, for fastsimcoal2 results and all 95% confidence interval [CI] values). Our demographic analyses show that a 3-epoch model was the best fit for the ENP population (Figures 2.3A, 2.3B, S2.12A; Tables S2.4, S2.5) and revealed an expansion starting ~115 thousand years ago (kya; 4,424 generations), from an ancestral N_e of 16,479 to 23,913. This was followed by a severe decline only 26 (one generation ago for fastsimcoal2 estimate; 95% CI: 0 – 2) or 52 years before present (two generations ago for ∂a∂i estimate; 95% CI: 1.82 – 2.19) to a current N_e= 305 individuals (95% CI: 0 – 1183; Figure 2.3A, 2.3B; Table S2.6), representing an approximately 99% reduction. To further verify the timing and size of this recent population reduction, we implemented a grid search (Figure S2.13; Supplemental Methods and Supplemental Results), performed additional inference runs varying the time for the whaling reduction (Tables S2.4, S2.6), used different optimization methods (Table S2.7), confirmed our power to detect such recent decline using coalescent SFS simulations under this model (Figure S2.14), and ran supplementary inferences under a SFS without filtering on genotype calls to avoid bias against rare alleles (Tables S2.8, S2.9; Supplemental Methods and Supplemental Discussion). These additional analyses demonstrated that our findings reflect a drastic recent reduction one or two generations ago. Since the average collection year for samples from this population was 2006 (Table S2.1), the estimated times of the reduction correspond to the years 1954 to 1980, coinciding with the most intense whaling period this population suffered between 1940 and 1980[26,29].

For the Gulf of California population, none of the inferred SFS for the single-population models had a good fit to the data (Figure S2.12B). Additionally, the models with the best likelihood did not show convergence or concordant parameter estimation between inference methods (Tables S2.4 – S2.6), which can indicate an overparameterization of the models (Supplemental Results). Therefore, we inferred the demographic history of the Gulf whales using a two-population model (described below) because they have shown to contain more information than single-population models and improve demographic inference[44].

The time of divergence and migration rates between both populations were estimated by testing several two-population models based on the joint SFS between ENP and GOC (Figures S2.15, S2.16; Table S2.4). The model of an ancestral size change before the populations diverged fits our data well (Figures 2.3C, S2.16; Table S2.4), is consistent among inference methods (Tables S2.10, S2.11) and is biologically feasible, therefore it was chosen as our best model (Supplemental Results). This model predicted that before the populations separated, the ancestral population expanded from ~16,000 effective individuals to ~25,000, more than 100 kya (4,322 generations). Then, the populations split between 16 and 25 kya (616 and 960 generations, ∂a∂i and fastsimcoal2 estimates, respectively). Thereafter, the ENP population remained at N_e = 17,386 until it recently crashed due to whaling, as shown by the single-population model. By contrast, the GOC effective population size remained small after the divergence at N_e = 114. The model also inferred asymmetrical gene flow, with higher migration rate from the Pacific into the Gulf population (3.42E-03; fraction of individuals that are migrants) than in the opposite direction (9.24E-05; Table S2.10). However, when scaled by the receiving population's effective size, these rates represent a long-term effective migration of 0.39 immigrants per generation into the Gulf and 1.61 into the Pacific population (Figure 2.3C).

To test if unsampled (ghost) populations contributed to migration into the GOC, we ran additional two-population models incorporating feasible ghost populations, the South Pacific and the western North Pacific (WNP). The ghost western North Pacific had a higher log-likelihood (Table S2.12) but did not considerably increase the total migration into the Gulf of California (the migration rate and effective migration from the ghost WNP into the GOC were 2.09E-04 and 0.01, respectively; Figure S2.17; Table S2.13), demonstrating that migration from ghost populations into the GOC is negligible and do not affect our estimates. However, ghost population models revealed that the divergence between the ancestral ENP and ghost WNP populations match the expansion observed in both the single-population ENP and two-population models, around 4,300 generations ago (Supplemental Discussion; Figures 2.3A, 2.3C, S2.17; Tables S2.6, S2.10, S2.13).

Our results suggest the GOC population was founded at the end of the Wisconsin glaciation during the Last Glacial Maximum[45] and remained small and highly isolated since then, receiving < 1 migrant per generation (Figure 2.3C). These findings are substantially different from estimates based on mitochondrial and microsatellite loci that predicted more recent divergence times, ~2,300 or 9,300 years before present (123 or 360 generations ago, respectively) and ~1 migrant per generation[28,31] (see Supplemental Discussion). Therefore, our results emphasize the greater resolution of whole genome resequencing data for demographic inference empowered by the sheer availability of independent genealogies sampled[20] compared with only a handful of microsatellite loci[28] and a maternally inherited non-recombining marker.

Patterns of deleterious variation and genetic load

Our demographic inference analysis suggests a historically large population size and a recent contraction for the ENP population and a high degree of isolation for the GOC population. To assess how these demographic trajectories have impacted fitness, we examined variants in coding regions, which are more likely to have functional impacts. The derived alleles were classified into four mutation types: synonymous, tolerated nonsynonymous (SIFT score ≥ 0.05), deleterious nonsynonymous (SIFT score < 0.05), and loss-of-function (LOF; identified using snpEff, details in Methods). The synonymous and tolerated nonsynonymous mutations serve as a proxy for neutral variants whereas the deleterious nonsynonymous and LOF mutations are proxies for putatively deleterious variants[46]. Since the dominance for variants in natural populations is poorly quantified, we assumed two extreme scenarios. Specifically, dominance of all variants is fully recessive (h = 0), or fully additive (h = 0.5).

For all four mutation types, heterozygosity is significantly depleted and homozygosity is significantly elevated in the GOC population (MWU tests P = 2.9E-12 in all comparisons; Table S2.14), consistent with reduced genome-wide heterozygosity and small population size. The number of homozygous derived deleterious nonsynonymous genotypes per individual was on average 39.68% higher in the GOC (2079) compared to the ENP population (1488). Similarly, the number of homozygous derived LOF genotypes was on average 28.98% higher in the Gulf (140) compared with the Pacific population (108; Figure 2.4A). Assuming that these deleterious mutations are also at least partially recessive, this increased homozygosity in the GOC is predicted to result in reduced fitness[47].

When deleterious mutations act in an additive manner, the genetic load is determined by counts of derived alleles per genome. We found that the ENP and GOC populations showed a similar number of derived neutral alleles as expected[48] (Table S2.14). For the putatively deleterious class of mutations, only nonsynonymous alleles showed a significant 2.03% elevation in the GOC population (GOC average = 5983, ENP average = 5864, MWU test P = 1.20E-07), whereas the number of LOF alleles were similar in the two populations (P = 0.87; Figure 2.4B). Assuming that these nonsynonymous alleles are slightly deleterious, the small population size of the GOC population likely decreased the efficacy of selection compared to the larger ENP population, allowing the persistence of deleterious variants in the Gulf. By contrast, the similar number of LOF alleles indicates that, in spite of the GOC population’s small size, purifying selection has remained effective at eliminating the most deleterious mutations. Overall, these results imply a slight increase in the genetic load in the GOC population if deleterious mutations are additive.

Finally, we computed the R_{XY} (relative accumulation of derived alleles) and R^2_{XY} (relative accumulation of derived homozygotes) statistics that compares the expected number of the derived alleles or homozygotes occurring only in one population[49] (Figure 2.4C). Among the four mutation types, only the deleterious nonsynonymous alleles showed a relative accumulation of derived alleles in GOC ($R_{GOC/ENP}$ = 1.04, Z-score $P = 0.02$), similar to the allele counts pattern (Figure 2.4B). However, the R^2_{XY} was significantly elevated for all mutation types in the GOC population (Z-score $P < 0.001$ for all comparisons), consistent with their higher homozygosity values in GOC (Figure 2.4A). We repeated these analyses using snpEff's mutation impact categories (i.e., high, moderate and low) to rule out software bias (see Methods), and found similar results (Figure S2.18). In summary, these results suggest an increase in genetic load in the GOC population, both due to a shift towards higher homozygosity among all protein-coding variants, as well as an overall accumulation of putatively deleterious nonsynonymous alleles compared to the ENP population. However, the magnitude of the effect on fitness is unclear, given uncertainties about the selection and dominance coefficients of these mutations[47].

Simulations of deleterious variation and genetic load

To further explore how fin whale demographic history and the recent whaling-induced decline has shaped patterns of deleterious variation and accumulation of genetic load, we ran forward-in-time genetic simulations using SLiM v.3.3.2[50]. We simulated a 10 Mb chromosomal segment with a combination of intergenic, intronic, and exonic regions. Selection coefficients for nonsynonymous deleterious mutations were drawn from a distribution estimated from humans[51], and dominance coefficients were set such that the most deleterious mutations were highly recessive, though nearly neutral mutations were closer to additive (see Methods for details).

Using this simulation framework, we first investigated the extent to which the recent whaling bottleneck may have led to an increase in genetic load in the ENP population. Specifically, we simulated under our best-fit ENP demographic model, which includes a contraction to N_e =305 two generations ago (Figure 2.3A). After two generations at N_e = 305, we did not observe any changes in genetic load, heterozygosity, or levels of inbreeding, as expected given the short duration of this decline (Figure 2.5A). To explore how various potential recovery scenarios may impact the viability of the ENP population in the future, we continued these simulations for an additional 18 generations following the decline, during which we observed increasing trends for genetic load and levels of inbreeding, though minimal impacts on genetic diversity (Figure 2.5A). To test the impacts of a partial recovery in the ENP, we also ran simulations where we increased the effective population size to N_e = 1000 after two generations at N_e =305. Here, we observe minimal increases in genetic load and inbreeding, suggesting that even a modest recovery would stave off any deleterious genetic effects (Figure 2.5A). In conclusion, these results highlight the importance of a prompt recovery for the fin whale to minimize deleterious genetic impacts from the whaling bottleneck.

Our next aim for these simulations was to assess the importance of low levels of migration (0.39 effective migrants/gen from ENP to GOC) for maintaining genetic diversity and fitness in the small GOC population (N_e = 114) despite long-term isolation (~16 kya). We simulated under our best-fit two population demographic model, running simulations that included the estimated rates of migration between the ENP and GOC (Figure 2.3C) as well as simulations where no migration was allowed. When carrying out simulations that include the empirically inferred rate of migration from ENP to GOC, we observe a 26.7% reduction in heterozygosity and increase in $F_{ROH > 1Mb}$ from 0 to 0.10 in the GOC population compared to the ENP population (Figure 2.5B), in good agreement with the trends from our empirical dataset (35.7% empirical reduction; Figure 2.2). Additionally, we find that average genetic load in the GOC population is elevated to 7.75% compared to 2.87% in the ENP population (Figure 2.5B). However, this increase in genetic load appears to be counteracted by purging of recessive strongly deleterious mutations ($s < -0.01$), which are reduced in frequency by 22.9% in the GOC population (Figure S2.19). By contrast, we observe minimal differences in the numbers of moderately ($-0.01 < s \leq -0.001$) or weakly ($-0.001 < s \leq -0.00001$) deleterious alleles per individual (Figure S2.19), suggesting that migration has helped keep these mutations from drifting to high frequency in the GOC population. In summary, these results suggest that isolation and small population size in the GOC may have resulted in a lowered fitness, though these fitness reductions have apparently not been substantial enough to impact population viability.

When simulating without migration, we observed far more dramatic changes in the genetic composition of the GOC population. Specifically, we found a near-complete loss of genetic diversity, higher levels of inbreeding ($F_{ROH>1Mb} = 0.11$), and a substantial increase in genetic load to 10.3% in the GOC population (Figure 2.5B). The loss of diversity is also confirmed in theoretical calculations (Supplemental Results). This increase in genetic load appears to be driven primarily by fixation of moderately deleterious alleles (9.22% gain in the isolated GOC population compared with the migration scenario; Figure S2.19). Thus, these simulations suggest that, in the absence of migration, the GOC population would have experienced a much more substantial increase in genetic load, which may have been substantial enough to drive extinction. In conclusion, these results highlight the importance of low levels of migration in maintaining viability in the GOC population over its long period of isolation.

Discussion

Detecting recent population bottlenecks in endangered species using estimates of genetic diversity in contemporary samples has been challenging[19,20], especially in long-lived species with long generation times, such as the great whales[21,52]. Specifically, the influence of changes in population size on genetic diversity is slow relative to temporal scale of human-induced events[19] and the overall loss of genetic variation depends on the duration of the bottleneck relative to the life history traits[53,54] such as life-span and generation time. Although genomic data can improve our ability to detect the impact of bottlenecks, studies analyzing whole genome data have failed to detect signals of whaling in blue[36] and gray whales[55], presumably due to small sample sizes. Here, we show that using high coverage genome re-sequencing (~27X), sampling a high number of individuals (~30 per population) at a single timepoint and SFS-based demographic inference approaches, it is possible to identify recent anthropogenic population contractions, such as the one imposed by the 20th century whaling on fin whales[26,29] (Supplemental Discussion). Besides our sampling and methodological approaches, the combination of a high pre-whaling genetic variation possessed by the fin whales in the Eastern North Pacific[28,31,32,56] together with an extreme reduction of two orders of magnitude, even if short, likely caused a deficit in low-frequency variants in present-day individuals that we were able to detect[20] (Figure 2.3B), a similar signal also found in the North Atlantic fin whales[25]. Therefore, our research demonstrates that even very recent human-driven population bottlenecks leave a detectable signal in the SFS derived from genome-wide data of contemporary individuals, and this signal can be used to identify the demographic and genetic effects of recent anthropogenic exploitation and model current and future impacts on populations.

Despite a 99% decline in effective population size, the Eastern North Pacific fin whales have retained most of their pre-whaling genetic diversity (Figures 2.2, 2.5A). They do not exhibit a substantial decrease in genome-wide heterozygosity nor an increase in inbreeding or genetic load (Figures 2.2, 2.4 and 2.5A). Since genetic diversity declines exponentially with the number of generations passed from the contraction, this lagging impact on genetic diversity is likely a consequence of the long generation time of fin whales[43] (~25.9 yrs) relative to the duration of the whaling bottleneck (~70 years) and relatively prompt recovery following the whaling moratorium that came into effect in 1985[30,54,57]. The contraction, although severe, only lasted for two generations (Supplemental Results). However, other detrimental effects remain alarming. The reduction in 99% of pre-whaling effective size has likely had strong ecological consequences[15,18,58]. Additionally, if the ENP population remains small, it may experience a loss of adaptive potential to resist future climate change or disease[59]. Furthermore, this reduced condition in the ENP could also imperil the viability of the Gulf of California population by further diminishing or completely halting the migration into this population, which our simulations have shown can accelerate the accumulation of deleterious load and loss of genetic diversity in GOC individuals. Both empirical and simulation findings show that continuing the current moratorium and enhancing population size remains essential for fin whale recovery and long-term persistence[17,26].

Regarding the Gulf of California fin whale population, our results show that as few as 0.39 migrants per generation have been sufficient to maintain genetic diversity and fitness in this population over ~16,000 years of isolation (Figure 2.5B), whereas migration from unsampled ghost populations are negligible (Supplemental Discussion). By contrast, when omitting migration from our simulations, we observe a near-complete loss of genetic diversity and substantial increase in levels of inbreeding and genetic load (Figure 2.5B). Thus, these results highlight the importance of gene flow for maintaining population viability over long evolutionary timescales[11,60], even when levels of migration are far lower than the classic rule of thumb of 'one migrant per generation'[10]. This rule has been widely applied in conservation, however it is based on a neutral model that makes numerous simplifying assumptions and does not consider deleterious variation[12]. Here, we combine empirical observations with more realistic models including deleterious variation to demonstrate that small populations can be maintained by exceedingly low levels of migration, even when modest levels of genetic load may accumulate[61]. These results have important implications for conserving other small and isolated populations, where maintaining high levels of migration may not be feasible.

In addition to migration, population persistence in the GOC also appears to be enabled in part by purging of strongly deleterious mutations, as has been shown in other small vertebrate populations[62,63] including marine mammals[34]. Specifically, our simulations suggest a 22.9% reduction in the frequency of these mutations in the GOC (Figure S2.19) due to its long-term small population size, occurring despite the impact of gene flow continually reintroducing these mutations[13]. However, we were unable to detect this purging in our empirical dataset, where we observed similar numbers of putatively deleterious loss-of-function (LOF) mutations in the GOC and ENP populations (Figure 2.4). This discrepancy could be explained by LOF mutations being an imperfect proxy of strongly deleterious variation[64,65].

Here, we have assessed the genomic impacts of both natural and anthropogenic bottlenecks on the second-largest mammal. We demonstrate that it is possible to confidently estimate the magnitude and timing of recent human-driven population bottlenecks, and to determine the key role that gene flow and purging of deleterious variants play in the persistence of small isolated populations by analyzing whole-genome resequencing data from contemporary samples together with individual-based simulations. From a conservation perspective, our findings expose the severity of whaling and indicate that it is necessary to reassess the recovery goals for the ENP fin whales and the regional threatened status of the GOC population, which may warrant specific conservation actions to maintain gene flow and avert additional impacts from climate change, mortality by entanglement[66] or microplastic contamination[67]. Therefore, our study contributes to fulfilling the overdue promise of genomics to conservation biology concerning the genetic effects of very recent population reductions caused by anthropogenic activities and identifying the evolutionary and ecological processes that promote the viability of small populations[68]. Finally, we demonstrate the importance of using both genomic and simulated data to inform the conservation of intensely exploited species.

Methods

Samples and sequencing

Tissue samples from 50 fin whales (*Balaenoptera physalus*) were collected using a standard protocol to obtain skin biopsies from free-ranging cetacean species, which use a small stainless-steel biopsy dart deployed from a crossbow or rifle[70,71]. These samples were collected throughout the Eastern North Pacific (ENP; N=30, represented by individuals from the coasts of California [9], Oregon [4], Washington [2], British Columbia [3] and Alaska [12]; Table S2.1), and the Gulf of California (GOC; N=20, from seven different localities; Bahía de La Paz [3], Loreto [6], Bahía de los Angeles [5], Bahía Kino [3], North of Tiburon Island [1], Puerto Refugio [1] and out of Bahía Los Frailes [1]). All samples from the Gulf of California were obtained under the appropriate collecting permits issued by the Mexican Wildlife Agency (Dirección General de Vida Silvestre, Subsecretaría de Gestión para la Protección Ambiental, Secretaría del Medio Ambiente y Recursos Naturales; permit numbers: D0070(2)-0598, D00700(2)-14093, D00750-1537 and SGPA/DGVS/-0576). Samples from the Eastern North Pacific were collected by the Southwest Fisheries Science Center (California, USA) under US Marine Mammal Protection Act permit. DNA from the samples were extracted using the QIAGEN Mini Prep Kit (Qiagen; California, USA). The genomic libraries were prepared from extracted DNA using the Illumina TruSeq DNA PCR-free standard kit (Illumina; California, USA) following the manufacturer instructions. Whole genome sequencing was performed using the 150-bp paired-end protocol on Illumina HiSeqX or NovaSeq6000 platforms. Library preparation and sequencing were performed in Fulgent genetics' sequencing core facility (Fulgent genetics LLC; California, USA).

To compare the fin whales' genomic characteristics within Mysticeti, previously generated genomic data from four representative Mysticeti species were downloaded from the NCBI Sequence Read Archive: the minke whale (*Balaenoptera acutorostrata*), the stable and abundant rorqual; the humpback whale (*Megaptera novaeangliae*), the closest relative with fin whales; the North Atlantic right whale (*Eubalaena glacialis*) and the blue whale (*Balaenoptera musculus*), the most endangered baleen whales (Table S2.1).

Read processing and alignment

We followed the sequence reads processing and genotyping pipeline adapted from the Genome Analysis Toolkit (GATK) Best Practices Guide[72] similar to ref.[46]. Read quality was first checked using FastQC v.0.11.8[73]. Illumina adapters were removed from the paired-end sequence reads using picard (v.2.20.3) MarkIlluminaAdapters. The adapter-free paired-end reads were aligned against the minke whale (*Balaenoptera acutorostrata scammoni*) reference genome (GCF_000493695.1 [BalAcu1.0]; Scaffold N50: 12,843,668, Downloaded on November 12, 2019) using BWA-MEM v.0.7.17[74]. Mapping statistics were generated using QUALIMAP v.2.2[75]. We used the minke whale genome as a reference because the available fin whale genome assembly is much more fragmented and poorly annotated (GCA_008795845.1; Scaffold N50: 871,016) and the blue whale genome (GCF_009873245.2) did not have genome annotation in 2019 (Supplemental Methods; Figure S2.20; Table S2.15). The fin whale and minke whale are in the same genus, with a divergence time of approximately 10 million years ago[36]. The average mapping rate of fin whale reads is 99.09± 0.21% (Table S2.1), suggesting that the divergence time with minke whales did not impact read alignment.

Genotype calling and filtration

Joint genotype calling at all sites (including invariant positions) across the reference genome was performed using GATK[76] (v.3.8). We removed PCR duplicates from the bam files using picard *MarkDuplicates*. Raw variant calling was performed for each individual using GATK's *HaplotypeCaller* using the default settings for removing low-quality reads (min_mapping_quality_score=20; min_base_quality_score=20). Joint genotype calls for the 50 fin whales were generated from the raw variants using GATK *GenotypeGVCF*, excluding scaffolds shorter than 1 Mbp. The total scaffold length used for genotyping was 2,324,429,847 bp, with the excluded scaffolds constitutes only 4.4% of the total genome length (2,431,687,698 bp).

Since we do not have a database of known variants, we did not perform base quality recalibration (BQSR) or variant quality score recalibration (VQSR). Instead, we performed a stringent set of quality and depth filters for the genotype calls, keeping only high-quality biallelic

SNPs and monomorphic genotypes (Figure S2.21). Sites that **1)** had low Phred score (QUAL < 30), **2)** failed GATK recommended hard filters (QD < 2.0 || FS > 60.0 || MQ < 40.0 || MQRankSum < -12.5 || ReadPosRankSum < -8.0 || SOR > 3.0), or **3)** fell within repeat regions identified by WindowMasker[77], RepeatMasker or CpG islands identified by UCSC genome browser (total length: 1,247,900,490 bp), were marked as failed filtration (Figure S2.21A). For each individual, the sites that passed the above filters were subjected to genotype-level filtration: only genotypes with a minimum depth of eight reads and maximum depth of 2.5x mean depth; a minimum Phred score of 20 and expected allele balance (≥ 0.9 for homozygous reference genotypes; ≥ 0.2 & ≤ 0.8 for heterozygous genotypes and ≤ 0.1 for homozygous alternative genotypes) were kept. Genotypes failing these filters were converted to missing (Figure S2.21B). Afterwards, sites were further filtered if they had more than 20% missing genotypes or more than 75% heterozygous genotypes (Figure S2.21A). We repeated the genotype calling and filtration pipeline with four additional baleen whales included with 50 fin whale samples. The derived dataset ("f50b4" in the following text) was only used in the construction of neighbor-joining tree and generation of genome-wide heterozygosity comparison. An additional variant dataset ("genotype-filter-free" dataset) for the ENP individuals without any genotype-level filters was generated and used in confirmatory demographic inference (Supplemental Methods).

Variant annotations and identification of neutral regions

We annotated variant sites using two softwares, snpEff v.4.3.1[78] and SIFT4G v.6.0[79]. We used the minke whale genome annotation gtf file to build custom snpEff and SIFT4G databases with default settings. We then annotated and predicted the effects of variants with *-canon* option in snpEff and *-t* option in SIFT4G. The most deleterious effect was selected per site.

We used the minke whale as an outgroup to classify the allele ancestral states, and considered the sites in the minke whale reference sequence as ancestral. Because the minke whale has evolved since the common ancestor with these two populations of fin whales, the ancestral alleles identified may not represent the true ancestral state. However, this error is not expected to bias the relative comparison of variants between the ENP and GOC fin whales since they are equally diverged from the minke whale. To detect the putatively neutral regions for demographic modeling, we first extracted sites that passed all filters and are at least at 20 kb distance from exons or coding regions and not in CpG islands or repetitive regions. The identified regions were aligned to the zebra fish genome, using BLAST v.2.7.1[80], regions with a hit with e-value lower than 1E-10 were further removed, as they could represent conserved regions and not evolving neutrally. 397,627,899 sites were defined as neutral.

Evaluation of population structure

Population structure analyses were performed using the R package SNPRelate v.1.16.0[81]. We selected biallelic sites in the vcf that passed variant filtration criteria and converted them to gds format using function *snpgdsVCF2GDS*. Linkage disequilibrium pruning was implemented (*snpgdsLDpruning*) with an r^2 cutoff of 0.2, and a minor allele frequency cutoff of 0.10. A total of 30,350 SNPs were kept for PCA, kinship and F_{ST} analyses.

We performed the PCA analysis using the function *snpgdsPCA*. After observing the overall population structure, an additional PCA was performed within ENP individuals to inspect variation among locations. The kinship between sample pairs was assessed using PLINK's identity-by-descent method of moments approach (*snpgdsIBDMoM*). We calculated kinship at three different levels: 1) populations (groups: ENP and GOC), 2) sampling locations (groups: AK, BC, OR, WA, CA and GOC); and 3) merged middle ENP locations combining samples from BC, WA and OR (groups: AK, MENP and GOC). The two-tailed MWU test was used to compare the average kinship coefficients among groups. F_{ST} between populations, sampling locations and merged ENP locations were calculated using the Weir and Cockerham estimator[82], with a SNP missing rate at 20% (function *snpgdsFst*, missing.rate = 0.2). The significance of F_{ST} was estimated using 999 permutations described in ref.[83]. Due to the low sample size in BC, OR and WA locations, we only estimated the significance of F_{ST} between populations and merged ENP locations. To determine the potential influence from population substructure within ENP on *Ne* estimates, we calculated the population size inflation factor by 1/(1- F_{ST}) (Alter et al. 2007), using the highest F_{ST} value found in the ENP.

The LD pruned SNP set was converted to PLINK ped format using function *seqGDS2VCF* in R package SeqArray v.1.26.2[84] and PLINK v.1.90[85]. ADMIXTURE[86] (v.1.3.0) analyses were performed using values of *K* from two to six, with 10 iterations per *K*. Mean cross-validation (CV) error for each *K* was used to select the best number of ancestral populations (*K*). To further test a substructure in the ENP, additional ADMIXTURE analyses were performed within ENP individuals, using values of *K* from one to six, with the same settings described above. A neighbor joining phylogenetic tree was constructed from 32,191 LD pruned SNPs in the "f50b4" dataset using function *nj* in R package ape v.5.3[87]. 1000 bootstraps were performed, and the North Atlantic right whale ("EubGla01") was designated as the outgroup (Figure S2.4).

Calculation of heterozygosity and identification of runs of homozygosity

We defined heterozygosity as the number of heterozygous genotypes divided by the total number of called genotypes, including monomorphic sites, that passed variant filtration standards[46]. We first calculated the genome-wide heterozygosity for all scaffolds used for genotyping. Two-tailed MWU tests were used to evaluate if the genome-wide heterozygosity varied significantly between the ENP and GOC populations. We also calculated the per-site heterozygosity in non-overlapping 1 Mb windows across the scaffolds. Windows with more than 80% missing data were excluded. The missing data in these windows derive from regions that failed site filter criteria described above.

For identifying ROH, we first separated the vcf file for ENP and GOC individuals and reestimated allele frequencies within each population. ROH were identified using *bcftools roh -G30* in bcftools v.1.9[37]. Three individuals were excluded from bcftools ROH analyses to avoid biasing allele frequency estimations [ENPCA09 and GOC010 due to admixture proportion > 0.25 (Figure 2.1C); ENPOR12 due to low genotyping rate (Figure S2.21)]. Additional ROH analysis was performed using R package RZooRoH v.0.2.3[38], which can classify ROH segments into different age classes. A model with ten classes (9 ROH and 1 non-ROH) and a successive rate of three was applied (*zoomodel, K=10, base=3*). A minor allele frequency cutoff of 0.05 was used but no individual was excluded. For both methods, ROH segments less than 100 kb were discarded. The rest of the segments were divided in three length categories, short (0.1 Mb ≤ ROH < 1 Mb), intermediate (1 Mb ≤ ROH < 5 Mb) and long (≥ 5 Mb). The concordance of the two methods was confirmed (Figure S2.8) and the output from the RZooRoH analysis is shown in the main text. The proportion of genomes with ROH (F_{ROH}) was calculated as the total length of ROH passing a certain length threshold (e.g. ROH > 100 kb) within an individual divided by the total scaffold length used for genotyping (2,324,429,847 bp). We used the two-tailed MWU test to compare total number of ROH segments in all length categories obtained in the two populations.

To determine if the inbreeding observed in both fin whale populations were due to recent or older events, we estimated the average time at which two haplotypes would coalesce in each of the ROH categories (short, intermediate and long). The length of ROH associated with inbreeding decreases reciprocally due to recombination in each generation[88–90]. This relationship can be written as: $L = 100/2tr$, where L is the mean ROH length (in Mb), the constant 100 represents large segments belonging to the common ancestor in cM, t is the number of generations to the common ancestor and r is the assumed constant recombination rate of 1 cM/1Mb[40,91]. Therefore, we calculated how many generations ago two haplotypes shared a common ancestor in each of the ROH categories as $t=100/2Lr$[40].

Projected site frequency spectra

A vcf file comprising only putatively neutral SNPs was used to obtain the site frequency spectrum (SFS) within and between populations. To avoid introducing bias to our demographic inferences from known contributing factors, such as uneven read depths[92], admixture proportions[42] and highly related individuals[93], six individuals were discarded in SFS projection (Low genotype depth: "ENPOR12"; Admixture proportion > 0.25: "ENPCA01", "ENPCA09", "GOC010"; Kinship > 0.15: "GOC080", "GOC111"). To avoid uncertainties in ancestral state classifications, we computed a folded SFS. This SFS was calculated based on a hypergeometric projection implemented using easySFS (https://github.com/isaacovercast/easySFS), which minimizes the effects of missing genotypes[94] (https://dadi.readthedocs.io/en/latest/user-guide/manipulating-spectra/#projection). From this projection, an optimal number of haploid individuals with a maximized number of SNPs are identified and this number is then used to construct the folded SFS. Both the single-population SFS for each population (projected haploid size: ENP = 44, GOC = 30; projected number of SNPs: ENP = 3,410,730, GOC = 1,532,968) and the joint two-population SFS were generated (projected number of SNPs: ENP-GOC = 3,418,226). Thereafter, the count of monomorphic sites was calculated and incorporated as follows: for the single-population SFS, monomorphic sites in the neutral regions that were called in no less than the number of diploid individuals in the projection were added to the 0-bin already calculated by the projection. For the two-population SFS, monomorphic sites were computed by counting the number of monomorphic sites that were called in at least 44 haploid individuals in the ENP population and at least 30 haploid individuals in the GOC population. These sites were added up to the previous 0-0-bin of the projection.

Demographic history reconstruction

We utilized the projected neutral SFS generated above to reconstruct the demographic history of fin whales surveyed in this study using two methods: ∂a∂i[42] (v.2.2.1; Diffusion Approximations for Demographic Inference) and fastsimcoal2[41] (v.2.6; fast sequential Markov coalescent simulation).

To explore a variety of possible demographic scenarios, we first tested the following single-population models on the ENP and GOC populations separately (Figure S2.11; Table S2.6). All the models are described forward in time. For population size parameters (N_{ANC}, N_{CUR}, etc.), all values are in units of numbers of diploids. For time parameters (T, T_{CUR}, etc.), all values are in units of generations. For the ENP population, we explored two additional 3Epoch models fixing the T_{CUR} to two generations (3EpochTcur2) or three generations (3EpochTcur3).

1. 1Epoch: single epoch model with no population size change. This model provides a “null model” that estimates ancestral population size (N_{ANC}).
2. 2Epoch: two epoch model with one size change event, from the ancestral size (N_{ANC}) to the current size (N_{CUR}) occurring T generations ago.
3. 3Epoch: three epoch model with two size change events. The first event changed from the ancestral size (N_{ANC}) to a bottleneck size (N_{BOT}) and lasted for T_{BOT} generations. The

second event changed from the bottleneck size (N_{BOT}) to the current size (N_{CUR}) occurring T_{CUR} generations ago.

4. 4Epoch: four epoch model with three size change events. The first event changed from the ancestral size (N_{ANC}) to a bottleneck size (N_{BOT}) and lasted for T_{BOT} generations. The second event changed from the bottleneck size (N_{BOT}) to a recovery size (N_{REC}) and lasted for T_{REC} generations. The third event changed from the recovery size (N_{REC}) to the current size (N_{CUR}) occurring T_{CUR} generations ago. For the 3Epoch and 4Epoch models, we note that despite the population sizes were named as a "bottleneck size" or "recovery size", we did not restrict the direction of size changes (expansion or contraction) for any events.

Next, we tested the following two-population models (Figure S2.15; Table S2.10) to elucidate the divergence time and gene flow in the ENP and GOC populations:

1. Split-NoMigration: a simple population split model with no migrations. The ancestral population (N_{ANC}) diverged into the ENP (N_{ENP}) and GOC (N_{GOC}) populations occurring T generations ago. Two populations remained isolated since then.
2. Split-SymmetricMigration: an isolation-migration model. The ancestral population (N_{ANC}) diverged into the ENP (N_{ENP}) and GOC (N_{GOC}) populations occurring T generations ago. The ENP and GOC populations maintained a symmetric migration rate of m.
3. Split-AsymmetricMigration: another isolation-migration model. This model is similar to model 2 (Split-SymmetricMigration), but the ENP and GOC populations were allowed to have different values of migration rate, with $m_{ENP\text{->}GOC}$ measured as the fraction of individuals each generation in the GOC population that are new migrants from ENP, and vice versa for $m_{GOC\text{->}ENP}$.

4. Split-AsymmetricMigration-ENPChangeTw2: this model is based on model 3 (Split-AsymmetricMigration), but an ENP population size change event to N_{ENP2} is introduced after population divergence, with a fixed $T_W = 2$ generations before present. This size change event after divergence is used to model the impact of whaling bottleneck.
5. AncestralSizeChange-Split-AsymmetricMigration: this model is based on model 3 (Split-AsymmetricMigration), but an ancestral size change event from N_{ANC} to N_{ANC2} that lasted for T_A generations was introduced before population divergence.
6. AncestralSizeChange-Split-Isolation-AsymmetricMigration: this model is based on model 5 (AncestralSizeChange-Split-AsymmetricMigration), but after population divergence, an isolation period lasted for T_D, during which there is no migration between the ENP and GOC populations. Asymmetric migrations between two populations occurred T_C generations before present.
7. AncestralSizeChange-Split-AsymmetricMigration-GOCChange: this model is based on model 5 (AncestralSizeChange-Split-AsymmetricMigration), but after population divergence, the GOC population remained at N_{GOC} for T_D generations. The GOC population then experienced a size change event from N_{GOC} to N_{GOC2} that occurred T_C generations before present.

To evaluate if unsampled (ghost) populations contribute to the total migration into the GOC population, we included two feasible ghost populations into the selected two-population model, the South Pacific (SP), which diverged from the North Pacific around 1.8 Mya according to mtDNA data[31]; and the Western North Pacific (WNP) population, which has been suggested to breed separately from the ENP[29] potentially since the recent Pleistocene's interglacial periods[23]. For our demographic inference with ∂a∂i, we ran only one ghost model using the same initial parameters as in our chosen model. The initial parameter for the divergence time of ghost population was set at the expansion time in the ENP population 3Epoch model, and the size of the ghost population was fixed to the size of the ancestral population before divergence to find the best parameter space. In contrast, for fastsimcoal2 we constrained the lower and upper bounds for the divergence time of the ghost populations based on the previous knowledge mentioned above to 35000 ~ 200000 generations ago for the SP population and 100 ~ 10000 generations ago for the WNP. We also fixed the size of the ghost populations to 30000 haploids, approximately the same size of the ancestral population before the divergence.

Fastsimcoal

The coalescent simulation approach fastsimcoal2 was employed to infer parameters and composite likelihoods for the demographic models specified above, using settings adapted from ref.[94]. Each inference was performed using the Expectation-Conditional Maximization (ECM) algorithm[95], using 60 ECM cycles (-*L* 60), in which each E-step consisted of 1,000,000 coalescent trees (-*n* 1000000), computing only the SFS for the minor allele (-*m*) with the following command line.

```
fsc26 -t $header.tpl -e $header.est -n 1000000 -m -M -L 60 -q
```

The starting parameters were chosen from a uniform distribution with an imposed minimum value and flexible upper boundary. The expected SFS under the fastsimcoal2 model parameters were compared to the empirical SFS and the multinomial log-likelihood was calculated. For single-population and joint populations models, we performed 100 and 50 replicates of the inference, respectively, to confirm that both parameters and log-likelihoods converged and parameters with the maximum log-likelihood were chosen. This difference in the number of replicates is due to inference of two-population model parameters is more computationally expensive and time consuming. All estimated size parameters were obtained as the number of haploids and converted to diploids, whereas time parameters were inferred as the number of generations before present day. To control for inflations in log-likelihood estimates in models with more parameters, we performed a likelihood ratio test (LRT) for nested models with its more immediate complex model (e.g. 2Epoch vs. 1Epoch, 3Epoch vs. 2Epoch) using the equation: *–2 * [loglikelihood (simple) – loglikelihood (complex)]*. The LRT significance was evaluated with a chi-square test (χ^2) with one or two degrees of freedom, depending on the number of parameter differences between models.

The parameter confidence intervals were obtained using a parametric bootstrap[41] following the simulation functionality described in fastsimcoal2's manual (http://cmpg.unibe.ch/software/fastsimcoal26/man/fastsimcoal26.pdf page. 56). For each model, we simulated 100 SNP-based SFS from the best-fit parameters in the observed data with approximately 4 million (3,927,079 for ENP single-population models, 3,908,444 for GOC single-population models and 3,864,185 for two-population models) non-recombining segments of 100 bp, mimicking the same number of observed sites. Parameters were estimated from 20 random starting conditions for the 100 bootstrapped SFS datasets using the same settings as described above for the empirical data. 95% confidence intervals of the best-fit parameters were obtained adding and subtracting two standard deviations of the 100 bootstrap estimated parameters from the empirical best-fit parameters.

∂a∂i

For demographic inference using ∂a∂i[94], haploid sample sizes plus 5,15 and 25 were used as extrapolation grid points as recommended in ref.[42]. Lower and upper bounds of model parameters were imposed based on prior knowledge of population history, and starting parameters under these boundaries were chosen from previous knowledge or outputs from nested runs and permuted with a fold=1. We used the *optimize_log* function as our optimization algorithm, and calculated the multinomial log-likelihood for the expected SFS obtained from each optimization.

Best-fit parameter sets of each model were scaled using N_{anc} calculated by the equation $\theta = 4N_{anc}\mu L$, where L is the total sequence length of the neutral region (392,707,916 bp for ENP single-population models, 390,844,414 bp for GOC single-population models and 386,418,461 bp for two-population models), μ is the fin whale mutation rate (2.77E-08 mutations/generation/bp)[35], and θ is the optimal value of theta for the given model. Population size parameters were adjusted by N_{anc} into diploids and time parameters were re-scaled by $2N_{anc}$ into generations. The model uncertainty was assessed by estimating 95% confidence intervals of the best-fit parameters using a Godambe Information Matrix (GIM) with bootstrapped data[96]. The bootstrapped data was obtained by dividing the genome into fragments of 2Mb and generating 1000 bootstrap pseudo-replicate datasets by resampling from those, which in total amounts for sampling 2Gb that approximate the length of the reference genome. To be conservative, we chose to resample 2Gb instead of 300Mb (the size of the neutral regions we analyzed) since using a larger sample size will result in larger confidence intervals.

One hundred replicates of each model were performed with permuted starting parameters to assess convergence of the inferred parameters and composite likelihood. Parameters with the maximum log-likelihood among replicates from each model were selected and the expected SFS under these parameters was compared with the empirical SFS. LRT was calculated as previously described.

Additionally, to ensure that the results from the ENP population 3-epoch model were in fact reflecting the recent bottleneck caused by whaling, we simulated the SFS under ∂a∂i's inferred demographic scenario using msprime v.0.7.4[97]. The simulated SFS were generated using a recombination rate of 1E-8 cross-over events per base pair per generation and a mutation rate of 2.77E-8 per base pair per generation[35], with 1000 replicates and a chunk size of 2Mb. Visual inspection was performed to validate the fit of simulated SFS to the empirical data. We also performed ∂a∂i inference on msprime simulated SFS using the same settings for empirical SFS and tested if we could obtain similar parameter estimates as the empirical data to confirm that we had the power to detect a recent population contraction.

To account for the correlations of current population size (N_{CUR}) and time of most recent contraction (T_{CUR}), we carried out grid searches to find the range of possible parameter pairs that are within two log-likelihood units of the maximum likelihood estimate (MLE; Supplemental Methods).

Model selection

We determined the models that more likely represent the demographic history of the populations using the demographic models without any constraints (i.e., not fixing any of the parameters to a certain value). To select the best demographic model, we considered several features of our demographic inference results. First, the log-likelihood of the models should be the highest given the satisfaction of the following criteria. Second, a good fit of the expected SFS to the empirical SFS. Third, the estimated parameter values between the two inference methods that we used (i.e., fastsimcoal2 and ∂a∂i) should be consistent, especially the direction of population size change (expansion vs contraction). Fourth, the composite likelihood of the top 10 replicated runs for each model should converge. We consider that a model has good convergence if the log-likelihood difference between the best run and the 10th best run of the model was no more than 25 log-likelihood units. Fifth, the model had a significantly better LRT than the more parsimonious model and that this LRT significance was consistent in both fastsimcoal2 and ∂a∂i. Sixth, the range of the confidence intervals should not be unrealistically large. Models meeting the above criteria, were chosen as the ones representing the demographic history of fin whale populations. After choosing the best demographic model according to the previous criteria, we try to confirm the findings of the chosen unconstrained models by running these models with some parameters fixed at different values, specifically the time of the bottleneck for the ENP one-population three epoch model and the divergence time for the two-population model. Results show that models with fixed parameters have better log-likelihoods and do not significantly change the parameter values obtained with the unconstraint models, demonstrating that the estimations of the unconstrained models are a good representation of the demographic history.

Quantifying putatively deleterious variation

Two lines of evidence were used to quantify relative levels of putatively deleterious variation in the ENP and GOC populations. We focused on mutations within protein-coding regions, which are more likely to have direct fitness impacts and identified derived alleles within four mutation types: synonymous, tolerated nonsynonymous, deleterious nonsynonymous and loss-of-function (LOF). The nonsynonymous mutations were classified as putatively tolerated (SIFT score ≥ 0.05) or deleterious (SIFT score < 0.05) based on phylogenetic constraints using SIFT4G[79]. The LOF mutations are predicted to eliminate or severely inhibit gene function and include splice acceptor, splice donor, start lost and stop gained mutations. LOF mutations were identified using the default settings in snpEff[78], which utilized the LOF definition in ref.[65]. We normalized for differences in missing data across individuals by the average number of called genotypes. Since the dominance for variants in natural populations is poorly quantified, we assumed two extreme scenarios when the dominance of all variants is recessive (h = 0), and the fitness is only reduced in homozygous derived genotypes or when variants are additive (h = 0.5), and the reduction in fitness is linear to the number of derived alleles. The real-life fitness impact probably lies between these two scenarios. We did not assume dominant variants ($0.5 < h \leqslant 1$) given that segregating deleterious variations are very unlikely to be dominant[47].

First, two-tailed MWU tests were used to evaluate if the normalized count of derived alleles and homozygotes varied significantly between the ENP and GOC populations in these four mutation types[46]. The count of derived putatively deleterious alleles, including the deleterious nonsynonymous and LOF alleles, are considered a proxy for additive genetic load, while the count of derived homozygotes provides a proxy for recessive load[98,99].

Second, we calculated the relative accumulation of mutations R_{XY} and homozygous mutations $R^2{}_{XY}$ for the four mutation types using methods adapted from ref.[49]. Here we designated the GOC population as population X and the ENP population as population Y. At each polymorphic site i, we defined d_X^i as the count of derived alleles at that site in a sample of n_X^i haploid genomes from population X and d_Y^i as the count of derived alleles in a sample of n_Y^i haploid genomes from population Y. The expected number of derived mutations observed only in population X but not in population Y is defined as:

$$L_{X,notY} = \sum_i \quad (d_X^i/n_X^i)(1 - d_Y^i/n_Y^i)$$

And the expected number of homozygous derived mutations observed only in X but not in Y is defined as:

$$L^2_{X,notY} = \sum_i \quad (1 - \frac{2d^i_X(n^i_X - d^i_X)}{n^i_X(n^i_X - 1)})(\frac{2d^i_Y(n^i_Y - d^i_Y)}{n^i_Y(n^i_Y - 1)})$$

The ratio statistics is further defined as:

$$R_{XY} = L_{X,notY}/L_{Y,notX}$$

$$R^2_{XY} = L^2_{X,notY}/L^2_{Y,notX}$$

The standard errors of R_{XY} and R^2_{XY} were estimated from a weighted-block jackknife[49]. If selection has been equally effective and mutation rates remain the same in both populations, the R_{XY} and R^2_{XY} statistics are expected to be 1. Z-score test was used to evaluate the significance of the deviation from the null expectation.

Lastly, we assessed the robustness of the four mutation types across the genome using an additional mutation impact scoring system implemented by snpEff. SnpEff classifies variants' impact severity into HIGH, MODERATE, LOW and MODIFIER categories based on their effect types. We excluded the MODIFIER category because these mutations are mostly non-protein coding. We additionally limited the MODERATE and LOW categories within the gtf identified coding sequence (CDS) region to exclude non-protein coding mutations as well. Two-tailed MWU tests and R_{XY} analyses were performed as described above to evaluate the variation in the count of derived alleles and homozygotes (Figure S2.18). For all above analyses, we removed the six individuals that were also discarded in the demographic inference.

Genetic load simulations

We conducted forward-in-time population genetic simulations using SLiM v.3.3.2[50]. For our simulations, we assumed a 10 Mb chromosomal segment with a uniform recombination rate of 1E-8 cross-over events per base pair per generation and randomly-generated intergenic, intronic, and exonic regions, following ref.[100]. Within this chromosomal segment, mutations occurred at a rate of 2.77E-8 per base pair per generation[35], with deleterious (nonsynonymous) mutations occurring only in exonic regions at a ratio of 2.31:1 to neutral (synonymous) mutations[101]. Selection coefficients for deleterious mutations were drawn from a distribution estimated from human data[51]. We assumed an inverse relationship between selection coefficients and dominance coefficients, given empirical evidence that strongly deleterious mutations also tend to be highly recessive[47,102]. Specifically, we assumed that strongly deleterious mutations ($s < -0.01$) were fully recessive ($h = 0.0$), moderately deleterious mutations ($-0.01 \leq s < -0.001$) were partially recessive ($h = 0.1$), and weakly deleterious mutations ($-0.001 < s \leq -0.00001$) were nearly additive ($h = 0.4$).

Using this simulation framework, we simulated under our two best-fit demographic models, including a single-population model for the ENP population, and a two-population divergence model for the ENP and GOC populations (see above for details). For both models, we assumed a burn-in duration of 10x the ancestral population size. During the simulation, we kept track of several quantities for each simulated population, including mean genetic load (the reduction in individual fitness, calculated multiplicatively across sites), mean genome-wide heterozygosity, mean inbreeding coefficient (here measured as F_{ROH}, where the minimum ROH length was 1Mb), and the mean number of strongly deleterious alleles ($s < -0.01$), moderately deleterious alleles ($-0.01 \leq s < -0.001$), and weakly deleterious alleles ($-0.001 < s \leq -0.00001$) per individual. These quantities were estimated using a sample size of 40 individuals. For all simulations, we ran 25 replicates and averaged these quantities across replicates.

Figures

Figure 2.1

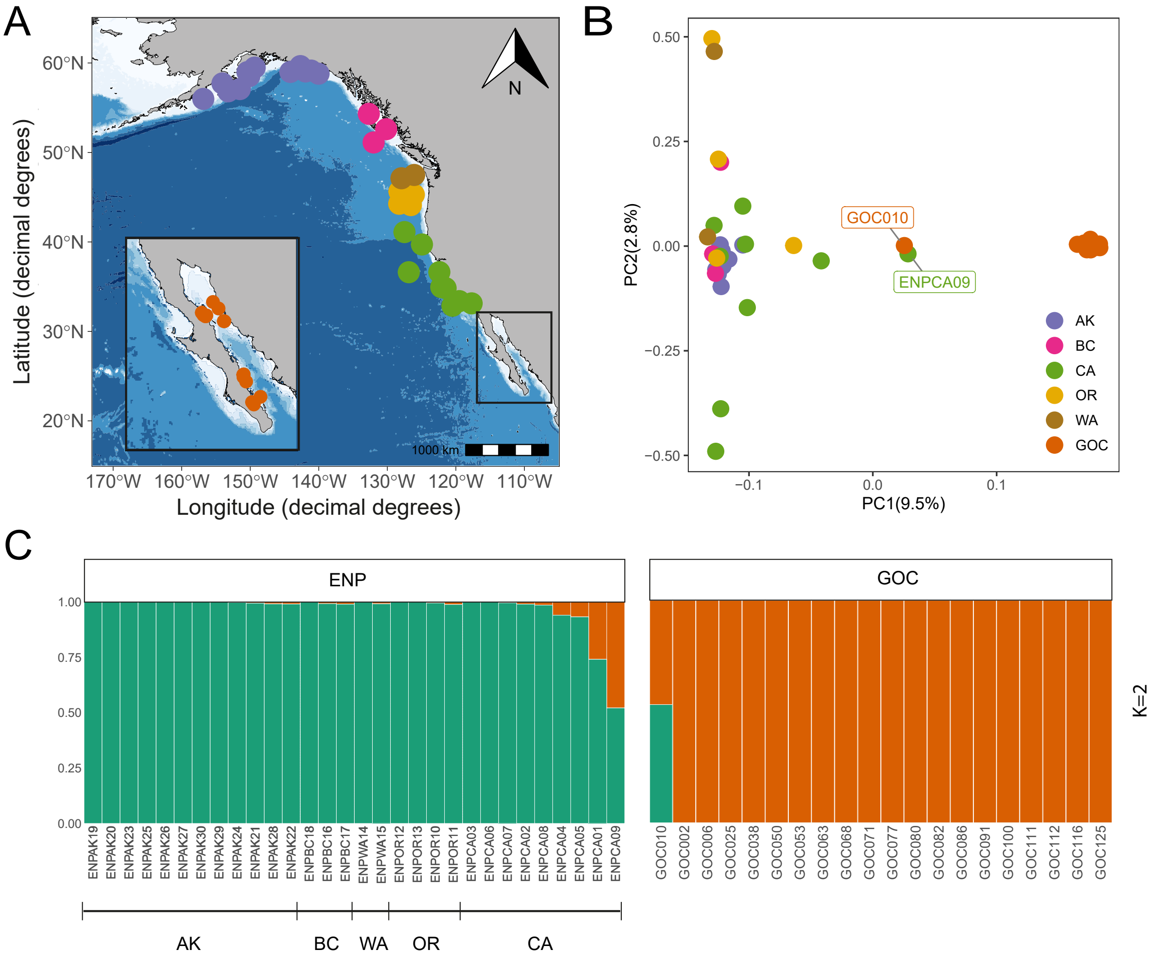

Figure 2.1. Population structure and sample origins for the fin whale genomes obtained in this study. **(A)** Thirty skin samples were collected along Eastern North Pacific (ENP) locations near Alaska (AK), British Columbia (BC), Washington (WA), Oregon (OR) and California (CA) from 1995 to 2017. Twenty samples were collected in seven sites within the Gulf of California (GOC) from Bahía de La Paz and Los Frailes in the southern Gulf to Bahía de los Ángeles, Puerto Refugio and Bahía Kino around the Midriff islands (Table S2.1). **(B)** PCA for 50 samples are colored by their location origin. The admixed individuals are labelled. **(C)** Admixture analyses supported two ancestral populations ($K = 2$).

Figure 2.2

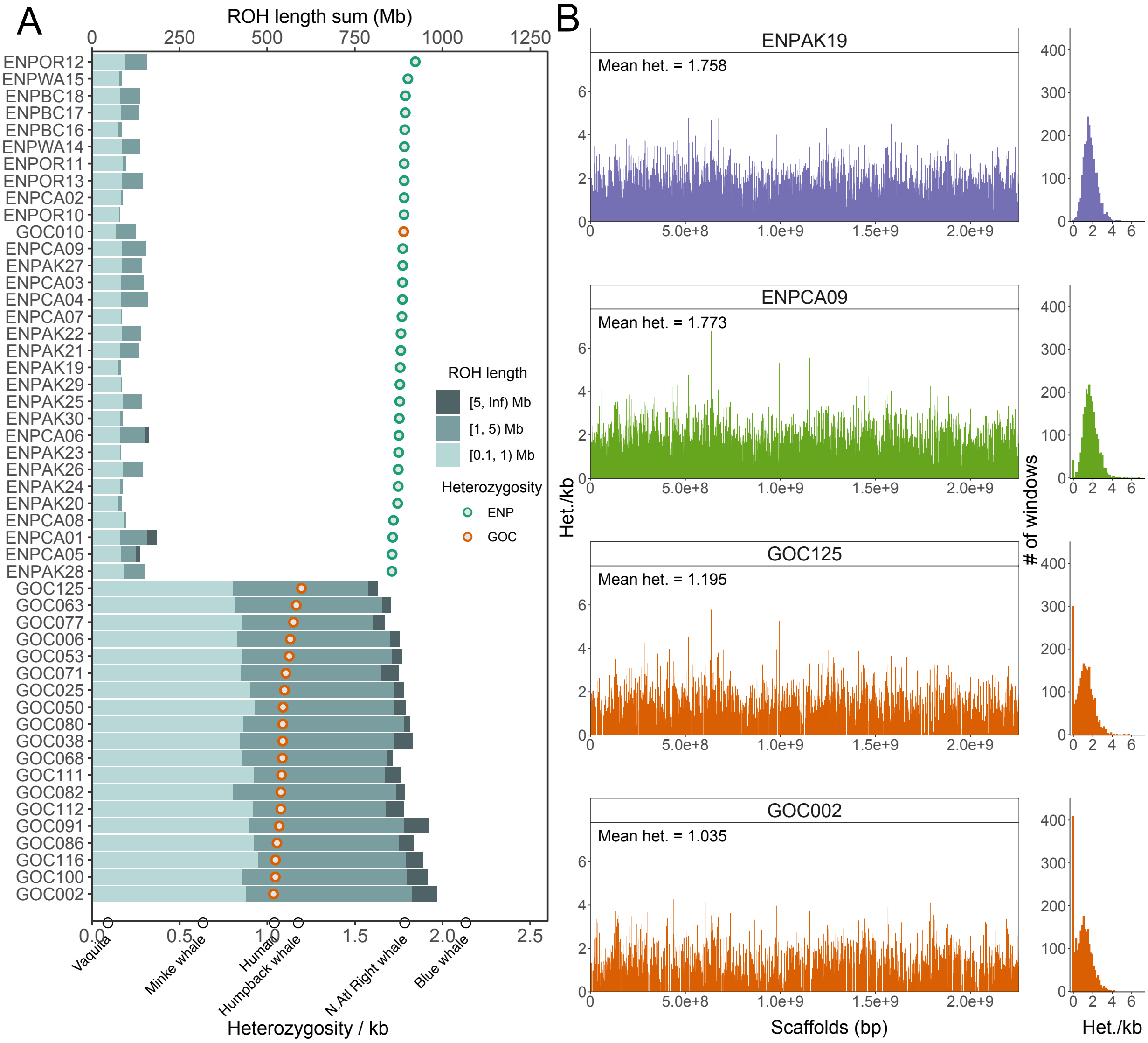

Figure 2.2. ROH and distribution of heterozygosity across the genome. **(A)** Points of genome-wide heterozygosity for each sample are ranked by decreasing heterozygosity from top to bottom. Circles at the bottom axis denote heterozygosity in other mammals. Barplots present summed lengths of short (0.1 Mb ≤ ROH < 1 Mb) to long (> 5 Mb) ROH per individual (top axis). **(B)** The left panel shows per-site heterozygosity in non-overlapping 1-Mb windows across called scaffolds. The genome-wide heterozygosity value is annotated as “Mean het”. The right

panel summarizes the distribution of per-window heterozygosity. Individuals with divergent demographic histories were selected as an example. ENPAK19 represents the large outbred Eastern North Pacific population that recently experienced whaling. ENPCA09 is an admixed individual. GOC002 and GOC125 belong to the small, isolated Gulf of California population.

Figure 2.3

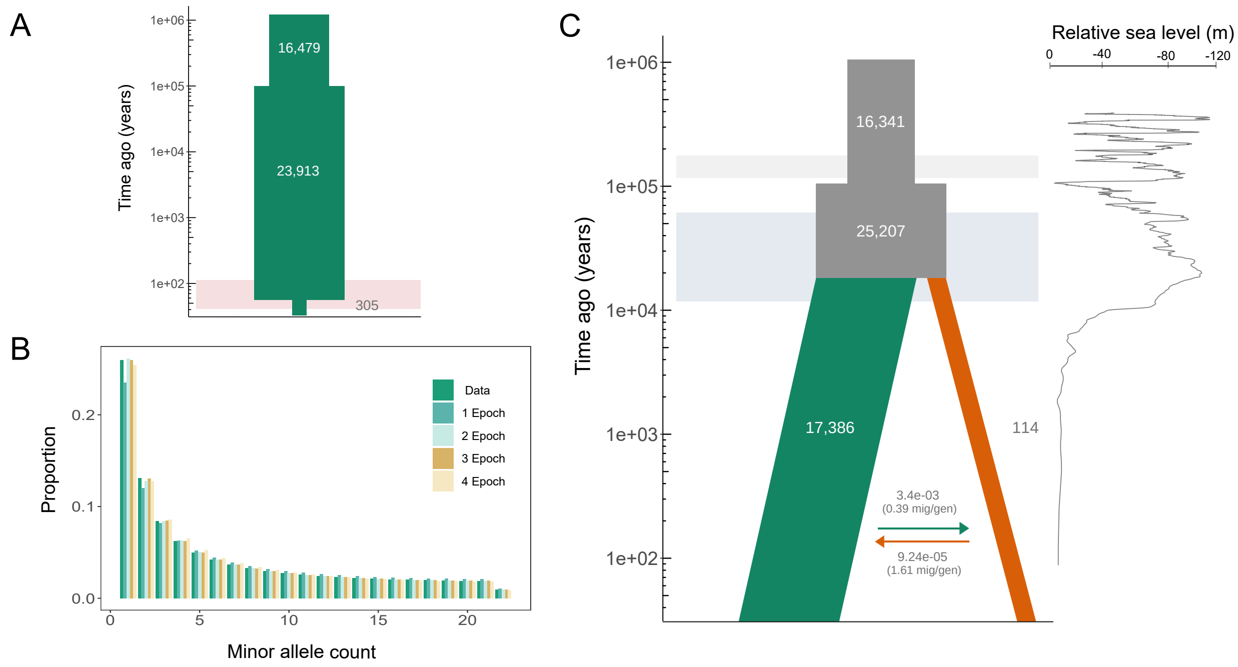

Figure 2.3. Demographic history inferred for fin whale populations. **(A)** The historical demography of the Eastern North Pacific (ENP) population is best represented by a single-population 3-epoch model. This model has an initial expansion, occurring around 115 thousand years ago (kya; 4,424 generations) followed by an approximately 99% reduction only 26 to 52 years ago (one or two generations), during the whaling period for this species in the North Pacific (red horizontal bar). **(B)** Fit of the SFS from each demographic model (1- to 4-epoch) obtained with ∂a∂i for the ENP population to the SFS from the empirical data (Data). The SFS distribution for the 3-epoch model represented in (A) shows the best fit to the data. **(C)** Two-population model showing an ancestral effective population size expansion from approximately 16,000 to 25,000 individuals during the Eemian interglacial period more than 100 kya (between the Illinois [gray bar] and Wisconsin [light blue bar] glaciations). The two populations diverged around 16 kya, during the Last Glacial Maximum. After the divergence, the ENP population remained at around 17,000 individuals, whereas the Gulf of California (GOC) population has

remained small at around 114 effective individuals. These populations have maintained low levels of asymmetrical gene flow, with higher migration rates going from ENP into GOC (3.42E-03), than vice-versa (9.24E-05). However, when scaled by the receiving population's effective size, the GOC is only receiving 0.39 effective migrants/generation, while the ENP is getting 1.61 effective migrants/gen. The black line to the right shows the relative sea level[69].

Figure 2.4

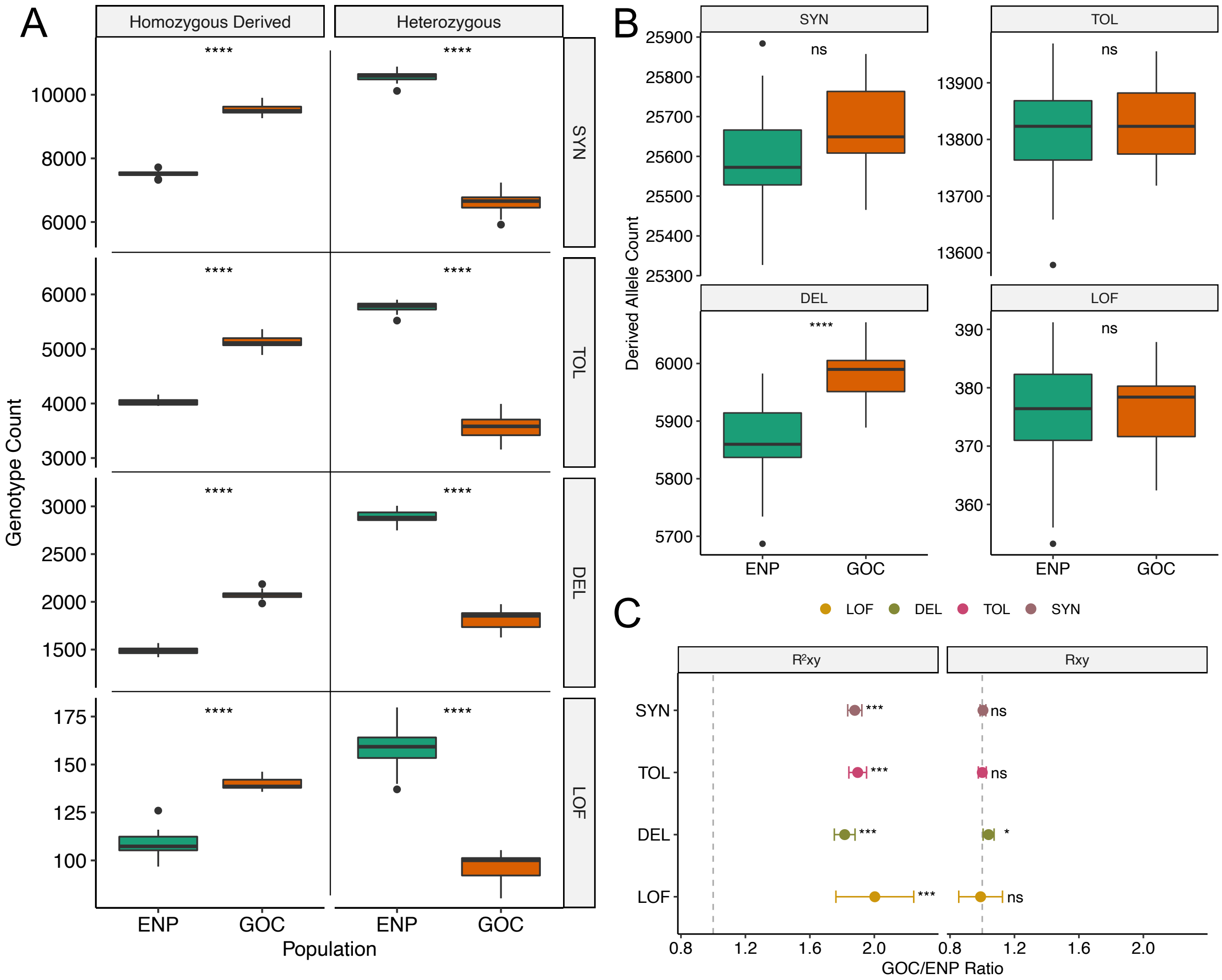

Figure 2.4. Increase in putatively deleterious variations in the GOC compared to the ENP fin whales. **(A)** The GOC fin whales contain significantly fewer heterozygous and more homozygous derived genotypes in all four functional categories of variants. **(B)** Only putatively deleterious nonsynonymous alleles (DEL) are significantly elevated (MWU test $P < 0.001$; Table S2.14) in the GOC compared with the ENP population. The ENP and GOC fin whales contain similar numbers of derived neutral alleles (SYN: synonymous and TOL: tolerated nonsynonymous), and putatively deleterious loss-of-function (LOF) alleles. In the boxplots, the notch indicates the median, and the boxes represent the 25th and 75th percentiles. The whiskers

extend to data points no more than 1.5 * IQR (inter-quantile range) from the hinges and the points show outliers beyond the whiskers. **(C)** R_{XY} and R^2_{XY} statistics in GOC (*X*) and ENP (*Y*) populations. $R_{XY} > 1$ (dashed gray line) indicates a relative accumulation of the corresponding mutation category in the GOC population. Similarly, $R^2_{XY} > 1$ indicates relative accumulation of homozygous mutations. The 2x standard error based on jackknife distribution is denoted as error bar. Significance levels: ns, not significant; * *P* < 0.01; *** *P* < 0.001; **** *P* < 0.0001.

Figure 2.5

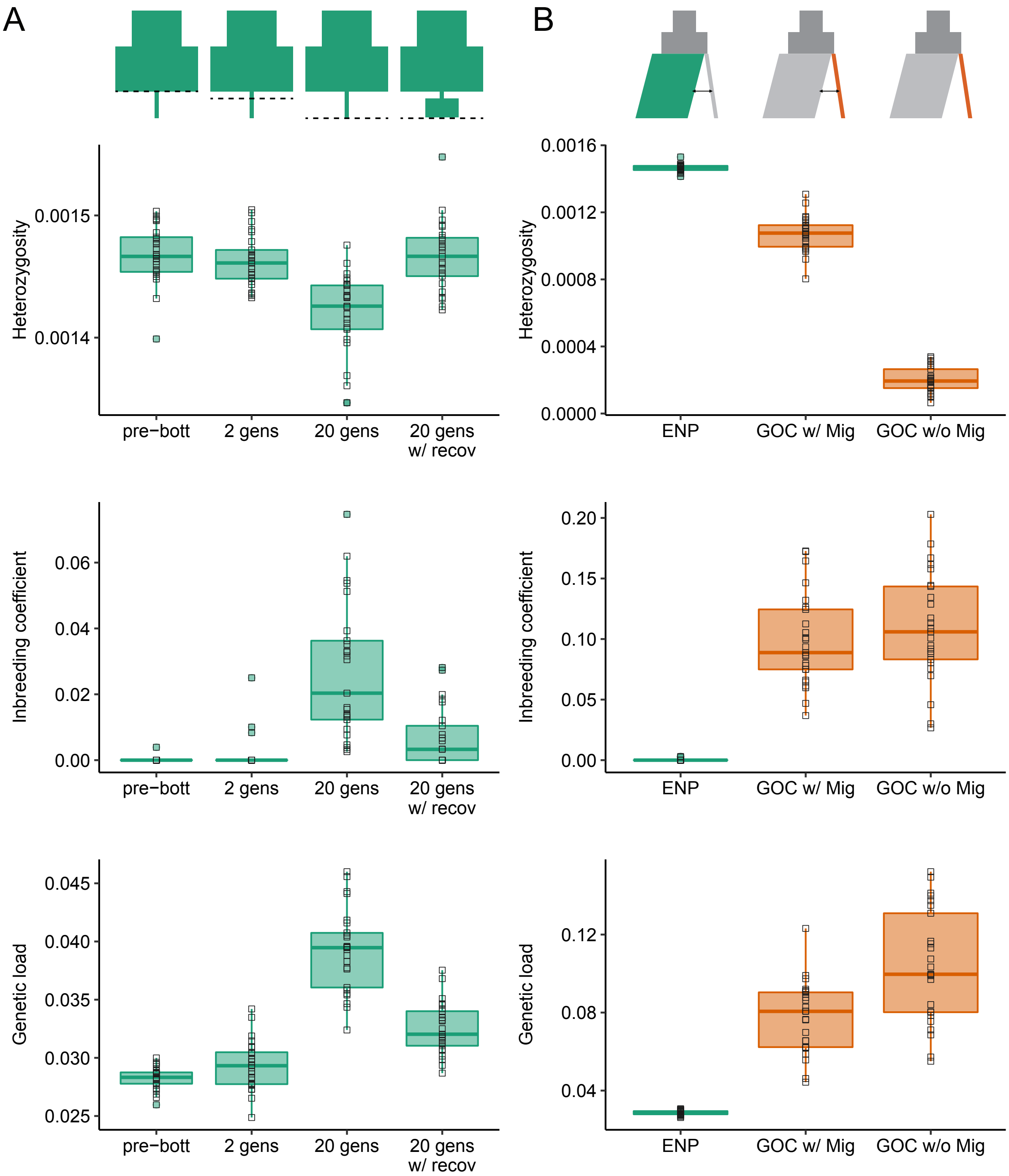

Figure 2.5. Simulations of heterozygosity, inbreeding coefficient and genetic load.

Representations of the demographic scenarios under which the simulations were performed are shown at the top. **(A)** Results for simulations under single-population 3-epoch model for the ENP population, including mean heterozygosity, levels of inbreeding ($F_{ROH>1Mb}$), and mean genetic load. Each quantity was measured prior to the onset of the whaling bottleneck (pre-bott), after two generations at the bottleneck N_e=305 (2 gens), after 20 generations at the bottleneck N_e=305 (20 gens), and 20 generations following the onset bottleneck where recovery to N_e=1000 occurred after just two generations at N_e=305 (20 gens w/ recov). In the demographic representations, the dashed line indicates the timing of sampling. **(B)** Results for simulations under our chosen two-population model. Each quantity is shown for the ENP and GOC (GOC w/mig) populations at the end of the simulation. We also simulated under a no migration demographic scenario (GOC w/o mig). Note the much lower heterozygosity, higher inbreeding, and higher genetic load in the GOC population in the absence of migration. In the demographic representations the sampled population (ENP or GOC) is shown in green or orange, and the presence/absence of migration indicated with the black arrows. For all boxplots, the notch indicates the median, and the boxes represent the 25th and 75th percentiles. The whiskers extend to data points no more than 1.5 * IQR (inter-quantile range) from the hinges and the solid squares show outliers beyond the whiskers. Hollow squares denote each simulation's value.

References

1. Ceballos, G. & Ehrlich, P. R. Mammal population losses and the extinction crisis. *Science* **296**, 904–907 (2002).

2. Pimm, S. L. *et al.* The biodiversity of species and their rates of extinction, distribution, and protection. *Science* **344**, 1246752–1246752 (2014).

3. Waters, C. N. *et al.* The Anthropocene is functionally and stratigraphically distinct from the Holocene. *Science* **351**, aad2622–aad2622 (2016).

4. Lande, R. Risks of Population Extinction from Demographic and Environmental Stochasticity and Random Catastrophes. *The American Naturalist* **142**, 911–927 (1993).

5. Reed, D. H. & Frankham, R. Correlation between Fitness and Genetic Diversity. *Conservation Biology* **17**, 230–237 (2003).

6. Melbourne, B. A. & Hastings, A. Extinction risk depends strongly on factors contributing to stochasticity. *Nature* **454**, 100–103 (2008).

7. Frankham, R. Genetics and extinction. *Biological Conservation* **126**, 131–140 (2005).

8. Willi, Y., Van Buskirk, J. & Hoffmann, A. A. Limits to the Adaptive Potential of Small Populations. *Annual Review of Ecology, Evolution, and Systematics* **37**, 433–458 (2006).

9. Wright, S. Evolution in Mendelian populations. *Genetics* **16**, 97 (1931).

10. Mills, L. S. & Allendorf, F. W. The One-Migrant-per-Generation Rule in Conservation and Management. *Conservation Biology* **10**, 1509–1518 (1996).

11. Frankham, R. Genetic rescue of small inbred populations: meta-analysis reveals large and consistent benefits of gene flow. *Mol Ecol* **24**, 2610–2618 (2015).

12. Wang, J. Application of the one-migrant-per-generation rule to conservation and management. *Conservation Biology* **18**, 332–343 (2004).

13. Kyriazis, C. C., Wayne, R. K. & Lohmueller, K. E. Strongly deleterious mutations are a primary determinant of extinction risk due to inbreeding depression. *Evolution Letters* **5**, 33–47 (2021).

14. Díez-del-Molino, D., Sánchez-Barreiro, F., Barnes, I., Gilbert, M. T. P. & Dalén, L. Quantifying Temporal Genomic Erosion in Endangered Species. *Trends in Ecology & Evolution* **33**, 176–185 (2018).

15. Springer, A. M. *et al.* Sequential megafaunal collapse in the North Pacific Ocean: An ongoing legacy of industrial whaling? *PNAS* **100**, 12223–12228 (2003).

16. Clapham, P. J., Young, S. B. & Brownell, Jr. R. L. Baleen whales: conservation issues and the status of the mostendangered populations. *Mammal Review* **29**, 35–60 (1999).

17. Baker, C. S. & Clapham, P. J. Modelling the past and future of whales and whaling. *Trends in Ecology & Evolution* **19**, 365–371 (2004).

18. Jackson, J. A., Patenaude, N. J., Carroll, E. L. & Baker, C. S. How few whales were there after whaling? Inference from contemporary mtDNA diversity. *Molecular Ecology* **17**, 236–251 (2008).

19. Palsbøll, P. J., Peery, M. Z., Olsen, M. T., Beissinger, S. R. & Bérubé, M. Inferring recent historic abundance from current genetic diversity. *Mol Ecol* **22**, 22–40 (2013).

20. Beichman, A. C., Huerta-Sanchez, E. & Lohmueller, K. E. Using genomic data to infer historic population dynamics of nonmodel organisms. *Annual Review of Ecology, Evolution, and Systematics* (2018).

21. Beland, S. L., Frasier, B. A., Darling, J. D. & Frasier, T. R. Using pre- and postexploitation samples to assess the impact of commercial whaling on the genetic characteristics of eastern North Pacific gray and humpback whales and to compare

methods used to infer historic demography. *Marine Mammal Science* vol. 36 398–420 (2020).

22. Roman, J. & Palumbi, S. R. Whales Before Whaling in the North Atlantic. *Science* **301**, 508–510 (2003).

23. Alter, S. E., Rynes, E. & Palumbi, S. R. DNA evidence for historic population size and past ecosystem impacts of gray whales. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 15162–15167 (2007).

24. Ruegg, K. *et al.* Long-term population size of the North Atlantic humpback whale within the context of worldwide population structure. *Conserv Genet* **14**, 103–114 (2013).

25. Wolf, M., de Jong, M., Halldórsson, S. D., Árnason, Ú. & Janke, A. Genomic Impact of Whaling in North Atlantic Fin Whales. *Molecular Biology and Evolution* **39**, msac094 (2022).

26. Rocha, R. C., Clapham, P. J. & Ivashchenko, Y. V. Emptying the oceans: A summary of industrial Whaling catches in the 20th century. *Marine Fisheries Review* **76**, 37–48 (2014).

27. Nigenda-Morales, S., Flores-Ramirez, S., Urban-R, J. & Vazquez-Juarez, R. MHC DQB-1 Polymorphism in the Gulf of California Fin Whale (Balaenoptera physalus) Population. *Journal of Heredity* **99**, 14–21 (2008).

28. Rivera-León, V. E. *et al.* Long-term isolation at a low effective population size greatly reduced genetic diversity in Gulf of California fin whales. *Scientific Reports* **9**, 12391 (2019).

29. Mizroch, S. A., Rice, D. W., Zwiefelhofer, D., Waite, J. & Perryman, W. L. Distribution and movements of fin whales in the North Pacific Ocean. *Mammal Review* **39**, 193–227 (2009).

30. Moore, J. E. & Barlow, J. Bayesian state-space model of fin whale abundance trends from a 1991-2008 time series of line-transect surveys in the California Current: Bayesian trend analysis from line-transect data. *Journal of Applied Ecology* **48**, 1195–1205 (2011).

31. Pérez-Álvarez, Mj. *et al.* Contrasting phylogeographic patterns among Northern and Southern Hemisphere fin whale populations with new data from the Southern Pacific. *Front. Mar. Sci.* **8**, (2021).

32. Bérubé, M., Urbán, J., Dizon, A. E., Brownell, R. L. & Palsbøll, P. J. Genetic identification of a small and highly isolated population of fin whales (Balaenoptera physalus) in the Sea of Cortez, México. *Conservation Genetics* **3**, 183–190 (2002).

33. Morin, P. A. *et al.* Reference genome and demographic history of the most endangered marine mammal, the vaquita. *Mol Ecol Resour* **21**, 1008–1020 (2021).

34. Robinson, J. A. *et al.* The critically endangered vaquita is not doomed to extinction by inbreeding depression. *Science* **376**, 635–639 (2022).

35. Yim, H.-S. *et al.* Minke whale genome and aquatic adaptation in cetaceans. *Nature Genetics* **46**, 88–92 (2014).

36. Árnason, Ú., Lammers, F., Kumar, V., Nilsson, M. A. & Janke, A. Whole-genome sequencing of the blue whale and other rorquals finds signatures for introgressive gene flow. *Science Advances* **4**, eaap9873 (2018).

37. Narasimhan, V. *et al.* BCFtools/RoH: a hidden Markov model approach for detecting autozygosity from next-generation sequencing data. *Bioinformatics* **32**, 1749–1751 (2016).

38. Bertrand, A. R., Kadri, N. K., Flori, L., Gautier, M. & Druet, T. RZooRoH: An R package to characterize individual genomic autozygosity and identify homozygous-by-descent segments. *Methods Ecol Evol* **10**, 860–866 (2019).

39. Kirin, M., Mcquillan, R., Franklin, C. S., Campbell, H. & Mckeigue, P. M. Genomic Runs of Homozygosity Record Population History and Consanguinity. *PLoS ONE* **5**, 13996 (2010).

40. Browning, S. R. Estimation of Pairwise Identity by Descent From Dense Genetic Marker Data in a Population Sample of Haplotypes. *Genetics* **178**, 2123–2132 (2008).

41. Excoffier, L., Dupanloup, I., Huerta-Sánchez, E., Sousa, V. C. & Foll, M. Robust Demographic Inference from Genomic and SNP Data. *PLOS Genetics* **9**, e1003905 (2013).

42. Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H. & Bustamante, C. D. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS genet* **5**, e1000695 (2009).

43. Taylor, B. L., Chivers, S. J., Larese, J. & Perrin, W. F. *Generation length and percent mature estimates for IUCN assessments of cetaceans*. http://swfsc.noaa.gov/BarbTaylorPubs.aspx (2007).

44. McCoy, R. C., Garud, N. R., Kelley, J. L., Boggs, C. L. & Petrov, D. A. Genomic inference accurately predicts the timing and severity of a recent bottleneck in a nonmodel insect population. *Molecular Ecology* **23**, 136–150 (2014).

45. Clark, P. U. *et al.* The Last Glacial Maximum. *Science* **325**, 710–714 (2009).

46. Robinson, J. A. *et al.* Genomic signatures of extensive inbreeding in Isle Royale wolves, a population on the threshold of extinction. *Science Advances* **5**, eaau0757 (2019).

47. Huber, C. D., Durvasula, A., Hancock, A. M. & Lohmueller, K. E. Gene expression drives the evolution of dominance. *Nature Communications* **9**, 2750 (2018).

48. Simons, Y. B., Turchin, M. C., Pritchard, J. K. & Sella, G. The deleterious mutation load is insensitive to recent population history. *Nat Genet* **46**, 220–224 (2014).

49. Do, R. *et al.* No evidence that selection has been less effective at removing deleterious mutations in Europeans than in Africans. *Nature genetics* **47**, 126 (2015).

50. Haller, B. C. & Messer, P. W. SLiM 3: Forward Genetic Simulations Beyond the Wright–Fisher Model. *Molecular Biology and Evolution* **36**, 632–637 (2019).

51. Kim, B. Y., Huber, C. D. & Lohmueller, K. E. Inference of the Distribution of Selection Coefficients for New Nonsynonymous Mutations Using Large Samples. *Genetics* **206**, 345–361 (2017).

52. Baker, C. S. *et al.* Abundant mitochondrial DNA variation and world-wide population structure in humpback whales. *Proceedings of the National Academy of Sciences* **90**, 8239–8243 (1993).

53. Nei, M., Maruyama, T. & Chakraborty, R. The bottleneck effect and genetic variability in populations. *Evolution* 1–10 (1975).

54. Amos, B. Levels of genetic variability in cetacean populations have probably changed little as a result of human activities. *Report of the International Whaling Commission* **46**, 657–658 (1996).

55. Brüniche-Olsen, A. *et al.* The inference of gray whale (Eschrichtius robustus) historical population attributes from whole-genome sequences. *BMC Evolutionary Biology* **18**, 87 (2018).

56. Archer, F. I. *et al.* Mitogenomic Phylogenetics of Fin Whales (Balaenoptera physalus spp.): Genetic Evidence for Revision of Subspecies. *PLOS ONE* **8**, e63396 (2013).

57. Aguilar, A. & García-Vernet, R. Fin whale: Balaenoptera physalus. in *Encyclopedia of marine mammals* 368–371 (Elsevier, 2018).

58. Essington, T. E. 5. Pelagic Ecosystem Response To A Century Of Commercial Fishing And Whaling. in *Whales, Whaling, and Ocean Ecosystems* (eds. Estes, J. A., DeMaster, D. P., Doak, D. F., Williams, T. M. & Brownell, R. L.) 38–49 (University of California Press, 2007). doi:10.1525/9780520933200-009.

59. Hoffmann, A. A., Sgrò, C. M. & Kristensen, T. N. Revisiting Adaptive Potential, Population Size, and Conservation. *Trends in Ecology & Evolution* **32**, 506–517 (2017).

60. Slatkin, M. Gene flow and the geographic structure of natural populations. *Science* **236**, 787–792 (1987).

61. Hedrick, P. W. & Garcia-Dorado, A. Understanding Inbreeding Depression, Purging, and Genetic Rescue. *Trends in Ecology & Evolution* **31**, 940–952 (2016).

62. Xue, Y. *et al.* Mountain gorilla genomes reveal the impact of long-term population decline and inbreeding. *Science* **348**, 242–245 (2015).

63. Grossen, C., Guillaume, F., Keller, L. F. & Croll, D. Purging of highly deleterious mutations through severe bottlenecks in Alpine ibex. *Nature Communications* **11**, 1–12 (2020).

64. Cooper, G. M. & Shendure, J. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. *Nature Reviews Genetics* **12**, 628–640 (2011).

65. MacArthur, D. G. *et al.* A Systematic Survey of Loss-of-Function Variants in Human Protein-Coding Genes. *Science* **335**, 823–828 (2012).

66. López, M. E. J., Palacios, D. M., Legorreta, A. J., R, J. U. & Mate, B. R. Fin whale movements in the Gulf of California, Mexico, from satellite telemetry. *PLOS ONE* **14**, e0209324 (2019).

67. Fossi, M. C. *et al.* Are baleen whales exposed to the threat of microplastics? A case study of the Mediterranean fin whale (Balaenoptera physalus). *Marine Pollution Bulletin* **64**, 2374–2379 (2012).

68. Shafer, A. B. *et al.* Genomics and the challenging translation into conservation practice. *Trends in ecology & evolution* **30**, 78–87 (2015).

69. Grant, K. M. *et al.* Sea-level variability over five glacial cycles. *Nat Commun* **5**, 5076 (2014).

70. Lambertsen, R. H. A Biopsy System for Large Whales and Its Use for Cytogenetics. *Journal of Mammalogy* **68**, 443–445 (1987).

71. Harlin, A. D., Würsig, B., Baker, C. S. & Markowitz, T. M. Skin swabbing for genetic analysis: Application to dusky dolphins (Lagenorhynchus obscurus). *Marine Mammal Science* **15**, 409–425 (1999).

72. Van der Auwera, G. A. *et al.* From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. *Current protocols in bioinformatics* **43**, 11–10 (2013).

73. Andrews, S. FastQC: A Quality Control Tool for High Throughput Sequence Data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (2010).

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

75. Okonechnikov, K., Conesa, A. & García-Alcalde, F. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics* **32**, 292–294 (2016).

76. McKenna, A. *et al.* The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* **20**, 1297–1303 (2010).

77. Morgulis, A., Gertz, E. M., Schäffer, A. A. & Agarwala, R. WindowMasker: window-based masker for sequenced genomes. *Bioinformatics (Oxford, England)* **22**, 134–141 (2006).

78. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly* **6**, 80–92 (2012).

79. Vaser, R., Adusumalli, S., Leng, S. N., Sikic, M. & Ng, P. C. SIFT missense predictions for genomes. *Nature protocols* **11**, 1 (2016).

80. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).

81. Zheng, X. *et al.* A High-performance Computing Toolset for Relatedness and Principal Component Analysis of SNP Data. *Bioinformatics* **28**, 3326–3328 (2012).

82. Weir, B. S. & Cockerham, C. C. Estimating F-statistics for the analysis of population structure. *evolution* 1358–1370 (1984).

83. Hudson, R. R., Boos, D. D. & Kaplan, N. L. A statistical test for detecting geographic subdivision. *Mol Biol Evol* **9**, 138–151 (1992).

84. Zheng, X. *et al.* SeqArray – A storage-efficient high-performance data format for WGS variant calls. *Bioinformatics* (2017) doi:10.1093/bioinformatics/btx145.

85. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American journal of human genetics* **81**, 559–575 (2007).

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

87. Paradis, E. & Schliep, K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35**, 526–528 (2019).

88. Pool, J. E. & Nielsen, R. Inference of Historical Changes in Migration Rate From the Lengths of Migrant Tracts. *Genetics* **181**, 711–719 (2009).

89. Thompson, E. A. Identity by Descent: Variation in Meiosis, Across Genomes, and in Populations. *Genetics* **194**, 301–326 (2013).

90. Hooper, R. *et al.* Runs of homozygosity in killer whale genomes provide a global record of demographic histories. *bioRxiv* 2020.04.08.031344 (2020) doi:10.1101/2020.04.08.031344.

91. Dumont, B. L. & Payseur, B. A. Evolution of the Genomic Rate of Recombination in Mammals. *Evolution* **62**, 276–294 (2008).

92. Han, E., Sinsheimer, J. S. & Novembre, J. Characterizing Bias in Population Genetic Inferences from Low-Coverage Sequencing Data. *Molecular Biology and Evolution* **31**, 723–735 (2014).

93. Blischak, P. D., Barker, M. S. & Gutenkunst, R. N. Inferring the Demographic History of Inbred Species from Genome-Wide SNP Frequency Data. *Molecular Biology and Evolution* **37**, 2124–2136 (2020).

94. Beichman, A. C. *et al.* Genomic analyses reveal range-wide devastation of sea otter populations. *Molecular Ecology* mec.16334 (2022) doi:10.1111/mec.16334.

95. Meng, X.-L. & Rubin, D. B. Maximum Likelihood Estimation via the ECM Algorithm: A General Framework. *Biometrika* **80**, 267–278 (1993).

96. Coffman, A. J., Hsieh, P. H., Gravel, S. & Gutenkunst, R. N. Computationally Efficient Composite Likelihood Statistics for Demographic Inference. *Molecular Biology and Evolution* **33**, 591–593 (2016).

97. Kelleher, J., Etheridge, A. M. & McVean, G. Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes. *PLOS Computational Biology* **12**, e1004842 (2016).

98. Lohmueller, K. E. *et al.* Proportionally more deleterious genetic variation in European than in African populations. *Nature* **451**, 994–997 (2008).

99. Beichman, A. C. *et al.* Aquatic Adaptation and Depleted Diversity: A Deep Dive into the Genomes of the Sea Otter and Giant Otter. *Molecular Biology and Evolution* **36**, 2631–2655 (2019).

100. Mooney, J. A. *et al.* Understanding the Hidden Complexity of Latin American Population Isolates. *The American Journal of Human Genetics* **103**, 707–726 (2018).

101. Huber, C. D., Kim, B. Y., Marsden, C. D. & Lohmueller, K. E. Determining the factors driving selective effects of new nonsynonymous mutations. *Proceedings of the National Academy of Sciences* **114**, 4465–4470 (2017).

102. Agrawal, A. F. & Whitlock, M. C. Inferences About the Distribution of Dominance Drawn From Yeast Gene Knockout Data. *Genetics* **187**, 553–566 (2011).

Chapter 3: Variation of the Distribution of Fitness Effects Across Animals

In preparation for submission

Supplementary materials available online as this dissertation's Supplementary Materials: Chapter3_Supplementary_Information.pdf and Chapter3_Supplementary_Tables.xlsx

Abstract

The distribution of fitness effects (DFE) describes the selection coefficients (*s*) of newly arising mutations and fundamentally influences population genetics processes. Despite only been inferred in a handful of organisms, the DFE varies across species. However, the extent and mechanisms of DFE variation in natural populations have not been systematically investigated across species with divergent phylogenetic histories and distinct ecological functions. Here, we inferred the DFE in natural populations of eight animals, including human, mice, fin whales, vaquitas, wolves, flycatchers, *drosophila*, and mosquitos. We developed new software, *varDFE,* to facilitate robust and flexible comparisons. We find that the DFE is more similar in more closely related species. Additionally, mammals have a higher proportion of strongly deleterious mutations and lower proportion of weakly deleterious mutations than insects. Population size is strongly correlated with the average impact of new deleterious mutations, but by itself, does not explain all the variation in the DFE across species. We next tested several models underlying how the DFE may vary across species, including the protein stability model, the mutation robustness model, and the Fisher's Geometric Model (FGM). Of these, the FGM was the most supported by the data. This study provides new insights into the long-standing question concerning the evolutionary stability of the DFE across species.

Key words: distribution of selection coefficients, DFE, varDFE, fisher's geometric model, comparative population genomics, deleterious variations, organism complexity, effective population size

Introduction

The distribution of fitness effects (DFE) is a fundamental concept in the study of evolutionary genetics. The fitness effect of new mutations (selection coefficient, *s*) measures whether the new mutation is deleterious ($s < -1e-5$), (nearly) neutral ($-1e-5 \leq s \leq 0$) or beneficial ($s > 0$) to the host. The DFE quantifies the relative proportion of mutations having different selection coefficients (Eyre-Walker and Keightley 2007). In addition to its intrinsic values for understanding how mutation affects fitness, the DFE plays an important role in shaping genetic variation and complex traits in natural populations (Eyre-Walker and Keightley 2007; Chen et al. 2022).

Considerable efforts have been put into estimation of the DFE in natural populations through comparing neutrally evolving and selected sites in genomics data sets (Keightley and Eyre-Walker 2007; Boyko et al. 2008). Until recently, the DFE had been inferred only in several model organisms, such as humans (Boyko et al. 2008; Kim et al. 2017), *drosophila* (Keightley and Eyre-Walker 2007; Eyre-Walker and Keightley 2009), and mice (Huber et al. 2017; Zhen et al. 2021). The paucity of high-quality whole genome resequencing (WGS) data for other organisms, the lack of a computationally efficient approach for DFE inference (Galtier 2016; Kim et al. 2017; Tataru et al. 2017), and other confounding factors, such as mutation rate estimations, have all prevented extending DFE inference to more species. The recent increase in available genomic resources and software for non-model organisms now offers unprecedented opportunities to investigate this fundamental question (Bourgeois and Warren 2021; Formenti et al. 2022). Currently, the DFE for the nematodes (Gilbert et al. 2022), flycatchers (Bolívar et al. 2018), nine great apes populations (Castellano et al. 2019), the Hawaiian monk seals (Gaughran 2021), the *Arabidopsis* (Hämälä and Tiffin 2020), seven cotton wood species (Liu et al. 2022), three oak species (Liang et al. 2022), and wild tomatoes (Huang et al. 2021) have been inferred. These estimates agree that most amino-acid changing mutations are nearly neutral or weakly deleterious, whereas beneficial and strongly deleterious mutations are rare (Eyre-Walker and Keightley 2007). However, the DFE is variable across species to some extent (Chen et al. 2017; Huber et al. 2017).

Understanding the diversity of mutational effects and correlating DFE with biological features of diverse species, is crucial for testing theoretical models of evolution (Huber et al. 2017; Chen et al. 2022) and has practical implications for conserving endangered species (Kyriazis et al. 2021; Wade et al. 2022). Several theoretical models proposed that organism complexity and long-term population sizes (N_a) are key drivers of DFE evolution but their predictions vary (Lourenço et al. 2011; Siegal and Leu 2014). Various definitions of organism complexity exist, such as the number of genes (Martin and Lenormand 2006) or unique cell types (Valentine et al. 1994; Quake 2022). These metrics, although straightforward, are often hard to quantify and sometimes incongruent (Tenaillon et al. 2007). In this study, we define the organism complexity as the dimensionality of phenotypic space, in other words, the number of genetically uncorrelated phenotypic traits under selection, as outlined in the Fisher's Geometric Model (FGM; Tenaillon et al. 2007; Tenaillon 2014). The protein stability model proposes that the distribution of fitness effects at population-scale ($2N_as$) is the same across species, therefore, selection is more effective in larger populations and organism complexity does not impact fitness effects. The mutation robustness model postulates that in complex organisms, more highly connected networks lead to more robustness, therefore, more complex organisms have fewer deleterious mutations. On the contrary, the Fisher's Geometric Model predicts that as organisms become more complex, random mutations will have larger effect sizes. Therefore, in the FGM, more complex organisms will have more deleterious mutations.

Despite its importance, the extent and mechanisms of DFE variation have rarely been surveyed to evaluate competing evolution models or across divergent phylogenetic groups due to previous technical challenges to inferring the DFE. Existing single lineage studies inferred DFE using various assumptions and methods, making direct comparisons of published DFE less tractable. Only the DFE in humans and *drosophila* were tested against evolution models. Humans were found to bear a higher proportion of strongly deleterious mutations than *drosophila,* lending support for the FGM model (Huber et al. 2017; Zhen et al. 2021). A larger comparison fitted a gamma distribution of DFE for 62 animal and plant species, and found that the shape parameter is less variable in closer related species (Chen et al. 2017), an observation also reported in nine great ape species (Castellano et al. 2019). However, mean selection effect (E[*s*]) were not estimated with confidence (Chen et al. 2017). With the increase in genomic resource in non-model organisms, a comprehensive comparison has recently become possible.

Here, we evaluated the extent of DFE variation in eight animal species with diverse phylogenetic relationships, life histories, organism complexity and long-term population size to test predictions from competing evolutionary genetic models. We implemented our comparative workflow in a software *varDFE,* an extension of ∂a∂i package (Gutenkunst et al. 2009). Mammals were found to overall have more strongly deleterious mutations than insects. The stability and variation of DFE within and among major evolutionary lineages, although intuitive, has not been systematically surveyed prior to our study. When evaluating the DFE with the protein stability model, the mutation robustness model and the Fisher's Geometric Model, the

FGM aligned best with our observations, although the protein stability model cannot be completely ruled out.

Results

varDFE package*: scaling up inferences and comparisons of DFE*

To implement a robust but flexible workflow to test DFE variations, we introduce a python API *varDFE* (https://github.com/meixilin/varDFE) based on the ∂a∂i (Gutenkunst et al. 2009) and Fit∂a∂i (Kim et al. 2017; currently the *DFE* module in ∂a∂i v.2.1.1) packages. *varDFE* automates DFE inference in four steps using each dataset's synonymous and nonsynonymous site frequency spectra (SFS) as inputs (Figure S3.1). In addition, our package extends the functionality of ∂a∂i with ability to explore positive selection coefficients, fit more functional forms of the DFE, and automate a grid search for optimal parameter values. Our package also provides automatic quality control features, such as plotting, convergence testing, and quantifying uncertainty in the estimates by the Fisher's Information Matrix.

Polymorphism data from eight animal species

Taking advantages of the increasing genomic resources in non-model organisms, we retrieved high-quality population-level polymorphism dataset for eight animal species. Two insect species (mosquitos and *drosophila*), one bird species (pied flycatchers) and five mammal species (arctic wolves, vaquitas, fin whales, mice and humans) were included (Table 3.1). All datasets were filtered with the same standard to retain only high-confidence genotype calls and at least eight diploid high-quality samples in coding regions (see Methods for details). We tallied the number of variants at different minor allele frequencies and generated folded SFS in synonymous and nonsynonymous/missense regions (SYN-SFS and MIS-SFS respectively) for each species (Figure S3.2).

To estimate long-term population size for each species, and control for population history and linked selection in the coding regions for downstream DFE inference (Kim et al. 2017; Tataru and Bataillon 2019), we inferred demographic parameters from the putatively neutral synonymous SFS using the *Demog1D_sizechangeFIM* component in *varDFE* package. Overall, the most parsimonious *two_epoch* demographic model based on unmasked SYN-SFS fit well for most species except for the mosquito (AC136) dataset (Figure S3.3, Table S3.1). For the AC136 dataset based on the unmasked SYN-SFS, the *three_epoch* demographic model improved the fit compared to the *two_epoch* model (*three_epoch* log-likelihood = -1405.30; *two_epoch* log-likelihood = -1684.36) but still did not fit the data well (data to data log-likelihood = -341.32). Therefore, we tried masking the singletons in the SYN-SFS and found that the singletons-masked SYN-SFS provided a dramatic improvement in model fit (Δlog-likelihood = -78 for masked *three_epoch* model compared to Δlog-likelihood = -1063.98 for the unmasked *three_epoch* model). All the inferred demographic models are qualitatively consistent with previous estimates (Table S3.2). We utilized the best-fit demographic parameters from the *three_epoch* masked model for the AC136 dataset and the *two_epoch* unmasked model for other datasets (Table 3.1) in the downstream analyses of the DFE reported below.

The eight species we evaluated are distinctive in the potential biological features that affect the DFE, such as phylogenetic positions, effective population sizes and life history strategies (Table 3.1). The divergence time from human ranged from 900 million years ago (Mya) for insects (Peterson et al. 2004) to 40.7 Mya for mice (Kumar and Hedges 1998). The long-term population sizes (N_a) varied from 6.2E+03 (vaquitas) to 2.77E+06 (*drosophila*). The generation time varied from 0.09 year per generation in mosquitos (Miles et al. 2017) to 25.9 years in fin whales (Taylor et al. 2007).

Mutations are more deleterious in mammals compared with insects

Conditional on the inferred demography, we estimated the DFE for new nonsynonymous mutations for each species. We assumed that the DFE follows a gamma distribution and mutations are neutral or deleterious ($s < 0$). Given that more deleterious mutations segregate in lower frequencies and lower numbers compared with neutral mutations, we fitted the gamma DFE to the observed differences between the MIS-SFS and the SYN-SFS. Gamma distributions provided a good fit to all MIS-SFS (Table 3.2; Figure S3.4).

On average, mutations are 17 to ~5000 times more deleterious in mammals (n = 5; E[|s|] = 1.02E-02 in mice to 7.14E-01 in wolves), compared to insects (n = 2; E[|s|] = 1.38E-04 in *drosophila*, 5.92E-04 in mosquitos; Figure 3.1A). The average mutation effects for pied flycatchers, the only bird species included, is similar to insects (E[|s|] = 5.21E-04). Noticeably, the scale (*β'*) parameter for wolves reached the upper boundary during inference for unknown reasons (Table S3.3), an observation also found in seals (Gaughran 2021). When excluding E[|s|] estimates from only the wolf dataset, mutations are at most ~200 times more deleterious in fin whales compared to *drosophila*.

When evaluating the shape (α) parameter in the gamma distribution independently, there are differences between species groups as well. On average, mammals have lower shape values (n = 5; α = 0.11 in wolves to α = 0.21 in mice) compared to insects (n = 2; α = 0.29 in mosquitos to α = 0.36 in *drosophila*). The shape parameter in the flycatcher is similar to insects (α = 0.29). The variations in α found in this study are in line with the previous evaluations (Chen et al. 2017).

We also observed that mammals have much higher proportions of strongly deleterious mutations compared to *drosophila,* mosquitos and flycatchers. The maximum-likelihood gamma distribution (Figure 3.1B) for each species demonstrated that 22.4% (vaquita) to 47.4% (wolves) of mutations in mammals are very strongly deleterious ($|s| > 10^{-2}$), compared to 0.00% mutations in *drosophila,* 0.07% in mosquitos and 0.03% in flycatchers (Table S3.4). On the other hand, the proportions of weakly deleterious mutations ($10^{-5} \leq |s| < 10^{-3}$) are more abundant in *drosophila,* mosquitos and flycatchers (68.37%, 58.09%, 58.72%, respectively), compared to mammals (15.99% in wolves to 29.78% in mice). The proportions of neutral to nearly neutral mutations ($0 \leq |s| < 10^{-5}$) are similar in all eight species (Figure 3.1B).

To further test that inferred differences in the DFE are not due to statistical uncertainty as a result of having limited data, we compared the DFE estimates for each species to a null model where the DFE is constrained to be the same across species (Figure 3.1C). The likelihood ratio test (LRT) demonstrated that the model where each species has its own shape and scale parameters from the gamma distribution fits the data (MIS-SFS) significantly better than the null model (LRT statistics $\Lambda = 14093.69$, df = 14, $P < 10^{-16}$; Figures 3.1C – 3.1E, S3.5). Therefore, the variation of DFE across taxa is statistically supported in our results after controlling for demography in each species, consistent with previous estimates (Chen et al. 2017; Huber et al. 2017). To rule out phylogenetic dependency (Felsenstein 1985), we calculated pairwise LRT statistics for all species pairs. The species pairs with lower LRT statistics, in other words, that are more likely to share the same DFE, are phylogenetically more closely related (Figure 3.1F). A hierarchical clustering of pairwise LRT statistics largely recovered the phylogenetic relationships of the eight species as well, with mammals and insects clustered into two distinct lineages by their LRT statistics (Figure 3.1F). In summary, phylogenetically close species have a similar DFE, with mammals harboring more strongly deleterious and less weakly deleterious variations compared with *drosophila,* mosquitos and flycatchers.

Robustness of DFE inference

To evaluate the robustness of our inference, we sought to test if different functional forms of DFE affect the patterns observed above. We used two additional commonly assumed distribution functions (Boyko et al. 2008; Kim et al. 2017): 1) a mixture of gamma distribution with point mass at neutrality (neugamma); and 2) a log-normal distribution (lognormal). Both neugamma and lognormal distributions fit the data similarly well as the gamma distribution (Table 3.2; Figures S3.6, S3.7). Comparing the three function forms (gamma, neugamma and lognormal) within each dataset, the neugamma distribution had the highest maximum likelihood in five datasets (DM100, FH18, BP44, MM16, HS100). Lognormal distribution had the highest maximum likelihood in the other three datasets (AC136, CL26, PS24). However, the lognormal distribution fit the *drosophila* (ΔLL = -217.82), flycatchers (ΔLL = -28.6) and humans (ΔLL = -32.09) datasets much worse than gamma or neugamma distributions. Given that the gamma distribution had less parameters than the neugamma distribution and performed more stably across datasets compared with the lognormal distribution by achieving the second highest maximum likelihood in all datasets except for wolves (Table S3.3), we confirmed that the gamma distribution is a good candidate function form for DFE comparisons.

Overall, average mutation effects remained more deleterious in mammals compared with insects and birds (Figures S3.8, S3.9). Mutation effects for vaquitas (PS24) are impacted qualitatively by the DFE function assumed. However, only when assuming the DFE follows a lognormal distribution, does the average mutation effects for vaquitas become close to that of mosquitos (Table 3.2; Figure S3.9). The other mammals, however, still harbor more deleterious variation than insects, regardless of the functional form of the DFE assumed.

We also evaluated the potential impacts of demographic model misspecification, given that only the mosquito dataset utilized the *three_epoch* model from SFS with singletons removed. Assuming the DFE follows a gamma distribution, we repeated the DFE inference for the mosquito dataset using the *two_epoch* demographic model with full SFS and for the other seven datasets, using the *three_epoch* model with singleton-masked SFS (Table S3.5). The average mutation effects and proportion of each mutation categories remain unchanged qualitatively (Figures S3.10, S3.11). When assuming *two_epoch*, *full SFS* model for all datasets, the average mutation effects (E[|s|]) in mosquitos reduced from 5.92E-04 (Figure 3.1A) to 8.60E-05 (Figure S3.10).

Overall, our examinations on different functional forms of DFE and demographic model specifications further validated the robustness of previous observations in this study. Consistently, mutations are more deleterious in the mammal datasets compared to insect and bird datasets.

Multiple biological candidate features are correlated with the DFE

Having established the robustness of our inference, together with our preliminary observations that mutations are more deleterious in mammals compared with insects and birds (Figure 3.1A), we tested candidate features that could be correlated with DFE variation. Given that parameters for the wolf dataset reached the preset upper boundary with uncertain biological significance (Figure 3.1A, 3.1B; Table S3.3), we excluded the wolf dataset in this analysis to rule out potential inference artifacts. As the long-term population size (N_a) decreased (Figure 3.2A), or the divergence time from human lineage decreased (Figure 3.2C), on average, mutations become more deleterious in a population (E[|s|]). On the other hand, the generation time (Figure 3.2B) and mutation rates (Figure 3.2D) are positively correlated with E[|s|]. All candidate features were significantly correlated with E[|s|] (P < 0.05), with the strongest correlations observed with long-term population size (R^2_{adj} = 0.72, P = 0.0094), and mutation rates (R^2_{adj}= 0.72, P = 0.0097).

We assumed that mammals are more complex than insects, in other words, having more phenotypes under selection, because of their larger genomes, larger number of genes and more protein-protein interactions (Huber et al. 2017). Therefore, the divergence times from human, generation times and mutation rates tested above could be reflective of an organism's complexity given their correlations with an organism's phylogenetics placement. Our results are in support of the Fisher's Geometric Model, which predicts that more complex organisms will have more deleterious variations. In contrast, the mutation robustness model, which predicts the opposite is not supported. However, the candidate features examined for organism complexities and the long-term population size are often correlated (Figure S3.12). Therefore, the more deleterious mutations found in mammals might also be caused from the usually low population sizes in mammals alone, as proposed by the protein stability model.

Population size is not the sole predictor of fitness effects

To distinguish the predictions from the protein stability model and the FGM, we evaluated the population-scaled mutation effects $\gamma = 2N_a s$. The protein stability model predicts that although the individual mutation effect s varies, the population-scaled mutation effect $2N_a s$ is the same across species, while the FGM does not predict this constancy. Assuming the population-scaled DFE also follows a gamma distribution, we found that average mutation effects are variable across species (Figure 3.3A). The LRT confirmed that the models where each species has its own scaled shape and scale parameters (Figures 3.3D, 3.3E, S3.13, S3.14) fit the data significantly better than if assuming a constant γ in different species (LRT statistics Λ =

59698.1, df = 14, P < 10^{-16}; Figure 3.3C). A hierarchical clustering of pairwise LRT statistics again recovered the phylogenetic relationships of the eight species, with insects forming a distinct DFE group from the mammals and the flycatcher (Figure S3.14). Interestingly, when examining the maximum-likelihood estimates, we no longer observe similar expected mutation effects (E[$|2N_a s|$]) within lineages. The average population-scaled mutation effects (E[$|2N_a s|$]) are the highest in wolves (1.06E+05), mice (4.22E+03) and mosquitos (2.70E+03). The proportion of each type of mutations fluctuated for each species as well (Figure 3.3B; Table S3.6). Therefore, although population size is strongly correlated with fitness effects (Figure 3.2), it is not the sole predictor for the DFE (Figure 3.3).

Fisher's Geometric Model explains the DFE variation in eight animal species

Since population size alone does not account for all the variation in DFE across taxa, we evaluated the Fisher's Geometric Model in more detail. We already found that biological candidate features associated with organism complexities are positively correlated with the average deleteriousness of mutations in diverse species, which agrees with the FGM predictions (Figure 3.2). To directly measure the complexity, we implemented an FGM-based DFE for populations at mutation-selection-drift equilibrium (Lourenço et al. 2011, eq.15). When the population is under mutation-selection-drift balance, the increase in fitness from beneficial mutations should counteract the drift load, i.e. the decrease in fitness caused from fixed deleterious mutations, and the population has an equilibrium phenotypic distance (z_{eq}) to the fitness optimum. Therefore, this Lourenço DFE relaxed the simplifying assumption made in the gamma DFE, where $z_{eq} = 0$ (Martin and Lenormand 2006), and considered beneficial mutations' impacts. This DFE incorporates long-term population size (N_a), mutation pleiotropy (m) and scale of mutation effects (σ) as parameters to estimate (see methods for details). The mutation pleiotropy (m) describes the number of phenotypic traits affected by a single mutation and the scale (σ) describes the net size of a mutation's phenotypic effects (Lourenço et al. 2011). Both m (Lourenço et al. 2011) and σ (Martin and Lenormand 2006; Huber et al. 2017) could be a direct measurement of organism complexity from different perspectives.

The Lourenço DFE provides an equally good or better fit to the gamma-distribution DFE (Figure S3.15; Table S3.3). In addition, the inferred long-term population size from the Lourenço DFE is largely in agreement with the N_a derived from demographic inference (R^2_{adj}= 0.45, P = 0.042; Figure 3.4A). Both suggested that the populations are likely under mutation-selection-drift equilibrium and the FGM-based Lourenço DFE describes their mutation effects well. However, in the flycatcher dataset, the inferred Lourenço DFE's cumulative probability integrated to 1.27 but not one, suggesting that this DFE function is less accurate for larger m, as pointed out in Lourenço et al. 2011.

We found that mutation effect scale (σ) is reflective of species phylogenetic position, with mammals having larger σ compared to insects and birds (Figure 3.4B). The mutation pleiotropy (m), however, is not conserved within lineages and tends to increase as σ decreases, especially within mammals (Figure 3.4B). The FGM also predicts that as the population size (N_a) decreases, the proportion of beneficial mutations increases to counteract the increased drift load (Lourenço et al. 2011). Smaller populations were found to harbor more beneficial mutations (R^2_{adj}= 0.74, P = 0.0036; Figure 3.4C) in the Lourenço DFE, confirming this additional FGM prediction. Overall, these observations further supported that the Fisher's Geometric Model is the best model to explain the DFE variations we observed.

Discussion

To our knowledge, our study is one of the first to evaluate the long-standing question on the evolutionary stability of the DFE across species under a model testing framework. The DFE estimates we obtained are consistent with previous studies. We reproduced prior DFE inference derived from the same data sets (humans and *drosophila*: Huber et al. 2017; mice: Zhen et al. 2021; vaquita: Robinson et al. 2022; Table S3.3). For some species, DFE estimates using different datasets are available, and our inference is similar despite variations in methods and populations analyzed. In humans, a shape parameter of 0.19 for a gamma DFE inferred in this study (Table 3.2) fall into the range of previous studies: from 0.12 ~ 0.16 (Chen et al. 2017), 0.16 (Castellano et al. 2019), 0.18 ~ 0.21 (Boyko et al. 2008), 0.17 ~ 0.21 (Kim et al. 2017) to 0.2 (Keightley and Eyre-Walker 2007). In *drosophila*, our inferred shape parameter of 0.36 is consistent with previous reports using different datasets: from 0.35 (Keightley and Eyre-Walker 2007) to 0.32 ~ 0.41 (Chen et al. 2017). Even in non-model organisms, the DFE inferred for a Sweden pied flycatcher population in this study is comparable with a previously inferred scaled DFE for an Italian collared flycatcher (*F. albicollis*) population. We inferred 62.78%, 17.8% and 19.3% mutations with $|N_a s| > 10$, $1 < |N_a s| \leq 10$ and $|N_a s| \leq 1$, respectively (Table S3.3), similar to the 77.8%, 8.9% and 13.3% values inferred in Bolívar et al. (2018). Similar shape parameters were obtained for mosquitos (0.32 in Chen et al. 2017, 0.29 in this study). The consistency of DFE estimates with literature further confirmed the DFE variation observed across species is likely of biological significance.

Inference of the DFE is challenging and we made several simplifying assumptions. First, all mutations are assumed to be additive ($h = 0.5$) in current DFE inference approaches, such as DFE-alpha (Keightley and Eyre-Walker 2007), Fit∂a∂i (Kim et al. 2017) or polyDFE (Tataru et al. 2017; Tataru and Bataillon 2019). However, dominance effects are more complex in wild populations (Huber et al. 2018) and are often correlated with a mutation's selection coefficients (Fuller et al. 2019). Incorrect assumptions on dominance likely bias DFE parameter estimates (Huang et al. 2021; Wade et al. 2022). New methods that jointly infer *h* and *s* through SFS and linkage disequilibrium (Ragsdale 2022) or family trios (Barroso and Lohmueller 2021) are promising in this regard. We also restricted the DFE inference to nonsynonymous SNP mutations in coding regions. Variation of the DFE across species in other mutation types such as insertion/deletions (Barton and Zeng 2018) and non-coding regions (Bergman and Kreitman 2001; Kousathanas et al. 2011) remains to be explored. In addition, GC-biased gene conversions (gBGC) likely impact inference of the DFE (Bolívar et al. 2018). Since these biases of dominance, mutation types and gBGC are likely impacting all the species we included similarly, the comparison on mutation effects is likely unbiased (Huang et al. 2021). Second, several parameters required for DFE inference, especially mutation rates (μ) and nonsynonymous to synonymous mutation rate ratio (r_L), are challenging to estimate by themselves (Figure S3.1). In wild populations where only phylogenetic based mutation rates estimates are available, this uncertainty can be large. For example, in baleen whales, published mutation rates ranged from 5.70E-09/bp/gen (Dornburg et al. 2012) to 3.31E-08 (Tollis et al. 2019). This uncertainty is variable across species and further investigation is warranted. Lastly, some biological features that were associated with DFE variation (Figure 3.2), namely the mutation rates and long-term population sizes, were inevitably used to parameterize the DFE inference. Independent proxies for complexity such as the number of unique cell types from growing cell atlas efforts (Quake 2022) could be adopted for future studies.

Despite the challenges, we propose that the DFE could be constrained by phylogeny with more closely related species having more similar DFE, by demonstrating that mutations are on average more deleterious in mammals compared with the two insects and one bird species analyzed (Figure 3.1). It is intuitive to assume DFE correlations in related species (Chen et al. 2017), and several previous studies supported this. High levels of DFE correlation had been reported between populations within humans, *drosophila* and wild tomato (Huang et al. 2021), between species within lineages, such as in great apes (Castellano et al. 2019) and cotton wood (Liu et al. 2022), although species-specific patterns are also evident. In addition, stark DFE differences have been observed between species with early divergences, namely humans and *drosophila* (Keightley and Eyre-Walker 2007; Huber et al. 2017). Experiments in *E. coli* showed that the DFE is largely unchanged during 50,000 generations of evolution (Limdi et al. 2022), further supporting our hypothesis that the evolutionary stability of the DFE might have been maintained at a larger time-scale. Evaluations of the DFE across diverse groups of animals had been conducted, and among-taxa variations were reported (Chen et al. 2017; Galtier and Rousselle 2020), similar to our findings. Chen et al. (2017) acknowledged the differences in DFE, whereas the uncertainties in the average deleterious effects (E[|s|]) prevented their further investigations. However, Galtier and Rousselle (2020) considered this variation a methodological artifact. In this study, we compiled high-quality whole-genome based polymorphism data with a larger sample size per species, and our results suggest that the DFE may be varying on a phylogenetic timescale, and stable within lower taxonomic levels.

Our findings on the DFE variation and its correlations with phylogeny, population sizes and organism complexities have several theoretical and practical implications. In evolutionary genetics, quantifying genetic variation, mutational load and the proportion of adaptive evolution in diverse species are of tremendous interest (Galtier and Rousselle 2020; Zhen et al. 2021). We emphasize that the Fisher's Geometric Model remains a powerful framework in evolution. With expanding population genomics data sets, more parameters and predictions from the FGM will become quantifiable. In conservation biology, genomics-informed simulations are increasingly popular to estimate the genetic health of small and fragmented populations (Kyriazis et al. 2021; Beichman et al. 2022; Robinson et al. 2022). The relative conservation of the DFE parameters within mammals suggest that it is reasonable to assume a human-like DFE parameter in other mammalian species when the study species' DFE parameter has not been inferred. In summary, our study provides new insights into the long-standing question concerning the evolutionary stability of DFE across species.

Methods

Data sets and initial processing

We compiled high-quality population-level polymorphism datasets from whole genome resequencing projects for eight animal species using the following criteria (Figure S3.1): 1) the average sequencing depth is at least 15x and genotypes should be called using GATK Best Practice Pipeline; 2) for each species, one natural population with no detectable population substructure or admixture is included; 3) for each population, at least eight diploid, un-related individuals are included; 4) for each dataset, only monomorphic or biallelic SNP sites are included. The datasets for *drosophila* (Huber et al. 2017), vaquitas (Robinson et al. 2021), mice (Zhen et al. 2021) and humans (Huber et al. 2017), had been quality controlled previously using similar methods described below and not described here.

For mosquitos (*Anopheles gambiae coluzzii*), we queried the Ag1000G phase 3 (Ag3) release using the python package malariagen-data (v. 0.15.0) on 2021-12-22 (Miles et al. 2017;

Consortium et al. 2020) with custom python scripts. We limited our search to a single-species *An. coluzzii* sample set (AG1000G-AO; n = 81) which were collected from breeding sites in Luanda, Angola in 2009 (Troco 2012). We first obtained coordinates of canonical transcripts in autosomes (2R, 2L, 3R, 3L regions) by identifying the longest transcript of each gene. We obtained single nucleotide polymorphism (SNP) calls using the *ag3.snp_calls* method and filtered the sites with the following criteria: 1) the sites should pass *variant_filter_pass_gamb_colu* filters provided by the Ag3 release; 2) the sites should fall into canonical CDS regions identified before; 3) the sites should not be in soft-masked or unknown regions of the genome (i.e. *a/t/c/g/N* bases in the genome sequence); 4) the sites should be invariant or biallelic. Variants were annotated using the canonical transcripts' annotations provided in *ag3.snp_effects*. The most deleterious annotation is retained for each variant. Only SYNONYMOUS_CODING (SYN) and NON_SYNONYMOUS_CODING (MIS) variants are downloaded as variant call format (VCF) files. We performed principal component analyses (PCA; function *snpgdsPCA*) and kinship analyses (function *snpgdsIBDKING*) for the LD-pruned SYN-VCF using the R package SNPRelate v.1.16.0 (Zheng et al. 2012) to evaluate population structure as described in Nigenda-Morales et al. (2022). In total, 69 individuals (n = 69) SNP-only SYN-VCF and MIS-VCF were retained for downstream site frequency spectrum (SFS) projection, and 12 individuals were excluded due to high kinship (n = 9; *kinship > 0.15*) or outliers in PCA analyses [n = 3; outliers defined as *abs(x - median(x)) > (6 * sd(x))* in PCA].

For flycatchers (*Ficedula hypoleuca*), we downloaded filtered SNP-only VCF files for four flycatcher species from dryad on 2022-03-24 (Chase et al. 2021a). This dataset derived from a study assessing the genomic differentiation landscapes in *Ficedula* species pairs (Chase et al. 2021b). In this dataset, hard filter thresholds, repeat region masks and genotype filters (minDP =

5, maxDP = 200, minGQ=30) had been applied (Chase et al. 2021a). We lifted the scaffold names and positions (*CHR* and *POS* fields) from the provided VCF file to the NCBI RefSeq chromosome coordinates using custom python scripts (FicAlb1.5; https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/247/815/GCF_000247815.1_FicAlb1.5/GCF_000247815.1_FicAlb1.5_assembly_structure/, accessed 2022-03-26). The lifted REF and ALT bases were verified using *BioPython* and unconcordant sites were removed (0.01%). We validated the population structure by performing PCA and kinship analyses described above using all LD-pruned SNPs. A custom snpEff (v. 5.1) database for FicAlb1.5 was built from the gtf file using default settings. We annotated and predicted the effects of variants using default options in snpEff. The most deleterious annotation is retained for each variant. Only synonymous_variant (SYN) and missense_variant (MIS) variants in assembled autosomes (chr1 to chr28) are retained. We additionally masked sites (13.2%) that fall into repeat regions identified by WindowMasker (soft-masked bases in RefSeq genome; Morgulis et al. 2006) and CpG islands identified by UCSC genome browser (Gardiner-Garden and Frommer 1987). In total, SNP-only SYN-VCF and MIS-VCF for nine individuals (n = 9) from the pied flycatcher population in mainland Sweden (Nadachowska-Brzyska et al. 2016) were retained for downstream SFS projection and no individual was excluded.

For wolves (*Canis lupus*), we previously obtained VCF files from the arctic wolf population (n = 15) in Canada (Phung et al. 2019; Robinson et al. 2019). Only sites passed variant quality filtrations and did not fall in repeat regions or CpG islands were retained for this study. Variants were annotated using snpEff v.4.3.1 (Cingolani et al. 2012), based on the dog reference genome annotation build CanFam3.1.75 available with snpEff installation. We considered sites where all three potential SNPs were annotated as either synonymous_variant

(SYN) or missense_variant (MIS) exclusively, and discarded sites with additional types of annotations (e.g., splicing sites and protein truncating variants, such as stop-gained variants). We estimated kinship and excluded relatives up to 4th degree (e.g. first cousins) using the KING-robust estimator (Manichaikul et al. 2010) implemented in PLINK (Purcell et al. 2007). In total, SNP-only SYN-VCF and MIS-VCF for 14 individuals (n = 14) were retained for downstream SFS projection, and one individual (n = 1) was excluded due to high kinship.

For fin whales (*Balaenoptera physalus*), we previously obtained VCF files from a historically stable population (n = 30) in the Eastern North Pacifics (Nigenda-Morales et al. 2022). Only sites passed variant quality filtrations and did not fall in repeat regions or CpG islands were retained for this study. Variants were annotated using SIFT4G v.6.0 (Vaser et al. 2016) and snpEff v.4.3.1 (Cingolani et al. 2012). Concordance of variant annotations were confirmed across the two software, and the most deleterious annotation derived from SIFT4G is used per SNP. We performed PCA and kinship analyses using all LD-pruned SNPs to evaluate population structure (Nigenda-Morales et al. 2022). In total, biallelic SNP-only SYN-VCF and MIS-VCF for 27 individuals (n = 27) were retained for downstream SFS projection, and three individuals were excluded due to admixture with another population (n = 2) or low genotyping rate (n = 1).

Calculation of Site Frequency Spectra and sequence lengths

We summarized polymorphism data using synonymous and nonsynonymous/missense site frequency spectra (SYN-SFS and MIS-SFS, respectively). Here we describe the SFS generation procedures for the mosquitos, flycatchers, wolves and fin whales. SFS generation for humans (Huber et al. 2017), *drosophila* (Huber et al. 2017), mice (Zhen et al. 2021) and vaquitas (Robinson et al. 2021) utilized similar methods as described below. For mosquitos, flycatchers, wolves and fin whales, each SNP-only SYN-VCF and MIS-VCF file that passed the quality control steps described above, were additionally filtered using the following criteria: 1) sites with more than 20% missing genotypes are removed; 2) sites with more than 75% heterozygous genotypes are removed using GATK (v.3.8) *SelectVariants* (McKenna et al. 2010) or bcftools (v.1.9) *filter* functions (Li 2011).

We projected down the sample size and computed folded SYN-SFS and MIS-SFS from SYN-VCF and MIS-VCF for each species using a modified easySFS module (https://github.com/isaacovercast/easySFS) and *make_data_dict_vcf, from_data_dict* methods in ∂a∂i (v.2.1.1) package (Gutenkunst et al. 2009). Computing a folded SFS avoids uncertainties in ancestral state classifications. We projected each SFS to the sample size that maximize the number of SNPs available to account for the sporadic missing genotypes (Table 3.1).

To calculate the total synonymous and nonsynonymous sequence lengths (L_{SYN} and L_{MIS}), we first obtained the length of coding regions (L_{CDS}) for each species. For mosquitos, fin whales and wolves, variants had been called for all sites in the genome. We intersected the coordinates of coding regions with the sites that passed filters and had allele counts (*INFO/AC*) no less than the projected SFS sample size in each species' all-sites VCF (e.g. *INFO/AC* ≥ *136* for mosquitos). For flycatchers, all-sites VCF is not available, we calculated an approximated L_{CDS} (11.8Mb) by multiplying the total CDS length (24.6Mb) from gtf file, with the proportion of callable sites (76.8%) reported for the autosomes (Table S1 in Chase et al. 2021b), and the proportion of SNPs (62.7%) passed the additional filters applied during dataset QC. For all species, we then calculated the synonymous and nonsynonymous sequence lengths (L_{SYN} and L_{MIS}; $L_{SYN} + L_{MIS} = L_{CDS}$) from previous estimates of nonsynonymous to synonymous mutation rate ratio $r_L = \frac{\theta_{MIS}}{\theta_{SYN}} = \frac{L_{MIS}}{L_{SYN}}$. We used an r_L of 2.31 for vertebrates (L_{MIS} = 2.31* L_{SYN}; $L_{MIS} =$

$\frac{2.31}{2.31+1}L_{CDS}$), and an r_L of 2.85, (L_{MIS}= 2.85 * L_{SYN}; $L_{MIS} = \frac{2.85}{2.85+1}L_{CDS}$) for invertebrates (Huber et al. 2017).

varDFE package components

We developed a robust DFE comparison software, *varDFE*, as an extension for the ∂a∂i and Fit∂a∂i packages. The *varDFE* package can be installed through *pip* in any python environments (version ≥ 3.10). Our workflow comprises of four main components:

1. *Demog1D_sizechangeFIM*: Demographic inference on putatively neutral SFS data.

2. *DFE1D_refspectra*: Precomputations of reference SFS database under selection, given all possible selection coefficients *s* and the inferred demographic scenarios in step one.

3. *DFE1D_inferenceFIM*: DFE inference on selected SFS data, given the expected DFE functional forms and the precomputed reference SFS database in step two.

4. *DFE1D_gridsearch*: DFE variation tests through grid search, given the expected DFE functional forms and the precomputed reference SFS database in step two.

We illustrate the usage of *varDFE* workflow and specific settings used in this study in the sections below.

Demographic inference

We first inferred demographic parameters from the putatively neutral synonymous SFS using the *Demog1D_sizechangeFIM* component in *varDFE* package. For each species, we fitted three models:

1. *snm*: standard neutral model with no population size change. This model estimates ancestral population size (N_a).

2. *two_epoch*: single population model with one size change event. Ratio of the current population size, i.e. the first size change, to ancestral population size (*nua*) and duration of size change (*Ta,* in units of $2*N_a$ generations) are inferred.
3. *three_epoch*: single population model with two size change events. Ratio of the first size change population size, to ancestral population size (*nua*), ratio of the current population size, i.e. the second size change, to ancestral population size (*nub*), and duration of two size change events (*Ta* and *Tb*) are inferred.

To account for the uncertainties in genotype calling methods, we also repeated the demographic inference by masking the singleton entries in the synonymous SFS. In total, six demographic inference runs (three demographic models in SYN-SFS with/without singleton masks) for each species were conducted.

To minimize setting variations across the species we surveyed, we only allowed the parameters starting positions to differ according to prior demographic inference for each species, i.e. humans, *drosophila* and mice: Zhen et al. 2021; mosquitos: Miles et al. 2017; flycatchers: Nadachowska-Brzyska et al. 2016; fin whales: Nigenda-Morales et al. 2022; vaquitas: Robinson et al. 2022. For each run, we run 100 replicates from a permuted starting parameter (fold=1). We set extrapolation grid points as sample sizes plus 5, 15 and 25 (Gutenkunst et al. 2009), maximum iteration as 100, and performed parameter optimization using the *optimize_log* function based on multinomial log-likelihood calculated from the expected SFS in each iteration. We calculated the population-scaled synonymous mutation rate θ_{SYN}, and estimated ancestral population size using $N_{anc} = \frac{\theta_{SYN}}{4\mu L_{SYN}}$, where μ is the exon mutation rates for each species and L_{SYN} is the previously estimated synonymous region sequence length.

The best-fit parameters with the maximum multinomial log-likelihood in the 100 replicates were chosen for each run. We evaluated the convergence of the inference across replicates with different starting values by calculating the difference in log-likelihood in the 20 replicates with the highest log-likelihoods. We estimated the best-fit parameters' uncertainties through Fisher's Information Matrix (FIM, *Godambe.FIM_uncert* function). To evaluate the best-fit parameters in the six runs per species, we plotted each best-fit expected SFS with the observed SYN-SFS using ggplot2 v.3.3.2 (Wickham 2016) in R v.3.6.2 (R Core Team 2019).

For the DFE inference presented in the main text, we chose the *three_epoch* demographic model based on masked SYN-SFS for the mosquito (AC136) dataset because of poor fit for alternative inferences (Figure S3.3; Table S3.2). We chose the *two_epoch* demographic models based on unmasked SYN-SFS for all seven other species because of good fit in this most parsimonious model. This is a simpler model that fits the data well (Figure S3.3; Table S3.2).

DFE inference

Conditional on the inferred demographic scenarios, we estimated the DFE for new nonsynonymous mutations based on the MIS-SFS for each species using methods from Fit∂a∂i. Briefly, our methods take advantage of the pattern that more deleterious mutations segregate at lower frequencies and at lower numbers compared with neutral mutations. Assuming the MIS-SFS is under selection while the SYN-SFS is putatively neutral, the best DFE parameters should fit the differences of MIS-SFS and SYN-SFS in each minor allele frequency bin. Two components in the *varDFE*, *DFE1D_refspectra* and *DFE1D_inferenceFIM*, provide flexible workflow to estimate both the full and deleterious-only DFE with any given demographic and DFE models. To ensure consistency of inference across species, we only allowed species-specific settings including demographic parameters, mutation rates (μ), length of nonsynonymous region (L_{MIS}), population-scaled mutation rates (θ) to vary across species.

We first computed and stored the expected SFS for a range of population-scaled selection coefficients $\gamma = 2N_a s$, given the best fit demographic scenarios using the *DFE1D_refspectra* component in *varDFE* package. Here we calculated the reference spectra given γ but not s, because the SFS reflects the effect of selection at population level, which depends on $\gamma = 2N_a s$ but not s alone. We implemented a slightly modified *Cache1D_mod* module, *Cache1D_mod2*, with a method to integrate over continuous positive gamma space. The range of positive gammas was set from +1e-5 to +100, with 701 points evenly distributed on the logscale, and negative gammas from -10000 to -1e-5, with 901 points evenly distributed on the logscale. The extrapolation points (*pts_l*) were set as [1000,1200,1400]. The reference spectra, from which the expected SFS given any γ can be computed, were cached to save time for recalculations and improve consistency.

To infer the deleterious-only DFE, we assumed that the DFE follows a gamma distribution and performed the inference using the *DFE1D_inferenceFIM* component. We wrote the DFE for each species as $\gamma \sim Gamma(\alpha, \beta)$, where α (α > 0 shape) and β (β > 0; scale) are parameters to be inferred. To avoid finding local maxima, for each species, we ran 100 replicates from a permuted starting parameter (fold=1) of α = 0.2 and β = 4000, informed from prior DFE estimations (Huber et al. 2017). We set parameter upper bounds at $\alpha_{max} = 2.0$ and $\beta_{max} = 1E+6$, lower bounds at $\alpha_{min} = 1E-3$ and $\beta_{min} = 1E-2$, maximum iteration as 100. We calculated population-scaled nonsynonymous mutation rate $\theta_{MIS} = 2.31 \times \theta_{SYN}$ for vertebrates and $\theta_{MIS} = 2.85 \times \theta_{SYN}$ for invertebrates. We performed parameter optimization using the *optimize_log* function based on Poisson log-likelihood calculated from the expected SFS, generated using the *integrate* method from the precomputed reference spectra, in each iteration.

The best-fit parameters with the maximum Poisson log-likelihood in the 100 replicates were chosen for each run. We evaluated replicates convergence across starting values by calculating the difference in log-likelihood in the 20 replicates with the highest log-likelihoods. We estimated the best-fit parameters' uncertainties through Fisher's Information Matrix (FIM, *Godambe.FIM_uncert* function). Recall that we optimized the gamma distribution parameters for the population-scaled selection coefficient γ. To obtain the distribution of s, we unscaled the gamma distribution by $\frac{1}{2N_a}$, therefore, $s \sim Gamma(\alpha, \beta')$, where $\beta' = \frac{\beta}{2N_a}$ and N_a had been inferred from demography estimation.

Average mutation effects E[|s|] for each species were calculated from the best-fit parameters given gamma distribution's property that $\mathrm{E}[|s|] = \alpha\beta'$. The confidence interval for E[|s|] was approximated from the FIM-derived parameter CI, where $E[|s|]_{min} = \alpha_{min}\beta'_{min}$ and $E[|s|]_{max} = \alpha_{max}\beta'_{max}$. To compute the proportion of mutations with different values of s in the maximum-likelihood gamma distribution, we found the cumulative probability for s ranged in [0, 1e-5), [1e-5, 1e-4), [1e-4, 1e-3), [1e-3, 1e-2) and [1e-2, 1] using the *pgamma* function in R.

The average population-scaled mutation effects (E[|$2N_as$|]) were calculated as $\mathrm{E}[|2N_a s|] = \alpha\beta$. The cumulative probability for $2N_as$ ranged in [0,1), [1, 10), [10, 100), [100, 1000) and [1000, Inf] was calculated using the *pgamma* function in R as well.

Likelihood ratio tests to compare s *or* $2N_a$s across species

To test whether the variation of DFE across species was statistically significant at both individual (s) and population ($2N_as$) level, assuming a gamma-distributed DFE, we used the *DFE1D_gridsearch* component in the *varDFE* package.

To test for differences in *s* across species, we evenly spaced the same 250 grid points in biologically meaningful ranges of shape (α = 0.1 ~ 0.5) and scale (β' = 1E-4 ~ 0.5) parameters. We obtained the expected SFS for each α-β' pair for each species using the *integrate* method from reference spectra, and calculated the Poisson log-likelihood relative to empirical MIS-SFS. From 62500 α-β' pairs, we were able to explore the full Poisson log-likelihood surface for each species simultaneously. To obtain the null model where the DFE is constrained to be the same across species, we assumed that each species' SFS are independent of each other given their distant phylogenetic divergence. Therefore, the log-likelihood for the null model can be calculated by summing the log-likelihood at each α-β' pair for each species and the MLE can be found from within the 62500 α-β' parameter grids. To formally test whether the shape (α) and scale (β') are different in any *x* number of species, we adopt the LRT from Huber et al. (2017). The LRT was constructed by subtracting the log-likelihood for the alternate model, where each species is allowed to have their its own DFE parameters ($\widehat{\alpha_1}, \widehat{\alpha_2}, \ldots, \widehat{\alpha_x}, \widehat{\beta_1}, \widehat{\beta_2}, \ldots, \widehat{\beta_x}$; df = *2x*, in our data, 16 parameters to infer) from the log-likelihood for the null model, with two DFE parameters inferred ($\alpha_1 = \alpha_2 = \ldots = \alpha_x$ and $\beta_1 = \beta_2 = \ldots = \beta_x$; df = 2) given the inferred demographic parameters for each species ($\widehat{\Theta_{D,1}}, \widehat{\Theta_{D,2}}, \ldots, \widehat{\Theta_{D,x}}$). Asymptotically, Λ should follow a χ^2 distribution with df = 2x-2 = 14.

$$\Lambda = LogLikelihood\left(\alpha_1 = \widehat{\alpha_2 = \ldots} = \alpha_x, \beta_1 = \widehat{\beta_2 = \ldots} = \beta_x \middle| \widehat{\Theta_{D,1}}, \widehat{\Theta_{D,2}}, \ldots, \widehat{\Theta_{D,x}}\right)$$
$$- LogLikelihood(\widehat{\alpha_1}, \widehat{\alpha_2}, \ldots, \widehat{\alpha_x}, \widehat{\beta_1}, \widehat{\beta_2}, \ldots, \widehat{\beta_x} | \widehat{\Theta_{D,1}}, \widehat{\Theta_{D,2}}, \ldots, \widehat{\Theta_{D,x}})$$

To provide a phylogenetic independent DFE comparison (Felsenstein 1985), we calculated the pairwise LRT in all possible pairs of species. For example, the LRT comparing humans (population 1) and *drosophila* (population 2) DFE can be written as the following.

Asymptotically, Λ should follow a χ^2 distribution with df = 2x-2 = 2. We hierarchically clustered the pairwise LRT statistics using the *hclust* function in R.

$$\Lambda = LogLikelihood\left(\widehat{\alpha_1 = \alpha_2}, \widehat{\beta_1 = \beta_2} \middle| \widehat{\Theta_{D,1}}, \widehat{\Theta_{D,2}}\right) - LogLikelihood(\widehat{\alpha_1}, \widehat{\alpha_{2,}} \widehat{\beta_1}, \widehat{\beta_2} | \widehat{\Theta_{D,1}}, \widehat{\Theta_{D,2}})$$

To test the variations of $\gamma = 2N_a s$ for each species, we conducted the grid search by evenly spacing the same 250 grid points in biologically meaningful ranges of scaled shape (α = 0.1 ~ 0.5) and scaled scale (β = 100 ~ 2.5E+4) parameters for population-scaled DFE, and repeated the overall and pairwise LRT analyses as outlined above.

Robustness of the DFE inference

Functional forms of DFE

To examine the robustness of our DFE inference, we implemented two additional functional forms of the DFE: 1) a mixture distribution with point mass at neutrality combined with the remaining mutations having a gamma-distribution of selection coefficients (neugamma), and 2) a log-normal distribution (lognormal), assuming the best-fit demographic scenarios using the same settings as described above.

For the neugamma distribution, we write the DFE for each species as $\gamma \sim Neugamma(\alpha, \beta, p)$. α (α > 0; shape) and β (β > 0; scale) are parameters from a gamma distribution, and *p* (*p* ≥ 0) is the proportion of neutral mutations ($0 \leq |\gamma| < 10^{-5}$). During inference, we set the parameter start positions, upper bounds and lower bounds the same as the gamma distribution for the α and β parameter. The starting position for *p* is 0.3, the upper bound is p_{max} = 1, the lower bounds is p_{min} = 1E-5.

$$f(\gamma) = \begin{cases} \frac{p}{10^{-5}} + (1-p) * Gamma(\gamma|\alpha,\beta), & 0 \leq |\gamma| < 10^{-5} \\ (1-p) * Gamma(\gamma|\alpha,\beta), & |\gamma| \geq 10^{-5} \end{cases}$$

To obtain the distribution of s, we unscaled the neugamma distribution by $\frac{1}{2N_a}$, therefore, $s{\sim}Neugamma(\alpha, \beta', p)$, where $\beta' = \frac{\beta}{2N_a}$ and N_a had been inferred from demography estimation. Average mutation effects E[|s|] for each species were calculated from the best-fit parameters using $E[|s|] = \frac{p}{2} * 10^{-5} + (1-p)\alpha\beta'$.

For the lognormal distribution, we write the DFE for each species as $\gamma{\sim}\ Lognormal(\mu_s, \sigma^2)$. During inference, we set the starting positions before permutation as $\mu_s = 1, \sigma = 0.1$, upper bounds as $\mu_s = 100, \sigma = 100$, lower bounds as $\mu_s = -100, \sigma =$ 1e-5. Since μ_s could take a negative value, we used the *optimize* instead of *optimize_log* function for parameter optimization. To obtain the distribution of s, we unscaled the lognormal distribution, therefore, $s{\sim}Lognormal(\mu_s', \sigma^2)$, where $\mu_s' = \mu_s - log(2N_a)$ and N_a had been inferred from demography estimation. Median mutation effects E[|s|] for each species were calculated using $E[|s|] = exp(\mu_s')$.

Sensitivity to the assumed demographic model

Since we used different demographic models for different species, we explored whether different models or masking singletons affected inferences. To do this, we calculated the reference spectra for the mosquito dataset using the *two_epoch* model with full SFS, and for the other seven dataset, using the *three_epoch* model with singleton-masked SFS. Assuming a gamma distributed DFE, we repeated the DFE inference process, using the reference spectra for each species either from the *two_epoch* model with full SFS or the *three_epoch* model with singleton-masked SFS.

Investigating mechanisms underlying variation in the DFE across species

To correlate the candidate explanatory variable (*X*) with the observed E[|s|], we performed linear regressions using the lm function in R. The candidate explanatory variables (*X*) included are 1) long term population size (N_a); 2) divergence time in million years from human lineage; 3) generation time in years per generation; 4) mutation rates in mutations per bp per generation. In general, the regression was specified as $log(E[|s|]) \sim log(X)$ to normalize the data. For divergence time, it is $log(E[|s|]) \sim log(X+1)$. Given that parameters for the wolf dataset reached upper boundaries during inference (Figure 3.1A, 3.1B; Table S3.3), we excluded the wolf data point, leaving seven species' DFE for regression. We also evaluated the correlations of N_a with other candidate explanatory variables using the *lm* function with the following formula: $log(X) \sim log(N_a)$.

Fitting the FGM-derived DFE

To further test whether the FGM can explain the variation in the DFE variations and contribution of beneficial mutations to SFS across species, we implemented a functional form of the DFE (Lourenço DFE) introduced in Lourenço et al. 2011 eq. 15, which is directly derived from FGM assumptions. Briefly, the FGM proposes that populations can be seen as a collection of phenotypes (*n*) under stabilizing selection around a local maxima and complexity (*n*) is defined as the total number of phenotypes under selection. Fitness (*w*) can be described as a Gaussian function of the distance to the optimum (*z*), $w(z) = exp(-z^2)$. Random mutations do not impact all phenotypes equally, but will likely affect a subset of *m* phenotypes with fitness effect size *r*. Here, m is defined as mutation pleiotropy (Lourenço et al. 2011) or effective complexity (Martin and Lenormand 2006). Effect size *r* follows a zero mean Guassian distribution with scale σ. When the population is under mutation-selection-drift balance, the increase in fitness from beneficial mutations should counteract the drift load, the decrease in fitness caused from deleterious mutations and the population will have an equilibrium phenotypic distance (z_{eq}) to the fitness optimum. Therefore, the DFE for the well-adapted population (Lourenço et al. 2011; eq.15) is correlated with the mutation pleiotropy (m), scale of mutation effects (σ) and long-term population size (N_a), $\gamma \sim Lourenço(m, \sigma, N_a)$. Here $\Gamma(.)$ is the gamma function and $K(.)$ is the modified Bessel function of the second kind (Lourenço et al. 2011).

$$f(\gamma) = \frac{2^{\frac{1-m}{2}}\sqrt{N_a}(|\gamma|)^{\frac{m-1}{2}}(1+\frac{1}{N_a\sigma^2})exp(-N_a\gamma)}{\sqrt{\pi}\sigma^m\Gamma(\frac{m}{2})} \times K_{\frac{m-1}{2}}(N_a|s|\sqrt{1+\frac{1}{N_a\sigma^2}})$$

We inferred the maximum likelihood estimates for the m, σ and N_a parameters using the *DFE1D_inferenceFIM* component as described above. Since this DFE also includes beneficial mutations, we calculated the expected SFS using the *integrate_continuous_pos* method implemented in *varDFE* package (Figure S3.16).

Tables

Table 3.1

Table 3.1. Polymorphism dataset summary. *Pop,* population. The first two letters denote the species name (e.g. DM), the numbers following are the projected SFS sample size (e.g. 100). T_{gen}, generation time per year. μ, mutation rates. N_a, long-term effective population size. T_{div}, approximated divergence time from humans in million years ago. *Demog Model*, the demographic model used for N_a inference.

Pop	Order	Scientific Name	Common Name	T_{gen}	μ	N_a	T_{div}	Demog Model
AC136	Diptera	*Anopheles coluzzii*	mosquito	0.09	1.50E-09	2.28E+06	900	3epoch, masked SFS
DM100	Diptera	*Drosophila melanogaster*	drosophila	0.10	1.50E-09	2.77E+06	900	2epoch, full SFS
FH18	Passeriformes	*Ficedula hypoleuca*	pied flycatcher	2.00	2.80E-09	2.47E+05	310	2epoch, full SFS
CL26	Carnivora	*Canis lupus*	arctic wolf	3.00	5.63E-09	7.41E+04	74.0	2epoch, full SFS
PS24	Artiodactyla	*Phocoena sinus*	vaquita	11.90	5.83E-09	6.27E+03	58.2	2epoch, full SFS
BP44	Artiodactyla	*Balaenoptera physalus*	fin whale	25.90	2.77E-08	1.24E+04	58.2	2epoch, full SFS
MM16	Rodentia	*Mus musculus*	mouse	0.33	5.40E-09	2.08E+05	40.7	2epoch, full SFS
HS100	Primates	*Homo sapiens*	human	25.00	2.50E-08	7.04E+03	0	2epoch, full SFS

Table 3.2

Table 3.2. Mutation fitness effects are variable across species. Function, the functional form of DFE. ll_model, the Poisson log-likelihood of the best-fit model. ll_data, the Poisson log-likelihood of the data. df, the degree of freedom, i.e., the number of parameters estimated. In *gamma* functional form, Parameter1 = shape, Parameter2 = scale. In *neugamma*, Parameter1 = shape, Parameter2 = scale, Parameter3 = proportion of neutral mutations. In *lognormal* functional form, Parameter1 = mean mutation effects, Parameter2 = variance of mutation effects. E[|s|], the expected mean (*gamma* and *neugamma*) or median (*lognormal*) fitness effects.

	Pop	Function	ll_model	ll_data	df	Parameter1	Parameter2	Parameter3	E[|s|]
1	AC136	gamma	-485.08	-296.16	2	0.2915	0.0020		5.92E-04
1	AC136	neugamma	-486.78	-296.16	3	0.2866	0.0023	0.0002	6.54E-04
1	AC136	lognormal	-429.52	-296.16	2	-7.9322	5.0464		3.59E-04
2	DM100	gamma	-347.03	-224.90	2	0.3570	0.0004		1.38E-04
2	DM100	neugamma	-343.49	-224.90	3	0.3844	0.0003	0.0110	1.21E-04
2	DM100	lognormal	-561.31	-224.90	2	-9.5317	3.9991		7.25E-05
3	FH18	gamma	-69.27	-41.13	2	0.2862	0.0018		5.21E-04
3	FH18	neugamma	-52.44	-41.13	3	1.1955	0.0001	0.1592	1.07E-04
3	FH18	lognormal	-81.04	-41.13	2	-8.7764	4.3027		1.54E-04
4	CL26	gamma	-81.12	-58.94	2	0.1068	6.6909		7.14E-01
4	CL26	neugamma	-72.86	-58.94	3	0.1577	6.7181	0.1454	9.05E-01
4	CL26	lognormal	-71.40	-58.94	2	-1.3662	15.5870		2.55E-01
5	PS24	gamma	-47.70	-43.95	2	0.1334	0.0984		1.31E-02
5	PS24	neugamma	-47.71	-43.95	3	0.1327	0.1014	0.0005	1.34E-02
5	PS24	lognormal	-47.17	-43.95	2	-8.1588	5.2064		2.86E-04
6	BP44	gamma	-123.16	-98.68	2	0.1425	0.1871		2.67E-02
6	BP44	neugamma	-120.04	-98.68	3	0.3126	0.0352	0.2169	8.63E-03
6	BP44	lognormal	-129.70	-98.68	2	-7.0754	6.5251		8.46E-04
7	MM16	gamma	-54.59	-41.96	2	0.2061	0.0493		1.02E-02
7	MM16	neugamma	-50.07	-41.96	3	0.3888	0.0047	0.0853	1.66E-03
7	MM16	lognormal	-59.66	-41.96	2	-5.5925	6.8391		3.73E-03
8	HS100	gamma	-241.88	-183.69	2	0.1896	0.0725		1.38E-02
8	HS100	neugamma	-215.74	-183.69	3	0.6755	0.0092	0.2731	4.53E-03
8	HS100	lognormal	-273.97	-183.69	2	-6.8572	4.8631		1.05E-03

Figures

Figure 3.1

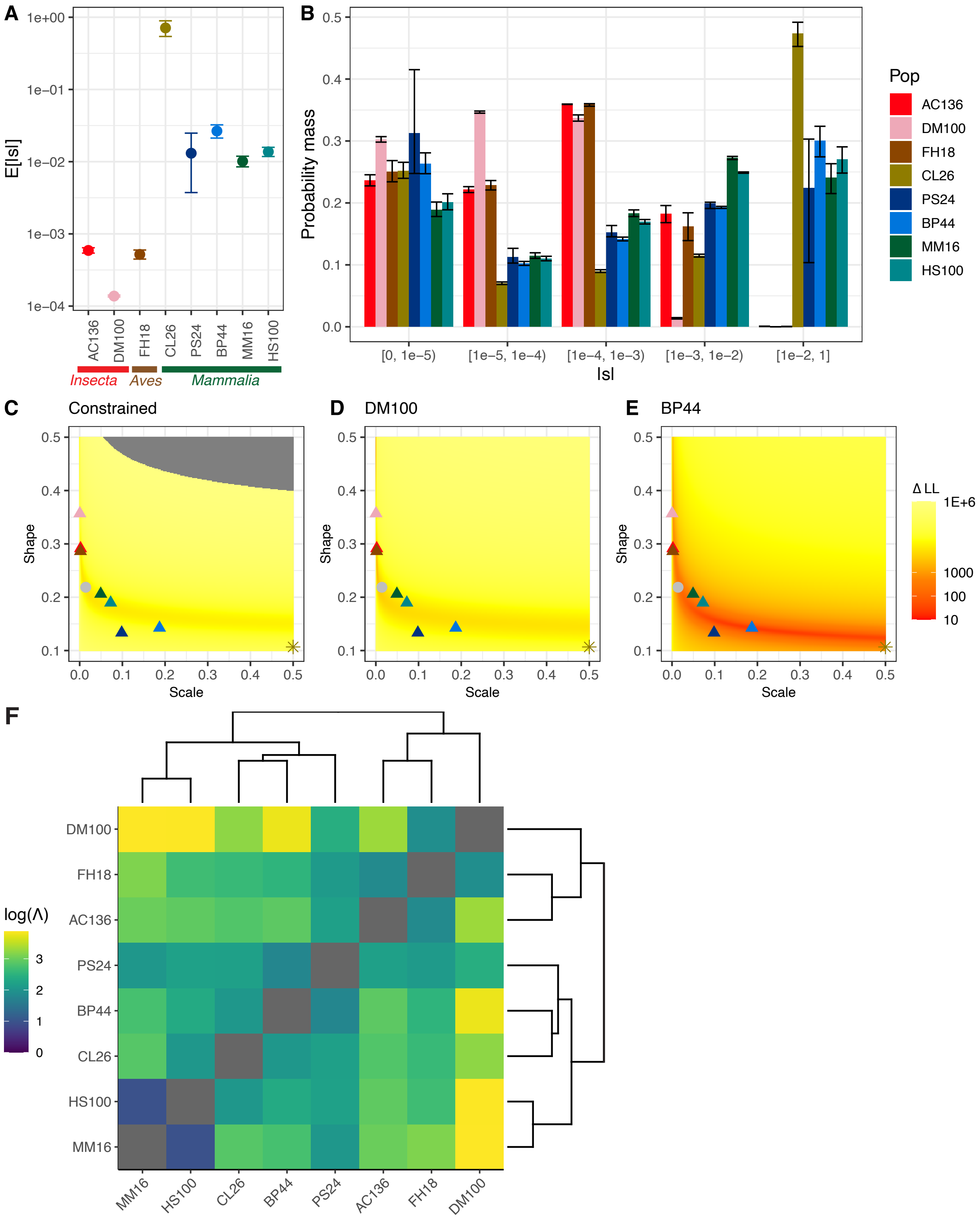

Figure 3.1. Mutations are more deleterious in mammals. Assuming a gamma-distributed DFE, **(A)** the average deleterious mutation effects (E[|s|]) for eight species. From left to right, species are increasingly divergent from humans: mosquitos (AC136, red), *drosophila* (DM100, pink), flycatchers (FH18, dark orange), wolves (CL26, khaki), vaquitas (PS24, navy), fin whales (BP44, blue), mice (MM16, green) and humans (HS100, turquoise). **(B)** Proportions of mutations with various ranges of |s|. From left to right, mutations ranged from (nearly) neutral ($-10^{-5} < s \leq 0$) to very strongly deleterious ($s \leq -10^{-2}$). Error bars represent FIM derived confidence interval. **(C-E)** The log-likelihood surfaces for the shape (α) and scale (*β'*) parameters **(C)** under the constrained model where all species have the same parameters or under alternative models allowing species' parameters to vary in representative **(D)** *drosophila* and **(E)** fin whales. For full log-likelihood surfaces, see Figure S3.5. Background colors from yellow to red indicate the differences of log-likelihood for given parameters to data, with gray backgrounds represent ΔLL > 1E+6. On each log-likelihood surface, the maximum likelihood estimate (MLE) for the null model is overlayed as a gray point. The MLEs for each species are overlayed as triangles using the same color code. The wolf dataset (CL26)'s MLE exceeds the grid search range and is plotted as a khaki asterisk. **(F)** Pairwise LRT statistics (Λ) for each species pair are colored on a log scale and hierarchically clustered. Darker cells represent more similar DFE estimates in the species compared, such as HS100 – MM16, whereas lighter cells represent more variable DFE, such as HS100 – DM100. The dendrogram derived from hierarchical clustering is annotated.

Figure 3.2

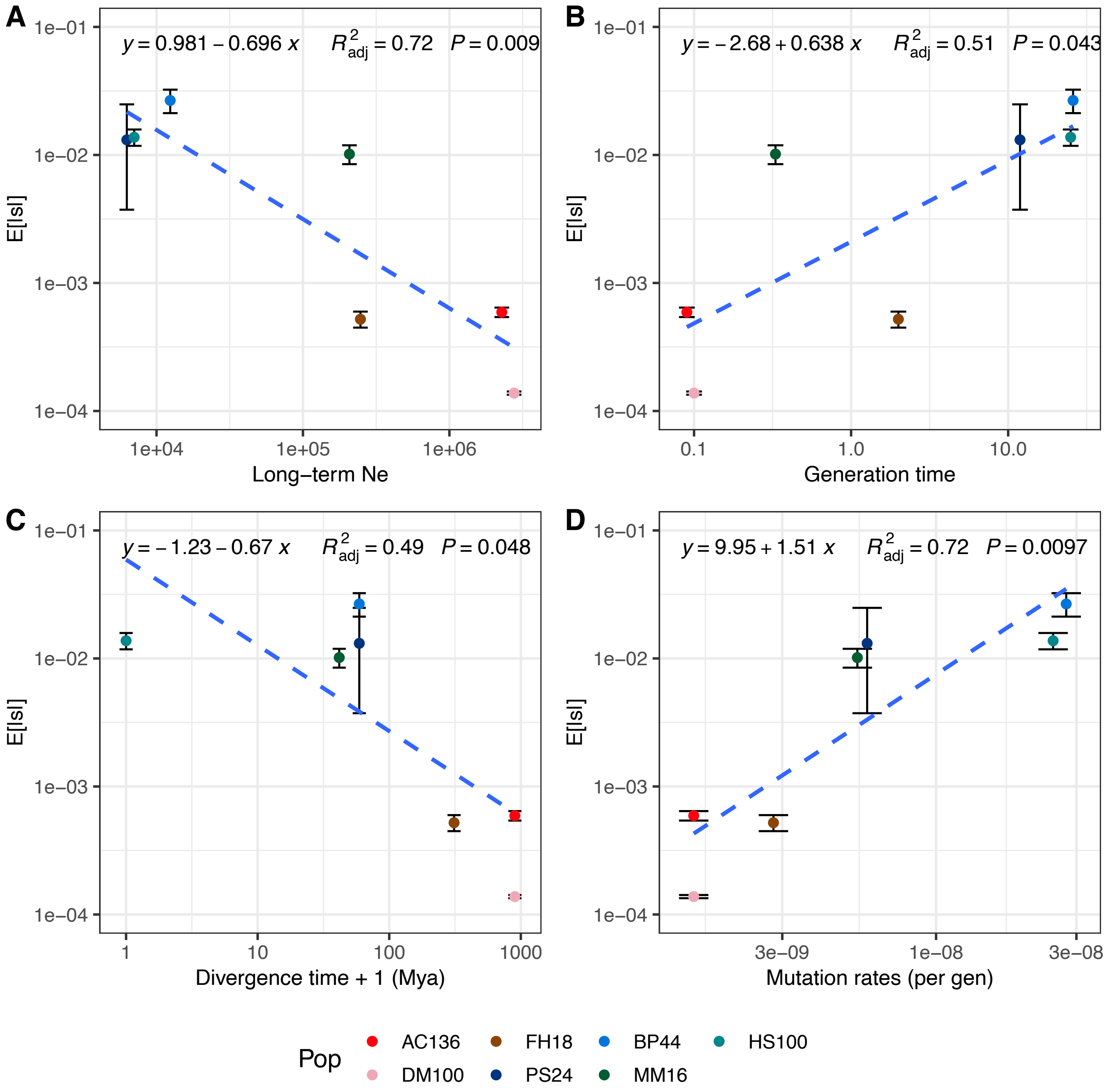

Figure 3.2. Multiple candidate factors are correlated with the expected selection coefficient assuming a gamma distributed DFE. The wolf dataset is excluded because parameters reached upper boundaries during inference. Each point represents one species, with error bars represent confidence interval. The regression result is overlayed as a dashed blue line, with equations, adjusted R^2, and p-value annotated at top. All axes are in log10 scale. **(A)** Long-term population

size is negatively correlated with deleterious mutation effects (E[|s|]). **(B)** Generation time is positively correlated with E[|s|]. **(C)** Divergence time with human lineage is negatively correlated with E[|s|]. **(D)** Mutation rate per bp per generation is positively correlated with E[|s|].

Figure 3.3

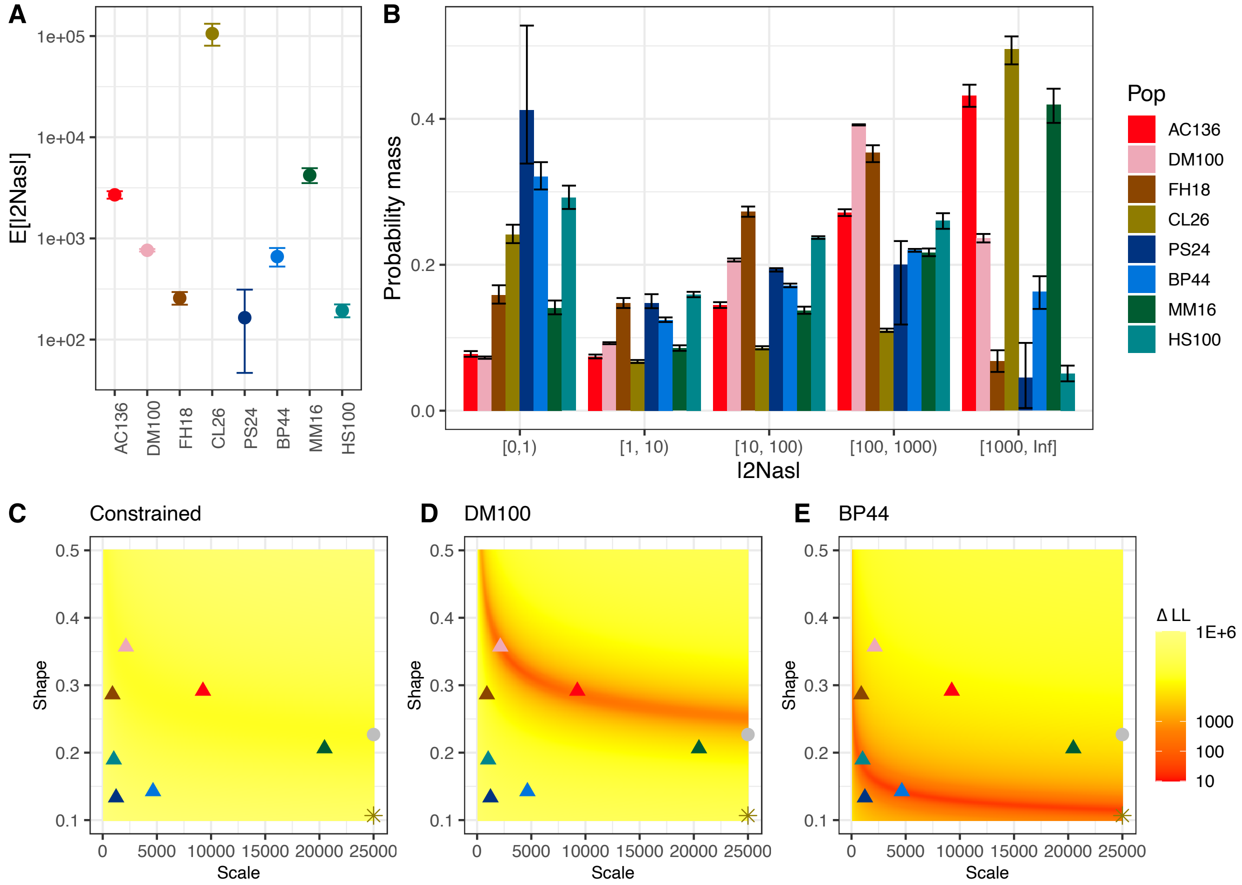

Figure 3.3. Population-scaled mutation effects ($2N_as$) are not conserved across species. Assuming a gamma-distributed DFE**, (A)** the average population-scaled deleterious mutation effects (E[|$2N_as$|]) for eight species. **(B)** Proportions of mutations with various range of |$2N_as$|. From left to right, mutations ranged from (nearly) neutral ($-1 < 2N_as \leq 0$) to very strongly deleterious ($2N_as \leq -1000$). Error bars represent FIM derived confidence interval. **(C-E)** The log-likelihood surface for the population level shape (α) and scale (β) parameters **(C)** under the constrained model where all species have the same parameters or under alternative models allowing species' parameters to vary in representative **(D)** *drosophila* and **(E)** fin whales. For full

log-likelihood surfaces, see Figure S3.13. Background colors from yellow to red indicate the differences of log-likelihood for given parameters to data, with gray backgrounds represent $\Delta LL > 1E+6$. On each log-likelihood surface, the maximum likelihood estimate (MLE) for the null model is overlayed as a gray point. The MLEs for each species are overlayed as triangles using the same color code. The wolf dataset (CL26)'s MLE exceeds the grid search range and is plotted as a khaki asterisk.

Figure 3.4

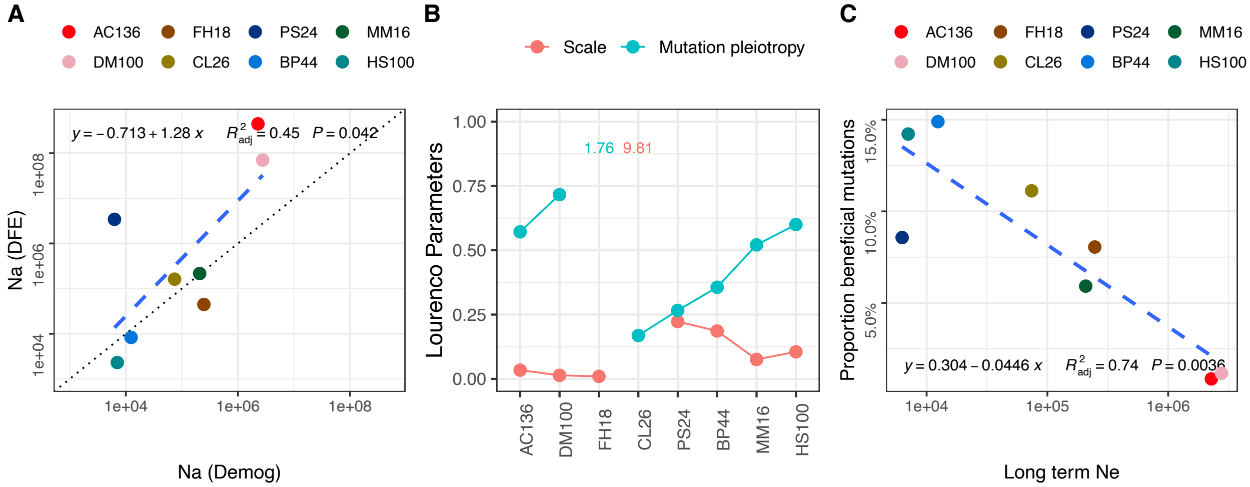

Figure 3.4. Fitting a Fisher's Geometric Model derived DFE. **(A)** The inferred long term effective population size (N_a) from the DFE (y-axis) is correlated with the effective size inferred from genetic variation data using ∂a∂i inference (x-axis). Axes are in log10 scale. **(B)** The estimated scale of mutation effects (σ) and mutation pleiotropy (m) parameters for eight species. The mutation pleiotropy for flycatchers (FH18, 1.76) and scale for wolves (CL26, 9.81) exceeded axes limit and their values are annotated as numbers. **(C)** The estimated proportion of beneficial mutations ($s > 0$) is negatively correlated with long-term population size inferred from ∂a∂i. The x-axis is in log10 scale.

References

Barroso GV, Lohmueller KE. 2021. Inferring the mode and strength of ongoing selection. :2021.10.08.463705. Available from: https://www.biorxiv.org/content/10.1101/2021.10.08.463705v2

Barton HJ, Zeng K. 2018. New Methods for Inferring the Distribution of Fitness Effects for INDELs and SNPs. *Molecular Biology and Evolution* 35:1536–1546.

Beichman AC, Kalhori P, Kyriazis CC, DeVries AA, Nigenda-Morales S, Heckel G, Schramm Y, Moreno-Estrada A, Kennett DJ, Hylkema M, et al. 2022. Genomic analyses reveal range-wide devastation of sea otter populations. *Molecular Ecology*:mec.16334.

Bergman CM, Kreitman M. 2001. Analysis of Conserved Noncoding DNA in Drosophila Reveals Similar Constraints in Intergenic and Intronic Sequences. *Genome Res.* 11:1335–1345.

Bolívar P, Mugal CF, Rossi M, Nater A, Wang M, Dutoit L, Ellegren H. 2018. Biased Inference of Selection Due to GC-Biased Gene Conversion and the Rate of Protein Evolution in Flycatchers When Accounting for It. *Molecular Biology and Evolution* 35:2475–2486.

Bourgeois YXC, Warren BH. 2021. An overview of current population genomics methods for the analysis of whole-genome resequencing data in eukaryotes. *Mol Ecol* 30:6036–6071.

Boyko AR, Williamson SH, Indap AR, Degenhardt JD, Hernandez RD, Lohmueller KE, Adams MD, Schmidt S, Sninsky JJ, Sunyaev SR, et al. 2008. Assessing the Evolutionary Impact of Amino Acid Mutations in the Human Genome. *PLOS Genetics* 4:e1000083.

Castellano D, Macià MC, Tataru P, Bataillon T, Munch K. 2019. Comparison of the Full Distribution of Fitness Effects of New Amino Acid Mutations Across Great Apes. *Genetics* 213:953–966.

Chase MA, Ellegren H, Mugal CF. 2021a. Positive selection plays a major role in shaping signatures of differentiation across the genomic landscape of two independent Ficedula flycatcher species pairs. :69099845542 bytes. Available from: http://datadryad.org/stash/dataset/doi:10.5061/dryad.n2z34tmw6

Chase MA, Ellegren H, Mugal CF. 2021b. Positive selection plays a major role in shaping signatures of differentiation across the genomic landscape of two independent Ficedula flycatcher species pairs. *Evolution* 75:2179–2196.

Chen J, Bataillon T, Glémin S, Lascoux M. 2022. What does the distribution of fitness effects of new mutations reflect? Insights from plants. *New Phytologist* 233:1613–1619.

Chen J, Glémin S, Lascoux M. 2017. Genetic Diversity and the Efficacy of Purifying Selection across Plant and Animal Species. *Molecular Biology and Evolution* 34:1417–1428.

Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly* 6:80–92.

Consortium TA gambiae 1000 G, Clarkson CS, Miles A, Harding NJ, Lucas ER, Battey CJ, Amaya-Romero JE, Kern AD, Fontaine MC, Donnelly MJ, et al. 2020. Genome variation and population structure among 1142 mosquitoes of the African malaria vector species Anopheles gambiae and Anopheles coluzzii. *Genome Res.* 30:1533–1546.

Dornburg A, Brandley MC, McGowen MR, Near TJ. 2012. Relaxed Clocks and Inferences of Heterogeneous Patterns of Nucleotide Substitution and Divergence Time Estimates across Whales and Dolphins (Mammalia: Cetacea). *Molecular Biology and Evolution* 29:721–736.

Eyre-Walker A, Keightley PD. 2007. The distribution of fitness effects of new mutations. *Nat Rev Genet* 8:610–618.

Eyre-Walker A, Keightley PD. 2009. Estimating the Rate of Adaptive Molecular Evolution in the Presence of Slightly Deleterious Mutations and Population Size Change. *Molecular Biology and Evolution* 26:2097–2108.

Felsenstein J. 1985. Phylogenies and the Comparative Method. *The American Naturalist* 125:1–15.

Formenti G, Theissinger K, Fernandes C, Bista I, Bombarely A, Bleidorn C, Ciofi C, Crottini A, Godoy JA, Höglund J, et al. 2022. The era of reference genomes in conservation genomics. *Trends in Ecology & Evolution* 37:197–202.

Fuller ZL, Berg JJ, Mostafavi H, Sella G, Przeworski M. 2019. Measuring intolerance to mutation in human genetics. *Nat Genet* 51:772–776.

Galtier N. 2016. Adaptive Protein Evolution in Animals and the Effective Population Size Hypothesis. *PLOS Genetics* 12:e1005774.

Galtier N, Rousselle M. 2020. How Much Does Ne Vary Among Species? *Genetics* 216:559–572.

Gardiner-Garden M, Frommer M. 1987. CpG Islands in vertebrate genomes. *Journal of Molecular Biology* 196:261–282.

Gaughran SJ. 2021. Patterns of Adaptive and Purifying Selection in the Genomes of Phocid Seals. Available from: https://www.proquest.com/docview/2557237113/abstract/646EE9058CAC4484PQ/1

Gilbert KJ, Zdraljevic S, Cook DE, Cutter AD, Andersen EC, Baer CF. 2022. The distribution of mutational effects on fitness in Caenorhabditis elegans inferred from standing genetic variation. *Genetics* 220:iyab166.

Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. 2009. Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data. *PLOS Genetics* 5:e1000695.

Hämälä T, Tiffin P. 2020. Biased Gene Conversion Constrains Adaptation in Arabidopsis thaliana. *Genetics* 215:831–846.

Huang X, Fortier AL, Coffman AJ, Struck TJ, Irby MN, James JE, León-Burguete JE, Ragsdale AP, Gutenkunst RN. 2021. Inferring Genome-Wide Correlations of Mutation Fitness Effects between Populations. *Molecular Biology and Evolution* 38:4588–4602.

Huber CD, Durvasula A, Hancock AM, Lohmueller KE. 2018. Gene expression drives the evolution of dominance. *Nature Communications* 9:1–11.

Huber CD, Kim BY, Marsden CD, Lohmueller KE. 2017. Determining the factors driving selective effects of new nonsynonymous mutations. *PNAS* 114:4465–4470.

Keightley PD, Eyre-Walker A. 2007. Joint Inference of the Distribution of Fitness Effects of Deleterious Mutations and Population Demography Based on Nucleotide Polymorphism Frequencies. *Genetics* 177:2251–2261.

Kim BY, Huber CD, Lohmueller KE. 2017. Inference of the Distribution of Selection Coefficients for New Nonsynonymous Mutations Using Large Samples. *Genetics* 206:345–361.

Kousathanas A, Oliver F, Halligan DL, Keightley PD. 2011. Positive and Negative Selection on Noncoding DNA Close to Protein-Coding Genes in Wild House Mice. *Molecular Biology and Evolution* 28:1183–1191.

Kumar S, Hedges SB. 1998. A molecular timescale for vertebrate evolution. *Nature* 392:917–920.

Kyriazis CC, Wayne RK, Lohmueller KE. 2021. Strongly deleterious mutations are a primary determinant of extinction risk due to inbreeding depression. *Evolution Letters* 5:33–47.

Li H. 2011. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* 27:2987–2993.

Liang Y, Shi Y, Yuan S, Zhou B, Chen X, An Q, Ingvarsson PK, Plomion C, Wang B. 2022. Linked selection shapes the landscape of genomic variation in three oak species. *New Phytologist* 233:555–568.

Limdi A, Owen SV, Herren C, Lenski RE, Baym M. 2022. Parallel changes in gene essentiality over 50,000 generations of evolution. :2022.05.17.492023. Available from: https://www.biorxiv.org/content/10.1101/2022.05.17.492023v1

Liu S, Zhang L, Sang Y, Lai Q, Zhang X, Jia C, Long Z, Wu J, Ma T, Mao K, et al. 2022. Demographic History and Natural Selection Shape Patterns of Deleterious Mutation Load and Barriers to Introgression across Populus Genome. *Molecular Biology and Evolution* 39:msac008.

Lourenço J, Galtier N, Glémin S. 2011. Complexity, Pleiotropy, and the Fitness Effect of Mutations. *Evolution* 65:1559–1571.

Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen W-M. 2010. Robust relationship inference in genome-wide association studies. *Bioinformatics* 26:2867–2873.

Martin G, Lenormand T. 2006. A General Multivariate Extension of Fisher's Geometrical Model and the Distribution of Mutation Fitness Effects Across Species. *Evolution* 60:893–907.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, et al. 2010. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20:1297–1303.

Miles A, Harding NJ, Bottà G, Clarkson CS, Antão T, Kozak K, Schrider DR, Kern AD, Redmond S, Sharakhov I, et al. 2017. Genetic diversity of the African malaria vector Anopheles gambiae. *Nature* 552:96–100.

Morgulis A, Gertz EM, Schäffer AA, Agarwala R. 2006. WindowMasker: window-based masker for sequenced genomes. *Bioinformatics* 22:134–141.

Nadachowska-Brzyska K, Burri R, Smeds L, Ellegren H. 2016. PSMC analysis of effective population sizes in molecular ecology and its application to black-and-white Ficedula flycatchers. *Molecular Ecology* 25:1058–1072.

Nigenda-Morales SF, Lin M, Nuñez-Valencia PG, Kyriazis CC, Beichman AC, Robinson JA, Ragsdale AP, R JU, Archer FI, Viloria-Gómora L, et al. 2022. The Genomic Footprint of Whaling and Isolation in Fin Whale Populations. *In prep*.

Peterson KJ, Lyons JB, Nowak KS, Takacs CM, Wargo MJ, McPeek MA. 2004. Estimating metazoan divergence times with a molecular clock. *Proceedings of the National Academy of Sciences* 101:6536–6541.

Phung TN, Wayne RK, Wilson MA, Lohmueller KE. 2019. Complex patterns of sex-biased demography in canines. *Proceedings of the Royal Society B: Biological Sciences* 286:20181976.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, De Bakker PI, Daly MJ, et al. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American journal of human genetics* 81:559–575.

Quake SR. 2022. A decade of molecular cell atlases. *Trends in Genetics* [Internet]. Available from: https://www.sciencedirect.com/science/article/pii/S016895252200004X

R Core Team. 2019. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing Available from: https://www.R-project.org/

Ragsdale AP. 2022. Local fitness and epistatic effects lead to distinct patterns of linkage disequilibrium in protein-coding genes. :2021.03.25.437004. Available from: https://www.biorxiv.org/content/10.1101/2021.03.25.437004v2

Robinson JA, Kyriazis CC, Nigenda-Morales SF, Beichman AC, Rojas-Bracho L, Robertson KM, Fontaine MC, Lohmueller KE, Wayne RK, Taylor BL, et al. 2021. Genomics reveals high recovery potential in the critically endangered vaquita. *Submitted*.

Robinson JA, Kyriazis CC, Nigenda-Morales SF, Beichman AC, Rojas-Bracho L, Robertson KM, Fontaine MC, Wayne RK, Lohmueller KE, Taylor BL, et al. 2022. The critically endangered vaquita is not doomed to extinction by inbreeding depression. *Science* 376:635–639.

Robinson JA, Räikkönen J, Vucetich LM, Vucetich JA, Peterson RO, Lohmueller KE, Wayne RK. 2019. Genomic signatures of extensive inbreeding in Isle Royale wolves, a population on the threshold of extinction. *Science Advances* 5:eaau0757.

Siegal ML, Leu J-Y. 2014. On the Nature and Evolutionary Impact of Phenotypic Robustness Mechanisms. *Annual Review of Ecology, Evolution, and Systematics* 45:495–517.

Tataru P, Bataillon T. 2019. polyDFEv2.0: testing for invariance of the distribution of fitness effects within and across species. *Bioinformatics* 35:2868–2869.

Tataru P, Mollion M, Glémin S, Bataillon T. 2017. Inference of Distribution of Fitness Effects and Proportion of Adaptive Substitutions from Polymorphism Data. *Genetics* 207:1103–1119.

Taylor BL, Chivers SJ, Larese J, Perrin WF. 2007. Generation length and percent mature estimates for IUCN assessments of cetaceans. NOAA Southwest Fisheries Science Center Available from: http://swfsc.noaa.gov/BarbTaylorPubs.aspx

Tenaillon O. 2014. The Utility of Fisher's Geometric Model in Evolutionary Genetics. *Annual Review of Ecology, Evolution, and Systematics* 45:179–201.

Tenaillon O, Silander OK, Uzan J-P, Chao L. 2007. Quantifying Organismal Complexity using a Population Genetic Approach. *PLOS ONE* 2:e217.

Tollis M, Robbins J, Webb AE, Kuderna LFK, Caulin AF, Garcia JD, Bèrubè M, Pourmand N, Marques-Bonet T, O'Connell MJ, et al. 2019. Return to the Sea, Get Huge, Beat Cancer: An Analysis of Cetacean Genomes Including an Assembly for the Humpback Whale (Megaptera novaeangliae). *Mol Biol Evol* 36:1746–1763.

Troco ADA. 2012. Resistência a insecticidas em Anopheles gambiae s.l. na região de Luanda, Angola. Available from: https://run.unl.pt/handle/10362/11354

Valentine JW, Collins AG, Meyer CP. 1994. Morphological complexity increase in metazoans. *Paleobiology* 20:131–142.

Vaser R, Adusumalli S, Leng SN, Sikic M, Ng PC. 2016. SIFT missense predictions for genomes. *Nat Protoc* 11:1–9.

Wade EE, Kyriazis CC, Cavassim MIA, Lohmueller KE. 2022. Quantifying the fraction of new mutations that are recessive lethal. :2022.04.22.489225. Available from: https://www.biorxiv.org/content/10.1101/2022.04.22.489225v1

Wickham H. 2016. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York Available from: https://ggplot2.tidyverse.org

Zhen Y, Huber CD, Davies RW, Lohmueller KE. 2021. Greater strength of selection and higher proportion of beneficial amino acid changing mutations in humans compared with mice and Drosophila melanogaster. *Genome Res.* 31:110–120.

Zheng X, Levine D, Shen J, Gogarten S, Laurie C, Weir B. 2012. A High-performance Computing Toolset for Relatedness and Principal Component Analysis of SNP Data. *Bioinformatics* 28:3326–3328.

Appendix Chapter: A Reference Genome Assembly of the Bobcat, *Lynx rufus*

Submitted to *Journal of Heredity*

Supplemental Tables available online as this dissertation's Supplementary Materials: AppendixChapter_Supplementary_Tables.xlsx

Abstract

The bobcat (*Lynx rufus*) is a medium-sized carnivore well adapted to various environments and an indicator species for landscape connectivity. It is one of the four species within the extant *Lynx* genus in the family *Felidae*. Because of its broad geographic distribution and central role in food webs, the bobcat is important for conservation. Here we present a high quality *de novo* genome assembly of a male bobcat located in Mendocino County (California, USA) as part of the California Conservation Genomics Project (CCGP). The assembly was generated using the standard CCGP pipeline from a combination of Omni-C and HiFi technologies. The primary assembly comprises 76 scaffolds spanning 2.4 Gb, represented by a scaffold N50 of 142 Mb, a contig N50 of 66.2 Mb and a BUSCO completeness score of 95.90%. The bobcat genome will be an important resource for the effective management and conservation of this species and comparative genomics exploration.

Keywords: California Conservation Genomics Project, CCGP, carnivores, comparative genomics, felidae, long-read assembly

Introduction

The bobcat (*Lynx rufus*) is one of the most adaptable and widespread carnivores in the Western Hemisphere (Figure A.1A, Riley et al. 2003; Reding et al. 2012). They prefer rocky terrain in brushy forest or chaparral, but are also habitat generalists that can persist in anthropogenically altered areas (Figure A.1B, Ahlborn and White 1990). For these reasons, the bobcat is an exemplary study species for functional landscape connectivity, urbanization effects, and local adaptations (Smith et al. 2020). With a large home range and central role in the food web, they are also considered an umbrella species for conserving diverse ecological communities (Kozakiewicz et al. 2019).

The bobcat shares a common ancestor with three other species in the *Lynx* genus, the Canada lynx (*Lynx canadensis*), the Eurasian lynx (*L. lynx*) and the Iberian lynx (*L. pardinus*), that diverged approximately 3.2 million years ago (Figure A.1C, Johnson et al. 2006). Currently, the genomes for the Canada lynx (GCF_007474595.2; scaffold N50 = 147 Mb) and the Iberian lynx (GCA_900661375.1; scaffold N50 = 1.5 Mb) are available on NCBI Genbank (Table SA.1). These four lynx species vary greatly in their ecological traits and demographic histories, as well as abundance and conservation status (Broderick 2020). Obtaining a high-quality bobcat reference genome will enable comparative genomics analyses in this lineage.

In California, the bobcat is a native mesocarnivore species that is crucial for ecosystem health (Ahlborn and White 1990). Regional studies using microsatellite or RADseq (Restriction site Associated DNA Sequencing) markers showed that habitat fragmentation, disease transmission and rodenticide exposure increasingly pose threats to urban bobcats in southern California (Serieys et al. 2015; Fraser et al. 2018; Kozakiewicz et al. 2019). However, studies at a broader geographic scale examining the entire genome are still lacking. A bobcat reference genome would expand the genetic toolkit available to wildlife scientists responsible for the conservation and management of bobcat, both in California and throughout its native range.

The California Conservation Genomics Project (CCGP) is generating a genomics variation database for hundreds of species with broad statewide distributions to help guide conservation of species and ecosystems under anthropogenic changes (Shaffer et al. 2022). As one of the study species focused on by the CCGP, here we present a high quality *de novo* genome assembly of the bobcat, with high contiguity, base accuracy and minimal gaps. The assembly derives from genomic DNA extracted from fresh whole blood samples taken from a male bobcat that was treated at a wildlife rehabilitation facility. With high molecular weight DNA, we leveraged the advantages of Omni-C proximity-ligation technologies and PacBio long read sequencing to generate a high quality assembly with chromosome-length scaffolds that is comparable to, or better than, existing lynx reference genomes (Abascal et al. 2016; Rhie et al. 2021). The bobcat assembly we present here, has a total length of 2.44 Gb, a scaffold N50 of 142 Mb and a contig N50 of 66.2 Mb. The bobcat genome will provide reference for high-resolution mapping of short read data and genomic variation discovery in ongoing CCGP landscape genomics surveys and serve as a useful resource for comparative analyses.

Methods

Biological Materials

Whole blood was sampled from a male bobcat admitted to a wildlife rehabilitation facility for treatment of injuries sustained during a vehicle collision. The sample was collected under a Memorandum of Understanding with the California Department of Fish and Wildlife per CCR Title 14, Section 679. This male bobcat was found near Redwood Valley, Mendocino County (GPS: 39.26564 N, 123.15892 W, WGS84), CA, USA. Whole blood samples were drawn in EDTA blood collection tubes, refrigerated overnight, flash-frozen in liquid nitrogen and transferred on dry ice to the sequencing facilities within 24 hours of collection. Samples were then stored at -80º C until DNA extraction and sequencing. A voucher subsample is stored in the CCGP archive at the University of California - Los Angeles at -80º C.

Nucleic acid library preparation and DNA sequencing

Pacific Biosciences HiFi library preparation and sequencing

High molecular weight (HMW) genomic DNA (gDNA) was isolated from whole blood preserved in EDTA. Three ml of RBC lysis solution (Qiagen Cat # 158445) was added to 1 ml of whole blood and the reaction was incubated at room temperature for 5 minutes (min). The sample was centrifuged at 2000 x g for 2 min to pellet white blood cells. The supernatant was discarded and 2 ml of lysis buffer containing 10 mM Tris-HCl pH 8.0, 25 mM EDTA, 0.5% (w/v) SDS and 100µg/ml Proteinase K was added to the cell pellet. The reaction was incubated at room temperature until the solution was homogenous. The lysate was then treated with 20µg/ml RNase A at 37º C for 30 min and cleaned with equal volumes of phenol/chloroform using phase lock gels (Quantabio Cat # 2302830). The DNA was precipitated by adding 0.4X volume of 5M ammonium acetate and 3X volume of ice cold ethanol. The DNA pellet was washed twice with 70% ethanol and resuspended in an elution buffer (10mM Tris, pH 8.0). Purity of gDNA was measured on a NanoDrop 1000 spectrophotometer (Thermo Scientific, Waltham, MA, USA) using the 260/280 and 260/230 ratios. The gDNA sample with a 260/280 ratio between 1.8 to 2.0 and a 260/230 ratio no less than 2.0 was considered pure (Pacific Biosciences 2021). The integrity of the HMW gDNA was verified on a Femto pulse system (Agilent Technologies, Santa Clara, CA).

The HiFi SMRTbell library was constructed using the SMRTbell Express Template Prep Kit v2.0 (Pacific Biosciences - PacBio; Menlo Park, CA; Cat. #100-938-900) according to the manufacturer’s instructions. This entailed HMW gDNA shearing to a target DNA size distribution between 15 – 20 kb. The sheared gDNA was concentrated using 0.45X of AMPure PB beads (PacBio, Cat. #100-265-900) for the removal of single-strand overhangs at 37º C for 15 min, followed by further enzymatic steps of DNA damage repair at 37º C for 30 min, end repair and A-tailing at 20º C for 10 min and 65º C for 30 min, ligation of overhang adapter v3 at 20º C for 60 min and 65º C for 10 min to inactivate the ligase, then nuclease treated at 37º C for one hour. The SMRTbell library was purified and concentrated with 0.45X Ampure PB beads (PacBio, Cat. #100-265-900) for size selection using the BluePippin system (Sage Science, Beverly, MA; Cat #BLF7510) to collect fragments greater than 9 kb. The 15 – 20 kb average HiFi SMRTbell library was sequenced at the University of California - Davis DNA Technologies Core (Davis, CA) using three 8M SMRT cells, Sequel II sequencing chemistry 2.0, and 30-hour movies each on a PacBio Sequel II sequencer.

Omni-C library preparation and sequencing

The Omni-C library was prepared using the Dovetail^{TM} Omni-C^{TM} Kit according to the manufacturer’s protocol with slight modifications. Briefly, chromatin was fixed in place in the nucleus. Fixed chromatin was digested with DNase I then extracted. Chromatin ends were repaired and ligated to a biotinylated bridge adapter followed by proximity ligation of adapter containing ends. After proximity ligation, crosslinks were reversed and the DNA purified from proteins. Purified DNA was treated to remove biotin that was not internal to ligated fragments. A sequencing library was generated using the NEB Ultra II DNA Library Prep kit (New England Biolabs, Ipswich, MA) with an Illumina compatible y-adaptor. Biotin-containing fragments were then captured using streptavidin beads. The post-capture product was split into two replicates prior to PCR enrichment to preserve library complexity with each replicate receiving unique dual indices. The library was sequenced at Vincent J. Coates Genomics Sequencing Lab (Berkeley, CA) on an Illumina NovaSeq platform (Illumina, San Diego, CA) to generate approximately 100 million paired end 150-bp reads per GB of genome size.

Genome assembly

Nuclear genome assembly

We assembled the genome of the bobcat following the CCGP assembly protocol Version 3.0, an improvement from Todd et al. (submitted). The main difference between versions is the use of an updated version of the *de novo* assembler HiFiasm [Version 0.16.1-r375] (Cheng et al. 2021, see Table A.1 for assembly pipeline and relevant software). The final output corresponds to a diploid assembly that consists of two pseudo haplotypes (primary and alternate). The primary assembly is more complete and consists of longer phased blocks. The alternate consists of haplotigs (contigs of clones with the same haplotype) in heterozygous regions and is not as complete and more fragmented. Given the characteristics of the latter, the alternate assembly cannot be considered on its own but as a complement of the primary assembly (https://lh3.github.io/2021/04/17/concepts-in-phased-assemblies; https://www.ncbi.nlm.nih.gov/grc/help/definitions/)

We removed remnant adapter sequences from the PacBio HiFi dataset using HiFiAdapterFilt [Version 1.0] (Sim 2021) and generated the initial diploid assembly with the filtered PacBio reads using HiFiasm. Next, we identified sequences corresponding to haplotypic duplications and contig overlaps on the primary assembly with purge_dups [Version 1.2.6]

(Guan et al. 2020) and transferred them to the alternate assembly. We scaffolded both assemblies using the Omni-C data with SALSA [Version 2.2] (Ghurye et al. 2017; 2019).

The primary assembly was manually curated by generating and analyzing Omni-C contact maps and breaking the assembly where major misassemblies were found. No further joins were made after this step. To generate the contact maps, we aligned the Omni-C data against the corresponding reference with bwa mem [Version 0.7.17-r1188, options -5SP] (Li 2013), identified ligation junctions, and generated Omni-C pairs using pairtools [Version 0.3.0] (Goloborodko et al. 2019). We generated a multi-resolution Omni-C matrix with cooler [Version 0.8.10] (Abdennur and Mirny 2020) and balanced it with hicExplorer [Version 3.6] (Ramírez et al. 2018). We used HiGlass [Version 2.1.11] (Kerpedjiev et al. 2018) and the PretextSuite (https://github.com/wtsi-hpag/PretextView; https://github.com/wtsi-hpag/PretextMap; https://github.com/wtsi-hpag/PretextSnapshot) to visualize the contact maps.

We closed the remaining gaps generated during scaffolding with the PacBio HiFi reads and YAGCloser [commit 20e2769] (https://github.com/merlyescalona/yagcloser). We then checked for contamination using the BlobToolKit Framework [Version 2.3.3] (Challis et al. 2020). Finally, we trimmed remnants of sequence adaptors and mitochondrial contamination based on NCBI contamination screening.

Mitochondrial genome assembly

We assembled the mitochondrial genome of the bobcat from the PacBio HiFi reads using the reference-guided pipeline MitoHiFi [https://github.com/marcelauliano/MitoHiFi] (Allio et al. 2020). The mitochondrial sequence of *Lynx lynx* (MH706704.1) was used as the starting reference sequence. After completion of the nuclear genome, we searched for matches of the resulting mitochondrial assembly sequence in the nuclear genome assembly using BLAST+

[Version 2.10] (Camacho et al. 2009) and filtered out contigs and scaffolds from the nuclear genome with a percentage of sequence identity >99% and size smaller than the mitochondrial assembly sequence.

Genome size estimation and quality assessment

We generated k-mer counts (k = 21) from the PacBio HiFi reads using meryl [Version 1] (https://github.com/marbl/meryl). The generated k-mer database was then used in GenomeScope2.0 [Version 2.0] (Ranallo-Benavidez et al. 2020) to estimate genome features including sequencing error, genome size, heterozygosity, and repeat content. To obtain general contiguity metrics, we ran QUAST [Version 5.0.2] (Gurevich et al. 2013). To evaluate genome quality and completeness we used BUSCO [Version 5.0.0] (Simão et al. 2015; Seppey et al. 2019) with the mammalia database (mammalia_odb10) which contains 9,226 genes. Assessment of base level accuracy (QV) and k-mer completeness was performed using the previously generated meryl database and merqury (Rhie et al. 2020). We further estimated genome assembly accuracy via BUSCO gene set frameshift analysis using the pipeline described in Korlach et al. (2017).

Assembly comparisons

We compared basic statistics with the other two existing nuclear assemblies and three mitochondrial assemblies in the *Lynx* genus (Figure A.1C). For nuclear assemblies, we downloaded the *Lynx canadensis* genome RefSeq assembly (GCF_007474595.2_mLynCan4.pri.v2; accessed on 2021-08-13) and the *Lynx pardinus* assembly (GCA_900661375.1_LYPA1.0; accessed on 2022-02-16). To compare basic statistics in the nuclear assemblies, we compiled information from the NCBI Genome Assembly Reports and individual publications (Table SA.1, Abascal et al. 2016; Rhie et al. 2021). To standardize the BUSCO scores, we repeated the BUSCO analyses described above for the other assemblies (Table SA.2). The divergence time plot was generated using the ggtree package [Version 2.0.4] (Yu et al. 2017) in R [Version 3.6.2] (R Core Team 2019). The coverage by scaffold length (NGx) plot was generated based on the scaffold lengths in the NCBI Genome Assembly Reports for each species using ggplot2 [Version 3.3.2] (Wickham 2016) in R.

For mitochondrial assemblies, we downloaded the sequences for Genbank accessions: CM017348.2 (*Lynx canadensis*), MH706704.1 (*Lynx lynx*) and NC_028319.1 (*Lynx pardinus*) on 2022-02-20. The base compositions and sequence lengths were summarized using biopython [Version 1.79] (Cock et al. 2009).

To count available whole genome assemblies in the Felidae, we queried the NCBI Assembly database using the Felidae taxonomy ID on 2022-02-21 (https://www.ncbi.nlm.nih.gov/assembly/?term=txid9681%5BOrganism%3Aexp%5D). The assembly species names were matched to the Felidae taxonomy described in Kitchener et al. (2017).

Results

Nuclear assembly

We generated a de novo nuclear genome assembly of the bobcat (mLynRuf1) using 247.6 million read pairs of Omni-C data and 6.7 million PacBio HiFi reads. The latter yielded ~ 40 fold coverage (N50 read length 14,593 bp; minimum read length 45 bp; mean read length 14,504 bp; maximum read length of 52,209 bp) based on the final assembled genome size of 2.4 Gb (Figure A.2A). Assembly statistics are reported in tabular and graphical form in Table A.2 and Figure A.2B, respectively.

The primary assembly consists of 76 scaffolds spanning 2.4 Gb with contig N50 of 66.2 Mb, scaffold N50 of 142.1 Mb, largest contig of 202.7 Mb, and largest scaffold of 239.8 Mb. Using BlobToolKit and BLAST+, we identified and removed one contig from the primary assembly corresponding to mitochondrial contamination, and seven contigs from the alternate assembly, six contigs corresponding to mitochondrial contamination and one contig to an arthropod contaminant. The Omni-C contact map suggests that the primary assembly is highly contiguous (Figure A.2C). As expected, the alternate assembly, which consists of sequence from heterozygous regions, is less contiguous (Figure A.2D). Because the primary assembly is not fully phased, we have deposited scaffolds corresponding to the alternate haplotype in addition to the primary assembly.

The final genome size (2.4 Gb) is close to the estimated values from the Genomescope2.0 k-mer spectra. The k-mer spectrum output shows a bimodal distribution with two major peaks, at ~ 19- and ~ 38-fold coverage, where peaks correspond to homozygous and heterozygous states respectively (Figure A.2A).

Based on PacBio HiFi reads, we estimated a 0.15% sequencing error rate and 0.59% nucleotide heterozygosity rate. The assembly has a BUSCO completeness score of 95.9% using the mammalia gene set, a per base quality (QV) of 66, a k-mer completeness of 96.3 and a frameshift indel QV of 47.15.

Mitochondrial assembly

The mitochondrial genome assembled with MitoHiFi has a final size of 17,097 bp. The base composition of the final assembly version is A = 32.43%, C = 26.64%, G = 14.24%, T = 26.69%, and consists of 22 transfer RNAs and 13 protein coding genes. Within the *Lynx* genus, the mitochondrial genome size is conserved (16,806 bp – 17,097 bp). The mitochondrial base compositions vary little across species as well (A = 32.31% – 32.43%, C = 26.64% – 27.06%, G =14.16% – 14.29%, T =26.35% – 26.69%; Table SA.3).

Discussion

Here we presented a high-quality bobcat reference genome assembly, with some scaffold lengths reaching chromosome levels. The five longest scaffolds in the bobcat assembly have almost identical lengths compared with the assigned A1, C1, B1, A2 and C2 chromosomes in the Canada lynx assembly (Figure A.1D). This bobcat assembly is highly continuous, complete and accurate. With a contig N50 of 66 Mb, scaffold N50 of 142 Mb, total gap length of 2.4 kb and a base pair QV of 66, our assembly greatly exceeds the best available standards of a minimum contig N50 of 1Mb, scaffold N50 of 10 Mb and QV of 40 proposed by the Vertebrate Genome Project (VGP, Rhie et al. 2021).

Compared with the other two available *Lynx* assemblies for Canada lynx (LYCA) and Iberian lynx, our bobcat assembly (LYRU) is of similar quality to the VGP Canada lynx assembly, which also utilized both long-read and short-read sequences (Figure A.1D; Table SA.1). We achieved a higher accuracy (QV=66) compared with the Canada lynx assembly (QV=36.8), a slightly higher BUSCO completeness score (95.9% for LYRU and 94.4% for LYCA; Table SA.2), less total gap length (2.4 kb for LYRU and 2.8 Mb for LYCA) and equivalent k-mer completeness (96.3% for LYRU and 96.4% for LYCA). The Canada lynx assembly generated chromosome assignments, which are not included in this current bobcat assembly release. Both bobcat and Canada lynx genomes were annotated using the NCBI Eukaryotic Genome Annotation Pipeline (Table SA.1). BUSCO analysis of gene annotations for our bobcat assembly using the carnivora_odb10 lineage dataset showed a 98.5% completeness, suggesting a high annotation quality (https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Lynx_rufus/100/). Both bobcat and Canada lynx genomes are superior in assembly metrics compared to the Iberian lynx assembly that was generated using only short-read sequencing techniques (Figure A.1D).

This bobcat assembly will provide resources for comparative genomic studies within the *Lynx* lineage, and more broadly the *Felidae* family. Of the 41 living felid species in eight *Felidae* lineages (Kitchener et al. 2017), 17 species have at least one genome assembly available in NCBI (Table SA.4). Assembled genome sizes in the *Lynx* lineage, measured in total genome assembly length, are approximately 2.4 Gb for all three existing assemblies (Table SA.1). At a larger scale, the genome sizes in the *Felidae* family are also conserved (2.30 Gb for *Panthera leo*, GCF_018350215.1 to 2.58 Gb for *Panthera pardus*, GCF_001857705.1; Table SA.4). The assembled *Lynx rufus* genome size is smaller than the flow cytometry measured size of 2.92 Gb for the *Lynx lynx* (Vinogradov 1998; Gregory 2005), a pattern observed in other species possibly caused by the repetitive regions (Elliott and Gregory 2015). Species within the *Lynx* lineage vary in abundance and conservation status, which is reflected in the nucleotide heterozygosity. The more abundant bobcat and Canada lynx have 0.59% and 0.19% heterozygosity in their assemblies (Rhie et al. 2021), while only 0.01% heterozygosity was reported for the endangered Iberian lynx (Abascal et al. 2016).

In addition to evolutionary studies, the bobcat reference genome will be an essential resource for genetics-informed conservation management. Currently, a bobcat hunting ban is in place in California until 2025, at which time the Fish and Game Commission must re-evaluate the appropriateness of a hunting season based on the best available science (California Assembly Bill No.1254, 2019). To support this evaluation, the California Statewide Bobcat Population Monitoring project is underway to assess population status (CDFW 2021). The bobcat genome will provide reference for ongoing CCGP whole genome resequencing projects that aim to identify statewide Management Units, evaluate genomic health and assess the outcomes of various hunting scenarios through genomics-informed simulations. Across its geographic range from southern Canada to Mexico (Kelly et al. 2016), researchers have studied bobcats to characterize patterns of genetic variation on a continental scale (Reding et al. 2012; Broderick 2020), and to assess impacts of habitat fragmentation on gene flow (Serieys et al. 2015; Janecka et al. 2016), as well as urbanization associated disease and toxins (Fraser et al. 2018; Kozakiewicz et al. 2020). The availability of a high-quality genome assembly will further advance research topics such as these as well.

In summary, this highly contiguous, complete, and accurate assembly for bobcat is a part of the larger goal of the California Conservation Genomics Project to build the most comprehensive conservation genomics dataset known to date. The availability of such high-quality assemblies will serve as an important tool for both fundamental evolutionary studies and conservation applications.

Tables

Table A.1

Table A.1. Assembly pipeline and software usage. Software citations are listed in the text.

Assembly	**Software**	**Version**
Filtering PacBio HiFi adapters	HiFiAdapterFilt https://github.com/sheinasim/HiFiAdapterFilt	Commit 64d1c7b
K-mer counting	Meryl	1
Estimation of genome size and heterozygosity	GenomeScope	2
De novo assembly (contiging)	HiFiasm	0.16.1-r375
Long read, genome-genome alignment	minimap2	2.16
Remove low-coverage, duplicated contigs	purge_dups	1.2.6
Scaffolding		
Omni-C mapping for SALSA	Arima Genomics mapping pipeline https://github.com/ArimaGenomics/mapping_pipeline	Commit 2e74ea4
Omni-C Scaffolding	SALSA	2
Gap closing	YAGCloser https://github.com/merlyescalona/yagcloser	Commit 20e2769
Omni-C Contact map generation		
Short-read alignment	bwa	0.7.17-r1188
SAM/BAM processing	samtools	1.11
SAM/BAM filtering	pairtools	0.3.0
Pairs indexing	pairix	0.3.7
Matrix generation	Cooler	0.8.10
Matrix balancing	hicExplorer	3.6
Contact map visualization	HiGlass	2.1.11
	PretextMap	0.1.4
	PretextView	0.1.5
	PretextSnapshot	0.0.3

Table A.1. Assembly pipeline and software usage. Software citations are listed in the text (cont.).

Organelle assembly		
Mitogenome assembly	MitoHiFi	2 Commit c06ed3e
Genome quality assessment		
Basic assembly metrics	QUAST	5.0.2
Assembly completeness	BUSCO	5.0.0
	Merqury	1
Contamination screening		
Local alignment tool	BLAST+	2.10
General contamination screening	BlobToolKit	2.3.3

Table A.2

Table A.2. Sequencing and assembly statistics, and accession numbers.

BioProjects and vouchers	CCGP NCBI BioProject		PRJNA720569	
	Genera NCBI BioProject		PRJNA765621	
	Species NCBI BioProject		PRJNA777191	
	NCBI BioSample		SAMN23391104	
	Specimen identification		CCGP_SWC_20201006	
Genome sequence	NCBI Genome accessions		**Primary**	**Alternate**
	Assembly accession		GCA_022079265.1	GCA_022079275.1
	Genome sequences		JAJSDN000000000	JAJSDO000000000
Sequencing data	PacBio HiFi reads	Run	3 PACBIO_SMRT (Sequel II) runs: 6.7 M spots, 97.2 G bases, 65.5Gb	
		Accession	SRR17978068	
	Omni-C Illumina reads	Run	2 Illumina NovaSeq 6000 runs: 247.6 M spots, 74.8 G bases, 24.9 Gb	
		Accession	SRR17978066-67	
Genome assembly quality metrics	Assembly identifier (Quality code *)		mLynRuf1 (7.8.Q66)	
	HiFi Read coverage §		40X	
			Primary	**Alternate**
	Number of contigs		100	72,828
	Contig N50 (bp)		66,217,191	97,441
	Longest Contigs		202,776,522	1,850,898
	Number of scaffolds		76	68,112
	Scaffold N50 (bp)		142,134,035	107,693
	Largest scaffold		239,891,529	11,854,743
	Size of final assembly (bp)		2,439,256,471	3,320,187,595
	Gaps per Gbp		8	1,420
	Indel QV (Frame shift)		47.15	47.15

Base pair QV				66.5		58.25
						Full assembly = 60.19
k-mer completeness				96.31		88.12
						Full assembly = 99.74
BUSCO		**C**	**S**	**D**	**F**	**M**
completeness (mammalia) n=9,226	P‡	95.90%	95.20%	0.70%	1.20%	2.90%
	A‡	81.10%	74.60%	6.50%	5.10%	13.80%
Organelles		1 Complete mitochondrial sequence				CM039064.1

** Assembly quality code x.y.Q derived notation, from Rhie et al. 2021. x = log10[contig NG50]; y = log10[scaffold NG50]; Q = Phred base accuracy QV (Quality value). BUSCO Scores. (C)omplete and (S)ingle; (C)omplete and (D)uplicated; (F)ragmented and (M)issing BUSCO genes. n, number of BUSCO genes in the set/database. bp: base pairs*

§ Read coverage has been calculated based on a genome size of 2.4 Gb.

‡ P(rimary) and (A)lternate assembly values

Figures

Figure A.1

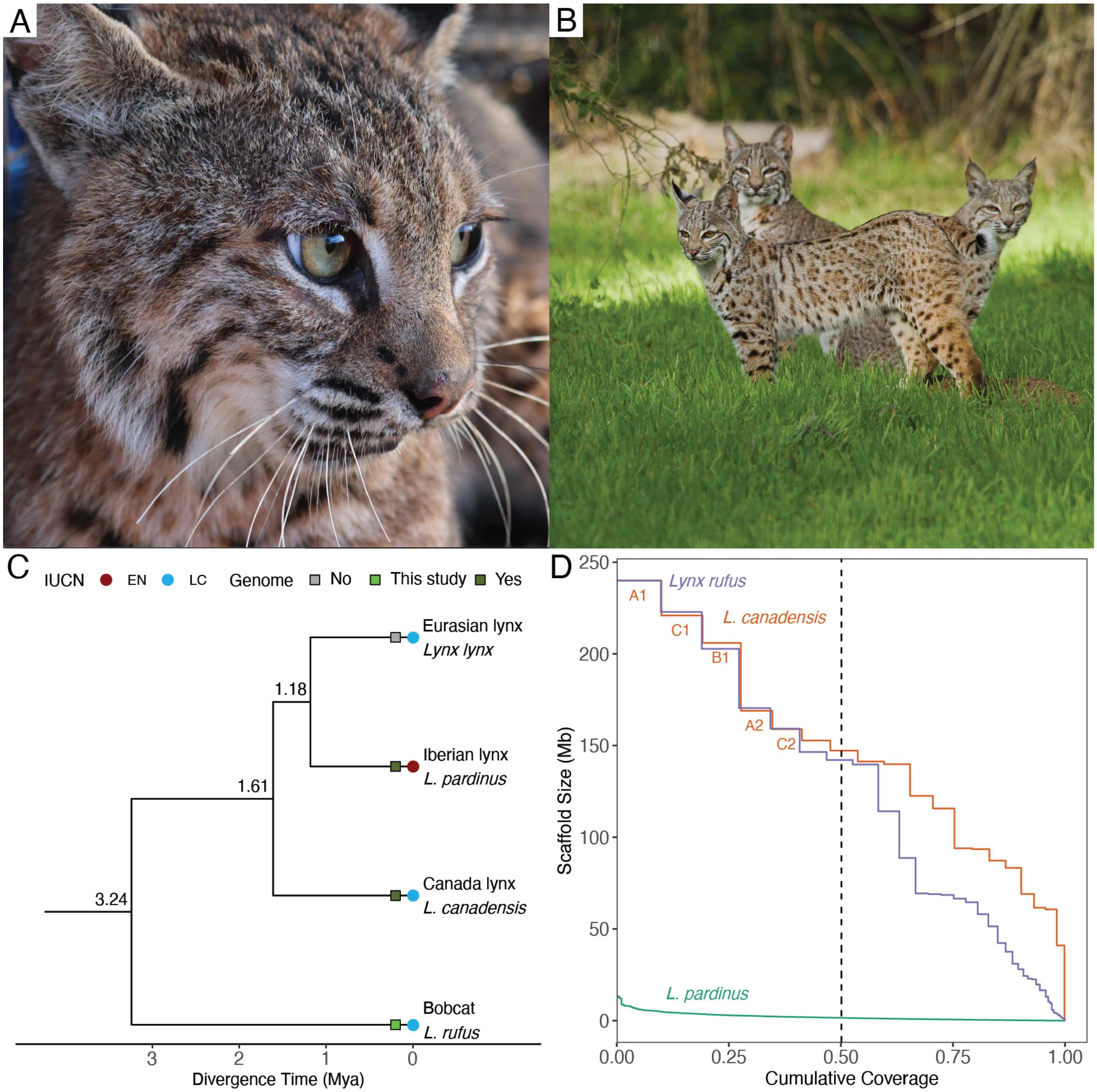

Figure A.1. A bobcat reference genome assembly. **A)** A bobcat, *Lynx rufus* (Photo Credit: Laurel Serieys). **B)** Representative habitats for bobcats (Photo Credit: Barry Rowan). **C)** Phylogenetic relationships in the *Lynx* genus. The IUCN Red List status (EN: Endangered, LC: Least Concern) and genome assembly availability (Yes: available on NCBI, No: unavailable on

NCBI) are denoted. Divergence time estimates are in units of million years ago (Mya, Johnson et al. 2006). **D)** NGx plot comparing the three available *Lynx* genome assemblies. This plot shows the *x* fraction of genome assembly that is represented by scaffolds of at least *y* Mb. The N50 value is represented by the dashed vertical line. Our bobcat assembly (purple) has similar scaffold-level contiguity with the Canada lynx assembly (orange). The Iberian lynx assembly (green) has lower scaffold-level contiguity. The names for five longest Canada lynx chromosomes are annotated.

Figure A.2

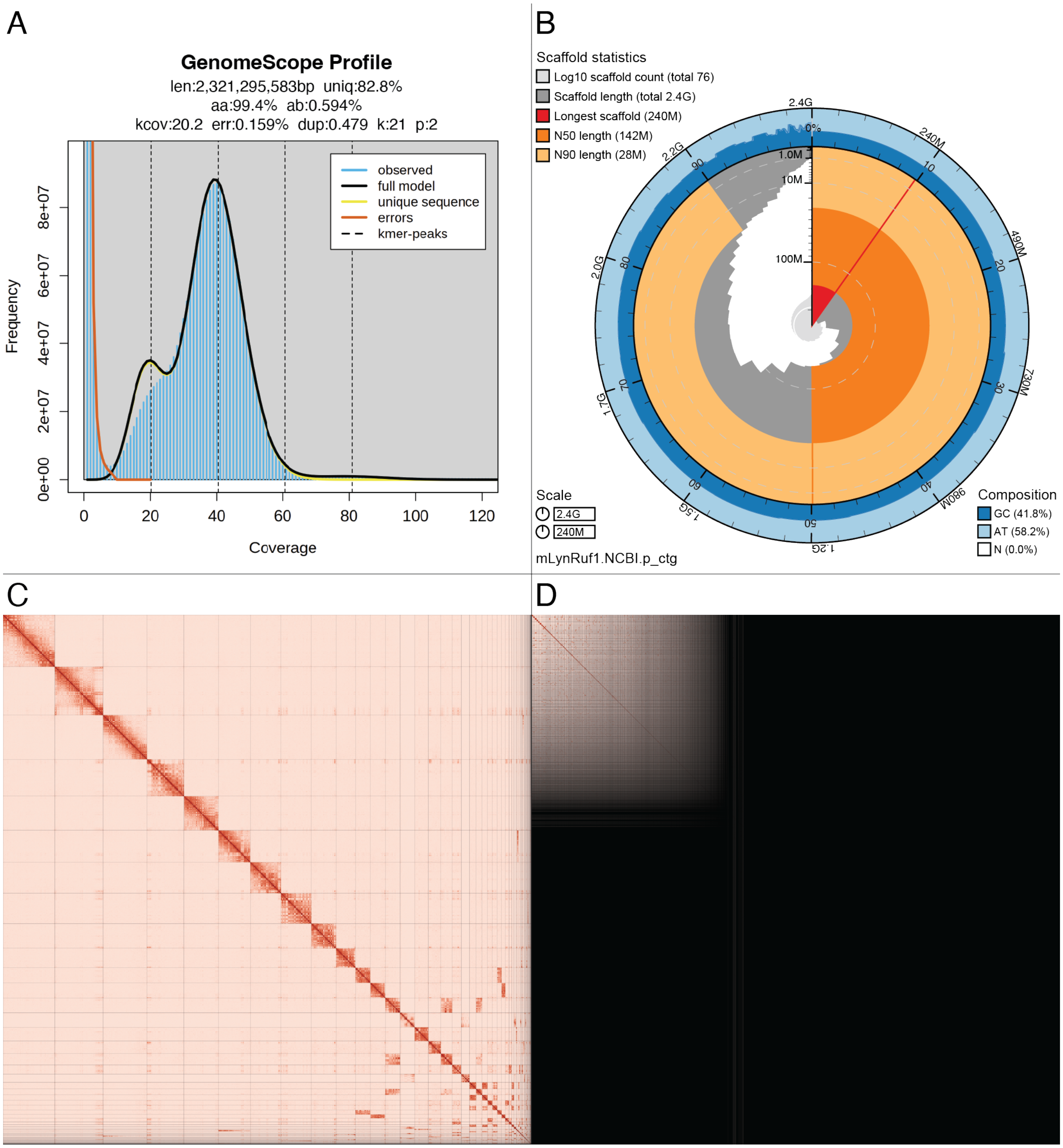

Figure A.2. Visual overview of genome assembly metrics. **A)** K-mer spectra output generated from PacBio HiFi data without adapters using GenomeScope2.0. The bimodal pattern observed corresponds to a diploid genome. K-mers covered at lower coverage and lower frequency correspond to differences between haplotypes, whereas the higher coverage and higher frequency

k-mers correspond to the similarities between haplotypes. **B)** BlobToolKit Snail plot showing a graphical representation of the quality metrics presented in Table A.2 for the *Lynx rufus* primary assembly (mLynRuf1). The plot circle represents the full size of the assembly. From the inside-out, the central plot covers length-related metrics. The red line represents the size of the longest scaffold; all other scaffolds are arranged in size-order moving clockwise around the plot and drawn in gray starting from the outside of the central plot. Dark and light orange arcs show the scaffold N50 and scaffold N90 values. The central light gray spiral shows the cumulative scaffold count with a white line at each order of magnitude. White regions in this area reflect the proportion of Ns in the assembly. The dark vs. light blue area around it shows mean, maximum and minimum GC vs. AT content at 0.1% intervals (Challis et al. 2020). **C-D)** Omni-C contact maps for the primary (2C) and alternate (2D) genome assembly generated with PretextSnapshot. Omni-C contact maps translate proximity of genomic regions in 3-D space to contiguous linear organization. Each cell in the contact map corresponds to sequencing data supporting the linkage (or join) between two of such regions. Scaffolds are separated by black lines and higher density corresponds to higher levels of fragmentation.

References

Abascal, Federico, André Corvelo, Fernando Cruz, José L. Villanueva-Cañas, Anna Vlasova, Marina Marcet-Houben, Begoña Martínez-Cruz, et al. 2016. "Extreme Genomic Erosion after Recurrent Demographic Bottlenecks in the Highly Endangered Iberian Lynx." *Genome Biology* 17 (1): 251. https://doi.org/10.1186/s13059-016-1090-1.

Abdennur, Nezar, and Leonid A Mirny. 2020. "Cooler: Scalable Storage for Hi-C Data and Other Genomically Labeled Arrays." Edited by Jonathan Wren. *Bioinformatics* 36 (1): 311–16. https://doi.org/10.1093/bioinformatics/btz540.

Ahlborn, G, and M White. 1990. "California's Wildlife, Bobcat." https://nrm.dfg.ca.gov/FileHandler.ashx?DocumentID=2609&inline=1.

Allio, Rémi, Alex Schomaker-Bastos, Jonathan Romiguier, Francisco Prosdocimi, Benoit Nabholz, and Frédéric Delsuc. 2020. "MitoFinder: Efficient Automated Large-scale Extraction of Mitogenomic Data in Target Enrichment Phylogenomics." *Molecular Ecology Resources* 20 (4): 892–905. https://doi.org/10.1111/1755-0998.13160.

Broderick, Jennifer. 2020. "A Genomic Analysis of Bobcat Populations in North America with a Comparison to the Canada Lynx: An Assessment of Local Adaptation to Unique Ecoregions and Phylogeography." https://dsc.duq.edu/etd/1886.

California Assembly Bill No.1254, 2019. "Bill Text - AB-1254 Bobcats: Take Prohibition: Hunting Season: Management Plan." 2019. https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201920200AB1254.

Camacho, Christiam, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, Kevin Bealer, and Thomas L Madden. 2009. "BLAST+: Architecture and Applications." *BMC Bioinformatics* 10 (1): 421. https://doi.org/10.1186/1471-2105-10-421.

CDFW. 2021. “Science Institute News | CDFW Begins Statewide Bobcat Monitoring Project.” May 14, 2021. https://wildlife.ca.gov/Science-Institute/News/cdfw-begins-statewide-bobcat-monitoring-project.

Challis, Richard, Edward Richards, Jeena Rajan, Guy Cochrane, and Mark Blaxter. 2020. “BlobToolKit – Interactive Quality Assessment of Genome Assemblies.” G3 Genes|Genomes|Genetics 10 (4): 1361–74. https://doi.org/10.1534/g3.119.400908.

Cheng, Haoyu, Erich D. Jarvis, Olivier Fedrigo, Klaus-Peter Koepfli, Lara Urban, Neil J. Gemmell, and Heng Li. 2021. “Robust Haplotype-Resolved Assembly of Diploid Individuals without Parental Data.” *ArXiv:2109.04785 [q-Bio]*, September. http://arxiv.org/abs/2109.04785.

Cock, P. J. A., T. Antao, J. T. Chang, B. A. Chapman, C. J. Cox, A. Dalke, I. Friedberg, et al. 2009. “Biopython: Freely Available Python Tools for Computational Molecular Biology and Bioinformatics.” *Bioinformatics* 25 (11): 1422–23. https://doi.org/10.1093/bioinformatics/btp163.

Elliott, Tyler A., and T. Ryan Gregory. 2015. “What’s in a Genome? The C-Value Enigma and the Evolution of Eukaryotic Genome Content.” Philosophical Transactions of the Royal Society B: Biological Sciences 370 (1678): 20140331. https://doi.org/10.1098/rstb.2014.0331.

Fraser, Devaughn, Alice Mouton, Laurel E. K. Serieys, Steve Cole, Scott Carver, Sue Vandewoude, Michael Lappin, Seth P. D. Riley, and Robert Wayne. 2018. “Genome-Wide Expression Reveals Multiple Systemic Effects Associated with Detection of Anticoagulant Poisons in Bobcats (Lynx Rufus).” *Molecular Ecology* 27 (5): 1170–87. https://doi.org/10.1111/mec.14531.

Ghurye, Jay, Mihai Pop, Sergey Koren, Derek Bickhart, and Chen-Shan Chin. 2017. “Scaffolding of Long Read Assemblies Using Long Range Contact Information.” *BMC Genomics* 18 (1): 527. https://doi.org/10.1186/s12864-017-3879-z.

Ghurye, Jay, Arang Rhie, Brian P. Walenz, Anthony Schmitt, Siddarth Selvaraj, Mihai Pop, Adam M. Phillippy, and Sergey Koren. 2019. “Integrating Hi-C Links with Assembly Graphs for Chromosome-Scale Assembly.” Edited by Ilya Ioshikhes. *PLOS Computational Biology* 15 (8): e1007273. https://doi.org/10.1371/journal.pcbi.1007273.

Goloborodko, Anton, Nezar Abdennur, Sergey Venev, Hbbrandao, and Gfudenberg. 2019. Mirnylab/Pairtools v0.3.0. Zenodo. https://doi.org/10.5281/zenodo.2649383.

Gregory, T.R. (2005). Animal Genome Size Database. http://www.genomesize.com. Accessed 2022-05-13.

Guan, Dengfeng, Shane A McCarthy, Jonathan Wood, Kerstin Howe, Yadong Wang, and Richard Durbin. 2020. “Identifying and Removing Haplotypic Duplication in Primary Genome Assemblies.” Edited by Alfonso Valencia. *Bioinformatics* 36 (9): 2896–98. https://doi.org/10.1093/bioinformatics/btaa025.

Gurevich, Alexey, Vladislav Saveliev, Nikolay Vyahhi, and Glenn Tesler. 2013. “QUAST: Quality Assessment Tool for Genome Assemblies.” *Bioinformatics* 29 (8): 1072–75. https://doi.org/10.1093/bioinformatics/btt086.

Janecka, Jan E., Michael E. Tewes, Imogene A. Davis, Aaron M. Haines, Arturo Caso, Terry L. Blankenship, and Rodney L. Honeycutt. 2016. “Genetic Differences in the Response to Landscape Fragmentation by a Habitat Generalist, the Bobcat, and a Habitat Specialist, the Ocelot.” *Conservation Genetics* 17 (5): 1093–1108. https://doi.org/10.1007/s10592-016-0846-1.

Johnson, Warren E., Eduardo Eizirik, Jill Pecon-Slattery, William J. Murphy, Agostinho Antunes, Emma Teeling, and Stephen J. O'Brien. 2006. "The Late Miocene Radiation of Modern Felidae: A Genetic Assessment." *Science*, January. https://doi.org/10.1126/science.1122277.

Kelly, M., D. Morin, and C.A. Lopez-Gonzalez. 2016. "Lynx Rufus. The IUCN Red List of Threatened Species 2016: E.T12521A50655874." International Union for Conservation of Nature. https://doi.org/10.2305/IUCN.UK.2016-1.RLTS.T12521A50655874.en.

Kerpedjiev, Peter, Nezar Abdennur, Fritz Lekschas, Chuck McCallum, Kasper Dinkla, Hendrik Strobelt, Jacob M. Luber, et al. 2018. "HiGlass: Web-Based Visual Exploration and Analysis of Genome Interaction Maps." *Genome Biology* 19 (1): 125. https://doi.org/10.1186/s13059-018-1486-1.

Kitchener, A. C., C. Breitenmoser-Würsten, E. Eizirik, A. Gentry, Lars Werdelin, A. Wilting, N. Yamaguchi, et al. 2017. "A Revised Taxonomy of the Felidae : The Final Report of the Cat Classification Task Force of the IUCN Cat Specialist Group." http://repository.si.edu/xmlui/handle/10088/32616.

Korlach, Jonas, Gregory Gedman, Sarah B. Kingan, Chen-Shan Chin, Jason T. Howard, Jean-Nicolas Audet, Lindsey Cantin, and Erich D. Jarvis. 2017. "De Novo PacBio Long-Read and Phased Avian Genome Assemblies Correct and Add to Reference Genes Generated with Intermediate and Short Reads." *GigaScience* 6 (10). https://doi.org/10.1093/gigascience/gix085.

Kozakiewicz, Christopher P., Christopher P. Burridge, W. Chris Funk, Meggan E. Craft, Kevin R. Crooks, Robert N. Fisher, Nicholas M. Fountain-Jones, et al. 2020. "Does the Virus Cross the Road? Viral Phylogeographic Patterns among Bobcat Populations Reflect a

History of Urban Development." *Evolutionary Applications* 13 (8): 1806–17. https://doi.org/10.1111/eva.12927.

Kozakiewicz, Christopher P., Christopher P. Burridge, W. Chris Funk, Patricia E. Salerno, Daryl R. Trumbo, Roderick B. Gagne, Erin E. Boydston, et al. 2019. "Urbanization Reduces Genetic Connectivity in Bobcats (Lynx Rufus) at Both Intra– and Interpopulation Spatial Scales." *Molecular Ecology* 28 (23): 5068–85. https://doi.org/10.1111/mec.15274.

Li, Heng. 2013. "Aligning Sequence Reads, Clone Sequences and Assembly Contigs with BWA-MEM." *ArXiv:1303.3997 [q-Bio]*, May. http://arxiv.org/abs/1303.3997.

Pacific Biosciences. 2021. "Technical Overview: Ultra-Low DNA Input Library Preparation Using SMRTbell Express Template Prep Kit 2.0." https://www.pacb.com/wp-content/uploads/Ultra-Low-DNA-Input-Library-Preparation-Using-SMRTbell-Express-TPK-2.0-Customer-Training-01.pdf. Accessed 2022-05-13.

R Core Team. 2019. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Ramírez, Fidel, Vivek Bhardwaj, Laura Arrigoni, Kin Chung Lam, Björn A. Grüning, José Villaveces, Bianca Habermann, Asifa Akhtar, and Thomas Manke. 2018. "High-Resolution TADs Reveal DNA Sequences Underlying Genome Organization in Flies." *Nature Communications* 9 (1): 189. https://doi.org/10.1038/s41467-017-02525-w.

Ranallo-Benavidez, T. Rhyker, Kamil S. Jaron, and Michael C. Schatz. 2020. "GenomeScope 2.0 and Smudgeplot for Reference-Free Profiling of Polyploid Genomes." *Nature Communications* 11 (1): 1432. https://doi.org/10.1038/s41467-020-14998-3.

Reding, Dawn M., Anne M. Bronikowski, Warren E. Johnson, and William R. Clark. 2012. "Pleistocene and Ecological Effects on Continental-Scale Genetic Differentiation in the

Bobcat (Lynx Rufus).” *Molecular Ecology* 21 (12): 3078–93. https://doi.org/10.1111/j.1365-294X.2012.05595.x.

Rhie, Arang, Shane A. McCarthy, Olivier Fedrigo, Joana Damas, Giulio Formenti, Sergey Koren, Marcela Uliano-Silva, et al. 2021. “Towards Complete and Error-Free Genome Assemblies of All Vertebrate Species.” *Nature* 592 (7856): 737–46. https://doi.org/10.1038/s41586-021-03451-0.

Rhie, Arang, Brian P. Walenz, Sergey Koren, and Adam M. Phillippy. 2020. “Merqury: Reference-Free Quality, Completeness, and Phasing Assessment for Genome Assemblies.” *Genome Biology* 21 (1): 245. https://doi.org/10.1186/s13059-020-02134-9.

Riley, Seth P. D., Raymond M. Sauvajot, Todd K. Fuller, Eric C. York, Denise A. Kamradt, Cassity Bromley, and Robert K. Wayne. 2003. “Effects of Urbanization and Habitat Fragmentation on Bobcats and Coyotes in Southern California.” *Conservation Biology* 17 (2): 566–76. https://doi.org/10.1046/j.1523-1739.2003.01458.x.

Seppey, Mathieu, Mosè Manni, and Evgeny M. Zdobnov. 2019. “BUSCO: Assessing Genome Assembly and Annotation Completeness.” In *Gene Prediction*, edited by Martin Kollmar, 1962:227–45. Methods in Molecular Biology. New York, NY: Springer New York. https://doi.org/10.1007/978-1-4939-9173-0_14.

Serieys, Laurel E. K., Amanda Lea, John P. Pollinger, Seth P. D. Riley, and Robert K. Wayne. 2015. “Disease and Freeways Drive Genetic Change in Urban Bobcat Populations.” *Evolutionary Applications* 8 (1): 75–92. https://doi.org/10.1111/eva.12226.

Shaffer, H Bradley, Erin Toffelmier, Russ B Corbett-Detig, Merly Escalona, Bjorn Erickson, Peggy Fiedler, Mark Gold, et al. 2022. “Landscape Genomics to Enable Conservation

Actions: The California Conservation Genomics Project." *Journal of Heredity*, April, esac020. https://doi.org/10.1093/jhered/esac020.

Sim, Sheina. 2021. *Sheinasim/HiFiAdapterFilt: First Release* (version v1.0.0). Zenodo. https://doi.org/10.5281/ZENODO.4716418.

Simão, Felipe A., Robert M. Waterhouse, Panagiotis Ioannidis, Evgenia V. Kriventseva, and Evgeny M. Zdobnov. 2015. "BUSCO: Assessing Genome Assembly and Annotation Completeness with Single-Copy Orthologs." *Bioinformatics* 31 (19): 3210–12. https://doi.org/10.1093/bioinformatics/btv351.

Smith, Julia G., Megan K. Jennings, Erin E. Boydston, Kevin R. Crooks, Holly B. Ernest, Seth P. D. Riley, Laurel E. K. Serieys, Shaelynn Sleater-Squires, and Rebecca L. Lewison. 2020. "Carnivore Population Structure across an Urbanization Gradient: A Regional Genetic Analysis of Bobcats in Southern California." *Landscape Ecology* 35 (3): 659–74. https://doi.org/10.1007/s10980-020-00971-4.

Todd, Brian D., Thomas S. Jenkinson, Merly Escalona, Eric Beraut, Oanh Nguyen, Ruta Sahasrabudhe, Peter A. Scott, Erin Toffelmier, Ian J. Wang, and H. B. Shaffer. submitted. "Reference Genome of the Northwestern Pond Turtle, Actinemys Marmorata." *Journal of Heredity*.

Vinogradov, Alexander E. 1998. "Genome Size and GC-Percent in Vertebrates as Determined by Flow Cytometry: The Triangular Relationship." Cytometry: The Journal of the International Society for Analytical Cytology 31 (2): 100–109.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Yu, Guangchuang, David Smith, Huachen Zhu, Yi Guan, and Tommy Tsan-Yuk Lam. 2017. “Ggtree: An R Package for Visualization and Annotation of Phylogenetic Trees with Their Covariates and Other Associated Data.” *Methods in Ecology and Evolution* 8 (1): 28–36. https://doi.org/10.1111/2041-210X.12628.