RESEARCH ARTICLE SUMMARY

ZOONOMIA

The contribution of historical processes to contemporary extinction risk in placental mammals

Aryn P. Wilder*†, Megan A. Supple*†, Ayshwarya Subramanian, Anish Mudide, Ross Swofford, Aitor Serres-Armero, Cynthia Steiner, Klaus-Peter Koepfli, Diane P. Genereux, Elinor K. Karlsson, Kerstin Lindblad-Toh, Tomas Marques-Bonet, Violeta Munoz Fuentes, Kathleen Foley, Wynn K. Meyer, Zoonomia Consortium, Oliver A. Ryder*‡, Beth Shapiro*‡

INTRODUCTION: The Anthropocene is marked by an accelerated loss of biodiversity, widespread population declines, and a global conservation crisis. Given limited resources for conservation intervention, an approach is needed to identify threatened species from among the thousands lacking adequate information for status assessments. Such prioritization for intervention could come from genome sequence data, as genomes contain information about demography, diversity, fitness, and adaptive potential. However, the relevance of genomic data for identifying at-risk species is uncertain, in part because genetic variation may reflect past events and life histories better than contemporary conservation status.

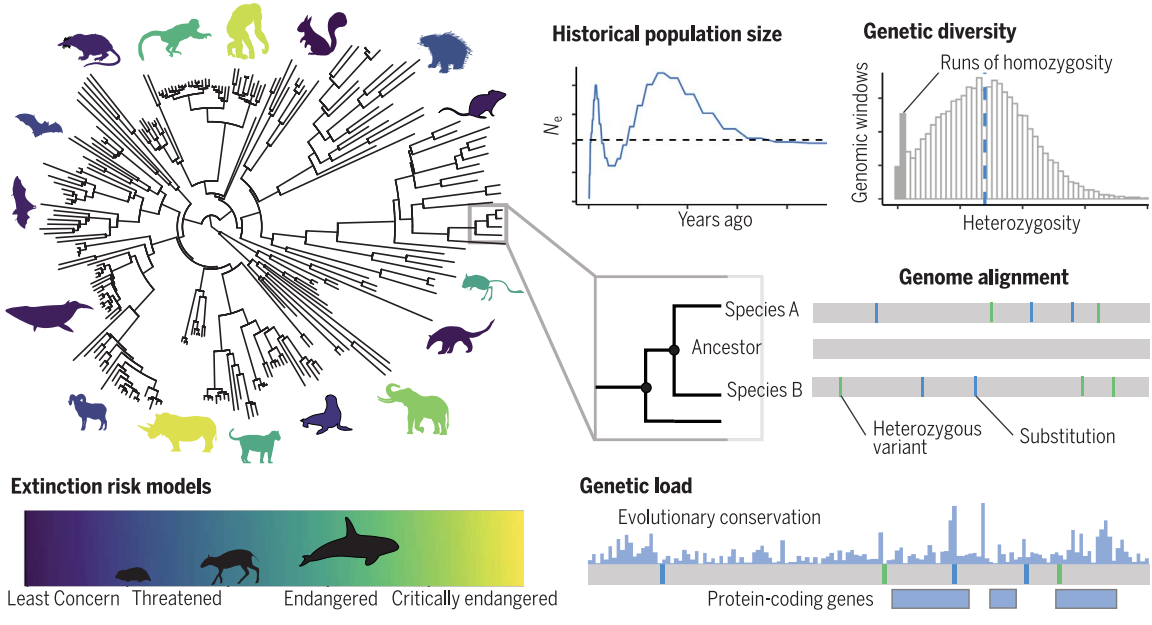
RATIONALE: The Zoonomia multispecies alignment presents an opportunity to systematically compare neutral and functional genomic diversity and their relationships to contemporary extinction risk across a large sample of diverse mammalian taxa. We surveyed 240 species spanning from the "Least Concern" to "Critically Endangered" categories, as published in the International Union for Conservation of Nature's Red List of Threatened Species. Using a single genome for each species, we estimated historical effective population sizes (N_e) and distributions of genome-wide heterozygosity. To estimate genetic load, we identified substitutions relative to reconstructed ancestral sequences, assuming that mutations at evolutionarily conserved sites and in protein-coding sequences, especially in genes essential for viability in mice, are predominantly deleterious. We examined relationships between the conservation status of species and metrics of heterozygosity, demography, and genetic load and used these data to train and test models to distinguish threatened from nonthreatened species.

RESULTS: Species with smaller historical N_e are more likely to be categorized as at risk of extinction, suggesting that demography, even from periods more than 10,000 years in the past, may be informative of contemporary resilience. Species with smaller historical N_e also carry proportionally higher burdens of weakly and moderately deleterious alleles, consistent with theoretical expectations of long-term accumulation and fixation of genetic load under strong genetic drift. We found weak support for a causative link between fixed drift load and extinction risk; however, other types of genetic load not captured in our data, such as rare, highly deleterious alleles, may also play a role. Although ecological (e.g., physiological, life-history, and behavioral) variables were the best predictors of extinction risk, genomic variables nonrandomly distinguished threatened from nonthreatened species in regression and machine learning models. These results suggest that information encoded within even a single genome can provide a risk assessment in the absence of adequate ecological or population census data.

CONCLUSION: Our analysis highlights the potential for genomic data to rapidly and inexpensively gauge extinction risk by leveraging relationships between contemporary conservation status and genetic variation shaped by the long-term demographic history of species. As more resequencing data and additional reference genomes become available, estimates of genetic load, estimates of recent demographic history, and accuracy of predictive models will improve. We therefore echo calls for including genomic information in assessments of the conservation status of species. ■

The list of author affiliations is available in the full article online.
*Corresponding author. Email: awilder@sdzwa.org (A.P.W.); msupple@ucsc.edu (M.A.S.); oryder@sdzwa.org (O.A.R.); bashapir@ucsc.edu (B.S.)
†These authors contributed equally to this work.
‡These authors contributed equally to this work.
Cite this article as A. P. Wilder *et al.*, *Science* **380**, eabn5856 (2023). DOI: 10.1126/science.abn5856

READ THE FULL ARTICLE AT
https://doi.org/10.1126/science.abn5856



Genomic information can help predict extinction risk in diverse mammalian species. Across 240 mammals, species with smaller historical N_e had lower genetic diversity, higher genetic load, and were more likely to be threatened with extinction. Genomic data were used to train models that predict whether a species is threatened, which can be valuable for assessing extinction risk in species lacking ecological or census data. [Animal silhouettes are from PhyloPic]

RESEARCH ARTICLE

ZOONOMIA

The contribution of historical processes to contemporary extinction risk in placental mammals

Aryn P. Wilder[1]*†, Megan A. Supple[2,3]*†, Ayshwarya Subramanian[4], Anish Mudide[5], Ross Swofford[4], Aitor Serres-Armero[6], Cynthia Steiner[1], Klaus-Peter Koepfli[7,8,9], Diane P. Genereux[4], Elinor K. Karlsson[4,10], Kerstin Lindblad-Toh[4,11], Tomas Marques-Bonet[6,12,13,14], Violeta Munoz Fuentes[15], Kathleen Foley[16,17], Wynn K. Meyer[17], Zoonomia Consortium‡, Oliver A. Ryder[1,18]*§, Beth Shapiro[2,3]*§

Species persistence can be influenced by the amount, type, and distribution of diversity across the genome, suggesting a potential relationship between historical demography and resilience. In this study, we surveyed genetic variation across single genomes of 240 mammals that compose the Zoonomia alignment to evaluate how historical effective population size (N_e) affects heterozygosity and deleterious genetic load and how these factors may contribute to extinction risk. We find that species with smaller historical N_e carry a proportionally larger burden of deleterious alleles owing to long-term accumulation and fixation of genetic load and have a higher risk of extinction. This suggests that historical demography can inform contemporary resilience. Models that included genomic data were predictive of species' conservation status, suggesting that, in the absence of adequate census or ecological data, genomic information may provide an initial risk assessment.

The current rate of biodiversity loss amounts to a sixth mass extinction (*1*) and is compounded by substantial population declines across nearly one-third of vertebrate species (*2*). Many species need immediate conservation intervention, but identifying them from the >20,000 species currently categorized as "Data Deficient" by the International Union for Conservation of Nature (IUCN) is a challenge. Fortunately, genomic data, which are increasingly available for a broad taxonomic range of species, may hold promise for helping to identify at-risk species by providing readily accessible information on demography and fitness-relevant genetic variation (*3*, *4*). It remains poorly explored, however, to what extent genomic data on their own are sufficient to help triage endangered species for conservation intervention.

Population genetic diversity and individual heterozygosity are long-recognized correlates of fitness-relevant functional variation (*5*, *6*). Our previous analysis of 124 placental mammalian genomes showed that lower heterozygosity and increased stretches of homozygosity are more common in species in threatened IUCN Red List categories (*7*). However, functional diversity, including estimates of adaptive variation and deleterious genetic load, may also be useful correlates of population resiliency. Such measures are increasingly accessible with emerging genomic tools (*8*) and comparative genomics resources such as the Zoonomia alignment of placental mammalian genomes (table S1) (*7*). The Zoonomia alignment provides high-resolution constraint scores and reconstructed ancestral sequences that can help to identify deleterious alleles at functionally important sites (*7*, *9*).

In this study, we surveyed the distribution of neutral and functional genetic variation across 240 species in the Zoonomia alignment to determine how historical effective population sizes (N_e) have influenced heterozygosity and deleterious genetic load (fig. S1). We tested the value of genomic data to more precisely target species for conservation efforts by comparing the outcome of predictive models of conservation status that use ecological data, genomic data, or both. While we acknowledge the limitations of assuming that single genomes are representative of an entire species, our approach capitalizes on the singular resource provided by the Zoonomia Consortium to explore whether genomic data can provide initial risk assessments that may be useful to triage data-deficient species and guide resource allocation for conservation intervention.

Historical population size is relevant to contemporary extinction risk

Species with historically small N_e tend to be classified into threatened IUCN Red List categories (Fig. 1). Species classified as "Near Threatened" (NT), "Vulnerable" (VU), "Endangered" (EN), or "Critically Endangered" (CR) had significantly smaller harmonic mean N_e ($\text{mean}_{\text{threatened}}$ = 18,950) compared with nonthreatened species ["Least Concern" (LC); $\text{mean}_{\text{nonthreatened}}$ = 27,839; $P < 3.3 \times 10^{-5}$ when accounting for relationships across the phylogeny; Fig. 1B and fig. S2]. N_e was also significantly smaller in threatened species than in nonthreatened species within two of three taxonomic orders with sufficient numbers of species to test (Cetartiodactyla: $\text{mean}_{\text{threatened}}$ = 18,336, $\text{mean}_{\text{nonthreatened}}$ = 22,648, P = 0.023; and Carnivora: $\text{mean}_{\text{threatened}}$ = 9636, $\text{mean}_{\text{nonthreatened}}$ = 26,195, $P = 2.4 \times 10^{-5}$; but not Primates: $\text{mean}_{\text{threatened}}$ = 22,508, $\text{mean}_{\text{nonthreatened}}$ = 24,373, P = 0.31) (fig. S3). Within these two orders in particular, large-bodied herbivores and carnivores have declined in both geographic range and population size during the Anthropocene (*10*, *11*). Smaller populations are expected to have higher extinction risk, yet these historical N_e estimates reflect periods more than 10,000 years in the past, suggesting that long-term characteristics of ancestral populations can be informative about present-day population size and extinction risk. These results support the utility of metrics of genome-wide diversity in conservation assessments, a topic that is currently being debated (*12*, *13*).

Estimates of historical N_e can also identify previously large populations that have experienced contemporary declines. Specifically, if the estimate of historical N_e is large while the population census size (N_c) is small, this inflates the N_e/N_c ratio. In a study of pinnipeds, for example, most species that had undergone recent declines had smaller N_c than expected given their historical N_e (*14*). To test this hypothesis across the taxonomic range of the Zoonomia alignment, we examined the ratio of deep historical N_e to contemporary N_c for 89 species with population census information available in PanTHERIA (*15*). Species in

[1]Conservation Genetics, San Diego Zoo Wildlife Alliance, Escondido, CA 92027, USA. [2]Department of Ecology and Evolutionary Biology, University of California, Santa Cruz, CA 95064, USA. [3]Howard Hughes Medical Institute, University of California, Santa Cruz, CA 95064, USA. [4]Broad Institute of MIT and Harvard, Cambridge, MA 02139, USA. [5]Phillips Exeter Academy, Exeter, NH 03833, USA. [6]Institute of Evolutionary Biology, Department of Experimental and Health Sciences, Universitat Pompeu Fabra, Barcelona 08003, Spain. [7]Smithsonian-Mason School of Conservation, George Mason University, Front Royal, VA 22630, USA. [8]Center for Species Survival, Smithsonian Conservation Biology Institute, National Zoological Park, Washington, DC 30008, USA. [9]Computer Technologies Laboratory, ITMO University, St. Petersburg 197101, Russia. [10]Program in Bioinformatics and Integrative Biology, University of Massachusetts Medical School, Worcester, MA 01605, USA. [11]Science for Life Laboratory, Department of Medical Biochemistry and Microbiology, Uppsala University, Uppsala 751 32, Sweden. [12]Catalan Institution of Research and Advanced Studies, Barcelona 08010, Spain. [13]Centre for Genomic Regulation, Barcelona Institute of Science and Technology, Barcelona 08028, Spain. [14]Institut Català de Paleontologia Miquel Crusafont, Universitat Autònoma de Barcelona, Barcelona 08193, Spain. [15]European Molecular Biology Laboratory–European Bioinformatics Institute, Wellcome Genome Campus, Hinxton CB10 1SD, UK. [16]College of Law, University of Iowa, Iowa City, IA 52242, USA. [17]Department of Biological Sciences, Lehigh University, Bethlehem, PA 18015, USA. [18]Department of Evolution, Behavior and Ecology, Division of Biology, University of California, San Diego, La Jolla, CA 92039, USA.
***Corresponding author. Email: awilder@sdzwa.org (A.P.W.); msupple@ucsc.edu (M.A.S.); oryder@sdzwa.org (O.A.R.); bashapir@ucsc.edu (B.S.)**
†These authors contributed equally to this work.
‡Zoonomia Consortium collaborators and affiliations are listed at the end of this paper.
§These authors contributed equally to this work.

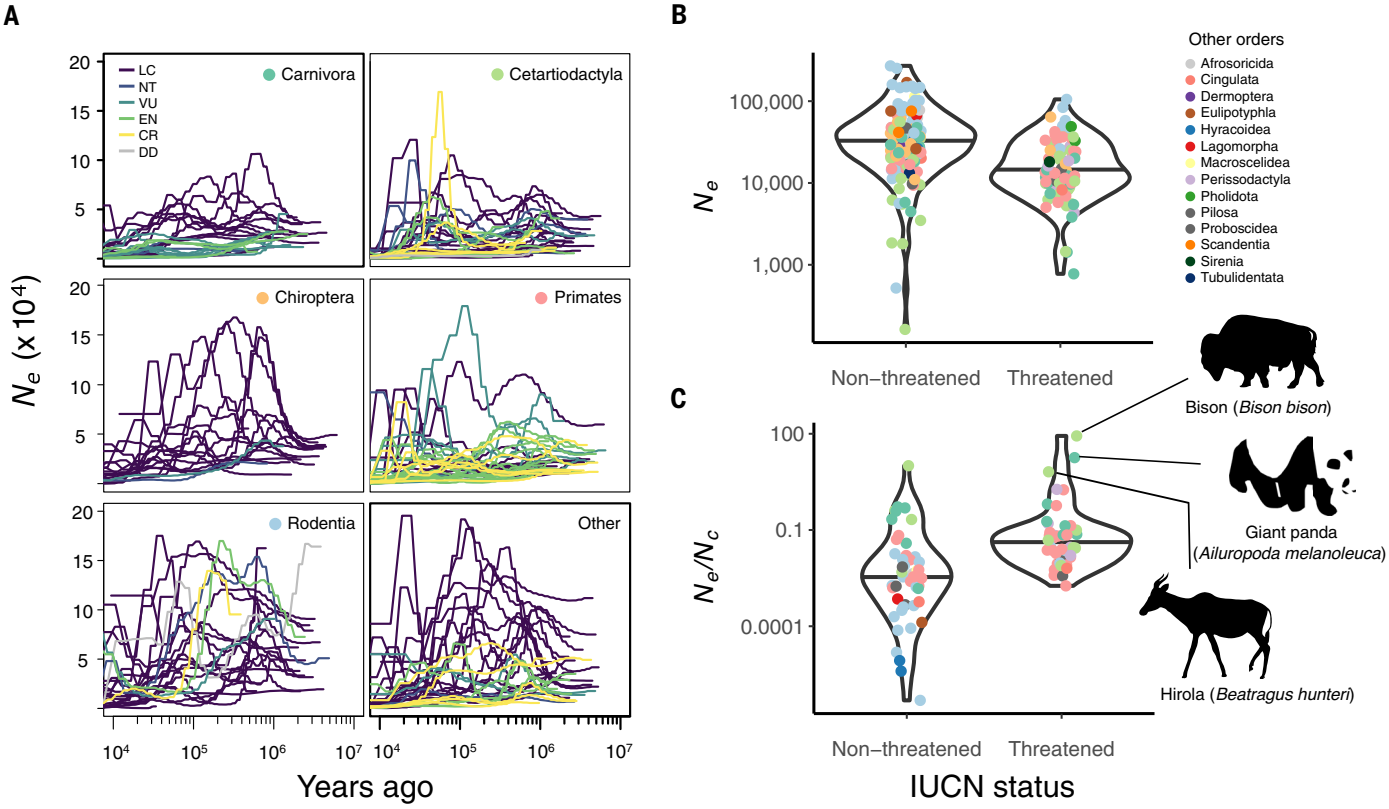


Fig. 1. Demographic history across mammalian orders and IUCN Red List categories. (**A**) Estimates of effective population sizes (N_e) over time, displayed by taxonomic order. Lines represent individual species, colored by IUCN status (LC, Least Concern; NT, Near Threatened; VU, Vulnerable; EN, Endangered; CR, Critically Endangered; DD, Data Deficient). Colored dots correspond to the taxonomic order of species depicted in (B) and (C). For visualization, only species with N_e estimates of <200,000 for every time point are shown. (**B**) Harmonic mean N_e was significantly lower in threatened IUCN categories relative to nonthreatened (phylolm, $P < 3.3 \times 10^{-5}$). (**C**) The ratio of historical N_e to contemporary census population size (N_e/N_c) can identify species with smaller N_c than expected from historical N_e (phylolm, $P = 0.012$). Points in (B) and (C) show individual species, colored by taxonomic order. [Animal silhouettes are from PhyloPic]

threatened IUCN categories had larger N_e/N_c ratios, that is, smaller contemporary N_c relative to historical N_e ($\text{mean}_{\text{threatened}} = 1.07 \times 10^{-3}$; $\text{mean}_{\text{nonthreatened}} = 4.29 \times 10^{-4}$; $P = 0.012$; Fig. 1C). The relationship was also significant within Primates (phylolm, $\text{mean}_{\text{threatened}} = 3.46 \times 10^{-3}$; $\text{mean}_{\text{nonthreatened}} = 1.11 \times 10^{-3}$; $P = 0.029$), the only order with available N_e/N_c estimates for a sufficient number of taxa in the two threat categories, indicating that the pattern holds among species with similar life-history traits. Across taxa, the largest N_e/N_c ratios included American bison (*Bison bison*), giant panda (*Ailuropoda melanoleuca*), and hirola (*Beatragus hunteri*), all of which have declined because of recent human activities (*16*–*18*).

Historically smaller populations carry proportionally larger burdens of genetic load

Historical N_e is correlated with the proportion of deleterious substitutions in mammalian genomes, reflecting the accumulation and fixation of genetic load over long evolutionary time periods. We called derived, single-nucleotide substitutions for each species relative to the reconstructed sequence of the nearest ancestral phylogenetic node and called heterozygous sites from short-read data mapped to the focal genome. We inferred the impacts of derived substitutions and heterozygous variants, assuming that mutations at sites that are conserved across taxa (phyloP > 2.27) (*9*) and nonsynonymous mutations are predominantly deleterious (fig. S1) (*19*). Assuming most substitutions are fixed and mutation rates are similar across the phylogeny (*20*, *21*), the proportion of substitutions that are deleterious should be correlated with the total number of fixed deleterious mutations in the genome. Deleterious substitutions should therefore largely reflect fixed drift load that reduces the mean fitness of the population, whereas heterozygous deleterious variants reflect segregating mutational load (*22*).

We found that species with smaller N_e had proportionally more substitutions at evolutionarily conserved sites genome-wide (phylolm, $P = 9.65 \times 10^{-3}$) and proportionally more missense substitutions in genes (phylolm, $P = 7.76 \times 10^{-5}$; fig. S4). PhyloP kurtosis, which describes the extreme phyloP outliers in the tail of the distribution across substitutions, was positively correlated with N_e (phylolm, $P = 0.014$). This correlation means that species with smaller N_e had smaller right tails and therefore fewer substitutions at extremely conserved sites. To further parse potential fitness impacts of mutations in protein-coding regions, we examined genes with associated viability phenotypes in single-gene knockout mouse lines classified by the International Mouse Phenotyping Consortium (IMPC), assuming that, when aggregated across many genes, viability classifications are correlated to their fitness impacts in other species (*23*). Species with smaller N_e had proportionally more missense mutations relative to coding mutations in nearly all categories (phylolm, $P < 3.00 \times 10^{-5}$; Fig. 2 and figs. S5 and S6). We observed proportionally fewer missense mutations in IMPC lethal genes relative to IMPC viable genes (analysis of variance, $P < 4.42 \times 10^{-9}$; fig. S7), reflecting stronger purifying selection in the lethal gene class, but the negative correlation was nonetheless consistent for both lethal and viable categories (Fig. 2). This relationship supports theoretical predictions that smaller populations experiencing strong drift accumulate and

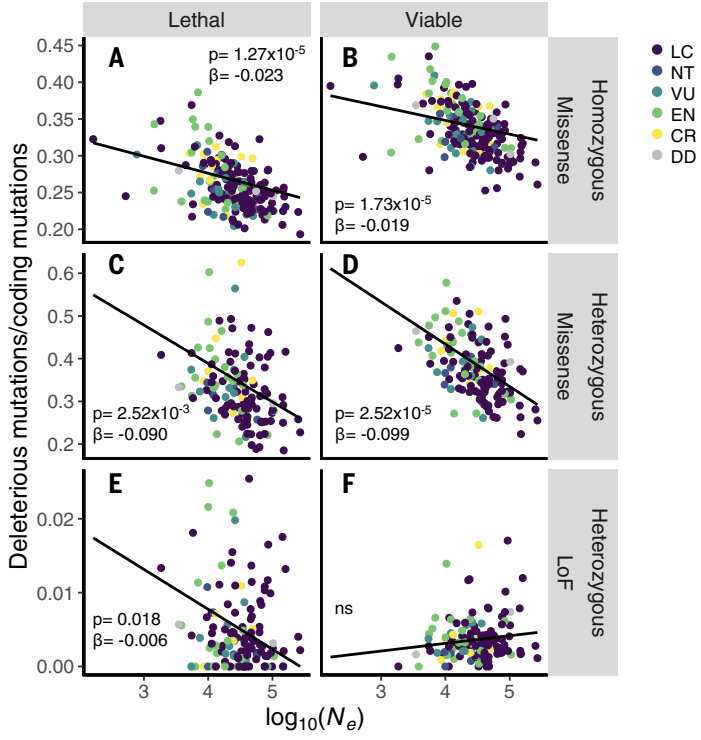


Fig. 2. Historically small populations have higher deleterious genetic load in protein-coding genes. Proportion of homozygous missense substitutions (**A** and **B**), heterozygous missense variants (**C** and **D**), and heterozygous loss-of-function (LoF) variants (**E** and **F**) in genes as a function of historical N_e across species. Genes were classified by associated lethal or viable phenotypes in knockout mice. Proportions of heterozygous and homozygous missense mutations were negatively correlated with N_e (all $P < 0.052$), whereas heterozygous loss-of-function alleles were not consistently correlated with N_e. Phylogenetically corrected P values and coefficients (phylolm) are reported. ns, not significant.

fix weakly and moderately deleterious alleles (drift load) (*12*, *24*) and supports empirical studies involving fewer or single taxa (*25*–*27*).

The correlations between N_e and conservation status and between N_e and drift load suggest that historical demography may influence contemporary extinction risk by shaping genome-wide diversity and genetic load. We found inconsistent relationships, however, between a species' proportional genetic load and its odds of being threatened. Species with proportionally more missense substitutions were more likely to be threatened when considering all genes (phyloglm, $P = 0.002$; fig. S4D) and when considering genes in lethal and viable IMPC categories (phyloglm, $P < 0.023$; fig. S6), as observed in other taxa (*28*). Drift load estimated from evolutionary constraint across the genome, however, showed the opposite pattern: Species with proportionally fewer substitutions at evolutionarily conserved sites were more likely to be threatened (phyloglm, $P = 1.38 \times 10^{-5}$; fig. S4C). This latter result contrasts with expectations, given that threatened species have smaller N_e on average (Fig. 1) and smaller N_e is associated with proportionally more substitutions at conserved sites (phylolm, $P = 9.6 \times 10^{-3}$; fig. S4A). Notably, a previous study of 100 mammal genomes also found that threatened species had lower mean conservation scores across mutations (*29*). The authors suggested that the pattern may reflect fewer recessive deleterious alleles because of purging or the loss of these rare alleles to drift. The conflicting relationships between conservation status and metrics of drift load thus do not provide strong support for a mechanistic link between fixed drift load as measured in this study and species' resilience against extinction.

Genomic information can help predict extinction risk

Historical N_e was the most consistent genomic predictor of conservation status across regression models, whereas the predictive value of genetic load metrics varied with phylogenetic context (Fig. 3 and tables S2 and S3). Ordinal and logistic regression models incorporating genomic variables with taxonomic order and dietary trophic level showed that the effect of N_e varied by ecological context. For example, an herbivore with a given N_e was more likely to be threatened than a carnivore or omnivore with the same N_e (Fig. 3B), supporting findings of elevated extinction risk in herbivores despite larger populations (*30*). Similarly, Carnivora and Primates both had increased risk with lower levels of severely deleterious genetic load. However, the specific metric of load that predicted conservation status differed among taxonomic orders, perhaps reflecting differences in natural history or ecological flexibility (figs. S8 to S10). Principal components regression of demographic and genetic load variables showed that, overall, threatened species tended to have proportionally more deleterious mutations in coding regions, lower heterozygosity, and smaller N_e (PC1; $P = 0.0038$), as well as proportionally more missense substitutions (PC3; $P = 5.6 \times 10^{-4}$; Fig. 3A and table S3). Although no single genomic variable unambiguously discriminated threatened from nonthreatened species (fig. S2), many have predictive value, which will be particularly relevant for species lacking adequate ecological or census data.

Although ecological data were more powerful than genomic data in predicting extinction risk in our predictive models, models using only information from single genomes nonetheless identified species at risk of being threatened. We generated random forest models to predict conservation status from ecological traits (*31*, *32*) and genomic features, using area under the receiver operating characteristic (AUROC) to evaluate performance. A model with AUROC of 0.5 has no predictive ability, whereas a model with AUROC of 1.0 has perfect predictive performance. We selected predictive variables from among 13 genome-wide summary statistics including demographic history, genetic diversity, and genetic load variables; ~57,000 window-based metrics per genome; and 39 ecological variables from PanTHERIA (*15*), including physiological, life-history, and behavioral variables (table S4). Models including only genomic features and no ecological variables (17 models; AUROC ranging from 0.69 to 0.82) performed worse than models including only ecological variables (one model; AUROC of 0.88) and performed similarly to models including both genomic and ecological variables (17 models; AUROC ranging from 0.68 to 0.83; table S5). Models with only genomic features, however, were consistently better able to distinguish threatened from nonthreatened species (tables S5 and S6 and figs. S11 to S13) compared with random chance (i.e., AUROC of 0.5). Models including only genomic variables performed similarly to other studies that predicted IUCN status from ecological or morphological data with comparable sample sizes (e.g., AUC ranging from 0.67 to 0.90 for $n = 171$ to 430 species) (*33*–*35*).

The number of species with values for ecological variables, genome-wide summary statistics, and genomic window-based metrics differed, which may affect model performance.

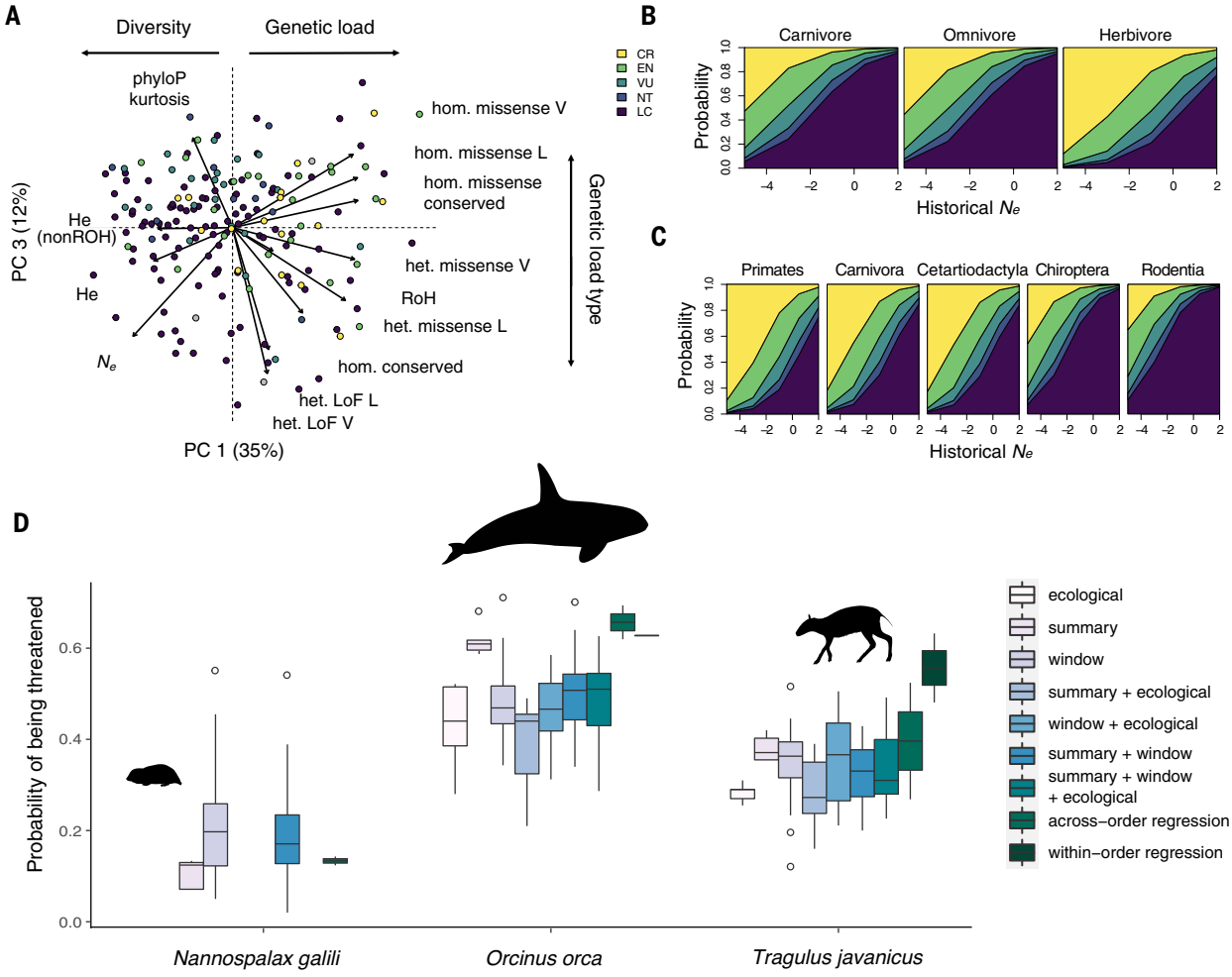


Fig. 3. Prediction of conservation status of species using genomic information. (**A**) Principal components (PCs) that significantly predict threatened status. PC1 describes heterozygosity, N_e, and deleterious variation, and PC3 distinguishes types of deleterious variation. Loadings of genomic variables (arrows; table S3) are labeled as described in table S2 (L, IMPC lethal genes; V, IMPC viable genes). Points indicate species, colored by IUCN status as shown in (B). hom., homozygous; het. heterozygous. (**B** and **C**) Probability of assignment to IUCN categories by diet and scaled values of historical N_e (B) and by taxonomic order and historical N_e of species (C). Decreased historical N_e is consistently associated with increased risk, but the magnitude varies by diet and taxonomic order. (**D**) Conservation status predictions for three data-deficient species using random forest models with genomic window-based metrics ("window"), ecological variables ("ecological"), and/or genome-wide summary variables ("summary") and predictions from regression models within and across taxonomic orders. *N. galili* lacked ecological data and adequate within-order data, so only predictions from across-order regression and windows models are shown for this species. Boxes extend from the first to third quartiles. Whiskers show first and third quartiles ± 1.5 times the interquartile range. [Animal silhouettes are from PhyloPic]

To compare the predictive value of genomic and ecological features directly, we next tested models in a set of 210 species for which both data types were available (tables S4 and S6). Again, the model with genome-wide summary statistics alone was predictive of threatened status (AUROC of 0.71) but performed more poorly than the model with ecological variables (AUROC of 0.83). Combining genomic summary statistics with ecological variables led to a modest improvement in distinguishing threatened from nonthreatened species (AUROC of 0.85) compared with genomic variables alone, with N_e as the fourth most important predictor in the model after weaning age, age at first birth, and age of sexual maturity (fig. S14). Models including genomic window–based features never outperformed models with ecological variables alone (table S6), suggesting that complementary information provided by genomic versus ecological data may be better captured by summary or transformed variables (e.g., principal components) than by numerous weakly informative window features that may overwhelm the predictive models. Overall, our evaluation suggests that while genomic information from a single individual is not better than ecological data for predicting threatened status, these data do have predictive value, especially when ecological variables are unavailable.

As a demonstration of their utility, we applied our regression and random forest models to predict the status of three species considered "Data Deficient" by the IUCN (Fig. 3D). The models suggest the Upper Galilee Mountains blind mole rat (*Nannospalax galili*), which lacks ecological data, is least likely to be threatened (11 to 44% probability), whereas the killer whale (*Orcinus orca*), for which both ecological and genomic data are available, is more likely to be threatened (35 to 68% probability), consistent with the identification of some at-risk populations (*36*). Predictions for the Java lesser chevrotain (*Tragulus javanicus*) depend on model specifications, with the highest threat prediction from the within-order regression model (67% probability), and other models suggesting it is less likely to be threatened (24 to 49% probability). The results indicate that, among the three species, the killer

whale should be prioritized for further study, and they demonstrate how genomic data can provide a rapid and inexpensive initial conservation assessment.

Discussion

Our results provide empirical support for theoretical predictions that small populations accumulate and fix weakly and moderately deleterious alleles, and they demonstrate a correlation between historical effective population size and contemporary extinction risk. We found little evidence, however, that species with historically small effective population sizes have higher risks of extinction because of elevated drift load. Alternatively, historically small populations may have an elevated extinction risk simply because these populations are small and thus more vulnerable to other threats, such as habitat loss or change, the introduction of infectious disease, competition with invasive species, and new hunting or predation pressures.

Despite the limitations of assuming that a single genome is representative of the diversity within a species, our comparative genomics approach allowed us to maximize the number of species analyzed to explore the power to detect genomic correlates of endangerment. Empirical studies suggest that a single individual can represent a species for characteristics shaped by long-term evolutionary history; variation in the proportion of deleterious mutations is typically smaller within species than between them (*29*, *37*), and historical N_e estimates are consistent across conspecifics (*38*, *39*). The analysis of multiple resequenced individuals per species, however, will increase accuracy and resolution by capturing intraspecific variation in genetic diversity, heterozygosity, and inbreeding (especially for species with strong population structure), enabling estimation of allele frequencies, improving inference of more recent demographic history, and allowing better detection of rare and segregating variants [e.g., inbreeding load (*22*)]. The latter may be particularly important for estimating extinction risk, as segregating variants tend to be enriched for deleterious alleles (*40*, *41*) and may disproportionately affect extinction risk from population bottlenecks (*12*). In the future, larger datasets comprising multiple individuals per species may shed light on longstanding questions about the relative impact on fitness of many weakly deleterious alleles versus a few strongly deleterious alleles (*22*, *25*, *37*, *42*, *43*).

Inferring real-world fitness from genomic data includes caveats. Evolutionary constraint may, for example, reflect past selection on loci that no longer affect fitness (*44*). Loci that seem functionally important in model species may be irrelevant to the species of interest, compensatory mutations may ameliorate the impact of deleterious mutations, and factors such as dominance, epistasis, pleiotropy, and purging may also complicate the relationship between genetic load and fitness. Finally, local differences in habitat may mean that the impact of deleterious mutations differs among individuals or populations (*25*, *45*, *46*). For these reasons, the impact of the observed proportionally higher load in smaller populations will be challenging to know in the absence of direct fitness data, such as reproductive success and the frequencies of genetic diseases and congenital abnormalities (*26*, *43*, *47*).

As additional genomes and population resequencing data become available (*48*), the power and accuracy of predictions of extinction risk from genomes will improve (*8*). Our analyses of the genomes of single individuals, which can be generated rapidly and inexpensively (*49*), demonstrate the potential for using genomic estimates of demography, diversity, and genetic load to triage species in need of immediate management intervention, and we join in the calls for including genomics in conservation status assessments (*50*–*53*).

Materials and methods summary

We provide a summary of our materials and methods below. Refer to the full materials and methods in the supplementary materials for further details.

Mammal genomes and metadata

We examined genomic variation in 240 species represented by 241 reference genomes in the Zoonomia multispecies alignment. The genome assemblies varied in quality, with contig N50 values ranging from 1 KB to 56 MB (table S1). Short-read sequence data, usually from the reference individual, were used to estimate metrics related to historical demography, heterozygosity, and heterozygous deleterious variants from single genomes. Homozygous deleterious genetic load was estimated relative to reconstructed ancestral sequences from the multispecies alignment (fig. S1).

For all species, we compiled metadata on conservation status, diet, and generation time (table S1). We assigned a conservation status [Least Concern (LC), Near Threatened (NT), Vulnerable (VU), Endangered (EN), or Critically Endangered (CR)] to the lowest known taxonomic level of the sequenced sample, using the IUCN Red List of Threatened Species (IUCN Red List API version 3) as a proxy for extinction risk. We classified each species as carnivore, herbivore, or omnivore according to (*54*), using information for the genus when species-specific information was unavailable. From available metadata, we categorized the sample used for both the reference genome and short-read data as a wild, captive, or domesticated individual. We tested correlations between all genomic metrics, and between genomic metrics and extinction risk, using a statistical framework that accounts for phylogenetic relationships across species. Phylogenetic linear regressions and phylogenetic logistic regressions were conducted in the R package phylolm (*55*), incorporating the phylogenetic tree with branch lengths (*56*) to account for non-independence. Using regression and machine learning models, we tested the potential for genomic data to predict the conservation status of species.

Estimating historical effective population sizes and genome-wide heterozygosity

We called heterozygous positions in all genomes with short-read data using the GATK pipeline, as described previously (*7*). Briefly, we mapped paired-end sequencing data to the respective genome assemblies using BWA mem (version 0.7.15) (*57*), marked and removed optical duplicates, and called heterozygous variants using the HaplotypeCaller module of the GATK software suite (version 3.6) (*58*).

We inferred the history of effective population sizes (N_e) for each species using PSMC (version 0.6.5-r67) (*59*). We called variants in each genome from scaffolds >50 KB in length, filtered for sequence read coverage and base quality score, and used these as input for PSMC. We rescaled the PSMC output using species-specific generation times (*60*) and a mammalian mutation rate (*21*) and calculated the harmonic mean across temporal estimates from periods >10 thousand years ago. To compare contemporary population sizes to historical N_e, we obtained census population estimates (N_c) for 89 species from the PanTHERIA database (*15*), estimating N_c as the product of population density and geographic area from census data (*15*, *61*).

We identified runs of homozygosity (RoH) using our previously described method (*7*). For every assembly, we calculated the ratio of heterozygous to callable positions in nonoverlapping 50-kb windows and fit a two-component Gaussian mixture model to the joint distribution, which is expected to be bimodal with a peak at the lower tail of the distribution corresponding to RoH (fig. S1B). Windows were then assigned as RoH or non-RoH and used to calculate the proportion of the genome in RoH (fRoH), genome-wide heterozygosity, and outbred heterozygosity (i.e., heterozygosity in non-RoH regions; figs. S2 and S15).

Deleterious genetic load

We called heterozygous variants from single-sample short-read data mapped to the reference genome of each species. Homozygous substitutions were estimated from each reference genome relative to the closest reconstructed ancestral sequence in the phylogeny using the halBranchMutations tool in the Comparative

Genomics Toolkit (62). Because new alleles become fixed or lost on the order of <4 N_e generations (63), most homozygous substitutions between species are likely fixed. We assessed the potential functional impact of mutations by (i) evolutionary conservation of the site (phyloP) and (ii) the estimated impact of the mutation on protein-coding genes. Mutations at evolutionarily conserved sites [phyloP > 2.27 (9)] and those that cause nonsynonymous changes in protein-coding genes were assumed to be predominantly harmful (19). Variant sites in each genome were assigned human-based phyloP scores estimated from the multispecies alignment (9). To infer functional impacts on protein-coding genes, each genome was annotated with human orthologs by lifting over human exon intervals to the target species. Synonymous, missense, and loss-of-function variants were then estimated in the program SnpEff v.5.0e (64). We also examined mutations in single-copy genes with associated viability phenotypic data in knockout mice as classified by the IMPC (23), using IMPC categories (e.g., lethal or viable) as proxies for gene essentiality and the potential fitness impacts of mutations in these genes (23).

Predicting threat from genomic variables

To predict whether a species is threatened (NT, VU, EN, and CR categories) or nonthreatened (LC category), we modeled conservation status across species from genomic variables using both regression and machine learning models.

We took two main approaches in our regression models of conservation status across species, using (i) phylogenetic logistic regression to model threatened versus nonthreatened status, which allowed us to test the significance of predictor variables, but not make predictions for species with unknown threat status, and (ii) ordinal regression models of specific IUCN categories, which allowed us to test significance and make predictions for species with unknown threat status. Unlike logistic regression, ordinal regression did not inherently incorporate the phylogeny, so we included taxonomic order as a factor in the models. We tested 13 genomic variables (table S2), modeled individually and as principal components, and included taxonomic order and dietary trophic level, a previously described correlate of extinction risk (65). We estimated model error by fitting parameters on 80% of the data and testing the remaining 20% of the data across 100 runs with different data subsets.

We used random forest–based classification to estimate the likelihood that a species is threatened from 13 genome-wide summary statistics of heterozygosity, demographic history, and genetic load and from five genomic metrics within homologous 50-KB windows (table S4). We trained models using the two genomic data types (windows-based and genome-wide), separately and combined, and incorporated 39 ecological variables from the PanTHERIA database (table S4). We used the scikit-learn 1.0.2 package for fitting all the models (66).

We first split our dataset into a 75% training set and a 25% test set. For each model, we performed preprocessing and imputation steps using only the training data, then we trained the model on the training set and evaluated it on the test set. We ran fivefold cross-validation on the training set to determine the optimal set of hyperparameters, tuning the number of decision trees, the maximum depth of the trees, and the number of features used at each decision to optimize a performance metric. We used AUROC to estimate how well a model predicts the correct output class. AUROC is designed to be more robust to class imbalance in comparison to a metric such as accuracy.

To leverage all available data, we first ran models using all species with data for a given data type (table S5). The number of species with values for ecological, genome-wide summary statistics, and window-based metrics differed however, which may affect the results. To compare the performance of ecological and genomic variables and their combination across the same set of species, we also trained and tested models in the set of species for which both data types were available (table S6).

The Zoonomia alignment included three species classified as Data Deficient by the IUCN, the Upper Galilee Mountains blind mole rat (*N. galili*), the Java lesser chevrotain (*T. javanicus*), and the killer whale (*O. orca*). The blind mole rat lacked ecological data on PanTHERIA. We used the within-order and across-order ordinal regression models and all random forest models to predict the probability that these species are threatened.

REFERENCES AND NOTES

1. A. D. Barnosky *et al*., Has the Earth's sixth mass extinction already arrived? *Nature* **471**, 51–57 (2011). doi: 10.1038/nature09678; pmid: 21368823
2. G. Ceballos, A. H. Ehrlich, P. R. Ehrlich, *The Annihilation of Nature: Human Extinction of Birds and Mammals* (Johns Hopkins Univ. Press, 2015).
3. M. A. Supple, B. Shapiro, Conservation of biodiversity in the genomics era. *Genome Biol.* **19**, 131 (2018). doi: 10.1186/s13059-018-1520-3; pmid: 30205843
4. B. Hansson, H. E. Morales, C. van Oosterhout, Comment on "Individual heterozygosity predicts translocation success in threatened desert tortoises". *Science* **372**, eabh1105 (2021). doi: 10.1126/science.abh1105; pmid: 34083458
5. B. Hansson, L. Westerberg, On the correlation between heterozygosity and fitness in natural populations. *Mol. Ecol.* **11**, 2467–2474 (2002). doi: 10.1046/j.1365-294X.2002.01644.x; pmid: 12453232
6. J. A. DeWoody, A. M. Harder, S. Mathur, J. R. Willoughby, The long-standing significance of genetic diversity in conservation. *Mol. Ecol.* **30**, 4147–4154 (2021). doi: 10.1111/mec.16051; pmid: 34191374
7. Zoonomia Consortium, A comparative genomics multitool for scientific discovery and conservation. *Nature* **587**, 240–245 (2020). doi: 10.1038/s41586-020-2876-6; pmid: 33177664
8. C. van Oosterhout, Mutation load is the spectre of species conservation. *Nat. Ecol. Evol.* **4**, 1004–1006 (2020). doi: 10.1038/s41559-020-1204-8; pmid: 32367032
9. M. J. Christmas *et al*., Evolutionary constraint and innovation across hundreds of placental mammals. *Science* **380**, eabn3943 (2023). doi: 10.1123/science.abn3943
10. W. J. Ripple *et al*., Status and ecological effects of the world's largest carnivores. *Science* **343**, 1241484 (2014). doi: 10.1126/science.1241484; pmid: 24408439
11. W. J. Ripple *et al*., Collapse of the world's largest herbivores. *Sci. Adv.* **1**, e1400103 (2015). doi: 10.1126/sciadv.1400103; pmid: 26601172
12. M. Kardos *et al*., The crucial role of genome-wide genetic variation in conservation. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2104642118 (2021). doi: 10.1073/pnas.2104642118; pmid: 34772759
13. J. C. Teixeira, C. D. Huber, The inflated significance of neutral genetic diversity in conservation genetics. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2015096118 (2021). doi: 10.1073/pnas.2015096118; pmid: 33608481
14. C. R. Peart *et al*., Determinants of genetic variation across eco-evolutionary scales in pinnipeds. *Nat. Ecol. Evol.* **4**, 1095–1104 (2020). doi: 10.1038/s41559-020-1215-5; pmid: 32514167
15. K. E. Jones *et al*., PanTHERIA: A species-level database of life history, ecology, and geography of extant and recently extinct mammals. *Ecology* **90**, 2648 (2009). doi: 10.1890/08-1494.1
16. IUCN SSC Antelope Specialist Group, *Beatragus hunteri*, IUCN SSC Antelope Specialist Group, e.T6234A50185297 (2017); https://dx.doi.org/10.2305/IUCN.UK.2017-2.RLTS.T6234A50185297.en.
17. S. Zhao *et al*., Whole-genome sequencing of giant pandas provides insights into demographic history and local adaptation. *Nat. Genet.* **45**, 67–71 (2013). doi: 10.1038/ng.2494; pmid: 23242367
18. P. W. Hedrick, Conservation genetics and North American bison (*Bison bison*). *J. Hered.* **100**, 411–420 (2009). doi: 10.1093/jhered/esp024; pmid: 19414501
19. B. M. Henn, L. R. Botigué, C. D. Bustamante, A. G. Clark, S. Gravel, Estimating the mutation load in human genomes. *Nat. Rev. Genet.* **16**, 333–343 (2015). doi: 10.1038/nrg3931; pmid: 25963372
20. M. Kimura, Evolutionary rate at the molecular level. *Nature* **217**, 624–626 (1968). doi: 10.1038/217624a0; pmid: 5637732
21. S. Kumar, S. Subramanian, Mutation rates in mammalian genomes. *Proc. Natl. Acad. Sci. U.S.A.* **99**, 803–808 (2002). doi: 10.1073/pnas.022629899; pmid: 11792858
22. P. W. Hedrick, A. Garcia-Dorado, Understanding inbreeding depression, purging, and genetic rescue. *Trends Ecol. Evol.* **31**, 940–952 (2016). doi: 10.1016/j.tree.2016.09.005; pmid: 27743611
23. V. Muñoz-Fuentes *et al*., The International Mouse Phenotyping Consortium (IMPC): A functional catalogue of the mammalian genome that informs conservation. *Conserv. Genet.* **19**, 995–1005 (2018). doi: 10.1007/s10592-018-1072-9; pmid: 30100824
24. M. Kimura, T. Maruyama, J. F. Crow, The mutation load in small populations. *Genetics* **48**, 1303–1312 (1963). doi: 10.1093/genetics/48.10.1303; pmid: 14071753
25. C. Grossen, F. Guillaume, L. F. Keller, D. Croll, Purging of highly deleterious mutations through severe bottlenecks in Alpine ibex. *Nat. Commun.* **11**, 1001 (2020). doi: 10.1038/s41467-020-14803-1; pmid: 32081890
26. J. A. Robinson *et al*., Genomic signatures of extensive inbreeding in Isle Royale wolves, a population on the threshold of extinction. *Sci. Adv.* **5**, eaau0757 (2019). doi: 10.1126/sciadv.aau0757; pmid: 31149628
27. K. Yoshida *et al*., Accumulation of deleterious mutations in landlocked threespine stickleback populations. *Genome Biol. Evol.* **12**, 479–492 (2020). doi: 10.1093/gbe/evaa065; pmid: 32232440
28. J. Rolland, D. Schluter, J. Romiguier, Vulnerability to fishing and life history traits correlate with the load of deleterious mutations in teleosts. *Mol. Biol. Evol.* **37**, 2192–2196 (2020). doi: 10.1093/molbev/msaa067; pmid: 32163146
29. T. van der Valk, M. de Manuel, T. Marques-Bonet, K. Guschanski, Estimates of genetic load suggest frequent purging of deleterious alleles in small populations. bioRxiv 696831 [Preprint] (2021). https://doi.org/10.1101/696831.
30. T. B. Atwood *et al*., Herbivores at the highest risk of extinction among mammals, birds, and reptiles. *Sci. Adv.* **6**, eabb8458 (2020). doi: 10.1126/sciadv.abb8458; pmid: 32923612

31. L. M. Bland, B. Collen, C. D. L. Orme, J. Bielby, Predicting the conservation status of data-deficient species. *Conserv. Biol.* **29**, 250–259 (2015). doi: 10.1111/cobi.12372; pmid: 25124400
32. A. D. Davidson, M. J. Hamilton, A. G. Boyer, J. H. Brown, G. Ceballos, Multiple ecological pathways to extinction in mammals. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 10702–10705 (2009). doi: 10.1073/pnas.0901956106; pmid: 19528635
33. R. H. L. Walls, N. K. Dulvy, Eliminating the dark matter of data deficiency by predicting the conservation status of Northeast Atlantic and Mediterranean Sea sharks and rays. *Biol. Conserv.* **246**, 108459 (2020). doi: 10.1016/j.biocon.2020.108459
34. D. B. Miles, Can morphology predict the conservation status of iguanian lizards? *Integr. Comp. Biol.* **60**, 535–548 (2020). doi: 10.1093/icb/icaa074; pmid: 32559284
35. R. K. Kopf, C. Shaw, P. Humphries, Trait-based prediction of extinction risk of small-bodied freshwater fishes. *Conserv. Biol.* **31**, 581–591 (2017). doi: 10.1111/cobi.12882; pmid: 27976421
36. E. Jourdain *et al*., North Atlantic killer whale *Orcinus orca* populations: A review of current knowledge and threats to conservation. *Mammal Rev.* **49**, 384–400 (2019). doi: 10.1111/mam.12168
37. J. A. Robinson *et al*., The critically endangered vaquita is not doomed to extinction by inbreeding depression. *Science* **376**, 635–639 (2022). doi: 10.1126/science.abm1742; pmid: 35511971
38. N. F. Saremi *et al*., Puma genomes from North and South America provide insights into the genomic consequences of inbreeding. *Nat. Commun.* **10**, 4769 (2019). doi: 10.1038/s41467-019-12741-1; pmid: 31628318
39. W. K. Meyer *et al*., Evolutionary history inferred from the de novo assembly of a nonmodel organism, the blue-eyed black lemur. *Mol. Ecol.* **24**, 4392–4405 (2015). doi: 10.1111/mec.13327; pmid: 26198179
40. G. Bertorelle *et al*., Genetic load: Genomic estimates and applications in non-model animals. *Nat. Rev. Genet.* **23**, 492–503 (2022). doi: 10.1038/s41576-022-00448-x; pmid: 35136196
41. J. B. W. Wolf, A. Künstner, K. Nam, M. Jakobsson, H. Ellegren, Nonlinear dynamics of nonsynonymous (d_N) and synonymous (d_S) substitution rates affects inference of selection. *Genome Biol. Evol.* **1**, 308–319 (2009). doi: 10.1093/gbe/evp030; pmid: 20333200
42. A. Khan *et al*., Genomic evidence for inbreeding depression and purging of deleterious genetic variation in Indian tigers. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2023018118 (2021). doi: 10.1073/pnas.2023018118; pmid: 34848534
43. L. Smeds, H. Ellegren, From high masked to high realized genetic load in inbred Scandinavian wolves. *Mol. Ecol.* **32**, 1567–1580 (2023). doi: 10.1111/mec.16802; pmid: 36458895
44. C. D. Huber, B. Y. Kim, K. E. Lohmueller, Population genetic models of GERP scores suggest pervasive turnover of constrained sites across mammalian evolution. *PLOS Genet.* **16**, e1008827 (2020). doi: 10.1371/journal.pgen.1008827; pmid: 32469868
45. J. A. Mee, S. Yeaman, Unpacking conditional neutrality: Genomic signatures of selection on conditionally beneficial and conditionally deleterious mutations. *Am. Nat.* **194**, 529–540 (2019). doi: 10.1086/702314; pmid: 31490722
46. Y. Zhang, A. J. Stern, R. Nielsen, Evolution of the genetic architecture of local adaptations under genetic rescue is determined by mutational load and polygenicity. bioRxiv 2020.11.09.374413 [Preprint] (2020). https://doi.org/10.1101/2020.11.09.374413.
47. J. A. Robinson, C. Brown, B. Y. Kim, K. E. Lohmueller, R. K. Wayne, Purging of strongly deleterious mutations explains long-term persistence and absence of inbreeding depression in island foxes. *Curr. Biol.* **28**, 3487–3494.e4 (2018). doi: 10.1016/j.cub.2018.08.066; pmid: 30415705
48. H. B. Shaffer, E. Toffelmier, "California Conservation Genomics Project First Year Annual Report" (University of California, Los Angeles, 2020); https://escholarship.org/content/qt2sc7s29z/qt2sc7s29z.pdf.
49. O. Dudchenko *et al*., The Juicebox Assembly Tools module facilitates *de novo* assembly of mammalian genomes with chromosome-length scaffolds for under $1000. bioRxiv 254797 [Preprint] (2018). https://doi.org/10.1101/254797.
50. F. W. Allendorf, P. A. Hohenlohe, G. Luikart, Genomics and the future of conservation genetics. *Nat. Rev. Genet.* **11**, 697–709 (2010). doi: 10.1038/nrg2844; pmid: 20847747
51. B. J. McMahon, E. C. Teeling, J. Höglund, How and why should we implement genomics into conservation? *Evol. Appl.* **7**, 999–1007 (2014). doi: 10.1111/eva.12193; pmid: 25553063
52. P. Brandies, E. Peel, C. J. Hogg, K. Belov, The value of reference genomes in the conservation of threatened species. *Genes* **10**, 846 (2019). doi: 10.3390/genes10110846; pmid: 31717707
53. C. van Oosterhout *et al*., Genomic erosion in the assessment of species extinction risk and recovery potential. bioRxiv 2022.09.13.507768 [Preprint] (2022). https://doi.org/10.1101/2022.09.13.507768.
54. R. M. Nowak, E. P. Walker, *Walker's Mammals of the World* (Johns Hopkins Univ. Press, 1999).
55. L.s. T. Ho, C. Ané, A linear-time algorithm for Gaussian and non-Gaussian trait evolution models. *Syst. Biol.* **63**, 397–408 (2014). doi: 10.1093/sysbio/syu005; pmid: 24500037
56. N. M. Foley *et al*., A genomic time scale for placental mammal evolution. *Science* **380**, eabl8189 (2023) doi: 10.1123/science.abl8189
57. H. Li, Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997 [q-bio.GN] (2013).
58. A. McKenna *et al*., The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010). doi: 10.1101/gr.107524.110; pmid: 20644199
59. H. Li, R. Durbin, Inference of human population history from individual whole-genome sequences. *Nature* **475**, 493–496 (2011). doi: 10.1038/nature10231; pmid: 21753753
60. M. Pacifici *et al*., Generation length for mammals. *Nat. Conserv.* **5**, 89–94 (2013). doi: 10.3897/natureconservation.5.5734
61. A. B. Roddy, D. Alvarez-Ponce, S. W. Roy, Mammals with small populations do not exhibit larger genomes. *Mol. Biol. Evol.* **38**, 3737–3741 (2021). doi: 10.1093/molbev/msab142; pmid: 33956142
62. G. Hickey, B. Paten, D. Earl, D. Zerbino, D. Haussler, HAL: A hierarchical format for storing and analyzing multiple genome alignments. *Bioinformatics* **29**, 1341–1342 (2013). doi: 10.1093/bioinformatics/btt128; pmid: 23505295
63. S. P. Otto, M. C. Whitlock, "Fixation probabilities and times" in *Encyclopedia of Life Sciences* (Wiley, 2006); https://doi.org/10.1038/npg.els.0005464.
64. P. Cingolani *et al*., A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain *w*^{1118}; *iso-2; iso-3*. *Fly* **6**, 80–92 (2012). doi: 10.4161/fly.19695; pmid: 22728672
65. A. Purvis, J. L. Gittleman, G. Cowlishaw, G. M. Mace, Predicting extinction risk in declining species. *Proc. Biol. Sci.* **267**, 1947–1952 (2000). doi: 10.1098/rspb.2000.1234; pmid: 11075706
66. A. Abraham *et al*., Machine learning for neuroimaging with scikit-learn. *Front. Neuroinform.* **8**, 14 (2014). doi: 10.3389/fninf.2014.00014; pmid: 24600388

ACKNOWLEDGMENTS

We thank M. Diekhans for technical assistance and I. Kaplow, H. Lewin, members of the Conservation Genetics Lab at San Diego Zoo Wildlife Alliance, and members of the Paleogenomics Lab at the University of California, Santa Cruz, for discussions. We thank M. Kardos and three other reviewers for insightful feedback. We gratefully acknowledge the MIT PRIMES program and the lab of V. Kuchroo at the Broad Institute for support of A.M. and A.S. **Funding:** Funding was provided by NIH grant R01 HG008742 (E.K.K.); the Swedish Research Council Distinguished Professor Award (K.L.-T.); the Wallenberg Foundation (K.L.-T.); European Research Council European Union's Horizon 2020 864203 (T.M.-B.); MINECO/FEDER, UE grant BFU2017-86471-P (T.M.-B.); Agencia Estatal de Investigación "Unidad de Excelencia María de Maeztu" CEX2018-000792-M (T.M.-B.); a Howard Hughes International Early Career award (T.M.-B.); Secretaria d'Universitats i Recerca (T.M.-B.); and CERCA Programme del Departament d'Economia i Coneixement de la Generalitat de Catalunya (T.M.-B.). **Author contributions:** Conceptualization: A.P.W., M.A.S., A.S.-A., C.S., K.-P.K., D.P.G., E.K.K., K.L.-T., T.M.-B., Z.C., O.A.R., and B.S. Data analysis: A.P.W., M.A.S., A.S., A.M., R.S., A.S.-A., V.M.F., K.F., and W.K. M. Interpretation of results: A.P.W., M.A.S., A.S., A.M., O.A.R., and B.S., with input from all authors. Writing – original draft: A.P.W., M.A.S., and B.S. Writing – review & editing: All authors. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** The data presented in this paper are detailed in supplementary materials. Summary data and analysis scripts are available at https://github.com/apwilder/Zoonomia_biodiversity, https://github.com/ayshwaryas/zoonomia_biodiversityML_paper, and https://github.com/LaMariposa/zoonomia_biodiversity. NCBI accession numbers for sequence data used in analyses are given in table S1.

Zoonomia Consortium

Gregory Andrews[1], Joel C. Armstrong[2], Matteo Bianchi[3], Bruce W. Birren[4], Kevin R. Bredemeyer[5], Ana M. Breit[6], Matthew J. Christmas[3], Hiram Clawson[2], Joana Damas[7], Federica Di Palma[8,9], Mark Diekhans[2], Michael X. Dong[3], Eduardo Eizirik[10], Kaili Fan[1], Cornelia Fanter[11], Nicole M. Foley[5], Karin Forsberg-Nilsson[12,13], Carlos J. Garcia[14], John Gatesy[15], Steven Gazal[16], Diane P. Genereux[4], Linda Goodman[17], Jenna Grimshaw[14], Michaela K. Halsey[14], Andrew J. Harris[5], Glenn Hickey[18], Michael Hiller[19,20,21], Allyson G. Hindle[11], Robert M. Hubley[22], Graham M. Hughes[23], Jeremy Johnson[4], David Juan[24], Irene M. Kaplow[25,26], Elinor K. Karlsson[1,4,27], Kathleen C. Keough[17,28,29], Bogdan Kirilenko[19,20,21], Klaus-Peter Koepfli[30,31,32], Jennifer M. Korstian[14], Amanda Kowalczyk[25,26], Sergey V. Kozyrev[3], Alyssa J. Lawler[4,26,33], Colleen Lawless[23], Thomas Lehmann[34], Danielle L. Levesque[6], Harris A. Lewin[7,35,36], Xue Li[1,4,37], Abigail Lind[28,29], Kerstin Lindblad-Toh[3,4], Ava Mackay-Smith[38], Voichita D. Marinescu[3], Tomas Marques-Bonet[39,40,41,42], Victor C. Mason[43], Jennifer R. S. Meadows[3], Wynn K. Meyer[44], Jill E. Moore[1], Lucas R. Moreira[1,4], Diana D. Moreno-Santillan[14], Kathleen M. Morrill[1,4,37], Gerard Muntané[24], William J. Murphy[5], Arcadi Navarro[39,41,45,46], Martin Nweeia[47,48,49,50], Sylvia Ortmann[51], Austin Osmanski[14], Benedict Paten[2], Nicole S. Paulat[14], Andreas R. Pfenning[25,26], BaDoi N. Phan[25,26,52], Katherine S. Pollard[28,29,53], Henry E. Pratt[1], David A. Ray[14], Steven K. Reilly[38], Jeb R. Rosen[22], Irina Ruf[54], Louise Ryan[23], Oliver A. Ryder[55,56], Pardis C. Sabeti[4,57,58], Daniel E. Schäffer[25], Aitor Serres[24], Beth Shapiro[59,60], Arian F. A. Smit[22], Mark Springer[61], Chaitanya Srinivasan[25], Cynthia Steiner[55], Jessica M. Storer[22], Kevin A. M. Sullivan[14], Patrick F. Sullivan[62,63], Elisabeth Sundström[3], Megan A. Supple[59], Ross Swofford[4], Joy-El Talbot[64], Emma Teeling[23], Jason Turner-Maier[4], Alejandro Valenzuela[24], Franziska Wagner[65], Ola Wallerman[3], Chao Wang[3], Juehan Wang[16], Zhiping Weng[1], Aryn P. Wilder[55], Morgan E. Wirthlin[25,26,66], James R. Xue[4,57], Xiaomeng Zhang[4,25,26]

[1]Program in Bioinformatics and Integrative Biology, UMass Chan Medical School, Worcester, MA 01605, USA. [2]Genomics Institute, University of California Santa Cruz, Santa Cruz, CA 95064, USA. [3]Department of Medical Biochemistry and Microbiology, Science for Life Laboratory, Uppsala University, Uppsala 751 32, Sweden. [4]Broad Institute of MIT and Harvard, Cambridge, MA 02139, USA. [5]Veterinary Integrative Biosciences, Texas A&M University, College Station, TX 77843, USA. [6]School of Biology and Ecology, University of Maine, Orono, ME 04469, USA. [7]The Genome Center, University of California Davis, Davis, CA 95616, USA. [8]Genome British Columbia, Vancouver, BC, Canada. [9]School of Biological Sciences, University of East Anglia, Norwich, UK. [10]School of Health and Life Sciences, Pontifical Catholic University of Rio Grande do Sul, Porto Alegre 90619-900, Brazil. [11]School of Life Sciences, University of Nevada Las Vegas, Las Vegas, NV 89154, USA. [12]Biodiscovery Institute, University of Nottingham, Nottingham, UK. [13]Department of Immunology, Genetics and Pathology, Science for Life Laboratory, Uppsala University, Uppsala 751 85, Sweden. [14]Department of Biological Sciences, Texas Tech University, Lubbock, TX 79409, USA. [15]Division of Vertebrate Zoology, American Museum of Natural History, New York, NY 10024, USA. [16]Keck School of Medicine, University of Southern California, Los Angeles, CA 90033, USA. [17]Fauna Bio Incorporated, Emeryville, CA 94608, USA. [18]Baskin School of Engineering, University of California Santa Cruz, Santa Cruz, CA 95064, USA. [19]Faculty of Biosciences, Goethe-University, 60438 Frankfurt, Germany. [20]LOEWE Centre for Translational Biodiversity Genomics, 60325 Frankfurt, Germany. [21]Senckenberg Research Institute, 60325 Frankfurt, Germany. [22]Institute for Systems Biology, Seattle, WA 98109, USA. [23]School of Biology and Environmental Science, University College Dublin, Belfield, Dublin 4, Ireland. [24]Department of Experimental and Health Sciences, Institute of Evolutionary Biology (UPF-CSIC), Universitat Pompeu Fabra, Barcelona 08003, Spain. [25]Department of Computational Biology, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA. [26]Neuroscience Institute, Carnegie Mellon University, Pittsburgh, PA 15213, USA. [27]Program in Molecular Medicine, UMass Chan Medical School, Worcester, MA 01605, USA. [28]Department of Epidemiology & Biostatistics, University of California San Francisco, San Francisco, CA 94158, USA. [29]Gladstone Institutes, San Francisco, CA 94158, USA. [30]Center for Species Survival, Smithsonian's National Zoo and Conservation Biology Institute, Washington, DC 20008, USA.

[31]Computer Technologies Laboratory, ITMO University, St. Petersburg 197101, Russia. [32]Smithsonian-Mason School of Conservation, George Mason University, Front Royal, VA 22630, USA. [33]Department of Biological Sciences, Mellon College of Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA. [34]Senckenberg Research Institute and Natural History Museum Frankfurt, 60325 Frankfurt am Main, Germany. [35]Department of Evolution and Ecology, University of California Davis, Davis, CA 95616, USA. [36]John Muir Institute for the Environment, University of California Davis, Davis, CA 95616, USA. [37]Morningside Graduate School of Biomedical Sciences, UMass Chan Medical School, Worcester, MA 01605, USA. [38]Department of Genetics, Yale School of Medicine, New Haven, CT 06510, USA. [39]Catalan Institution of Research and Advanced Studies (ICREA), Barcelona 08010, Spain. [40]CNAG-CRG, Centre for Genomic Regulation, Barcelona Institute of Science and Technology (BIST), Barcelona 08036, Spain. [41]Department of Medicine and Life Sciences, Institute of Evolutionary Biology (UPF-CSIC), Universitat Pompeu Fabra, Barcelona 08003, Spain. [42]Institut Català de Paleontologia Miquel Crusafont, Universitat Autònoma de Barcelona, 08193 Cerdanyola del Vallès, Barcelona, Spain. [43]Institute of Cell Biology, University of Bern, 3012 Bern, Switzerland. [44]Department of Biological Sciences, Lehigh University, Bethlehem, PA 18015, USA. [45]BarcelonaBeta Brain Research Center, Pasqual Maragall Foundation, Barcelona 08005, Spain. [46]CRG, Centre for Genomic Regulation, Barcelona Institute of Science and Technology (BIST), Barcelona 08003, Spain. [47]Department of Comprehensive Care, School of Dental Medicine, Case Western Reserve University, Cleveland, OH 44106, USA. [48]Department of Vertebrate Zoology, Canadian Museum of Nature, Ottawa, ON K2P 2R1, Canada. [49]Department of Vertebrate Zoology, Smithsonian Institution, Washington, DC 20002, USA. [50]Narwhal Genome Initiative, Department of Restorative Dentistry and Biomaterials Sciences, Harvard School of Dental Medicine, Boston, MA 02115, USA. [51]Department of Evolutionary Ecology, Leibniz Institute for Zoo and Wildlife Research, 10315 Berlin, Germany. [52]Medical Scientist Training Program, University of Pittsburgh School of Medicine, Pittsburgh, PA 15261, USA. [53]Chan Zuckerberg Biohub, San Francisco, CA 94158, USA. [54]Division of Messel Research and Mammalogy, Senckenberg Research Institute and Natural History Museum Frankfurt, 60325 Frankfurt am Main, Germany. [55]Conservation Genetics, San Diego Zoo Wildlife Alliance, Escondido, CA 92027, USA. [56]Department of Evolution, Behavior and Ecology, School of Biological Sciences, University of California San Diego, La Jolla, CA 92039, USA. [57]Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138, USA. [58]Howard Hughes Medical Institute, Harvard University, Cambridge, MA 02138, USA. [59]Department of Ecology and Evolutionary Biology, University of California Santa Cruz, Santa Cruz, CA 95064, USA. [60]Howard Hughes Medical Institute, University of California Santa Cruz, Santa Cruz, CA 95064, USA. [61]Department of Evolution, Ecology and Organismal Biology, University of California Riverside, Riverside, CA 92521, USA. [62]Department of Genetics, University of North Carolina Medical School, Chapel Hill, NC 27599, USA. [63]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden. [64]Iris Data Solutions, LLC, Orono, ME 04473, USA. [65]Museum of Zoology, Senckenberg Natural History Collections Dresden, 01109 Dresden, Germany. [66]Allen Institute for Brain Science, Seattle, WA 98109, USA.

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.abn5856
Materials and Methods
Figs. S1 to S15
Tables S1 to S6
References (*67*–*85*)
MDAR Reproducibility Checklist

View/request a protocol for this paper from *Bio-protocol*.

Submitted 7 December 2021; accepted 8 February 2023
10.1126/science.abn5856