Title

Evolutionary and Conservation Genomics of California Rodents

ISBN

9798297601383

Author

Voss, Erin Rebecca

Publication Date

2025-08-01

Peer reviewed|Thesis/dissertation

Evolutionary and Conservation Genomics of California Rodents

By

Erin Voss

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Integrative Biology

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Michael W. Nachman, Chair
Professor Rasmus Nielsen
Professor Rauri C. K. Bowie

Summer 2025

Abstract

Evolutionary and Conservation Genomics of California Rodents

by

Erin Voss

Doctor of Philosophy in Integrative Biology

University of California, Berkeley

Professor Michael W. Nachman, Chair

Understanding how organisms interact with each other and their environment has long been a central goal of evolutionary biology. Interactions between organisms, environmental adaptation, and response to landscape change leave signatures in genomic data, which in turn can be parsed to uncover the nature and details of these processes. For example, we can learn about natural selection and resulting adaptation by measuring changes in gene expression and comparing evolutionary rates across different genes and biological contexts. We can also evaluate genetic diversity, population connectivity, and deleterious mutational burden to assess whether species are in decline and provide context for biodiversity management. This is a key goal of the California Conservation Genomics Project (CCGP), a statewide effort to survey genetic variation of which I was a part, and which formed the basis for the research presented in chapters one and three through six.

For my dissertation, I used a combination of field work, laboratory work, and bioinformatic analysis to explore these questions in three California rodents: deer mice (*Peromyscus*), kangaroo rats (*Dipodomys*), and pocket gophers (*Thomomys*). My thesis takes the form of six chapters which comprise new genome assemblies paired with analysis of population- and species-level genomic data for each taxon. In chapters one, three, and five, I described new genome sequences for focal taxa that have been assembled in collaboration with the CCGP. In chapters two, four, and six, I explored questions regarding sexual selection and reproductive evolution, conservation genomic diversity and mutational load, and population connectivity and isolation across the landscape. Ultimately, my work focuses on generating and analyzing genomic data with the goal of understanding the impact of genetic variation for organisms.

In chapters one and two, I presented a new genome assembly for *Peromyscus maniculatus* from California and explore gene expression in the male reproductive tract for three species of *Peromyscus* mice with different mating systems. Firstly, I asked if gene expression differs in the testis, epididymis, or seminal vesicle in the monogamous California mouse (*Peromyscus californicus*) as compared with polygynandrous, multiple-mating North American deer mouse (*Peromyscus maniculatus*) or brush mouse (*Peromyscus boylii*). Secondly, I measured the rate of evolution of 4,816 genes across eight rodent species with broad variation in mating behavior to ask whether genes evolve via changes in expression, gene sequence, or both. Though these two sets of genes did not appear to overlap, suggesting distinct modes of response

to sexual selection, I identified a set of eleven reproductive genes, including seminal fluid proteins *Svs2* and *Pate4*, that displayed significant changes in both gene expression and evolutionary rate. These genes are strong candidates for involvement in sexual selection in rodents.

In chapters three and four, I focused on kangaroo rats (*Dipodomys*), a unique genus of desert-adapted rodents endemic to western North America. I first shared a genome assembly for Merriam's kangaroo rat (*D. merriami*), which is extremely widespread and ecologically flexible. I then compared *D. merriami* with five other species of kangaroo rats, three of which are endangered (Giant kangaroo rat *D. ingens*, San Joaquin Valley kangaroo rat *D. nitratoides*, and Stephens' kangaroo rat *D. stephensi*) and two of which have moderately large ranges and stable population trajectories (Heermann's kangaroo rat *D. heermanni*, Panamint kangaroo rat *D. panamintinus*). I compared measures of genetic diversity and inbreeding across species and inferred past effective population size before modeling the distribution of fitness effects for each species. Though genetic diversity is not associated with conservation status in kangaroo rats, endangered species appeared to carry a higher load of mildly to moderately deleterious variants. Further, presumed extinct subspecies *D. n. exilis* displayed strong evidence of inbreeding prior to extinction. Hence, endangered species with high background genetic diversity are still at risk of inbreeding due population isolation. I concluded that while genetic diversity is not a primary concern for endangered kangaroo rats, population isolation and subsequent genomic erosion is a significant threat to species in the fragmented landscapes of Southern California and the Central Valley.

In chapters five and six, I explored the phylogeographic history of one of America's most widely recognized rodents, Botta's pocket gopher (*Thomomys bottae*). I shared a genome assembly for *T. bottae* and report on repetitive elements in the genome, as classic karyotype studies have demonstrated extensive genomic variation in heterochromatin content across the species range. I then used whole-genome resequencing data for 95 individuals from across their California range to identify populations and geographic units, measure structure and gene flow between populations, and survey genetic diversity and inbreeding across the landscape. Whole genome resequencing data confirmed classic allozyme studies in identifying three to four major geographic units in the state, though the placement of populations in the Sierra Nevada differed. Further, I reported that populations are highly structured and inbred but maintain moderate levels of genetic variation despite isolation. This unique genomic profile provides a baseline for comparison with endangered species in California, many of which have similar heterozygosity and lower inbreeding than I observed in *T. bottae*.

Acknowledgements

Getting a Ph.D. takes a village, and I am deeply grateful to mine. First, thank you to my advisor, Dr. Michael Nachman, for sharing his enthusiasm for biology, his rigorous experimental design, his keen eye as an editor, and the incredible depth of his scholarship on the topics of evolution, population genetics, and mammalogy with me. I am especially grateful to Michael for encouraging me and everyone in his research group to find their own way and discover their passions in science.

Thank you also to my dissertation committee members, Dr. Rauri Bowie and Dr. Rasmus Nielsen, for sharing their time, scientific advice, and encouragement as I found my way. I am also grateful to my qualifying exam committee members, Dr. Jose Vazquez-Medina, Dr. Noah Whiteman, and Dr. Rosemary Gillespie, and my advisory committee member Dr. Bree Rosenblum. Museum of Vertebrate Zoology mammalogists Dr. Chris Conroy and Dr. Jim Patton taught me about field work, museum specimen preparation, deer mice, gophers, and kangaroo rats. I also learned so much from MVZ community members Lydia Smith and Terri Barclay – thank you.

To my Nachman lab mates: Sylvia Durkin, David Manahan, Noëlle Bittner, Mal Ballinger, Sarah Banker, Katya Mack, Isaac Linn, Tommy Herrera, Yocelyn Gutierrez-Guerrero, Libby Beckman, and Chris Kozak, thank you for sharing the highs and lows of graduate school with me. I could not have done this without you! Thank you also to my office mate David Tian and to my fellow MVZ and Integrative Biology graduate students. It has been a joy to be a part of this community with you for the past seven years.

Thank you to my parents, John Voss and Regina O'Donnell, for taking Rachel and I into the woods and onto the river. Your love of nature has become a core part of my identity and my reason for pursuing science. To my sister, Rachel, thank you for tolerating me and sharing those experiences with me. Thank you beyond words to my partner and now husband, Mitchell Breitbart. To all four of you, from the bottom of my heart, thank you for supporting me through the first three very hard years of my Ph.D., and the next four, which weren't easy either! I love you.

Thank you to the California Conservation Genomics Project, the National Institutes of Health Kirschstein NRSA Fellowship, the Museum of Vertebrate Zoology Graduate Student Fellowship, the American Society of Mammalogy, and the Department of Integrative Biology for funding and financial support. Thank you also to Hastings Natural History Reservation for supporting field work undertaken for my thesis.

I would like to end with a land acknowledgement. The research presented here is focused on native mammals of California and was conducted at the University of California, Berkeley, which sits on the territory of xučyun, the ancestral and unceded land of the Chochenyo speaking Ohlone people, the successors of the historic and sovereign Verona Band of Alameda County. I am deeply grateful for the opportunity to live and work here and recognize that I have benefitted from the use and occupation of this land.

Table of Contents

INTRODUCTION

Evolutionary genetics is an interdisciplinary field that weaves together elements of organismal biology and ecology with population genetics and molecular techniques. Similarly, my dissertation brings together several different questions and research topics. My goal as a biologist is to learn about how organisms interact with each other and the ecosystems around them, and how these ecological and environmental contexts drive evolution. I make inferences about evolutionary processes using genomic data gathered from wild-caught individuals from my own field work and museum collections.

The organisms and landscapes of California are a focal point of my dissertation research. California is home to more endangered species (n = 294) than any other state in the continental U.S. (U.S. Fish and Wildlife 2025), and the California Floristic Province has been designated a hotspot of both biodiversity and potential species loss (Mittermeier et al. 2004). For over one hundred years, the region has inspired natural historians and conservationists including Joseph Grinnell, Annie Alexander, and John Muir to establish institutions devoted to the documentation and preservation of California landscapes and their biodiversity.

Continuing in this tradition, the state government established two biodiversity initiatives in 2020: the California 30x30 Project and the California Conservation Genomics Project (California Natural Resources Agency 2022). The goal of the CCGP is to survey genomic diversity across the landscape of California and the variety of organisms native to the American west (Toffelmier et al. 2022). The CCGP is an umbrella for 153 'species projects', which are comprised of a genome assembly and whole-genome resequencing for 100 to 150 individuals from one or several closely related species (Shaffer et al. 2022). Each project is designed to follow a landscape genetics sampling scheme to maximize geographic coverage of the species' range (Manel et al. 2003; Balkenhol et al. 2015). All species projects will then be combined into a meta-analysis so that the CCGP can make conservation recommendations regarding which geographic regions will best preserve genetic uniqueness and resilience (Chambers et al. 2025).

Powered by the expansion of genome sequencing to non-model organisms, conservation genomics is entering a new multi-species era of ecosystem-level genetic diversity assessment (Thomassen et al. 2011; Formenti et al. 2022). However, large scale analyses inevitably miss the details and nuance that may be most useful for conservation managers working to protect individual species or local habitats (e.g. Scott et al. 2020). Hence, individual contributors within the CCGP are also tasked with more deeply exploring the rich data generated for their study systems (e.g. Wooldridge et al. 2024; Benham et al., in press).

Five of the six chapters presented here were conducted with funding and resources provided by the CCGP to the Museum of Vertebrate Zoology at the University of California, Berkeley to explore population genetics, phylogeography, and local adaptation in California rodents. The remaining chapter explores mating system variation and sexual selection in addition to other chapters' focus on natural selection and genetic drift. I explore these questions in *Peromyscus* deer mice, *Dipodomys* kangaroo rats, and *Thomomys* pocket gophers. However, before proceeding with any of the topics outlined above or approaches outlined below, working with genomic data for non-model organisms requires resource development. Thus, Chapters One,

Three and Five of my dissertation present new genome assemblies for the three organismal groups represented in my thesis, while Chapters Two, Four and Six investigate different questions regarding the evolutionary genetics of each group.

In Chapter Two, I compare gene expression and gene sequence evolution in the context of sexual selection and mating behavior. Comparing sequence evolution at synonymous and nonsynonymous sites is a tried-and-true method for identifying genes that are under purifying, relaxed, or positive selection (McDonald and Kreitman 1991; Nielsen and Yang 2003). Classic studies of gene sequence evolution found that genes involved in reproduction and immune function often show high numbers of nonsynonymous changes in protein-coding genes (Tanaka and Nei 1989; Metz and Palumbi 1996; Wyckoff 2000; Swanson and Vacquier 2002). Thirty years later, measuring DNA sequence evolution remains a powerful and unbiased way to identify targets of natural selection because it can be done across the genome without identifying candidate genes *a priori*. Across taxa, reproductive and immune genes still appear to evolve more rapidly than most other sequences in the genome (e.g., Bustamante et al. 2005; Shultz and Sackton 2019; Morales et al. 2025). However, measuring changes in gene sequence fails to account for changes in gene regulation, which is a second major component of molecular evolution (King and Wilson 1975).

Gene regulation provides an alternative mode of evolution that does not alter protein structure or function (Brawand et al. 2011). Resulting gene expression serves as a 'molecular' phenotype that represents an intermediate step between DNA sequence and cellular or organismal phenotype (Enard et al. 2002; Emilsson et al. 2008) and identifying specific genes that are expressed at a higher or lower level across different environments, tissues, or species can provide information about the ways organisms respond to the world around them (Whitehead and Crawford 2006). For example, gene expression evolution has been implicated in the evolution of heat stress tolerance in corals (Kenkel and Matz 2017), climactic adaptation in house mice (Mack et al. 2018; Bittner et al. 2021); and mating behavioral variation in in cichlid fishes (York et al. 2018). Here, I compare gene expression and gene sequence evolution in the male reproductive tract across three species of *Peromyscus* deer mice with variation in mating behavior to explore how sexual selection drives gene sequence evolution in comparison with gene expression evolution.

In the fourth and sixth chapters of my dissertation, I situate evolution in the context of the landscape by exploring how genetic variation is partitioned within and between populations and how patterns of variation compare between nonsynonymous and synonymous sites in the genome. These comparisons provide information about how populations have diverged and adapted across the landscape and how vulnerable they are to human-induced changes.

Conservation genetics took off in the 1990s when Kimura's neutral theory (1985) was applied to endangered species. If, as neutral theory predicts, genetic drift is relatively strong compared with natural selection in small, isolated populations, then mildly deleterious mutations are more likely to be fixed due to random chance (Lande 1994; Lynch et al. 1995). Relatedly, inbreeding depression occurs when deleterious recessive mutations are unmasked by mating between relatives (Crnokrak and Roff 1999). Both inbreeding depression and genetic load can result in reduced fitness for already vulnerable populations, further increasing their risk of extinction (Spielman et al. 2004).

Early conservation genetic studies of charismatic Florida panthers and Isle Royale wolves demonstrated the risks of inbreeding depression in endangered species (Roelke et al. 1993; Hedrick et al. 2014). Maintenance of sufficient genetic diversity to avoid 'mutational meltdown' due to inbreeding depression and genetic load has become a core tenet of endangered species conservation. However, thirty years of genetic data for endangered species has demonstrated that there is no simple relationship between conservation status and genetic diversity (DeWoody et al. 2021; Teixeira and Huber 2021; Schmidt et al. 2023). Some endangered species have high background genetic diversity and suffer from inbreeding depression, while others have extremely low genetic diversity and appear unaffected (Robinson et al. 2019; 2022). In Chapter Four, I focus on six species of kangaroo rats (*Dipodomys*) in California, three of which are endangered and three of which are widespread. I investigate how much genetic variability threatened species show compared with non-threatened species and ask whether inbreeding or mutational load appear to threaten vulnerable kangaroo rats.

In Chapter Six, I investigate the phylogeography of a non-threatened species, Botta's pocket gopher (*Thomomys bottae*), which was included in the California Conservation Genomics Project as a point of comparison for endangered species. Pocket gophers have many unique population genetic attributes that stem from their subterranean habitat specialization: gopher populations tend be isolated and disjunct, but locally dense (Patton and Feder 1981; Daly and Patton 1990). Allozyme studies suggest a history of incomplete lineage sorting and genetic drift within these structured populations (Patton and Yang 1977; Patton and Smith 1990). Further, *T. bottae* pocket gophers provide a useful contrast to endangered species because populations are isolated and fragmented but not endangered. In addition to exploring the population structure and phylogenetic relationships among pocket gophers from different regions of California, I measure inbreeding and genetic diversity within each population.

Kangaroo rats, pocket gophers, and deer mice all display unique behaviors, population genetics, evolutionary history, and degree of ecological specialization. In the course of my dissertation, I have used a combination of techniques including field work, laboratory work, and bioinformatic analysis to answer questions related to conservation genomics, phylogeography, and sexual selection. Ultimately, my goal is to combine these approaches to understand the consequences of genetic variation for organisms in the context of their environment.

CHAPTER 1

A highly contiguous genome assembly for *Peromyscus maniculatus* from California

ABSTRACT

The North American deer mouse, *Peromyscus maniculatus*, is widely distributed and one of the best-studied rodents in the world. Over the past 20 years, numerous studies have explored the physiology, ecology, and genetic basis of adaptive evolution in this species. In 2017, a reference genome was assembled for the eastern subspecies, *P. maniculatus bairdii*, and this has served as the basis for most studies of genetic variation. However, a reference genome from the western portion of the range would be useful given the species' extremely broad geographic distribution and extensive morphological variation. Although *P. maniculatus* is generally common, there are also several island populations of conservation concern in California. As part of the California Conservation Genomics Project (CCGP), we report the assembly of a reference genome for the western subspecies, *P. maniculatus sonoriensis*. This genome was sequenced with a combination of long-read PacBio and short-read Omni-C data and assembled *de novo*. The resulting assembly is 2.84 Gb in length and contains 1,838 scaffolds, with a scaffold N50 of 43.8 Mb and a BUSCO completeness score of 95.4%. This resource will assist scientists in answering questions related to deer mouse population genetics, systematics, structural genome evolution, epidemiology, developmental biology, and conservation.

1.1 INTRODUCTION

Peromyscus is a genus of small mouse-like rodents with 56 recognized species endemic to North America and northern Central America. This group has been an important focus of North American mammalogy for over a century, starting with the publication of W. H. Osgood's taxonomy of the genus in 1909 and continuing with the seminal work of Francis Sumner. In 1913, Sumner spent a year at UC Berkeley in the Museum of Vertebrate Zoology studying geographic variation among specimens collected by Joseph Grinnell and then brought deer mice into captivity in his La Jolla 'mouse house' at the Scripps Institute (Sumner 1923). Sumner documented variation in color and other quantitative traits among populations and performed crosses between them (Sumner 1917, 1918, 1929). Similar investigations were subsequently pursued by Dice (1933, 1940). Collectively, this work showed that deer mice are well suited for studies of both adaptation and speciation: deer mice can be easily bred in captivity, show remarkable morphological and genetic diversity, and vary in the extent of reproductive isolation between taxa. Further, their work influenced the early development of the field of population genetics. Dobzhansky, Wright, and Haldane drew from the empirical results of Sumner and Dice to explore the inheritance of quantitative traits in natural populations (Wright 1932; Dobzhansky 1937) and the strength of selection in the wild (Haldane 1948).

Following Sumner's lead, mammalogists brought six species of *Peromyscus* into the laboratory between the 1930s and 1980s (Joyner et al. 1998). Since then, researchers have

explored variation within and between different species of *Peromyscus* in a diverse array of traits including those involved in life history (Gubernick and Alberts 1987; Sohal et al. 1993), behavior (Dawson et al. 1988), immune function (Botten et al. 2000), and adaptation to high elevation (Snyder 1981; Chappell et al. 1988). These and other studies of the physiology, behavior, evolution, and disease ecology of *Peromyscus* prompted Dewey and Dawson (2001) to term the deer mouse the '*Drosophila* of North American mammalogy.' To this day, multiple captive populations are maintained at the Peromyscus Genetic Stock Center and several other laboratories (Bedford and Hoekstra 2015), and the genomes of five species have been sequenced (Table 1).

Of the different species in the genus, the most intensively studied is *Peromyscus maniculatus*. This species is widely distributed across most of North America and can be found in many different habitats (Fig. 1). In the past two decades, biologists have utilized laboratory and field work to study coat color (Hoekstra et al. 2006, Steiner et al. 2007; Linnen et al. 2013; Barrett et al. 2019; Wooldridge et al. 2022), tail length (Kingsley et al. 2017; 2024; Hager et al. 2022), burrowing behavior (Weber et al. 2013; Hu et al. 2022), biparental care (Turner et al. 2010; Bendesky et al. 2017), postcopulatory sexual selection (Turner and Hoekstra 2006; Fisher et al. 2016; Meléndez-Rosa et al. 2019; Voss and Nachman 2024), adaptation to high altitude (Storz et al. 2007; 2009; 2010; Cheviron et al. 2012; 2014; Natarajan et al. 2015; Scott et al. 2015), adaptation to deserts (MacManes and Eisen 2014, Kordonowy and MacManes 2017; Tigano et al. 2020; Colella et al. 2021) and adaptation to urban environments (Harris et al. 2013; Harris and Munshi-South 2017). Lassance and Hoekstra (2017) assembled the first reference genome for *P. maniculatus* from a captive individual originating from the eastern subspecies *P. m. bairdii* in Ann Arbor, Michigan. This chromosome-level reference genome has enabled scientists to pinpoint specific genes, loci, and inversions that underpin natural variation in *Peromyscus* (e.g., Harringmeyer and Hoekstra 2022; Wooldridge et al. 2022).

Given the extremely broad range and extensive variation observed in *P. maniculatus,* a reference genome from a distinct part of the species range would be a useful resource. Moreover, while the species is generally common, there are several subspecies of conservation concern. As part of the California Conservation Genomics Project (CCGP), we present a new reference genome for a wild-caught individual collected in Kern County, California, which falls within the range of the western subspecies *P. maniculatus sonoriensis*. This genome was sequenced using long-read PacBio and short read Omni-C data and will serve as the reference for a landscape genetics study of *Peromyscus* across California, including the endangered Channel Islands subspecies *P. m. anacapae* and *P. m. clementis*.

1.2 METHODS

Sample collection

One male *Peromyscus maniculatus sonoriensis* was caught by JLP in a Sherman trap at Cameron Creek (N 35.09131, W 118.30937) in the Tehachapi Mountains, Kern County, California, USA (permit DS-192560001, California Department of Fish and Wildlife). The animal was sacrificed following the guidelines of the American Society of Mammalogists (Sikes et al. 2016), tissue was flash-frozen, and a voucher specimen (skin and skull) was deposited in the mammal collection of the UC Berkeley Museum of Vertebrate Zoology (MVZ:Mamm:240117; JLP collection number 29048).

DNA extraction

We used the Nanobind Tissue Big DNA kit in accordance with the manufacturer's instructions (Pacific BioSciences - PacBio, Menlo Park, CA) to extract high molecular weight (HMW) genomic DNA (gDNA) from 50 mg of liver tissue. After purification with phenol-chloroform, we measured the quality of the extracted HMW DNA on the NanoDrop ND-1000 spectrophotometer in terms of absorbance ratios (260/280 = 1.82 and 260/230 = 2.02), and used a Quantus Fluorometer (QuantiFluor ONE dsDNA Dye assay; Promega, Madison, WI) to gauge DNA yield (21 μg). Eighty-four percent of the fragments were 150 kilobases (kb) or longer (Femto Pulse system, Agilent, Santa Clara, CA).

DNA Sequence Library Preparation

To create a HiFi SMRTbell library, we used the SMRTbell Express Template Prep Kit v2.0 (PacBio) with HMW gDNA sheared to 15–18 kb using Diagenode's Megaruptor 3 system (Diagenode, Belgium; cat. B060100003). We concentrated the sheared gDNA with 0.45X of AMPure PB beads (PacBio) and removed single-strand overhangs at 37˚C for 15 minutes, followed by further enzymatic steps including DNA damage repair at 37˚C for 30 minutes, end repair and A-tailing at 20˚C for 10 minutes and 65˚C for 30 minutes, and ligation of overhang adapter v3 at 20˚C for 60 minutes. We purified and concentrated the library with 1X AMPure PB beads, treated it with nuclease at 37˚C for 30 minutes, and selected fragments greater than 7 kb on the PippinHT system (Sage Science, Beverly, MA), resulting in a 15–20 kb library. Sequencing on the PacBio Sequel IIe machine was carried out at UC Davis DNA Technologies Core (Davis, CA) using four 8M SMRT cells, Sequel II sequencing chemistry 2.0, and 30-hour movies.

We prepared the Omni-C library using the Omni-C Kit (Dovetail Genomics, Scotts Valley, CA) following the manufacturer's protocol with some modifications. The liver (ID: JLP29048) was chilled with liquid nitrogen and simultaneously ground with mortar and pestle. Subsequently, chromatin was fixed in place in the nucleus and large debris was removed with 100 μm and 40 μm cell strainers. We then carried out DNase I digests under various conditions until the distribution of DNA fragment lengths was appropriate. Chromatin ends were repaired and ligated to a biotinylated bridge adapter followed by proximity ligation of adapter containing ends. Crosslinks were then reversed, the DNA was purified from proteins and treated to remove biotin that was not internal to ligated fragments. We constructed the Illumina library using the NEB Ultra II DNA Library Prep kit (New England Biolabs, Ipswich, MA) with a compatible y-adaptor. Streptavidin beads were used to capture biotin-containing fragments. To preserve library complexity, prior to PCR enrichment we split the library into two replicates each receiving unique dual indices. The Vincent J. Coates Genomics Sequencing Lab (Berkeley, CA) seqenced the library on the NovaSeq 6000 platform (Illumina, CA) resulting in approximately 100 million 2 x 150 bp read pairs per gigabase (Gb) of genome size.

Nuclear genome assembly

After trimming remnants of sequence adaptors with HiFiAdapterFilt (Sim et al. 2022), we assembled the genome following the CCGP pipeline Version 4.0 (Supplementary Table 2). PacBio HiFi reads and Omni-C data were used to assemble with minimal manual curation. The initial dual (partially phased) diploid assembly was generated from HiFi and Omni-C data in HiFiasm (Cheng et al. 2021, Cheng et al. 2022). We then aligned the Omni-C data to each assembly with the Arima Genomics Mapping Pipeline and scaffolded them with SALSA

(Ghurye et al. 2017, 2019). We generated Omni-C contact maps for both assemblies by aligning the Omni-C data with BWA-MEM (Li 2013), identified ligation junctions, and generated Omni-C pairs in pairtools (Open2C et al. 2024). A multi-resolution Omni-C matrix was produced in cooler (Abdennur and Mirny 2020) and balanced with hicExplorer (Ramírez et al. 2018). We checked for major mis-assemblies by visualizing the contact maps in HiGlass (Kerpedjiev et al. 2018) and the PretextSuite (https://github.com/wtsi-hpag/PretextView). If we identified a strong off-diagonal signal in the proximity of a join made by the scaffolder, and a lack of signal in the consecutive genomic region, we marked the join. Afterwards, all marked joins were dissolved by cutting the scaffolds at the join coordinates. After this process, no further manual joins were made. Some of the remaining gaps (joins) were closed using the PacBio HiFi reads and YAGCloser (https://github.com/merlyescalona/yagcloser). Finally, we screened for contamination with the BlobToolKit (Challis et al. 2020).

We counted k-mers in the PacBio reads with meryl (https://github.com/marbl/meryl) and used the resulting database to assess base level accuracy (QV) and k-mer completeness in merqury (Rhie et al. 2020). Genome size and heterozygosity were estimated from the k-mer database in GenomeScope2.0 (Ranallo-Benavidez et al. 2020). We calculated contiguity metrics with QUAST (Gurevich et al. 2013), evaluated genome completeness using BUSCO (Manni et al. 2021) with the Glires ortholog database (glires_odb10; 13,798 genes), and estimated genome assembly accuracy via BUSCO gene set frameshift analysis (Korlach et al. 2017). We considered output haplotype 1 to be the primary assembly based on genome quality metrics and BUSCO completeness.

Mitochondrial genome assembly

We were unable to assemble the mitogenome from HiFi long reads because mitochondrial DNA was over-digested during HMW DNA extraction, resulting in the absence of long mitochondrial DNA fragments. The same problem was observed in other rodent genomes sequenced by the CCGP. Instead, we mapped short-read Omni-C Illumina data to the *P. m. bairdii* reference to assemble a *P. m. sonoriensis* mitogenome (Lassance and Hoekstra 2017; HU_Pman_2.1; NCBI: NC_039921.1). Reads were trimmed with Trim Galore (Andrews 2010; Martin 2011), duplicates removed with Dedupe (Duplicate Read Remover), and nuclear encoded mitochondrial pseudogenes (numts) filtered using Numt Parser (de Flamingh et al. 2023). Numt references were characterized by including all BLAST (Camacho et al. 2009) hits for the *P. m. bairdii* mitogenome against the nuclear assembly (NCBI: GCF_003704035.1). We mapped filtered reads to the *P. m. bairdii* mitogenome in Geneious Prime (Kearse et al. 2012). The resulting mitogenome was annotated with MITOS2 (Bernt et al. 2013).

Genome annotation

To identify repetitive sequences in the genome, we used RepeatModeler2 (Flynn et al. 2020) with LTRStruct to detect Long Terminal Repeats (LTRs) (Ellinghaus et al. 2008) and DeepTE to classify interspersed repeats (Yan et al. 2020). To prevent masking of protein-coding regions, we used transposonPSI (Haas 2010) and removed sequences homologous to the *P. m. bairdii* proteome before collapsing redundant repetitive sequences with *cd-hit-est* (Li and Godzik 2006). We then annotated the genome sequence with RepeatMasker under default settings in three iterations: 1) simple repeats; 2) predicted interspersed repeats identified with RepeatModeler2 above; and 3) curated *P. m. bairdii* and other Rodentia repeats from Dfam

(Osmanski et al. 2023). Lastly, we used scripts from ParseRM (Kapusta et al. 2017) to identify nested transposable elements and estimate the age of repeats.

We annotated gene features in the repeat-masked *P. m. sonoriensis* draft sequence by lifting over genes from the NCBI annotation for *P. m. bairdii* (GCF_003704035.1; NCBI Eukaryotic Annotation Pipeline) with Liftoff (Shumate and Salzberg 2021), including steps to polish intron/exon boundaries (*-polish*) and detect gene duplicates (*-copies*).

Re-scaffolding with the P. m. bairdii reference

Following *de novo* genome assembly, we wanted to take advantage of existing genetic resources in *Peromyscus* to improve the contiguity of our assembly and explore potential instances of structural evolution. To assign scaffolds to chromosomes, we re-scaffolded our *P. m. sonoriensis* genome against the chromosome-level *P. m. bairdii* assembly (RagTag, Alonge et al. 2022). We then filtered out low quality alignments with a C-score of less than 0.1, grouped alignments into blocks of 100 genes, and performed a Quota merger of syntenic blocks to account for structural variation between genomes (Tang et al. 2008). To visualize the impact of re-scaffolding on our genome assembly, we aligned the *P. m. sonoriensis* genome to the *P. m. bairdii* genome before and after RagTag re-scaffolding with LAST (Kiełbasa et al. 2011). To evaluate synteny between *P. m. sonoriensis* and other *Peromyscus* species, we also aligned the re-scaffolded (v.1.1) genome to the published genome assemblies for *P. leucopus* (Long et al. 2019) and *P. californicus* (Trainor et al. 2022).

1.3 RESULTS

The PacBio HiFi sequencing libraries yielded 6.86 x 10^6 reads, resulting in 32.2-fold coverage (N50 read length 12,987 bp; minimum read length 74 bp; mean read length 12,656 bp; maximum read length of 53,663 bp), assuming a genome size of 2.7 Gb (estimated by Genomescope2.0). The reads had a 0.139% sequencing error rate and nucleotide heterozygosity of 0.016. The k-mer spectrum is bimodal, with peaks at 16 and 31, where peaks correspond to heterozygous and homozygous states of a diploid species, respectively (Fig. 2A). The observed profile, with a higher peak at 16 than at 31, suggests a highly heterozygous species.

The final sequence (mPerMan1.0) consists of two partially phased assemblies (*sensu* Cheng et al. 2021), tagged as primary and alternate, both similar in size to the value estimated by Genomescope2.0 (Fig. 3A). These assemblies are not maternal and paternal haplotypes in the strict sense and are likely to contain haplotigs and switch errors. The primary assembly is 2.85 Gb long and comprises 1,838 scaffolds with an N50 of 43.8 Mb. The BUSCO Glires completeness score is 95.4%, with per-base quality (QV) of 61.8, k-mer completeness of 79.67% and frameshift indel QV of 41.9. During manual curation, 28 of the SALSA joins were identified as mis-joins and broken, and 14 gaps were closed in each of the two assemblies. For detailed assembly statistics see Table 2, Fig. 2 (primary) and Supplementary Fig. 1 (alternate).

The final *P. m. sonoriensis* mitochondrial assembly included 14,764 mapped Illumina reads, with a mean coverage of 136.7 (range: 5 - 213). The mitogenome is 16,324 bp long, with a base composition of A= 34.8%, C = 23.8%, G =12.8%, T = 28.6%, and consists of 22 unique transfer RNAs and 13 protein-coding genes. Liftoff nuclear gene annotation identified 21,837 of 22,079 protein-coding genes (98.9%) present in the *P. m. bairdii* reference. 19,153 of these had at least one valid open reading frame. We also annotated and masked repetitive elements and

identified 51.86% of the genome as repetitive: 21.3% of the genome consisted of Long Terminal Repeat sequences, 12.8% Long Interspersed Nuclear Elements, 11.9% Short Interspersed Nuclear Elements, and 0.89% Class II DNA transposable elements. The remainder was a combination of simple and unclassified repetitive elements (Supplementary Table 2). We observed a major pulse of Long Terminal Repeat (LTR) activity between three and five million years ago, whereas LINE and SINE elements were most active closer to ten million years ago (Fig. 3).

RagTag assembly-guided scaffolding assigned 1,612 of 1,838 scaffolds to 24 chromosomes from the *P. m. bairdii* assembly for a final 1,039 scaffolds in mPerMan1.1. The N50 increased from 43.8 Mb to 111 Mb, comparable with the *P. m. bairdii* N50 of 115 Mb (Supp. Fig. 2). A comparison of LAST alignments of *P. m. sonoriensis* versus *P. m. bairdii* before and after re-scaffolding demonstrates increased synteny and linearity between the genomes (Supp. Figs. 3, 4), while alignments with *P. leucopus* and *P. californicus* suggest several potential chromosomal rearrangements since these species diverged (Supp. Figs. 5,6,7). The re-scaffolded assembly is available on the Dryad data repository.

1.4 Discussion

We report a *de novo* genome assembly for the western deer mouse *Peromyscus maniculatus sonoriensis* comprising 1,838 scaffolds with a scaffold N50 of 43.8 Mb and a BUSCO completeness score of 95.4%. Though the sequencing data underlying our *de novo* assembly are insufficient to reach chromosome-level resolution, we addressed this limitation by re-scaffolding the *P. m. sonoriensis* assembly against the published reference genome for *P. m. bairdii* to assign scaffolds to chromosomes, reaching a scaffold N50 of 111 Mb for mPerMan v1.1. We hope this genome will be a useful resource for studies of *Peromyscus* population genetics, systematics, structural evolution, development, epidemiology, behavior, and conservation.

Relationships within *P. maniculatus* and among the 56 species in the *Peromyscus* genus have mostly been defined with morphological and mitochondrial characters (e.g., Bradley et al. 2007; Platt et al. 2015; Greenbaum et al. 2019), and some fundamental aspects of deer mouse systematics remain unresolved. For example, the taxonomic status and relationship between the western subspecies *P. m. gambelii* and *P. m. sonoriensis* are poorly understood. The individual whose genome was sequenced here was sampled from a region traditionally associated with *P. m. sonoriensis* (Hall and Kelson 1959), but mitochondrial DNA studies suggest that the *P. m. gambelii* clade extends throughout southern California (Greenbaum et al. 2019; Boria and Blois 2023).

Current work funded by the California Conservation Genomics Project may be well-placed to clarify these relationships. The CCGP has generated sequence data for 96 individuals from across the state, which, combined with published data, will enable scientists to more accurately identify clades and geographic boundaries between *P. maniculatus* lineages. Despite morphological assignment to *P. m. sonoriensis,* preliminary whole genome sequencing results unambiguously assign the reference individual to *P. m. gambelii* (T. Herrera, pers. comm.). However, further work is needed to confirm whether taxonomic updates are warranted.

Additional genomic data, including structural rearrangements such as inversions, may also be useful in exploring relationships among *Peromyscus* lineages, and new results suggest that structural genomic rearrangements are common within *P. maniculatus* (Gozashti et al. 2025). To explore this possibility, we aligned the re-scaffolded *P. m. sonoriensis* genome to the *P. m. bairdii, P. leucopus,* and *P. californicus* reference genomes (Lassance and Hoekstra 2017; Long et al. 2019; Trainor et al. 2022). We recovered a known inversion on chromosome 15 that drives adaptive phenotypic variation between *P. maniculatus* forest and prairie ecotypes (Hager et al. 2022), as well as putative inversions on chromosomes 6, 10 and 22 (Supp. Fig. 4). We also annotated repetitive regions of the genome according to transposable element class and age. We observed a spike in LTR activity between 3 and 5 million years ago (Fig. 3), in keeping with a recent study reporting that endogenous retrovirus (ERV)-type LTR elements have been unusually active in the *P. maniculatus* lineage in the last five to ten million years (Gozashti et al. 2023; Osmanski et al. 2023).

While we have performed preliminary analyses regarding structural evolution in *Peromyscus*, our main objective in assembling this genome is to contribute to the California Conservation Genomics Project's goal of identifying shared hotspots of genetic diversity and geographic barriers to gene flow in the state (Shaffer et al. 2022; Toffelmier et al. 2022). Conservation genetic research often focuses on narrowly distributed species or small, vulnerable populations, but widespread species such as *P. maniculatus* provide necessary context regarding the spatial distribution of genetic diversity across the landscape and the continuum of adaptive potential in the face of human-induced climate change. For example, *P. maniculatus* has expanded its range to higher elevations in the Sierra Nevada since the early 1900s (Rowe et al. 2015), but little is known about how the population genetics of the species has changed with warming temperatures and shifting habitat availability (e.g., Bi et al. 2019; Benham et al. 2024). The CCGP has funded genome assembly and broad geographic whole genome resequencing for several widespread species, including *P. maniculatus*, the western fence lizard *Sceloporus occidentalis* (Bishop et al. 2023), and the California quail *Callipepla californica* (Benham et al. 2023). Exploring landscape genetic diversity in organisms across the spectrum of range size and conservation risk will allow us to better understand when and where conservation actions should focus on genetic diversity versus other aspects of biodiversity protection and recovery.

1.5 FIGURES

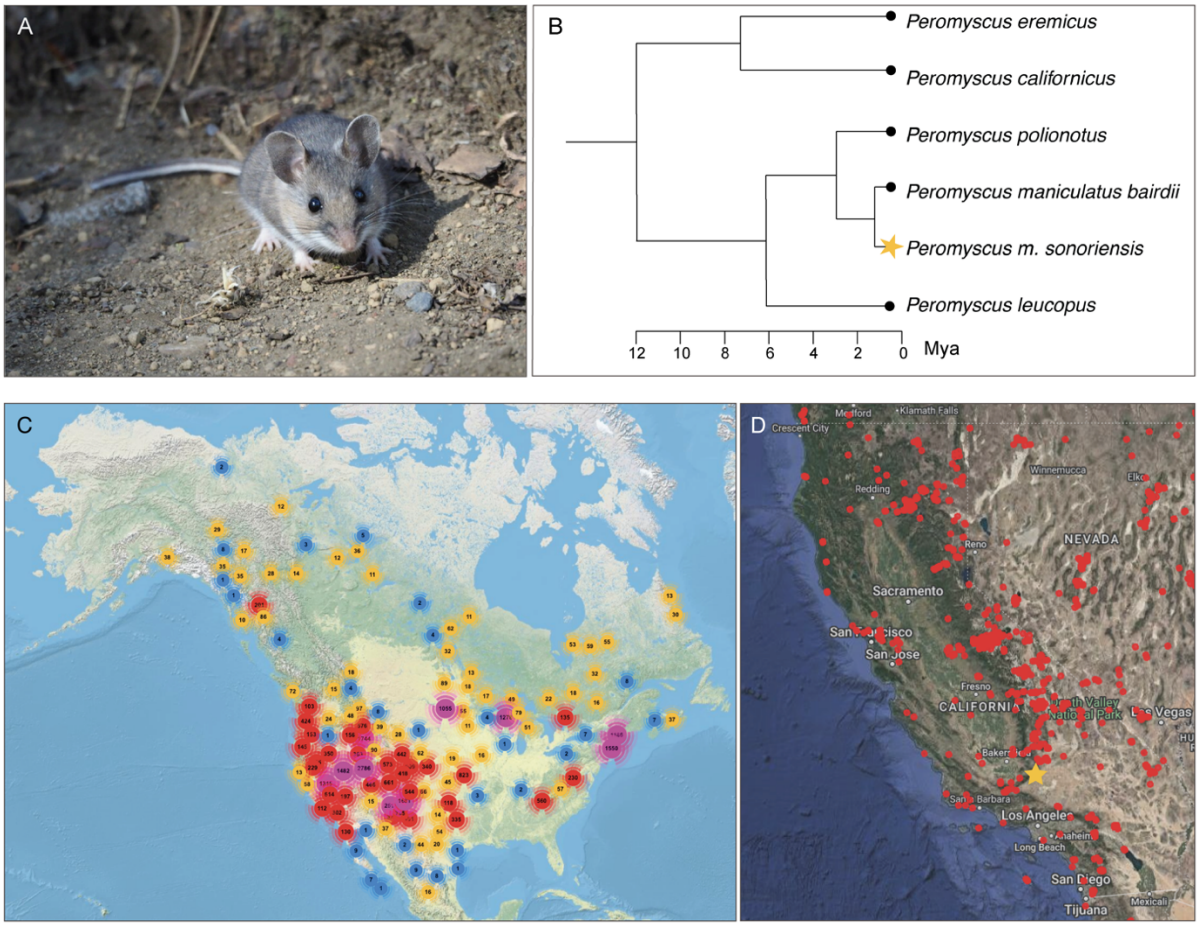


Figure 1. A) *Peromyscus maniculatus sonoriensis* (Photo: Fern Wexler via iNaturalist). B) Scaffolded genome assemblies of *Peromyscus* available publicly as of April 2025, including the *P. m. sonoriensis* assembly described here (indicated with a star). Phylogenetic tree adapted from Upham et al. (2019). C) Holdings of *Peromyscus maniculatus* tissues in 39 mammalogy collections catalogued by ARCTOS (accessed 6/21/2023; Supp. Table 1). D) *P. maniculatus* samples from California accessioned at the Museum of Vertebrate Zoology, UC Berkeley. The star indicates the individual whose genome was sequenced here (MVZ:Mamm:240117).

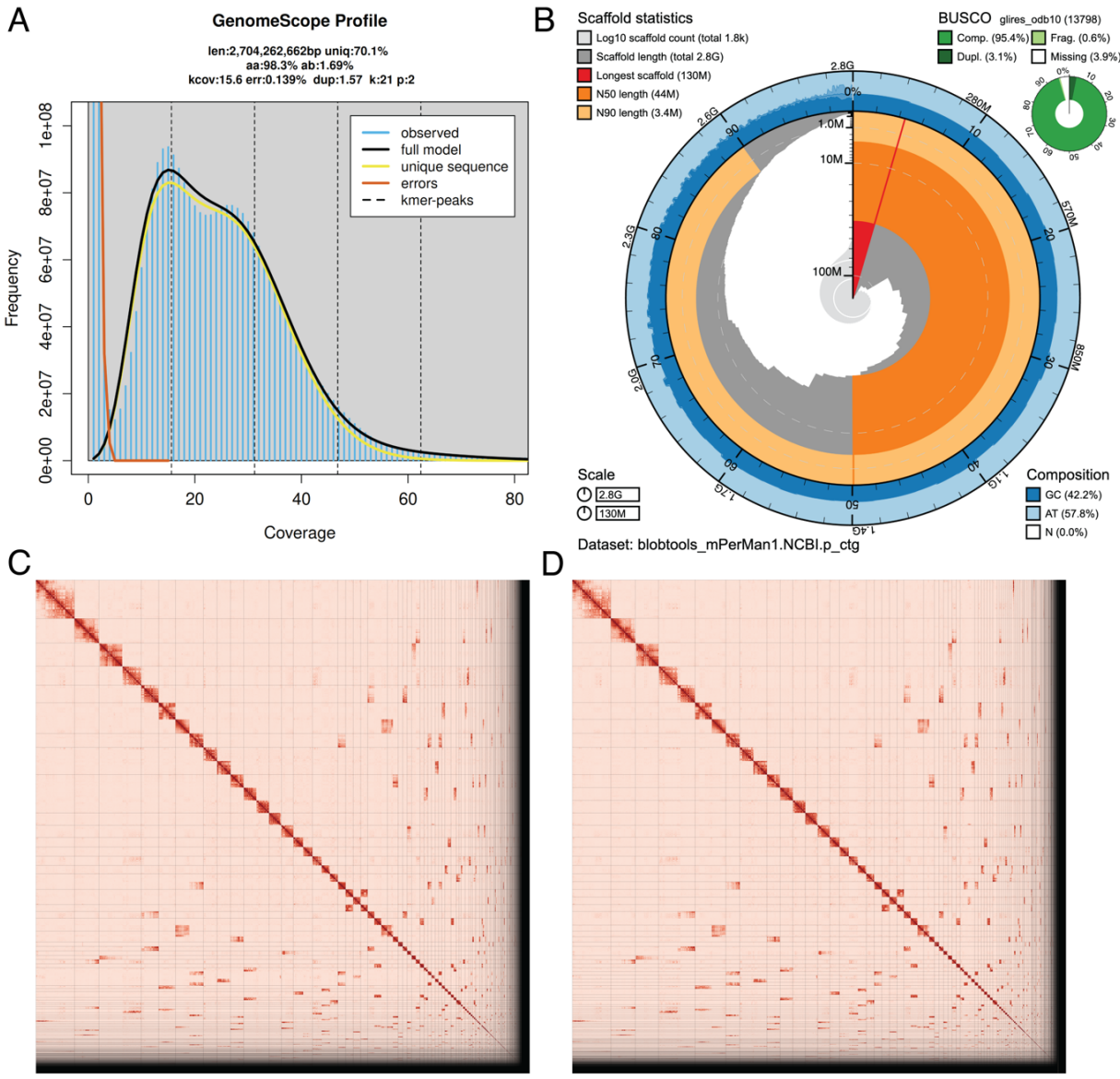


Figure 2. Quality assessment of the final genome assembly of *Peromyscus maniculatus sonoriensis* (mPerMan1.0). (A) The bimodal distribution of the k-mer spectrum estimated from the adapter-trimmed PacBio HiFi reads is an expected property of a diploid genome. (B) Snail plot of quality metrics. The circumference represents the length of the assembly: scaffolds are drawn clockwise in order of size, and the red line indicates the longest. The middle arcs represent N50 (dark orange) and N90 (light orange). Completeness of the BUSCO core gene set assembly is shown in the top right panel. (C) PretextSnapshot Omni-C contact maps of the primary (mPerMan1.0.p) and alternate (D; mPerMan1.0.a) assemblies. Every cell represents data supporting linkage between genomic regions found in proximity in 3D space.

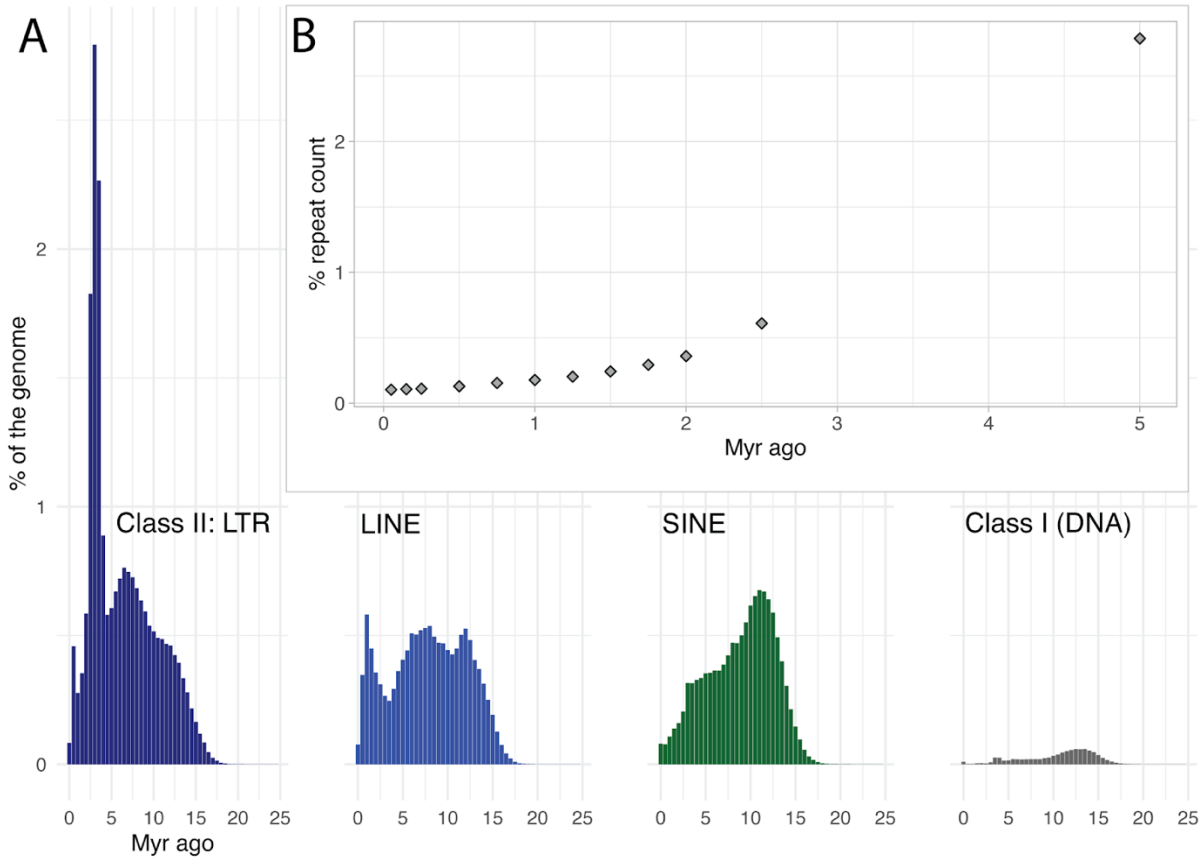


Figure 3. Patterns of repeatable element accumulation in the genome of *P. m. sonoriensis*. (A) Accumulation landscapes of the major transposon subfamilies since the divergence of *Peromyscus* from *Onychomys* approximately 21 MYA (Upham et al. 2019). (B) Percentage of all interspersed repeatable elements evolved over the timescale of evolution of the *P. maniculatus* species group.

1.6 TABLES

Table 1. Assembly metrics for six *Peromyscus* genome assemblies available on NCBI GenBank as of April 2025.

Common name	Scientific name	Length	# Scaffolds	Scaff. N50	Scaff. L50	BUSCO score	GC content	Citation (GenBank ID)
North American deer mouse (prairie)	*Peromyscus maniculatus bairdii*	2.5 Gb	8,523*	115.0 Mb	9	98.2%	42.28 %	Lassance and Hoekstra 2017 GCF_003704035.1
North American deer mouse (western)	*Peromyscus maniculatus sonoriensis*	2.8 Gb	1,837	43.8 Mb	21	95.4%	42.19 %	This study GCA_026229955.1
White-footed mouse	*Peromyscus leucopus*	2.5 Gb	1,856*	114.3 Mb	9	97.6%	42.23 %	Long et al. 2019 GCF_004664715.2
Oldfield mouse	*Peromyscus polionotus*	2.6 Gb	5,292*	117.6 Mb	9	98.3%	42.20 %	Lassance and Hoekstra 2018 GCA_003704135.2
California mouse	*Peromyscus californicus*	2.5 Gb	237	115.0 Mb	9	98.7%	42.21 %	Trainor et al. 2022 GCF_007827085.1
Cactus mouse	*Peromyscus eremicus*	2.9 Gb	4,296*	97.2 Mb	11	97.3%	42.40 %	Keane 2023 GCF_949786415.1

*contains 24 assembled chromosomes and additional unplaced scaffolds.

Table 2. Genome sequencing and assembly quality metrics for *Peromyscus maniculatus sonoriensis*.

Metric	Value	
Genome Sequence		
PacBio HiFi reads		
Run	1 PACBIO_SMRT (Sequel II) run: 6.9M spots, 86.9G bases, 52.1Gb	
Accession	SRX19054208	
Omni-C Illumina reads		
Run	2 ILLUMINA (Illumina NovaSeq 6000) runs: 244.8M spots, 73.9G bases, 23.7Gb	
Accession	SRX19054209, SRX19054210	
Assembly identifier (Quality code*)	mPerMan1 (7.7.P7.Q61.C55)	
HiFi Read coverage §	34.76X	
Genome Assembly Quality Metrics	**Primary**	**Alternate**
Number of contigs	2,082	1,235
Contig N50	10,714,300 bp	12,308,238 bp
Contig NG50§	11,246,172 bp	11,942,166 bp
Longest contig	53,767,646 bp	108,452,771 bp
Number of Scaffolds	1,838	1,006
Scaffold N50	43,824,230 bp	63,956,138 bp
Scaffold NG50§	43,824,230 bp	63,956,138 bp
Largest scaffold	129,284,778 bp	204,820,094 bp
Size of final assembly	2,847,741,276 bp	2,625,495,430 bp
Phased block NG50§	11,678,182 bp	12,308,238 bp
Gaps per Gbp (#Gaps)	86 (244)	87 (229)
Indel QV (Frameshift):	41.9031	41.9550
Base pair QV	61.7598	61.6518
	Full assembly = 61.7077	
k-mer completeness	79.6659	76.3852
	Full assembly = 99.3783	
BUSCO score‡ (glires_odb10)		
Complete genes	95.40%	93.80%
Complete: single copy	92.30%	91.50%
Complete: duplicated	3.10%	2.30%
Fragmented	0.60%	0.70%
Missing	4.00%	5.50%
Organelles	1 complete mitochondrial assembly (17,626 bp)	

***** Assembly quality code x.y.P.Q.C derived notation, from (Rhie et al. 2021). x = log10[contig NG50]; y = log10[scaffold NG50]; P = log10 [phased block NG50]; Q = Phred base accuracy QV (Quality value); C = % genome represented by the first 'n' scaffolds, following a known karyotype for *P. maniculatus* of 2n = 48 (Deaven et al. 1977). Quality code for all the assembly denoted by primary assembly (mPerMan1.0.p). § Read coverage and NGx statistics have been calculated based on the estimated genome size of 2.6 Gb. ‡ (P)rimary and (A)lternate assembly values.

CHAPTER 2

Mating system variation and gene expression in the male reproductive tract of *Peromyscus* mice

This chapter has been previously published and is reproduced here in accordance with the journal's article sharing policy:

Voss, ER, Nachman MW. 2024. Mating system variation and gene expression in the male reproductive tract of *Peromyscus* mice. *Molecular Ecology* 2024:e17433:1-17.

DOI: 10.1111/mec.17433

ABSTRACT

Genes involved in reproduction often evolve rapidly at the sequence level due to postcopulatory sexual selection (PCSS) driven by male-male competition and male-female sexual conflict, but the impact of PCSS on gene expression has been under-explored. Further, though multiple tissues contribute to male reproductive success, most studies have focused on the testes. To explore the influence of mating system variation on reproductive tract gene expression in natural populations, we captured adult males from monogamous Peromyscus californicus and polygynandrous P. boylii and P. maniculatus. We generated RNAseq libraries, quantified gene expression in the testis, seminal vesicle, epididymis, and liver, and identified 3627 mating system-associated differentially expressed genes (MS-DEGs), where expression shifted in the same direction in P. maniculatus and P. boylii relative to P. californicus. Gene expression variation was most strongly associated with mating behaviour in the seminal vesicles, where 89% of differentially expressed genes were MS-DEGs, including the key seminal fluid proteins Svs2 and Pate4. We also used published rodent genomes to test for positive and relaxed selection on Peromyscus-expressed genes. Though we did not observe more overlap than expected by chance between MS-DEGs and positively selected genes, 203 MS-DEGs showed evidence of positive selection. Fourteen reproductive genes were under tree-wide positive selection but convergent relaxed selection in P. californicus and Microtus ochrogaster, a distantly related monogamous species. Changes in transcript abundance and gene sequence evolution in association with mating behaviour suggest that male mice may respond to sexual selection intensity by altering aspects of sperm motility, sperm-egg binding and copulatory plug formation.

2.1 INTRODUCTION

Genes involved in reproduction often evolve rapidly due to male-female sexual conflict and male-male competition, collectively referred to as "postcopulatory sexual selection" (Fig.

1A; Eberhard 1996, reviewed in Birkhead and Pizzari 2002; Swanson and Vacquier 2002). Many studies of the rapid evolution of reproductive genes have focused on sequence evolution of genes involved in sperm production (e.g., Torgerson et al.; 2002; Dorus et al. 2010), but in most animals, the male reproductive tract includes accessory glands that produce additional proteins for transfer to females during mating. For example, many *Drosophila* seminal fluid proteins (Sfps) have extremely high rates of non-synonymous to synonymous substitutions (i.e. high d_N/d_S values; Begun et al., 2000; Swanson et al., 2001a; Haerty et al., 2007; Wigby et al., 2020), suggesting that they are under positive selection. Experimental work has shown that Sfps play key roles in sperm storage (Bertram et al., 1996, Avila and Wolfner 2017) and sperm competition (Clark et al., 1995; Patlar and Civetta 2022). Sfps create a mating plug that may prevent female remating (Lung and Wolfner 2001; Brown et al., 2023) and can temporarily alter female mating receptivity and egg laying (Heifetz et al., 2000; Prout and Clark 2000; Findlay et al., 2014). *Drosophila* female reproductive proteins, such as egg-surface proteins that interact with sperm and regulate fertilization, also evolve rapidly (Swanson et al., 2004; McDonough-Goldstein et al., 2021).

The patterns of sexual selection and rapid reproductive protein evolution seen in *Drosophila* have been observed across a broad range of organisms from abalone and sea urchins to humans (e.g., Metz and Palumbi 1996; Yang et al., 2000; Vicens et al., 2014). As in *Drosophila*, mammals have a complex male reproductive tract (Fig. 1B) that includes the testes as well as the epididymis, which plays roles in sperm storage and maturation (Dean et al., 2008; Sullivan and Mieusset 2016), and accessory glands such as the seminal vesicle, prostate gland, and bulbourethral gland, which contribute carbohydrates, lipids, and proteins to the seminal fluid (Linzey and Lane 1969; Ramm et al., 2005; Claw et al., 2018). In some species, mammalian seminal fluid forms a copulatory plug when it coagulates in the vagina (Voss 1979; Schneider et al., 2016; Sutter et al., 2016), but can also mediate sperm-female reproductive tract interactions (Kawano et al., 2014; Noda and Ikawa 2019), stimulate female immune responses to mating (Sharkey et al., 2007; Tollner et al., 2011; Schjenken and Robertson 2020), and interact with the neuroendocrine system to alter the timing of ovulation (Ratto et al., 2012; Berland et al., 2016).

Variation in mating behavior can modulate the strength of postcopulatory sexual selection. Studies of gene sequence evolution in primates (Wyckoff et al. 2000; Dorus et al. 2004) and rodents (Ramm et al., 2008; 2009) found a positive correlation between the rate of evolution of sperm protamines and semenogelin and the level of female polyandry and male-male competition. The strength of postcopulatory sexual selection also affects reproductive function and physiology. In species with sperm competition, males have larger testes and seminal vesicles (Harcourt et al., 1995; Ramm et al., 2005) and produce larger, faster sperm with more mitochondria (Snook 2005; Anderson et al., 2005; Gomendio and Roldan 2008; Tourmente et al., 2013; Lüpold et al., 2020).

Although most previous studies of the evolution of reproductive genes in mammals have focused on nonsynonymous changes in protein-coding gene sequences, changes in gene expression are known to underlie much of evolution (King and Wilson 1975, Carroll 2005). Differential expression provides an alternative route to phenotypic change that may avoid the negative pleiotropic effects that can accompany gene sequence evolution (Brawand et al., 2011; Vicens et al., 2017).

Mice in the genus *Peromyscus* provide an excellent opportunity to study the influence of mating system on both gene expression and protein sequence evolution for male reproductive

genes. Polygynandry (loosely affiliative multimale – multifemale mating system) is the ancestral state in *Peromyscus*, and different species in this genus exhibit broad variation in mating ecology (Bedford and Hoekstra 2015; Meléndez-Rosa et al., 2019; Weber and Fisher 2023). Here, we compare gene expression and gene sequence evolution across the male reproductive tract in three species of *Peromyscus* (Fig. 1C). Using wild-caught, reproductively mature mice, we quantified gene expression in the testis, seminal vesicle, epididymis, and liver in monogamous *P. californicus* (PECA), polygynandrous *P. maniculatus* (PEMA) and polygynandrous *P. boylii* (PEBO) (Birdsall and Nash 1973; Ribble 1991; Ribble and Stanley 1998).

We address five major questions. First, are different patterns of gene expression associated with different mating systems? Second, which male reproductive tissues show the greatest changes in gene expression in different mating systems? Third, which specific genes are expressed at very high levels in polygynandrous species and at low levels in monogamous species? Since monogamy is the derived condition and does not involve sperm competition, the identification of highly expressed genes in polygynandrous species that are lowly expressed in monogamous species may identify genes that are particularly important in postcopulatory sexual selection. Fourth, what phenotypes are associated with the most rapidly evolving male reproductive genes? And finally, which genes, if any, show both rapid protein evolution and changes in gene expression associated with different mating systems?

We discovered many genes that show differences in expression associated with mating system, most notably in the seminal vesicles. The most rapidly evolving genes included those involved in sperm motility and sperm-egg binding, though for the most part, rapidly evolving genes did not overlap with mating-system associated differentially expressed genes. Together, these results illustrate the distinct roles that gene expression and protein sequence evolution play in male reproductive tissues and identify a set of genes likely to be important in postcopulatory sexual selection.

2.2 METHODS

Field Methods

To explore gene expression variation in the context of naturally occurring postcopulatory sexual selection, we trapped reproductively mature *P. californicus californicus* and *P. boylii rowleyi* at Hastings Natural History Reserve (Carmel Valley, CA), and *P. maniculatus gambellii* at the Field Station for the Study of Behavior Ecology and Reproduction (Berkeley, CA). Mice were captured using Sherman traps following trapping procedures outlined in the American Society of Mammalogists Guidelines (Sikes et al., 2016).

Scrotal adult males were sacrificed via isoflurane and cervical dislocation. To control for circadian variation in gene expression, all mice were sacrificed and all tissues were sampled between 9 am and 1 pm. First, we took standard body measurements (total length, tail length, hind-foot length, ear length, and weight) and recorded testis length, width, and mass. Tissues were sampled in a consistent order within 10 minutes of euthanasia to minimize RNA degradation and were preserved in RNAlater before transfer to a -80˚C freezer. Museum skins and skulls were prepared and deposited in the mammal collection of the UC Berkeley Museum of Vertebrate Zoology (MVZ); specimens and MVZ catalog numbers are provided in Supp. Table 1. We sacrificed and sampled tissues from 10 scrotal males per species with a goal of

obtaining at least five samples per tissue from each species. Trapping, tissue collection, and specimen preparation were approved under UC Berkeley ACUC protocol AUP-2019-03-11939.

Field observation-based species identification was confirmed by amplifying and sequencing the cytochrome B mitochondrial locus via PCR and Sanger sequencing (as in Meléndez-Rosa et al., 2019). One animal identified in the field as *P. maniculatus* was reassigned to *P. boylii* following cytB amplification.

RNA extraction and library preparation

We sampled the following tissues from each individual for mRNA sequencing: a single whole cauda epididymis, one half of a single testis, one half of a single seminal vesicle and approximately 10 mg of liver. RNA was extracted using the Qiagen RNEasy PowerLyzer kit. RNA concentration was quantified and quality assessed using an Agilent 2100 Bioanalyzer RNA Pico 6000 chip. Only samples with RIN scores greater than 7 were used for library preparation, and four low-concentration samples with high RIN scores were reconcentrated using a bead-based mRNA capture step following extraction.

We prepared five samples per tissue per species for RNAseq (or 5 samples per tissue x 3 species x 4 tissues = 60 samples total). Due to difficulty in extracting RNA from seminal vesicles and epididymes, we sampled the following numbers of individuals per species: *P. californicus*: 9 individuals; *P. maniculatus*: 7 individuals; *P. boylii*: 7 individuals (Supp. Table 1).

RNA libraries were prepared using a KAPA Hyperprep RNAseq kit with a bead-based mRNA capture step to exclude ribosomal and non-coding RNAs. mRNA was fragmented for 5 minutes, and 15 cycles of PCR were used to amplify cDNA following the RNA-cDNA conversion step; all other steps were standard as per the Hyperprep protocol. Samples were bioanalyzed again to check for appropriate fragment length and amplified and sequenced to an average depth of 30 million reads per sample on a 150 PE NovaSeq Illumina S4 at the Vincent J Coates Genome Sequencing Laboratory at UC Berkeley.

RNAseq Data Processing

Raw reads were demultiplexed by the sequencing facility and returned as paired end .fastq files for each individual tissue sample (Voss and Nachman 2024). We used fastQC v.0.11.9 (Andrews 2010) to assess raw read quality and fastp v.0.23.2 (Chen et al., 2018) to trim adapter sequences and remove short reads and low-quality bases.

Transcriptome Assembly and Annotation

Since high-quality genomes are not available for all species, we separately assembled a de novo transcriptome for each species to minimize bias in assembly. We combined reads to include all tissues for one sample from each species. As testis RNAseq samples had the highest complexity and lowest duplication rates, we included two testis samples from each species to maximize transcriptome completeness. We used Trinity v.2.15.1 (Haas et al., 2013) to assemble transcriptomes, followed by CD-HIT v.4.8.1 (Li and Godzik 2006; Fu et al., 2012) to cluster sequences with greater than 95% sequence similarity and rnaquast v.2.2.2 (Bushmanova et al., 2016) to remove chimeric or misassembled sequences. We obtained 906,044 to 924,059 transcripts in each transcriptome (Supp. Table 2). This number of transcripts is typical due to the sensitivity of the Trinity assembler.

We assessed transcriptome completeness using Busco v.5.4.5 (Simão et al., 2015) with Euarchontoglires odb10 database. For each of the three transcriptomes, we then used TransDecoder v.5.7.0 to predict open reading frames, followed by Trinotate v.3.2.2 (Bryant et al., 2017) with the *P. maniculatus* Ensembl protein database (Lassance and Hoekstra 2020; HU_Pman_2.1.3 GenBank:GCA_003704035.3) to identify genes and annotate transcriptomes.

Using the three annotated transcriptomes (Supp. Table 2), we used Salmon v.1.10.0 (Patro et al., 2017) to quantify transcript expression from raw RNAseq reads for each sample and converted transcript counts to gene counts using tximport. At this point, we removed two samples (one *P. boylii* seminal vesicle and one *P. boylii* epididymis) due to atypically low percent reads mapping back to the transcriptome. We then used DeSeq2 v.1.40.1 (Love et al., 2014) in R (v.4.2.2) to apply a negative binomial generalized linear model (GLM) to raw read counts outputted from Salmon. This approach normalizes library size variation across samples and applies a correction for differences in transcript length across transcriptomes, minimizing potential bias towards longer genes or any differential annotation due to variation in sequence similarity to the reference species (*P. maniculatus*). As all samples were sequenced on the same machine at the same time, there were no batch effects, and environmental variability was minimized as much as possible during field sampling. We filtered out genes with a mean expression level of fewer than 10 counts per sample.

We used Biomart to obtain the external gene name, gene description, and gene ontology terms for all genes. We also curated a list of GO terms related to reproduction, sperm, reproductive tract development, seminal vesicle, epididymis, and testis for comparison and reference during downstream statistical testing and identification of genes of interest (Supp. Table 3).

Differential Gene Expression Analyses

Using DeSeq2, we separately analyzed each tissue for differences in gene expression between species, then combined all tissues, applied a variance stabilizing transformation, and performed a principal components analysis on the full dataset.

For each tissue, we used a Wald test to identify genes with significant differences in expression between species and applied a Benjamini-Hochberg procedure for multiple testing to correct for false discovery rate (p-adj < 0.01 and $\log_2$-fold change > 1.0). As we were specifically interested in the impact of differences in mating behavior on gene expression, we performed three pairwise comparisons: two monogamous - polygynandrous (*P. californicus* - *P. maniculatus*, *P. californicus* - *P. boylii*) comparisons, and one polygynandrous – polygynandrous (*P. maniculatus – P. boylii*) comparison. We then looked for genes with significant differences in expression in the same direction in both polygynandrous - monogamous comparisons (i.e., up- or down-regulated in both *P. maniculatus* and *P. boylii* relative to *P. californicus*), hereafter referred to as mating system-associated differentially expressed genes (MS-DEGs).

We overlapped genes that met these criteria with the curated reproductive gene ontology term list (Supp. Table 3) to find genes with known reproductive function whose differential expression might relate to differences in postcopulatory sexual selection. We also queried the Mouse Genome Informatics (MGI) Mammalian Phenotype database with all significant DEGs from testis, seminal vesicle, and epididymis that are linked to known reproductive phenotypic variation. Lastly, we performed a Gene Ontology overrepresentation test with PANTHER v.17.0

(Mi et al., 2019; Thomas et al., 2022) to ask if any molecular functions were over-represented among DEGs.

Gene expression profiles across tissues

To explore how overall gene expression varies across different tissues, we compared the number of genes expressed in each tissue and ranked genes by expression level. To rank genes for each tissue, we calculated the mean expression of each gene (from transcript counts) across individuals from all three species. We also identified the most highly expressed genes from each tissue and asked if those genes differed significantly in expression across species in the focal tissue.

Weighted Gene Correlation Network Analysis

We performed a weighted gene correlation network analysis (WGCNA, Langfelder and Horvath 2008) for each tissue individually in R. WGCNA identifies clusters or 'modules' of genes with correlated expression patterns and produces an 'eigengene' score that summarizes the module value for each sample. We ran WGCNA and then identified modules whose summary 'eigengene' score correlated with mating behavior. We further identified genes present in these modules that had reproduction GO terms.

Examining rates of gene sequence evolution

To compare gene expression with gene sequence evolution, we used publicly available data to formally test genes for evidence of positive or relaxed selection. We included two of the focal taxa, *P. maniculatus* (Lassance and Hoekstra 2020; GenBank:GCA_003704035.3 Refseq CDS), and *P. californicus* (Trainor et al., 2022; GenBank:GCA_007827085.3 Refseq CDS), as well as a broader set of muroid rodents, including *P. leucopus* (Long et al., 2019; GCF_004664715.2), *Microtus ochrogaster* (prairie vole; Di Palma et al., 2012; GCF_000317375.1), *Microtus oregoni* (creeping vole; Couger et al., 2021; GCF_018167655.1), *Mesocricetus auratus* (golden hamster; Harris et al., 2021; GCF_017639785.1), *Mus musculus* (Genome Reference Consortium 2020; GCF_000001635.27) and *Rattus norvegicus* (Howe 2020; GCF_015227675.2). Since a reference genome is not available for *P. boylii*, we excluded it from these analyses.

We filtered CDS files for the longest transcript of each gene expressed in the differential expression dataset and used OrthoFinder (Emms and Kelly 2015) to identify, align, and construct gene trees for a set of 4,816 single copy orthologs present in all eight species. We then used two HyPhy analysis methods: BUSTED (Murrell et al., 2015), which tests for evidence of positive selection acting upon a gene in a test branch set or across the whole tree, and RELAX (Wertheim et al., 2015), which identifies instances of relaxed selection on a test branch or set of branches. Both methods identify genes where a model with a high d_N/d_S best explains patterns observed across the tree at some proportion of sites.

We performed both analyses three times, with *P. maniculatus* as the test branch, *P. californicus* as the test branch, and *P. californicus* and *M. ochrogaster* (an independently monogamous rodent) as joint test branches. We also ran BUSTED on the whole tree to test for widespread positive selection, which might be expected if a reproductive gene is under postcopulatory sexual selection across multiple lineages. We then applied a Benjamini-Hochberg correction for multiple testing and asked which genes showed significant evidence of positive or

relaxed selection, with a focus on differentially expressed genes or genes with reproductive GO terms.

Since the set of eight species included only 4,816 single-copy orthologs, we were also interested in exploring sequence evolution in a broader set of Peromyscus genes. To do this, we used orthologr to estimate pairwise *Peromyscus maniculatus – Mus musculus* d_N/d_S values for 13,869 one-to-one orthologs present in gene expression data.

2.3 RESULTS

Reproductive phenotypes

To capture variation in gene expression in a natural setting, we sampled testes, cauda epididymes, seminal vesicles, and liver from ten wild-caught, reproductively mature *P. californicus, P. boylii*, and *P. maniculatus* (Fig. 2A, Supp. Table 1). When scaled by body mass, *P. maniculatus* and *P. boylii* had significantly larger testes than *P. californicus* (n = 30, ANOVA, Tukey HSD: PECA-PEBO p = .00016, PECA-PEMA p = 0.000011, PEBO-PEMA p = 0.58, Figure 2B). *P. maniculatus* had significantly larger seminal vesicles relative to body mass than either *P. boylii* or *P. californicus*, but *P. boylii* and *P. californicus* did not differ significantly from each other (n = 28, ANOVA, Tukey HSD: PECA-PEBO p = 0.93, PECA-PEMA p = 0.0038, PEBO-PEMA p = 0.0097, Supp. Fig. 1).

De novo transcriptome assemblies

We assembled and annotated three de novo *Peromyscus* transcriptomes with the Trinity-Trinotate pipeline (Haas et al., 2013; Bryant et al., 2017). For each species, we used a combined set of two testis samples, one seminal vesicle, one epididymis and one liver sample to maximize transcriptome completeness. Transcriptomes assembled for *P. maniculatus*, *P. californicus*, and *P. boylii* contained an average of 917,847 transcripts, 18,616 *P. maniculatus*-annotated genes, and 87.8% BUSCO completeness (Supp. Table 2).

Differential gene expression associated with different mating systems

A total of 16,183 genes were expressed in all three species at a minimum of 10 mean counts per sample (normalized for transcript length and library size). In a principal components analysis (PCA), samples clustered first by tissue (Fig. 3A) along PCs 1 (33.3% of variation explained) and 2 (17% of variation explained) and then by species (Fig. 3B), with monogamous *P. californicus* separating from polygynandrous *P. boylii* and *P. maniculatus* along PC3 (9.3% variation explained).

We identified differentially expressed genes for each tissue individually and found 3,627 mating system-associated differentially expressed genes (MS-DEGs) where expression shifts in the same direction in both polygynandrous species relative to monogamous *P. californicus* using a multiple testing-adjusted p-value $p < 0.01$ and $\log_2$-fold change > 1 across the three reproductive tissues (Supp. Figs. 2-5). Of these, 295 MS-DEGs are associated with reproductive gene ontology terms or phenotypes from the Mouse Genome Informatics database (Supp. Table 4).

We performed a gene ontology (GO) term analysis to ask whether any molecular function categories were over-represented among the 3,627 reproductive tract MS-DEGs. Endopeptidase,

peptidase inhibitor, and signaling receptor GO terms were all enriched when compared to the background set of genes expressed in the male reproductive tract (Supp. Table 5).

More MS-DEGs in seminal vesicle than in other reproductive tissues

For each tissue, we calculated the proportion of DEGs that were MS-DEGs (Table 1). For example, 1,522 of 3,050 testis differentially expressed genes changed in the same direction (gene expression increased or decreased in both PEBO and PEMA compared with PECA), or approximately 50% of DEGs were MS-DEGs. A similar proportion was seen in the epididymis. In contrast, a far greater proportion (1,993 of 2,232 DEGs or 89.3%) of seminal vesicle DEGs shifted in the same direction (χ^2 test $p < 0.0001$, Supp. Table 6). When a reduced set of genes with reproductive functions was tested, the higher proportion of MS-DEGs in the seminal vesicle persisted (χ^2 test $p < 0.0001$). MS-DEGs could arise from parallel changes in the same direction in both *P. boylii* and *P. maniculatus*, or perhaps more likely, from changes arising in *P. californicus*. Of the 1,993 seminal vesicle MS-DEGs, a slightly greater proportion of genes decreased (n = 1,040 genes) than increased (n = 953 genes) in expression in *P. californicus* relative to *P. boylii* and *P. maniculatus* (χ^2 test $p = 0.0513$).

We explored whether these differences in gene expression might be a result of a small number of coordinated shifts in gene expression using a weighted gene co-expression network analysis (WGCNA). One to two modules per tissue had per-sample WGCNA summary 'eigengene' scores that varied significantly with mating behavior (Supp. Fig. 6). These modules contain a mix of reproductive and non-reproductive genes, but seminal vesicle module 2 contains twice as many reproductive genes that are significant MS-DEGs as other mating behavior correlated-modules (Supp. Table 7).

To compare variation in gene expression profiles among tissues, we ordered and plotted genes from lowest to highest mean expression in each tissue (Fig. 4). We observed broad variation in expression profiles: the seminal vesicle stands out for having a few genes that are very highly expressed (Fig. 4D), while testis had more genes expressed at an intermediate to high level and the most total genes expressed (Fig. 4B), and epididymis showed lower overall gene expression (Fig. 4C).

Copulatory plug genes Svs2 and Pate4 are expressed at a lower level in monogamous mice

The two most highly expressed genes in the seminal vesicle, Seminal vesicle secretory protein 2 (*Svs2*) and Prostate and testis expressed 4 (*Pate4*) were expressed at a more than two log-fold lower level in monogamous *P. californicus* compared to polygynandrous *P. maniculatus* and *boylii* (*Svs2*: $p = 0.000575$, ANOVA, Fig. 4E; *Pate4*: $p = 0.000885$, ANOVA, Fig. 4F). To investigate differential expression in seminal fluid proteins more broadly, we identified 96 seminal fluid proteins (Sfps) from a proteomic study of *Mus* seminal fluid (Smyth et al., 2022) that are expressed in the *Peromyscus* male reproductive tract (Supp. Table 8). Many Sfps (n = 84) are present in all three reproductive tissues, but 60 are most abundantly expressed in the epididymis, and 20 are MS-DEGs (Supp. Table 9). Five of the most abundant Sfps, including *Svs2* and *Pate4*, were primarily expressed in the seminal vesicle.

Rapidly evolving genes are associated with sperm motility and sperm-egg binding

Across a phylogeny of eight muroid rodents, 958 of 4,816 single copy orthologs analyzed displayed significant evidence of positive selection (Fig. 5A; Table 2). Significantly more genes showed evidence of positive selection in *P. maniculatus* than *P. californicus* ($p = 0.0016$,

Fisher's exact test). Gene ontology terms related to peptidase and protease function, ATP-dependent activity, cilia, and cell surface receptor signaling were over-represented in genes showing evidence of tree-wide positive selection (PantherDB: *p-adj* < 0.05 with FDR correction). Moreover, 64 genes under tree-wide positive selection had functions related to sperm motility and sperm-egg binding, including *Ovch2*, *Spaca1*, *Acrbp* and female-derived *Zp2* (Supp. Table 10).

Overlap between rapidly evolving and differentially expressed genes

Across the male reproductive tract, we identified 203 MS-DEGs with evidence of positive selection, which did not exceed the amount of overlap expected by chance (Fig. 5B, Table 2, $p = 0.337$, Fisher's exact test). However, seventeen genes with reproductive functions were both rapidly evolving and MS-DEGs in the male reproductive tract. In a separate analysis of pairwise *P. maniculatus – M. musculus* d_N/d_S, MS-DEGs had higher d_N/d_S than the genome-wide average (MS-DEG mean d_N/d_S: 0.170, genome-wide mean d_N/d_S: 0.158, $p = 1.178 \times 10^{-5}$, Mann-Whitney U Test bootstrapped with 1,000 replicates, Supp. Fig. 7).

Genes can evolve rapidly due to either positive selection or relaxation of purifying selection (Dapper and Wade 2020). We therefore also looked for overlap between MS-DEGs and those genes showing relaxed selection. Overall, fewer genes showed evidence of relaxed selection *(*$n = 458$) than positive selection ($n = 523$) on *Peromyscus* branches ($p = 0.031$, Fisher's exact test), but we identified more genes with evidence of relaxed selection in *P. maniculatus* ($n = 260$) than in *P. californicus* ($n = 198$). Though there was not more overlap than expected by chance between genes under relaxed selection and reproductive tract MS-DEGs (Fig. 5B, $p = 0.416$, Fisher's exact test), several reproductive genes ($n = 14$) showed evidence of relaxed selection in both monogamous lineages tested (*P. californicus* and *M. ochrogaster*) despite significant tree-wide positive selection (Table 3). Six of these genes are MS-DEGs, which is more than expected by chance (Fisher's exact test, $p = 0.000055$). Genes that showed relaxed constraint in association with monogamy displayed functions such as sperm-egg binding, sperm flagellar motility, and regulation of meiosis and gamete production.

2.4 DISCUSSION

Postcopulatory sexual selection is mediated by mating behavior: if a female mates with multiple males per estrous cycle, her male mating partners must compete to fertilize eggs and exclude the sperm of rival males, while monogamous males do not contend with male-male sperm competition. Rapid evolution of reproductive protein-coding gene sequences has been thoroughly documented (e.g., Wyckoff et al., 2000; Swanson et al., 2001a; Dean et al., 2009), but gene expression has been under-explored in the context of mating ecology and postcopulatory sexual selection. Here, we captured adult males in the wild to investigate differential expression and rates of gene sequence evolution across the testis, seminal vesicle, and epididymis, all of which contribute to male fertility and reproductive success.

In monogamous *P. californicus*, male-male competition is reduced and postcopulatory sexual selection should be weak in comparison with polygynandrous *P. maniculatus* and *P. boylii*. A recent study showed that traits under postcopulatory sexual selection, such as sperm production efficiency and sperm swimming speed, are increased and show a stronger correlation with female traits such as oviduct length in more polygynandrous *Peromyscus* (Weber and Fisher

2023). This suggests that the extent of both male-male competition and male-female sexual conflict varies with mating behavior. In keeping with these predictions, we found that wild-caught *P. californicus* had smaller testes than *P. maniculatus* and *P. boylii* relative to body size, suggesting that monogamous males invest less in sperm production.

Broad patterns of differential expression in the male reproductive tract

Peromyscus mice exhibit differential gene expression in association with variation in postcopulatory sexual selection: we identified 3,627 mating system-associated differentially expressed genes (MS-DEGs) in the male reproductive tract. Several gene co-expression modules followed this pattern as well, suggesting that some of these genes may be co-regulated by shared transcription factors. *P. maniculatus* and *P. californicus* are more closely related to each other than either is to *P. boylii* (Fig. 1C), so MS-DEGs could result from convergent changes in gene expression in both polygynandrous species, or from a single change in monogamous *P. californicus*. The latter explanation is simpler since it only involves changes in one lineage. The comparisons between the two polygynandrous species and *P. californicus* are not phylogenetically independent, but they do provide a more stringent test for differential expression than a single pairwise polygynandrous – monogamous comparison.

As sperm production is phenotypically plastic and males are known to alter investment based on perceived risk of male-male competition (Ramm et al., 2015; Firman et al., 2018), most studies of postcopulatory sexual selection take place in a laboratory setting with unmated, singly housed males. Here, we measured gene expression in wild mice. This undoubtedly increases noise since wild mice likely vary in reproductive status, time since last mating, age, nutrition, and pathogen exposure. However, one advantage of field studies is that gene expression variation can be studied in a natural reproductive context. Here, we were able to identify genes with strong differences in expression between species and mating systems despite the "noisiness" of gene expression data collected in the field.

Numerous genes with functions related to spermatogenesis, sperm motility, and sperm – egg binding were MS-DEGs, including the four most abundant reproductive genes in the testis: *Spata6*, *Crisp2*, *Ropn1*, and *Zpbp* (Fig. 6), which all show reduced expression in *P. californicus*. These decreases in gene expression may reflect decreased investment in sperm production, which is reflected in the relatively smaller testes of *P. californicus*. *Prkar1a*, which has been implicated in sperm midpiece length and swimming speed in *P. maniculatus* and *P. polionotus* (Fisher et al., 2016), also shows reduced expression in the *P. californicus* testis.

MS-DEGs constituted roughly half of all DEGs in the testis and epididymis, but nearly 90% in the seminal vesicle. This implies that changes in gene expression in the seminal vesicles might be particularly important in species with different mating systems. If most MS-DEGs arose from changes in *P. californicus* associated with a transition to monogamy, and if the change in mating system entailed a relaxation of selection for particular gene products, we might expect that most MS-DEGs would be downregulated in *P. californicus* seminal vesicles compared to the other two species. A slightly greater but non-significant proportion of seminal vesicle MS-DEGs were downregulated than upregulated in *P. californicus* compared with the polygynandrous species, suggesting that expression differs broadly, but not uniformly, in the *P. californicus* seminal vesicle, though a handful of key genes show strongly reduced expression.

The seminal vesicle and decreased copulatory plug investment in monogamous mice

The seminal vesicle also had a distinct gene expression profile from the testis and epididymis, with a small number of very highly expressed genes. *Svs2* is one of 95 known *Mus* seminal fluid proteins (Sfps) expressed in *Peromyscus* (Noda and Ikawa 2019; Shindo et al., 2019). *Svs2* and *Pate4* are the most abundantly expressed genes in the seminal vesicle, but we observed a two log-fold reduction in expression in *Svs2* and similar reduction in *Pate4* in monogamous *P. californicus*. In the laboratory, *Svs2*$^{-/-}$ sperm can fertilize eggs, but do not make a copulatory plug and undergo a premature acrosome reaction or die in the uterus, suggesting multiple roles including sperm-uterine fluid interaction and sperm acrosome reaction timing (Kawano et al., 2014; Shindo et al., 2019). *Pate4* knockout mice are subfertile and form severely reduced copulatory plugs (Noda et al., 2019).

The seminal fluid and copulatory plug are costly physiological investments (Mangels et al., 2016), Male laboratory mice subject to sexual selection show increased *Svs1* and *Svs2* expression (Simmons et al., 2020) despite energetic costs, suggesting that high *Svs* expression confers fitness benefits to polygynandrous males. It is possible that selection has favored a significant reduction in *Svs2* and *Pate4* expression in *P. californicus* to decrease physiological investment in seminal fluid proteins. However, Gubernick (1988) reported that in a laboratory setting, *P. californicus* often forms a copulatory plug even though females are not known to mate multiply. The size and function of the *Peromyscus* copulatory plug has not been quantified, but we hypothesize that *P. californicus* still makes a copulatory plug and expresses *Svs2* and *Pate4* at lower levels due to the roles that these genes play in protecting sperm and preventing premature sperm capacitation (Schneider et al., 2016).

Positive and relaxed selection in the male reproductive tract

There are inherent challenges to identifying genes that are involved in reproduction and may be under positive selection: rapidly evolving genes often lack one-to-one orthologs across larger clades, so they cannot be included in formal tests of selection, and gene function annotations may be missing or incomplete. However, comparing patterns of positive and relaxed selection across species and tissues can still be informative. We speculated that *P. maniculatus* might have more reproductive genes with evidence of positive selection and *P. californicus* more genes under relaxed selection due to reduced male-male competition. Instead, we observed that *P. maniculatus* showed more evidence of both positive and relaxed selection.

In rodents, which most often show loosely affiliative multiple-mating systems, positive selection drives the evolution of faster sperm, effective sperm-egg recognition and fertilization, and copulatory plug formation (Tourmente et al., 2011; 2013; Firman et al., 2014; Mangels et al., 2016; Hook et al., 2021). Consistent with these patterns, we observed evidence of tree-wide positive selection in sperm-egg binding proteins such as *Ovch2*, *Spaca1*, and *Acrbp*. However, Dapper and Wade (2020) argue that reproductive genes may evolve rapidly even in the absence of positive selection. Since genes that promote male fertility are only under selection fifty percent of the time, and vice versa for female fertility genes, selection acting upon them may be relaxed compared with genes that are expressed in non-reproductive contexts in both sexes. Thus, relaxed selection may be the most appropriate null model for sequence evolution in the context of postcopulatory sexual selection. Though we observe more genes under positive than relaxed selection in both *P. maniculatus* and *P. californicus*, a substantial proportion (11.67%) of genes with reproductive gene ontology terms showed evidence of relaxed selection. This is slightly but not significantly higher than the proportion of non-reproductive genes under relaxed selection (8.78%), which is consistent with Dapper and Wade's (2020) hypothesis. Reproductive

genes in both *P. californicus* and *P. maniculatus* may be under relaxed selection in comparison with genes involved in development and organismal function in both sexes.

Though relaxed selection may be relevant in both polygynandrous and monogamous contexts, we sought to identify genes that might be responding to changes in the strength of postcopulatory sexual selection. In keeping with the ancestral state of loosely affiliative polygynandry, sperm-egg binding proteins *Hexb* and *Catsper1* and sperm flagellum proteins *Cfap43* and *Dcdc2c* (Table 3) show evidence of tree-wide positive selection, but relaxed selection on both monogamous branches (*P. californicus* and *M. ochrogaster*). These genes may be under positive selection when postcopulatory sexual selection is strong but experience relaxed selection in monogamous species where male-male competition is absent.

Integrating gene sequence and expression evolution

Altering gene expression can be a way to avoid negative pleiotropic effects associated with gene sequence evolution (Brawand et al., 2011). We explored whether the same genes are rapidly evolving and differentially expressed in the male reproductive tract. If the same genes are both evolving rapidly and differentially expressed, postcopulatory sexual selection might act strongly on a few genes. If instead, separate sets of genes show rapid gene sequence evolution versus differential expression, they may be responding to postcopulatory sexual selection in different ways because of distinct advantages or constraints on expression and protein function.

We did not observe more overlap than expected by chance between genes under positive selection and MS-DEGs (Fig. 5B, Table 2). The lack of overlap between these two gene categories suggests a broad partitioning of gene sequence evolution and gene expression evolution in the male reproductive tract. However, seventeen MS-DEGs with known reproductive functions showed evidence of positive selection, and six MS-DEGs showed evidence of tree-wide positive selection and *P. californicus*-specific relaxed selection. These genes may be important for male reproductive success when male-male competition is strong and experience relaxed selection on both gene sequence and expression when male-male competition is reduced.

Sperm motility genes suggest ways in which gene sequence and gene expression might both be relevant to changes in PCSS. Sperm swimming speed depends on sperm size, sperm tail protein efficacy, and midpiece and flagellum gene expression (Tourmente et al., 2013; Vicens et al., 2014), and is correlated with male-male competition intensity (Gomendio and Roldan, 2008; Tourmente et al., 2011; Firman and Simmons 2011). Sperm flagellar genes such as *Ropn1* and *Spata6* showed decreased expression in the *P. californicus* testis, and sperm flagellum genes *Cfap65*, *Hoatz*, and *Dnai1* were under tree-wide positive selection but showed both decreased expression and relaxed selection in *P. californicus*, together suggesting that selection on sperm swimming speed may be relaxed in monogamy. Observed decreases in sperm gene expression could reflect lower investment in sperm production more generally, but it is important to also note that many testis MS-DEGs increase in abundance in *P. californicus*.

Though we could not formally test seminal vesicle MS-DEGs *Pate4* and *Svs2* for evidence of positive selection since they were not recovered in all eight species with genome data, they had among the highest pairwise *P. maniculatus – M. musculus* d_N/d_S values we observed (*Pate4* $d_N/d_S = 1.052$, *Svs2* $d_N/d_S = 1.276$). Prior work in *Peromyscus* and *Mus* has shown that *Svs2* is under positive selection in rodents and suggests that an accumulation of glutamine residues, which are targets for copulatory plug catalyst *Tgm4* (Tseng et al., 2009;

2012; Dean 2013), might lead to formation of a more effective copulatory plug (Ramm et al., 2008; 2009). *Svs2* may thus respond to postcopulatory sexual selection via changes in both protein abundance and gene sequence: polygynandrous rodents express more, longer *Svs2* transcripts.

This study represents a thorough but incomplete evaluation of the association between mating ecology and gene expression in focal *Peromyscus* species. Since we have excluded the prostate, coagulating gland, and bulbourethral gland, we did not comprehensively survey gene expression across the male reproductive tract. Further, the testis and epididymis are complex tissues with multiple cell types that contribute to different stages of spermatogenesis (Hermann et al., 2018; Shi et al., 2021; Majane et al., 2022). We assessed variation in gene expression across whole tissues or tissue compartments (as in the cauda epididymis), but RNA sequencing of cell-type-enriched cell populations as in Kopania et al., (2022) could clarify which cell types show differential expression in response to postcopulatory sexual selection in *Peromyscus*.

Additionally, without exploring female reproductive tract gene expression, our understanding of the relevance of gene expression to male-female sexual conflict is one-sided (Firman et al., 2017; Firman 2018; McDonough-Goldstein et al., 2021). For example, there is strong evidence that some female reproductive proteins are evolving rapidly in *Peromyscus* and other mammals (Swanson et al., 2001b; Turner and Hoekstra 2006; 2008), but we do not know whether gene expression varies across females that experience different levels of polygynandry (e.g., Veltsos et al., 2022). Recent work in *Drosophila* also suggests that some genes that were thought to be male-specific Sfps are expressed in the female reproductive tract, bringing into question the roles of male-female conflict versus male-female cooperation in reproduction (Cridland and Begun 2023). Exploring female reproductive tract gene expression would address whether any 'male fertility' genes are expressed in the female reproductive tract or vice versa and would contribute to our understanding of female defense traits such as egg extracellular matrix and uterine fluid composition (Hook et al., 2022; Weber and Fisher 2023).

Conclusion

Exploring gene expression in monogamous and polygynandrous mice yielded insights into PCSS dynamics across the male reproductive tract: the seminal vesicle showed the most mating system-associated differential expression of any tissue, suggesting that gene expression is more flexible and selection perhaps more relaxed in the context of monogamy. Gene expression in the testis and epididymis is likely more constrained and under stronger stabilizing selection, as both are involved in sperm production and gene expression is conserved during meiosis and early spermatogenesis (Good and Nachman 2005; Burgoyne et al., 2009; Kopania et al., 2022).

Several gene-specific findings also suggest that *P. californicus* males may have reduced investment in male-male competition strategies (Fig. 6); these genes may be responding to male-male competition and/or male-female sexual conflict. In sum, male *Peromyscus* mice exhibit changes in gene expression, gene sequence evolution, and sometimes both in association with the varying demands on sperm function, sperm – egg binding, and seminal fluid – uterine fluid interactions that constitute postcopulatory sexual selection.

2.5 FIGURES

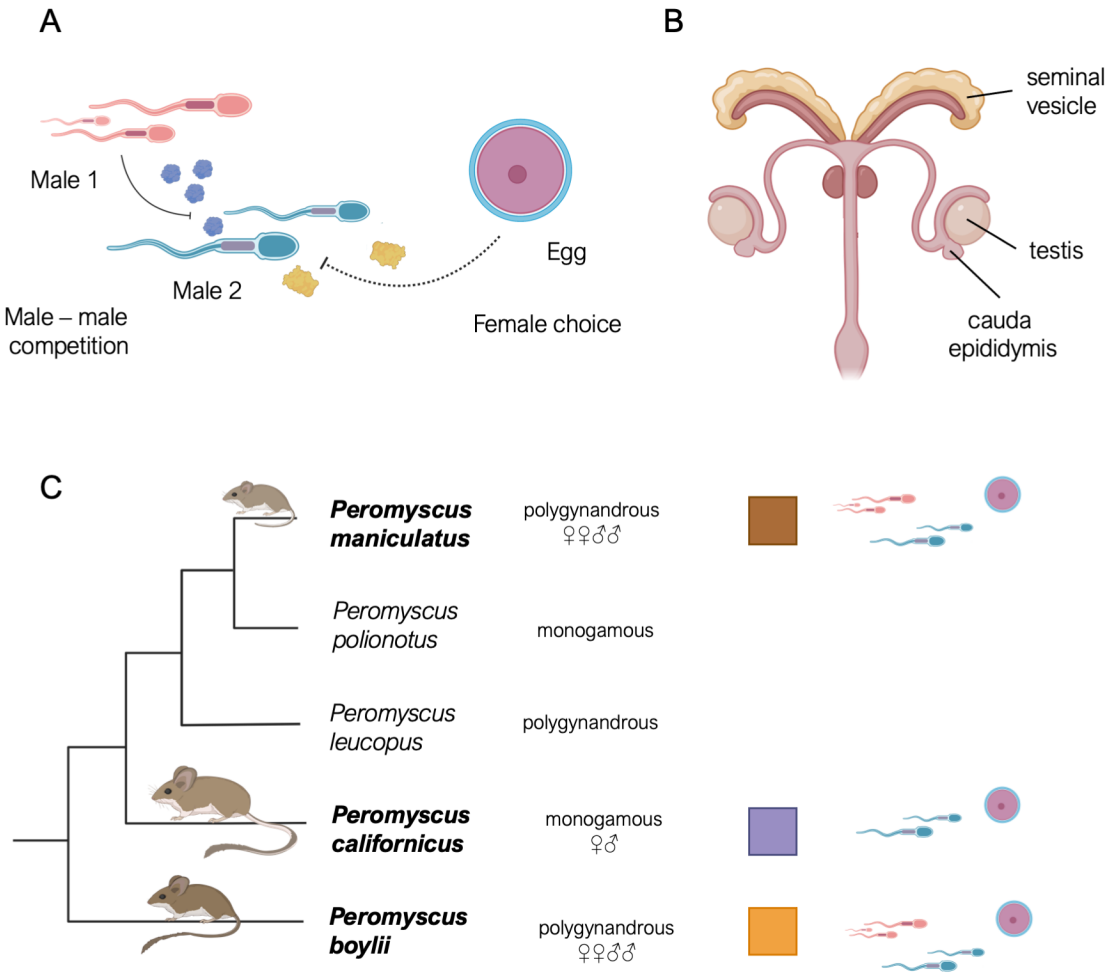


Figure 1. A: Postcopulatory sexual selection consists of two potential biological interactions: male-male competition and female choice or sexual conflict. As male-male competition is only relevant when females mate with multiple males per estrous cycle, postcopulatory sexual selection is mediated by variation in mating behavior. B. The mouse male reproductive tract contains multiple tissues that contribute to sperm development, seminal fluid composition, and mating outcome, all of which are potentially subject to postcopulatory sexual selection. C: Abridged *Peromyscus* phylogeny (per Platt et al., 2015), with relevant variation in mating behavior and the corresponding strength of postcopulatory sexual selection males experience. Squares indicate colors that are used in all figures to indicate results for each focal species. Focal species may interact where they co-occur, but are reproductively isolated from each other and represent distinct species groups within the *Peromyscus* genus.

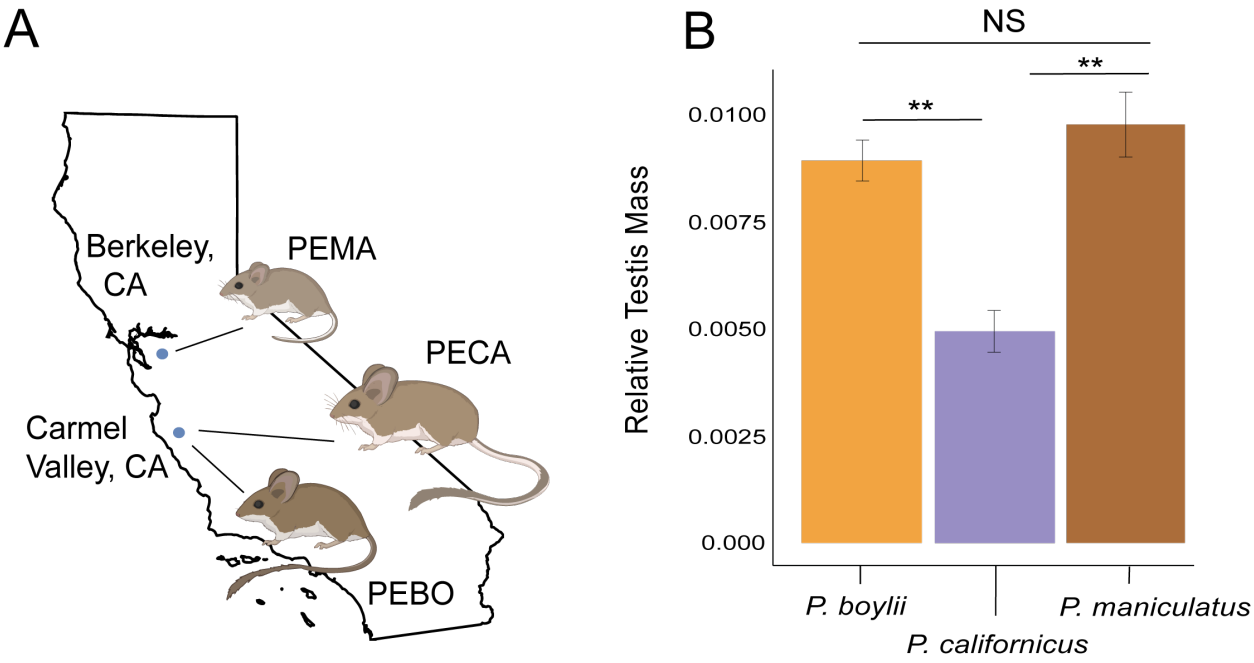


Figure 2. A: Mice were collected and testes, seminal vesicles, cauda epididymes, and liver were preserved from two locations: Hastings Natural History Reserve, Carmel Valley, CA (*P. californicus* abbrev. PECA and *P. boylii* abbrev. PEBO) and the Field Station for the Study of Behavior, Ecology, and Reproduction, Berkeley, CA (*P. maniculatus* abbrev. PEMA). B: Polygynandrous species have larger testes than monogamous species. When scaled by body mass, testis mass differs significantly between both polygynandrous species and monogamous *P. californicus* (n = 30, ANOVA, Tukey HSD: PECA-PEBO $p = 0.00016$, PECA-PEMA $p = 0.000011$), but not between the two polygynandrous species *P. boylii* and *P. maniculatus* ($p = 0.58$). Error bars represent standard error.

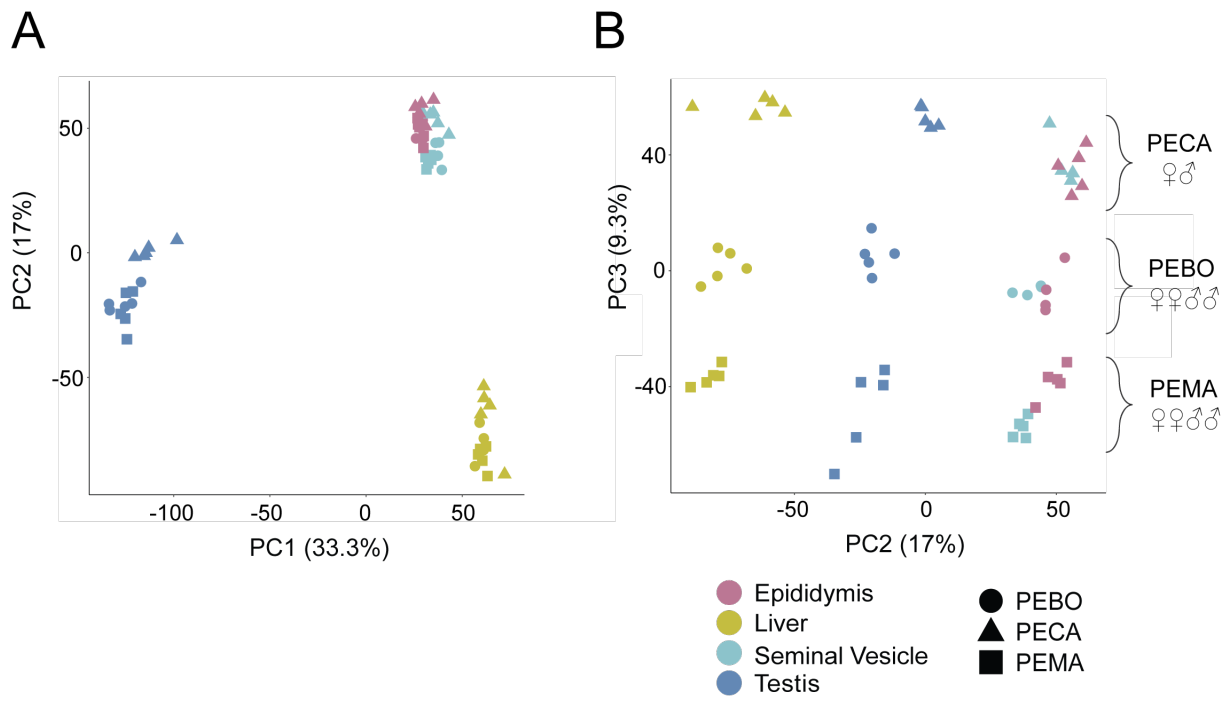


Figure 3. Principal components analysis of gene expression variation across all species and tissues after applying a variance-stabilizing transformation to all counts data (n = 16,183 genes). A: Principal components 1 and 2 explain a combined 50% of variation and cluster samples by tissue. B: Principal components 2 (17% variation explained) and 3 (9.3% variation explained) separate samples by tissue, then by mating behavioral variation (Monogamy ♂♀; Polygynandry ♂♂♀♀).

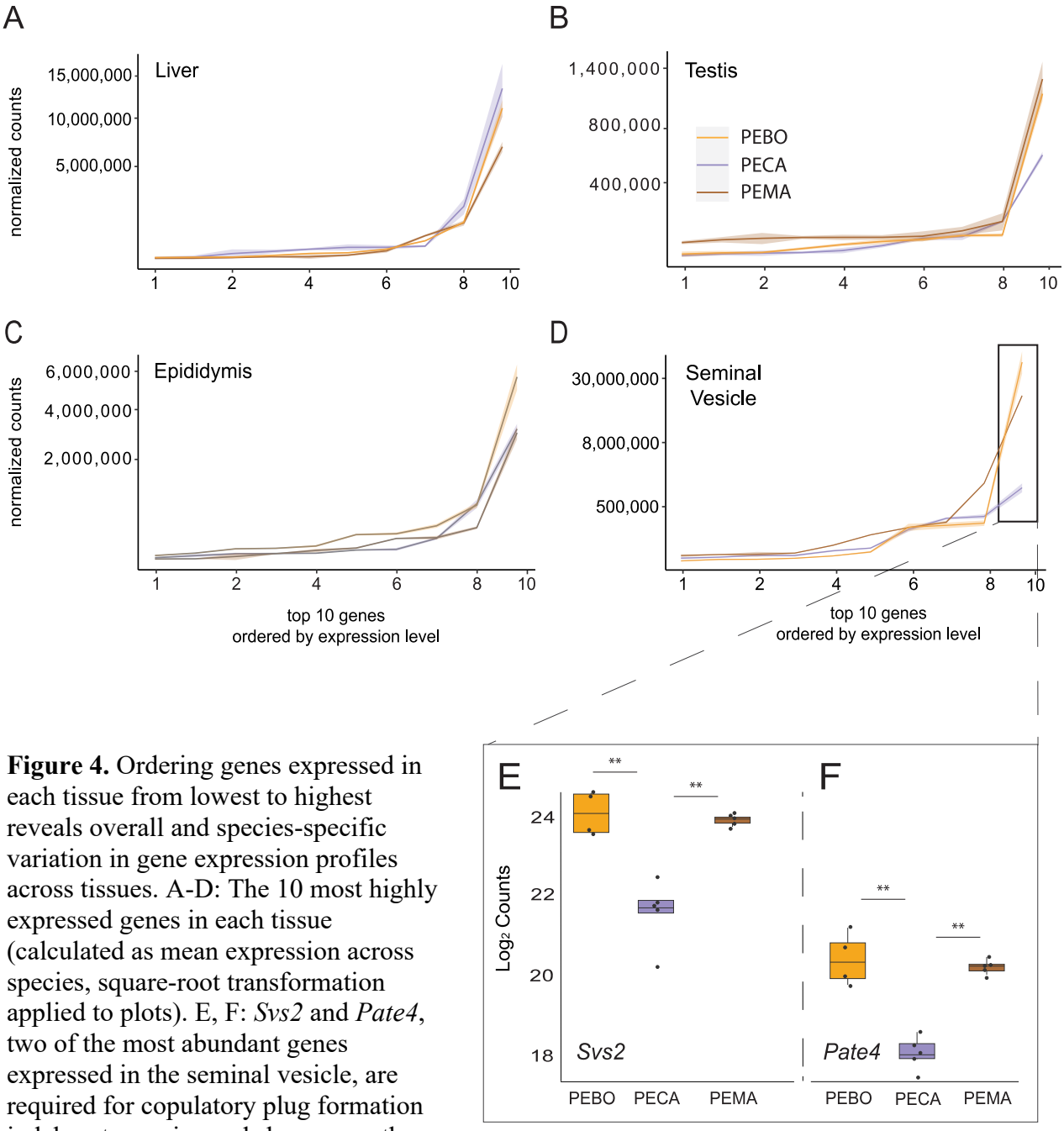


Figure 4. Ordering genes expressed in each tissue from lowest to highest reveals overall and species-specific variation in gene expression profiles across tissues. A-D: The 10 most highly expressed genes in each tissue (calculated as mean expression across species, square-root transformation applied to plots). E, F: *Svs2* and *Pate4*, two of the most abundant genes expressed in the seminal vesicle, are required for copulatory plug formation in laboratory mice and show more than two log-fold lower expression in mongamous *P. californicus* (*Svs2*: $p = 0.000575$, *Pate4*: $p = 0.000885$, ANOVA), joining a category we describe as mating system associated differentially expressed genes (MS-DEGs). Confidence intervals represent standard error and points indicate expression level in each individual sampled.

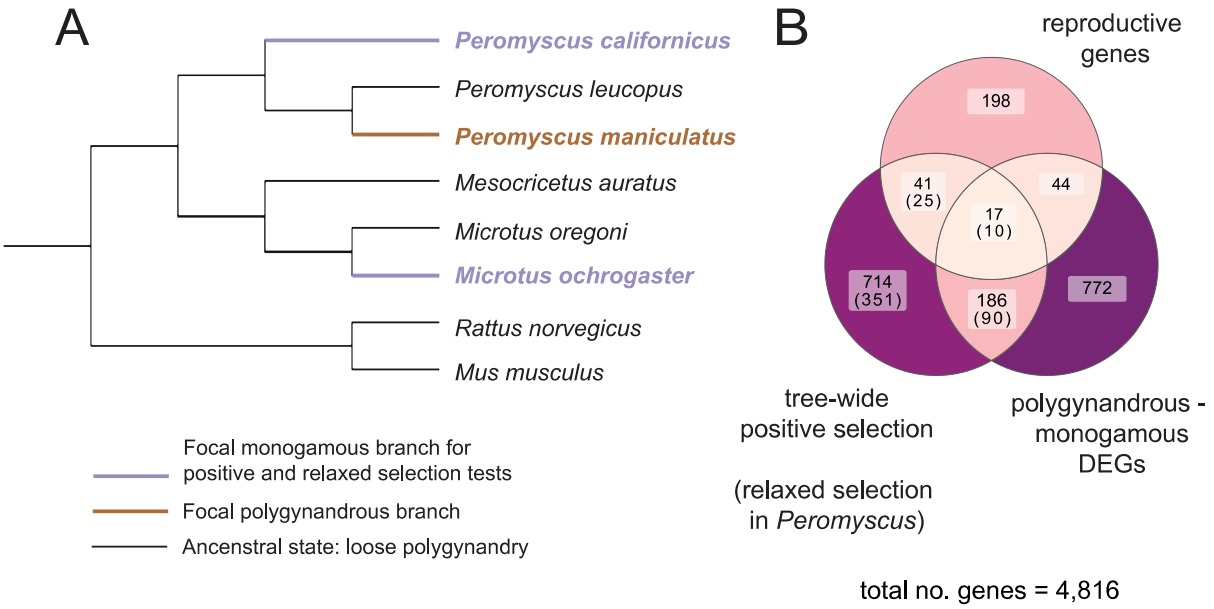


Figure 5. We tested 4,816 genes across eight muroid rodents for evidence of positive or relaxed selection (HyPhy BUSTED and RELAX methods). A: phylogenetic relationships between taxa included in selection analyses (phylogeny from Platt et al., 2015; Steppan and Schenk 2017). B: Of the 4,816 genes tested for positive selection, 958 had significant evidence of positive selection, 1,019 of genes tested were male reproductive tract MS-DEGs, and 300 had known reproductive functions. 203 genes were both differentially expressed and best described by an evolutionary rate model including positive selection, which is not more overlap than expected by chance ($p = 0.337$, Fisher's exact test). Additionally, 458 genes had evidence of relaxed selection in *Peromyscus*. Overlap between genes with evidence of relaxed selection and other gene categories is indicated in parentheses.

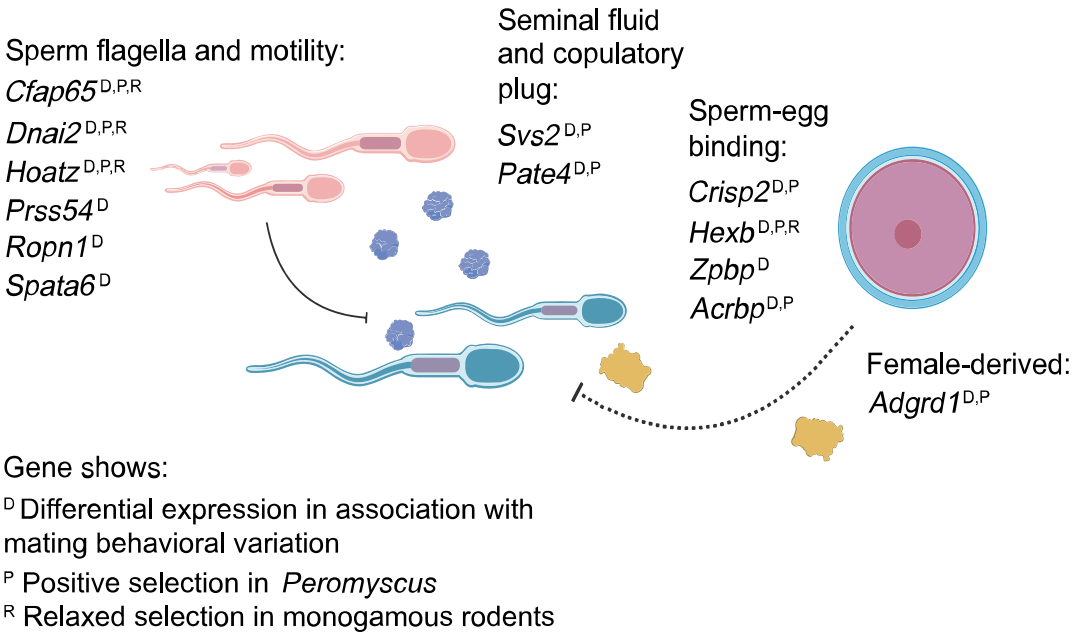


Figure 6. Postcopulatory sexual selection (PCSS) combines two potential biological interactions: male-male competition and male-female sexual conflict. *Peromyscus* mice respond to PCSS with a combination of differential gene expression and molecular sequence evolution; changes in gene expression or gene sequence may occur in response to male-male competition, male-female conflict, or a combination of the two. Genes involved in sperm motility, seminal fluid composition, and sperm-egg binding that display evidence of mating system-associated differential expression and positive selection or relaxed constraint are depicted in the context of the postcopulatory interactions that occur within the female reproductive tract.

2.6 TABLES

Table 1. Summary of differential expression results by tissue (testis, seminal vesicle, epididymis, and liver), with emphasis on mating system-associated differentially expressed genes (MS-DEGs).

Significance Threshold	Testis no. genes	Seminal vesicle no. genes	Epididymis no. genes	Liver no. genes
Genes expressed	15,893	14,178	15,557	13,505
Genes expressed with reproductive gene ontology (GO) terms†	792	638	732	581
DEGS*	3,050	2,232	2,000	2,053
MS-DEGs	1,522	1,993	1,027	1,048
Proportion of MS-DEGs out of all DEGs	49.9	89.3**	51.4	51.0
MS-DEGs with reproductive GO terms	88	137**	72	54

† Gene ontology terms via Mouse Genome Informatics, Uniprot and NCBI databases.

*Differentially expressed genes identified using a Wald test with Benjamini Hochberg correction for multiple testing ($p\text{-}adj < 0.01$, $LFC > 1$).

**Proportion of differentially expressed genes associated with mating system differs significantly in seminal vesicle compared with other tissues (χ^2 test $p < 0.0001$) for all MS-DEGs and the subset of genes with reproductive GO terms.

Table 2. Summary of positive results including tree-wide and branch-specific tests. Below, overlap between positively selected genes and MS-DEGs across male reproductive tissues.

	Positive Selection Analyses	No. genes	No. reproductive genes
All results	Single-copy orthologous genes analyzed	4,816	300
	Tree-wide significant	958	64
	P. maniculatus test branch significant	297	19
	P. californicus test branch significant	226	19
Overlap with MS-DEGs	All reproductive tissues – tree-wide significant	203	17
	Testis – tree-wide significant	75	3
	Seminal Vesicle – tree-wide significant	120	12
	Epididymis – tree-wide significant	54	4
	Liver – tree-wide significant	63	6

Positive selection tests performed using HyPhy BUSTED, p-values adjusted for multiple testing with a Benjamini-Hochberg correction. Gene ontology terms via Mouse Genome Informatics, Uniprot and NCBI databases.

Table 3. Genes with reproductive function gene ontology terms that show evidence of relaxed selection in monogamous *P. californicus* and *M. ochrogaster*, but positive selection across the gene tree (Fig. 5A). †Indicates male reproductive tract MS-DEGs.

Gene Name	Gene Description	RELAX adj. p-value	Gene Function
Cfap43	Cilia and flagella associated protein 43	0	Flagellated sperm motility, sperm axoneme assembly
Mov10l1†	Move10 like RISC complex helicase I	0	Male meiotic nuclear division, male gamete generation, spermatogenesis
Dcdc2c	Doublecortin domain containing 2c	2.72×10^{-6}	Sperm flagellum
Hexb†	Hexosaminidase B	5.48×10^{-6}	Acrosomal vesicle, penetration of zona pellucida
Gli2	GLI-Kruppel family member GLI2	0.00043	Sperm cilia axoneme, prostatic bud formation
Cct3	Chaperonin containing TCP1 subunit 3	0.0015	Sperm-zona pellucida binding, protein folding chaperone
Cfap65†	Cilia and flagella associated protein 65	0.0050	flagellated sperm motility, sperm axoneme assembly, sperm midpiece, acrosomal vesicle
Dnai1†	Dynein axonemal intermediate chain 1	0.0097	Cilia dynein complex, spermatozoa and testis-expressed
Hoatz†	HOATZ cilia and flagella associated protein	0.0126	Spermatogenesis, flagellated sperm motility
Cav1†	Caveolin 1	0.0120	Sperm acrosomal membrane, sperm cilia
Catsper1	Cation channel, sperm associated 1	0.0247	Flagellated sperm motility, CatSper complex, fusion of sperm to egg plasma membrane, fertilization, sperm principal piece

Majin	Membrane anchored junction protein	0.0251	Homologous chromosome pairing, meiotic recombination, spermatogenesis
Dmc1	DNA meiotic recombinase 1	0.0263	Male meiosis I, homologous chromosome pairing, spermatid development, oocyte maturation
Brdt†	Bromodomain testis associated	0.0459	Sperm DNA condensation, male meiotic nuclear division

Relaxed selection tests performed using HyPhy RELAX, p-values adjusted for multiple testing with a Benjamini-Hochberg correction. Gene descriptions via Mouse Genome Informatics, Uniprot and NCBI databases.

Supplementary figures and tables are available online at DOI: 10.1111/mec.17433.

CHAPTER 3

A high-quality genome assembly for a desert-adapted rodent, Merriam's kangaroo rat (*Dipodomys merriami*)

This chapter has been previously published and is reproduced here in accordance with the journal's article sharing policy:

Voss ER, Escalona M, Nguyen O, Marimuthu MPA, Chumchim N, Fairbairn CW, Seligmann W, Beraut E, Conroy CJ, Patton JL, Bowie RCK, Nachman MW. 2025. A high-quality genome assembly for a desert-adapted rodent, Merriam's kangaroo rat (*Dipodomys merriami*). *Journal of Heredity* esaf023.

DOI: 10.1093/jhered/esaf023

ABSTRACT

Merriam's kangaroo rat (*Dipodomys merriami*) is a member of a unique family of primarily desert-adapted North American rodents (Heteromyidae). Of the 20 species in the genus, *D. merriami* is one of the most wide-ranging and ecologically flexible, inhabiting desert scrub, grassland, sagebrush steppe, and juniper-piñon woodland in the southwestern deserts of the United States and Mexico. We present a *de novo* reference genome for *D. merriami* generated from PacBio HiFi long-read and Omni-C chromatin proximity sequencing as a part of the California Conservation Genomics Project. The primary pseudo-haplotype assembly comprises 3,110 scaffolds, with a contig N50 of 8.6 Mb, scaffold N50 of 49.1 Mb, and a total length of 3.57 Gb. Further, a BUSCO completeness score of 97.8% suggests that the assembly is highly complete. This reference genome will serve as a resource for future studies of *Dipodomys* conservation genomics, desert adaptation, and phylogeography.

3.1 INTRODUCTION

The deserts of the United States southwest and northern Mexico are home to many organisms that have adapted to challenging hot and dry environments. Kangaroo rats (Castorimorpha: Heteromyidae: Dipodomyinae), which include 20 species in the genus *Dipodomys*, provide a classic example of desert adaptation. They can survive indefinitely without free water via physiological, morphological, and behavioral adaptations to limit water loss (Fig. 1A, B; Howell and Girsh 1935; Schmidt-Nielsen and Schmidt-Nielsen 1952; Tracy and Walsberg 2002). As desert herbivores, kangaroo rats are considered ecosystem engineers and keystone species, and the ecology and physiology of several *Dipodomys* species has been deeply explored (Hinds and MacMillen 1983; Brown and Heske 1990; Brock and Kelt 2004). Genomic perspectives of Heteromyid evolution, population genetics and desert adaptation are emerging:

whole genomes have been assembled for *D. spectabilis* (Harder et al. 2022)*, D. stephensi* (Johnson et al. 2019)*, D. ordii* (Liu et al. 2014), and *Perognathus longimembris* (Wilder et al. 2022, Kozak et al. 2024) and these are available on GenBank.

Rocha et al. (2021) characterize kangaroo rats as 'evaders', desert organisms that behaviorally avoid extreme heat exposure and are physiologically adapted to reduce water loss. To survive under arid conditions, kangaroo rats reduce respiratory water loss via physiological and morphological adaptations including highly concentrated urine and elongated nasal passages with highly convoluted turbinal bones (Jackson and Schmidt-Nielsen 1964; MacMillen and Hinds 1983). They construct complex burrow systems, which they use for seed caching and reproduction, and only venture out at night when temperatures are cool (Kenagy 1973).

Though these Heteromyid rodents are uniquely adapted to their arid environments, some species are more flexible than others. Species with more limited ranges or closer dependence on particular biotic communities are at risk: five species of kangaroo rats are listed as vulnerable and two as endangered by the USFWS and IUCN due to severe habitat reduction and fragmentation. Further, the arid biomes of the American southwest are in the midst of a prolonged, climate-change induced drought, and changes in temperature and precipitation may cause the spatial distribution of ecosystems and key food sources across the landscape to shift unpredictably going forward (Archer and Predick 2008; Wilkening et al. 2019; Riddell et al. 2021).

Here, we present a genome assembly for Merriam's kangaroo rat, *Dipodomys merriami*. The *D. merriami* genome was assembled as a part of the California Conservation Genomics Project (CCGP), which seeks to quantify and assess the distribution of organismal genetic variation across the state of California (Shaffer et al. 2022; Toffelmier et al. 2022). One subspecies of Merriam's kangaroo rat, *D. merriami parvus*, the San Bernardino kangaroo rat, is of conservation concern due to urbanization, dam construction, and changes to water regimes in the regions north of Los Angeles and east of San Bernardino (USFWS 1998, 2020; Chock et al. 2020; Hendricks et al. 2020). Though Merriam's kangaroo rat more broadly is not at risk, it forms a species complex with two narrowly distributed species in Baja California (*D. insularis* and *D. margaritae*) as well as the endangered Fresno kangaroo rat *D. nitratoides*. The Fresno kangaroo rat most likely originated from a northwestern population of *D. merriami* that crossed the Tehachapi mountains to colonize the San Joaquin Valley in California and subsequently diverged from *D. merriami* via allopatric speciation (Patton et al. 2019). This reference genome will provide a starting point from which to examine the landscape and conservation genomics of several vulnerable species of kangaroo rats in California, including *D. m. parvus* and *D. nitratoides* as well as the more distantly related Stephens' kangaroo rat in Southern California (*D. stephensi*), and the giant kangaroo rat in the western San Joaquin Valley (*D. ingens*).

3.2 METHODS

Biological materials

A male *D. merriami merriami* individual was captured at the mouth of Freeman Canyon, Kern County, California (35.6091, -117.92995) on 12 August 2020 with authorization from the California Department of Fish and Wildlife (Scientific Collecting permit S-192560001-19256-002 issued to JLP) and euthanized following the Guidelines of the American Society of

Mammalogists (Sikes et al. 2016). The study skin, skull, and tissues have been deposited as a vouchered specimen at the Museum of Vertebrate Zoology, Berkeley, California (https://arctos.database.museum/guid/MVZ:Mamm:240054).

High molecular weight (HMW) genomic DNA (gDNA) was extracted from 78 mg of liver tissue (male, MVZ:Mamm:240054, JLP29074) using the Nanobind Tissue Big DNA kit as per the manufacturer's instructions (Pacific BioSciences - PacBio, Menlo Park, CA). The DNA purity was estimated using absorbance ratios (260/280 = 1.81 and 260/230 = 1.98) on the NanoDrop ND-1000 spectrophotometer. The final DNA yield (25 µg) was quantified using the Quantus Fluorometer (QuantiFluor ONE dsDNA Dye assay; Promega, Madison, WI). The size distribution of the HMW DNA was estimated using the Femto Pulse system (Agilent, Santa Clara, CA) and 85% of the fragments were 100 kb or longer.

Nucleic acid library preparation

The HiFi SMRTbell library was constructed using the SMRTbell Express Template Prep Kit v2.0 (PacBio, Cat. #100-938-900) according to the manufacturer's instructions. HMW gDNA was sheared to a target DNA size distribution between 15 kb – 18 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 37˚C for 15 minutes, followed by further enzymatic steps of DNA damage repair at 37˚C for 30 minutes, end repair and A-tailing at 20˚C for 10 minutes and 65˚C for 30 minutes, and ligation of overhang adapter v3 at 20˚C for 60 minutes. The SMRTbell library was purified and concentrated with 1X Ampure PB beads for nuclease treatment at 37˚C for 30 minutes and size selection using the BluePippin/PippinHT system (Sage Science, Beverly, MA; Cat #BLF7510/HPE7510) to collect fragments greater than 7 – 9 kb. The 15 – 20 kb average HiFi SMRTbell library was sequenced at UC Davis DNA Technologies Core (Davis, CA) using eleven 8M SMRT cells, Sequel II sequencing chemistry 2.0, and 30-hour movies each on PacBio Sequel II & IIe sequencers.

The Omni-C library was prepared using the Dovetail™ Omni-C™ Kit (Dovetail Genomics, Scotts Valley, CA) according to the manufacturer's protocol with slight modifications. First, specimen tissue (liver, MVZ:Mamm:240054, JLP29074) was thoroughly ground with a mortar and pestle while cooled with liquid nitrogen. Subsequently, chromatin was fixed in place in the nucleus. The suspended chromatin solution was then passed through 100 µm and 40 µm cell strainers to remove large debris. Fixed chromatin was digested under various conditions of DNase I until a suitable fragment length distribution of DNA molecules was obtained. 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. An NGS library was generated using an 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 6000 platform (Illumina, CA) to generate approximately 100 million 2 x 150 bp read pairs per GB genome size.

Genome Assembly

We assembled the genome following the CCGP assembly pipeline version 5.0 with modifications (details, including software versions, available in Supplementary Table 2). After removing remnant adapters from the PacBio HiFi long-read sequences using HiFi AdapterFilt (Sim et al. 2022), we generated an initial phased diploid assembly with HiFiasm (Cheng et al. 2022) in HiC mode with the filtered HiFi long-reads in combination with Omni-C short-read sequences. We aligned the Omni-C data to both assemblies following the Arima Genomics Mapping Pipeline and then scaffolded both assemblies with YaHS (Zhou et al 2023).

We generated and analyzed Omni-C contact maps to curate both assemblies. To generate the contact maps, we used BWA-MEM (Li 2013) to align the Omni-C data and then identified ligation junctions and generated Omni-C pairs with pairtools (Open2C et al. 2024). Then, we generated multi-resolution Omni-C matrices with cooler (Abdennur and Mirny 2020) and balanced them with hicExplorer (Ramírez et al. 2018). We visualized the contact maps with HiGlass (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). We identified and later broke joins where major mis-assemblies and mis-joins were found. Some of the remaining gaps were closed using the PacBio HiFi reads and YAGCloser (https://github.com/merlyescalona/yagcloser). Lastly, we checked for contamination using the BlobToolKit (Challis et al. 2020).

Genome quality assessment

We used meryl (https://github.com/marbl/meryl) to obtain k-mer counts from the PacBio HiFi reads, which were then used in GenomeScope2.0 (Ranallo-Benavidez et al. 2020) to estimate genome features including genome size, heterozygosity, and repeat content. We ran QUAST to obtain general contiguity metrics (QUality ASsessment Tool; Gurevich et al. 2013) and BUSCO (Benchmarking Single Copy Orthologs; Manni et al. 2021) to evaluate genome quality and functional completeness in comparison with the 13,798 gene Glires ortholog database (glires_odb10). We also assessed base-level accuracy (QV) and k-mer completeness with the previously generated meryl database and merqury (Rhie et al. 2020), and further estimated genome assembly accuracy via BUSCO gene set frameshift analysis (Korlach et al. 2017). To check the difference in metrics and genome size between haplotypes, we compared both haplotypes by generating a pairwise sequence alignment between haplotypes (primary and alternate) using nucmer (from mummer; Marçais et al. 2018) and visualized the alignment on the web platform dot (https://dot.sandbox.bio/; see Supplementary Figure 2).

To contextualize the new genome assembly's completeness and contiguity, we searched for published Heteromyid genome resources using the NCBI Genome database. We queried the database for Heteromyidae reference genomes and found six species with published genome assemblies. References, accession number, and assembly statistics including genome size, number of scaffolds, scaffold N50, and BUSCO score for each genome are reported in Table 2.

Mitochondrial genome assembly

We assembled the *D. merriami merriami* mitochondrial genome from the PacBio HiFi reads using the reference-guided pipeline MitoHiFi (Allio et al. 2020; Uliano-Silva et al. 2023). The mitochondrial sequence of the closely related *Castor canadensis* (NCBI:NC_033912.1; Lok et al. 2017) was used as the starting sequence. After completion of the nuclear genome, we

searched for matches of the resulting mitochondrial assembly sequence in the nuclear genome assembly using BLAST+ (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.

Preliminary Annotation

We assembled the *D. merriami* mitogenome using MitoHiFi and used MITOS2 for annotation. We lifted over genome features from *D. spectabilis*, the closest relative with a high-quality genome annotation (estimated time since divergence: 9.3 million years, Timetree.org). We used liftoff (Shumate and Salzberg 2021) to lift over gene coding sequences, exons, and mRNAs from *D. spectabilis* (NCBI: GCF_019054845.1) to *D. merriami*, polish feature boundaries and identify putative instances of gene duplication.

3.3 RESULTS

Sequencing data

PacBio sequencing libraries yielded 173.7 Gb of HiFi data for a total of 15.1 million reads and an average 24.1x coverage genome-wide based on the primary assembly genome size of 3.57 Gb (N50 read length 11,649 bp; minimum read length 85 bp; mean read length 11,504 bp and maximum read length 40,213 bp; see Supplementary Figure 3 for read length distribution). We estimated a sequencing error rate of 0.153% from the PacBio HiFi reads, and the k-mer spectrum shows a bimodal distribution with a major peak at ~57-fold coverage and a minor peak at ~30-fold coverage (Fig. 2A). The Omni-C sequencing libraries generated 227.1 million read pairs.

Assembly metrics

The final assembly (mDipMer1) consists of two phased haplotypes that have been tagged primary and alternate based on our assessment of their completeness and contiguity (Fig. 2C, Fig. S1). The primary assembly (mDipMer1.1.p) contains 3,110 scaffolds, with a contig N50 of 8.6 Mb, scaffold N50 of 49.1 Mb, and a total assembly length of 3.57 Gb (Fig. 2B). An overall BUSCO score of 97.8% based on 13,798 single copy orthologs in the glires_odb10 database suggests a highly complete genome. The alternate assembly is 2.74 Gb in length, contains 2,258 scaffolds, and has a contig N50 of 12.9 Mb, scaffold N50 of 129.5 Mb, and BUSCO score of 95.6%. More detailed metrics for both the primary and alternate haplotypes are listed in Table 1. The difference in total length of the two assemblies is partly attributable to the sex chromosomes: the larger primary assembly contains scaffolds corresponding to both sex chromosomes whereas the alternate does not. We identified these scaffolds by looking at the sequencing depth of coverage variation across scaffolds in both haplotypes (Palmer et al. 2019). An alignment of the primary and alternate assemblies is provided in Supplementary Figure 2.

Preliminary Annotation

Using Liftoff, we succeeded in transferring 25,198 gene annotations (of 33,494; 75.2%) from *D. spectabilis* to *D. merriami*. 20,208 of these contained at least one valid open reading frame and 99% or greater sequence coverage. We also assembled a 17,159 bp mitochondrial genome containing 13 protein-coding genes, 2 rRNAs, and 22 unique tRNAs. The mitochondrial

genome and preliminary nuclear genome feature annotation from *D. spectabilis*, are available on the Dryad Data Repository (further details available in Data Availability Statement).

3.4 DISCUSSION

The genome assembly for Merriam's kangaroo rat is highly contiguous. We report a total length of 3.57 Gb, a scaffold N50 of 49.1 Mb, 3,110 scaffolds, and a BUSCO score of 97.8%. The assembly contains several mammalian chromosome-length scaffolds (maximum scaffold length: 191.5 Mb; 13 scaffolds > 50 Mb). As a preliminary step towards genome annotation, we have also lifted over gene feature annotations from *D. spectabilis* to *D. merriami*.

We add this genome to those of four Heteromyid species with assemblies available on GenBank: the banner-tailed kangaroo rat, *D. spectabilis* (Harder et al. 2022: GCA_019054845.1), Stephens' kangaroo rat, *D. stephensi* (Johnson et al. 2019:GCA_004024685.1), Ord's kangaroo rat, *D. ordii* (Liu et al. 2014: GCA_000151885.2), and the little pocket mouse, *Perognathus longimembris* (Wilder et al. 2022: GCA_023159225.12; Kozak et al. 2024: GCA_024363575.2). Both published genomes for *P. longimembris* are highly contiguous, with scaffold N50 values approaching chromosome-level resolution. This assembly falls short of that metric but is comparable with the recently published *D. spectabilis* assembly and improves upon earlier short-read sequencing-based assemblies generated for *D. stephensi* and *D. ordii* in terms of total number of scaffolds, scaffold N50, and BUSCO score (Table 2). The *D. merriami* genome is also the largest Heteromyid genome in terms of total length (3.57 Gb) published thus far; the next largest reported assembly is for *D. spectabilis* (2.8 Gb).

The difference in size between the *D. merriami* genome and other Heteromyid genomes is substantial but not surprising. Classic studies of *Dipodomys* genome evolution based on C-value and karyotype report chromosome numbers ranging from 2N=52 to 2N=74 across the genus (Stock 1974) and generally high but varied heterochromatin content (Hatch et al. 1976). *D. merriami* has a fundamental chromosome number of 100, while the *D. spectabilis* fundamental number is 70. Some of the variation in genome size is likely to be driven by this variation in karyotype and heterochromatic repetitive element content.

Dipodomys merriami has a degree of flexibility in habitat choice, diet and reproductive timing that is not shared by all kangaroo rat species (Zeng and Brown 1987; Nagy and Gruchacz 1994; Tracy and Walsberg 2001). Their species distributional range covers the western Great Basin, Mojave, Colorado, Sonoran, and Chihuahuan deserts of the southwestern United States and northern Mexico (Fig. 1C) where they occupy arid habitats, including sagebrush steppe, desert scrub, desert grassland, and juniper-piñon woodland.

In contrast, the endangered giant kangaroo rat *D. ingens* and vulnerable Fresno kangaroo rat (*D. nitratoides*) and Stephens' kangaroo rat (*D. stephensi*) are threatened by habitat loss to agriculture in the southern San Joaquin Valley and development in coastal southern California (Price and Endo 1989; Lortie et al. 2018; US Fish and Wildlife Service 1987; 1988a; 1988b; 2022). Within the context of the origin of the genus *Dipodomys* approximately 12 million years ago (MYA), *D. merriami* diverged from *D. ingens* and *D. stephensi* between 6 and 10 MYA (TimeTree; Kumar et al. 2017) and from its close relative *D. nitratoides* more recently (3 MYA).

We plan to integrate the *D. merriami* reference genome with whole genome resequencing from several California *Dipodomys* species, including the closely related *D. nitratoides*, as well as the more distantly related *D. ingens*, *D. stephensi*, *D. panamintinus* and *D. heermanni*, to deepen our understanding of *Dipodomys* demography and genetic diversity. Endangered species often show reduced genetic diversity and inbreeding depression because of their small population sizes (e.g. Robinson et al. 2021; Tian et al. 2022), but some seem able to maintain higher levels of genetic variation or lower mutational load (e.g. Morin et al. 2021). Thus, widespread, ecologically flexible *D. merriami* may have higher genome-wide genetic diversity than other kangaroo rat species with smaller contemporary ranges and census population sizes. As current genetic variation depends on both present population size and past demographic history, we can also infer *Dipodomys* historical effective population size (N_e) from whole-genome data (as in Harder et al. 2022) and ask if *D. merriami* has a historically large N_e. In contrast, species of concern such as *D. nitratoides* and *D. ingens* might have stable, long-term low N_e or show evidence of recent bottlenecks (e.g., Robinson et al. 2022).

More broadly, the *D. merriami* genome might shed light on the genomic underpinnings of desert adaptation, which is complex, polygenic, and requires many physiological and behavioral changes. Genetic adaptations to desert life have been explored in multiple mammalian taxa, including North American rodents such as the cactus mouse *Peromyscus eremicus* (Colella et al. 2021), the pocket mouse *Chaetodipus intermedius* (Bittner et al. 2022), as well as foxes (Rocha et al. 2023), camels (Wu et al. 2014), and sheep (Yang et al. 2016). Strikingly, even some organisms that have adopted different strategies for surviving the stresses of desert life show changes in the same gene pathways, including fat storage and metabolism, insulin resistance, and kidney arachidonic acid metabolism, which may stimulate water re-uptake (Rocha et al. 2021). The *D. merriami* genome could also be used to identify signatures of desert adaptation and ask whether there is evidence of convergent adaptation in these pathways with other arid-adapted animals, or if kangaroo rats have evolved and adapted to the desert in ways that are unique to the Heteromyid lineage.

In sum, this new genome assembly for Merriam's kangaroo rat will enable scientists to explore the genomics of kangaroo rat adaptation to the harsh desert environment and provide context for conservation efforts in southern California and the fragile arid ecosystems of southwestern North America.

3.5 FIGURES

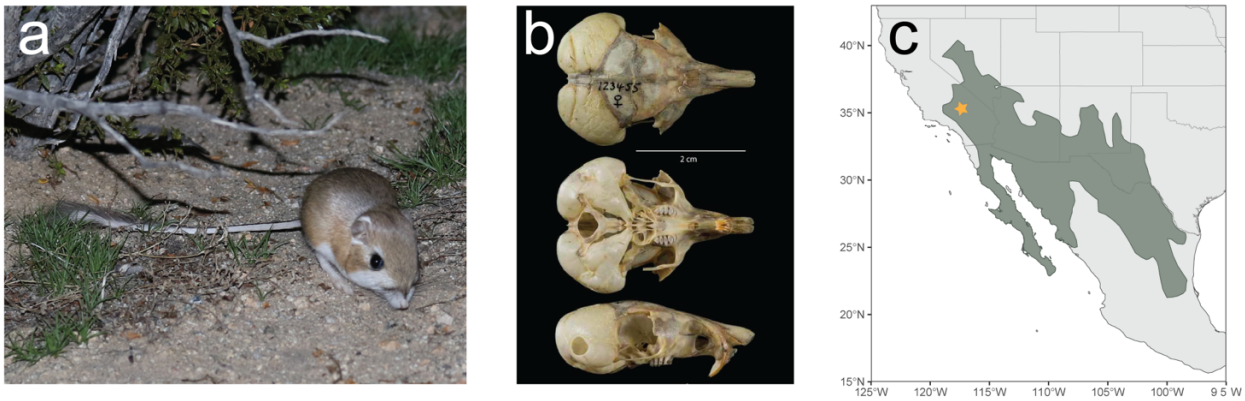


Figure 1. A: *Dipodomys merriami* inhabits arid environments across the southwestern United States and Mexico. Kangaroo rats display long hind limbs used in saltatorial locomotion, a long tail for balance, and external cheek pouches they fill with seeds during foraging bouts. B: *Dipodomys spp.* have unique cranial features, including a long, narrow rostrum to reduce evaporative water loss and enlarged auditory bullae to enhance hearing. C: species range map; star indicates sample locality of the reference individual. (A: Photo credit: Joshua Doby, iNaturalist. (https://www.inaturalist.org/observations/200486409, San Bernardino County, California). B: type specimen for genome assembly, *D. merriami collinus*, Museum of Vertebrate Zoology (MVZ:Mamm:123455, photograph by William Stone), UC Berkeley. C: Range data retrieved from IUCN).

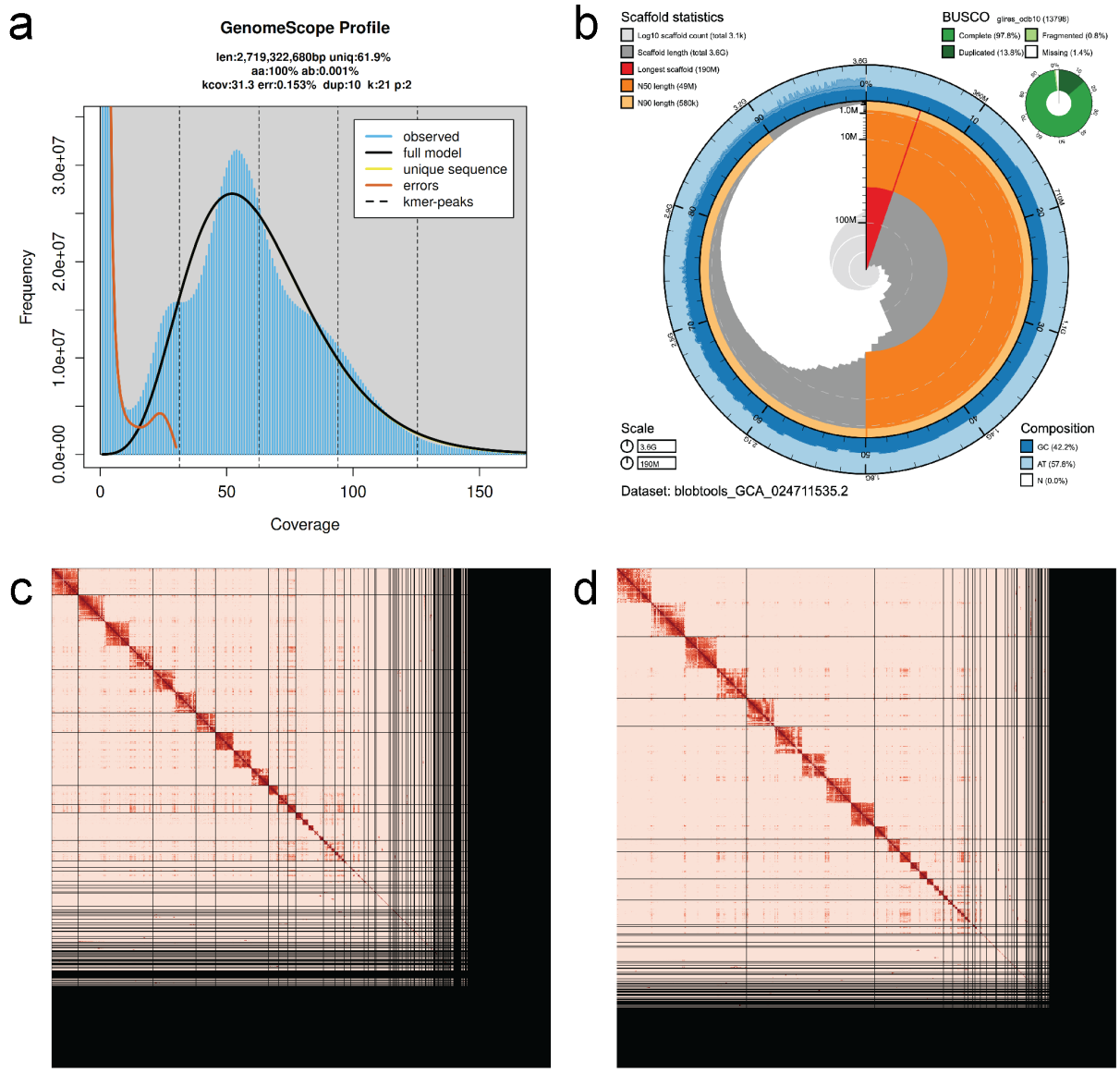


Figure 2. Visual overview of *Dipodomys merriami* genome assembly composition and quality. A: GenomeScope2.0 k-mer spectrum from adapter-trimmed HiFi sequence data. B: Snail plot summary visual of *D. merriami* primary assembly quality generated with BlobToolKit (Challis et al. 2020). Central circular plot represents the entire length of the genome, with scaffolds added in clockwise fashion from longest to shortest. The outer ring indicates relative AT/CG content, and scaffold summary statistics and BUSCO scores are shown at top left and right. Snail plot for alternate assembly is shown in Supplementary Figure 1. C, D: Primary and alternate assembly contiguous linear arrangement visualized as Omni-C contact maps (PreTextSnapShot). Darker areas represent regions of the genome that are close to each other in 3D space as identified by chromatin-proximity sequencing.

3.6 TABLES

Table 1. Metrics for the primary and alternate assemblies of Merriam's kangaroo rat (*Dipodomys merriami*) genome.

Metric	Primary	Alternate
Number of contigs	3,382	2,523
Contig N50 (bp)	8,649,182	12,858,602
Contig NG50§	15,011,649	12,858,602
Longest contig	63,678,358	58,418,348
Number of Scaffolds	3,110	2,258
Largest scaffold	191,530,317	189,654,535
Scaffold N50	49,097,427	129,497,925
Scaffold NG50§	129,005,595	129,497,925
Scaffold L50	14	9
Scaffold LG50	9	9
Size of final assembly	3,568,728,228	2,738,440,652
Gaps per Gbp (#Gaps)	76(272)	97(265)
BUSCO‡		
Complete genes	97.8%	95.6%
Complete: single copy	84.0%	85.0%
Complete: duplicated	13.8%	10.6%
Fragmented	0.8%	0.8%
Missing	1.4%	3.6%
Base pair QV	66.7747	606.5546
	Full assembly = 66.6778	
k-mer completeness	95.1101	89.6307
	Full assembly = 99.1517	

§ Read coverage and NGx statistics have been calculated based on the estimated genome size of 2.7 Gb.
‡ (P)rimary and (A)lternate assembly values.

Table 2. Assembly metrics for other Heteromyidae genome assemblies available on NCBI GenBank as of September 2024.

English name	Scientific name	Length (Gb)	# Scaffs.	Scaff. N50 (Mb)	Scaff. L50	BUSCO score (%)	GC content (%)	Citation (GenBank ID)
Merriam's kangaroo rat	*Dipodomys merriami*	3.57	3,110	49.1	14	97.8	42	Voss et al. 2025 GCA_024711535
Spectacled kangaroo rat	*Dipodomys spectabilis*	2.8	2,026	9.6	69	98.2	41	Harder et al. 2022 GCA_019054845
Ord's kangaroo rat	*Dipodomys ordii*	2.2	65,193	11.9	56	93.8	42	Liu et al. 2014 GCA_000151885
Stephens' kangaroo rat	*Dipodomys stephensi*	2.3	1,269, 643	0.037	16,981	75.4	42	Johnson et al. 2019 GCA_004024685
Pacific pocket mouse	*Perognathus longimembris pacificus*	2.2	6,180*	72.7	11	96.8	42	Wilder et al. 2022 GCA_023159225
Little pocket mouse	*Perognathus longimembris longimembris*	2.3	982	74.3	11	93.8	42	Kozak et al. 2024 GCA_024363575

*chromosome-level assembly contains 28 assembled chromosomes and 6,152 additional scaffolds.

Supplementary figures and tables are available online at DOI: 10.1093/jhered/esaf023

CHAPTER 4

Levels of genetic variation, distribution of fitness effects, and conservation of kangaroo rats inferred from whole-genome sequences

ABSTRACT

Population fragmentation, contraction, and inbreeding can lead to decreased genetic diversity, reducing evolutionary potential and options for recovery for species in decline. However, despite numerous examples of highly inbred species with genetic defects, the relationship between neutral genetic diversity and risk of extinction is not straightforward, and there has been debate over the importance of neutral genetic diversity in the conservation genetics community. To compare neutral and selectively relevant genetic diversity, we sequenced the genomes of 142 kangaroo rats (*Dipodomys*) from six species in California, three of which are threatened or endangered (*D. ingens*, *nitratoides*, and *stephensi*) and three of which are not of conservation concern (*D. merriami*, *heermanni*, and *panamintinus*). Using these whole-genome data, we quantified overall genetic diversity (heterozygosity) and inbreeding (runs of homozygosity), inferred historic effective population size over the past one million years (using PSMC), and estimated the distribution of fitness effects (DFE) for each species. We report no significant difference in overall genetic diversity or frequency of runs of homozygosity between threatened and non-threatened species of kangaroo rats. Instead, historic effective population size was strongly correlated with present-day genetic diversity. The endangered San Joaquin kangaroo rat had a larger historical population size and higher genetic diversity compared with the endangered Giant (*D. ingens*) and Stephens' (*D. stephensi*) kangaroo rats. While endangered kangaroo rats did not harbor longer runs of homozygosity overall, the extinct Fresno kangaroo rat, (*D. n. exilis*) displayed extremely long (20 Mb) runs of homozygosity suggestive of strong inbreeding in samples collected five years prior to their last sighting in the wild. Furthermore, the endangered kangaroo rats *D. nitratoides* and *D. ingens* appear to carry greater loads of mildly to moderately deleterious mutations than the non-threatened *D. merriami* and *D. heermanni*. Taken together, these results suggest that while overall levels of heterozygosity may not be good predictors of extinction risk, mutational load and levels of inbreeding are still relevant concerns for genetically diverse kangaroo rats.

4.1 INTRODUCTION

The field of conservation genetics posits that genetic diversity and extinction risk are intertwined: the neutral theory predicts that as effective population size decreases and drift increases, genetic diversity will decrease (1983). By definition, endangered species have small population sizes and may thus have lower genetic diversity, reducing the amount of raw genetic material available to facilitate adaptation and recovery, and increasing the likelihood of chance fixation of deleterious mutations (Lande 1994, Lynch et al. 1995). Small populations are

generally also subject to a higher mutational load (Kimura et al. 1963). Preservation of genetic diversity is a key tenet of endangered species management and guides decisions regarding habitat protection, captive breeding, and individual translocations for genetic rescue (Hedrick 1992; Crandall et al. 2000, Moritz 2002; DeWoody et al. 2021; Rossetto et al. 2021).

Multiple meta-analyses of conservation genetic data have confirmed that, on balance, endangered species do have lower genetic diversity than non-threatened species (Willoughby et al. 2015; Li et al. 2016; Brüniche-Olsen et al. 2021; Leroy et al. 2022). Further, the expansion of population genomic resources to non-model organisms has enabled conservation geneticists to probe the genomes of endangered species more deeply (Hohenlohe et al. 2021). With these new data, scientists have reported numerous examples of critically endangered species with extremely low genetic diversity and genomic signatures of inbreeding (e.g. Abascal et al. 2016; Robinson et al. 2016; Tian et al. 2022), some of which are linked to negative health consequences. Genetic load, or the reduction in mean population fitness due to deleterious mutations (van Oosterhout 2020; Bertorelle et al. 2022), can be observed in small, isolated populations such as the Isle Royale wolves in Lake Superior and endangered Florida panthers (Robinson et al. 2019; Ochoa et al. 2022). These populations have a high prevalence of atypical phenotypes such as kinked tails and malformed vertebrae as well as reduced fertility and high susceptibility to infectious disease (Roelke et al. 1993; Johnson et al. 2010; Hedrick et al. 2014).

However, some endangered species have high genetic diversity and some widespread species have low genetic diversity. For example, California condors (*Gymnogyps californianus*) display long genomic runs of homozygosity (ROH), which arise due to inbreeding: individuals are more likely to inherit the same allele from each parent when parents are closely related (Curik et al. 2014; Shafer and Kardos 2025). Despite their extensive ROH and critically endangered status, condors have higher background levels of heterozygosity than the closely related and abundant turkey vulture (*Cathartes aura*; Robinson et al. 2021). On the other hand, some abundant species such as the barn owl (*Tyto alba*), harp seal (*Pagophilus groenlandicus*), and Norway rat (*Rattus norvegicus*) have relatively low genetic diversity (Leffler et al. 2012; Peart et al. 2020; Brüniche-Olsen et al. 2021). Furthermore, the relationship between neutral and selectively relevant (i.e. nonsynonymous) variation is not always straightforward; low genetic diversity does not always correlate with high genetic load (Hedrick and Garcia-Dorado 2016; Dussex et al. 2023). The endangered Iberian lynx (*Lynx pardinus*) and vaquita (*Phocoena sinus*) both have extremely low genetic diversity but no evidence of excess genetic load, suggesting that long-term small effective population sizes have facilitated purging of deleterious mutations (Kleinman-Ruiz et al., 2022; Robinson et al. 2022).

This complex relationship between genetic diversity, genetic load, and population viability can make it challenging to apply genomic data to practical conservation actions. Some have contended that neutral genetic diversity has been over-valued in conservation genetics (Teixeira and Huber 2021). In rebuttals to this argument, others state that this stance misses the bigger picture that genetic diversity and extinction risk are linked as outcomes shaped by multiple biological factors. These include landscape-level dynamics such as degree of habitat fragmentation and species-specific variables such as historic effective population size, demography, mutation rate, fecundity, and breeding system (DeWoody et al. 2021).

Here, test these competing points of view by comparing neutral and selectively relevant genetic diversity using whole genome resequencing data from six species of kangaroo rats (*Dipodomys*), three of which are imperiled and three of which are abundant and not of

conservation concern. Kangaroo rats belong to a unique clade of rodents (Heteromyidae: Dipodomyinae) that have adapted physiologically and behaviorally to life in the hot, dry North American southwest (Schmidt-Nielsen and Schmidt-Nielsen 1952; Tracy and Walsberg 2002; Rocha et al. 2021). The twenty-two species in the genus have mixed demographic trajectories in the Anthropocene: some are widespread, common, and ecologically flexible, while other species with limited distributions, experiencing habitat loss and fragmentation, or sensitivity to disturbance are declining (Goldingay et al. 1997; Longland and Dmitri 2021). The IUCN categorizes eight species of kangaroo rats as decreasing and six of these as near threatened (*D. spectabilis*), vulnerable (*D. elator*, *D. nitratoides*, *D. stephensi*), or endangered (*D. ingens*, *D. gravipes* – possibly extinct).

We generated whole genome resequencing (WGS) data for 142 individuals across six species of kangaroo rats (Figure 1). These include the endangered Giant kangaroo rat (*D. ingens*) and threatened San Joaquin kangaroo rat (*D. nitratoides*) and Stephens' kangaroo rat (*D. stephensi*), as well as non-threatened Merriam's kangaroo rat (*D. merriami*), Panamint kangaroo rat (*D. panamintinus*) and Heermann's kangaroo rat (*D. heermanni*). With these data, we asked 1) Do whole-genome resequencing data recapitulate previously reported phylogenetic relationships between species? 2) Do endangered species show reduced genetic diversity (measured as heterozygosity) or increased runs of homozygosity compared with non-threatened species? 3) How does historic effective population size compare across species, and which better predicts genetic diversity, contemporary range size or past effective population size? 4) Does the distribution of fitness effects and the efficacy of selection differ between endangered and abundant species?

We report no consistent difference in genome-wide genetic diversity between threatened and non-threatened kangaroo rats: we observed high heterozygosity in the vulnerable San Joaquin kangaroo rat and strikingly low heterozygosity in the non-threatened Panamint kangaroo rat considering their abundance in the Great Basin Desert. Instead, patterns of genetic diversity are correlated with inferred historic effective population size. However, closer examination at the subspecies level revealed reduced heterozygosity and strong evidence of inbreeding in a now-extinct population of the San Joaquin kangaroo rat, *D. nitratoides exilis*, despite high background heterozygosity in this species. Further, our comparison of the distribution of fitness effects across species suggests that both demographic history and present demographic status shape mutational load. Hence, we cannot rule out inbreeding or genetic load as concerns for vulnerable species with high neutral genetic diversity. Our results suggest that, by itself, the level of neutral genetic diversity is not a good indicator of extinction risk in kangaroo rats. However, comparing genetic diversity across closely related species or populations can provide useful information about current dynamics of natural selection and genetic drift in species of concern for endangered species management. We suggest that while loss of genetic diversity is not a primary risk factor for extinction in *D. ingens* and *D. nitratoides*, conservation actions are still needed to protect habitat for these declining species.

4.2 METHODS

Sample processing and library preparation

All individuals included in this manuscript (*D. merriami* n = 41, *D. panamintinus* n = 47, *D. heermanni* n = 21, *D. nitratoides* n = 21, *D. ingens* n = 8, *D. stephensi* n = 6) are vouchered

specimens in the mammal collection at the Museum of Vertebrate Zoology (MVZ), Berkeley, CA. MVZ accession numbers, sex, locality, and collecting date for each specimen are available in Supplementary Table 1. All DNA extraction and library preparation was performed in the Evolutionary Genetics Laboratory at the MVZ and sequencing was performed at the Vincent J. Coates Genome Sequencing Laboratory at UC Berkeley as part of the California Conservation Genomics Project (CCGP) (Shaffer et al. 2022).

For each specimen, flash frozen tissue stored in liquid nitrogen was subsampled to approximately 10 mg of tissue and DNA was extracted using the bead-based Mag-Bind® Blood & Tissue DNA HDQ96 Kit (Omega Bio-Tek). We assessed whether extraction was successful by running eluted DNA on a 1% agarose gel, assessed DNA for quality with a Nanodrop spectrophotometer, and measured double-stranded DNA concentration using the Biotium PicoGreen AccuClear® Ultra High Sensitivity dsDNA Quantitation Kit on a SpectraMax M2 fluorescence plate reader. DNA extraction was repeated for low concentration samples where possible. If multiple extractions were performed per sample, they were merged and all samples were purified and reconcentrated with a SPRI bead clean-up step before dilution to a standard concentration. To fragment DNA to an appropriate length (300-500 bp) for short-read library preparation, samples were sonicated with a qSonica instrument and assessed for fragment length on a 1.5% agarose gel. Lastly, we performed a size selection step with low ratio SeraMag SPRI beads.

We prepared 150 bp paired-end Illumina libraries using the Kapa HyperPrep Kit according to manufacturer's instructions. We verified whether library preparation was successful and quantified DNA concentration with a Nanodrop spectrophotometer or Invitrogen Qubit fluorometer, before assessing library fragment length with the Agilent DNA1000 Assay on a BioAnalyzer 2100. Libraries were sequenced to an average of 10x coverage across six lanes of an Illumina NovaSeq 6000 S4 and returned as fastq files.

Variant Calling

Raw reads were mapped to the *D. merriami* reference genome (Voss et al. 2025; NCBI: GCA_024711535.1) and variants called using snpArcher, a snakemake workflow developed by the California Conservation Genomics Project (Mirchandani et al. 2024; https://github.com/ccgproject/ccgpWorkflow). In brief, raw fastq files for each sample were provided to the snpArcher Snakemake environment along with the reference genome fasta file. The snpArcher workflow uses fastp to trim and filter reads, Sentieon bwa-mem to map reads to the reference genome and remove PCR duplicates, Sentieon haplotyper to call variants, and GATK VariantFiltration to filter out indels, low quality reads, and sites with extremely high or low coverage.

After variant calling and basic filtering steps were completed through snpArcher, we performed a second round of filtering with vcftools v0.1.16 (Danecek et al. 2011; 2021). We removed indels and multi-allelic sites in addition to any sites where read depth was less than 5 or greater than 30, where more than 20% of individuals were missing data, or where the quality score was lower than 40. We produced two VCF files: one with a minimum minor allele frequency (MAF) of 0.02 applied, and one with no MAF filter for use in demographic inference, where filtering out low frequency alleles can skew results. Scaffolds with no variants called were filtered out of all downstream analyses. We also ran quality checks on individuals included in the

VCF and removed two individuals with low mean depth of coverage and high missingness in the snpArcher quality control report.

Population Genetics Analyses

To confirm species assignment and data quality, we performed principal components analysis on 5,502,172 linkage-pruned SNPs using plink v1.90b6.21 (Chang et al. 2015). Linkage was assessed in 50 kb sliding windows with a 10 kb window step size, and SNPs were considered linked and pruned if the r^2 between them was greater than 0.1. The relationships between focal species are well resolved and no evidence of hybridization has been found between the taxa included in this study. However, genomic data for the genus is sparse, so we used ADMIXTURE v1.3.0 (Alexander and Lange 2011) for further data exploration with the same set of SNPs. At this point, we re-assigned two individuals that had been misidentified as *D. ingens* to *D. heermanni* based on admixture and PCA results. We also constructed a phylogenetic tree based on 2,159 BUSCO genes present in all 142 samples as well as the *D. spectabilis* reference genome (Harder et al. 2022: GCF_019054845.1) and two reference genomes for *Perognathus longimembris* (Wilder et al. 2022: GCF_023159225.1; Kozak et al. 2023: GCA_024363435.2), which served as outgroups.

As *D. spectabilis* and *P. longimembris* were not present in our VCF file, we built a rooted tree based on single-copy orthologous sequences as described in Robinson et al. (2022): for each individual in the WGS dataset, we created a whole genome fasta file based on the *D. merriami* reference genome and variants captured in the VCF file using bcftools consensus v1.21 (Danecek et al. 2021). We then used BUSCO compleasm v0.2.6 (Manni et al. 2021) to identify orthologs from the Glires odb10 database present in all individuals and seqkit subseq to extract BUSCO gene sequences (Shen et al. 2016). For every BUSCO ortholog that was identified in all 145 samples (n = 2,159 BUSCOs present in 142 *Dipodomys* WGS samples and three outgroups), we aligned sequence data from all individuals using MAFFT v7.525 (Katoh and Stanley 2013) and trimmed alignments with trimAl v1.4.1 (Capella-Gutiérrez et al. 2009). We then used IQ-TREE v. 2.1.4-beta (Minh et al. 2020) to construct a rooted tree with *P. longimembris* specified as the outgroup and rate model GTR+F+R2 to account for the use of multiple loci that may evolve at different rates. Lastly, we performed 1000 bootstraps to evaluate tree support and visualized the bootstrapped tree with R packages ape v5.8 (Paradis and Schliep 2018) and ggtree v 3.12.0 (Yu et al. 2017). All figures and visualizations included in this manuscript were created in R v4.4.0 (R Core Team 2024) with ggplot2 v3.5.1 (Wickham 2016) and formatted in Adobe Illustrator.

Conservation Genomic Analyses

Neutral theory predicts that small populations will have low genetic diversity due to the random loss of variation via genetic drift (Kimura 1983). To test whether this holds for endangered kangaroo rats, we first calculated genome-wide heterozygosity on a per-individual basis using bcftools het (Li et al. 2011) to count heterozygous sites observed for each individual and divided this value by the number of callable sites in the genome (1,970,346,594 bp). We then compared mean individual heterozygosity by species and subspecies and tested for differences between groups with an ANOVA followed by a Tukey post-hoc test in R. We also calculated genome-wide π for each species and subspecies with scikit-allel (https://github.com/cggh/scikit-allel) and Tajima's D in non-overlapping sliding 50kb windows with vcftools (Danecek et al. 2011).

Inbreeding between closely related individuals results in long runs of homozygosity (ROH). To explore the possibility of recent inbreeding in *Dipodomys* species of concern, we next used bcftools roh to identify runs of homozygosity longer than 100 kb, calculated the proportion of the genome contained in ROH (f_{ROH}) for each individual, and compared f_{ROH} across species and subspecies. We also calculated f_{ROH} for the subset of runs of homozygosity that were longer than 1 Mb. To visualize patterns of ROH across the genome, we used bcftools and a custom bash script to calculate heterozygosity in non-overlapping 500 kb sliding windows, which we plotted with R.

Inferring Historic Effective Population Size

As genetic diversity is influenced by both contemporary population size and demographic history, we used Pairwise Sequentially Markovian Coalescent (PSMC) v0.6.5 (Li and Durbin 2011) to infer past effective population size for each species. PSMC can be sensitive to missing data (Mather et al. 2020), so we selected the individual with the highest mean depth of coverage and lowest missingness from each species for analysis. Using other individuals for each species produced similar results. For these high coverage individuals, we selected variants with no missing data on autosomal scaffolds greater than 5 Mb in length. We used bcftools consensus and PSMC fq2psmcfa to create whole genome fasta sequences from the VCF file with no minor allele frequency threshold, as frequency-based filtering can bias demographic inference.

PSMC can sometimes produce artifactual peaks and collapses in recent effective population size (N_e) if default settings are used (Hilgers et al. 2025), so we tested several options for the parameter that sets the number and size of time windows in which recombination events are identified. A value of -p "2+2+20*2+6" appeared to avoid such false population size peaks while still maintaining large enough windows of time for sufficient recombination events to occur, so we selected this setting and ran PSMC separately for each species with parameters -N35 -t15 -r5 -p "2+2+20*2+6", with the exception of *D. stephensi*, for which we used -p "2+2+20*2+6". We performed 50 bootstraps for each PSMC run and plotted the results in R, using the average mammalian mutation rate of 2.2×10^{9} mutations/site/generation (as in Harder et al. 2022) and a generation time of 1.5 years (pers. comm., J. Patton).

To compare the impact of past and present effective population size on heterozygosity, we performed linear regressions to test for correlation between 1) present species range size and heterozygosity and 2) past effective population size and heterozygosity. We obtained species range sizes from IUCN range polygons (IUCN 2024) and extracted range area (km^2) using R package sf v.1.0-16. We estimated historic effective population size for each species by taking the PSMC average effective population size from time windows corresponding to 10,000 to 250,000 years ago.

Identifying synonymous and nonsynonymous variants

To identify selectively relevant genetic variation, we used the *D. merriami* gene annotation described in Voss et al. (2025) to identify CDS coding regions of the genome and built a custom *D. merriami* variant effect prediction database with SIFT4G_Create_Genomic_DB (https://github.com/pauline-ng/SIFT4G_Create_Genomic_DB/). As the annotation described above is lifted over from *D. spectabilis*, we only included coding sequences with at least one open reading frame and 99% or greater coverage in the liftover annotation. We then ran SIFT-4G (Kumar et al. 2009; Vaser et al. 2016) on our complete 142-individual VCF file with this database. SIFT-4G outputs an annotated VCF that classifies each

variant as non-coding, synonymous, nonsynonymous, or start/stop codon gain/loss and identifies which mutations are likely to be deleterious based on predicted changes to amino acid identity and protein structure.

Estimating the Distribution of Fitness Effects

The distribution of fitness effects (DFE) for segregating mutations can be estimated by comparing the synonymous and nonsynonymous site frequency spectra (SFS) for a population of interest (Tataru and Bataillon 2020). We used the fastDFE (Sendrowski and Bataillon 2024) package in python to construct folded SFS and estimate the DFE for populations from four of six species included in this study.

As CCGP sampling was evenly distributed to maximize landscape coverage, analyses based on population samples must be conducted with care. To identify populations suitable for estimating the site frequency spectrum, we sub-sampled each species to include specimens captured in close geographic proximity. We confirmed that they were valid population samples by calculating isolation by distance (IBD) with plink and proceeding only with populations with no evidence of IBD (Supplementary Table 3). There were no valid population samples available for *D. stephensi* or *D. panamintinus*, but we were able to construct SFS and estimate the DFE for *D. nitratoides* (VU), *D. merriami* (LC), *D. ingens* (EN), and *D. heermanni* (LC).

For each population, we first used fastDFE's built in vcf parser to construct synonymous and nonsynonymous site frequency spectra, which we folded and passed to fastDFE's BaseInference. FastDFE first uses the synonymous SFS to model demography and then compares nonysynonymous and synonymous SFS to infer the distribution of fitness effects. With folded site frequency spectra, the DFE is limited to deleterious variation, which is appropriate given that we are interested in the impact of genetic drift and potential accumulation of deleterious mutations in small populations. We ran fastDFE with a mixed gamma-exponential distribution and performed 1,000 runs and 500 bootstraps for each population before evaluating model fit based on log-likelihood values and similarity between observed and modeled SFS. For *D. ingens*, *D. nitratoides*, and *D. heermanni*, parameters *S_b* (overall S for beneficial selection), *p_b* (probability a mutation is beneficial), and *eps* (false positive identification as derived allele) were set at a fixed value to account for use of a folded site frequency spectrum. Those parameters were not fixed for *D. merriami* because the resulting model fit was poor.

4.3 RESULTS

Whole genome resequencing data

We generated short-read whole genome resequencing data to an average depth of 10x coverage for 142 individuals sampled from across central and southern California, including 41 *D. merriami*, 21 *D. nitratoides*, 8 *D. ingens*, 6. *D. stephensi*, 21 *D. heermanni*, and 47 *D. panamintinus* individuals (Fig. 1). After quality filtering, we retained a set of 98,490,369 SNPs, which was reduced to 55,047,119 SNPs when a minor allele frequency filter of 0.02 was applied. For admixture and principal components analyses, this set of variants was linkage pruned (window size = 50kb, sliding window = 10 kb, $r^2 = 0.1$) to a set of 5,502,172 SNPs.

Evolutionary relationships between and within species

Genome-wide principal components analysis with this set of 5,502,172 SNPs recapitulates species relationships and broadly agrees with published *Dipodomys* phylogenies (Alexander and Riddle 2005). *D. merriami* and *D. nitratoides* form a species complex with two insular species of kangaroo rats that are restricted to islands in Baja California (Álvarez-Castañeda et al. 2009), and these two species cluster tightly together when PC1 (35.6% variance explained) and PC2 (25.6% variance explained) are plotted (Fig. 2B). Similarly, *D. panamintinus* and *D. heermanni*, previously described as sister taxa (Alexander and Riddle 2005), cluster adjacent to each other, while *D. stephensi*, thought to form a clade with the *panamintinus-heermanni* pair, is closest to those species on both PCs 1 and 2, and *D. ingens* appears distant from other lineages. When higher-order principal components PC4 (4.68% of variance explained) and PC5 (2.29% of variance explained) are plotted, samples belonging to sister taxa pairs form distinct clusters by species as indicated in the phylogenetic topology recovered (Supp. Fig. 2).

To place principal components in an evolutionary context, we constructed a phylogenetic tree from 2,159 BUSCO genes that were present in all 142 samples in our WGS data as well as the spectacled kangaroo rat (*D. spectabilis*; Harder et al. 2022), and little pocket mouse *Perognathus longimembris longimembris* (Kozak et al. 2024) and *P. longimembris pacificus* (Wilder et al. 2023), which served as outgroup samples. The phylogeny is well-supported by bootstrapping and corroborates established relationships among the *Dipodomys* species present in our dataset. Major nodes separating species all have 100% bootstrap support (Fig. 2A), and most nodes within species show 85% or greater bootstrap support (full tree with bootstrap values available in Supplementary Figure 1). Though species relationships are well-resolved across *Dipodomys*, relationships between individuals within each species are less straightforward: while subspecies of *D. panamintinus* form distinct clades, *D. heermanni* samples do not group by subspecies. *D. nitratoides* samples also do not group by subspecies, with members of the extinct *D. n. exilis* population appearing in two different locations and the threatened *D. n. nitratoides* and *D. n. brevinasus* evenly distributed across the clade. *D. merriami parvus*, the endangered San Bernardino kangaroo rat, is nested within the *D. m. merriami* clade.

Admixture analyses also support each species as a unit and did not uncover any evidence of hybridization between species, though the best supported number of populations is k = 9 (CV error = 0.08709), suggesting that further genetic differentiation has accumulated between lineages within *D. merriami*, *D. nitratoides*, and *D. panamintinus* (Fig. 2B). Structure within *D. merriami* does not map clearly onto a particular subspecies, but in vulnerable San Joaquin kangaroo rat *D. nitratoides*, *D. n. brevinasus* and *D. n. nitratoides* form one population, while extinct *D. n. exilis* forms a second population. *D. panamintinus* has two subspecies, *D. p. caudatus* and *D. p. leucogenys*, that are disjunct from the remainder of the species range, and *D. p. leucogenys* forms a separate population within the Panamint kangaroo rat clade.

Surveying neutral genetic diversity and inbreeding

Across species, heterozygosity ranges from 0.049% per site in *D. stephensi* to 0.123% per site in *D. heermanni* (Table 1, Fig. 3A), which is approximately the same as in humans (1000 Genomes Project 2015). There is no difference in heterozygosity between endangered and non-endangered species (T-test $p = 0.48$). *D. panamintinus*, a species of least concern, shows lower heterozygosity (0.060% per site) than the endangered *D. ingens* (0.064% per site). *D. merriami*

displays moderate heterozygosity (0.086% per site), but less than the endangered *D. nitratoides* (0.102% per site). When further separated into subspecies, the extinct *D. n. exilis* (0.094% per site) has significantly lower heterozygosity than either *D. n. nitratoides* (0.106% per site, p = 0.0025) or *D. n. brevinasus* (0.104% per site, p = 0.0027, Tukey post-hoc test).

We report little evidence of inbreeding in *Dipodomys*. We calculated f_{ROH}, the proportion of the genome contained in runs of homozygosity longer than 100 kb, and no species has a mean f_{ROH} greater than 0.1 (Fig. 3B). F_{ROH} varies across species, but there is no difference in f_{ROH} between vulnerable and widespread species (p = 0.61). On a per-species basis, threatened *D. stephensi* has the highest proportion of the genome contained in ROH (f_{ROH} = 0.091), whereas *D. heermanni* has the lowest (f_{ROH} = 0.0406). However, non-threatened *D. merriami* (f_{ROH} = 0.0841) and *D. panamintinus* (f_{ROH} = 0.0761) had similar f_{ROH} to endangered *D. ingens* (f_{ROH} = 0.0820) and *D. nitratoides* (f_{ROH} = 0.0556). Further, most ROH we observed were relatively short (median ROH length = 282,498 bp), suggesting that inbreeding may have occurred in the more distant past, or between more distant relatives.

Dipodomys nitratoides exilis, the extinct subspecies, provides an exception to this pattern. Though a lower proportion of the genome is contained in runs of homozygosity for *D. nitratoides* (f_{ROH} = 0.0556) than *D. merriami*, some individuals from *D. n. exilis* display remarkably long runs of homozygosity. Three out of five individuals have multiple ROH greater than 10 Mb in length, one of which displays two ROH longer than 20 Mb (Fig. 3C). Further, when we eliminate short (less than 1Mb) runs of homozygosity, we report an f_{ROH} of 0.0207 for *D. n. exilis* versus an f_{ROH} of 0.0070 for *D. n. nitratoides*. Anecdotally, two additional endangered or extinct lineages that are closely related to more widespread species show reduced genetic diversity and a greater proportion of the genome captured by runs of homozygosity compared with a widespread counterpart. Our sampling is limited for these lineages, but two *D. h. berkeleyensis* (extinct) individuals have a mean heterozygosity of 0.094% per site (versus *D. heermanni* 0.12% per site), and one *D. m. parvus* (endangered) individual has a genome-wide heterozygosity of 0.062% per site (versus *D. merriami* 0.086% per site). Further, we report a long (1 Mb) f_{ROH} of 0.011 for extinct *D. h. berkeleyensis* compared with 1Mb-f_{ROH} = 0.0060 for related *D. h. tularensis*. Disjunct populations of the Panamint kangaroo rat also have more long ROH (*D. p. leucogenys* 1Mb f_{ROH} = 0.0171) than individuals in the core of the range (*D. p. mohavensis* 1 Mb f_{ROH} = 0.0048).

Historic effective population size, present range size, and genetic diversity

We inferred historic effective population size (N_e) over the past one million years for each species using PSMC and the highest coverage individual from each species (listed in Supplementary Table 2). Trajectories of historic effective population size varied extensively across species (Fig. 4A). Endangered species *D. ingens* and *D. stephensi* have had lower N_e over the past million years than *D. nitratoides*, while *D. heermanni* shows the largest overall historic population size, and its sister species *D. panamintinus*, may have had a much smaller effective population size. PSMC infers a population bottleneck in widespread *D. merriami* approximately 70,000 to 100,000 years ago, followed by a recent population expansion. Though range size is not correlated with genome-wide heterozygosity (Fig. 4B, Spearman correlation coefficient, R^2 = 0.24, p = 0.36), inferred historic effective population size, taken as an average from 10,000 to 250,000 years ago, is strongly correlated with heterozygosity (Fig. 4C, Spearman correlation coefficient, R^2 = 0.89, p = 0.017). This result is robust to the choice of individual used for PSMC inference.

Deleterious variation and the distribution of fitness effects

To estimate the distribution of fitness effects, we used the genome annotation from Voss et al. (2025) to parse nonsynonymous and synonymous sites and constructed folded site frequency spectra (SFS) for each. The shape of the synonymous site frequency spectrum is a function of the demographic history of the population, and the difference between the nonsynonymous and synonymous spectra is driven by the relative strength of genetic drift and natural selection in the population (Eyre-Walker et al. 2006; Keightley and Eyre-Walker 2007; Lanfear et al. 2014; Tataru et al. 2017). We constructed these paired SFS in a subset of individuals from four of six species: *D. merriami merriami* (n = 13; Fig. 5A), *D. nitratoides brevinasus* (n = 8; Fig. 5C), *D. heermanni swarthi* (n = 5; Fig. 5E) and *D. ingens* (n = 6; Fig. 5G). SFS are shown unfolded but were folded for DFE inference.

fastDFE was run with a mixed gamma-exponential distribution for all sample populations with best fit model parameters reported below. fastDFE reports fitness effects as S, which is equivalent to the selection coefficient times the effective population size ($N_e s$) and divides S into S-deleterious (S_d) and S-beneficial (S_b). In keeping with PSMC inference, we observed an L-shaped neutral SFS *D. merriami merriami* in San Bernardino County, suggesting that the population has expanded rapidly (Fig. 5A). The best fit model for the *D. merriami merriami* DFE had parameters S_d = -9.871, b = 0.139, p_b = 0.04, S_b = 8.09, *eps* = 0.13, likelihood = -100.88. Because we focused on negative fitness effects, we fixed parameters related to beneficial mutations p_b and b and derived allele identification error *eps* at p_b = 0, *eps* = 0, and b = 1.0 in the remaining three populations. As fastDFE models with fixed parameters fit poorly for *D. m. merriami*, the program was allowed to determine the best fit value for each parameter freely within standard bounds. For *D. n. brevinasus* the best fit model had an S_d = -3.42 and a likelihood = -509.31 (Fig. 5D), for *D. ingens* S_d = -0.167 with a likelihood = -448.16 (Fig. 5F), *D. h. swarthi*, S_d = -1.99 with likelihood = -492.35 (Fig. 5H).

The distribution of fitness effects has a complex, multi-modal shape (Eyre-Walker and Keightley 2007), which is impacted by demography. Since genetic drift is weaker in large populations, mildly deleterious mutations are purged more effectively by natural selection, resulting in a DFE centered at 0 (neutrality) with a very short tail. In small populations, a longer tail of weakly to moderately deleterious variations is predicted because genetic drift is stronger (Robinson et al. 2023). DFE estimated for four populations of California kangaroo rats are consistent with this expectation: non-threatened *D. m. merriami* from San Bernardino County and *D. h. swarthi* have very narrow DFE with most fitness effects (S) between 0 and -0.01 (Fig. 5B, H). In contrast, DFE estimated for threatened *D. n. brevinasus* and endangered *D. ingens* have more leftward skew and a long tail of mutations with $S < 0.01$ (Fig. 5D, F).

4.4 Discussion

We present whole-genome resequencing data for six species of kangaroo rats and 142 individuals in California. Previously reported relationships among species are supported by these data, and there is no evidence of admixture between species. However, in accordance with recent mitochondrial studies by Patton et al. (2019) and Benedict et al. (2019), samples designated as different subspecies in *D. nitratoides*, *D. heermanni*, and *D. merriami* do not form reciprocally monophyletic groupings, and divergence within these species is shallow.

Dipodomys genetic diversity does not reflect conservation status: there is no significant difference in genome-wide heterozygosity between threatened and non-threatened species, and contemporary range size is not correlated with genetic diversity (Fig. 4B). We observe similar genetic diversity in the threatened Stephens' kangaroo rat (*D. stephensi*), endangered Giant kangaroo rat (*D. ingens*), and non-threatened Panamint kangaroo rat (*D. panamintinus*). Despite their similar range size to *D. panamintinus*, individuals belonging to *D. heermanni* display twice as much heterozygosity, and the endangered San Joaquin kangaroo rat (*D. nitratoides*) is nearly as heterozygous (Fig. 3A).

Genetic diversity reflects historic, not contemporary, population size in kangaroo rats

Rather, patterns of intraspecific genetic diversity in kangaroo rats appear to be driven by historic effective population size (N_e), as individual heterozygosity is strongly correlated with average effective population size taken between 10,000 and 250,000 years ago (Fig. 4C). Per PSMC inference, genetically diverse *D. heermanni* had the greatest historic N_e over the past million years, and endangered *D. nitratoides* had a larger N_e than widespread sister species *D. merriami*, which has expanded dramatically in the last 50,000 years (Fig. 4A). *Dipodomys ingens*, *D. stephensi*, and *D. panamintinus* have had smaller long-term effective population sizes, which is reflected in their lower genetic diversity.

These results suggest that genetic diversity is not a useful metric for extinction risk in kangaroo rats. Not only is genetic diversity not associated with conservation status, it is not particularly low: we report genome-wide heterozygosity values on the order of magnitude observed in human populations (Leffler et al. 2012). These findings corroborate prior reports of high mitochondrial and microsatellite genetic diversity across threatened *D. nitratoides*, *D. ingens*, and *D. stephensi* (Metcalf et al. 2007; Loew et al. 2009; Patton et al. 2019; Statham et al. 2019). Encouragingly, this implies that genetic diversity is not a primary risk factor for extinction in these species, and further, that kangaroo rats may have the genetic potential for adaptation and resilience following human-induced population decline.

Inbreeding and genomic erosion in extinct and endangered lineages

However, we observe genomic signatures of population decline and extinction when we examine levels of heterozygosity and runs of homozygosity at the subspecies level. For example, the genetically diverse San Joaquin kangaroo rat (*D. nitratoides*) is restricted to the southern half of California's Central Valley, which has been intensively modified for agriculture and is often overlooked as a unique biological community (Germano et al. 2011). As early as the 1910s, the Fresno kangaroo rat (*D. n. exilis*), a distinctive eastern subspecies of *D. nitratoides*, was under threat due to habitat conversion (Grinnell 1920). Surveys of *D. n. exilis* identified a few small and isolated populations in the 1980s and early 1990s (Chesemore and Rhodehamel 1992; Morrison et al. 1996), but no surviving individuals have been documented since 1992 and the subspecies may be extinct (USFWS 1998b; 2010; Patton et al. 2019).

This study contains five *D. n. exilis* samples in addition to five Tipton kangaroo rats (*D. n. nitratoides*) and eleven short-nosed kangaroo rats (*D. n. brevinasus*). Though heterozygosity is relatively high across all three subspecies, it is significantly lower in *D. n. exilis* than *D. n. nitratoides* or *D. n. brevinasus* (Fig. 3A). Further, a greater proportion of the genomes of *D. n. exilis* individuals are captured by runs of homozygosity (f_{ROH}; Fig. 3B). Though f_{ROH} is higher in both *D. ingens* and *D. stephensi* than *D. n. exilis*, most runs of homozygosity in those species are short (< 1Mb), whereas we observe multiple very long ROH (> 10 Mb) in *D. n. exilis* (Fig. 3C).

Shorter runs of homozygosity can arise through matings between distant relatives or be retained from inbreeding events further back in the past, but the long ROH observed in *D. n. exilis* strongly imply the occurrence of inbreeding between close relatives shortly before extinction.

Another endangered heteromyid, the Pacific pocket mouse (*P. longimembris pacificus*), shows the same pattern of long runs of homozygosity on a background of high heterozygosity, and inbreeding depression has led to reduced reproductive success in Pacific pocket mice bred in captivity (Wilder et al. 2020; 2023). Though little is known about the impact of inbreeding depression in the Fresno kangaroo rat (*D. n. exilis*), the combination of large historic N_e and dramatic population decline, like that observed in the Pacific pocket mouse, suggests that this subspecies may have suffered from inbreeding depression too. Recessive, strongly deleterious alleles segregating in the formerly large population that would be purged in a long-term small population may have become fixed in small, isolated populations of *D. n. exilis* in the final years of their decline (Hedrick and Garcia-Dorado 2016; Kyriazis et al. 2021; Robinson et al. 2023).

Though evidence for inbreeding is limited outside of *D. n. exilis*, we report reduced heterozygosity compared with closely related taxa in several declining subspecies. Sample sizes are not sufficient for statistical testing, but anecdotally, *D. m. parvus*, the San Bernardino kangaroo rat, and *D. h. berkeleyensis*, which has been extirpated from the hills east of the San Francisco Bay, show lower genetic diversity and higher f_{ROH} compared with non-threatened lineages of the same species (Table 1). Other studies have reported decreased genetic diversity and increased inbreeding coefficients in the Giant kangaroo rat (*D. ingens*) and the San Bernardino kangaroo rat *D. m. parvus* (Blackhawk et al. 2016; Hendricks et al. 2020; USFWS 2024). Despite maintenance of high overall genetic diversity, other threatened kangaroo rats in fragmented landscapes, such as *D. n. nitratoides*, are likely experiencing genomic erosion, or reduction in genetic diversity due to population decline and fragmentation (Rubidge et al. 2012, Díez-del-Molino et al. 2018).

Genetic drift load and the distribution of fitness effects

Exploring mutational load and the efficacy of selection in endangered species has been a central focus of recent conservation genomic studies (e.g. Chen et al. 2016; Tian et al. 2022; Abascal et al. 2023; Hoffman et al. 2024). New mutations occur on a continuum of fitness impact (the distribution of fitness effects, or DFE) and can be advantageous, neutral, or deleterious (Eyre-Walker and Keightley 2007). Most segregating mutations are neutral, and the proportion of mildly and moderately deleterious mutations in a population is a function of the relative strength of drift and selection and thus of population size (Ohta 1992). If selection dominates, as is typical of large populations, nonsynonymous mutations will be rarer in the population and only weakly deleterious mutations will escape purging by natural selection, yielding a narrow distribution of fitness effects (Lanfear et al. 2014). If genetic drift is strong, as is predicted in small populations, then more moderately deleterious mutations will remain in the gene pool, and there will be a longer tail of negative fitness effects, resulting in a higher mutational load. In keeping with these predictions, Leroy et al. (2021) demonstrate that selection is weaker in narrowly distributed island songbird species than in continental species with larger ranges. However, more studies are needed to understand how selection functions in small populations with varied demographic histories (Robinson et al. 2023).

To assess the relative efficacy of selection at purging deleterious variation, we constructed site frequency spectra for synonymous and nonsynonymous sites for four *Dipodomys*

populations and used these SFS to infer the distribution of fitness effects. For several species in this study, we might predict different DFE based on historic and contemporary population size. For example, widespread Merriam's kangaroo rat has undergone rapid population growth in the past 50,000 years, and if contemporary population size is more important, we would predict a narrow, neutral DFE for *D. merriami*. If, by contrast, the genome is still recovering from genetic drift in the smaller historic population, we would predict a wider DFE with more segregating deleterious mutations. We observe a narrow DFE in *D. m. merriami*, and most segregating mutations appear to be neutral or nearly neutral, with most *S* between 0 and -0.01 (Fig. 5B). This suggests that genetic drift is weak and natural selection is able to effectively purge mildly deleterious mutations, as we might predict based on the large contemporary population size of this widespread, abundant species.

If the DFE were driven mostly by historic effective population size, we would expect to see a similar DFE in *D. nitratoides* as in *D. merriami*: genetic diversity is higher in *D. nitratoides* even though the contemporary range of *D. nitratoides* is one tenth the range size of *D. merriami*. However, in *D. n. brevinasus*, selection against mildly to moderately deleterious mutations appears weaker than in *D. m. merriami* despite large historic N_e: we observe a broader distribution of fitness effects and a long tail of mutations with more negative *S* (Fig. 5C). As such, the San Joaquin kangaroo rat is likely accumulating weakly deleterious mutations even as neutral genetic diversity remains high.

The DFE for the Giant kangaroo rat *D. ingens* is intermediate between *D. n. brevinasus* and *D. m. merriami*, with a moderate tail of mutations with negative *S* (Fig. 5D). Though *D. ingens* persists in two main populations 150 km apart in the western Central Valley, all samples in this study are derived from the southern population, which is larger and shows greater connectivity across subpopulations (Statham et al. 2019; USFWS 2020a; 2020c). This lower fragmentation may allow the southern giant kangaroo rat to avoid the genomic erosion and genetic load that we observe in more fragmented species, such as in *D. nitratoides*.

Unanswered questions regarding population fragmentation

Specimens for this study were selected from voucher specimens in museum collections. This provides an opportunity to connect genotypes with phenotypes in future work. However, the use of museum specimens also comes with intrinsic sampling limitations. As such, some important questions remain unanswered. For example, this study contains six Stephens' kangaroo rat individuals, several of which are historic samples from before 1990. While these data can provide a general sense of genetic diversity within *D. stephensi*, it would be helpful to have more modern sampling to assess contemporary inbreeding and genetic load. Additionally, the Giant kangaroo rat's range is divided into two primary population centers, but all samples included in this study are derived from the southern population. Statham et al. (2019) report that the northern and southern lineages diverged several thousand years ago and should potentially be managed as distinct units, but we are not able to address questions related to population structure or the partitioning of genetic variation within the Giant kangaroo rat with these data. Sampling for the San Joaquin Valley kangaroo rat (*D. nitratoides*) is more complete, but gaps remain in areas around putative transition zones between subspecies. More sampling might help clarify our findings that subspecies are not monophyletic. Further, we lack sufficient sampling from the Tipton kangaroo rat (*D. n. nitratoides*) to address concerns regarding population fragmentation within this subspecies. We hope future conservation genomics efforts will address more questions related to population structure in California *Dipodomys*.

Genetic diversity in the context of kangaroo rat conservation biology

Though genomic data from three species of threatened kangaroo rats suggest that limited genetic diversity is not a primary threat to their survival, two endangered lineages in California are at critically high risk of extinction: the Tipton kangaroo rat (*D. n. nitratoides*) and San Bernardino kangaroo rat (*D. m. parvus*). Both subspecies display relatively specialized habitat preferences: *D. n. nitratoides* is mostly found in open areas with alkali sink patches in the southern Central Valley, while *D. m. parvus* is restricted to river-adjacent alluvial plains in Southern California, and this ecological specialization contributes to extreme habitat fragmentation and reduction. Less than 1.5% of the historic Tipton kangaroo rat distributional range remains suitable for occupation (Patton et al. 2019), and urbanization and changing water regimes in the areas surrounding Los Angeles have reduced *D. m. parvus* available habitat such that only three populations persist, which display genetic signals of inbreeding (Hendricks et al. 2020). Unoccupied and undeveloped habitat for both species is scant, and conservation challenges for *D. n. nitratoides* are exacerbated by competition with other kangaroo rats: *D. nitratoides* often co-occurs with generalist *D. heermanni* but is smaller and avoids interspecific interactions, leading translocations of *D. n. nitratoides* to sites with *D. heermanni* present to fail (Tennant et al. 2013; USFWS 2020a).

The conservation outlook is slightly better for the Giant kangaroo rat (*D. ingens*) and Stephens' kangaroo rat (*D. stephensi*). Several large habitat preserves have been created to protect the Giant Kangaroo rat and other San Joaquin Valley endemic species (Widick and Bean 2019), and *D. ingens* outcompetes other kangaroo rats when they co-occur. However, there were just six *D. ingens* populations remaining when surveyed in 2020, making the species vulnerable to stochastic threats such as wildfire, drought, and disease (USFWS 2020a). *D. stephensi*, despite having the lowest individual heterozygosity of any species surveyed, is on a promising population size trajectory and has recently been downlisted from endangered to threatened (USFWS 2022). Successful conservation efforts have protected one third of suitable *D. stephensi* habitat in southern California, and populations of Stephens' kangaroo rat have been stable or increasing (USFWS 2021).

More generally, conservation efforts have met with mixed success in kangaroo rats. Some conservation biology strategies are difficult to implement in Heteromyid rodents: for example, captive breeding is extremely challenging because kangaroo rats are solitary, territorial animals. Translocation has been somewhat more successful, but social relationships and intraspecific competition create difficulties here too: kangaroo rats that were translocated together with neighbors were more likely to survive and be recaptured at new sites (Tennant et al. 2013; Shier and Swaisgood 2012). Translocations have been conducted for Stephens' kangaroo rat (*D. stephensi*), the Giant kangaroo rat (*D. ingens*), and the Tipton kangaroo rat (*D. n. nitratoides*) but are often last-ditch efforts to relocate animals from development sites, and survival is low (Germano et al. 2013; Tennant et al. 2013; Saslaw and Cypher 2020; Cypher et al. 2021; Shier et al. 2021). Ultimately, California has a combination of diverse habitats and landscapes and substantial pressure from human habitation and modification that is hard to balance. From a genomic perspective, it is not too late to change the trajectory of species such as the Tipton kangaroo rat and San Bernardino kangaroo rat, but this window will close as genomic erosion and fragmentation continues.

Conservation genomics for imperiled but not critically endangered species

In recent years, conservation geneticists have published several accounts of species on the brink of extinction, such as the Florida panther, the California condor, and the Devil's hole pupfish (Robinson et al. 2021; Ochoa et al. 2022; Tian et al. 2022). These species' genomes are dramatically altered by their declines: researchers report long runs of homozygosity, extremely low genetic diversity, and often an excess of fixed loss-of-function mutations. Case studies such as these point the way towards options for genetic and demographic rescue in critically endangered species, but conservation genomics can also play a role in protecting species that have not yet reached the point of needing a captive breeding program for recovery.

Genomic data can be a powerful tool for conservation but must be interpreted with caution. Recent meta-analyses demonstrate that genetic diversity cannot accurately predict whether a species will be on the IUCN Red List, which is based on species range, population size, habitat quality and fragmentation (IUCN 2012), nor does Red List membership identify species with low genetic diversity (Schmidt et al. 2023). As we demonstrate in this study, it is critical that conservation geneticists not write off threats to endangered species with high genetic diversity. Species such as the Fresno kangaroo rat are nearing extinction despite high levels of genetic diversity retained from large historic populations, while some lineages of common species, such as disjunct *D. panamintinus leucogenys* in southeastern California, have low genetic diversity and elevated runs of homozygosity. These genetically depauperate lineages may warrant flagging as taxa with low adaptive potential. Simultaneously, we affirm the utility of genomic data in understanding shifting population dynamics, population structure, and genetic load. Comparative sampling across closely related units or populations can provide context regarding demographic history and may not require costly whole genome resequencing; data presented here corroborate prior findings based on microsatellite and mitochondrial DNA studies in *Dipodomys*.

However, deep sequencing across the genome can identify instances of inbreeding via runs of homozygosity in the absence of large sample sizes, which are often impossible to acquire for endangered species. While neutral genetic diversity is mostly driven by historic effective population size in this study, the DFE seems to reflect a mixture of past and present demography, suggesting that regions of the genome that impact fitness may be more sensitive to changes in population size than overall genetic diversity. Modeling the DFE may thus be a way for conservation geneticists to investigate the impact of genetic drift in smaller populations before overall genetic diversity is affected and could be used as an indicator or warning sign to identify shifts in a population's demographic and evolutionary trajectory. Lastly, though not a focus of this study, temporal sampling using museum specimens collected over time can allow researchers to quantify genomic erosion and genetic diversity loss, which may be particularly useful for species with high background genetic diversity (Díez-del-Molino et al. 2018; Benham, Walsh and Bowie 2024).

Conclusion

Recent frameworks set out by the Convention on Biological Diversity formalize protection of intraspecific genetic diversity as a goal equal to the conservation of species and ecosystem diversity (Hoban et al. 2020; 2021). Multispecies genetic sequencing efforts such as the California Conservation Genomics Project represent a promising step towards integration of genetic diversity with larger biodiversity protection efforts (Shaffer et al. 2022; Heuertz et al.

2023). A growing body of literature suggests that conservation status and genetic diversity do not map neatly onto one another. Thus, genomics-informed management may play a key role in selecting among conservation strategies for endangered species with varied demographic histories and evolutionary potential for resilience to a changing landscape.

4.5 FIGURES

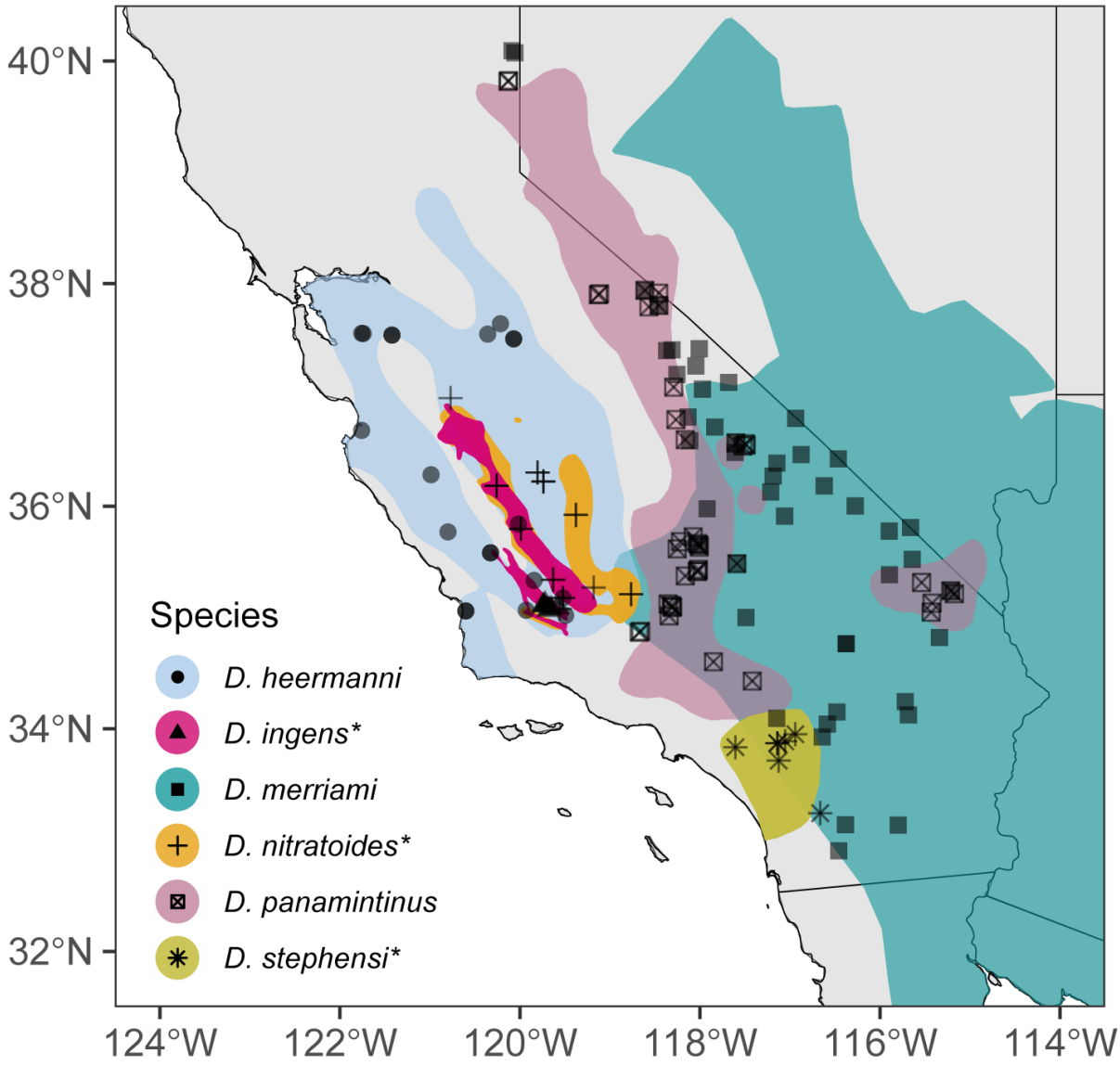


Figure 1. Species ranges and sampling locality map for 142 individuals from six species of kangaroo rats (*Dipodomys*) in California, three of which are endangered or threatened (indicated with asterisk), and three of which are of least conservation concern. We generated 10x-coverage whole genome resequencing short-read data for each individual, including *D. heermanni* (n = 21), *D. ingens** (n = 8), *D. merriami* (n = 44), *D. nitratoides** (n = 21); *D. panamintinus* (n = 47), and *D. stephensi** (n = 6). Species ranges are indicated by color extent on the map with sampling localities overlaid as black symbols; both color and symbol for each species are defined in the figure legend. All samples are vouchered specimens in the Museum of Vertebrate Zoology (University of California, Berkeley), and sample accession number and latitude and longitude for each collection locality are available in Supplementary Table 1.

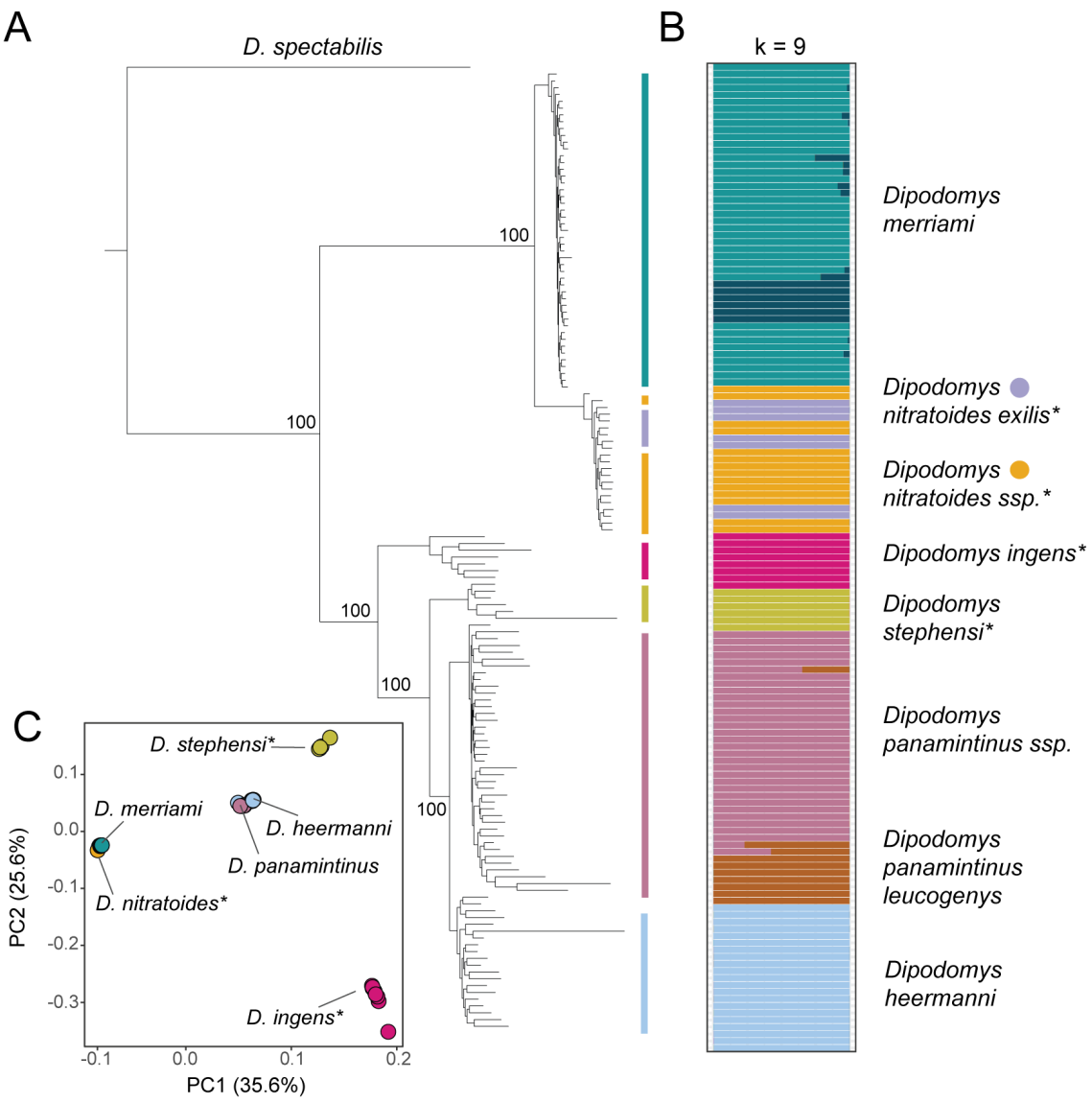
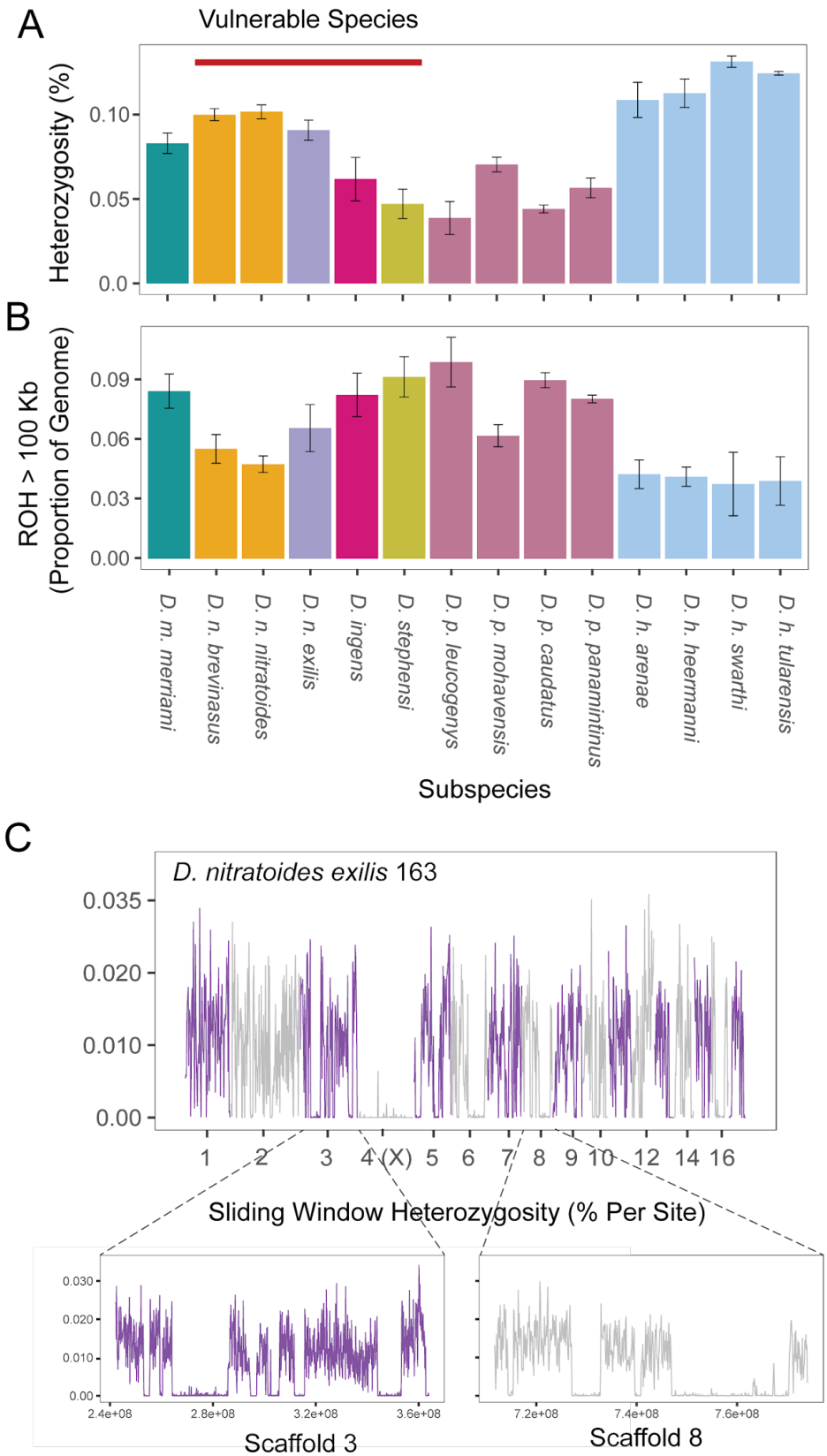


Figure 2. Phylogenetic tree including all whole-genome resequencing samples (n = 142) and principal components analysis. A: Rooted phylogeny made with 2,159 BUSCO genes present in all samples in addition to reference genomes for outgroups *D. spectabilis* and *P. longimembris* (not shown; tree constructed with IQ-TREE). Vulnerable species are indicated with an asterisk, and a full tree with bootstrap values for all nodes is available in Supplementary Figure 1. All major nodes received 100% bootstrap support. B: Admixture population assignment including all WGS samples and 5,502,172 linkage-pruned SNPs. Samples are ordered such that they match the placement of tips in the phylogenetic tree in A. K = 9 had the lowest cross-validation error of any number of populations between 3 and 12 (CV error = 0.08709) C: Principal components analysis with the same set of 5,502,172 SNPs used in admixture analyses; PC1 (35.6% of variance explained) and PC2 (25.6% of variance explained) are shown. Samples in PC space are colored by species according to the same color scheme as in A and B.



D. nitratoides exilis - extinct since 1992

Figure 3. A) Mean heterozygosity summarized by subspecies. Bar height indicates mean heterozygosity and error bars indicate standard error. Vulnerable species are indicated in red. Taken together, there is no significant difference in mean individual heterozygosity between threatened and non-threatened species ($p = 0.48$). However, the extinct *D. n. exilis* has significantly lower heterozygosity than extant close relatives *D. n. nitratoides* (het. = 0.106% per site, $p = 0.0025$) and *D. n. brevinasus* (het. = 0.104% per site, $p = 0.0027$, Tukey post-hoc test). Full results from ANOVA comparing heterozygosity across species is available in Supp. Table. 4. B) Proportion of genome (f_{ROH}) contained in runs of homozygosity greater than 100kb in length by subspecies. Bar height indicates mean f_{ROH} and error bars indicate standard error. Full results from ANOVA comparing f_{ROH} across species is available in Supp. Table 4. C) Sliding window heterozygosity measured in 1-Mb bins across the longest sixteen scaffolds for individual 163 from *D. n. exilis* (extinct), which has the greatest genome-wide proportion of long runs of homozygosity. Inset below, long ROH are visible on scaffolds 3 and 8; scaffold 4 represents the X chromosome. Long runs of homozygosity are indicative of recent inbreeding between close relatives.

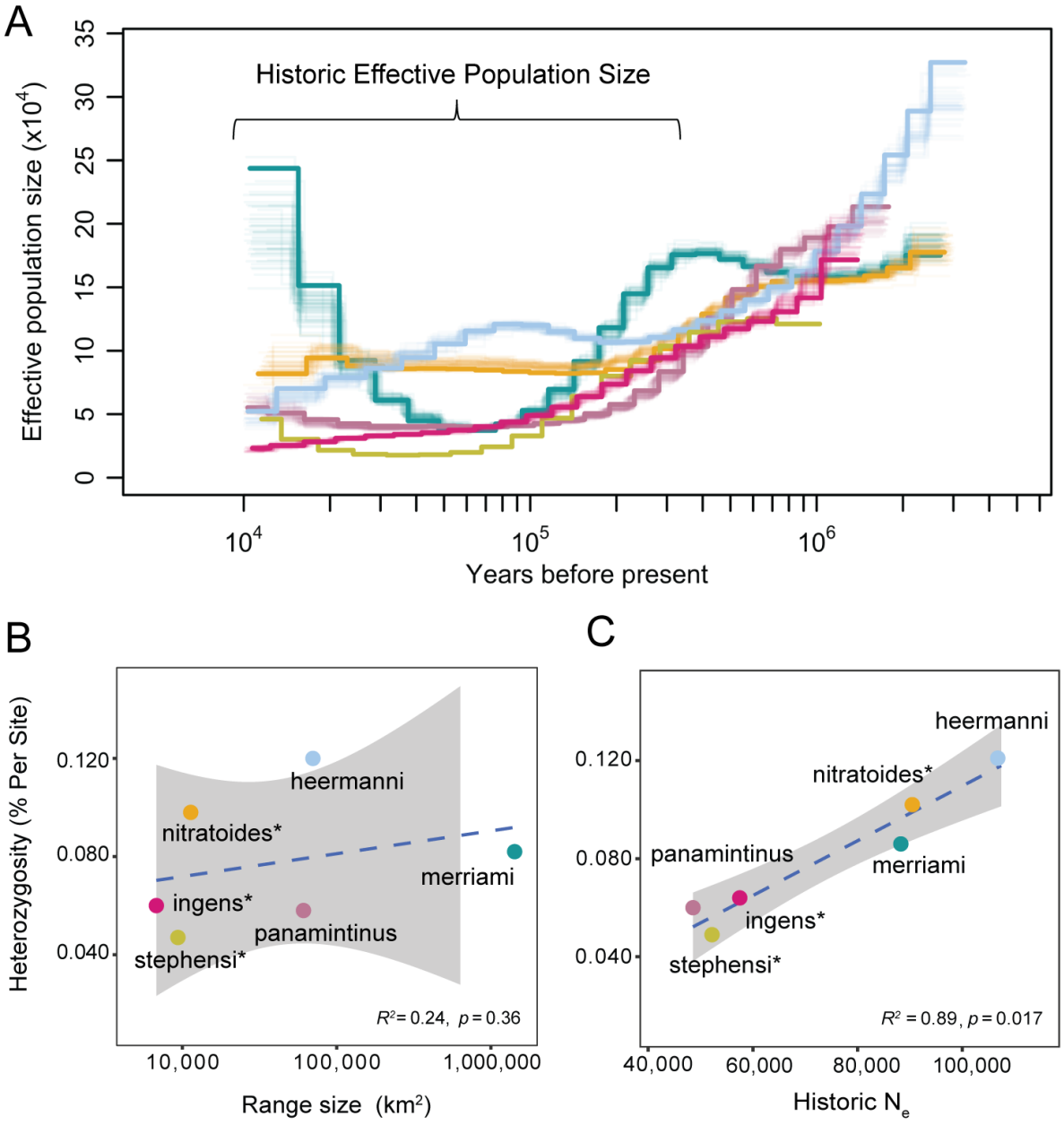


Figure 4. A) Inferred historic effective population size estimated using Pairwise Sequentially Markovian Coalescent (PSMC) analysis. Lighter colored lines represent 50 bootstrap runs of PSMC for each sample. PSMC was run using the highest coverage individual from each species and plotted with a generation time of 1.5 years and the average mammalian mutation rate of 2.2×10^{9} mutations/site/generation (as in Harder et al. 2022). B) Across species, individual heterozygosity is not correlated with contemporary range size in km^2 ($R^2 = 0.24$, $p = 0.36$), but C) is correlated with inferred historic effective population size from 10,000 to 250,000 years ago, shown with bracket ($R^2 = 0.89$, $p = 0.017$). Range data retrieved from IUCN and range size calculated with R sf package (Supp. Table 5).

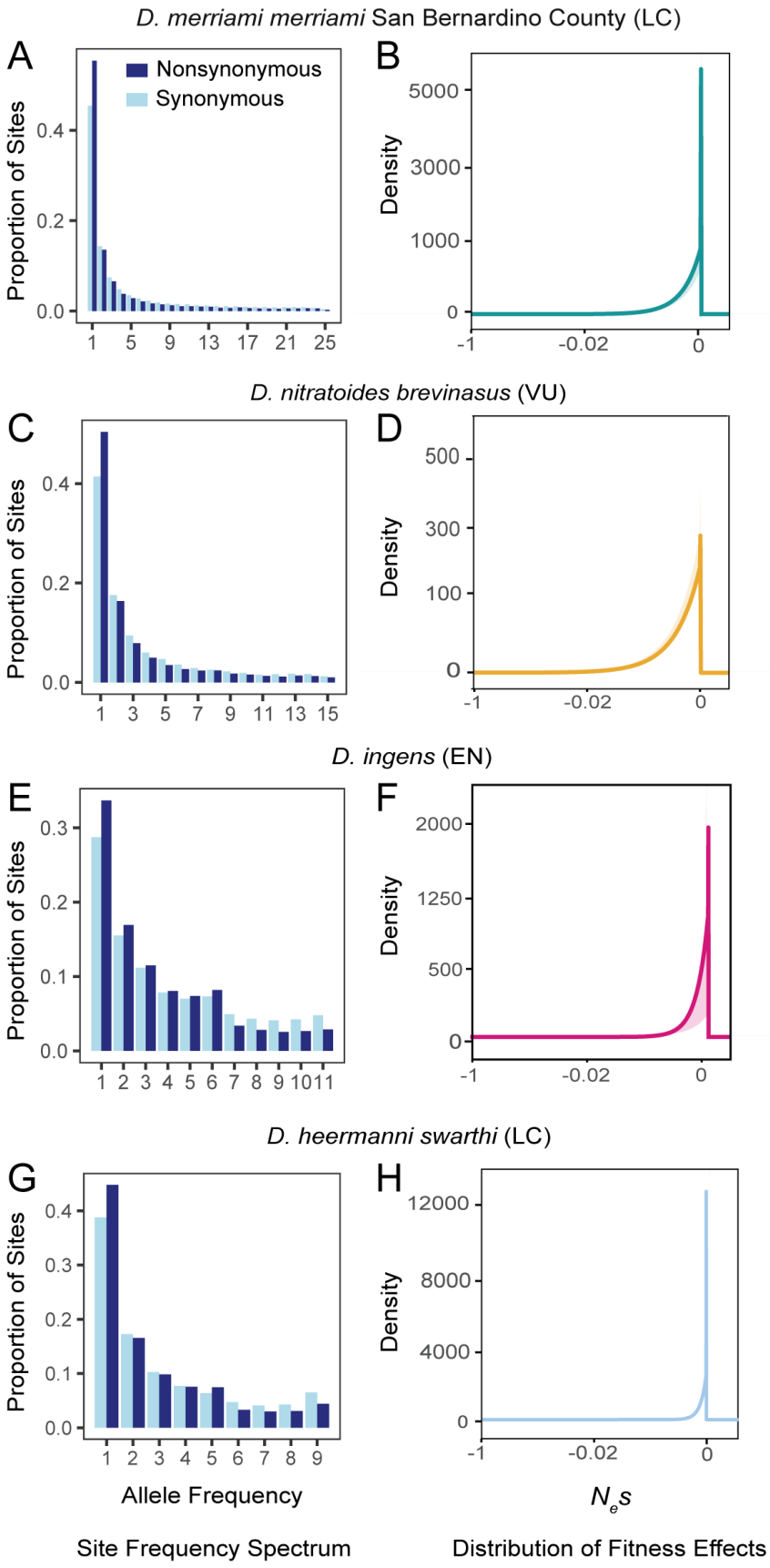

D. merriami merriami San Bernardino County (LC)
A
Nonsynonymous
Synonymous
Proportion of Sites
0.0
0.2
0.4
1 5 9 13 17 21 25
B
Density
0
1000
3000
5000
-1 -0.02 0
D. nitratoides brevinasus (VU)
C
Proportion of Sites
0.0
0.2
0.4
1 3 5 7 9 11 13 15
D
Density
0
100
300
500
-1 -0.02 0
D. ingens (EN)
E
Proportion of Sites
0.0
0.1
0.2
0.3
1 2 3 4 5 6 7 8 9 10 11
F
Density
0
500
1250
2000
-1 -0.02 0
D. heermanni swarthi (LC)
G
Proportion of Sites
0.0
0.1
0.2
0.3
0.4
1 2 3 4 5 6 7 8 9
Allele Frequency
Site Frequency Spectrum
H
Density
0
4000
8000
12000
-1 -0.02 0
Ns
Distribution of Fitness Effects

Figure 5. Site frequency spectrum (SFS) for synonymous and nonsynonymous sites (A, C, E, G) and inferred distribution of fitness effects (B, D, F, H) for four populations of *Dipodomys* kangaroo rats in California. Light blue bars represent the synonymous SFS, and dark blue bars represent the nonsynonymous SFS. A, B: *D. merriami merriami* in San Bernardino County, B, C: *D. nitratoides brevinasus*, D, E: *D. ingens*, and E, F: *D. heermanni swarthi*. Fitness effects are reported as *S*, which is equivalent to $N_e s$. For each population, the distribution of fitness effects was inferred with fastDFE by 1) using synonymous SFS to account for demography and 2) comparing nonsynonymous SFS with synonymous SFS to estimate the distribution of fitness effects from neutral ($N_e s = 0$) to weakly and moderately deleterious ($N_e s < 0$). SFS are shown unfolded, but DFE were estimated from folded site frequency spectra and the fitness effects of beneficial mutations ($N_e s > 0$) were not inferred. Model parameters and likelihood values are reported in the results section, and a list of samples included in each population is available in Supplementary Table 3.

4.6 TABLES

Table 1. Whole genome resequencing sampling for each species of kangaroo rat included in this study. Measures of genetic diversity.

Species	Common name	N	Conservation status	Hetero-zygosity (%)	f_{ROH} (> 100 kb)	Tajima's D[a]
Dipodomys m. merriami	Merriam's kangaroo rat	40	Least Concern	0.086	0.0839	-0.2125
D. m. parvus		1	Endangered	0.062	0.0970	--
Dipodomys panamintinus	Panamint kangaroo rat	47	Least Concern	0.060	0.0761	0.0519
D. p. leucogenys		7		0.0402	0.0984	0.3262
Dipodomys heermanni	Heermann's kangaroo rat	21	Least Concern	0.123	0.0406	0.5360
D. h. berkeleyensis		2	Extinct†	0.094	0.0569	--
Dipodomys ingens	Giant kangaroo rat	8	Endangered	0.064	0.0820	0.5987
Dipodomys stephensi	Stephens' kangaroo rat	6	Threatened	0.049	0.0910	-0.5211
Dipodomys nitratoides	San Joaquin	21	Threatened	0.102	0.0556	1.09637
D. n. brevinasus	Short-nosed	11	Threatened	0.104	0.0550	0.7886
D. n. nitratoides	Tipton	5	Endangered	0.106	0.0473	0.4226
D. n. exilis	Fresno	5	Extinct*	0.094	0.0654	0.5230

†Last observed 1986.

*Last observed 1992.

[a]Calculated in sliding 50-kb windows.

CHAPTER 5

De novo genome assembly of a Geomyid rodent, Botta's pocket gopher (*Thomomys bottae bottae*)

This chapter has been previously published and is reproduced here in accordance with the journal's article sharing policy:

Voss ER, Escalona M, Kozak KM, Seligmann W, Fairbairn CW, Nguyen O, Marimuthu MPA, Conroy CJ, Patton JL, Bowie RCK, Nachman MW. 2024. *De novo* genome assembly of a Geomyid rodent, Botta's pocket gopher (*Thomomys bottae bottae*). *Journal of Heredity* esae045: 1-11.

DOI: 10.1093/jhered/esae045

ABSTRACT

Botta's pocket gopher (*Thomomys bottae*) is a common and widespread subterranean rodent of the North American west. The species has been of long interest to evolutionary biologists due to the phenotypic diversity across its range and unusual levels of variation in chromosome number and composition. Here, we present a high-quality reference genome from a male *T. b. bottae* individual captured in the San Francisco Bay Area. The assembly is comprised of 2,792 scaffolds, with a scaffold N50 value of 23.6 Mb and a BUSCO score of 91.0%. This genome helps fill a significant taxonomic sampling gap in rodent genome resources. With this reference genome, we envision new opportunities to investigate questions regarding the genomics of adaptation to the belowground niche. Further, we can begin to explore the impact of associated life history traits, such as limited dispersal and low population connectivity, on intraspecific genetic and phenotypic variation, genome evolution, speciation, and phylogenetic relationships across the Geomyoidea.

5.1 INTRODUCTION

Pocket gophers (Geomyidae) are among the most recognizable rodents of the American west: they provide critical ecological services such as soil engineering, but they are also regarded as pests for their habit of digging up gardens and agricultural fields. Botta's pocket gopher (*Thomomys bottae*, Eydoux and Gervais 1836; Fig. 1A), has a broad geographic range, extending from the tip of Baja California and the northwestern coast of Mexico north to southern Oregon, east to Utah, Colorado, and Texas in the USA and south again into the Mexican Plateau (Fig. 1C). This range spans a wide array of environments: *T. bottae* occupies coastal forests in

northern California, alpine meadows of western mountain ranges, and shrub habitat in the Chihuahan, Sonoran, Mojave, and Great Basin deserts, spanning an elevational range from below sea level to above timberline (Fig. 1C). Patton and Smith (1990) recognized 15 subspecies in California, and up to 195 subspecies have been described (Jones and Baxter 2004), though the taxonomic status of many lineages remains uncertain (Álvarez-Castañeda 2010).

Many aspects of pocket gopher ecology are shaped by their subterranean life history. Gophers construct and dwell underground in complex burrow systems, and their morphology, physiology and behavior include physiological adaptations to low light and oxygen conditions, as well as skeletal adaptations such as chisel-teeth and forelimb claws modified for digging (Lessa and Thaeler 1989). Botta's pocket gophers inhabit a wide variety of substrates of different colors and hardness (Marcy et al. 2013); while *T. bottae* are primarily scratch-diggers, some populations display remodeled rostra and strongly anchored procumbent incisors that permit excavation of dense clay soils (Lessa and Patton 1989; Marcy et al. 2016). Pelage color also varies widely in association with substrate color (Grinnell 1927; Ingles 1950; Patton and Smith 1990; Wlasiuk and Nachman 2007), suggesting that substrate-matching may be under selection to reduce risk of predation while above ground during dispersal and foraging (Fassler and Leavitt 1975; Janes and Barss 1985).

Pocket gophers defend exclusive use territories. While gopher populations may be locally dense, population sizes tend to be small and geographically fragmented, leading to high population structure and an outsized role for genetic drift in determining the evolutionary trajectories of distinct populations (Patton and Feder 1981; Smith 1998). For example, Patton and Smith (1990) observed high levels of chromosomal variability across *T. bottae* populations, with chromosome number varying from 2n=76 to 2n=88 from west to east across the western USA. Additionally, chromosome size and composition in gophers vary dramatically even when diploid number is constant. Some variation likely stems from peri- and paracentric chromosomal inversions, but most likely primarily results from whole-arm additions or deletions of heterochromatin (Patton and Sherwood 1982). As a result, some populations have entirely bi-armed chromosomes, while others have up to 38 single-armed chromosomes. The origin, function, and mode of evolution of this second type of chromosomal rearrangement remain unknown but it has a large impact on genome size, which varies by as much as 35% among gopher populations (Sherwood and Patton 1982). Similarly, the phylogenetic relationships between species in *Thomomys* and *Geomys*, which together comprise the largest diversity within the seven genera in the family Geomyidae (25 of the approximately 42 species in the family; Mammal Diversity Database 2023), are complex and unresolved: they reflect frequent population isolation events, incomplete lineage sorting, and rapid speciation, and would benefit from further genomic investigation (but see Belfiore et al. 2008 for *Thomomys* phylogeny).

Here, we present a high-quality genome assembly for *T. bottae bottae*, which was generated as a part of the California Conservation Genomics Project (CCGP, Shaffer et al. 2022; Toffelmier et al. 2022). Though Botta's pocket gopher is not of conservation concern, its high levels of intraspecific genetic variation and wide geographic and ecological range provide an opportunity to study adaptation across geographically varied habitats. *Thomomys bottae* is also a useful exemplar for comparison to small mammal species with more limited geographic ranges that are of conservation concern. For example, the CCGP includes genome resequencing for threatened and endangered kangaroo rats *Dipodomys stephensi* and *D. ingens*, and the endangered pocket mouse *Perognathus longimembris pacificus* (Wilder et al. 2022), which are

members of the sister family Heteromyidae. Moreover, this new genomic resource will provide a framework for new studies of the genomic variability, biogeography, and adaptation of this iconic North American rodent to the landscapes of the western United States and Mexico.

5.2 Methods

Biological materials

Liver tissue was obtained from a male *T. b. bottae* collected on 21 November 2020 at the University of California Richmond Field Station, Contra Costa County, California (37.91663, -122.3322) under authorization of the California Department of Fish and Wildlife and euthanized following the Guidelines of the American Society of Mammalogists (Sikes et al. 2016). The voucher, comprised of a museum study skin, skull, and formalin-fixed carcass along with flash-frozen tissues, is deposited at the Museum of Vertebrate Zoology, Berkeley, California (https://arctos.database.museum/guid/MVZ:Mamm:240275).

Nucleic acid library preparation and DNA sequencing

High molecular weight (HMW) genomic DNA (gDNA) was extracted from 78 mg of liver tissue using the Nanobind Tissue Big DNA kit (Pacific BioSciences – PacBio, CA). For long-read sequencing, we constructed a HiFi SMRTbell library using the SMRTbell Express Template Prep Kit v2.0 (PacBio) according to the manufacturer's instructions, resulting in an average fragment size of 15 – 20 kb, which were sequenced at the UC Davis DNA Technologies Core. We additionally prepared an Omni-C library (Dovetail Genomics, Scotts Valley, CA) according to the manufacturer's protocol with slight modifications. The library was sequenced at Vincent J. Coates Genomics Sequencing Lab (Berkeley, CA) on an Illumina NovaSeq platform 6000 to generate 100 million 2x150 bp read pairs per Gb genome size. Specific details of laboratory protocols used to generate PacBio and Omni-C libraries are provided in Supplementary Methods.

Genome assembly

We assembled the genome following the CCGP assembly pipeline version 4.0 (Table 1). We removed remnant adapter sequences from PacBio HiFi data using HiFiAdapterFilt (Sim et al. 2022) and generated an initial phased diploid assembly using HiFiasm (Cheng et al. 2021) in HiC mode using the filtered PacBio HiFi reads and Omni-C short-reads. We aligned the Omni-C data to both assemblies following the Arima Genomics Mapping Pipeline and then scaffolded both assemblies with SALSA (Ghurye et al. 2017; Ghurye et al. 2019).

The assemblies were minimally curated by generating and analyzing their corresponding Omni-C contact maps. To generate contact maps, we aligned the Omni-C data with BWA-MEM (Li 2013), identified ligation junctions, and generated Omni-C pairs using pairtools (Open2C et al. 2024). Then, we generated multi-resolution Omni-C matrices with cooler (Abdennur and Mirny 2020) and balanced them with hicExplorer (Ramírez et al. 2018). We used HiGlass (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 identified and later broke joins where major mis-assemblies and misjoins were found. Some of the remaining gaps (joins generated during scaffolding and/or curation) were closed using the PacBio HiFi reads and YAGCloser

(https://github.com/merlyescalona/yagcloser). Finally, we checked for contamination using the BlobToolKit (Challis et al. 2020).

Genome quality assessment

We generated k-mer counts from the PacBio HiFi reads using meryl (https://github.com/marbl/meryl). K-mer counts were then used in GenomeScope2.0 (Ranallo-Benavidez et al. 2020) to estimate genome features including genome size, heterozygosity, and repeat content. We ran QUAST to obtain general contiguity metrics (Gurevich et al. 2013). To evaluate genome quality and functional completeness we used BUSCO (Manni et al. 2021) with the 13,798 gene Glires ortholog database (glires_odb10). We also assessed base-level accuracy (QV) and k-mer completeness with the previously generated meryl database and merqury (Rhie et al. 2020), and further estimated genome assembly accuracy via BUSCO gene set frameshift analysis (Korlach et al. 2017). Measurements of the size of the phased blocks is based on the size of the contigs generated by HiFiasm on HiC mode.

Mitochondrial genome

We assembled the *T. bottae* mitochondrial genome from the PacBio HiFi reads using the reference-guided pipeline MitoHiFi (Allio et al. 2020; Uliano-Silva et al. 2023) with the *Castor canadensis* mitochondrial sequence (NCBI:NC_033912.1; Lok et al. 2017). After assembling the nuclear genome, we searched for mitochondrial assembly sequence matches using BLAST+ (Camacho et al. 2009) and filtered out contigs from the nuclear genome with greater than 99% mitogenome sequence similarity and size smaller than the mitochondrial assembly sequence.

Repeat annotation

We used a combination of *de novo* and known element annotation to identify repetitive regions of the *T. bottae* genome. First, we masked simple repetitive regions of the genome using RepeatMasker (Tarailo-Graovac and Chen 2009) and then used a Dfam library (v.3.3) to identify known Glires repetitive elements in the genome (Jurka et al. 2005; Hubley et al. 2016). Second, we used RepeatModeler (Flynn et al. 2020) to perform *de novo* identification of transposable element motifs in the genome and DeepTE to classify them (Yan et al. 2020). Using this library, we re-masked the *T. bottae* genome and quantified the proportion of the genome inhabited by each repeat element family using RepeatMasker's ProcessRepeats. Lastly, we used the calcDivfromAlign.pl script included with RepeatMasker to calculate percent Kimura's 2-parameter distance between repetitive elements and consensus element sequences, which provides a relative estimate of element age (Benham et al. 2024).

5.3 RESULTS

Sequencing data

The Omni-C and PacBio HiFi sequencing libraries generated 897.73 million read pairs and 13.4 million reads, respectively. The latter yielded 57-fold coverage (N50 read length 12,072 bp; minimum read length 108 bp; mean read length 11,862 bp; maximum read length of 47,691 bp) based on the Genomescope 2.0 genome size estimation of 2.8 Gb. Initial estimates of genome size based on *T. bottae* C-values were much larger (4-5 Gb; Sherwood and Patton 1982), suggesting that 2.8 Gb may represent the mappable portion of the genome. Based on PacBio HiFi reads, we estimated a sequencing error rate of 0.0973% and 0.56% nucleotide

heterozygosity rate. The k-mer spectrum shows a bimodal distribution with two peaks at 28- and 56-fold coverage (Fig. 2A).

Assembly assessment

The final assembly (mThoBot1) consists of two phased haplotypes that vary in size compared to the estimated value from GenomeScope2.0 (Fig. 2A), as has been observed in other taxa (e.g. Pflug et al. 2020). The assemblies have been tagged as primary and alternate based on our assessment of contiguity. The primary assembly consists of 2,792 scaffolds spanning 2.84 Gb with a contig N50 of 14.54 Mb, a scaffold N50 of 23.6 Mb, a longest contig of 81.46 Mb and a largest scaffold of 81.46 Mb. The alternate assembly consists of 3,340 scaffolds, spanning 3.46 Gb with a contig N50 of 11.28 Mb, a scaffold N50 of 19.4 Mb, a largest contig of 85.75 Mb and a largest scaffold of 109.0 Mb.

During manual curation we generated a total of six breaks, five on the primary assembly and one on the alternate assembly. No further joins were made. In the gap closing step, we were able to close a total of 16 gaps, 10 on the primary assembly and six on the alternate assembly. We further filtered out a total of 66 contigs corresponding to contaminant sequences (see Supp. Methods for details) and a single contig corresponding to the mitochondrial genome, which is described below.

The primary assembly has a BUSCO completeness score of 91.0% using the Glires gene set, a per base quality (QV) of 64.43, a k-mer completeness of 87.70 and a frameshift indel QV of 42.01. The alternate assembly has a BUSCO completeness score of 92.9% using the same gene set, a per base QV of 93.53, a k-mer completeness of 94.96 and a frameshift indel QV of 42.2. Assembly statistics are reported in Table 2, and graphical representation for the primary assembly is presented in Figure 2B (alternate assembly: Supp. Fig. 1). The Omni-C contact maps show that both assemblies are highly contiguous (Figures 2C & 2D). We have deposited scaffolds corresponding to both primary and alternate assemblies with NCBI (Table S1; Data availability).

Initial annotation

We assembled a 16,962 bp mitochondrial genome for *T. bottae* with 21 unique transfer RNAs, 13 protein coding genes and 2 rRNAs.

For the primary nuclear assembly, we masked repetitive elements iteratively using a Glires library with 1,349 elements and a *de novo* repeat library with 1,944 elements. The *de novo*-identified *T. bottae* repeat element library is available on the Dryad data repository (doi:10.5061/dryad.wh70rxwwp).

Repeat masking identified 7,426,339 individual repetitive elements in the *T. bottae* genome, which together comprise 49.74% of the 2.8 Gb primary assembly (Table 3). Retrotransposons accounted for 36.36% of the genome: 18.08% is composed of long interspersed nuclear elements (LINEs), 12.5% is composed of short interspersed nuclear elements (SINEs) and 5.78% is composed of long terminal repeats (LTRs). L1/CIN4 retrotransposons were the single most abundant repeat type (17.6% of genome), followed by Alu SINEs at 2.09% and ERV-classII LTRs at 1.15%. In contrast, DNA transposons were rare at 1.12%. Satellites and simple repeats comprised 9.17% of the genome. Estimates of Kimura's 2-parameter (K2P) distances suggest that LINEs appear to have had an ancient pulse of activity (K2P 30-40%), but also have driven more recent repetitive element expansion (K2P 0-15%) in the genome

compared with SINEs (K2P peaks at 15-25%; Fig. 2E). LTRs also show several K2P peaks that suggest intermittent activity throughout the evolutionary history of the *T. bottae* genome.

5.4 Discussion

We present a high-quality *Thomomys* genome assembly with 2,792 scaffolds and an N50 of 23 Mb (Table 2), filling a notable taxonomic gap in genomic resources for rodents. The family Geomyidae belongs to the rodent suborder Castorimorpha, which includes beavers (Castoridae), kangaroo rats and allies (Heteromyidae), and gophers. Of approximately 100 species in this suborder, only six have genome assemblies available on NCBI: the American beaver *Castor canadensis* (Lok et al. 2017: GCF_001984765.1), three kangaroo rats, *Dipodomys ordii* (Liu et al. 2014: GCF_000151885.1), *D. stephensi* (Johnson et al. 2019: GCA_004024685.1), and *D. spectabilis* (Harder et al. 2022: GCF_019054845.1), and one pocket mouse, *Perognathus longimembris* (Wilder et al. 2022: GCF_023159225.1; Kozak et al. 2024: GCA_024363575.2) in addition to *T. bottae*.

Concurrently with this genome's initial release on NCBI, the Vertebrate Genomes Project (Genome 10K consortium; Rhie et al. 2021) published a second *T. bottae* genome (*T. b. perpallidus*, NCBI: mThoBot2hap1, GCA_031878675.1; Conroy et al. 2023). Though their primary haplotype assembly is slightly smaller at 1.9 Gb, the genome is far more contiguous, with 340 scaffolds and a scaffold N50 of 61.3 Mb. Notably, both genomes are smaller than karyotype and C-value observations would suggest (Patton 1972; Patton and Sherwood 1982; Patton and Smith 1990). We speculate that the relatively high proportion of repetitive DNA in the *T. bottae* genome (Table 3) may be under-represented in these genome assemblies. Repetitive elements are challenging to assemble due to the high degree of similarity between different regions of the genome (Tørresen et al. 2019).

These new genome assemblies represent the first step toward understanding *T. bottae* genomic evolution (Patton and Smith 1990). Most *T. bottae* individuals have the same complement of chromosomes, with a typical diploid chromosome number of 2n=76 (Patton 1972), but populations in Mexico and Texas have 2n=74-78 (Patton and Dingman 1970; Berry and Baker 1971) and some populations in New Mexico and Colorado have up to 2n=88 due to the presence of heterochromatic micro-chromosomes (Hafner et al. 1983). Many populations display multiple whole-arm chromosomal deletions, ranging from zero deletions along the coast of California up to 38 deletions moving from west to east across the western USA (Patton and Smith 1990). However, chromosomal arm variation does not seem to prevent hybridization between individuals with different numbers of deletions (Patton and Sherwood 1982). Though the karyotype of the specific individual sequenced is not known, gophers karyotyped in the San Francisco Bay Area, including from the same locality from which the animal whose genome we sequenced came, have 2n=76 chromosomes (Sherwood and Patton 1982, Patton and Smith 1990) with no whole-arm chromosomal deletions.

To better understand repeat content in *T. bottae*, we conducted known repetitive element annotation and *de novo* TE prediction on the reference genome and identified 48.61% of the genome as repetitive. As has been observed in other rodents (e.g. Platt et al. 2018; Osmanski et al. 2023), LINE and SINE elements were the most common TEs, followed by LTRs, while DNA transposons were rare. Repetitive element activity varies through time and estimates of the genetic distance between consensus sequences and specific elements in the genome can provide

a sense of the relative age of specific repetitive elements (e.g. Benham et al. 2024). Without a reasonable mutation rate estimate for Geomyidae, we cannot estimate the absolute time since element insertion, but LINEs display evidence of current and recent activity in the genome, while SINEs appear to be mostly intermediate in age and DNA transposons appear to be ancient.

Annotation of repetitive elements in the reference genome represents a first pass towards understanding genome size evolution in *T. bottae*. However, additional curation and long-read sequencing from multiple individuals would be useful for understanding the distribution of repetitive elements across the genome and the nature of chromosomal variation across *T. bottae* and the Geomyidae more generally. With such data, it would be feasible to identify which specific elements have driven recent gopher genome evolution and begin to explain large-scale patterns of chromosomal variation (e.g. Gozashti et al. 2023). Chromosome staining studies indicate that the chromosomal arms absent in some populations of pocket gophers are heterochromatic (Barros and Patton 1985), suggesting that repetitive DNA evolution drives variation in gopher genome size and karyotype. Further, even when two populations have the same number of chromosomes and whole-arm deletions, C-values suggest additional variation in genome size (Sherwood and Patton 1982). We hypothesize that repetitive element proliferation and excision, in combination with unique aspects of gopher biology, are responsible for variation in gopher genome size. Due to small local population sizes, limited dispersal, and resulting population fragmentation, genetic drift may drive rapid genetic divergence between populations and may allow repetitive elements to become locally fixed or lost.

In combination with whole-genome resequencing, the *T. bottae* genome will broaden our capacity to explore numerous questions related to pocket gopher biology, systematics, and population genomics. For example, how is genetic variation partitioned among populations? How many named taxa represent monophyletic lineages? Which loci are associated with phenotypic variation among populations? Due to their complex biogeographic history, it may be difficult to resolve broader relationships among *T. bottae*, *T. townsendii* and *T. umbrinus*, and the wider Geomyid family. However, additional genomes could begin to address questions regarding the extent of incomplete lineage sorting, ancient admixture, and recent introgression between these species. For example, the extent of gene movement across species boundaries may differ between hybrid zones: *T. bottae* and *T. umbrinus* hybridize but show reduced female fertility and male sterility in F1 hybrids, while *T. bottae* and *T. townsendii* are completely inter-fertile (Patton 1973; Patton et al. 1984).

This genome may also provide insight into the genomics of pocket gopher adaptation to a subterranean life history. For example, other independent lineages of subterranean mammals, such as the naked mole rat *Heterocephalus glaber* (Rodentia: Heterocephalidae), Cape golden mole *Chrysochloris asiatica* (Afrotheria: Chrysochloridae), and star-nosed mole *Condylura cristata* (Eulipotyphla: Talpidae) display genomic signatures of adaptation to the stressors associated with living underground, such as resistance to hypoxia, hypercapnia (high CO_2 levels), pathogen load, and morphological requirements for digging (Fang et al. 2014; Li et al. 2020). Further, genomic changes such as olfactory receptor gene family expansion (Courcelle et al. 2023) and relaxed constraint in vision genes (Partha et al. 2017) may underlie sensory adaptations such as enhanced smell and reduced eyesight. Explorations of gene family expansion, gene loss, and rates of substitution in *T. bottae* compared with independently evolved subterranean rodents will advance our knowledge of gopher adaptations to a subterranean lifestyle and will help identify genetic changes that are convergent versus those that are unique to

pocket gophers (e.g. Jiang et al. 2020). The Botta's pocket gopher genome assembly provides new opportunities to deepen our understanding of adaptation to the underground niche and the impact of a subterranean life history on speciation and genome evolution in vertebrates.

5.5 FIGURES

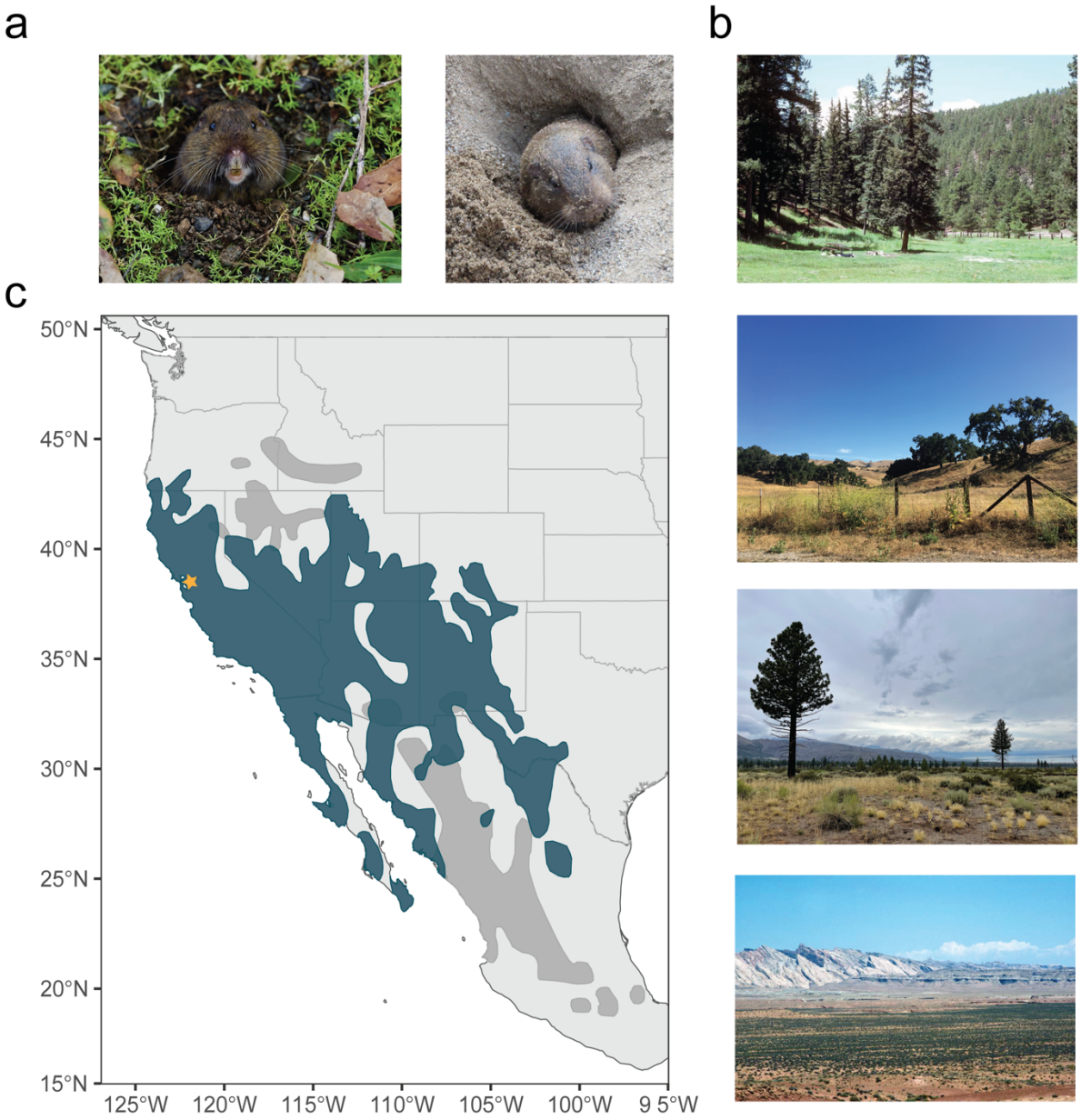


Figure 1. A: Botta's pocket gophers (*Thomomys bottae*) construct complex burrow systems and B: inhabit an extremely broad range of environments, including coniferous forests, alpine meadows, oak savannah, cultivated fields, chaparral, and desert. C: *T. bottae*'s geographic range stretches across southwestern North America. *Thomomys umbrinus* (southern gray range) and *Thomomys townsendii* (northern gray range) belong to the same species complex and hybridizes with *T. bottae* at contact zones. The star indicates the capture location of the *T. bottae* individual from which the genome in this study was constructed (Richmond Field Station, Contra Costa County, California, USA). Photo credits: Gopher image credits: left M.G. Pagano, right via iNaturalist; landscape images ERV, JLP; range data retrieved from IUCN.

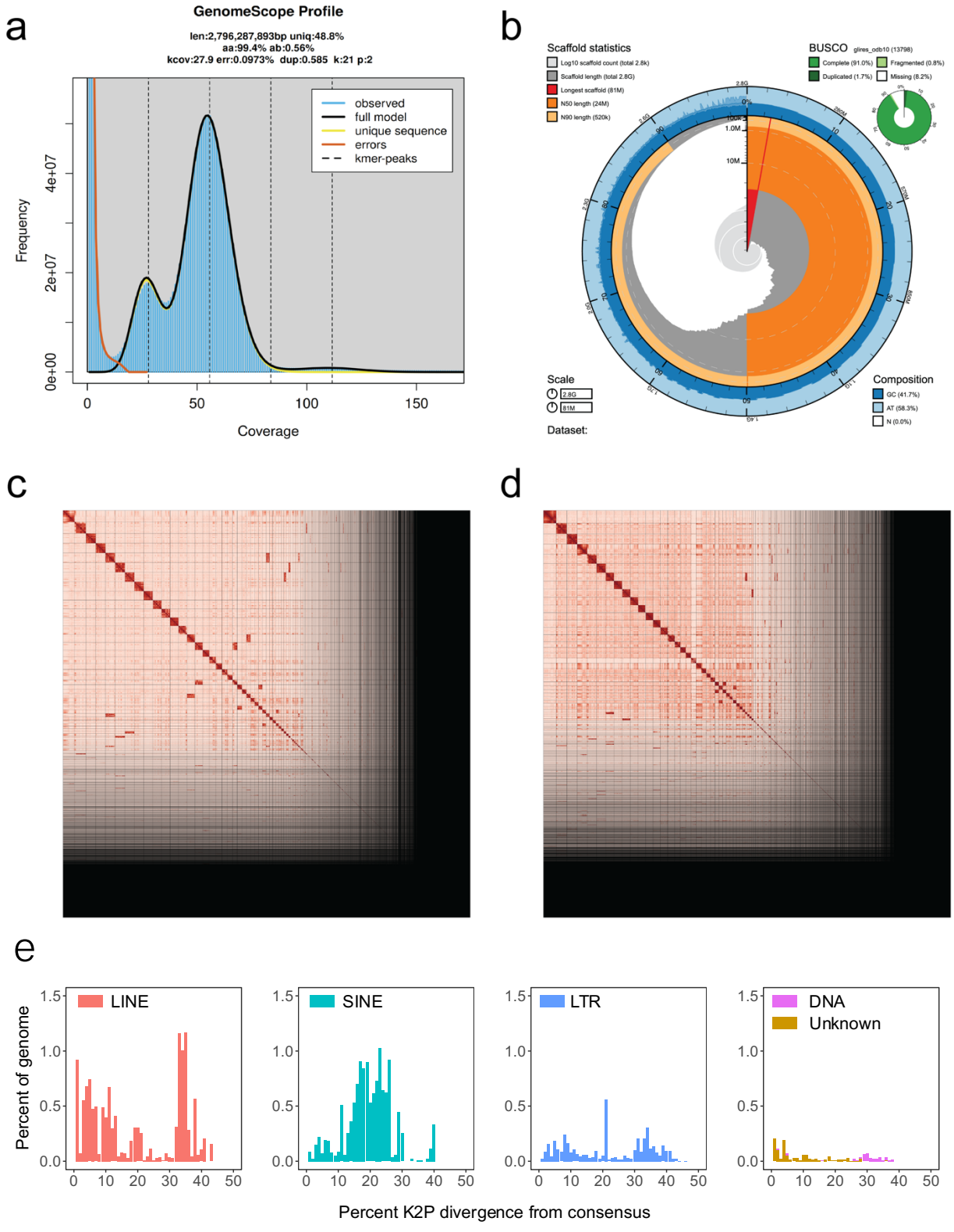

a
GenomeScope Profile
len:2,796,287,893bp uniq:48.8%
aa:99.4% ab:0.56%
kcov:27.9 err:0.0973% dup:0.585 k:21 p:2
observed
full model
unique sequence
errors
kmer-peaks
Frequency
0e+00
2e+07
4e+07
0
50
100
150
Coverage
b
Scaffold statistics
Log10 scaffold count (total 2.8k)
Scaffold length (total 2.8G)
Longest scaffold (81M)
N50 length (24M)
N90 length (520k)
BUSCO glires_odb10 (13798)
Complete (91.0%)
Fragmented (0.8%)
Duplicated (1.7%)
Missing (8.2%)
Scale
2.8G
81M
Dataset:
Composition
GC (41.7%)
AT (58.3%)
N (0.0%)
c
d
e
Percent of genome
LINE
SINE
LTR
DNA
Unknown
0.0
0.5
1.0
1.5
0
10
20
30
40
50
Percent K2P divergence from consensus

Figure 2. Visual overview of *Thomomys bottae* genome assembly composition and quality. A: GenomeScope2.0 k-mer spectrum from adapter-trimmed HiFi sequence data. The bimodal pattern is expected for a diploid genome, with the left-hand peak representing k-mers in heterozygous regions and the right-hand peak representing k-mers in homozygous regions. B: BlobToolKit Snail plot genome assembly summary. The central circle represents the length and number of scaffolds: the circumference of the circle represents the entire length of the genome and scaffolds are added in order of length starting from longest (red) scaffold to shortest (all scaffolds in gray). Orange bars represent scaffold N50 and N90, and surrounding blue bars represent GC and AT content. BUSCO score is given in the upper right corner and length-related statistics are at top left. C,D: Contiguous linear organization of the primary (C) and alternate (D) genome assemblies visualized as an Omni-C contact map (PreTextSnapshot). Red represents regions of the genome that are close to each other in 3D space as identified by chromatin-proximity sequencing. Dark regions represent short fragments of the genome that could not be assembled into larger scaffolds. The large proportion of unassembled regions may reflect high repeat content. E: Transposable element landscapes for LINE, SINE, LTR, DNA and unknown element types. Divergence from the consensus sequence is measured as percent Kimura's 2-parameter distance on the x-axis; more divergent TEs are likely ancient and have accumulated mutations over time. TE abundance is normalized as a percentage of the genome on the y-axis.

5.6 TABLES

Table 1. Assembly pipeline and software.

Assembly	Software and options*	Version
Filtering PacBio HiFi adapters	HiFiAdapterFilt	Commit 64d1c7b
K-mer counting	Meryl (k=21)	1
Estimation of genome size and heterozygosity	GenomeScope	2
De novo assembly (contiging)	HiFiasm (Hi-C Mode, –primary, output p_ctg.hap1, p_ctg.hap2)	0.16.1-r375
Scaffolding		
Omni-C data alignment	Arima Genomics Mapping Pipeline	Commit 2e74ea4
Omni-C Scaffolding	SALSA (-DNASE, -i 20, -p yes)	2
Gap closing	YAGCloser (-mins 2 -f 20 -mcc 2 -prt 0.25 -eft 0.2 -pld 0.2)	Commit 0e34c3b
Omni-C contact map generation		
Short-read alignment	BWA-MEM (-5SP)	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 (hicCorrectmatrix correct --filterThreshold -2 4)	3.6
Contact map visualization	HiGlass	2.1.11
	PretextMap	0.1.4
	PretextView	0.1.5
	PretextSnapshot	0.0.3
Genome quality assessment		
Basic assembly metrics	QUAST (--est-ref-size)	5.0.2
	BUSCO (-m geno, -l glires)	5.0.0
Assembly completeness	Merqury	2020-01-29
Contamination screening		
Local alignment tool	BLAST+ (-db nt, -outfmt ‘6 qseqid staxids bitscore std’ , -max_target_seqs 1, -max_hsps 1, -evalue 1e-25)	2.1
General contamination	BlobToolKit (HiFi coverage, BUSCO=glires, NCBI Taxa ID=10013)	2.3.3
Mitochondrial genome assembly	MitoHiFi (-r, -p 50, -o l, -a) Reference: *Castor canadensis* (NCBI:NC_033912.1)	2.2
Genome annotation		
Repetitive element annotation	DeepTE (-m M -prop_thr 0.8)	Commit babd65e950
	RepeatModeler (-LTRstruct)	2.0.3
	RepeatMasker2 (Dfam database)	4.1.2-p1 (Dfam v.3.3)

*Options detailed for non-default parameters. Software citations are listed in the text.

Table 2. Quality metrics for the primary and alternate assemblies of Botta's pocket gopher (*Thomomys bottae*) genome mThoBot1.0.

Metric	Primary	Alternate
Number of contigs	2,898	3,464
Contig N50 (bp)	14,543,669	11,283,157
Contig NG50§	14,576,268	19,020,183
Longest contig	81,460,496	85,754,623
Number of Scaffolds	2,792	3,340
Scaffold N50	23,603,529	19,408,772
Scaffold NG50§	26,618,111	37,562,823
Largest scaffold	81,460,496	109,030,123
Size of final assembly	2,845,630,888	3,469,049,410
Phased block NG50§	20,902,488	16,209,434
Gaps per Gbp (#Gaps)	37	36
BUSCO‡	91.0%	92.9%
Complete genes	89.3%	90.8%
Complete: single copy	1.7%	2.1%
Complete: duplicated	0.8%	0.8%
Fragmented	8.2%	6.3%
Missing		
Indel QV (Frameshift):	42.0132814	42.2074167
Base pair QV	64.4338	64.7871
	Full assembly = 64.6243	
k-mer completeness	87.7051	93.5386
	Full assembly = 99.1352	

§ NGx statistics have been calculated based on an estimated genome size of 2.8 Gb.
‡ BUSCO completeness assessed with glires_odb10 database (n=13,798 genes). NCBI BioProject: PRJNA777226; whole genome sequence accession numbers JANJXV000000000 (primary) and JANJXW000000000 (alternate). Additional information listed in supplementary table 1 and data availability statement.

Table 3. Repetitive element annotation of the number, length, and percent of the genome of Botta's pocket gopher occupied by repetitive elements*, which together comprise 49.74% of the genome.

Repetitive element class	Total Elements	Length (bp)	Percent of genome (%)
Retroelements	5,269,565	1,034,633,804	36.36
SINEs	3,367,810	355,807,977	12.50
Alu/B1	654,626	59,362,288	2.09
MIRs	118,403	14,430,923	0.51
Unknown	2,594,781	282,014,766	9.90
LINEs	1,422,224	514,415,651	18.08
L1/CIN4	1,339,511	500,904,020	17.60
L2/CR1/Rex	78,004	12,862,058	0.45
RTE/Bov-B	3,590	542,131	0.02
LTR elements	479,531	164,410,176	5.78
Ty1/Copia	1,581	254,739	0.01
Gypsy/DIRS1	14,460	4,489,501	0.16
ERVs	421,543	134,793,241	4.74
ERV-classI	12,284	2,998,877	0.11
ERV-classII	247,664	32,698,503	1.15
ERVL-MaLRs	113961	19,933,989	0.70
DNA transposons	177,474	31,979,635	1.12
hobo-Activator	72,022	13,252,042	0.47
Tc1-IS630-Pogo	38,830	7,400,619	0.26
hAT-Charlie	45,835	7,256,208	0.25
TcMar-Tigger	20,787	4,070,766	0.14
Other repetitive elements			
Rolling-circles	1,525	159,084	0.01
Small RNA	202,146	22,644,712	0.80
Satellites	19,080	1,655,288	0.06
Simple repeats	1,312,574	242,158,596	8.51
Low complexity	226,865	16,940,668	0.60
Unclassified interspersed	217,110	64,973,792	2.28
Total masked	7,426,339	1,415,145,579	49.74

Repetitive elements iteratively identified and masked with 1) Repbase simple repetitive elements, 2) Repbase and Dfam libraries of Glires and ancestral repetitive elements, 3) *de novo* identified transposable elements using RepeatModeler, and 4) *de novo* identified transposable elements using DeepTE. Software versions listed in Table 1.

Supplementary figures and tables are available online at DOI: 10.1093/jhered/esae045

CHAPTER 6

Genetic variation, phylogeography and population structure in pocket gophers (*Thomomys bottae*) in California

ABSTRACT

Botta's pocket gopher (*Thomomys bottae*) is one of the most widespread rodents in the North American southwest. This species has a complex biogeographic history characterized by rapid diversification combined with incomplete lineage sorting, followed by local isolation and drift. As part of the California Conservation Genomics Project, we sequenced the genomes of 95 individuals to study levels and patterns of genetic variation, population structure, and phylogeography of gophers from a diverse range of habitats across the state of California. While pocket gophers are not a species of conservation concern, they can provide a frame of reference for comparison with other species. We found that average levels of nucleotide diversity varied among populations from a low of $\pi = 0.00239$ for northern *T. b. perpallidus* in the Great Basin Desert to a high of $\pi = 0.00495$ for *T. b. mewa* in the Sierra Nevada foothills. Admixture analyses revealed nine distinct populations, in contrast to the 14 named subspecies. Thus, subspecies designations did not map directly to populations, mainly due to the absence of genetic differentiation between some subspecies. Admixture analysis revealed some individuals of mixed ancestry, providing evidence of gene flow between populations. Nonetheless, F_{ST} was extremely high among admixture-identified populations ranging from 0.244 between two populations in the Great Basin Desert to 0.746 between the Sierra-awahnee population and the northern of the two populations in the Great Basin. Inbreeding coefficients (F_{IS}) and the fraction of the genome with runs of homozygosity (F_{ROH}) suggest that most populations are moderately inbred, with most F_{IS} between 0.15 and 0.2. Principal component analysis and phylogenetic trees identified three major groups of California pocket gophers, corresponding to those mainly from Northern California, Central California, and the Great Basin and broadly consistent with groupings previously identified based on allozyme and mitochondrial data. These results highlight the biogeographic history and complex population structure in gophers. The discovery of moderate to high levels of inbreeding in a common species also underscores the challenges of drawing conclusions in conservation practices from simple summary statistics of genetic data.

6.1 INTRODUCTION

Botta's pocket gopher (*Thomomys bottae*; Eydoux and Gervais 1836), a member of the subterranean rodent family Geomyidae, is one of the most widespread mammals in North America. Early researchers noted that pocket gophers are both highly specialized and abundant; they have evolved a suite of adaptations to living underground and rarely leave their burrows, yet have a large geographic range and can live in many different habitats (Merriam 1895; Grinnell 1927). Botta's pocket gophers can be found from the Chihuahuan desert to the redwood forests

of Northern California, the Pacific coast to the Rocky Mountains, and below sea level in Death Valley to above the timberline in the Sierra Nevada (Jones and Baxter 2004). Despite this expansive range, Grinnell (1927) noted that populations are patchily distributed; pocket gophers cannot inhabit soils that are too hard for digging (Marcy et al. 2013) or too dry or alkaline to support plant life for food, nor do they often cross rivers or bodies of water. Additionally, early naturalists including Grinnell and Merriam (Merriam 1895; Grinnell 1927; Grinnell and Hill 1936; Hall 1981) observed extensive variation in body size, coat color, and skeletal characteristics, particularly in the incisors and skull, leading them to describe a combined 200 subspecies (Jones and Baxter 2004).

To infer evolutionary relationships among these numerous *T. bottae* forms, Patton and collaborators (Patton 1972; Patton and Yang 1977; Patton and Sherwood 1982; Patton and Smith 1990) sampled and analyzed allozymes from more than 170 populations across the range of *T. bottae* and karyotyped more than 1,500 individuals. Allozyme data indicate that California populations form four geographic units that likely diverged early in the history of the species, and individual populations within these units display extremely high levels of structure (Patton and Yang 1977). Further, pocket gophers from California to Texas display remarkable variation in karyotype, though individuals with different chromosomal arrangements are not reproductively isolated (Patton 1972; Thaeler 1980; Patton and Sherwood 1982). This combination of population structure and local fixation of chromosomal variation suggests that population biology characteristics associated with fossoriality facilitate population isolation within *T. bottae*, resulting in genetic drift and high levels of differentiation in short periods of time (Patton and Smith 1990).

Despite a century of study on the population biology, ecology, evolution, and systematics of *Thomomys bottae*, some questions remain. Patton and coauthors established that population structure and differentiation is high within *T. bottae*, that several divergent geographic units likely diversified early in the history of the species, and that some morphological variation is driven by local adaptation (coat color) while other variation (body size) is highly plastic (Patton and Brylski 1987; Patton and Smith 1990). However, a handful of studies conducted over the past twenty-five years using mitochondrial and several nuclear loci have yielded results inconsistent with prior allozyme studies (Smith 1998; Belfiore et al. 2008; Álvarez-Castañeda 2010). For example, mitochondrial studies of *T. bottae* populations collapse two major geographic units into one and identify at least one additional lineage extending north from Baja California (Smith 1998; Álvarez-Castañeda and Patton 2004; Álvarez-Castañeda 2010). An additional multi-locus nuclear study of *Thomomys* found low support for any node within the *T. bottae-T. townsendii* clade (Belfiore et al. 2008). Though the species' history of incomplete lineage sorting and genetic drift make it difficult to resolve the timing and order of divergence between lineages, more comprehensive genomic data could provide a measure of clarity regarding the phylogeographic history of Botta's pocket gopher.

Although pocket gophers are not of conservation concern, *Thomomys bottae* was included in the California Conservation Genomics Project, which seeks to survey statewide genetic diversity in over 200 species mainly focusing on taxa that are threatened or endangered (Shaffer et al. 2022). Because pocket gophers are widespread, but populations are isolated and structured, *T. bottae* provides a useful baseline for comparison with other California mammals that are endangered. In addition to exploring the evolutionary history of *T. bottae*, we can use genomic data generated by the CCGP to ask how *T. bottae* compares with species of

conservation concern with respect to genetic diversity, population structure, and inbreeding, with the goal of better understanding the conservation genomic status of endangered small mammals.

Here, we sequenced whole genomes of 95 pocket gophers across the state of California to 1) ask how *T. bottae* compares with other California rodents and small mammals more broadly in terms of genome-wide heterozygosity, genetic diversity and inbreeding, and 2) revisit phylogenetic relationships within *Thomomys bottae*. Do whole-genome sequences support relationships defined with allozyme data? How many lineages of pocket gophers exist within California, and where are the geographic boundaries between them? Is there gene flow among populations or units? To answer these questions, we measured genetic diversity (nucleotide diversity and individual heterozygosity), summarized the distribution of allele frequencies using Tajima's D, estimated the level of inbreeding (F_{IS} and the frequency of runs of homozygosity), summarized genetic distances using F_{ST}, conducted admixture and principal components analysis, and created a phylogenetic tree using maximum likelihood methods with genomic data for these 95 *T. bottae* individuals. We also took advantage of existing museum specimens associated with whole genome sequences to explore the genetic basis of variation in coat color.

We found that *T. bottae* displays moderate levels of genetic diversity. Populations in central California and the Sierra Nevada foothills show the highest genetic diversity (π = 0.00494), and northern California and Great Basin populations show the lowest genetic diversity (π = 0.00239). Levels of inbreeding were moderate in many populations. Phylogenetic analyses based on whole-genome sequences identified four clades that largely, but not completely, coincide with groupings identified in classic allozyme studies of *T. bottae* diversity.

6.2 METHODS

Sample processing and library preparation

All *T. bottae* individuals (n = 96) in this study were collected between 1974 and 2017, prepared as museum specimens, and deposited in the mammal collection of the Museum of Vertebrate Zoology (MVZ) at the University of California, Berkeley. Samples were drawn from across the range of *T. bottae* in California and include specimens from 14 named subspecies. Supplementary Table 1 contains specimen metadata for each sample, including catalog number, collecting locality, date, sex, and subspecies assignment. All laboratory work, including DNA extraction and library preparation, was conducted in the Evolutionary Genetics Laboratory at the MVZ.

DNA was first extracted from flash frozen tissue stored in liquid nitrogen with the Mag-Bind® Blood & Tissue DNA Kit (Omega Bio-Tek). We used a Nanodrop spectrophotometer to assess whether DNA extraction was successful and measured DNA concentration using the Biotium PicoGreen AccuClear® Ultra High Sensitivity dsDNA Quantitation Kit on a SpectraMax M2 fluorescence plate reader. Multiple extractions were performed for some low concentration samples, which were pooled and purified via a SPRI bead clean-up. We then used a qSonica instrument to sonicate extracted DNA to an appropriate length for short-read library sequencing (300-500 bp). We assessed fragment size with a 1.5% agarose gel and filtered out long and short DNA fragments using a low ratio SeraMag SPRI bead size selection step. At this point, we used a Kapa HyperPrep Kit according to manufacturer's instructions to prepare 150-bp paired end Illumina whole-genome libraries for sequencing. We verified whether library

preparation was successful with a Nanodrop spectrophotometer or Invitrogen Qubit fluorometer and assessed library fragment length with the Agilent DNA1000 Assay on a BioAnalyzer 2100. Libraries were sequenced to an average of 10x coverage across six lanes of an Illumina NovaSeq 6000 S4 at the Vincent J. Coates Genome Sequencing Laboratory at UC Berkeley and returned as fastq files.

Variant Calling

We mapped raw reads to the *T. bottae bottae* reference genome (Voss et al. 2024; NCBI: GCA_024803745.1) and called variants with the snpArcher pipeline (Mirchandani et al. 2024; https://github.com/ccgproject/ccgpWorkflow). Briefly, we provided raw fastq files for each sample to the snpArcher Snakemake environment along with the reference genome fasta file. The snpArcher pipeline first trims and filters raw reads with fastp, maps reads to the reference genome and removes PCR duplicates with Sentieon bwa mem, and calls variants with Sentieon haplotyper before filtering out indels, low quality reads, and sites with low or high coverage with GATK VariantFiltration.

After variant calling and initial filtering was completed, we used vcftools v0.1.16 and bcftools v.1.21 (Danecek et al. 2011; 2021) to filter out polymorphic sites with more than two alleles, average read depth of less than 5 or greater than 30, greater than 20% missing data, or a quality score of less than 40. After performing quality control checks, we removed one individual that appeared to belong to a different species in the snpArcher report (see Results). Quality filtering resulted in a set of 282,721,405 variant sites; we also created a second file with a minor allele frequency filter of 0.02 for estimating population structure with ADMIXTURE and principal components analysis (PCA) and for phylogenetic analyses.

Population Genetics Analyses

To assess initial population structure within California populations of *T. bottae*, we performed a principal components analysis (PCA) on 5,766,144 linkage-pruned single nucleotide polymorphisms (SNPs) with a minor allele frequency of 0.02. Linkage pruning was performed with plink v1.90b6.21 (Chang et al. 2015) in 50 kb sliding windows with a 10 kb step size (SNPs with $r^2 > 0.1$ were pruned). We then performed a population structure analysis with ADMIXTURE v1.3.0 (Alexander and Lange 2011) using the same set of LD-pruned sites and tested which number of populations had the lowest cross-validation error from k = 3 to k = 12.

To quantify genetic diversity and inbreeding within populations and differentiation between populations, we calculated a suite of population genetic statistics on the 9 populations identified by admixture (k = 9 had the lowest cross-validation error). We used SNPs with no minor allele frequency filter to avoid biasing results by removing rare variants. To reduce computing time, we subsampled 20 million sites from the full set of 282,721,405 and applied a second round of quality filtering based on per-site read depth and missingness, resulting in 15,621,715 SNPs. Using a combination of vcftools and bcftools (Danecek et al. 2011; 2021), we calculated nucleotide diversity (π) per site for each population in 100-kb non-overlapping windows. We also calculated the average observed heterozygosity per site for each individual across all individuals in each population. We estimated the distribution of allele frequencies using Tajima's D (Tajima 1989) to make inferences about population history. Negative values of Tajima's D are expected in recently expanded populations, while positive values are expected following population contractions. We measured the degree of inbreeding using two different estimators. We calculated F_{IS} from comparisons of observed heterozygosity with expected

heterozygosity based on SNP frequencies in individual populations. We also calculated f_{ROH}, the proportion of the genome that is contained in runs of homozygosity (McQuillan et al. 2008). Finally, we calculated pairwise F_{ST} between each of the nine populations in 100-kb non-overlapping windows.

Phylogeographic Analyses

To assess the phylogenetic relatedness between different *T. bottae* populations and individuals, we constructed an alignment using LD-pruned SNPs, which we further reduced from 5,766,144 to 1,228,667 sites by applying a 10% missing data cutoff. Since there are no whole-genome sequences from a closely related outgroup, we constructed an unrooted phylogenetic tree using RAxML-NG v. 1.2.2 (Kozlov et al. 2019) with model GTR+G. We performed 100 bootstraps to evaluate support and visualized the bootstrapped tree with R packages ape v5.8 (Paradis and Schliep 2018) and ggtree v 3.12.0 (Yu et al. 2017). We used R v4.4.0 (R Core Team 2024), ggplot2 v3.5.1 (Wickham 2016), and Adobe Illustrator to create all figures and visualizations included in this manuscript.

Color Analyses

We took advantage of the whole-genome sequences from population samples of gophers which differ markedly in color to conduct a genome wide-association study (GWAS) with the aim of discovering genes underlying color variation. To survey pocket gopher coat color, we photographed each specimen that had been prepared as a study skin and was available in the MVZ mammal collection (n = 61). To standardize lighting conditions across all images, we photographed specimens in a 16x16 inch light box with a white base and walls and included a Color Checker Classic Mini (Calibrite) for downstream white balancing. Each specimen was photographed three times and specimen position in the light box was varied to control for differences in light angle and reflection. Raw images were imported into Adobe Photoshop to standardize image exposure and adjust for white balance against the color checker. To summarize variation in coat color, a rectangle was drawn across the specimen's dorsal surface and color was averaged across all pixels within the rectangle. If any portions of the specimen were degraded or molt lines were visible, separate rectangles were drawn above and below the molt line to avoid biasing color, as the molt line is darker than mature hairs in the dorsal pelage. Color was quantified and recorded using the L*A*B color system. Each value was averaged across all three photographs of the specimen. If multiple rectangles were drawn per specimen, color was first averaged across all rectangles for a single photograph and then across all three photographs of the specimen.

Color measurements were imported into R for statistical analysis. We first used an ANOVA and Tukey post-hoc test to separately compare L*, A*, and B* and determine whether there was significant color variation across *T. bottae* subspecies. Second, we performed a principal components analysis to summarize overall color variation. We then combined phenotypic measures, including L*, A*, B*, and PC1 with SNP data using plink to test for associations between genetic loci and color variation. We performed genome-wide association (GWA) tests with GEMMA (https://github.com/genetics-statistics/GEMMA) for each color measurement of interest with 5,121,497 SNPs (linkage pruned as described above) and controlled for relatedness with a kinship matrix due to the high level of population structure within our dataset. GWA results were visualized in R with qqman v.0.1.9 (Turner 2014).

6.3 RESULTS

DNA sequence data

We sequenced 96 *T. bottae* individuals (Fig. 1) to an average depth of 10x coverage. By mapping the reads to the reference genome (Voss et al. 2024) and calling variants with stringent quality filters, we produced a set of 282,721,405 SNPs after removing one individual that likely belongs to the sister species, *T. townsendii* (Supp. Fig. 1).

Principal components analysis

For principal components analysis, we used 5,766,144 SNPs that had been filtered for a minor allele frequency greater than 0.02 and pruned for linkage disequilibrium (50-kb window, 10-kb step, $r^2 > 0.1$). These data yielded three primary sample clusters when PC1 (12.6% percent of variance explained) and PC2 (11.7% of variance explained) were plotted on the x- and y-axis (Fig. 2). Sample clusters represent geographic units as follows: one cluster represents samples from the Great Basin Desert, Death Valley, and southeastern California, one cluster contains all individuals captured from the San Francisco Bay northward, and the largest cluster contains all samples from the central and southern coasts of California inland to the Central Valley and into the Sierra Nevada. Individuals from different clusters (circled in purple and blue, Fig. 2) map very closely to each other in the southern Sierra Nevada and adjacent deserts. We further explored structure within each group by performing a principal components analysis on each cluster individually, which are shown in Supplementary Figures 2-4.

Phylogenetic relationships among California pocket gophers

Using 1,228,667 LD-pruned SNPs to construct a maximum-likelihood phylogenetic tree for these 95 samples with RAxML (Fig. 3A), the three clusters observed in PC space represent three primary clades of California gopher species. When sampling localities were plotted on a map of California, clades correspond to individuals from different geographic regions. (Fig. 3B). In keeping with Patton and Smith (1990) we refer to these groups as the Great Basin unit (purple, individuals mostly assigned to *T. b. perpallidus*), the Central California unit (blue; individuals mostly assigned to *T. b. bottae*) and the Northern California unit (red; individuals mostly assigned to *T. b. navus*). Three samples that cluster closely with the Great Basin group in PC space represent a distinct lineage, termed the Basin and Range unit (shown in yellow in Fig. 3), which correspond with two *T. b. albatus* samples and one *T. b. perpallidus* sample.

This tree was constructed without an outgroup and is mid-point rooted, making it difficult to draw conclusions about the position of the ancestor or the order of splits between clades. However, bootstrap support for the four primary clades is high, as is support for relationships between closely related individuals at branch tips. The central California unit appears to be comprised of several distinct lineages, though the relationships between these populations are not well-resolved and have low bootstrap support. The Great Basin unit is comprised of two lineages, each containing many *T. b. perpallidus* individuals, with *T. b. operarius* nested within one lineage and *T. b. riparius* in the other. The Northern California unit does not form distinct lineages, but rather appears as one large clade with *T. b. laticeps* and *T. b. saxatilis* samples nested within *T. b. navus*. Additionally, the three samples that fall between clusters 2 and 3 on PCs 1 and 2 are assigned as sister to the Northern California clade, but bootstrap support for this placement is low.

Admixture identifies nine populations

To identify distinct populations for further assessment of genetic structure and diversity, we performed admixture analysis and varied the number of populations from k = 3 to k = 12; k = 9 had the lowest cross-validation error (CV error = 0.3378). Admixture identified the Northern California clade as a population and separated the Great Basin and Central California clades into three and five populations respectively (Fig. 3C). We classified three *T. b. nigricans* samples characterized as admixed as a potential tenth unique population based on prior research (Patton and Smith 1990; Alvarez-Castañeda 2010). We named these populations according to geography and most common member subspecies as follows: Northern (n = 20), Sierra-awahnee (n = 3), Sierra-mewa (n = 5), bottae-Central Calif. (n = 21), bottae-N. Calif. (n = 8), bottae-S. Cal (n. = 9), Nigricans (n = 3), Perpallidus-North (n = 7), Perpallidus-South (n = 15), and Albatus (n = 3), shown in order from left to right in Figure 3C.

Several individuals show evidence of admixture. Three individuals (*T. b. navus*) from the northern cluster 2 that separate slightly from the rest of the northern samples in PC space appear admixed between the Northern and Sierra-awahnee populations. These individuals were captured in the northern Sierra Nevada, while the Sierra-awahnee population is in the vicinity of Yosemite National Park, suggesting that there is past or present gene flow between Northern and Central California units of pocket gophers in the Sierras. A further three *T. b. pascalis* individuals are identified as admixed between bottae-central California and Sierra-mewa, and five individuals bear varying proportions of ancestry from the northern and southern populations within the Great Basin unit. One individual identified as belonging to *T. b. perpallidus* appears admixed between that unit and the Albatus population. Lastly, several individuals bear trace amounts of admixture suggestive of past gene flow between populations.

Population structure, genetic diversity, and inbreeding

Individuals were assigned to the population for which they bore the greatest ancestry proportion, and we used these populations to explore variation in genetic diversity and inbreeding across *T. bottae*. We randomly subsampled the full set of 282,721,405 SNPs to 15,621,715 with no minor allele frequency applied to avoid filtering out alleles that are private to local populations. Measures of genetic diversity and inbreeding are listed in Table 2 and pairwise F_{ST} is available in Table 3.

Genetic structure between populations was remarkably high, with genome-wide F_{ST} (taken as an average across 100-kb non-overlapping windows) exceeding 0.50 in more than half of pairwise comparisons between populations and ranging from $F_{ST} = 0.244$ between Perpallidus-North and South populations to $F_{ST} = 0.746$ between Perpallidus-North and Sierra-Awahnee. Between phylogeographic units, $F_{ST} = 0.363$ between Great Basin and Central California, 0.368 between Central California and Northern California, and 0.609 between Great Basin and Northern California. As a baseline for comparison, we randomly divided the 21 individuals from Bottae-Central Cal. into two groups (n=10 and n=11) and measured F_{ST} between them, which was 0.0014. We also measured F_{ST} between two named subspecies within the Northern population, *T. b. navus* and *T. b. laticeps*, and obtained an F_{ST} of 0.013.

We observed moderate genetic diversity within populations, with observed heterozygosity per nucleotide ranging from 0.00185 in Perpallidus-North to 0.00393 in Nigricans. Among named subspecies, *T. b. leucodon*, found within the Northern population at the far northern edge of California, had the lowest heterozygosity at 0.00109. Nucleotide diversity (π) was consistently higher than the average observed heterozygosity per individual but showed

similar patterns across populations. The Sierra-Mewa population displayed the highest nucleotide diversity at $\pi = 0.00495$. Populations from central California (Bottae-N. Cal $\pi = 0.00449$, Bottae-central Cal $\pi = 0.00490$), and the Northern population ($\pi = 0.0047$) also had nucleotide diversity on the high end of the range observed.

We also calculated Tajima's D for each population to summarize the distribution of allele frequencies and make inferences about changes in population size. Negative values of Tajima's D are associated with population expansions, while positive values are associated with population contractions (Tajima 1989). None of the observed values for Tajima's D differed significantly from 0. Eight out of ten populations have slightly to moderately negative Tajima's D, with the most negative values observed in Bottae-N. Cal (Tajima's D = -0.750) and Bottae-S. Cal (Tajima's D = -0.731). Two populations have slightly positive Tajima's D (Albatus Tajima's D = 0.126, Sierra-Awahnee Tajima's D = 0.054).

Consistent with the observation that nucleotide diversity was consistently higher than observed heterozygosity, inbreeding coefficients (F_{IS}) and the frequency of runs of homozygosity (ROH) were high and correlated ($r^2 = 0.59$; $p = 0.0035$), with F_{IS} ranging from 0.101 in Nigricans to $F_{IS} = 0.500$ in the Northern population, with most populations falling between $F_{IS} = 0.15$ and $F_{IS} = 0.25$. Considering just *T. b. navus* individuals from the Northern population, F_{IS} was 0.668. F_{ROH}, which measures how much of the genome is captured by runs of homozygosity, ranged from 0.070 in Nigricans to 0.370 in the Northern population. Further, individual variability in f_{ROH} was high, ranging from 0.0274 (Sierra-Mewa) up to a remarkable 0.8593 in an individual from the Northern population. Although we have used stringent quality filters, we note that an under-counting of rare heterozygous sites could lead to upwardly biased estimates of F_{IS} and F_{ROH}. As F_{IS} is sensitive to how populations are defined, the use of admixture-identified populations that span large geographic areas (e.g., Northern California) could also result in potential F_{IS} inflation.

Exploring color variation and genetics

We quantified color variation in 61 *T. bottae* individuals in this study for which study skins were available in the MVZ mammal collection. When we quantified color using the L*a*b* color space, subspecies varied significantly in L* (ANOVA $p = 3.77 \times 10^{-10}$), which measures perceptual lightness, or how light or dark a color is. A* and b* together measure variation in human-visible color space (red, yellow, green, blue); though a* varied significantly across subspecies (ANOVA $p = 0.0028$); only subspecies pairs *T. b. perpallidus – bottae* (*p-adj* = 0.0076) and *T. b. perpallidus – navus* (*p-adj* = 0.012) differed significantly when a Tukey post-hoc test was applied. B* did not differ significantly by subspecies (ANOVA $p = 0.076$) This makes sense given that subspecies were named in part according to observed color variation (Fig. 1), but most color variation occurs along the light-dark (L*) axis. While there is variation in how reddish or grayish individuals are, all individuals are similarly hued (brown) compared with the full spectrum of color as captured by a* and b*.

We also used a principal components analysis to explore color variation independent of subspecies assignment (Fig. 4A). When we plotted PCs 1 (67.8% of variation explained) and 2 (29.1% of variation explained), we observed variation within and across subspecies, which we have visualized as different shades of brown for each subspecies. Though lighter subspecies tended to occupy different areas of PC space than darker subspecies, there was no clear separation by subspecies, suggesting that color variation is relatively continuous across *T. bottae*.

When we tested for associations between genotype and color phenotype, a few individual SNPs had weakly significant associations with L* (Fig. 4B). However, no peaks or regions of the genome rose to significance about the background pattern of association. Ultimately, we were not able to identify any genes or sites that were linked to color variation (Supp. Figs. 5-6).

6.4 DISCUSSION

Whole-genome resequencing data for 95 *Thomomys bottae* pocket gophers from across the state of California (Fig. 1) yielded a remarkable 282 million variants, compared with 163 million variants called for a comparable California Conservation Genomics Project dataset that consists of six species of kangaroo rats and 142 individuals (Voss et al., in prep.). Whole-genome data supported most of the populations defined with allozyme data, though the phylogenetic placement of populations in the Sierra Nevada fits better with mitochondrial genetic studies from the 2000s (Smith 1998; Alvarez-Castañeda 2010). Further, pocket gopher populations are highly structured and inbred, yet retain moderately high levels of genetic diversity, providing a reference point for conservation genomic studies of other fragmented, isolated, and small populations.

Whole genome resequencing confirms four Thomomys bottae *units in California*

Three primary sample clusters emerge from principal components analysis when PCs 1 and 2 are plotted (Fig. 2), which are further separated into four groups in a maximum likelihood phylogenetic tree from all 95 samples (Fig. 3A). These four groups align with Patton and Smith's (1990) description of *T. bottae* geographic units as follows: Northern California, Central California, Great Basin, and Basin and Range in southeastern California. The Basin and Range group in southeastern California was recovered in the phylogenetic analysis but not in the principal components analysis (Fig. 2, Fig. 3A).

Unfortunately, the absence of an outgroup precludes drawing strong conclusions about the placement of the root of the tree in Figure 4C. However, we do find that the geographic boundaries between the major groups identified with whole genome sequences closely match those identified with allozyme data (Fig. 5A, B). In contrast, some analyses based on mitochondrial sequences greatly expand the Central California unit to include much of the northern California coast, the Sacramento Valley and Great Basin Desert (Álvarez-Castañeda 2010). The patterns identified with whole genome sequences support a Northern California clade extending from the Sacramento River northward inland and along the coast as well as a distinct Great Basin unit, as seen with allozyme data (Patton and Smith 1990).

Discrepancies between whole-genome sequences and allozyme data arise in the Sierra Nevada, where whole-genome data better reflect relationships defined with mitochondrial data. Patton and Smith (1990) group *T. b. awahnee*, *mewa*, and *alpinus* with the Great Basin unit, but our analysis suggests that they belong to either the Northern or Central California units (Figs. 2, 3; Smith 1998; Álvarez-Castañeda 2010), with the additional possibility of admixture between those two groups in the central Sierra Nevada. This discordance between nuclear and mitochondrial phylogenies may reflect larger patterns of incomplete lineage sorting (e.g., DeRaad et al. 2023), as Patton and Smith (1990) hypothesize that the four main groups reported here (Basin and Range, Great Basin, Central California, Northern California) diverged rapidly shortly after the origin of the species in the past one to two million years (Belfiore et al. 2008).

Evidence of gene flow between populations and units

Admixture analysis further separates the four main *T. bottae* groups into nine populations with some evidence of admixture between them (Fig. 3C). We identify two likely instances of gene flow across units: three *T. b. navus* individuals from the Northern population appear admixed with the Sierra-Awahnee population, and one individual collected near the Great Basin-Basin and Range boundary appears to bear ancestry from both units. The phylogenetic position of the Sierra-Awahnee population itself is uncertain; it is placed with Northern California in the best RAxML tree but has low bootstrap support and appears intermediate between Northern and Central California clusters in principal components analysis (Fig. 2, 3A). We also observe instances of putative gene flow between populations within the same geographic unit: numerous individuals are assigned mixed ancestry between Sierra-Mewa and Bottae-Central California populations, and several individuals from Perpallidus-North and South show varying levels of ancestry from each (Fig. 3C). Lastly, three *T. b. nigricans* individuals are reported as admixed between Sierra-Mewa and Bottae-S. Cal, but it is more likely that these individuals bear ancestral genetic variation stemming from a Baja California unit that is not captured by this study (Alvarez-Castañeda and Patton 2004; Alvarez-Castañeda 2010).

Populations are isolated, yet moderately heterozygous

Despite evidence of gene flow between populations in some regions of California, measurements of genetic structure (F_{ST}) suggest that populations are strongly isolated from each other. No F_{ST} value between the nine populations is lower than 0.24, and the majority of pairwise values revealed $F_{ST} > 0.5$ (Table 3). Further, measures of inbreeding (F_{IS} and F_{ROH}) suggest that individual populations are slightly to significantly inbred. Nigricans, Sierra-Awahnee, and Bottae-S. Calif. populations are isolated but appear less inbred than Northern, Bottae-Central Calif., or Perpallidus-North, with the Northern population displaying an inbreeding coefficient (F_{IS}) of 0.500. Though F_{ST} is similar, if slightly higher, than levels reported by Patton and Smith (1990), inbreeding coefficients (F_{IS}) are much higher, with no F_{IS} greater than 0.1 in allozyme data or in other studies (Patton and Feder 1981, Daly and Patton 1990). However, we note that differences in sampling design between allozyme and whole-genome resequencing studies may contribute to these divergent F_{IS} measurements.

Despite evidence of population structure and inbreeding, most *T. bottae* populations maintain moderate levels of genetic diversity, and relative heterozygosity across populations is consistent with the findings of Patton and Smith (1990). Even populations with the lowest amount of genetic variation (Perpallidus-North; $\pi = 0.0024$, het. = 0.00185) have levels of nucleotide diversity comparable to those seen in other rodents (Teixeira and Huber 2021; Wilder et al. 2022). Moreover, the Northern population, which appears to have high levels of inbreeding, still displays moderately high levels of genetic diversity ($\pi = 0.00447$, het. = 0.00237). Populations in central and southwestern California tend to have the highest heterozygosity (*T. b. bottae*, *mewa, nigricans*), while isolated populations in the Great Basin and far northern California have the lowest heterozygosity (*T. b. operarius*, *riparius*, *saxatilis*, *leucodon*). As genetic diversity is a function of effective population size, Patton and Smith (1990) suggested that this reflects variability in habitat quality and resulting population numbers. In central California's agricultural fields and pastures, populations are likely larger and more connected over time compared with isolated desert valleys and mountaintops, which support smaller, more frequently isolated populations and hence display lower genetic diversity.

However, this does not explain the combination of high genetic diversity and inbreeding we observe in several populations. We speculate that large regional population size and micro-structure between local populations could lead to this outcome. Daly and Patton (1990) measured dispersal in a population at Hastings Natural History Reservation and found that more than seventy percent of adults first captured as juveniles were found within 40 meters of their juvenile trapping location, and only six percent of individuals dispersed more than 200 meters. Short dispersal distances could lead to mating between related individuals, resulting in high F_{IS} values despite moderate to high genetic diversity. Similar F_{IS} values can be observed in house mice (*Mus musculus*), which have huge global population sizes but live and reproduce in locally structured demes (Petras 1967; Selander 1970).

Conservation genomics of pocket gophers

One of the founding principles in the field of conservation genetics derives from Kimura's neutral theory (1983): since genetic diversity is a function of population size, measuring genetic diversity can be used to infer population size and thereby inform conservation practices. Calculating population genetic statistics such as heterozygosity, F_{IS}, and more recently, F_{ROH} have become popular ways to indirectly measure the impact of population decline on vulnerable or endangered species (e.g. Chen et al. 2016; Kleinman-Ruiz et al. 2022). Some endangered species have extremely low genetic diversity: Robinson et al (2019) measured heterozygosity per nucleotide in the Isle Royale wolves, a classic example of inbreeding depression, at 0.00096, an order of magnitude lower than in *T. bottae*. Heterozygosity in California condors, meanwhile, is only slightly lower than in some *T. bottae* populations at 0.00137 (Robinson et al. 2021), and some extremely genetically depauperate populations, such as the Channel Island foxes, with het. = 0.000142, display no evidence of inbreeding depression. Further, meta-analyses of conservation genetic data suggest that there is no relationship between genetic diversity and IUCN Red List status (Schmidt et al. 2023). These mixed findings regarding genetic diversity have led to debate over the utility of summary statistics in conservation genomics (e.g. Teixeira and Huber 2021; DeWoody et al. 2021).

Thomomys bottae displays levels of inbreeding on par with some highly endangered species. For example, the average F_{IS} observed in a captive population of Mexican wolves derived from seven founder individuals is 0.227 (Clement et al. 2024). The Devil's Hole pupfish, with the smallest known range of any vertebrate and a census population that has dropped as low as 35, has an F_{ROH} of 0.34–0.81 (Tian et al. 2022). If evaluated solely based on measures such as F_{ROH}, some populations of *T. bottae*, such as the Northern California unit, might be considered species of concern. Indeed, Halsey et al. (2025) suggest that small, isolated populations of pocket gophers in Texas may warrant population management for conservation. *T. bottae* populations can be ephemeral, and inbreeding could be a factor in which populations survive. However, the inbreeding observed in *T. bottae* serves as a reminder that moderate to high levels of inbreeding may not always be indicators of inbreeding depression, and that measures of inbreeding will be most useful for guiding conservation decisions when coupled with other data.

The California Conservation Genomics Project is conducting landscape-level whole genome resequencing for 200 species, providing the opportunity to compare genetic diversity and inbreeding across a wide array of organisms in a standardized manner. Additionally, the project was designed to include species that are widespread but have endangered populations, such as Merriam's kangaroo rat (*Dipodomys merriami*), and closely related species of different conservation status, such as California and Gambel's quail (Benham et al., in press). This

comparative perspective, both broadly across many taxa and specifically across closely related comparisons, offers an opportunity to understand whether there are particular contexts in which population genetic statistics are most informative for conservation biology.

Color varies within and across clades

Botta's pocket gopher varies in color from near white to dark brown, gray, and auburn (Fig. 1) Color variation broadly matches soil type and moisture: individuals in desert habitats are generally lighter than samples from northern California and the Sierra Nevada, with additional color variation reported outside of the California range that is not captured by this study. This background-matching dorsal coloration is presumed to be under selection to avoid visual predators, though this hypothesis has never been tested (Ingles 1950; Janes and Barss 1985).

Data generated for this study presented an opportunity to look for genetic variants associated with coat color in a genome-wide manner for the first time, as every individual whose genome was sequenced is a vouchered specimen at the Museum of Vertebrate Zoology, and 61 of 95 WGS samples had been prepared as study skins, so it was possible to measure coat color from study skins. We quantified color and tested for associations between genetic loci and color variation but were not able to recover any associations between genomic variation and color variation. Though we used a kinship matrix to control for population structure as much as possible, the high F_{ST} observed between populations suggests that the 'noise' of background genetic differentiation will likely swamp most signals of genotype-phenotype association in *T. bottae* unless much larger sample sizes become available. More generally, these findings underscore the challenges of identifying the genetic basis of polygenic traits through association in highly structured populations.

Conclusion

Initial divergence between four clades of Botta's pocket gophers in California was likely allopatric and may have been driven by large-scale changes to the landscape of the North American southwest that occurred during the early history of *Thomomys bottae* one to three million years ago (Patton and Smith 1990; Belfiore et al. 2008). Since then, population isolation, resulting genetic drift, and possibly local adaptation has driven high levels of structure and inbreeding within large-scale geographic units. However, population sizes and regional connectivity are sufficient to maintain reasonable levels of genetic diversity, which may be the key difference between inbred *T. bottae* populations and similarly inbred endangered species. It is an open question whether summary statistics of genetic diversity have predictive value for population management, but *T. bottae* provides a valuable data point as practitioners in this field evaluate possible next steps in conservation genomics. Ultimately, Botta's pocket gopher has a complex biogeographic history characterized by rapid divergence followed by genetic drift, incomplete lineage sorting, and periodic gene flow, which facilitate local variation and result in the wide array of forms observed by Merriam (1895) and Grinnell (1927).

6.5 FIGURES

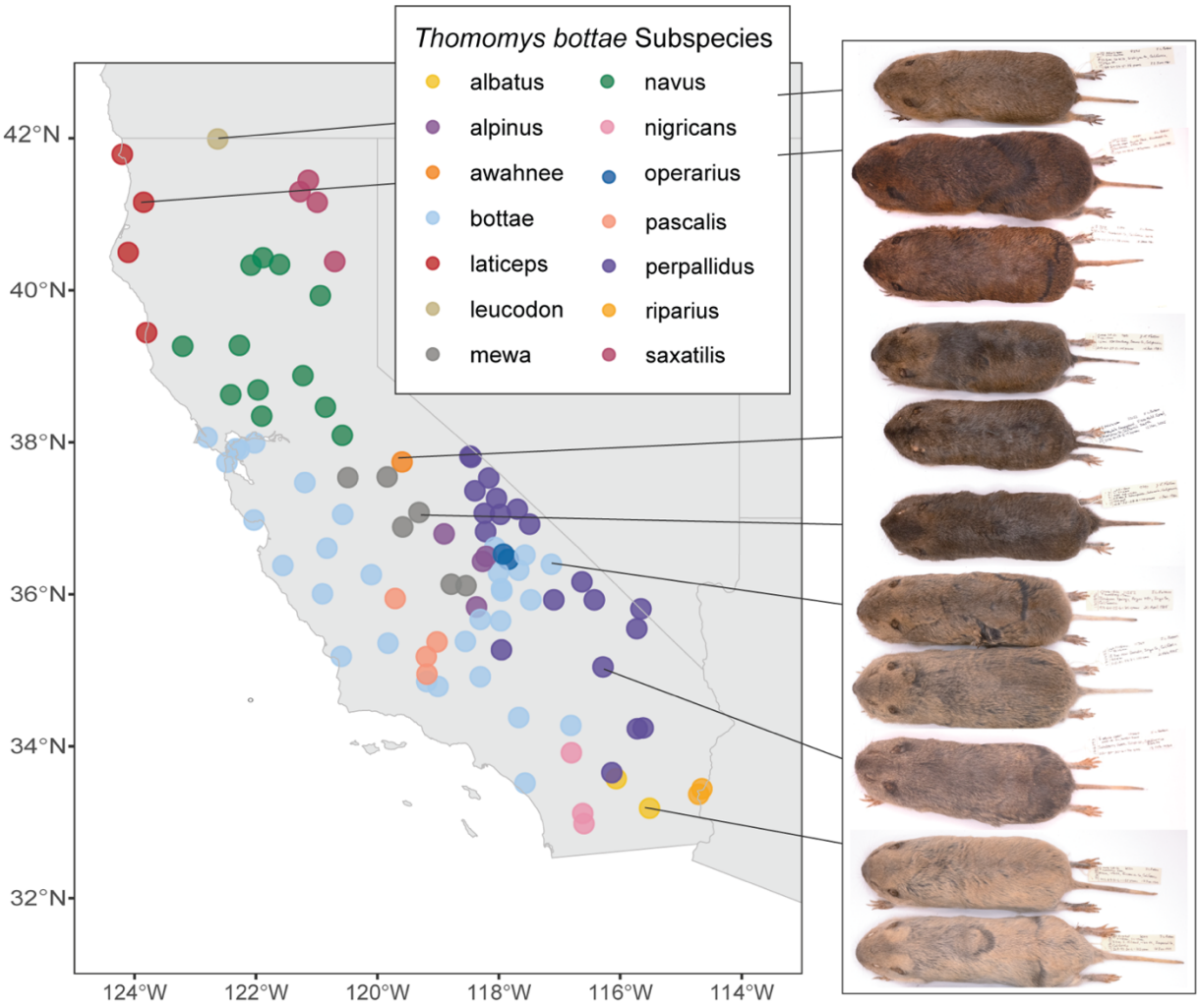


Figure 1. Left: Sampling localities for 96 vouchered specimens of Botta's pocket gopher (*Thomomys bottae)* from the mammal collection at the Museum of Vertebrate Zoology (University of California, Berkeley) whose genomes were sequenced as a part of the California Conservation Genomics Project. Samples are colored by subspecies, with 14 subspecies represented in this study. Sample sizes per subspecies are available in Table 1, and metadata and accession numbers for all individuals included in this study are available in Supplementary Table 1. Right: Museum specimens from this study demonstrate the extensive variation in coat color observed *T. bottae* populations across the state, which is presumed to be a local adaptation for camouflage against the wide variety of soil types that pocket gophers inhabit across their range.

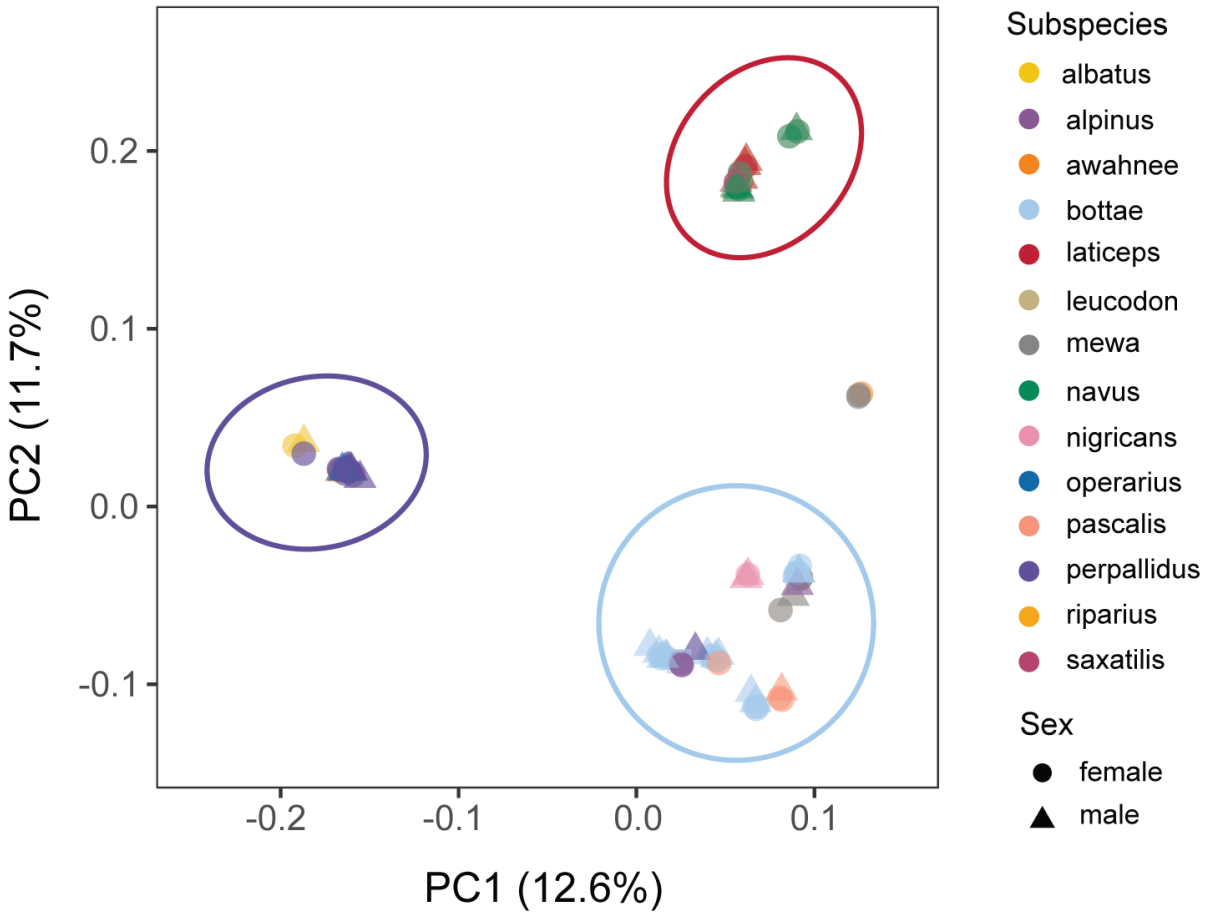


Figure 2. Principal components analysis of genetic variation among 95 Botta's pocket gophers sampled from across California populations of *Thomomys bottae*. We performed PCA using 5,766,144 LD-pruned SNPs, which clusters *T. bottae* individuals into three groups along PC axes 1 (12.6% of variation explained) and 2 (11.7% of variation explained). Each point represents one individual; point color represents subspecies assignment and shape indicates the sex of the individual.

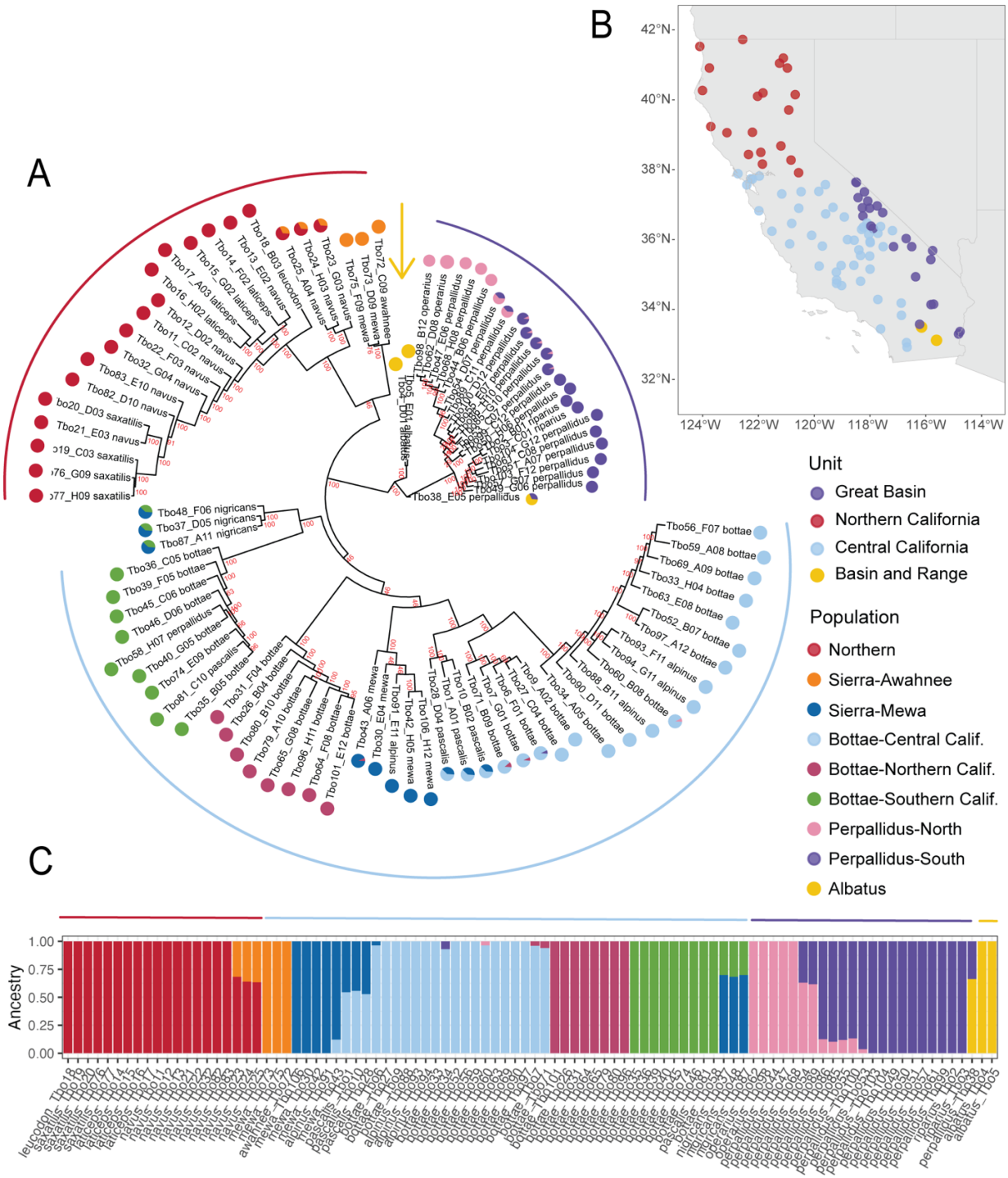


Figure 3. Population structure and phylogenetic relationships between *Thomomys bottae* populations across California. A) Three clusters identified in principal components analysis (Fig. 2) emerge as clades when we construct a maximum-likelihood phylogenetic tree using RAxML-ng with 1,228,667 variants. An additional fourth clade comprised of two samples appears in

phylogenetic but not principal components analysis. Bootstrap values indicate number of bootstraps out of 100 that supported each node. Population assignments with admixture (C) are indicated by the corresponding-colored circle at each tip. Where individuals are admixed between populations, ancestry proportions are indicated as a pie chart. Colored bars outside each clade correspond to colors in B), where sampling localities are plotted on a map of California. Each point represents a single individual and is colored according to phylogenetic unit. Units occupy distinct geographic regions of California, and we have named them in keeping with Patton and Smith's (1990) identifiers: Great Basin (purple), Northern California (red), Central California (light blue), and Basin and Range (yellow). C) Admixture analysis with the same set of 5,766,144 LD-pruned SNPs used for principal components analysis best supports a model of k = 9 populations (lowest cross-validation error; CV = 0.3378), which identify the Northern California unit as a single population and further divide units in Central California and the Great Basin Desert into multiple sub-populations with some evidence of gene flow. Each vertical bar represents one individual, and colors represent ancestry proportions from each of the nine populations. Populations are named and colored as follows from left to right: Northern (red), Sierra-Awahnee (orange), Sierra-Mewa (dark blue), Bottae-Central California (light blue), Bottae-Northern California (dark pink), Bottae-Southern California (green), Perpallidus-North (light pink), Perpallidus-South (purple), and Albatus (yellow). We treat three admixed (green-blue) individuals assigned to *T. b. nigricans* as a distinct population for summary statistics. Individuals are assigned to the population for which they have the highest ancestry proportion, and summary statistics related to population structure, genetic diversity, and inbreeding are reported in Tables 2 and 3. Assignment to geographic units shown in A and B is indicated by the colored bar over the admixture plot.

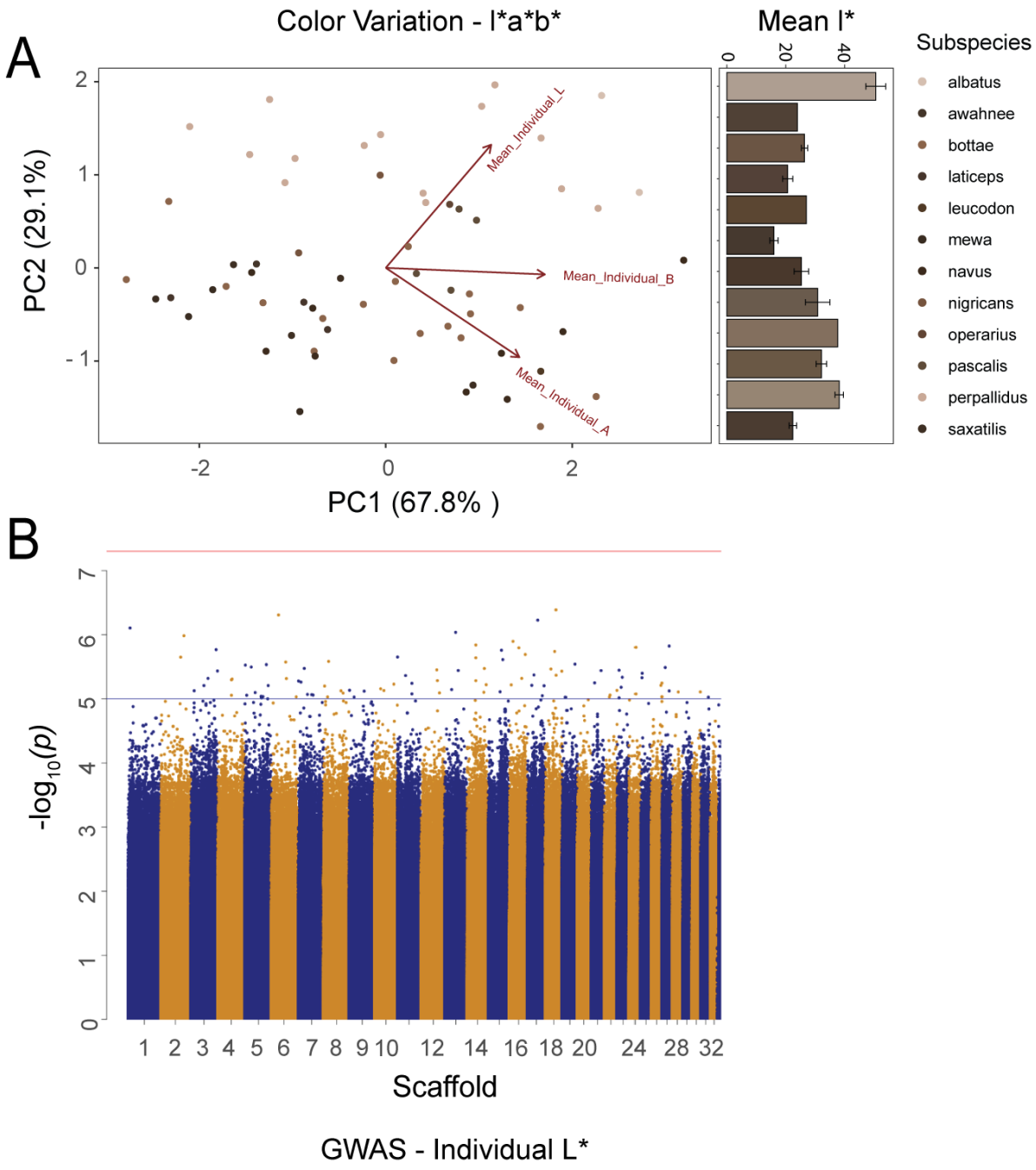


Figure 4. A) Left: principal components analysis summarizing coat color variation for n = 61 vouchered museum study skins who were included in whole-genome resequencing as a part of this study. We photographed each specimen and quantified average dorsal coat color using the L*a*b* color system; PC loadings for each color variable are indicated with arrows. Each point represents one individual in L*a*b* color space; subspecies have been colored to approximate color variation observed in specimens. Right: though in color variation is relatively continuous across individuals from different subspecies in the principal components analysis, with no apparent clusters by color, there is significant variation in lightness (l*) by subspecies (ANOVA

$p = 3.77 \times 10^{-10}$). B) Genome-wide association test for associations between variation in L* (lightness) and 5,121,497 SNPs from across the T. bottae genome. Scaffolds are indicated with alternating blue and yellow points, and association p-values are reported on a $-log_{10}p$ scale. Significance thresholds are indicated with red and blue horizontal lines. A few variants had weakly significant associations with L*, but we did not identify any peaks of significance with strong color-genotype associations. Manhattan plots for GWAS performed with a* and b* are available in Supplementary Figures 5 and 6.

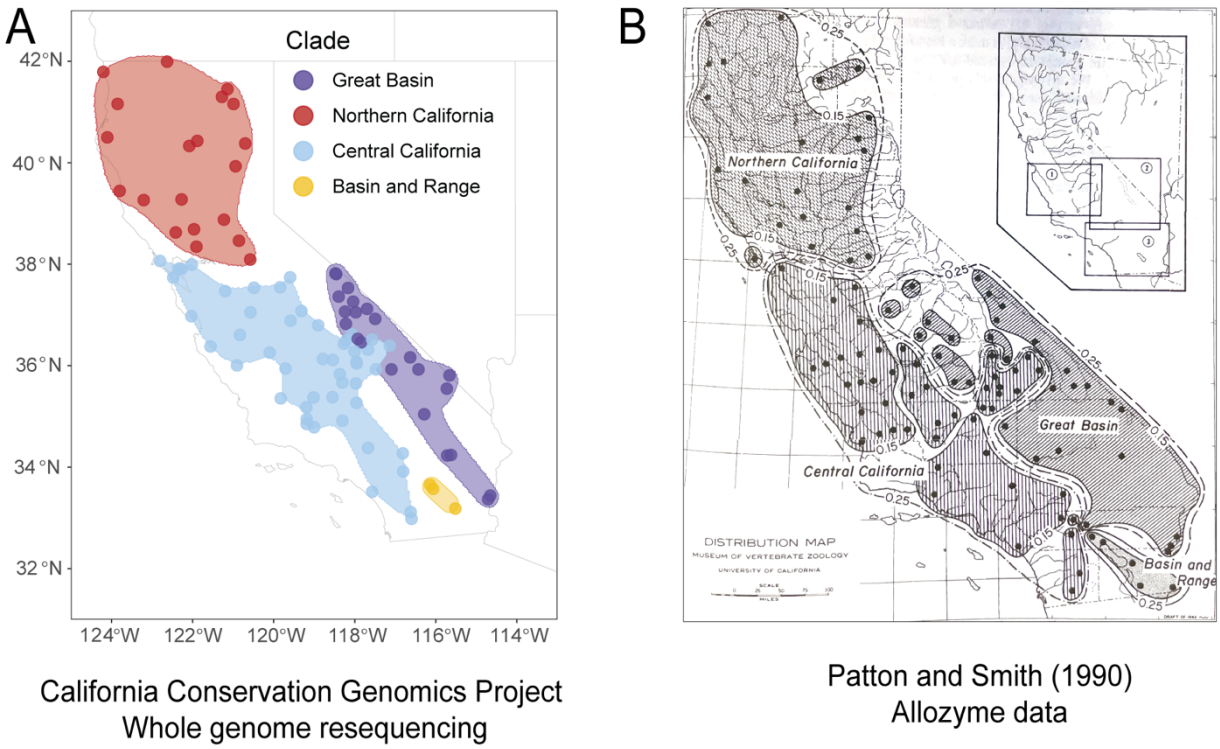


Figure 5. A) Geographic units of *T. bottae* in California as defined with whole-genome resequencing data include the Great Basin (purple), Northern California (red), Central California (blue) and Basin and Range (yellow). Sampling localities are indicated as points colored by clade assignment. Minimum geographic extent of each unit shown as a convex polygon encompassing all sampling localities for that group. We further explored structure within each clade by performing a principal components analysis for each geographic unit individually; results are shown in Supplementary Figures 2, 3, and 4. B) Patton and Smith (1990) identified four clades of Botta's pocket gophers within California as identified with allozyme data. Except for several populations in the central Sierra Nevada, allozyme and whole-genome resequencing data results are largely concordant.

6.6 TABLES

Table 1. Samples

Subspecies	n
T. b. albatus	2
T. b. alpinus	4
T. b. awahnee	1
T. b. bottae	31
T. b. laticeps	4
T. b. leucodon	1
T. b. mewa	6
T. b. navus	12
T. b. nigricans	3
T. b. operarius	2
T. b. pascalis	4
T. b. perpallidus	20
T. b. riparius	2
T. b. saxatilis	4

Table 2. Population genetic diversity and inbreeding statistics for admixture-defined populations of Botta's pocket gopher (*Thomomys bottae*) in California. Metrics are included for some named subspecies where they provide additional context for outlier populations. Π and Tajima's D were averaged across 100-kb non-overlapping windows.

Population	π	heterozygosity	Tajima's D	F_{IS}	f_{ROH}
Albatus	0.00380	0.00309	0.126	0.193	0.240
Sierra-Awahnee	0.00377	0.00237	0.054	0.165	0.126
Bottae N. Calif	0.00449	0.00390	-0.750	0.184	0.132
Bottae central Calif.	0.00490	0.00332	-0.465	0.369	0.202
Bottae S. Calif.	0.00390	0.00331	-0.731	0.176	0.125
Sierra-Mewa	0.00495	0.00368	-0.342	0.268	0.082
Nigricans	0.00434	0.00393	-0.120	0.101	0.071
Northern	0.00447	0.00237	-0.970	0.500	0.370
T. b. laticeps	0.00377	0.00211	0.353	0.442	0.390
T. b. navus	0.00752	0.00293	-0.511	0.668	0.302
T. b. saxatilis	-	0.00133	-	-	0.561
T. b. leucodon	-	0.00109	-	-	0.491
Perpallidus South	0.00320	0.00255	-0.378	0.212	0.191
T. b. riparius	-	0.00191	-	-	0.353
Perpallidus North	0.00239	0.00185	-0.346	0.201	0.204
T. b. operarius	-	0.00134	-	-	0.265

Table 3. Pairwise genome-wide F_{ST} between California *Thomomys bottae* populations.

Population	Albatus	Sierra-awahnee	Bottae N. Calif.	Bottae central Calif.	Bottae S. Calif.	Sierra-mewa	Nigricans	Northern	Perpallidus -S
Albatus	-								
Sierra-Awahnee	0.639	-							
Bottae-N. Calif.	0.514	0.570	-						
Bottae-central Cal.	0.440	0.509	0.395	-					
Bottae-S. Calif.	0.566	0.631	0.521	0.396	-				
Sierra-Mewa	0.472	0.509	0.439	0.363	0.505	-			
Nigricans	0.540	0.627	0.517	0.451	0.521	0.480	-		
Northern	0.493	0.524	0.473	0.466	0.530	0.458	0.501	-	
Perpallidus-S	0.443	0.673	0.580	0.492	0.608	0.567	0.621	0.561	-
Perpallidus-N	0.551	0.746	0.613	0.508	0.648	0.604	0.680	0.573	0.244

REFERENCES

1000 Genomes Project Consortium, Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., Korbel, J. O., Marchini, J. L., McCarthy, S., McVean, G. A., & Abecasis, G. R. (2015). A global reference for human genetic variation. *Nature* 526:68–74. https://doi.org/10.1038/nature15393

Abascal, F., Corvelo, A., Cruz, F., Villanueva-Cañas, J. L., Vlasova, A., Marcet-Houben, M., Martínez-Cruz, B., Cheng, J. Y., Prieto, P., Quesada, V., Quilez, J., Li, G., García, F., Rubio-Camarillo, M., Frias, L., Ribeca, P., Capella-Gutiérrez, S., Rodríguez, J. M., Câmara, F., … Godoy, J. A. (2016). Extreme genomic erosion after recurrent demographic bottlenecks in the highly endangered Iberian lynx. *Genome Biology* 17:251. https://doi.org/10.1186/s13059-016-1090-1

Abdennur, N., & Mirny, L. A. (2020). Cooler: scalable storage for Hi-C data and other genomically labeled arrays. *Bioinformatics* 36:311–316. https://doi.org/10.1093/bioinformatics/btz540

Alexander, D. H., & Lange, K. (2011). Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *BMC Bioinformatics* 12:246. https://doi.org/10.1186/1471-2105-12-246

Alexander, L. F., & Riddle, B. R. (2005). Phylogenetics of the New World rodent family Heteromyidae. *Journal of Mammalogy* 86:366–379. https://doi.org/10.1644/BER-120.1

Allio, R., Schomaker-Bastos, A., Romiguier, J., Prosdocimi, F., Nabholz, B., & Delsuc, F. (2020). MitoFinder: Efficient automated large-scale extraction of mitogenomic data in target enrichment phylogenomics. *Molecular Ecology Resources* 20:892–905. https://doi.org/10.1111/1755-0998.13160

Alonge, M., Lebeigle, L., Kirsche, M., Jenike, K., Ou, S., Aganezov, S., Wang, X., Lippman, Z. B., Schatz, M. C., & Soyk, S. (2022). Automated assembly scaffolding using RagTag elevates a new tomato system for high-throughput genome editing. *Genome Biology* 23:258. https://doi.org/10.1186/s13059-022-02823-7

Alvarez-Castañeda, S. T. (2010). Phylogenetic structure of the *Thomomys bottae-umbrinus* complex in North America. *Molecular Phylogenetics and Evolution* 54:671–679. https://doi.org/10.1016/j.ympev.2009.11.012

Álvarez-Castañeda, S. T., Lidicker, W. Z., & Rios, E. (2009). Revision of the *Dipodomys merriami* complex in the Baja California peninsula, Mexico. *Journal of Mammalogy* 90:992–1008. https://doi.org/10.1644/07-MAMM-A-398.1

Alvarez-Castaneda, S. T., & Patton, J. L. (2004). Geographic genetic architecture of pocket gopher (*Thomomys bottae*) populations in Baja California, Mexico. *Molecular Ecology* 13:2287–2301. https://doi.org/10.1111/j.1365-294X.2004.02243.x

Anderson, M. J., Nyholt, J., & Dixson, A. F. (2005). Sperm competition and the evolution of sperm midpiece volume in mammals. *Journal of Zoology* 267:135–142. https://doi.org/10.1017/S0952836905007284

Andrews, S., & Others. (2010). *FastQC: a quality control tool for high throughput sequence data*. Babraham Bioinformatics, Babraham Institute, Cambridge, United Kingdom.

Archer, S. R., & Predick, K. I. (2008). Climate change and ecosystems of the southwestern United States. *Rangelands*, *30*(3), 23–28. https://doi.org/10.2111/1551-501X(2008)30[23:CCAEOT]2.0.CO;2

Balkenhol, N., Cushman, S. A., Storfer, A., & Waits, L. P. (2015). Introduction to landscape genetics - concepts, methods, applications. In *Landscape Genetics* (pp. 1–8). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118525258.ch01

Barros, M. A., & Patton, J. L. (1985). Genome evolution in pocket gophers (genus *Thomomys*). III. Fluorochrome-revealed heterochromatin heterogeneity. *Chromosoma* 92:337–343. https://doi.org/10.1007/BF00327464

Bedford, N. L., & Hoekstra, H. E. (2015). *Peromyscus* mice as a model for studying natural variation. *eLife* 4:06813. https://doi.org/10.7554/eLife.06813

Begun, D. J., Whitley, P., Todd, B. L., Waldrip-Dail, H. M., & Clark, A. G. (2000). Molecular population genetics of male accessory gland proteins in *Drosophila*. *Genetics* 156:1879–1888. https://doi.org/10.1093/genetics/156.4.1879

Belfiore, N. M., Liu, L., & Moritz, C. (2008). Multilocus phylogenetics of a rapid radiation in the genus *Thomomys* (Rodentia: Geomyidae). *Systematic Biology* 57:294–310. https://doi.org/10.1080/10635150802044011

Bendesky, A., Kwon, Y.-M., Lassance, J.-M., Lewarch, C. L., Yao, S., Peterson, B. K., He, M. X., Dulac, C., & Hoekstra, H. E. (2017). The genetic basis of parental care evolution in monogamous mice. *Nature* 544:434–439. https://doi.org/10.1038/nature22074

Benedict, B. D., Castellanos, A. A., & Light, J. E. (2019). Phylogeographic assessment of the Heermann's kangaroo rat (*Dipodomys heermanni*). *Journal of Mammalogy* 100:72–91. https://doi.org/10.1093/jmammal/gyy166

Benham, P. M., Cicero, C., Escalona, M., Beraut, E., Marimuthu, M. P. A., Nguyen, O., Nachman, M., & Bowie, R. C. (2023). A highly contiguous genome assembly for the California quail (*Callipepla californica*). *The Journal of Heredity* 114:418–427. https://doi.org/10.1093/jhered/esad008

Benham, P. M., Walsh, J., & Bowie, R. C. K. (2024). Spatial variation in population genomic responses to over a century of anthropogenic change within a tidal marsh songbird. *Global*

Change Biology 30:e17126. https://doi.org/10.1111/gcb.17126

Bernt, M., Donath, A., Jühling, F., Externbrink, F., Florentz, C., Fritzsch, G., Pütz, J., Middendorf, M., & Stadler, P. F. (2013). MITOS: improved de novo metazoan mitochondrial genome annotation. *Molecular Phylogenetics and Evolution* 69:313–319. https://doi.org/10.1016/j.ympev.2012.08.023

Berry, D. L., & Baker, R. J. (1971). Apparent convergence of karyotypes in two species of pocket gophers of the genus *Thomomys* (Mammalia, Rodentia). *Cytogenetics* 10:1–9. https://doi.org/10.1159/000130121

Bertorelle, G., Raffini, F., Bosse, M., Bortoluzzi, C., Iannucci, A., Trucchi, E., Morales, H. E., & van Oosterhout, C. (2022). Genetic load: genomic estimates and applications in non-model animals. *Nature Reviews Genetics* 23:492-503. https://doi.org/10.1038/s41576-022-00448-x

Bertram, M. J., Neubaum, D. M., & Wolfner, M. F. (1996). Localization of the *Drosophila* male accessory gland protein Acp36DE in the mated female suggests a role in sperm storage. *Insect Biochemistry and Molecular Biology* 26:971–980. https://doi.org/10.1016/s0965-1748(96)00064-1

Bi, K., Linderoth, T., Singhal, S., Vanderpool, D., Patton, J. L., Nielsen, R., Moritz, C., & Good, J. M. (2019). Temporal genomic contrasts reveal rapid evolutionary responses in an alpine mammal during recent climate change. *PLoS Genetics* 15:e1008119. https://doi.org/10.1371/journal.pgen.1008119

Birdsall, D. A., & Nash, D. (1973). Occurrence of successful multiple insemination of females in natural populations of deer mice (*Peromyscus maniculatus*). *Evolution* 27:106–110. https://doi.org/10.1111/j.1558-5646.1973.tb05922.x

Birkhead, T. R., & Pizzari, T. (2002). Postcopulatory sexual selection. *Nature Reviews Genetics*, 3:262–273. https://doi.org/10.1038/nrg774

Bishop, A. P., Westeen, E. P., Yuan, M. L., Escalona, M., Beraut, E., Fairbairn, C., Marimuthu, M. P. A., Nguyen, O., Chumchim, N., Toffelmier, E., Fisher, R. N., Shaffer, H. B., & Wang, I. J. (2023). Assembly of the largest squamate reference genome to date: The western fence lizard, *Sceloporus occidentalis*. *The Journal of Heredity* 114:521–528. https://doi.org/10.1093/jhered/esad037

Bittner, N. K. J., Mack, K. L., & Nachman, M. W. (2021). Gene expression plasticity and desert adaptation in house mice. *Evolution* 75: 1477–1491. https://doi.org/10.1111/evo.14172

Blackhawk, N. C., Germano, D. J., & Smith, P. T. (2016). Genetic variation among populations of the endangered giant kangaroo rat, *Dipodomys ingens*, in the southern San Joaquin valley. *The American Midland Naturalist* 175:261–274. https://doi.org/10.1674/0003-0031-175.2.261

Boria, R. A., & Blois, J. L. (2023). Phylogeography within the *Peromyscus maniculatus* species

group: Understanding past distribution of genetic diversity and areas of refugia in western North America. *Molecular Phylogenetics and Evolution* 180:107701. https://doi.org/10.1016/j.ympev.2023.107701

Bradley, R. D., Durish, N. D., Rogers, D. S., Miller, J. R., Engstrom, M. D., & Kilpatrick, C. W. (2007). Towards a molecular phylogeny for *Peromyscus*: evidence from mitochondrial Cytochrome-B sequences. *Journal of Mammalogy* 88:1146–1159. https://doi.org/10.1644/06-MAMM-A-342R.1

Brawand, D., Soumillon, M., Necsulea, A., Julien, P., Csárdi, G., Harrigan, P., Weier, M., Liechti, A., Aximu-Petri, A., Kircher, M., Albert, F. W., Zeller, U., Khaitovich, P., Grützner, F., Bergmann, S., Nielsen, R., Pääbo, S., & Kaessmann, H. (2011). The evolution of gene expression levels in mammalian organs. *Nature* 478:343–348. https://doi.org/10.1038/nature10532

Brock, R. E., & Kelt, D. A. (2004). Keystone effects of the endangered Stephens' kangaroo rat (*Dipodomys stephensi*). *Biological Conservation* 116:131–139. https://doi.org/10.1016/S0006-3207(03)00184-8

Brown, J. H., & Heske, E. J. (1990). Control of a desert-grassland transition by a keystone rodent guild. *Science* 250:1705–1707. https://doi.org/10.1126/science.250.4988.1705

Brüniche-Olsen, A., Kellner, K. F., Belant, J. L., & DeWoody, J. A. (2021). Life-history traits and habitat availability shape genomic diversity in birds: implications for conservation. *Proceedings of the Royal Society of London. Series B, Biological Sciences* 288:2021.1441. https://doi.org/10.1098/rspb.2021.1441

Bryant, D. M., Johnson, K., DiTommaso, T., Tickle, T., Couger, M. B., Payzin-Dogru, D., Lee, T. J., Leigh, N. D., Kuo, T.-H., Davis, F. G., Bateman, J., Bryant, S., Guzikowski, A. R., Tsai, S. L., Coyne, S., Ye, W. W., Freeman, R. M., Jr, Peshkin, L., Tabin, C. J., … Whited, J. L. (2017). A tissue-mapped axolotl de novo transcriptome enables identification of limb regeneration factors. *Cell Reports* 18:762–776. https://doi.org/10.1016/j.celrep.2016.12.063

Burgoyne, P. S., Mahadevaiah, S. K., & Turner, J. M. A. (2009). The consequences of asynapsis for mammalian meiosis. *Nature Reviews Genetics* 10:207–216. https://doi.org/10.1038/nrg2505

Bushmanova, E., Antipov, D., Lapidus, A., Suvorov, V., & Prjibelski, A. D. (2016). rnaQUAST: a quality assessment tool for de novo transcriptome assemblies. *Bioinformatics* 32:2210–2212. https://doi.org/10.1093/bioinformatics/btw218

Bustamante, C. D., Fledel-Alon, A., Williamson, S., Nielsen, R., Hubisz, M. T., Glanowski, S., Tanenbaum, D. M., White, T. J., Sninsky, J. J., Hernandez, R. D., Civello, D., Adams, M. D., Cargill, M., & Clark, A. G. (2005). Natural selection on protein-coding genes in the human genome. *Nature* 437:1153–1157. https://doi.org/10.1038/nature04240

California Natural Resources Agency. Pathways to 30 × 30 California. (2022). https://canature.maps.arcgis.com/sharing/rest/content/items/8da9faef231c4e31b651ae6dff95254e/data. Accessed 05 August 2025.
Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: architecture and applications. *BMC Bioinformatics* 10:421. https://doi.org/10.1186/1471-2105-10-421

Capella-Gutiérrez, S., Silla-Martínez, J. M., & Gabaldón, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25:1972–1973. https://doi.org/10.1093/bioinformatics/btp348

Carroll, S. B. (2005). Evolution at two levels: on genes and form. *PLoS Biology* 3:e245. https://doi.org/10.1371/journal.pbio.0030245

Challis, R., Kumar, S., Sotero-Caio, C., Brown, M., & Blaxter, M. (2023). Genomes on a Tree (GoaT): A versatile, scalable search engine for genomic and sequencing project metadata across the eukaryotic tree of life. *Wellcome Open Research* 8:24. https://doi.org/10.12688/wellcomeopenres.18658.1

Challis, R., Richards, E., Rajan, J., Cochrane, G., & Blaxter, M. (2020). BlobToolKit - interactive quality assessment of genome assemblies. *G3* 10:1361–1374. https://doi.org/10.1534/g3.119.400908

Chambers, E. A., Bishop, A. P., & Wang, I. J. (2025). Individual-based landscape genomics for conservation: An analysis pipeline. *Molecular Ecology Resources* 25:e13884. https://doi.org/10.1111/1755-0998.13884

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* 4:7. https://doi.org/10.1186/s13742-015-0047-8

Chapman, T., Liddle, L. F., Kalb, J. M., Wolfner, M. F., & Partridge, L. (1995). Cost of mating in *Drosophila melanogaster* females is mediated by male accessory gland products. *Nature* 373:241–244. https://doi.org/10.1038/373241a0

Chen, P. S., Stumm-Zollinger, E., Aigaki, T., Balmer, J., Bienz, M., & Böhlen, P. (1988). A male accessory gland peptide that regulates reproductive behavior of female *D. melanogaster*. *Cell* 54:291–298. https://doi.org/10.1016/0092-8674(88)90192-4

Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 34:i884–i890. https://doi.org/10.1093/bioinformatics/bty560

Cheng, H., Concepcion, G. T., Feng, X., Zhang, H., & Li, H. (2021). Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nature Methods* 18:170–175. https://doi.org/10.1038/s41592-020-01056-5

Cheng, H., Jarvis, E. D., Fedrigo, O., Koepfli, K.-P., Urban, L., Gemmell, N. J., & Li, H. (2022). Robust haplotype-resolved assembly of diploid individuals without parental data. *Nature Biotechnology* 40:1332-1335. https://www.nature.com/articles/s41587-022-01261-x

Chesemore, D. L., Rhodehamel, W. M., Williams, D. F., Byrne, S., & Rado, T. A. (1992). Ecology of a vanishing subspecies: the Fresno kangaroo rat (*Dipodomys nitratoides exilis*). *Endangered and Sensitive Species of the San Joaquin Valley, California.* Calif. Energy Comm., Sacramento, USA.

Chock, R. Y., McCullough Hennessy, S., Wang, T. B., Gray, E., & Shier, D. M. (2020). A multi-model approach to guide habitat conservation and restoration for the endangered San Bernardino kangaroo rat. *Global Ecology and Conservation 21*:e00881. https://doi.org/10.1016/j.gecco.2019.e00881

Clark, A. G., Aguadé, M., Prout, T., Harshman, L. G., & Langley, C. H. (1995). Variation in sperm displacement and its association with accessory gland protein loci in *Drosophila melanogaster*. *Genetics* 139:189–201. https://doi.org/10.1093/genetics/139.1.189

Clark, A. G., Begun, D. J., & Prout, T. (1999). Female x male interactions in *Drosophila* sperm competition. *Science* 283:217–220. https://doi.org/10.1126/science.283.5399.217

Clement, M. J., Oakleaf, J. K., Heffelfinger, J. R., Gardner, C., deVos, J., Rubin, E. S., Greenleaf, A. R., Dilgard, B., & Gipson, P. S. (2024). An evaluation of potential inbreeding depression in wild Mexican wolves. *The Journal of Wildlife Management* 88:e22640. https://doi.org/10.1002/jwmg.22640

Courcelle, M., Fabre, P.-H., & Douzery, E. J. P. (2023). Phylogeny, ecology, and gene families covariation shaped the olfactory subgenome of rodents. *Genome Biology and Evolution* 15: evad197. https://doi.org/10.1093/gbe/evad197

Crandall, K. A., Bininda-Emonds, O. R., Mace, G. M., & Wayne, R. K. (2000). Considering evolutionary processes in conservation biology. *Trends in Ecology & Evolution* 15:290–295. https://doi.org/10.1016/s0169-5347(00)01876-0

Cridland, J. M., & Begun, D. J. (2023). Male-derived transcripts isolated from the mated female reproductive tract in *Drosophila melanogaster*. *G3* 13: jkad202. https://doi.org/10.1093/g3journal/jkad202

Crnokrak, P., & Roff, D. A. (1999). Inbreeding depression in the wild. *Heredity* 83:260–270. https://doi.org/10.1038/sj.hdy.6885530

Curik, I., Ferenčaković, M., & Sölkner, J. (2014). Inbreeding and runs of homozygosity: A possible solution to an old problem. *Livestock Science 166*:26–34. https://doi.org/10.1016/j.livsci.2014.05.034

Cypher, B. L., Phillips, S. E., Westall, T. L., Tennant, E. N., Saslaw, L. R., Kelly, E. C., & Job,

C. L. V. H. (2017). Conservation of endangered Tipton kangaroo rats (*Dipodomys nitratoides nitratoides*): status surveys, habitat suitability, and conservation strategies. Endangered Species Recovery Program, Turlock, CA. https://mail.esrp.org/publications/pdf/Cypher_etal_2017_TKR_Conservation.pdf

Cypher, B. L., Phillips, S. E., Westall, T. L., Tennant, E. N., Saslaw, L. R., Kelly, E. C., & VanHorn Job, C. L. (2021). Conservation of endangered Tipton kangaroo rats (*Dipodomys nitratoides nitratoides)*: status surveys, habitat suitability, and conservation recommendations. *California Fish and Wildlife Journal*, *CESA Special Issue*, 382–397. https://doi.org/10.51492/cfwj.cesasi.23

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., Durbin, R., & 1000 Genomes Project Analysis Group. (2011). The variant call format and VCFtools. *Bioinformatics* 27:2156–2158. https://doi.org/10.1093/bioinformatics/btr330

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience* 10. https://doi.org/10.1093/gigascience/giab008

Dapper, A. L., & Wade, M. J. (2020). Relaxed selection and the rapid evolution of reproductive genes. *Trends in Genetics* 36:640–649. https://doi.org/10.1016/j.tig.2020.06.014

Davis, B. L., Williams, S. L., & Lopez, G. (1971). Chromosomal studies of *Geomys*. *Journal of Mammalogy* 52:617–620. https://doi.org/10.2307/1378601

de Flamingh, A., Rivera-Colón, A. G., Gnoske, T. P., Kerbis Peterhans, J. C., Catchen, J., Malhi, R. S., & Roca, A. L. (2023). Numt Parser: Automated identification and removal of nuclear mitochondrial pseudogenes (numts) for accurate mitochondrial genome reconstruction in *Panthera*. *The Journal of Heredity* 114:120–130. https://doi.org/10.1093/jhered/esac065

Dean, M. D. (2013). Genetic disruption of the copulatory plug in mice leads to severely reduced fertility. *PLoS Genetics*, *9*(1), e1003185. https://doi.org/10.1371/journal.pgen.1003185

Dean, M. D., Clark, N. L., Findlay, G. D., Karn, R. C., Yi, X., Swanson, W. J., MacCoss, M. J., & Nachman, M. W. (2009). Proteomics and comparative genomic investigations reveal heterogeneity in evolutionary rate of male reproductive proteins in mice (*Mus domesticus*). *Molecular Biology and Evolution* 26:1733–1743. https://doi.org/10.1093/molbev/msp094

Dean, M. D., Findlay, G. D., Hoopmann, M. R., Wu, C. C., MacCoss, M. J., Swanson, W. J., & Nachman, M. W. (2011). Identification of ejaculated proteins in the house mouse (*Mus domesticus*) via isotopic labeling. *BMC Genomics* 12:306. https://doi.org/10.1186/1471-2164-12-306

Dean, M. D., Good, J. M., & Nachman, M. W. (2008). Adaptive evolution of proteins secreted during sperm maturation: an analysis of the mouse epididymal transcriptome. *Molecular Biology*

and Evolution 25:83–392. https://doi.org/10.1093/molbev/msm265

Deaven, L. L., Vidal-Rioja, L., Jett, J. H., & Hsu, T. C. (1977). Chromosomes of *Peromyscus* (Rodentia, Cricetidae). VI. The genomic size. *Cytogenetics and Cell Genetics*, 19:241–249. https://doi.org/10.1159/000130816

DeRaad, D. A., McCullough, J. M., DeCicco, L. H., Hime, P. M., Joseph, L., Andersen, M. J., & Moyle, R. G. (2023). Mitonuclear discordance results from incomplete lineage sorting, with no detectable evidence for gene flow, in a rapid radiation of *Todiramphus* kingfishers. *Molecular Ecology* 32:4844–4862. https://doi.org/10.1111/mec.17080

Dewey, M. J., & Dawson, W. D. (2001). Deer mice: "The *Drosophila* of North American mammalogy." *Genesis* 29:105–109. https://doi.org/10.1002/gene.1011

DeWoody, J. A., Harder, A. M., Mathur, S., & Willoughby, J. R. (2021). The long-standing significance of genetic diversity in conservation. *Molecular Ecology* 30:4147–4154. https://doi.org/10.1111/mec.16051

Dipodomys ordii genome assembly Dord_2.0. (n.d.). NCBI. Retrieved June 12, 2024, from https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000151885.1/

Dipodomys stephensi genome assembly DipSte_v1_BIUU. (n.d.). NCBI. Retrieved June 12, 2024, from https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_004024685.1/

Dixson, A. L., & Anderson, M. J. (2002). Sexual selection, seminal coagulation and copulatory plug formation in primates. *Folia Primatologica; International Journal of Primatology* 73:63–69. https://doi.org/10.1159/000064784

Dorus, S., Evans, P. D., Wyckoff, G. J., Choi, S. S., & Lahn, B. T. (2004). Rate of molecular evolution of the seminal protein gene SEMG2 correlates with levels of female promiscuity. *Nature Genetics* 36:1326–1329. https://doi.org/10.1038/ng1471

Dussex, N., Morales, H. E., Grossen, C., Dalén, L., & van Oosterhout, C. (2023). Purging and accumulation of genetic load in conservation. *Trends in Ecology & Evolution* 38:961–969. https://doi.org/10.1016/j.tree.2023.05.008

Eberhard, W. (1996). *Female Control: Sexual Selection by Cryptic Female Choice*. Princeton University Press, Princeton, USA.

Eizenga, J. M., Novak, A. M., Sibbesen, J. A., Heumos, S., Ghaffaari, A., Hickey, G., Chang, X., Seaman, J. D., Rounthwaite, R., Ebler, J., Rautiainen, M., Garg, S., Paten, B., Marschall, T., Sirén, J., & Garrison, E. (2020). Pangenome graphs. *Annual Review of Genomics and Human Genetics* 21:139–162. https://doi.org/10.1146/annurev-genom-120219-080406

Ellinghaus, D., Kurtz, S., & Willhoeft, U. (2008). LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics* 9:18.

https://doi.org/10.1186/1471-2105-9-18

Emilsson, V., Thorleifsson, G., Zhang, B., Leonardson, A. S., Zink, F., Zhu, J., Carlson, S., Helgason, A., Walters, G. B., Gunnarsdottir, S., Mouy, M., Steinthorsdottir, V., Eiriksdottir, G. H., Bjornsdottir, G., Reynisdottir, I., Gudbjartsson, D., Helgadottir, A., Jonasdottir, A., Jonasdottir, A., … Stefansson, K. (2008). Genetics of gene expression and its effect on disease. *Nature* 452:423–428. https://doi.org/10.1038/nature06758

Emms, D. M., & Kelly, S. (2015). OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biology* 16:157. https://doi.org/10.1186/s13059-015-0721-2

Enard, W., Khaitovich, P., Klose, J., Zöllner, S., Heissig, F., Giavalisco, P., Nieselt-Struwe, K., Muchmore, E., Varki, A., Ravid, R., Doxiadis, G. M., Bontrop, R. E., & Pääbo, S. (2002). Intra- and interspecific variation in primate gene expression patterns. *Science* 296:340–343. https://doi.org/10.1126/science.1068996

Eyre-Walker, A., & Keightley, P. D. (2007). The distribution of fitness effects of new mutations. *Nature Reviews. Genetics* 8:610–618. https://doi.org/10.1038/nrg2146

Fang, X., Nevo, E., Han, L., Levanon, E. Y., Zhao, J., Avivi, A., Larkin, D., Jiang, X., Feranchuk, S., Zhu, Y., Fishman, A., Feng, Y., Sher, N., Xiong, Z., Hankeln, T., Huang, Z., Gorbunova, V., Zhang, L., Zhao, W., … Wang, J. (2014). Genome-wide adaptive complexes to underground stresses in blind mole rats *Spalax*. *Nature Communications 5*:3966. https://doi.org/10.1038/ncomms4966

Fassler, D., & Leavitt, R. D. (1975). Terrestrial activity of the northern pocket gopher (Geomyidae) as indicated by owl predation. *The Southwestern Naturalist* 19:452-453. https://doi.org/10.2307/3670410

Firman, R. C. (2018). Postmating sexual conflict and female control over fertilization during gamete interaction. *Annals of the New York Academy of Sciences* 1422:48–64. https://doi.org/10.1111/nyas.13635

Firman, R. C., Gasparini, C., Manier, M. K., & Pizzari, T. (2017). Postmating female control: 20 Years of cryptic female choice. *Trends in Ecology & Evolution* 32:368–382. https://doi.org/10.1016/j.tree.2017.02.010

Firman, R. C., Gomendio, M., Roldan, E. R. S., & Simmons, L. W. (2014). The coevolution of ova defensiveness with sperm competitiveness in house mice. *The American Naturalist* 183:565–572. https://doi.org/10.1086/675395

Firman, R. C., & Simmons, L. W. (2011). Experimental evolution of sperm competitiveness in a mammal. *BMC Evolutionary Biology* 11:19. https://doi.org/10.1186/1471-2148-11-19

Fisher, H. S., Jacobs-Palmer, E., Lassance, J.-M., & Hoekstra, H. E. (2016). The genetic basis

and fitness consequences of sperm midpiece size in deer mice. *Nature Communications* 7:13652. https://doi.org/10.1038/ncomms13652

Flynn, J. M., Hubley, R., Goubert, C., Rosen, J., Clark, A. G., Feschotte, C., & Smit, A. F. (2020). RepeatModeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences USA* 117:9451–9457. https://doi.org/10.1073/pnas.1921046117

Formenti, G., Theissinger, K., Fernandes, C., Bista, I., Bombarely, A., Bleidorn, C., Ciofi, C., Crottini, A., Godoy, J. A., Höglund, J., Malukiewicz, J., Mouton, A., Oomen, R. A., Paez, S., Palsbøll, P. J., Pampoulie, C., Ruiz-López, M. J., Svardal, H., Theofanopoulou, C., … European Reference Genome Atlas (ERGA) Consortium. (2022). The era of reference genomes in conservation genomics. *Trends in Ecology & Evolution* 37:197–202. https://doi.org/10.1016/j.tree.2021.11.008

Fu, L., Niu, B., Zhu, Z., Wu, S., & Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 28:3150–3152. https://doi.org/10.1093/bioinformatics/bts565

Gahlay, G. K., & Rajput, N. (2020). The enigmatic sperm proteins in mammalian fertilization: an overview. *Biology of Reproduction* 103:1171–1185. https://doi.org/10.1093/biolre/ioaa140

Galindo, B. E., Vacquier, V. D., & Swanson, W. J. (2003). Positive selection in the egg receptor for abalone sperm lysin. *Proceedings of the National Academy of Sciences USA* 100:4639–4643. https://doi.org/10.1073/pnas.0830022100

Germano, D. J., Rathbun, G. B., Saslaw, L. R., Cypher, B. L., Cypher, E. A., & Vredenburgh, L. M. (2011). The San Joaquin Desert of California: Ecologically misunderstood and overlooked. *Natural Areas Journal* 31:138–147. https://doi.org/10.3375/043.031.0206

Germano, D. J., Saslaw, L. R., Smith, P. T., & Cypher, B. L. (2013). Survivorship and reproduction of translocated Tipton kangaroo rats in the San Joaquin Valley, California. *Endangered Species Research* 19:265–276. https://doi.org/10.3354/esr00470

Ghurye, J., Pop, M., Koren, S., Bickhart, D., & Chin, C.-S. (2017). Scaffolding of long read assemblies using long range contact information. *BMC Genomics*, 18:527. https://doi.org/10.1186/s12864-017-3879-z

Ghurye, J., Rhie, A., Walenz, B. P., Schmitt, A., Selvaraj, S., Pop, M., Phillippy, A. M., & Koren, S. (2019). Integrating Hi-C links with assembly graphs for chromosome-scale assembly. *PLoS Computational Biology* 15:e1007273. https://doi.org/10.1371/journal.pcbi.1007273

Gomendio, M., & Roldan, E. R. S. (2008). Implications of diversity in sperm size and function for sperm competition and fertility. *The International Journal of Developmental Biology* 52:439–447. https://doi.org/10.1387/ijdb.082595mg

Goldingay, R. L., A. Kelly, P., & F. Williams, D. (1997). The Kangaroo Rats of California: endemism and conservation of keystone species. *Pacific Conservation Biology: A Journal Devoted to Conservation and Land Management in the Pacific Region* 3:47-60. https://doi.org/10.1071/pc970047

Good, J. M., & Nachman, M. W. (2005). Rates of protein evolution are positively correlated with developmental timing of expression during mouse spermatogenesis. *Molecular Biology and Evolution* 22:1044–1052. https://doi.org/10.1093/molbev/msi087

Gozashti, L., Feschotte, C., & Hoekstra, H. E. (2023). Transposable element interactions shape the ecology of the deer mouse genome. *Molecular Biology and Evolution* 40: msad069. https://doi.org/10.1093/molbev/msad069

Gozashti, L., Harringmeyer, O. S., & Hoekstra, H. E. (2025). How repeats rearrange chromosomes in deer mice. *Cell Reports* 44:115644. https://doi.org/10.1016/j.celrep.2025.115644

Greenbaum, I. F., Honeycutt, R. L., & Chirhart, S. E. (2019). Taxonomy and phylogenetics of the *Peromyscus maniculatus* species group. https://webapps.fhsu.edu/ksmammal/bibFiles/554.pdf

Grinnell, J. (1920). A new kangaroo rat from the San Joaquin valley, California. *Journal of Mammalogy* 1:178-179. https://doi.org/10.2307/1373309

Grinnell, J. (1927). Geography and evolution in the pocket gophers of California. *Annual Report of the Board of Regents of the Smithsonian Institution. Smithsonian Institution. Board of Regents*, 343.

Grinnell, J., & Hill, J. E. (1936). Pocket gophers (*Thomomys*) of the lower Colorado valley. *Journal of Mammalogy* 17:1-10. https://doi.org/10.2307/1374540

Gubernick, D. J. (1988). Reproduction in the California mouse, *Peromyscus californicus*. *Journal of Mammalogy* 69:857–860. https://doi.org/10.2307/1381649

Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. *Bioinformatics* 29:1072–1075. https://doi.org/10.1093/bioinformatics/btt086

Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., Couger, M. B., Eccles, D., Li, B., Lieber, M., MacManes, M. D., Ott, M., Orvis, J., Pochet, N., Strozzi, F., Weeks, N., Westerman, R., William, T., Dewey, C. N., … Regev, A. (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols* 8:1494–1512. https://doi.org/10.1038/nprot.2013.084

Hager, E. R., Harringmeyer, O. S., Wooldridge, T. B., Theingi, S., Gable, J. T., McFadden, S., Neugeboren, B., Turner, K. M., Jensen, J. D., & Hoekstra, H. E. (2022). A chromosomal inversion contributes to divergence in multiple traits between deer mouse ecotypes. *Science*

377:399–405. https://doi.org/10.1126/science.abg0718

Halsey, M. K., Roberts, E. K., Wright, E. A., Soniat, T. J., Lindsay, L. L., Moreno-Santillan, D., Pitts, R. M., Dávalos, L. M., Bradley, R. D., Stevens, R. D., & Ray, D. A. (2024). Newly assembled pocket gopher genomes can facilitate conservation management of biodiversity. *Journal of Mammalogy* gyae138. https://doi.org/10.1093/jmammal/gyae138

Harcourt, A. H., Purvis, A., & Liles, L. (1995). Sperm competition: mating system, not breeding season, affects testes size of primates. *Functional Ecology* 9:468–476. https://doi.org/10.2307/2390011

Harder, A. M., Walden, K. K. O., Marra, N. J., & Willoughby, J. R. (2022). High- quality reference genome for an arid-adapted mammal, the banner-tailed kangaroo rat (*Dipodomys spectabilis*). *Genome Biology and Evolution* 14. https://doi.org/10.1093/gbe/evac005

Harris, S. E., & Munshi-South, J. (2017). Signatures of positive selection and local adaptation to urbanization in white-footed mice (*Peromyscus leucopus*). *Molecular Ecology* 26:6336–6350. https://doi.org/10.1111/mec.14369

Harshman, L. G., & Prout, T. (1994). Sperm displacement without sperm transfer in *Drosophila melanogaster*. *Evolution* 48:58–766. https://doi.org/10.1111/j.1558-5646.1994.tb01359.x

Haug-Baltzell, A., Stephens, S. A., Davey, S., Scheidegger, C. E., & Lyons, E. (2017). SynMap2 and SynMap3D: web-based whole-genome synteny browsers. *Bioinformatics* 33:2197–2198. https://doi.org/10.1093/bioinformatics/btx144

Hedrick, P. W. (1992). Genetic conservation in captive populations and endangered species. In *Applied Population Biology* (pp. 45–68). Springer Netherlands. https://doi.org/10.1007/978-0-585-32911-6_3

Hedrick, P. W., & Garcia-Dorado, A. (2016). Understanding inbreeding depression, purging, and genetic rescue. *Trends in Ecology & Evolution* 31:940–952. https://doi.org/10.1016/j.tree.2016.09.005

Hedrick, P. W., Peterson, R. O., Vucetich, L. M., Adams, J. R., & Vucetich, J. A. (2014). Genetic rescue in Isle Royale wolves: genetic analysis and the collapse of the population. *Conservation Genetics* 15:1111–1121. https://doi.org/10.1007/s10592-014-0604-1

Heifetz, Y., Lung, O., Frongillo, E. A., Jr, & Wolfner, M. F. (2000). The *Drosophila* seminal fluid protein Acp26Aa stimulates release of oocytes by the ovary. *Current Biology* 10:99–102. https://doi.org/10.1016/s0960-9822(00)00288-8

Hendricks, S., Navarro, A. Y., Wang, T., Wilder, A., Ryder, O. A., & Shier, D. M. (2020). Patterns of genetic partitioning and gene flow in the endangered San Bernardino kangaroo rat (*Dipodomys merriami parvus*) and implications for conservation management. *Conservation Genetics* 21:819–833. https://doi.org/10.1007/s10592-020-01289-z

Hermann, B. P., Cheng, K., Singh, A., Roa-De La Cruz, L., Mutoji, K. N., Chen, I.-C., Gildersleeve, H., Lehle, J. D., Mayo, M., Westernströer, B., Law, N. C., Oatley, M. J., Velte, E. K., Niedenberger, B. A., Fritze, D., Silber, S., Geyer, C. B., Oatley, J. M., & McCarrey, J. R. (2018). The mammalian spermatogenesis single-cell transcriptome, from spermatogonial stem cells to spermatids. *Cell Reports* 25:1650-1667.e8. https://doi.org/10.1016/j.celrep.2018.10.026

Herndon, L. A., & Wolfner, M. F. (1995). A *Drosophila* seminal fluid protein, Acp26Aa, stimulates egg laying in females for 1 day after mating. *Proceedings of the National Academy of Sciences USA* 92:10114–10118. https://doi.org/10.1073/pnas.92.22.10114

Heuertz, M., Carvalho, S. B., Galindo, J., Rinkevich, B., Robakowski, P., Aavik, T., Altinok, I., Barth, J. M. I., Cotrim, H., Goessen, R., González-Martínez, S. C., Grebenc, T., Hoban, S., Kopatz, A., McMahon, B. J., Porth, I., Raeymaekers, J. A. M., Träger, S., Valdecantos, A., … Garnier-Géré, P. (2023). The application gap: Genomics for biodiversity and ecosystem service management. *Biological Conservation* 278:109883. https://doi.org/10.1016/j.biocon.2022.109883

Hilgers, L., Liu, S., Jensen, A., Brown, T., Cousins, T., Schweiger, R., Guschanski, K., & Hiller, M. (2025). Avoidable false PSMC population size peaks occur across numerous studies. *Current Biology* 35: p927-930.e3. 10.1016/j.cub.2024.09.028

Hoban, S., Bruford, M., D'Urban Jackson, J., Lopes-Fernandes, M., Heuertz, M., Hohenlohe, P. A., Paz-Vinas, I., Sjögren-Gulve, P., Segelbacher, G., Vernesi, C., Aitken, S., Bertola, L. D., Bloomer, P., Breed, M., Rodríguez-Correa, H., Funk, W. C., Grueber, C. E., Hunter, M. E., Jaffe, R., … Laikre, L. (2020). Genetic diversity targets and indicators in the CBD post-2020 Global Biodiversity Framework must be improved. *Biological Conservation* 248:108654. https://doi.org/10.1016/j.biocon.2020.108654

Hoban, S., Bruford, M. W., Funk, W. C., Galbusera, P., Griffith, M. P., Grueber, C. E., Heuertz, M., Hunter, M. E., Hvilsom, C., Stroil, B. K., Kershaw, F., Khoury, C. K., Laikre, L., Lopes-Fernandes, M., MacDonald, A. J., Mergeay, J., Meek, M., Mittan, C., Mukassabi, T. A., … Vernesi, C. (2021). Global commitments to conserving and monitoring genetic diversity are now necessary and feasible. *Bioscience* 71:964–976. https://doi.org/10.1093/biosci/biab054

Hoffman, J. I., Vendrami, D. L. J., Hench, K., Chen, R. S., Stoffel, M. A., Kardos, M., Amos, W., Kalinowski, J., Rickert, D., Köhrer, K., Wachtmeister, T., Goebel, M. E., Bonin, C. A., Gulland, F. M. D., & Dasmahapatra, K. K. (2024). Genomic and fitness consequences of a near-extinction event in the northern elephant seal. *Nature Ecology & Evolution* 8:2309–2324. https://doi.org/10.1038/s41559-024-02533-2

Hoffmann, A. A., Sgrò, C. M., & Kristensen, T. N. (2017). Revisiting adaptive potential, population size, and conservation. *Trends in Ecology & Evolution* 32:506–517. https://doi.org/10.1016/j.tree.2017.03.012

Hohenlohe, P. A., Funk, W. C., & Rajora, O. P. (2021). Population genomics for wildlife

conservation and management. *Molecular Ecology* 30:62–82. https://doi.org/10.1111/mec.15720

Holderegger, R., & Wagner, H. H. (2008). Landscape genetics. *BioScience* 58:199–207. https://doi.org/10.1641/B580306

Hook, K. A., Liu, C., Joyner, K. A., Duncan, G. A., & Fisher, H. S. (2022). Female reproductive fluid composition differs based on mating system in Peromyscus mice. In *bioRxiv* (p. 2022.03.18.484937). https://doi.org/10.1101/2022.03.18.484937

Hook, K. A., Wilke, L. M., & Fisher, H. S. (2021). Apical sperm hook morphology is linked to sperm swimming performance and sperm aggregation in *Peromyscus* mice. *Cells* 10. https://doi.org/10.3390/cells10092279

Hubley, R., Finn, R. D., Clements, J., Eddy, S. R., Jones, T. A., Bao, W., Smit, A. F. A., & Wheeler, T. J. (2016). The Dfam database of repetitive DNA families. *Nucleic Acids Research*, 44:D81-9. https://doi.org/10.1093/nar/gkv1272

Ingles, L. G. (1950). Pigmental variations in populations of pocket gophers. *Evolution* 4:353–357. https://doi.org/10.2307/2405602

IUCN. (2012). Red List Criteria Version 3.1 Second edition. *International Union for Conservation of Nature and Natural Resources*.

IUCN. (2024) The IUCN Red List of Threatened Species. Version 2024-1. https://www.iucnredlist.org. Downloaded on July 10th 2024.

Jackson, D. C., & Schmidt-Nielsen, K. (1964). Countercurrent heat exchange in the respiratory passages. *Proceedings of the National Academy of Sciences* 51:1192–1197. https://doi.org/10.1073/pnas.51.6.1192

Janes, S. W., & Barss, J. M. (1985). Predation by three owl species on northern pocket gophers of different body mass. *Oecologia*, 67:76–81. https://doi.org/10.1007/BF00378454

Jiang, M., Shi, L., Li, X., Dong, Q., Sun, H., Du, Y., Zhang, Y., Shao, T., Cheng, H., Chen, W., & Wang, Z. (2020). Genome-wide adaptive evolution to underground stresses in subterranean mammals: Hypoxia adaption, immunity promotion, and sensory specialization. *Ecology and Evolution* 10:377–7388. https://doi.org/10.1002/ece3.6462

Johnson, W. E., Onorato, D. P., Roelke, M. E., Land, E. D., Cunningham, M., Belden, R. C., McBride, R., Jansen, D., Lotz, M., Shindle, D., Howard, J., Wildt, D. E., Penfold, L. M., Hostetler, J. A., Oli, M. K., & O'Brien, S. J. (2010). Genetic restoration of the Florida panther. *Science* 329:1641–1645. https://doi.org/10.1126/science.1192891

Jones, C. A., & Baxter, C. N. (2004). *Thomomys bottae*. *Mammalian Species* 742:1–14. https://doi.org/10.1644/742

Joyner, C. P., Myrick, L. C., Crossland, J. P., & Dawson, W. D. (1998). Deer mice as laboratory animals. *ILAR Journal* 39:322–330. https://doi.org/10.1093/ilar.39.4.322

Jurka, J., Kapitonov, V. V., Pavlicek, A., Klonowski, P., Kohany, O., & Walichiewicz, J. (2005). Repbase Update, a database of eukaryotic repetitive elements. *Cytogenetic and Genome Research* 110:462–467. https://doi.org/10.1159/000084979

Katoh, K., & Standley, D. (2013). MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution* 30:772–780. https://doi.org/10.1093/molbev/mst010

Kawano, N., Araki, N., Yoshida, K., Hibino, T., Ohnami, N., Makino, M., Kanai, S., Hasuwa, H., Yoshida, M., Miyado, K., & Umezawa, A. (2014). Seminal vesicle protein SVS2 is required for sperm survival in the uterus. *Proceedings of the National Academy of Sciences USA* 111:4145–4150. https://doi.org/10.1073/pnas.1320715111

Kawano, N., & Yoshida, M. (2007). Semen-coagulating protein, SVS2, in mouse seminal plasma controls sperm fertility. *Biology of Reproduction* 76:353–361. https://doi.org/10.1095/biolreprod.106.056887

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., Thierer, T., Ashton, B., Meintjes, P., & Drummond, A. (2012). Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* 28:1647–1649. https://doi.org/10.1093/bioinformatics/bts199

Keightley, P. D., & 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. https://doi.org/10.1534/genetics.107.080663

Kenagy, G. J. (1973). Daily and seasonal patterns of activity and energetics in a heteromyid rodent community. *Ecology* 54:1201–1219. https://doi.org/10.2307/1934184

Kenkel, C. D., & Matz, M. V. (2016). Gene expression plasticity as a mechanism of coral adaptation to a variable environment. *Nature Ecology & Evolution* 1:14. https://doi.org/10.1038/s41559-016-0014

Kerpedjiev, P., Abdennur, N., Lekschas, F., McCallum, C., Dinkla, K., Strobelt, H., Luber, J. M., Ouellette, S. B., Azhir, A., Kumar, N., Hwang, J., Lee, S., Alver, B. H., Pfister, H., Mirny, L. A., Park, P. J., & Gehlenborg, N. (2018). HiGlass: web-based visual exploration and analysis of genome interaction maps. *Genome Biology* 19:125. https://doi.org/10.1186/s13059-018-1486-1

Kiełbasa, S. M., Wan, R., Sato, K., Horton, P., & Frith, M. C. (2011). Adaptive seeds tame genomic sequence comparison. *Genome Research* 21: 487–493. https://doi.org/10.1101/gr.113985.110

Kim, J., Lee, C., Ko, B. J., Yoo, D. A., Won, S., Phillippy, A. M., Fedrigo, O., Zhang, G., Howe, K., Wood, J., Durbin, R., Formenti, G., Brown, S., Cantin, L., Mello, C. V., Cho, S., Rhie, A., Kim, H., & Jarvis, E. D. (2022). False gene and chromosome losses in genome assemblies caused by GC content variation and repeats. *Genome Biology* 23:204. https://doi.org/10.1186/s13059-022-02765-0

Kimura, M. (1985). *The neutral theory of molecular evolution*. Cambridge University Press. Cambridge, United Kingdom.

Kimura, M., Maruyama, T., & Crow, J. F. (1963). The mutation load in small populations. *Genetics* 48:1303–1312. https://doi.org/10.1093/genetics/48.10.1303

King, M. C., & Wilson, A. C. (1975). Evolution at two levels in humans and chimpanzees. *Science* 188:107–116. https://doi.org/10.1126/science.1090005

Kleinman-Ruiz, D., Lucena-Perez, M., Villanueva, B., Fernández, J., Saveljev, A. P., Ratkiewicz, M., Schmidt, K., Galtier, N., García-Dorado, A., & Godoy, J. A. (2022). Purging of deleterious burden in the endangered Iberian lynx. *Proceedings of the National Academy of Sciences USA* 119:e2110614119. https://doi.org/10.1073/pnas.2110614119

Kopania, E. E. K., Larson, E. L., Callahan, C., Keeble, S., & Good, J. M. (2022). Molecular evolution across mouse spermatogenesis. *Molecular Biology and Evolution* 39: msac023. https://doi.org/10.1093/molbev/msac023

Kordonowy, L., & MacManes, M. (2017). Characterizing the reproductive transcriptomic correlates of acute dehydration in males in the desert-adapted rodent, *Peromyscus eremicus*. *BMC Genomics* 18:473. https://doi.org/10.1186/s12864-017-3840-1

Korlach, J., Gedman, G., Kingan, S. B., Chin, C.-S., Howard, J. T., Audet, J.-N., Cantin, L., & Jarvis, E. D. (2017). De novo PacBio long-read and phased avian genome assemblies correct and add to reference genes generated with intermediate and short reads. *GigaScience* 6:1–16. https://doi.org/10.1093/gigascience/gix085

Kozak, K. M., Escalona, M., Chumchim, N., Fairbairn, C., Marimuthu, M. P. A., Nguyen, O., Sahasrabudhe, R., Seligmann, W., Conroy, C., Patton, J. L., Bowie, R. C. K., & Nachman, M. W. (2024). A highly contiguous genome assembly for the pocket mouse *Perognathus longimembris longimembris*. *The Journal of Heredity* 115:130–138. https://doi.org/10.1093/jhered/esad060

Kozlov, A. M., Darriba, D., Flouri, T., Morel, B., & Stamatakis, A. (2019). RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics* 35:4453–4455. https://doi.org/10.1093/bioinformatics/btz305

Kyriazis, C. C., Wayne, R. K., & Lohmueller, K. E. (2021). Strongly deleterious mutations are a primary determinant of extinction risk due to inbreeding depression. *Evolution Letters* 5:33–47. https://doi.org/10.1002/evl3.209

Lande, R. (1994). Risk of population extinction from fixation of new deleterious mutations. *Evolution* 48:1460–1469. https://doi.org/10.1111/j.1558-5646.1994.tb02188.x

Lanfear, R., Kokko, H., & Eyre-Walker, A. (2014). Population size and the rate of evolution. *Trends in Ecology & Evolution* 29:33–41. https://doi.org/10.1016/j.tree.2013.09.009

Langfelder, P., & Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 9:559. https://doi.org/10.1186/1471-2105-9-559

Leffler, E. M., Bullaughey, K., Matute, D. R., Meyer, W. K., Ségurel, L., Venkat, A., Andolfatto, P., & Przeworski, M. (2012). Revisiting an old riddle: what determines genetic diversity levels within species? *PLoS Biology* 10:e1001388. https://doi.org/10.1371/journal.pbio.1001388

Leroy, T., Rousselle, M., Tilak, M.-K., Caizergues, A. E., Scornavacca, C., Recuerda, M., Fuchs, J., Illera, J. C., De Swardt, D. H., Blanco, G., Thébaud, C., Milá, B., & Nabholz, B. (2021). Island songbirds as windows into evolution in small populations. *Current Biology* 31:1303-1310.e4. https://doi.org/10.1016/j.cub.2020.12.040

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. In *arXiv [q-bio.GN]*. arXiv. http://arxiv.org/abs/1303.3997

Li, H., & Durbin, R. (2011). Inference of human population history from individual whole-genome sequences. *Nature* 475:493–496. https://doi.org/10.1038/nature10231

Li, H., Xiang-Yu, J., Dai, G., Gu, Z., Ming, C., Yang, Z., Ryder, O. A., Li, W.-H., Fu, Y.-X., & Zhang, Y.-P. (2016). Large numbers of vertebrates began rapid population decline in the late 19th century. *Proceedings of the National Academy of Sciences USA* 113:14079–14084. https://doi.org/10.1073/pnas.1616804113

Li, K., Zhang, S., Song, X., Weyrich, A., Wang, Y., Liu, X., Wan, N., Liu, J., Lövy, M., Cui, H., Frenkel, V., Titievsky, A., Panov, J., Brodsky, L., & Nevo, E. (2020). Genome evolution of blind subterranean mole rats: Adaptive peripatric versus sympatric speciation. *Proceedings of the National Academy of Sciences USA* 117:32499–32508. https://doi.org/10.1073/pnas.2018123117

Li, W., & Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22:658–1659. https://doi.org/10.1093/bioinformatics/btl158

Linnen, C. R., Poh, Y.-P., Peterson, B. K., Barrett, R. D. H., Larson, J. G., Jensen, J. D., & Hoekstra, H. E. (2013). Adaptive evolution of multiple traits through multiple mutations at a single gene. *Science* 339:1312–1316. https://doi.org/10.1126/science.1233213

Linzey, A., & Layne, J. (1969). Comparative morphology of the male reproductive tract in the rodent genus *Peromyscus* (Muridae). *American Museum Novitiates* 2355:1–47.

Loew, S. S., Williams, D. F., Ralls, K., Pilgrim, K., & Fleischer, R. C. (2005). Population structure and genetic variation in the endangered Giant Kangaroo Rat (*Dipodomys ingens*). *Conservation Genetics* 6:495–510. https://doi.org/10.1007/s10592-005-9005-9

Logsdon, G. A., Vollger, M. R., & Eichler, E. (2020). Long-read human genome sequencing and its applications. *Nature Reviews Genetics* 21:597–614. https://doi.org/10.1038/s41576-020-0236-x

Lok, S., Paton, T. A., Wang, Z., Kaur, G., Walker, S., Yuen, R. K. C., Sung, W. W. L., Whitney, J., Buchanan, J. A., Trost, B., Singh, N., Apresto, B., Chen, N., Coole, M., Dawson, T. J., Ho, K., Hu, Z., Pullenayegum, S., Samler, K., … Scherer, S. W. (2017). De novo genome and transcriptome assembly of the Canadian beaver (*Castor canadensis*). *G3* 7:755–773. https://doi.org/10.1534/g3.116.038208

Long, A. D., Baldwin-Brown, J., Tao, Y., Cook, V. J., Balderrama-Gutierrez, G., Corbett-Detig, R., Mortazavi, A., & Barbour, A. G. (2019). The genome of *Peromyscus leucopus*, natural host for Lyme disease and other emerging infections. *Science Advances* 5:eaaw6441. https://doi.org/10.1126/sciadv.aaw6441

Longland, W. S., & Dimitri, L. A. (2021). Kangaroo rats: Ecosystem engineers on western rangelands. *Rangelands* 43:72–80. https://doi.org/10.1016/j.rala.2020.10.004

Lortie, C. J., Filazzola, A., Kelsey, R., Hart, A. K., & Butterfield, H. S. (2018). Better late than never: a synthesis of strategic land retirement and restoration in California. *Ecosphere* 9:e02367. https://doi.org/10.1002/ecs2.2367

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* 15:550. https://doi.org/10.1186/s13059-014-0550-8

Lung, O., & Wolfner, M. F. (2001). Identification and characterization of the major *Drosophila melanogaster* mating plug protein. *Insect Biochemistry and Molecular Biology* 31:543–551. https://doi.org/10.1016/s0965-1748(00)00154-5

Lüpold, S., de Boer, R. A., Evans, J. P., Tomkins, J. L., & Fitzpatrick, J. L. (2020). How sperm competition shapes the evolution of testes and sperm: a meta-analysis. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 375:2020.0064. https://doi.org/10.1098/rstb.2020.0064

Lynch, M., Conery, J., & Burger, R. (1995). Mutation accumulation and the extinction of small populations. *The American Naturalist* 146:489–518. https://doi.org/10.1086/285812

Lyons, E., & Freeling, M. (2008). How to usefully compare homologous plant genes and chromosomes as DNA sequences: how to usefully compare plant genomes. *The Plant Journal: For Cell and Molecular Biology* 53:661–673. https://doi.org/10.1111/j.1365-313X.2007.03326.x

Mack, K. L., Ballinger, M. A., Phifer-Rixey, M., & Nachman, M. (2018). Gene regulation underlies environmental adaptation in house mice. *Genome Research* 28:1636–1645. https://doi.org/10.1101/gr.238998.118

Mackay, T. F. C., Stone, E. A., & Ayroles, J. F. (2009). The genetics of quantitative traits: challenges and prospects. *Nature Reviews Genetics* 10:565–577. https://doi.org/10.1038/nrg2612

MacManes, M. D., & Eisen, M. B. (2014). Characterization of the transcriptome, nucleotide sequence polymorphism, and natural selection in the desert adapted mouse *Peromyscus eremicus. PeerJ*, 2:e642. https://doi.org/10.7717/peerj.642

MacMillen, R. E., & Hinds, D. S. (1983). Water regulatory efficiency in heteromyid rodents: a model and its application. *Ecology* 64:152–164. https://doi.org/10.2307/1937337

Majane, A. C., Cridland, J. M., & Begun, D. J. (2022). Single-nucleus transcriptomes reveal evolutionary and functional properties of cell types in the *Drosophila* accessory gland. *Genetics* 220: iyab213. https://doi.org/10.1093/genetics/iyab213

Manel, S., Schwartz, M. K., Luikart, G., & Taberlet, P. (2003). Landscape genetics: combining landscape ecology and population genetics. *Trends in Ecology & Evolution* 18:189–197. https://doi.org/10.1016/s0169-5347(03)00008-9

Mangels, R., Tsung, K., Kwan, K., & Dean, M. D. (2016). Copulatory plugs inhibit the reproductive success of rival males. *Journal of Evolutionary Biology* 29:2289–2296. https://doi.org/10.1111/jeb.12956

Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A., & Zdobnov, E. M. (2021). BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Molecular Biology and Evolution* 38: 4647–4654. https://doi.org/10.1093/molbev/msab199

Manni, M., Berkeley, M. R., Seppey, M., & Zdobnov, E. M. (2021). BUSCO: Assessing genomic data quality and beyond. *Current Protocols* 1:e323. https://doi.org/10.1002/cpz1.323

Marcy, A. E., Fendorf, S., Patton, J. L., & Hadly, E. A. (2013). Morphological adaptations for digging and climate-impacted soil properties define pocket gopher (*Thomomys spp*.) distributions. *PloS One* 8:e64935. https://doi.org/10.1371/journal.pone.0064935

Marcy, A. E., Hadly, E. A., Sherratt, E., Garland, K., & Weisbecker, V. (2016). Getting a head in hard soils: Convergent skull evolution and divergent allometric patterns explain shape variation in a highly diverse genus of pocket gophers (*Thomomys*). *BMC Evolutionary Biology* 16:207. https://doi.org/10.1186/s12862-016-0782-1

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.Journal*, 17:10. https://doi.org/10.14806/ej.17.1.200

Mather, N., Traves, S. M., & Ho, S. Y. W. (2020). A practical introduction to sequentially Markovian coalescent methods for estimating demographic history from genomic data. *Ecology and Evolution* 10:579–589. https://doi.org/10.1002/ece3.5888

McDonald, J. H., & Kreitman, M. (1991). Adaptive protein evolution at the *Adh* locus in *Drosophila*. *Nature* 351:652-654. https://www.nature.com/articles/351652a0.pdf?origin=ppub

McDonough-Goldstein, C. E., Borziak, K., Pitnick, S., & Dorus, S. (2021). *Drosophila* female reproductive tract gene expression reveals coordinated mating responses and rapidly evolving tissue-specific genes. *G3* 11: jkab020. https://doi.org/10.1093/g3journal/jkab020

McGraw, L. A., Suarez, S. S., & Wolfner, M. F. (2015). On a matter of seminal importance. *BioEssays* 37:142–147. https://doi.org/10.1002/bies.201400117

Meléndez-Rosa, J., Bi, K., & Lacey, E. A. (2019). Differential gene expression in relation to mating system in Peromyscine rodents. *Ecology and Evolution* 9:5975–5990. https://doi.org/10.1002/ece3.5181

Merriam, C. H. (1895). Monographic revision of the pocket gophers, family Geomyidae (excl. of the species of *Thomomys*). *North American Fauna* 8:1–258. https://archive.org/details/monographicrevis08merr

Metcalf, A. E., Nunney, L., & Hyman, B. C. (2007). Geographic patterns of genetic differentiation within the restricted range of the endangered Stephens' Kangaroo rat *Dipodomys stephensi*. *Evolution* 55:1233–1244. https://doi.org/10.1111/j.0014-3820.2001.tb00643.x

Metz, E. C., & Palumbi, S. R. (1996). Positive selection and sequence rearrangements generate extensive polymorphism in the gamete recognition protein bindin. *Molecular Biology and Evolution* 13:397–406. https://doi.org/10.1093/oxfordjournals.molbev.a025598

Metz, E. C., Robles-Sikisaka, R., & Vacquier, V. D. (1998). Nonsynonymous substitution in abalone sperm fertilization genes exceeds substitution in introns and mitochondrial DNA. *Proceedings of the National Academy of Sciences USA* 95:10676–10681. https://doi.org/10.1073/pnas.95.18.10676

Mi, H., Muruganujan, A., Huang, X., Ebert, D., Mills, C., Guo, X., & Thomas, P. D. (2019). Protocol Update for large-scale genome and gene function analysis with the PANTHER classification system (v.14.0). *Nature Protocols* 14:703–721. https://doi.org/10.1038/s41596-019-0128-8

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., von Haeseler, A., & Lanfear, R. (2020). IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Molecular Biology and Evolution* 37:1530–1534. https://doi.org/10.1093/molbev/msaa015

Mirchandani, C. D., Shultz, A. J., Thomas, G. W. C., Smith, S. J., Baylis, M., Arnold, B., Corbett-Detig, R., Enbody, E., & Sackton, T. B. (2024). A fast, reproducible, high-throughput variant calling workflow for population genomics. *Molecular Biology and Evolution* 41:msad270. https://doi.org/10.1093/molbev/msad270

Morales, A. E., Dong, Y., Brown, T., Baid, K., Kontopoulos, D.-G., Gonzalez, V., Huang, Z., Ahmed, A.-W., Bhuinya, A., Hilgers, L., Winkler, S., Hughes, G., Li, X., Lu, P., Yang, Y., Kirilenko, B. M., Devanna, P., Lama, T. M., Nissan, Y., … Hiller, M. (2025). Bat genomes illuminate adaptations to viral tolerance and disease resistance. *Nature* 638:449–458. https://doi.org/10.1038/s41586-024-08471-0

Morin, P. A., Archer, F. I., Avila, C. D., Balacco, J. R., Bukhman, Y. V., Chow, W., Fedrigo, O., Formenti, G., Fronczek, J. A., Fungtammasan, A., Gulland, F. M. D., Haase, B., Peter Heide-Jorgensen, M., Houck, M. L., Howe, K., Misuraca, A. C., Mountcastle, J., Musser, W., Paez, S., … Jarvis, E. D. (2021). Reference genome and demographic history of the most endangered marine mammal, the vaquita. *Molecular Ecology Resources* 21:1008–1020. https://doi.org/10.1111/1755-0998.13284

Moritz, C. (2002). Strategies to protect biological diversity and the evolutionary processes that sustain it. *Systematic Biology* 51:238–254. https://doi.org/10.1080/10635150252899752

Morrison, M. L., Mills, L. S., & Kuenzi, A. J. (1996). Study and management of an isolated, rare population: the Fresno kangaroo rat. *Wildlife Society Bulletin* 24:602–606. https://www.jstor.org/stable/3783147?seq=1

Mittermeier, R. A., Gil, P. R., Hoffman, M., Pilgrim, J., Brooks, T., Mittermeier, C. G., Lamoreux, J., & da Fonseca, G. A. B. (2004). *Hotspots revisited: earth's biologically richest and most endangered terrestrial ecoregions*. Washington (DC): Conservation International.

Muñoz, M. M., Frishkoff, L. O., Pruett, J., & Mahler, D. L. (2023). Evolution of a model system: New insights from the study of *Anolis* lizards. *Annual Review of Ecology, Evolution, and Systematics* 54:475–503. https://doi.org/10.1146/annurev-ecolsys-110421-103306

Munshi-South, J., & Richardson, J. L. (2017). *Peromyscus* transcriptomics: Understanding adaptation and gene expression plasticity within and between species of deer mice. *Seminars in Cell & Developmental Biology* 61:131–139. https://doi.org/10.1016/j.semcdb.2016.08.011

Murrell, B., Weaver, S., Smith, M. D., Wertheim, J. O., Murrell, S., Aylward, A., Eren, K., Pollner, T., Martin, D. P., Smith, D. M., Scheffler, K., & Kosakovsky Pond, S. L. (2015). Gene-wide identification of episodic selection. *Molecular Biology and Evolution* 32:1365–1371. https://doi.org/10.1093/molbev/msv035

Nagy, K. A., & Gruchacz, M. J. (1994). Seasonal water and energy metabolism of the desert-dwelling kangaroo rat (*Dipodomys merriami*). *Physiological Zoology* 67:1461–1478. https://doi.org/10.1086/physzool.67.6.30163907

Neubaum, D. M., & Wolfner, M. F. (1999). Mated *Drosophila melanogaster* females require a seminal fluid protein, Acp36DE, to store sperm efficiently. *Genetics* 153:845–857. https://doi.org/10.1093/genetics/153.2.845

Noda, T., Fujihara, Y., Matsumura, T., Oura, S., Kobayashi, S., & Ikawa, M. (2019). Seminal vesicle secretory protein 7, PATE4, is not required for sperm function but for copulatory plug formation to ensure fecundity. *Biology of Reproduction* 100:1035–1045. https://doi.org/10.1093/biolre/ioy247

Noda, T., & Ikawa, M. (2019). Physiological function of seminal vesicle secretions on male fecundity. *Reproductive Medicine and Biology* 18:241–246. https://doi.org/10.1002/rmb2.12282

Ochoa, A., Onorato, D. P., Roelke-Parker, M. E., Culver, M., & Fitak, R. R. (2022). Give and take: Effects of genetic admixture on mutation load in endangered Florida panthers. *The Journal of Heredity* 113:491–499. https://doi.org/10.1093/jhered/esac037

Open2C, Abdennur, N., Fudenberg, G., Flyamer, I. M., Galitsyna, A. A., Goloborodko, A., Imakaev, M., & Venev, S. V. (2024). Pairtools: From sequencing data to chromosome contacts. *PLoS Computational Biology* 20:e1012164. https://doi.org/10.1371/journal.pcbi.1012164

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

Osmanski, A. B., Paulat, N. S., Korstian, J., Grimshaw, J. R., Halsey, M., Sullivan, K. A. M., Moreno-Santillán, D. D., Crookshanks, C., Roberts, J., Garcia, C., Johnson, M. G., Densmore, L. D., Stevens, R. D., Zoonomia Consortium†, Rosen, J., Storer, J. M., Hubley, R., Smit, A. F. A., Dávalos, L. M., … Ray, D. A. (2023). Insights into mammalian TE diversity through the curation of 248 genome assemblies. *Science* 380:eabn1430. https://doi.org/10.1126/science.abn1430

Panhuis, T. M., & Swanson, W. J. (2006). Molecular evolution and population genetic analysis of candidate female reproductive genes in *Drosophila*. *Genetics* 173:2039–2047. https://doi.org/10.1534/genetics.105.053611

Paradis, E., & Schliep, K. (2018). ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* 35:526–528. https://doi.org/10.1093/bioinformatics/bty633

Partha, R., Chauhan, B. K., Ferreira, Z., Robinson, J. D., Lathrop, K., Nischal, K. K., Chikina, M., & Clark, N. L. (2017). Subterranean mammals show convergent regression in ocular genes and enhancers, along with adaptation to tunneling. *eLife 6*. https://doi.org/10.7554/eLife.25884

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., & Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods* 14:417–419. https://doi.org/10.1038/nmeth.4197

Patton, J. L. (1972). Patterns of geographic variation in karyotype in the pocket gopher, *Thomomys bottae* (Eydoux and Gervais). *Evolution* 26:574–586. https://doi.org/10.2307/2407054

Patton, J. L. (1973). An analysis of natural hybridization between the pocket gophers, *Thomomys bottae* and *Thomomys umbrinus*, in Arizona. *Journal of Mammalogy* 54:561–584. https://doi.org/10.2307/1378959

Patton, J. L., & Dingman, R. E. (1970). Chromosome studies of pocket gophers, genus *Thomomys*. II. Variation in *T. bottae* in the American southwest. *Cytogenetics* 9:139–151. https://www.ncbi.nlm.nih.gov/pubmed/5461082

Patton, J. L., Hafner, J. C., Hafner, M. S., & Smith, M. F. (1979). Hybrid zones in *Thomomys bottae* pocket gophers: genetic, phenetic, and ecologic concordance patterns. *Evolution* 33:860–876. https://doi.org/10.2307/2407651

Patton, J. L., & Sherwood, S. W. (1982). Genome evolution in pocket gophers (genus *Thomomys*). I. Heterochromatin variation and speciation potential. *Chromosoma* 85:149–162. https://doi.org/10.1007/BF00294962

Patton, J. L., & Smith, M. F. (1990). *The Evolutionary Dynamics of the Pocket Gopher Thomomys Bottae, with Emphasis on California Populations*. University of California Press. Berkeley, USA.

Patton, J. L., Smith, M. F., Price, R. D., & Hellenthal, R. A. (1984). Genetics of hybridization between the pocket gophers *Thomomys bottae* and *Thomomys townsendii* in northeastern California. *The Great Basin Naturalist*. https://www.jstor.org/stable/41712092

Patton, J. L., Williams, D. F., Kelly, P. A., Cypher, B. L., & Phillips, S. E. (2019). Geographic variation and evolutionary history of *Dipodomys nitratoides* (Rodentia: Heteromyidae), a species in severe decline. *Journal of Mammalogy* 100:1546–1563. https://doi.org/10.1093/jmammal/gyz128

Patton, J. L., & Yang, S. Y. (1977). Genetic variation in *Thomomys bottae* pocket gophers: macrogeographic patterns. *Evolution* 31:697–720. https://doi.org/10.1111/j.1558-5646.1977.tb01064.x

Peart, C. R., Tusso, S., Pophaly, S. D., Botero-Castro, F., Wu, C.-C., Aurioles-Gamboa, D., Baird, A. B., Bickham, J. W., Forcada, J., Galimberti, F., Gemmell, N. J., Hoffman, J. I., Kovacs, K. M., Kunnasranta, M., Lydersen, C., Nyman, T., de Oliveira, L. R., Orr, A. J., Sanvito, S., … Wolf, J. B. W. (2020). Determinants of genetic variation across eco-evolutionary scales in pinnipeds. *Nature Ecology & Evolution* 4:1095–1104. https://doi.org/10.1038/s41559-020-1215-5

Petras, M. L. (1967). Studies of natural populations of *Mus*. I. Biochemical polymorphisms and their bearing on breeding structure. *Evolution* 21:259–274. https://doi.org/10.1111/j.1558-

5646.1967.tb00154.x

Platt, R. N., 2nd, Amman, B. R., Keith, M. S., Thompson, C. W., & Bradley, R. D. (2015). What Is *Peromyscus*? Evidence from nuclear and mitochondrial DNA sequences suggests the need for a new classification. *Journal of Mammalogy* 96:708–719. https://doi.org/10.1093/jmammal/gyv067

Platt, R. N., 2nd, Vandewege, M. W., & Ray, D. A. (2018). Mammalian transposable elements and their impacts on genome evolution. *Chromosome Research* 26:25–43. https://doi.org/10.1007/s10577-017-9570-z

Price, M. V., & Endo, P. R. (1989). Estimating the distribution and abundance of a cryptic species, *Dipodomys stephensi* (Rodentia: Heteromyidae), and implications for management. *Conservation Biology* 3:293–301. https://doi.org/10.1111/j.1523-1739.1989.tb00089.x

Prout, T., & Clark, A. G. (2000). Seminal fluid causes temporarily reduced egg hatch in previously mated females. *Proceedings of the Royal Society of London. Series B, Biological Sciences* 267:201–203. https://doi.org/10.1098/rspb.2000.0988

Ramírez, F., Bhardwaj, V., Arrigoni, L., Lam, K. C., Grüning, B. A., Villaveces, J., Habermann, B., Akhtar, A., & Manke, T. (2018). High-resolution TADs reveal DNA sequences underlying genome organization in flies. *Nature Communications* 9:189. https://doi.org/10.1038/s41467-017-02525-w

Ramm, S. A., McDonald, L., Hurst, J. L., Beynon, R. J., & Stockley, P. (2009). Comparative proteomics reveals evidence for evolutionary diversification of rodent seminal fluid and its functional significance in sperm competition. *Molecular Biology and Evolution* 26:189–198. https://doi.org/10.1093/molbev/msn237

Ramm, S. A., Oliver, P. L., Ponting, C. P., Stockley, P., & Emes, R. D. (2008). Sexual selection and the adaptive evolution of mammalian ejaculate proteins. *Molecular Biology and Evolution* 25:207–219. https://doi.org/10.1093/molbev/msm242

Ramm, S. A., Parker, G. A., & Stockley, P. (2005). Sperm competition and the evolution of male reproductive anatomy in rodents. *Proceedings of the Royal Society of London. Series B, Biological Sciences* 272:949–955. https://doi.org/10.1098/rspb.2004.3048

Ranallo-Benavidez, T. R., Jaron, K. S., & Schatz, M. C. (2020). GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nature Communications* 11:1432. https://doi.org/10.1038/s41467-020-14998-3

Ratto, M. H., Leduc, Y. A., Valderrama, X. P., van Straaten, K. E., Delbaere, L. T. J., Pierson, R. A., & Adams, G. P. (2012). The nerve of ovulation-inducing factor in semen. *Proceedings of the National Academy of Sciences USA* 109:15042–15047. https://doi.org/10.1073/pnas.1206273109

Raymond Hall, E. (1981). *The mammals of North America* (2nd ed.). Wiley.

Reid, K., Bell, M. A., & Veeramah, K. R. (2021). Threespine stickleback: a model system for evolutionary genomics. *Annual Review of Genomics and Human Genetics* 22:357–383. https://doi.org/10.1146/annurev-genom-111720-081402

Rhie, A., McCarthy, S. A., Fedrigo, O., Damas, J., Formenti, G., Koren, S., Uliano-Silva, M., Chow, W., Fungtammasan, A., Kim, J., Lee, C., Ko, B. J., Chaisson, M., Gedman, G. L., Cantin, L. J., Thibaud-Nissen, F., Haggerty, L., Bista, I., Smith, M., … Jarvis, E. D. (2021). Towards complete and error-free genome assemblies of all vertebrate species. *Nature* 592:737–746. https://doi.org/10.1038/s41586-021-03451-0

Rhie, A., Walenz, B. P., Koren, S., & Phillippy, A. M. (2020). Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biology* 21:245. https://doi.org/10.1186/s13059-020-02134-9

Ribble, D. O. (1991). The monogamous mating system of *Peromyscus californicus* as revealed by DNA fingerprinting. *Behavioral Ecology and Sociobiology* 29:161–166. https://doi.org/10.1007/BF00166397

Ribble, D. O., & Stanley, S. (1998). Home ranges and social organization of syntopic *Peromyscus boylii* and *P. truei*. *Journal of Mammalogy* 79:932–941. https://doi.org/10.2307/1383101

Riordan, E. C., & Rundel, P. W. (2014). Land use compounds habitat losses under projected climate change in a threatened California ecosystem. *PloS One* 9:e86487. https://doi.org/10.1371/journal.pone.0086487

Robinson, J. A., Bowie, R. C. K., Dudchenko, O., Aiden, E. L., Hendrickson, S. L., Steiner, C. C., Ryder, O. A., Mindell, D. P., & Wall, J. D. (2021). Genome-wide diversity in the California condor tracks its prehistoric abundance and decline. *Current Biology* 31:2939-2946. https://doi.org/10.1016/j.cub.2021.04.035

Robinson, J. A., Kyriazis, C. C., Nigenda-Morales, S. F., Beichman, A. C., Rojas-Bracho, L., Robertson, K. M., Fontaine, M. C., Wayne, R. K., Lohmueller, K. E., Taylor, B. L., & Morin, P. A. (2022). The critically endangered vaquita is not doomed to extinction by inbreeding depression. *Science* 376:635–639. https://doi.org/10.1126/science.abm1742

Robinson, J. A., Ortega-Del Vecchyo, D., Fan, Z., Kim, B. Y., vonHoldt, B. M., Marsden, C. D., Lohmueller, K. E., & Wayne, R. K. (2016). Genomic flatlining in the endangered island fox. *Current Biology* 26:1183–1189. https://doi.org/10.1016/j.cub.2016.02.062

Robinson, J. A., Räikkönen, J., Vucetich, L. M., Vucetich, J. A., Peterson, R. O., Lohmueller, K. E., & Wayne, R. K. (2019). Genomic signatures of extensive inbreeding in Isle Royale wolves, a population on the threshold of extinction. *Science Advances* 5:eaau0757. https://doi.org/10.1126/sciadv.aau0757

Robinson, J., Kyriazis, C. C., Yuan, S. C., & Lohmueller, K. E. (2023). Deleterious variation in natural populations and implications for conservation genetics. *Annual Review of Animal Biosciences* 11:93–114. https://doi.org/10.1146/annurev-animal-080522-093311

Rocha, J. L., Godinho, R., Brito, J. C., & Nielsen, R. (2021). Life in deserts: the genetic basis of mammalian desert adaptation. *Trends in Ecology & Evolution* 36:637–650. https://doi.org/10.1016/j.tree.2021.03.007

Rocha, J. L., Silva, P., Santos, N., Nakamura, M., Afonso, S., Qninba, A., Boratynski, Z., Sudmant, P. H., Brito, J. C., Nielsen, R., & Godinho, R. (2023). North African fox genomes show signatures of repeated introgression and adaptation to life in deserts. *Nature Ecology & Evolution* 7:1267–1286. https://doi.org/10.1038/s41559-023-02094-w

Roelke, M. E., Martenson, J. S., & O'Brien, S. J. (1993). The consequences of demographic reduction and genetic depletion in the endangered Florida panther. *Current Biology* 3:340–350. https://doi.org/10.1016/0960-9822(93)90197-v

Rowe, K. C., Rowe, K. M. C., Tingley, M. W., Koo, M. S., Patton, J. L., Conroy, C. J., Perrine, J. D., Beissinger, S. R., & Moritz, C. (2015). Spatially heterogeneous impact of climate change on small mammals of montane California. *Proceedings of the Royal Society of London. Series B, Biological Sciences* 282:2014.1857. https://doi.org/10.1098/rspb.2014.1857

Rubidge, E. M., Patton, J. L., Lim, M., Burton, A. C., Brashares, J. S., & Moritz, C. (2012). Climate-induced range contraction drives genetic erosion in an alpine mammal. *Nature Climate Change* 2:285–288. https://doi.org/10.1038/nclimate1415

Saslaw, L., & Cypher, B. (2020). Strategies for translocating endangered giant kangaroo rats (*Dipodomys ingens*). *Western Wildlife* 7:30–37. https://www.esrp.org/publications/pdf/Saslaw_Cypher_2020_GKR%20translocation_WW.pdf

Sawyer, Y. E., Flamme, M. J., Jung, T. S., MacDonald, S. O., & Cook, J. A. (2017). Diversification of deer mice (Rodentia:Genus *Peromyscus*) at their north-western range limit: genetic consequences of refugial and island isolation. *Journal of Biogeography* 44:1572–1585. https://doi.org/10.1111/jbi.12995

Schein, M., Yang, Z., Mitchell-Olds, T., & Schmid, K. J. (2004). Rapid evolution of a pollen-specific oleosin-like gene family from *Arabidopsis thaliana* and closely related species. *Molecular Biology and Evolution* 21:659–669. https://doi.org/10.1093/molbev/msh059

Schmidt, C., Hoban, S., Hunter, M., Paz-Vinas, I., & Garroway, C. J. (2023). Genetic diversity and IUCN Red List status. *Conservation Biology* 37:e14064. https://doi.org/10.1111/cobi.14064

Schmidt-Nielsen, K., & Schmidt-Nielsen, B. (1952). Water metabolism of desert mammals. *Physiological Reviews* 32:135–166. https://doi.org/10.1152/physrev.1952.32.2.135

Schneider, M. R., Mangels, R., & Dean, M. D. (2016). The molecular basis and reproductive

function(s) of copulatory plugs. *Molecular Reproduction and Development* 83:755–767. https://doi.org/10.1002/mrd.22689

Scott, P. A., Allison, L. J., Field, K. J., Averill-Murray, R. C., & Shaffer, H. B. (2020). Individual heterozygosity predicts translocation success in threatened desert tortoises. *Science* 370:1086–1089. https://doi.org/10.1126/science.abb0421

Selander, R. (1970). Behavior and genetic variation in natural populations. *American Zoologist* 10:53–66. https://doi.org/10.1093/ICB/10.1.53

Sendrowski, J., & Bataillon, T. (2024). fastDFE: Fast and Flexible Inference of the Distribution of Fitness Effects. *Molecular Biology and Evolution* 41:msae070. https://doi.org/10.1093/molbev/msae070

Shafer, A. B. A., & Kardos, M. (2025). Runs of homozygosity and inferences in wild populations. *Molecular Ecology* 34:e17641. https://doi.org/10.1111/mec.17641

Shaffer, H. B., Toffelmier, E., Corbett-Detig, R. B., Escalona, M., Erickson, B., Fiedler, P., Gold, M., Harrigan, R. J., Hodges, S., Luckau, T. K., Miller, C., Oliveira, D. R., Shaffer, K. E., Shapiro, B., Sork, V. L., & Wang, I. J. (2022). Landscape genomics to enable conservation actions: The California Conservation Genomics Project. *The Journal of Heredity* 113:577-588. https://doi.org/10.1093/jhered/esac020

Sharkey, D. J., Macpherson, A. M., Tremellen, K. P., & Robertson, S. A. (2007). Seminal plasma differentially regulates inflammatory cytokine gene expression in human cervical and vaginal epithelial cells. *Molecular Human Reproduction* 13:491–501. https://doi.org/10.1093/molehr/gam028

Shen, W., Le, S., Li, Y., & Hu, F. (2016). SeqKit: A cross-platform and ultrafast toolkit for FASTA/Q file manipulation. *PloS One* 11:e0163962. https://doi.org/10.1371/journal.pone.0163962

Sherwood, S. W., & Patton, J. L. (1982). Genome evolution in pocket gophers (genus *Thomomys*). II. Variation in cellular DNA content. *Chromosoma* 85:163–179. https://doi.org/10.1007/BF00294963

Shi, J., Fok, K. L., Dai, P., Qiao, F., Zhang, M., Liu, H., Sang, M., Ye, M., Liu, Y., Zhou, Y., Wang, C., Sun, F., Xie, G., & Chen, H. (2021). Spatio-temporal landscape of mouse epididymal cells and specific mitochondria-rich segments defined by large-scale single-cell RNA-seq. *Cell Discovery* 7:34. https://doi.org/10.1038/s41421-021-00260-7

Shier, D. M., Navarro, A. Y., Tobler, M., Thomas, S. M., King, S. N. D., Mullaney, C. B., & Ryder, O. A. (2021). Genetic and ecological evidence of long-term translocation success of the federally endangered Stephens' kangaroo rat. *Conservation Science and Practice* 3:e478. https://doi.org/10.1111/csp2.478

Shier, D. M., & Swaisgood, R. R. (2012). Fitness costs of neighborhood disruption in translocations of a solitary mammal: Social effects in translocation. *Conservation Biology* 26:116–123. https://doi.org/10.1111/j.1523-1739.2011.01748.x

Shultz, A. J., & Sackton, T. B. (2019). Immune genes are hotspots of shared positive selection across birds and mammals. *eLife* 8:41815. https://doi.org/10.7554/eLife.41815

Shindo, M., Inui, M., Kang, W., Tamano, M., Tingwei, C., Takada, S., Hibino, T., Yoshida, M., Yoshida, K., Okada, H., Iwamoto, T., Miyado, K., & Kawano, N. (2019). Deletion of a seminal gene cluster reinforces a crucial role of SVS2 in male fertility. *International Journal of Molecular Sciences* 20:4557. https://doi.org/10.3390/ijms20184557

Sikes, R. S., & Animal Care and Use Committee of the American Society of Mammalogists. (2016). 2016 Guidelines of the American Society of Mammalogists for the use of wild mammals in research and education. *Journal of Mammalogy* 97:663–688. https://doi.org/10.1093/jmammal/gyw078

Sim, S. B., Corpuz, R. L., Simmonds, T. J., & Geib, S. M. (2022). HiFiAdapterFilt, a memory efficient read processing pipeline, prevents occurrence of adapter sequence in PacBio HiFi reads and their negative impacts on genome assembly. *BMC Genomics* 23:157. https://doi.org/10.1186/s12864-022-08375-1

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31:3210–3212. https://doi.org/10.1093/bioinformatics/btv351

Simmons, L. W., Sloan, N. S., & Firman, R. C. (2020). Sexual selection shapes seminal vesicle secretion gene expression in house mice. *Molecular Biology and Evolution* 37:1114–1117. https://doi.org/10.1093/molbev/msz295

Smith, M. F. (1998). Phylogenetic relationships and geographic structure in pocket gophers in the genus *Thomomys*. *Molecular Phylogenetics and Evolution* 9:1–14. https://doi.org/10.1006/mpev.1997.0459

Smyth, S. P., Nixon, B., Anderson, A. L., Murray, H. C., Martin, J. H., MacDougall, L. A., Robertson, S. A., Skerrett-Byrne, D. A., & Schjenken, J. E. (2022). Elucidation of the protein composition of mouse seminal vesicle fluid. *Proteomics* 22:e2100227. https://doi.org/10.1002/pmic.202100227

Snook, R. R. (2005). Sperm in competition: not playing by the numbers. *Trends in Ecology & Evolution* 20:46–53. https://doi.org/10.1016/j.tree.2004.10.011

Spielman, D., Brook, B. W., & Frankham, R. (2004). Most species are not driven to extinction before genetic factors impact them. *Proceedings of the National Academy of Sciences USA* 101:5261–15264. https://doi.org/10.1073/pnas.0403809101

Statham, M. J., Bean, W. T., Alexander, N., Westphal, M. F., & Sacks, B. N. (2019). Historical population size change and differentiation of relict populations of the endangered giant kangaroo rat. *The Journal of Heredity* 110:548–558. https://doi.org/10.1093/jhered/esz006

Sullivan, R., & Mieusset, R. (2016). The human epididymis: its function in sperm maturation. *Human Reproduction Update* 22:574–587. https://doi.org/10.1093/humupd/dmw015

Sumner, F. B. (1917). Several color "mutations" in mice of the genus *Peromyscus*. *Genetics* 2:291–300. https://doi.org/10.1093/genetics/2.3.291

Sutter, A., Simmons, L. W., Lindholm, A. K., & Firman, R. C. (2016). Function of copulatory plugs in house mice: mating behavior and paternity outcomes of rival males. *Behavioral Ecology* 27:185–195. https://doi.org/10.1093/beheco/arv138

Swanson, W. J., Clark, A. G., Waldrip-Dail, H. M., Wolfner, M. F., & Aquadro, C. F. (2001). Evolutionary EST analysis identifies rapidly evolving male reproductive proteins in *Drosophila*. *Proceedings of the National Academy of Sciences USA* 98:7375–7379. https://doi.org/10.1073/pnas.131568198

Swanson, W. J., & Vacquier, V. D. (2002). The rapid evolution of reproductive proteins. *Nature Reviews Genetics* 3:137–144. https://doi.org/10.1038/nrg733

Swanson, W. J., Wong, A., Wolfner, M. F., & Aquadro, C. F. (2004). Evolutionary expressed sequence tag analysis of *Drosophila* female reproductive tracts identifies genes subjected to positive selection. *Genetics* 168:1457–1465. https://doi.org/10.1534/genetics.104.030478

Swanson, W. J., Yang, Z., Wolfner, M. F., & Aquadro, C. F. (2001). Positive Darwinian selection drives the evolution of several female reproductive proteins in mammals. *Proceedings of the National Academy of Sciences USA* 98:2509–2514. https://doi.org/10.1073/pnas.051605998

Tanaka, T., & Nei, M. (1989). Positive Darwinian selection observed at the variable-region genes of immunoglobulins. *Molecular Biology and Evolution* 6:447–459. https://doi.org/10.1093/OXFORDJOURNALS.MOLBEV.A040569

Tang, H., Bowers, J. E., Wang, X., Ming, R., Alam, M., & Paterson, A. H. (2008). Synteny and collinearity in plant genomes. *Science* 320:486–488. https://doi.org/10.1126/science.1153917

Tarailo-Graovac, M., & Chen, N. (2009). Using RepeatMasker to identify repetitive elements in genomic sequences. *Current Protocols in Bioinformatics*, 4.10.1-4.10.14. https://doi.org/10.1002/0471250953.bi0410s25

Tataru, P., & Bataillon, T. (2020). polyDFE: Inferring the distribution of fitness effects and properties of beneficial mutations from polymorphism data. In J. Y. Dutheil (Ed.), *Statistical Population Genomics* (pp. 125–146). Springer US. https://doi.org/10.1007/978-1-0716-0199-0_6

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. https://doi.org/10.1534/genetics.117.300323

Teixeira, J. C., & Huber, C. D. (2021). The inflated significance of neutral genetic diversity in conservation genetics. *Proceedings of the National Academy of Sciences USA* 118: e2015096118. https://doi.org/10.1073/pnas.2015096118

Tennant, E. N. (2011). *Conservation of Tipton kangaroo rats (Dipodomys nitratoides nitratoides): effects of competition and potential for translocation*. California State University, Bakersfield. https://scholarworks.calstate.edu/downloads/ng451p80z

Tennant, E. N., Germano, D. J., Cypher, B. L. (2013). Translocating endangered kangaroo rats in the San Joaquin Valley of California: recommendations for future efforts. *California Fish and Game* 99:90–103. https://www.csub.edu/~dgermano/TKR-Tennant%20et%20al.-CDF.pdf

Thaeler, C. S. (1980). Chromosome numbers and systematic relations in the genus *Thomomys* (Rodentia: Geomyidae). *Journal of Mammalogy* 61:414–422. https://doi.org/10.2307/1379835

Thaeler, C. S. (1968). An analysis of three hybrid populations of pocket gophers (genus *Thomomys*). *Evolution* 22:543-555. https://doi.org/10.2307/2406879

Thomas, P. D., Ebert, D., Muruganujan, A., Mushayahama, T., Albou, L.-P., & Mi, H. (2022). PANTHER: Making genome-scale phylogenetics accessible to all. *Protein Science* 31:8–22. https://doi.org/10.1002/pro.4218

Thomassen, H. A., Fuller, T., Buermann, W., Milá, B., Kieswetter, C. M., Jarrín-V, P., Cameron, S. E., Mason, E., Schweizer, R., Schlunegger, J., Chan, J., Wang, O., Peralvo, M., Schneider, C. J., Graham, C. H., Pollinger, J. P., Saatchi, S., Wayne, R. K., & Smith, T. B. (2011). Mapping evolutionary process: a multi-taxa approach to conservation prioritization: Conservation of pattern and process. *Evolutionary Applications* 4:397–413. https://doi.org/10.1111/j.1752-4571.2010.00172.x

Tian, D., Patton, A. H., Turner, B. J., & Martin, C. H. (2022). Severe inbreeding, increased mutation load and gene loss-of-function in the critically endangered Devils Hole pupfish. *Proceedings of the Royal Society of London. Series B, Biological Sciences* 289:2022.1561. https://doi.org/10.1098/rspb.2022.1561

Toffelmier, E., Beninde, J., & Shaffer, H. B. (2022). The phylogeny of California, and how it informs setting multispecies conservation priorities. *The Journal of Heredity* 113:597–603. https://doi.org/10.1093/jhered/esac045

Tollner, T. L., Venners, S. A., Hollox, E. J., Yudin, A. I., Liu, X., Tang, G., Xing, H., Kays, R. J., Lau, T., Overstreet, J. W., Xu, X., Bevins, C. L., & Cherr, G. N. (2011). A common mutation in the defensin DEFB126 causes impaired sperm function and subfertility. *Science Translational Medicine* 3:92ra65. https://doi.org/10.1126/scitranslmed.3002289

Torgerson, D. G., Kulathinal, R. J., & Singh, R. S. (2002). Mammalian sperm proteins are rapidly evolving: evidence of positive selection in functionally diverse genes. *Molecular Biology and Evolution* 19:1973–1980. https://doi.org/10.1093/oxfordjournals.molbev.a004021

Tørresen, O. K., Star, B., Mier, P., Andrade-Navarro, M. A., Bateman, A., Jarnot, P., Gruca, A., Grynberg, M., Kajava, A. V., Promponas, V. J., Anisimova, M., Jakobsen, K. S., & Linke, D. (2019). Tandem repeats lead to sequence assembly errors and impose multi-level challenges for genome and protein databases. *Nucleic Acids Research* 47:10994–11006. https://doi.org/10.1093/nar/gkz841

Tourmente, M., Gomendio, M., & Roldan, E. R. S. (2011). Sperm competition and the evolution of sperm design in mammals. *BMC Evolutionary Biology* 11:12. https://doi.org/10.1186/1471-2148-11-12

Tourmente, M., Rowe, M., González-Barroso, M. M., Rial, E., Gomendio, M., & Roldan, E. R. S. (2013). Postcopulatory sexual selection increases ATP content in rodent spermatozoa. *Evolution* 67:1838–1846. https://doi.org/10.1111/evo.12079

Tracy, R. L., & Walsberg, G. E. (2001). Intraspecific variation in water loss in a desert rodent, *Dipodomys merriami*. *Ecology* 82:1130–1137. https://doi.org/10.1890/0012-9658(2001)082[1130:iviwli]2.0.co;2

Tracy, R. L., & Walsberg, G. E. (2002). Kangaroo rats revisited: re-evaluating a classic case of desert survival. *Oecologia* 133:449–457. https://doi.org/10.1007/s00442-002-1059-5

Tram, U., & Wolfner, M. F. (1999). Male seminal fluid proteins are essential for sperm storage in *Drosophila melanogaster*. *Genetics* 153:837–844. https://doi.org/10.1093/genetics/153.2.837

Tsaur, S. C., & Wu, C. I. (1997). Positive selection and the molecular evolution of a gene of male reproduction, Acp26Aa of *Drosophila*. *Molecular Biology and Evolution* 14:544–549. https://doi.org/10.1093/oxfordjournals.molbev.a025791

Tseng, H.-C., Lin, H.-J., Tang, J.-B., Gandhi, P. S. S., Chang, W.-C., & Chen, Y.-H. (2009). Identification of the major TG4 cross-linking sites in the androgen-dependent SVS I exclusively expressed in mouse seminal vesicle. *Journal of Cellular Biochemistry* 107:899–907. https://doi.org/10.1002/jcb.22190

Tseng, H.-C., Tang, J.-B., Gandhi, P. S. S., Luo, C.-W., Ou, C.-M., Tseng, C.-J., Lin, H.-J., & Chen, Y.-H. (2012). Mutual adaptation between mouse transglutaminase 4 and its native substrates in the formation of copulatory plug. *Amino Acids* 42:951–960. https://doi.org/10.1007/s00726-011-1009-9

Turner, L. M., Chuong, E. B., & Hoekstra, H. E. (2008). Comparative analysis of testis protein evolution in rodents. *Genetics* 179:2075–2089. https://doi.org/10.1534/genetics.107.085902

Turner, L. M., & Hoekstra, H. E. (2006). Adaptive evolution of fertilization proteins within a genus: variation in ZP2 and ZP3 in deer mice (*Peromyscus*). *Molecular Biology and Evolution* 23:1656–1669. https://doi.org/10.1093/molbev/msl035

Turner, L. M., & Hoekstra, H. E. (2008). Reproductive protein evolution within and between species: maintenance of divergent ZP3 alleles in *Peromyscus*. *Molecular Ecology* 17:2616–2628. https://doi.org/10.1111/j.1365-294X.2008.03780.x

Turner, L. M., Young, A. R., Römpler, H., Schöneberg, T., Phelps, S. M., & Hoekstra, H. E. (2010). Monogamy evolves through multiple mechanisms: evidence from V1aR in deer mice. *Molecular Biology and Evolution* 27:1269–1278. https://doi.org/10.1093/molbev/msq013

Turner, S. D. (2014). qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. In *bioRxiv* (p. 005165). bioRxiv. https://doi.org/10.1101/005165

Uliano-Silva, M., Ferreira, J. G. R. N., Krasheninnikova, K., Darwin Tree of Life Consortium, Formenti, G., Abueg, L., Torrance, J., Myers, E. W., Durbin, R., Blaxter, M., & McCarthy, S. A. (2023). MitoHiFi: a python pipeline for mitochondrial genome assembly from PacBio high fidelity reads. *BMC Bioinformatics* 24:288. https://doi.org/10.1186/s12859-023-05385-y

Upham, N. S., Esselstyn, J. A., & Jetz, W. (2019). Inferring the mammal tree: Species-level sets of phylogenies for questions in ecology, evolution, and conservation. *PLoS Biology* 17:e3000494. https://doi.org/10.1371/journal.pbio.3000494

U.S. Fish and Wildlife Service. (1998). Final Rule to List the San Bernardino Kangaroo Rat as Endangered (1998). *USFWS*. https://www.govinfo.gov/content/pkg/FR-1998-09-24/pdf/98-25545.pdf#page=1

U.S. Fish and Wildlife Service. (2002). Endangered and Threatened Wildlife and Plants; Final Designation of Critical Habitat for the San Bernardino Kangaroo Rat. *USFWS*. https://www.fws.gov/policy/library/2002/02fr19811.pdf

U.S. Fish and Wildlife Service. (2020a). 5-YEAR REVIEW Giant kangaroo rat *(Dipodomys ingens)*. https://ecosphere-documents-production-public.s3.amazonaws.com/sams/public_docs/species_nonpublish/3095.pdf

U.S. Fish and Wildlife Service. (2020b). 5-YEAR REVIEW San Bernardino Kangaroo Rat (*Dipodomys merriami parvus*). *USFWS*. https://ecosphere-documents-production-public.s3.amazonaws.com/sams/public_docs/species_nonpublish/2975.pdf

U.S. Fish and Wildlife Service. (2020c). 5-YEAR REVIEW Tipton Kangaroo Rat *(Dipodomys nitratoides nitratoides)*. https://ecosphere-documents-production-public.s3.amazonaws.com/sams/public_docs/species_nonpublish/2987.pdf

U.S. Fish and Wildlife Service. (2020d). Species Status Assessment Report for the Giant Kangaroo Rat *(Dipodomys ingens)*. https://iris.fws.gov/APPS/ServCat/DownloadFile/180415

U.S. Fish and Wildlife Service. (2021). Species Report for Stephens' Kangaroo Rat *(Dipodomys stephensi)*. https://iris.fws.gov/APPS/ServCat/DownloadFile/214387

U.S. Fish and Wildlife Service. (2022). Endangered and Threatened Wildlife and Plants; Reclassification of Stephens' Kangaroo Rat from Endangered to Threatened with a Section 4(d) Rule. In *Federal Register* (Nos. 2022–03317; Vol. 87, pp. 8967–8981). https://www.federalregister.gov/d/2022-03317

U.S. Fish and Wildlife Service. (2025). ECOS environmental conservation online system; listed species believed to or known to occur in each State. https://ecos.fws.gov/ecp/report/species-listings-by-state-totals?statusCategory=Listed

Veltsos, P., Porcelli, D., Fang, Y., Cossins, A. R., Ritchie, M. G., & Snook, R. R. (2022). Experimental sexual selection reveals rapid evolutionary divergence in sex-specific transcriptomes and their interactions following mating. *Molecular Ecology* 31:3374–3388. https://doi.org/10.1111/mec.16473

Vicens, A., Lüke, L., & Roldan, E. R. S. (2014). Proteins involved in motility and sperm-egg interaction evolve more rapidly in mouse spermatozoa. *PloS One* 9:e91302. https://doi.org/10.1371/journal.pone.0091302

Voss, E. R., & Nachman, M. W. (2025). Mating system variation and gene expression in the male reproductive tract of Peromyscus mice. *Molecular Ecology* 34:e17433. https://doi.org/10.1111/mec.17433

Voss, E. R., Escalona, M., Kozak, K. M., Seligmann, W., Fairbairn, C., Nguyen, O., Marimuthu, M. P. A., Conroy, C., Patton, J. L., Bowie, R. C., & Nachman, M. W. (2024). De novo genome assembly of a Geomyid rodent, Botta's pocket gopher (*Thomomys bottae bottae*). *The Journal of Heredity* 116:esae045. https://doi.org/10.1093/jhered/esae045

Voss, E. R., Escalona, M., Nguyen, O., Marimuthu, M. P. A., Chumchim, N., Fairbairn, C. W., Seligmann, W., Beraut, E., Conroy, C. J., Patton, J. L., Bowie, R. C. K., & Nachman, M. W. (2025). A high-quality genome assembly for a desert-adapted rodent, Merriam's kangaroo rat (*Dipodomys merriami*). *The Journal of Heredity* 117:esaf023. https://doi.org/10.1093/jhered/esaf023

Voss, R. C. (1979). Male accessory glands and the evolution of copulatory plugs in rodents. *OCCASIONAL PAPERS OF THE MUSEUM OF ZOOLOGY UNIVERSITY OF MICHIGAN* 689:1–27. https://deepblue.lib.umich.edu/bitstream/handle/2027.42/57125/OP689.pdf

Waberski, D., Claassen, R., Hahn, T., Jungblut, P. W., Parvizi, N., Kallweit, E., & Weitze, K. F. (1997). LH profile and advancement of ovulation after transcervical infusion of seminal plasma at different stages of oestrus in gilts. *Journal of Reproduction and Fertility* 109:29–34. https://doi.org/10.1530/jrf.0.1090029

Weber, J. N., Peterson, B. K., & Hoekstra, H. E. (2013). Discrete genetic modules are responsible for complex burrow evolution in *Peromyscus* mice. *Nature* 493:402–405.

https://doi.org/10.1038/nature11816

Weber, W. D., & Fisher, H. S. (2023). Sexual selection drives the coevolution of male and female reproductive traits in *Peromyscus* mice. *Journal of Evolutionary Biology* 36:67–81. https://doi.org/10.1111/jeb.14126

Wertheim, J. O., Murrell, B., Smith, M. D., Kosakovsky Pond, S. L., & Scheffler, K. (2015). RELAX: detecting relaxed selection in a phylogenetic framework. *Molecular Biology and Evolution* 32:820–832. https://doi.org/10.1093/molbev/msu400

Whitehead, A., & Crawford, D. L. (2006). Neutral and adaptive variation in gene expression. *Proceedings of the National Academy of Sciences USA* 103:5425–5430. https://doi.org/10.1073/pnas.0507648103

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J. L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O., Paux, E., SanMiguel, P., & Schulman, A. H. (2007). A unified classification system for eukaryotic transposable elements. *Nature Reviews Genetics* 8:973–982. https://doi.org/10.1038/nrg2165

Wickham, H. (2016). Programming with ggplot2. In *Use R!* (pp. 241–253). Springer International Publishing. https://doi.org/10.1007/978-3-319-24277-4_12

Widick, I. V., & Bean, W. T. (2019). Evaluating current and future range limits of an endangered, keystone rodent (*Dipodomys ingens*). *Diversity & Distributions* 25:1074–1087. https://doi.org/10.1111/ddi.12914

Wilder, A. P., Dudchenko, O., Curry, C., Korody, M., Turbek, S. P., Daly, M., Misuraca, A., Wang, G., Khan, R., Weisz, D., Fronczek, J., Aiden, E. L., Houck, M. L., Shier, D. M., Ryder, O. A., & Steiner, C. C. (2022). A chromosome-length reference genome for the endangered Pacific pocket mouse reveals recent inbreeding in a historically large population. *Genome Biology and Evolution* 14:evac122. https://doi.org/10.1093/gbe/evac122

Wilder, A. P., Navarro, A. Y., King, S. N. D., Miller, W. B., Thomas, S. M., Steiner, C. C., Ryder, O. A., & Shier, D. M. (2020). Fitness costs associated with ancestry to isolated populations of an endangered species. *Conservation Genetics* 21:589–601. https://doi.org/10.1007/s10592-020-01272-8

Wilkening, J., Pearson-Prestera, W., Mungi, N. A., & Bhattacharyya, S. (2019). Endangered species management and climate change: when habitat conservation becomes a moving target. *Wildlife Society Bulletin* 43:11–20. https://doi.org/10.1002/wsb.944

Willoughby, J. R., Sundaram, M., Wijayawardena, B. K., Kimble, S. J. A., Ji, Y., Fernandez, N. B., Antonides, J. D., Lamb, M. C., Marra, N. J., & DeWoody, J. A. (2015). The reduction of genetic diversity in threatened vertebrates and new recommendations regarding IUCN conservation rankings. *Biological Conservation* 191:495–503. https://doi.org/10.1016/j.biocon.2015.07.025

Wlasiuk, G., & Nachman, M. W. (2007). The genetics of adaptive coat color in gophers: coding variation at Mc1r is not responsible for dorsal color differences. *The Journal of Heredity* 98:567–574. https://doi.org/10.1093/jhered/esm059

Wooldridge, T. B., Kautt, A. F., Lassance, J.-M., McFadden, S., Domingues, V. S., Mallarino, R., & Hoekstra, H. E. (2022). An enhancer of Agouti contributes to parallel evolution of cryptically colored beach mice. *Proceedings of the National Academy of Sciences USA*, 119:e2202862119. https://doi.org/10.1073/pnas.2202862119

Wooldridge, B., Orland, C., Enbody, E., Escalona, M., Mirchandani, C., Corbett-Detig, R., Kapp, J. D., Fletcher, N., Cox-Ammann, K., Raimondi, P., & Shapiro, B. (2024). Limited genomic signatures of population collapse in the critically endangered black abalone (*Haliotis cracherodii*). *Molecular Ecology* e17362. https://doi.org/10.1111/mec.17362

Wyckoff, G. J., Wang, W., & Wu, C. I. (2000). Rapid evolution of male reproductive genes in the descent of man. *Nature* 403:304–309. https://doi.org/10.1038/35002070

Yan, H., Bombarely, A., & Li, S. (2020). DeepTE: a computational method for de novo classification of transposons with convolutional neural network. *Bioinformatics* 36:4269–4275. https://doi.org/10.1093/bioinformatics/btaa519

Yang, Z., Swanson, W. J., & Vacquier, V. D. (2000). Maximum-likelihood analysis of molecular adaptation in abalone sperm lysin reveals variable selective pressures among lineages and sites. *Molecular Biology and Evolution* 17:1446–1455. https://doi.org/10.1093/oxfordjournals.molbev.a026245

York, R. A., Patil, C., Abdilleh, K., Johnson, Z. V., Conte, M. A., Genner, M. J., McGrath, P. T., Fraser, H. B., Fernald, R. D., & Streelman, J. T. (2018). Behavior-dependent cis regulation reveals genes and pathways associated with bower building in cichlid fishes. *Proceedings of the National Academy of Sciences USA* 115:E11081–E11090. https://doi.org/10.1073/pnas.1810140115

Yu, G., Smith, D. K., Zhu, H., Guan, Y., & Lam, T. T.-Y. (2017). Ggtree: An R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution* 8:28–36. https://doi.org/10.1111/2041-210x.12628

Zeng, Z., & Brown, J. H. (1987). Population ecology of a desert rodent: *Dipodomys merriami* in the Chihuahuan desert. *Ecology* 68:1328–1340. https://doi.org/10.2307/1939217

Zhang, Z., Schwartz, S., Wagner, L., & Miller, W. (2000). A greedy algorithm for aligning DNA sequences. *Journal of Computational Biology* 7:203–214. https://doi.org/10.1089/10665270050081478

APPENDICES

Appendix 1. Data Availability

CHAPTER 1

Data generated for this study are available under NCBI BioProject PRJNA777211 and all associated accession numbers are listed in Table 2. Raw sequencing data for individual MVZ:Mamm:240117 (NCBI BioSample SAMN29044276) are deposited in the NCBI Short Read Archive under SRR23102654 for PacBio HiFi sequencing data, and SRR23102652 and SRR23102653 for the Omni-C Illumina sequencing data. GenBank accessions for primary and alternate assemblies are GCA_026229955.1 and GCA_026167925.1; and for genome sequences JAOPKW000000000 and JAOPKX000000000. Assembly scripts can be found at www.github.com/ccgproject/ccgp_assembly. Scripts for comparative analysis and figures: www.github.com/evo-eco-gen/CCGP_Peromyscus. Liftoff annotation and RagTag scaffolding files are available on Dryad.

CHAPTER 2

Illumina RNA sequencing data from this study are available through the NCBI Sequence Read Archive under BioProject ID PRJNA1068126, and assembled transcriptomes are accessible through the Dryad Data Repository (DOI: 10.5061/dryad.b5mk-kwhmw). Animal samples collected by ERV are deposited with the Museum of Vertebrate Zoology collection at the University of California, Berkeley, and metadata will be made available on the Arctos museum database.

CHAPTER 3

Genome assemblies generated for this study are available under NCBI BioProject IDs PRJNA851460 (principal) and PRJNA851459 (alternate). Raw sequencing data for sample MVZ:Mamm:240054 (NCBI BioSample SAMN29046532) are deposited in the NCBI Short Read Archive (SRA) under accessions SRX17304138 - SRX17304140. Assembly scripts and other data for the analyses presented can be found at the following GitHub repository: www.github.com/ccgproject/ccgp_assembly. Preliminary annotation and mitochondrial genome sequence are available on the Dryad Data repository at https://doi.org/10.5061/dryad.x0k6djhtc.

CHAPTER 4

Raw Illumina sequence data for all individuals is available on the NCBI Sequence Read Archive under BioProject ID PRJNA1211119. The *D. merriami* reference genome used as the basis for variant calling is available on NCBI under accession number GCA_024711535.1. MVZ specimen catalog numbers are listed in Supplementary Table 1; additional sample metadata is available on the Museum of Vertebrate Zoology Mammal specimen catalog on the Arctos museum database at https://arctos.database.museum/search.cfm?guid_prefix=MVZ:Mamm.

CHAPTER 5

Genome assemblies generated for this study are available under NCBI BioProject IDs PRJNA851166 and PRJNA851165. Raw sequencing data for sample MVZ:Mamm:240275 (NCBI BioSample SAMN29044214) are deposited in the NCBI Short Read Archive under

accessions SRR21383821for PacBio HiFi sequencing data, and SRR21383819-20 for the Omni-C Illumina sequencing data. GenBank accessions for both haplotypes are GCA_024803745.1 and GCA_024803775.1; and for genome sequences JANJXV000000000 and JANJXW000000000; mitochondrial genome accession is CM045757.1. Additional NCBI data identifiers are listed in Supplementary Table 1. The *de novo* repetitive element library for *T. bottae* is available on Dryad Data Repository at doi:10.5061/dryad.wh70rxwwp. Assembly scripts can be found at the following GitHub repository: www.github.com/ccgproject/ccgp_assembly.

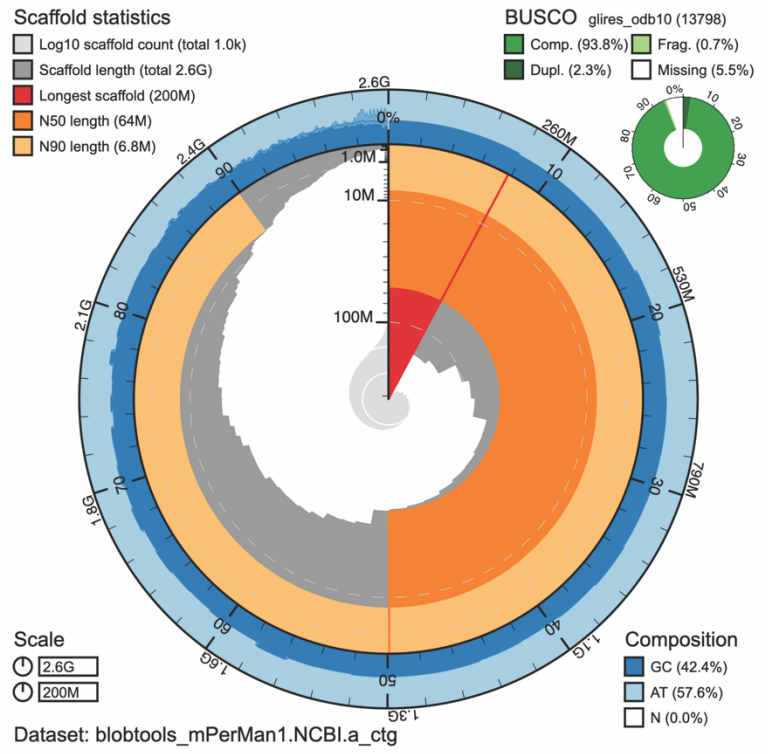
CHAPTER 6

Sequence data generated as a part of this study is available on NCBI under BioProject PRJNA993671. All sequenced individuals are specimens accessioned in the Museum of Vertebrate Zoology (University of California, Berkeley) and specimen identifiers are available in Supplementary Table 1. Additional metadata is available on the Arctos museum database at https://arctos.database.museum/search.cfm?guid_prefix=MVZ:Mamm.

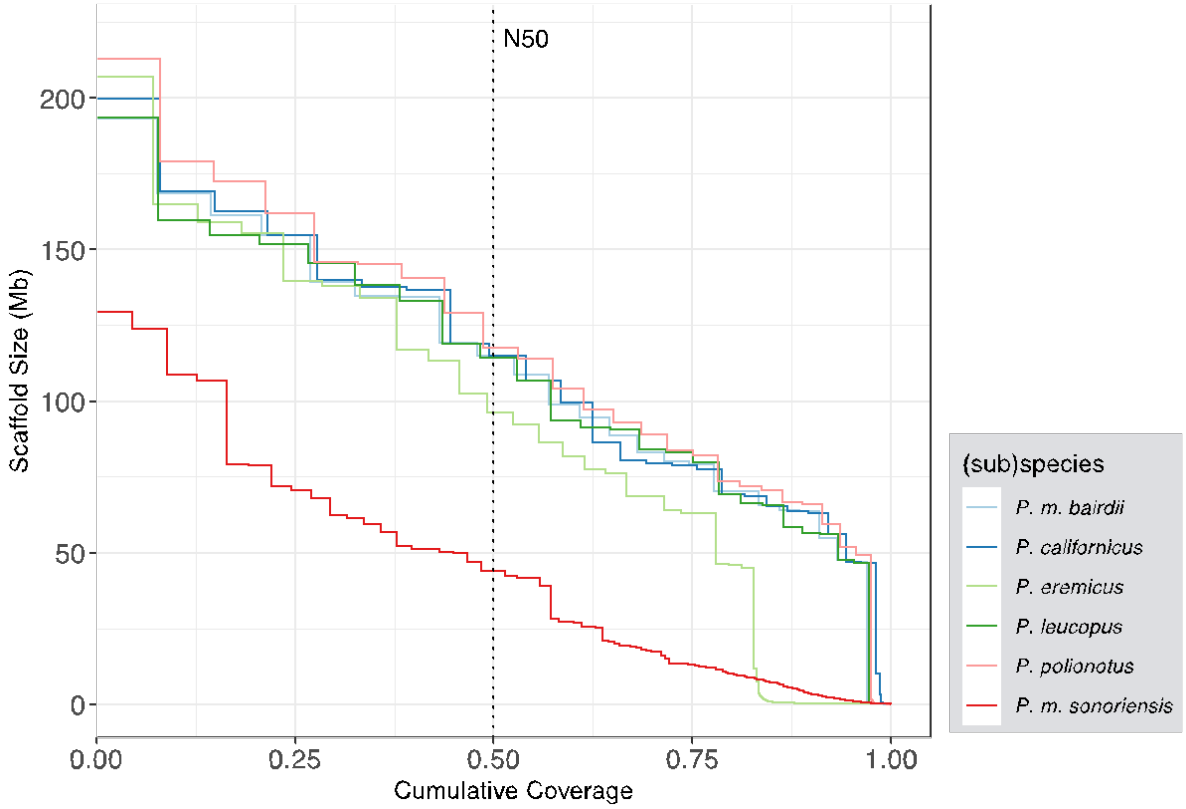
Appendix 2. Supplementary Materials for Chapter 1

Supplementary Table 1. Total content of interspersed repeats in the genome assembly grouped by major superfamilies, as determined by multiple rounds of masking with repeat sequences predicted *de novo* and available in curated databases. TEs nested within other TEs of the same type were not counted.

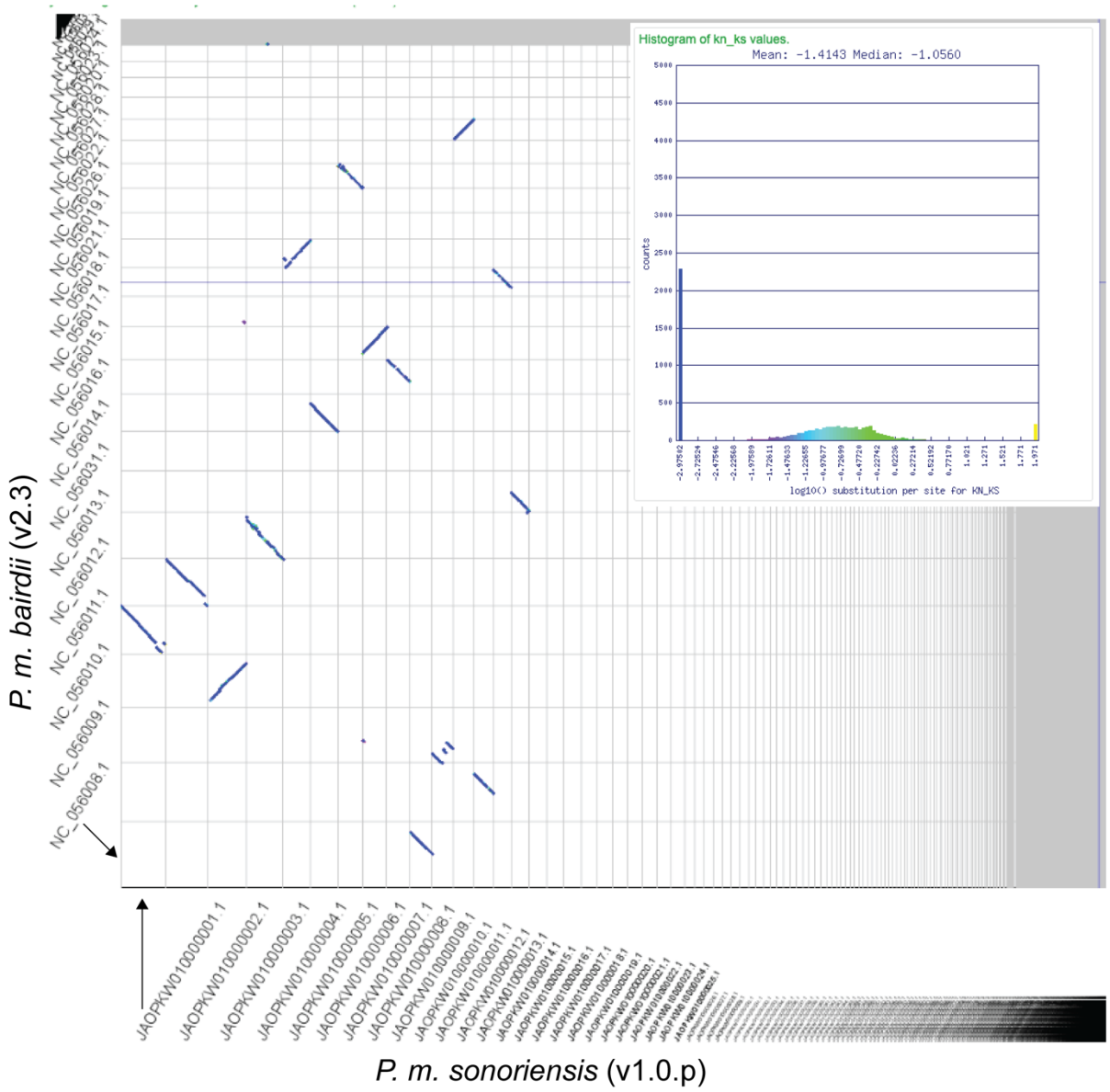
Class	Order	Base pairs	% Genome
Class I	LTR	607,403,844	21.3
	LINE	364,342,067	12.8
	SINE	339,317,730	11.9
	Other/Unknown	31,939,731	1.12
Class II		25,345,665	0.89
Simple		99,085,055	0.09
Unclassified		9,334,320	0.33
TOTAL		1,476,768,412	51.86



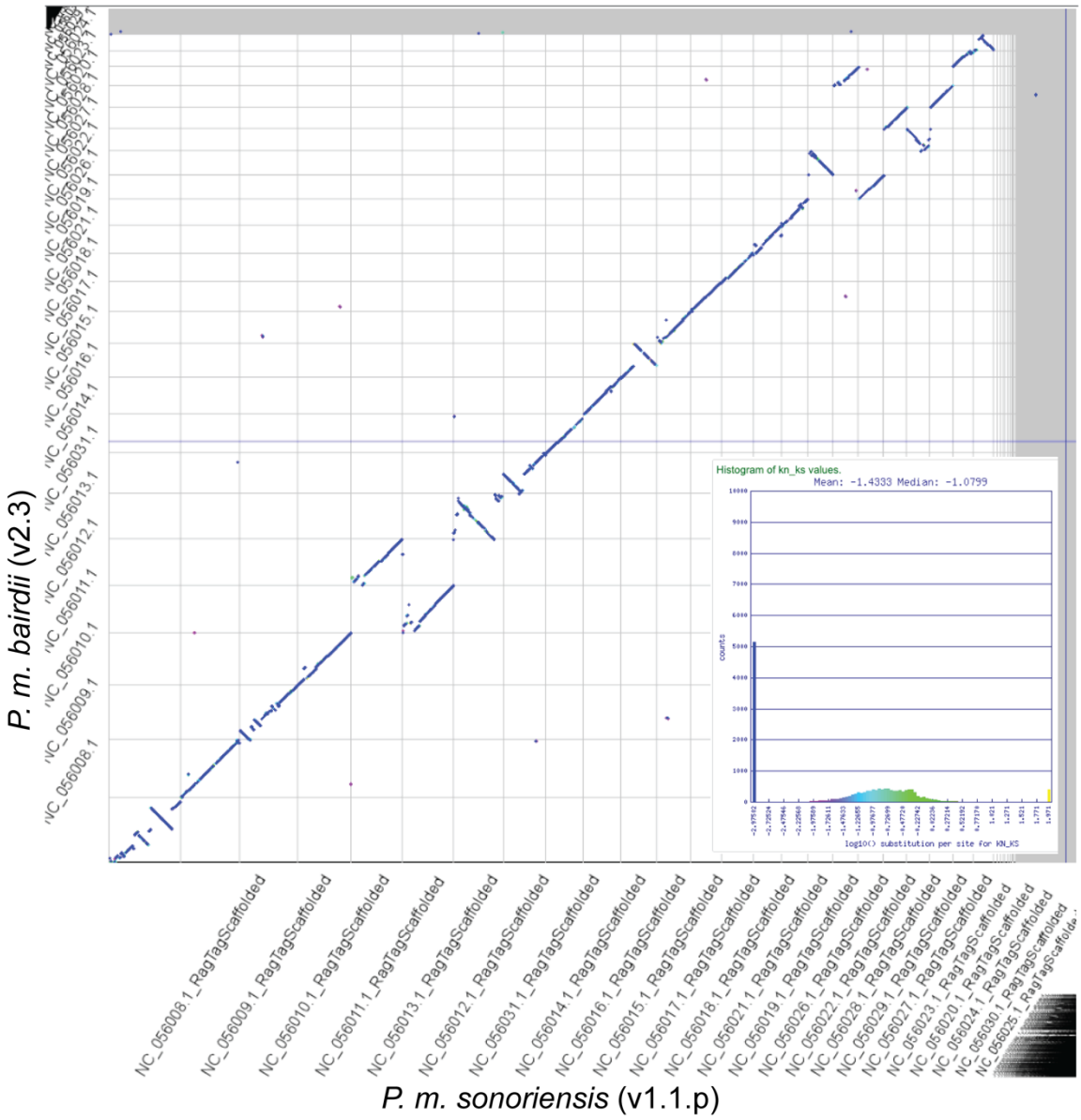
Supplementary Figure 1. Snail plot of quality metrics of the final alternate assembly (mPerLon1.1). The circumference represents the length of the assembly: scaffolds are drawn clockwise in order of size, and the red line indicates the longest. The middle arcs represent N50 (dark orange) and N90 (light orange). Completeness of BUSCO core gene set assembly is shown in the top right panel.



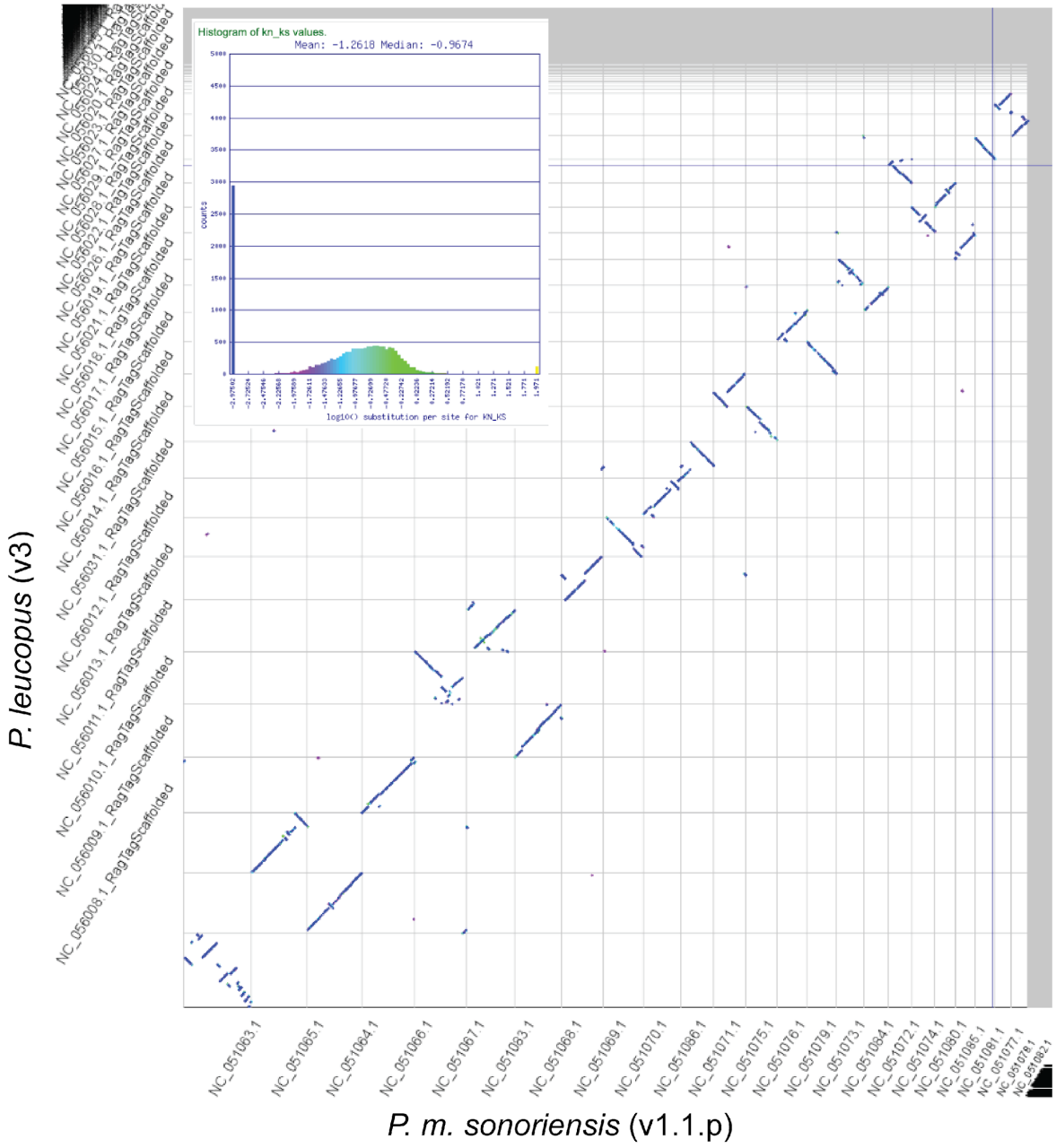
Supplementary Figure 2. An Nx50 plot shows the accumulation of genome coverage by the assembly, arranging the scaffolds from the longest to the shortest along the *x* axis. Included are the two best assemblies of *P. maniculatus*, and the scaffolded assemblies of other *Peromyscus* species available publicly in June 2023.



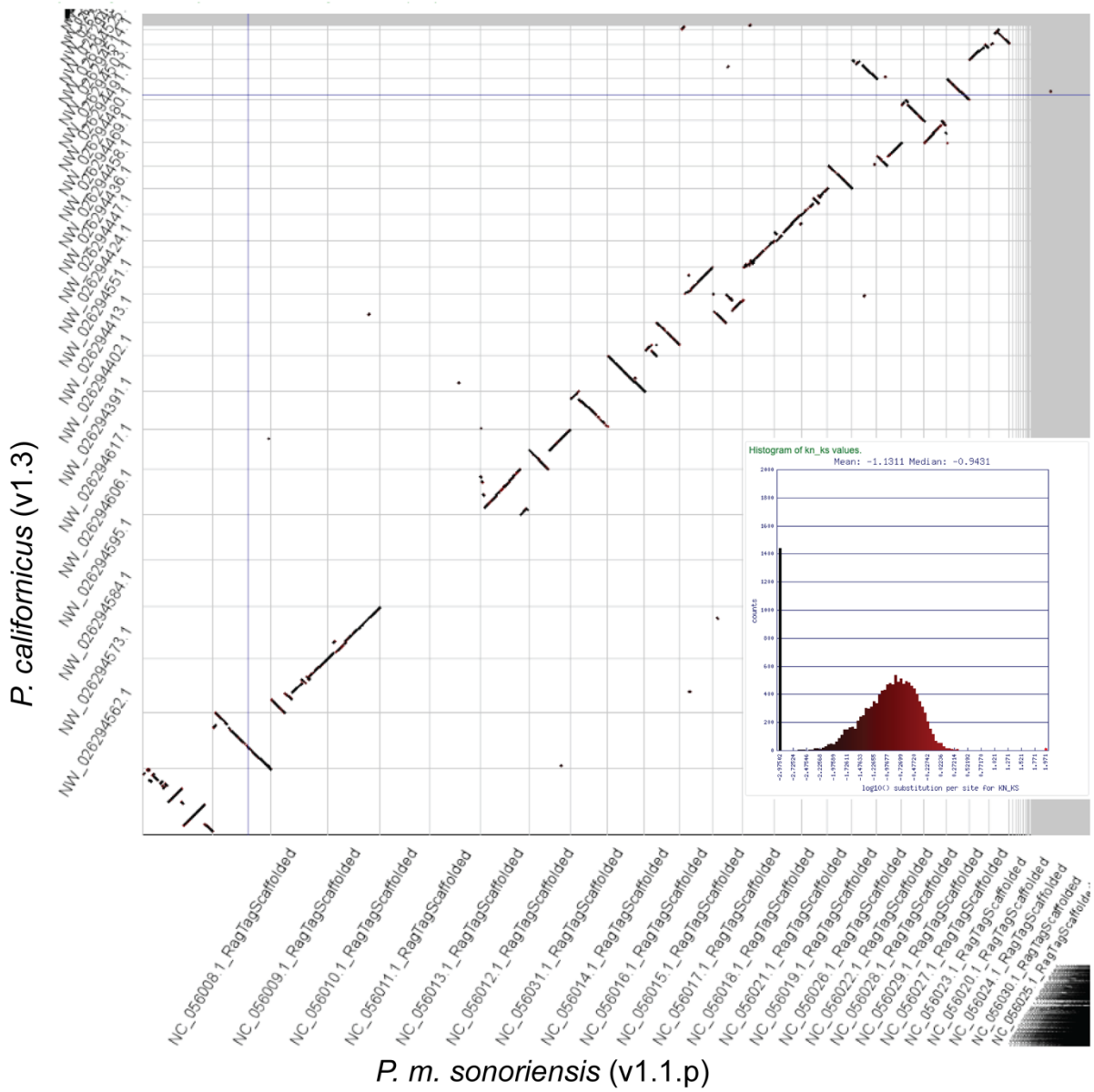
Supplementary Figure 3. A LAST whole genome alignment of the *P. m. sonoriensis* assembly v 1.0 (scaffolded only; GCA_026229955.1) versus that of *P. m. bairdii* v2.3. The non-linear distribution appearance reflects the fact that the *P. m. sonoriensis* scaffolds are not assigned to chromosomes in v1.0. Insert: the distribution of K_n/K_s ratios of individual CDS alignments.



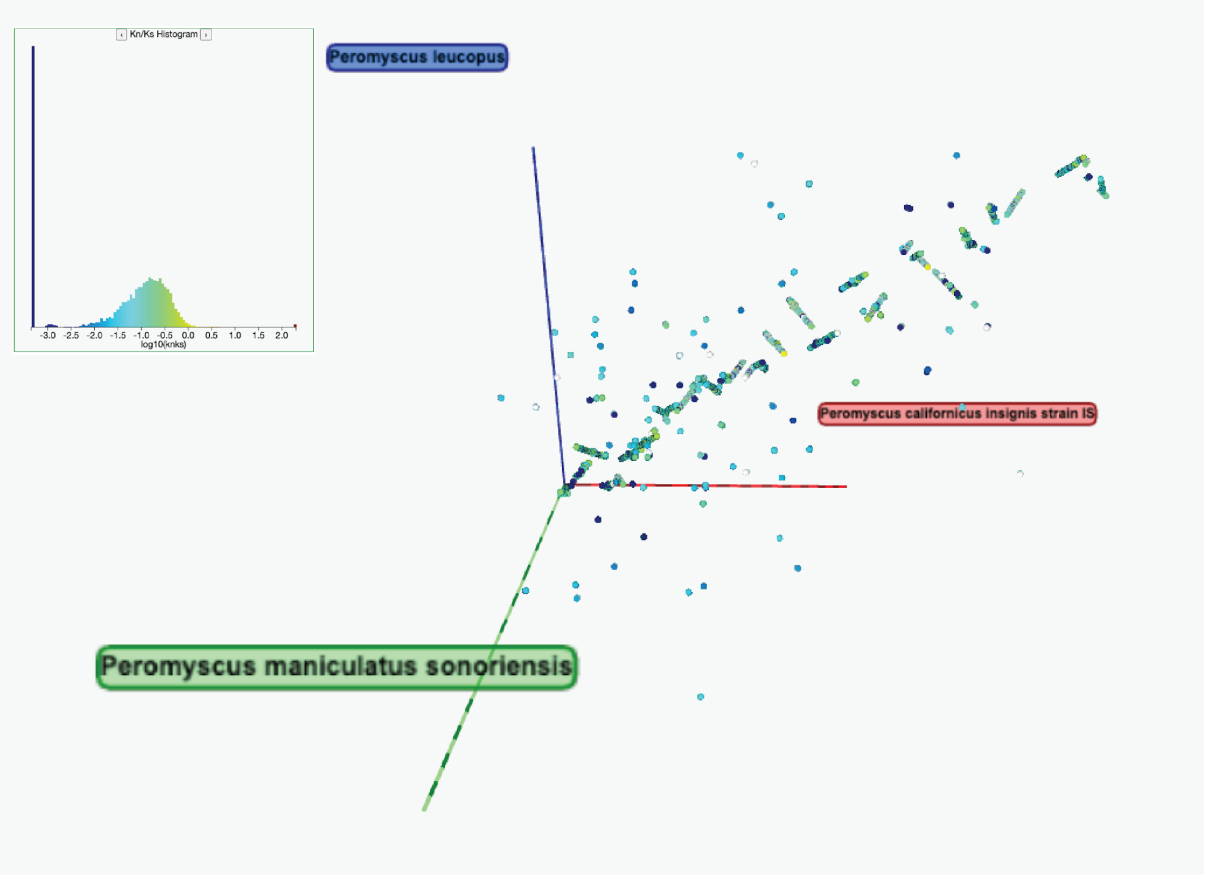
Supplementary Figure 4. A whole genome LAST alignment of the *P. m. sonoriensis* assembly v 1.1 (chromosomal) versus that of *P. m. bairdii* v2.3. Insert: the distribution of K_n/K_s ratios of individual CDS alignments.



Supplementary Figure 5. A whole genome LAST alignment of the *P. m. sonoriensis* assembly v 1.1 (chromosomal) versus that of *P. leucopus* v3. Insert: the distribution of K_n/K_s ratios of individual CDS alignments.

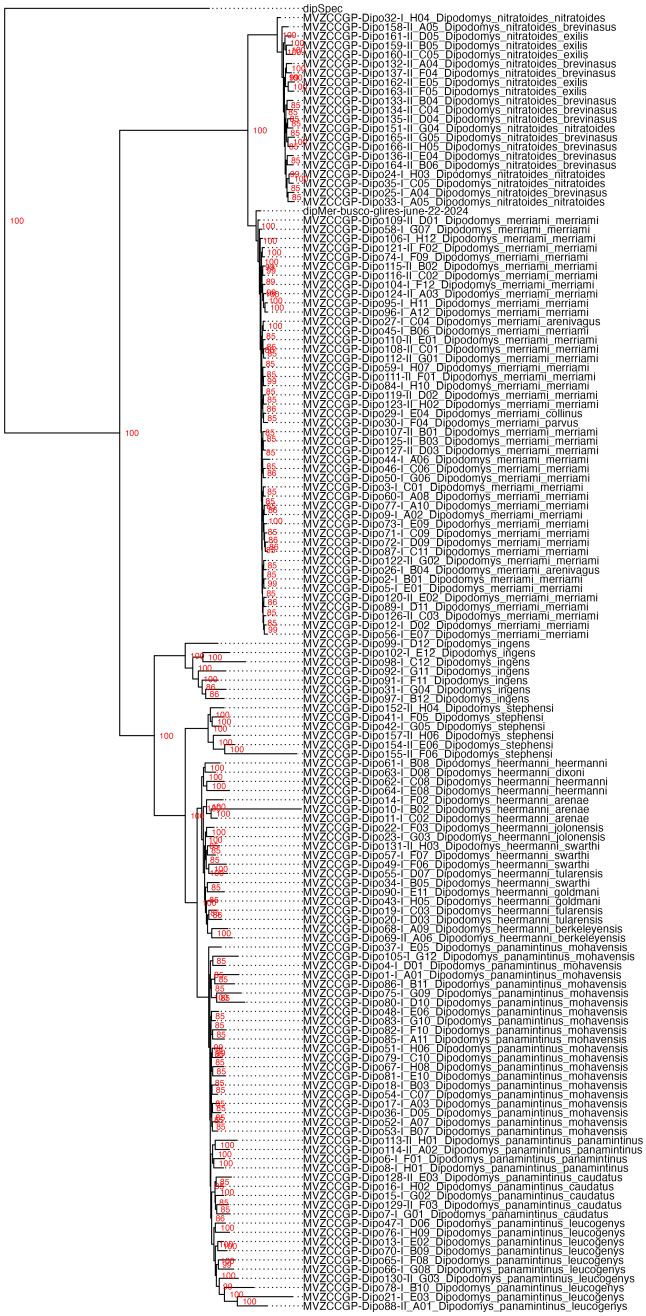


Supplementary Figure 6. A whole genome LAST alignment of the *P. m. sonoriensis* assembly v 1.1 (chromosomal) versus that of *P. californicus insignis* v1.3. Insert: the distribution of K_n/K_s ratios of individual CDS alignments. The plot can be replicated and viewed interactively at https://genomevolution.org/r/1pdyy.

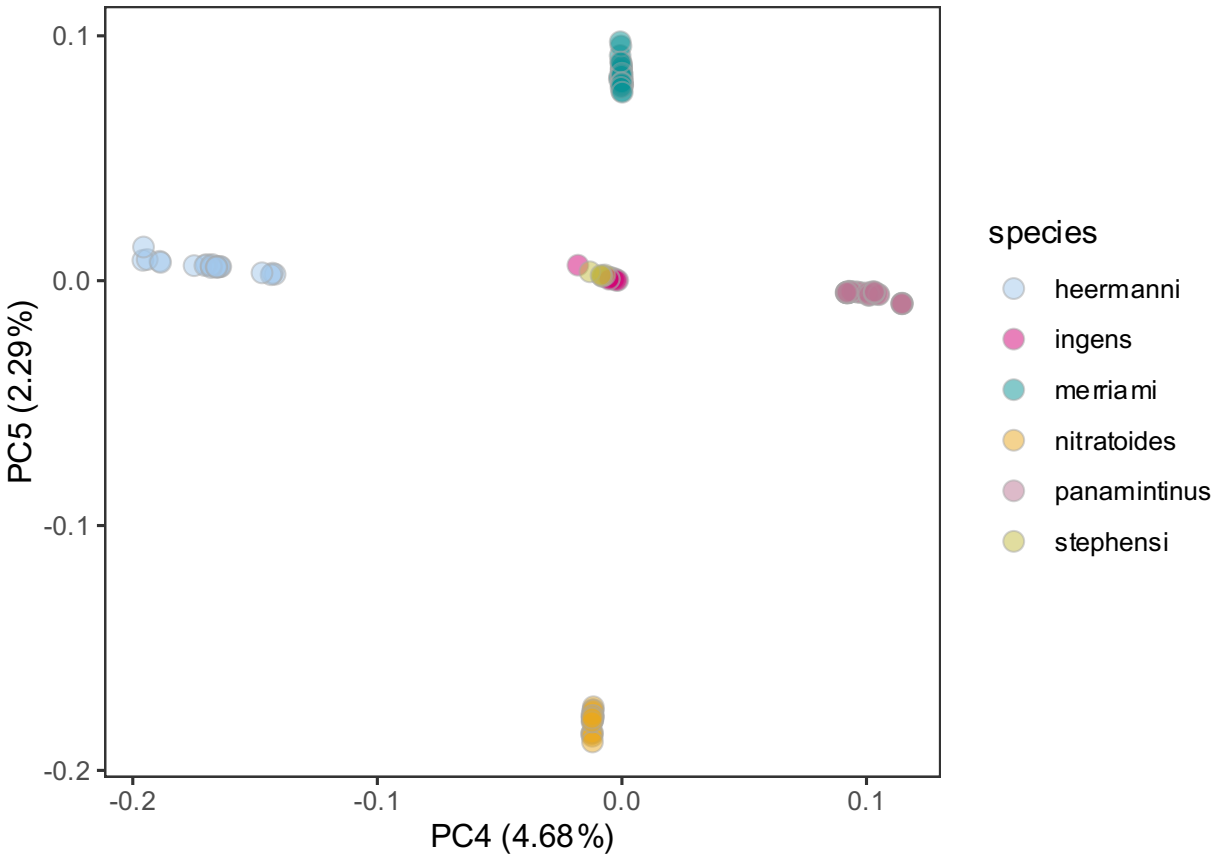


Supplementary Figure 7. A three-dimensional projection of synteny between the chromosomal assemblies of *P. maniculatus sonoriensis* (v1.1), *P. leucopus*, and *P. californicus insignis*. Individual dots represent LAST alignments of 13048 CDS genes, colored by their K_n/K_s ratios (insert).

Appendix 3. Supplementary Figures for Chapter 4.

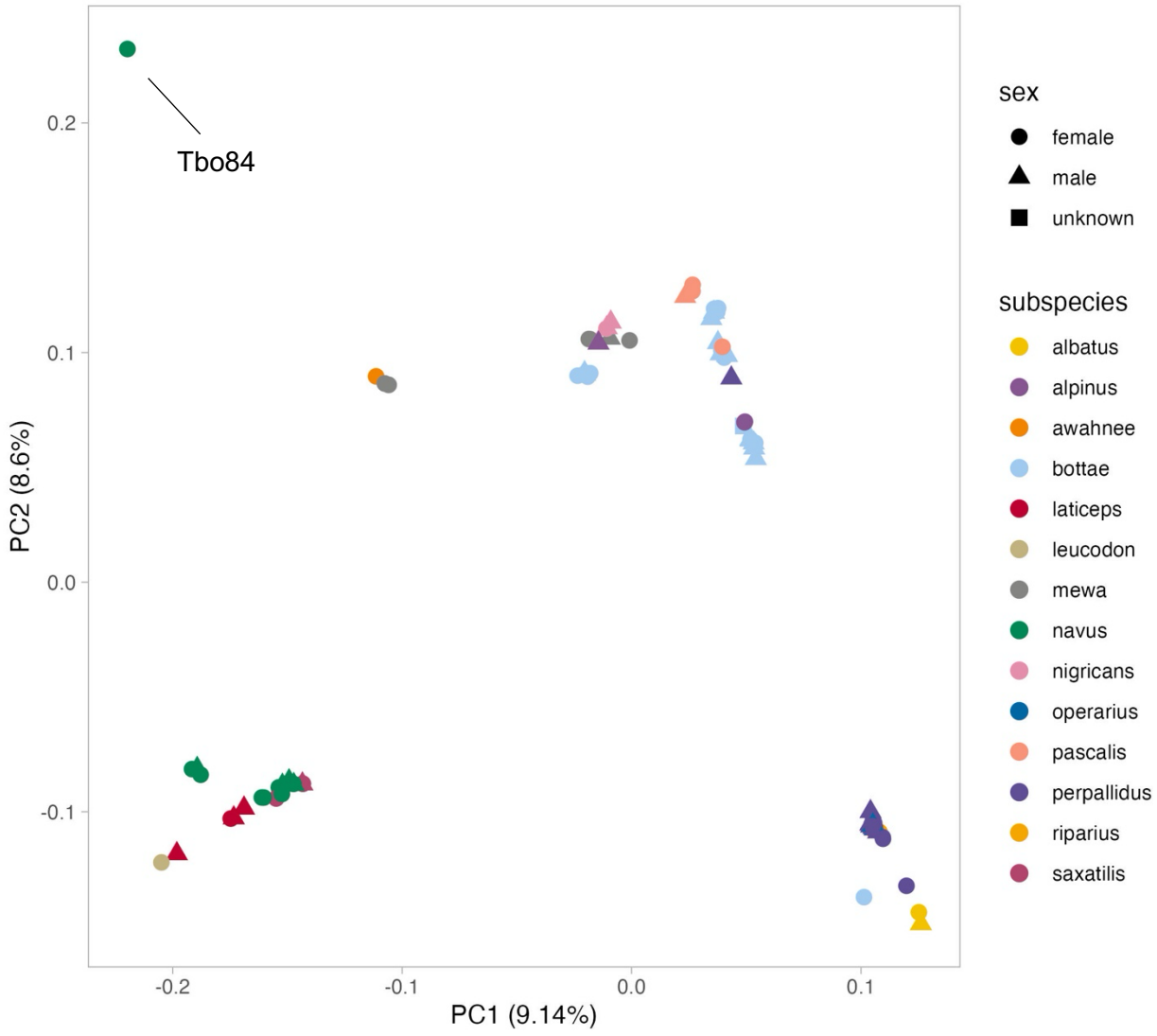


Supplementary Figure 1. Full phylogenetic tree created with 142 *Dipodomys* kangaroo rats with 2,159 BUSCO genes.

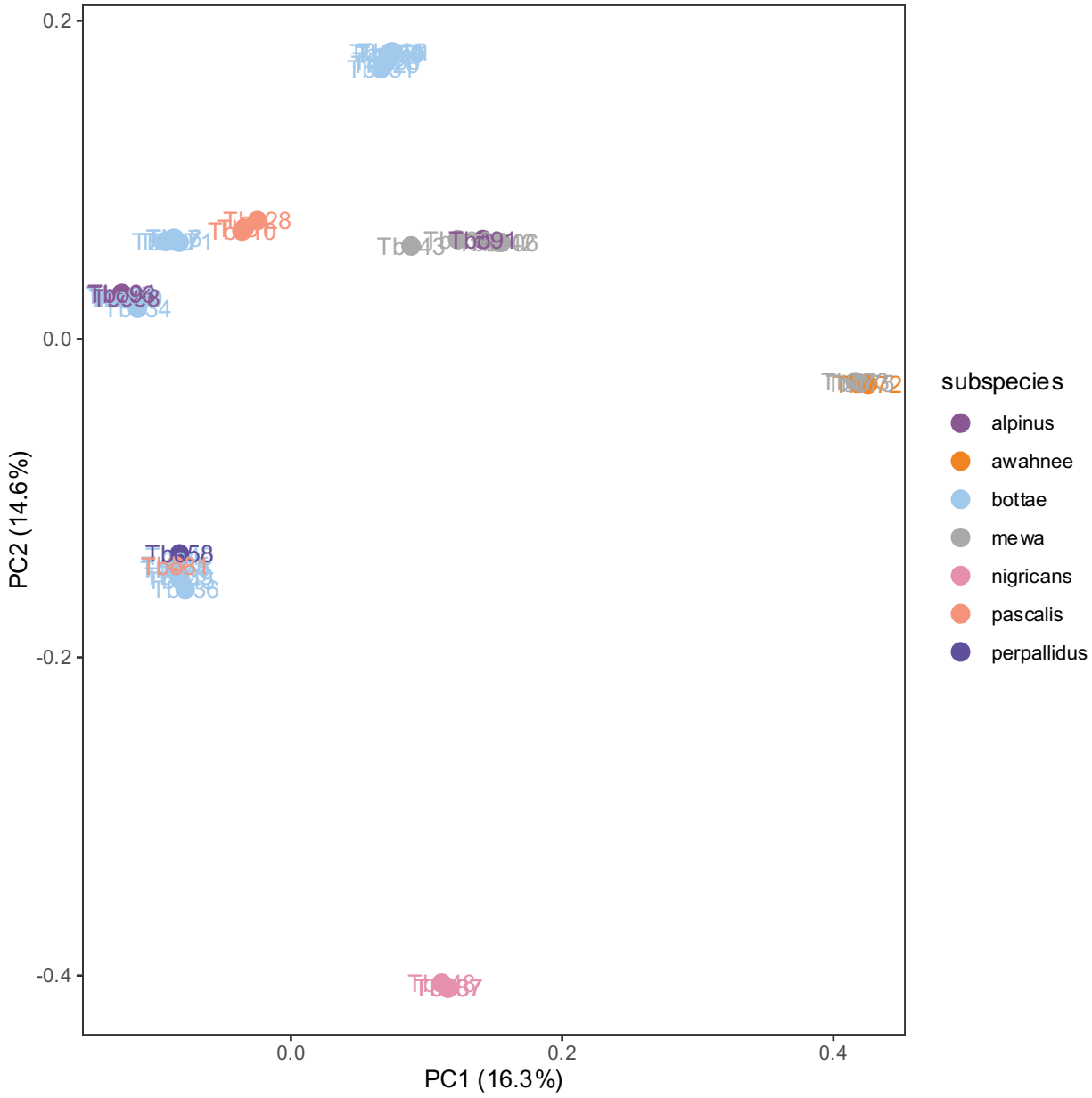


Supplementary Figure 2. Principal components 4 (4.68% variance explained) and 5 (2.29% variance explained) from a principal components analysis based on 5,502,172 linkage-pruned SNPs separate closely related species pairs into individual clusters.

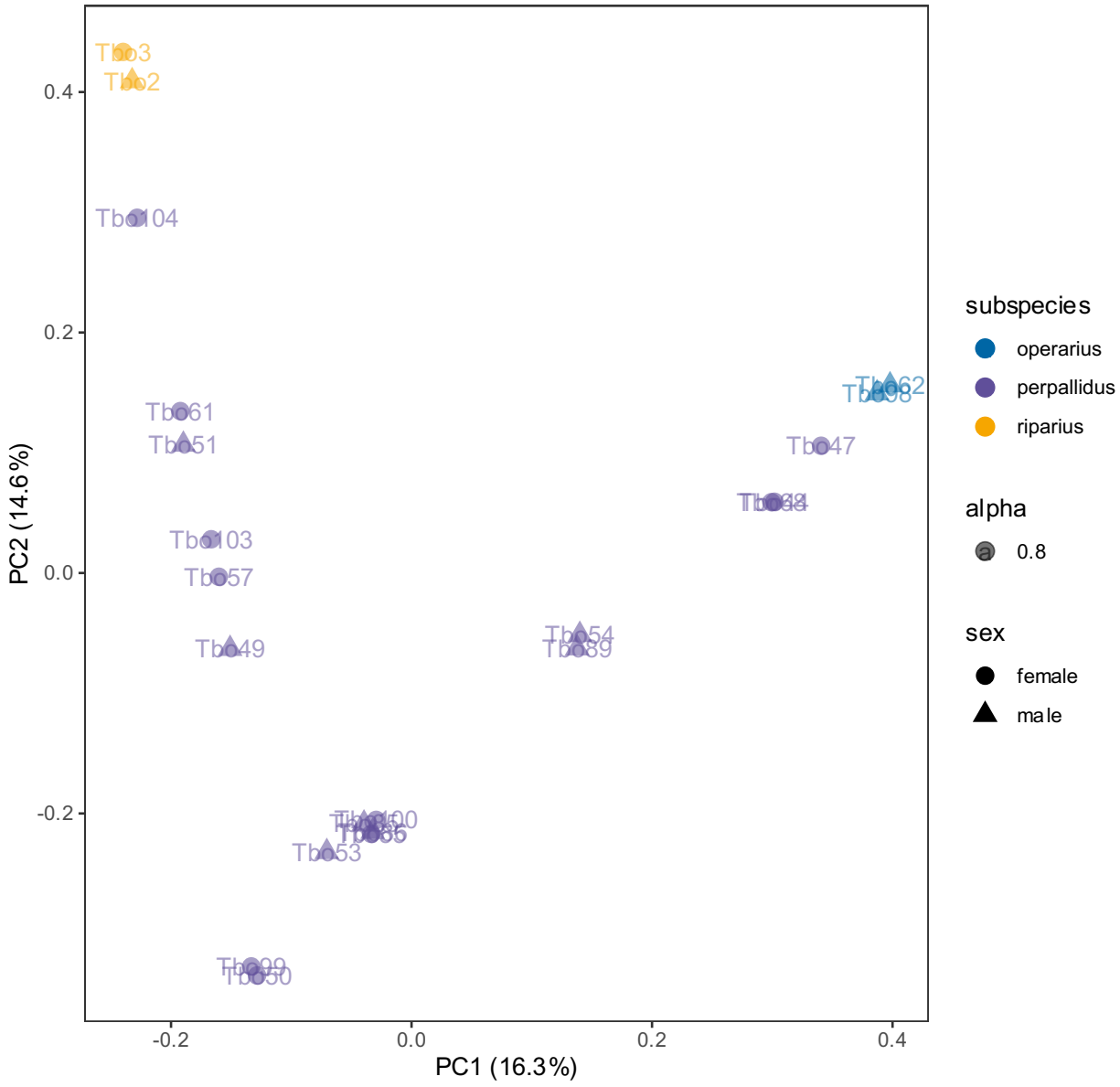
Appendix 4. Supplementary Figures for Chapter 6.



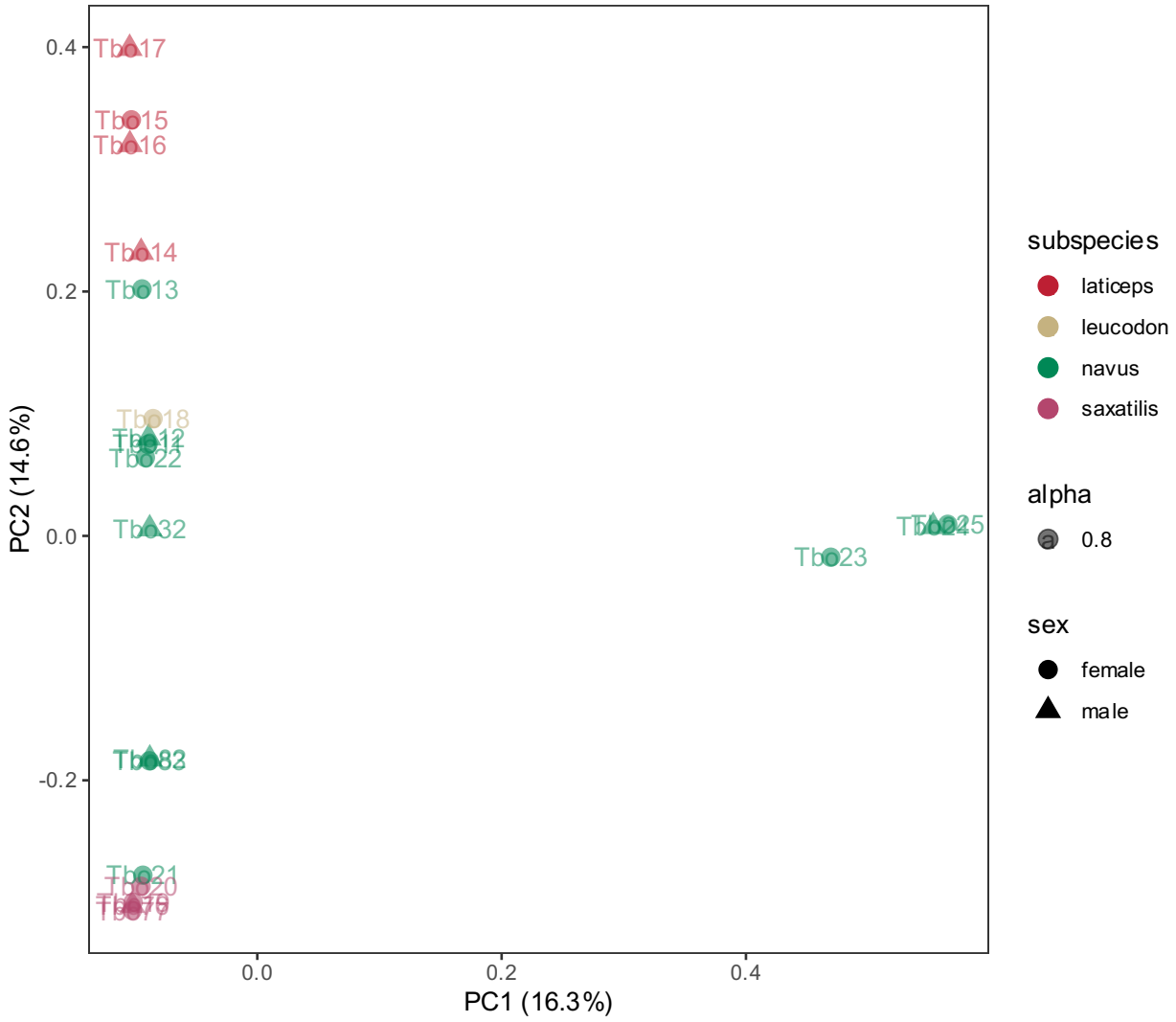
Supplementary Figure 1. Principal components analysis with 5,766,144 LD-pruned SNPs retained from whole genome sequencing data for 96 *Thomomys bottae* individuals sampled from across California. Sample Tbo84 in the upper left-hand corner was discarded from further analyses and may represent an individual of hybrid *T. bottae – T. townsendii* ancestry. Sample metadata and museum specimen accession numbers are available in Supplementary Table 1.



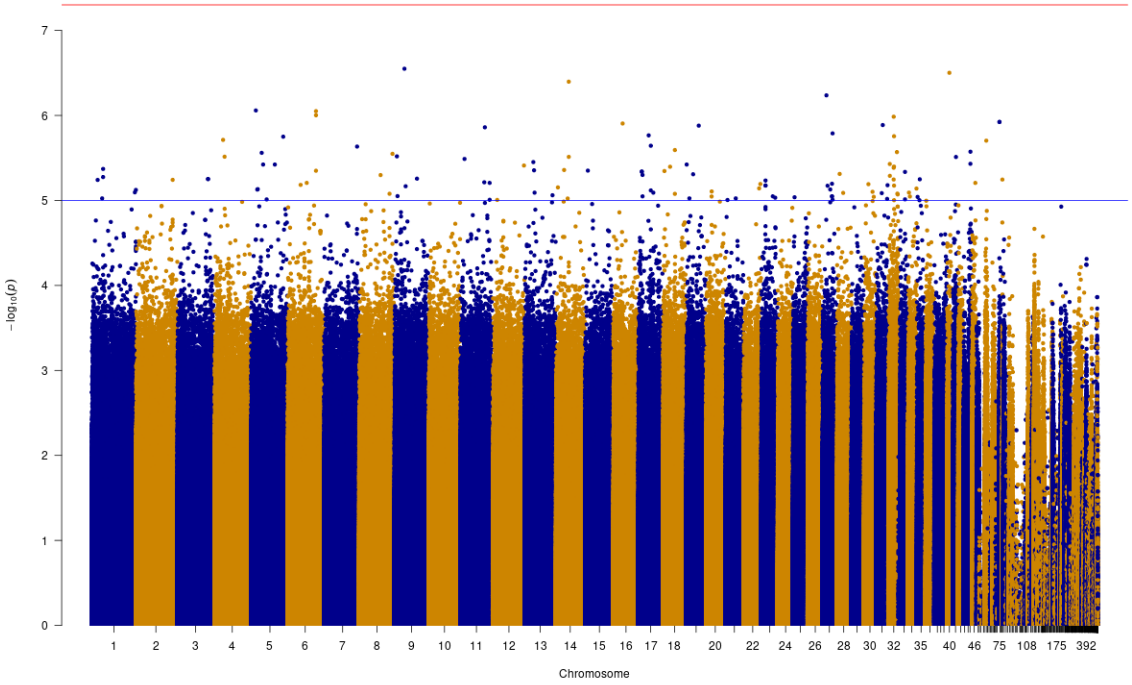
Supplementary Figure 2. Central California unit principal components analysis.



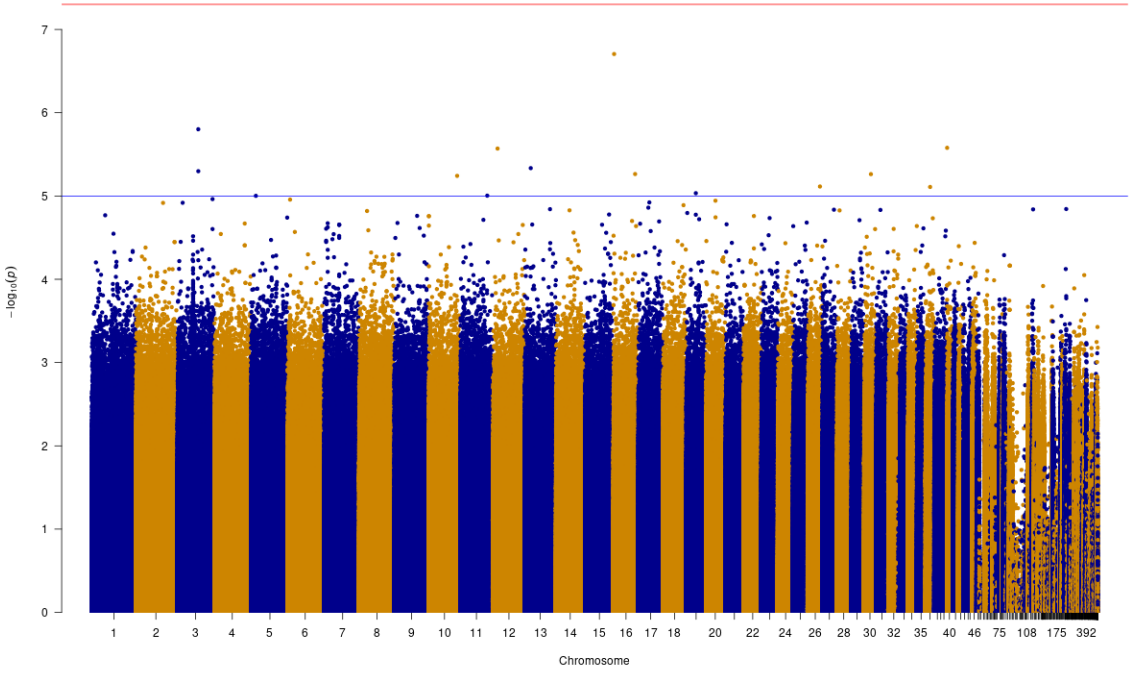
Supplementary Figure 3. Great Basin unit principal components analysis.



Supplementary Figure 4. Northern California unit principal components analysis.



Supplementary Figure 5. GWAS – A*.



Supplementary Figure 6. GWAS – B*.