

Genomic diversity as a key conservation criterion: proof-of-concept from mammalian whole-genome resequencing data

Jong Yoon Jeon

Purdue University

Andrew N. Black

Western Association of Fish and Wildlife Agencies

Erangi J. Heenkenda

Purdue University

Andrew J. Mularo

Purdue University

Gina F. Lamka

Auburn University

Safia Janjua

Purdue University

Anna Brüniche-Olsen

University of Copenhagen

John W. Bickham

Texas A&M University

Janna R. Willoughby

Auburn University

J. Andrew DeWoody

dewoody@purdue.edu


Purdue University

Research Article

Keywords: Genetic diversity, sustainability, evolutionary potential, heterozygosity, autozygosity

Posted Date: December 21st, 2023

DOI: <https://doi.org/10.21203/rs.3.rs-3761026/v1>

License:  This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Version of Record: A version of this preprint was published at Evolutionary Applications on September 10th, 2024. See the published version at <https://doi.org/10.1111/eva.70000>.

Abstract

Many international, national, state, and local organizations prioritize the ranking of threatened and endangered species to help direct conservation efforts. For example, the International Union for Conservation of Nature (IUCN) regularly publishes the influential Red List of Threatened Species. Unfortunately, current approaches to categorizing the conservation status of species do not explicitly consider genetic or genomic diversity (GD), even though GD is positively associated with both contemporary evolutionary fitness and with future evolutionary potential. To test if genome sequences can help improve conservation ranking efforts, we estimated GD metrics from publicly available mammalian population data and examined their statistical association with formal Red List conservation categories. We considered intrinsic biological factors that could impact GD and quantified their relative influences. Key population GD metrics are both reflective and predictive of IUCN conservation categories. Specifically, our analyses revealed that genome-wide heterozygosity and autozygosity (a product of inbreeding) are associated with the current Red List categorization, likely because demographic declines that lead to “listing” decisions also reduce levels of standing genetic variation. We argue that by virtue of this relationship, conservation organizations like IUCN can leverage genome sequence data to help infer conservation status in otherwise data-deficient species. This study 1) outlines the theoretical and empirical justification for a new GD criterion based on the mean loss of genome-wide heterozygosity over time; 2) provides a bioinformatic pipeline for estimating GD from population genomic data; and 3) provides an analytical framework and explicit recommendations for use by conservation authorities.

Main

Global biodiversity is declining rapidly as humans continually modify wild habitats and expand their environmental footprint. Habitat reduction and fragmentation, overharvesting, invasive species, and other anthropogenic impacts routinely lead to population declines, reduced gene flow, and subsequent increases in inbreeding and genetic drift^{1,2}. Collectively, these anthropogenic impacts lead to a loss of genetic/genomic diversity (GD) and a concomitant reduction in population fitness^{3,4}. The loss of GD and fitness can accelerate an extinction vortex^{5,6} and jeopardize the sustainability of a population or species because GD provides the evolutionary potential needed to adapt to a changing environment⁷⁻⁹. In this regard, the Convention on Biological Diversity (CBD) recently listed “maintaining at least 90% of GD of all species” as “Goal A” in Post-2020 Global Biodiversity Framework with support from the International Union for Conservation of Nature (IUCN). As one of three components of biodiversity, along with species and ecosystem diversity, GD is becoming central in conservation policies.

As an international entity comprised largely of academic, government, and private members, IUCN strives to help protect nature by using the best available science to prioritize conservation efforts. One of the most visible tasks of the IUCN is their production and regular updating of the “Red List”, which classifies species into one of nine categories (Extinct, Extinct in the Wild, Critically Endangered, Endangered, Vulnerable, Near Threatened, Least Concern, Data Deficient, and Not Evaluated). The Red

List often influences national and state authorities in their official listing decision for species under their supervision. For example, the IUCN Red List is used at the international level by the Convention on International Trade in Endangered Species (CITES), the Convention on Biological Diversity (CBD), and by the United Nations Sustainable Development Goals (SDGs). The IUCN Red List is also used at the national level by the National Institute of Biological Resources of South Korea and the U.S. Fish and Wildlife Service, and at the state or provincial level by the California Department of Fish and Wildlife and by the Indiana Department of Natural Resources (among many others). Decisions made by IUCN regarding the Red List can and do reverberate through the global conservation community.

The IUCN makes categorical assignments for each species considered according to four conservation criteria: 1) census population size; 2) demographic trajectory; 3) geographic range size, and 4) associated quantitative analyses of population viability. However, Red List assignments do not explicitly consider GD despite calls to do so^{10–13}. This is unfortunate because GD is an important component of population viability^{14,15}, and in many instances GD can provide insights into rare or elusive species whose population attributes are otherwise difficult to address^{16,17}. For example, Rice's whale is a newly described species of baleen whale endemic to the Gulf of Mexico¹⁸. Baleen whales are notoriously difficult to study at sea, but empirical GD estimates from only a few individuals (e.g., sourced from beached whales or noninvasively collected DNA) could provide critical demographic context for conservation plans.

Unfortunately, there is no well-established standard to determine when the loss of GD becomes a conservation concern (though see^{19–21}). Previous studies based on microsatellite genetic markers have suggested that threshold levels of GD can be used to help delimit conservation categories. For instance, Willoughby et al. (2015) proposed a conceptual framework that—based strictly on GD estimates from related species and recognizing that GD is but one component of population viability—designates IUCN conservation categories based on the estimated time (in generations) that a species or population is predicted to lose more GD than 75% of its taxonomic relatives. This conceptual framework was proposed at the twilight of the microsatellite era. Here, we extend it into the modern genomic era.

We assessed relationships among population-level GD metrics and formal IUCN conservation categories to determine whether the traditional four criteria employed by the Red List effectively captures mammalian GD. We reasoned that if it did so, Red List Threatened species (Critically Endangered, Endangered, and Vulnerable) should exhibit lower levels of GD than Non-Threatened species (Near Threatened and Least Concern) due to inbreeding, genetic drift, and reduced gene flow. If not, this would indicate that IUCN's four evaluation criteria insufficiently capture a key aspect of biological diversity (i.e., GD). We primarily focused on the idea that mean genome-wide heterozygosity, H , can serve as an effective metric of GD that is a useful addition to the current classification approach employed by the Red List. Unlike some lagging indicators of GD such as nucleotide diversity (which may reflect more ancient demographic events such as bottlenecks or expansions), H is a leading indicator of GD because it can change dramatically in only a few generations (e.g., due to inbreeding^{10,22}). Heterozygosity can

also be accurately estimated from only a few whole genome sequences^{23,24}, an important consideration with respect to Threatened populations or species.

We do not mean to suggest that a fixed GD threshold should be used to determine conservation categories (e.g., mean $H < 0.002$ = Endangered) because of the inherent variation in GD observed among taxa. Species vary in key biological attributes that are known to affect GD such as body sizes, generation times, and metabolic rates^{25–27}. Thus, we also examined associations among fundamental biological characteristics and GD to account for major biological factors that might otherwise confound the relationship between GD and Red List status. We did so using Class Mammalia as an example because many flagship species of conservation interest (e.g., pandas, tigers, and whales) are mammals. Furthermore, mammalian data are sufficiently dense in both the IUCN Red List and in sequence repositories to allow for robust analyses of our GD framework.

Results

Genetic diversity among mammalian species

Among 613 species and subspecies with reference genome assemblies available from National Center for Biotechnology Information (NCBI), 98 species also had population genomic whole-genome resequencing (WGR) datasets that met our criteria for inclusion. Fifteen species were subsequently dropped during the bioinformatic data analysis due to unsatisfied thresholds (e.g., low mapping rates, depths, and/or breadth), resulting in 83 species in our final WGR dataset (Supplementary Dataset S1). Our “IUCN” (Supplementary Dataset S2) and “EcoEvo” (Supplementary Dataset S3) datasets had 71 and 64 species, respectively, after reconciling taxonomy and pruning for phylogenetic pseudoreplication (the “IUCN dataset” included all the species having their own categorical Red List assessment but excluded those listed as “Data-Deficient” and the “EcoEvo dataset” included all the species having data of eco-evolutionary factors; see Statistical Analysis section in Methods for details). One species (*Odocoileus virginianus*, the white-tailed deer) did not yield a $F_{ROH>100kb}$ (F_{100kb}) estimate, perhaps because of the pooled sequence data²⁸ (<https://www.ncbi.nlm.nih.gov/search/all/?term=PRJNA576136>), and 21 species did not yield $F_{ROH>1Mb}$ (F_{1Mb}) estimates. Considering its high genomic diversity and unique absence of F_{100kb} ROHs, we imputed zero for F_{100kb} and F_{1Mb} for *O. virginianus* as the lack of runs of homozygosity (ROHs) could be real given the extensive history of hybridization, population bottlenecks/expansions, and reintroductions/translocation in this species²⁹.

Descriptive statistics for each GD metric are summarized in Table 1 and Supplementary Table S1 (see also Fig. 2 and Supplementary Figures S1–S12). Nucleotide diversity (π) and Watterson’s theta (θ_W) values were strongly correlated ($r > 0.9$) with H . In general, Non-Threatened species have higher GD than Threatened species. Individual and categorical H was effectively twice as high in Non-Threatened species compared to Threatened species (Table 1 and Fig. 1), and F_{1Mb} was doubled in Threatened

species (Table 1). The overall (cumulative) fraction of ROHs estimated with $F1Mb$ was generally higher among Threatened species (Table 1 or Supplementary Figure S12).

Table 1

Genomic diversity metrics grouped by IUCN full categories. Abbreviations: spp. No. = the number of species, H = observed genome-wide heterozygosity, θ_W = Watterson's theta, π = nucleotide diversity, D = Tajima's D , $F100kb = F_{ROH > 100kb}$, $F1Mb = F_{ROH > 1Mb}$, NA = not applicable.

IUCN category	spp. No.	mean (H)	sd (H)	mean (θ_W)	sd (θ_W)	mean (π)	sd (π)
DD	2	0.00152	0.00031	0.00137	0.00025	0.00151	0.00042
LC	31	0.00366	0.00429	0.00379	0.00484	0.00359	0.00392
NT	5	0.00255	0.00094	0.00201	0.00076	0.00237	0.00085
VU	15	0.00121	0.00089	0.00119	0.00091	0.00131	0.00093
EN	15	0.00176	0.00280	0.00187	0.00341	0.00198	0.00354
CR	15	0.00145	0.00081	0.00133	0.00079	0.00140	0.00081
Non-Threatened	36	0.00351	0.00401	0.00354	0.00453	0.00342	0.00367
Threatened	45	0.00148	0.00174	0.00147	0.00206	0.00157	0.00213
IUCN category	spp. No.	mean (D)	sd (D)	mean ($F100kb$)	sd ($F100kb$)	mean ($F1Mb$)	sd ($F1Mb$)
DD	2	0.36770	0.42422	0.07753	0.10954	0.01300	NA
LC	31	0.21913	0.73928	0.07883	0.08274	0.02477	0.03985
NT	5	0.65424	0.47784	0.17202	0.10911	0.01980	0.02357
VU	15	0.69864	0.45017	0.11886	0.08329	0.06752	0.07027
EN	15	0.45159	0.55982	0.12005	0.09415	0.04228	0.04155
CR	15	0.26068	0.88502	0.09922	0.06770	0.02603	0.04629
Non-Threatened	36	0.27956	0.71961	0.09177	0.09109	0.02378	0.03679
Threatened	45	0.47030	0.66805	0.11271	0.08111	0.04407	0.05430

Bats and rodents (Order Chiroptera and Rodentia, respectively) had the highest mean π , θ_W , and H values, almost double the next most genetically diverse Order (Artiodactyla; even-toed ungulates). Carnivores and whales (Orders Carnivora and Cetacea, respectively) are at the other end of the GD distribution. Mean Tajima's D (D) was lowest among Proboscidea (elephants), Cetacea, Rodentia and highest among Eulipotyphla (hedgehogs and relatives), Pholidota (pangolins), and Carnivora. Chiroptera, Dasyuromorphia (Australian carnivorous marsupials), and Proboscidea had the lowest mean $F100kb$

while Eulipotyphla, Carnivora, and Primates had the highest $F_{100\text{kb}}$. Dasyuromorphia, Pholidota, and Proboscidea had the lowest mean $F_{1\text{Mb}}$ while Carnivora, Primates, Rodentia had the highest.

Statistical associations between GD and the IUCN Red List

Detailed model results are presented in Supplementary Table S2 (see also Supplementary Figures S13–S17). The main Phylogenetic Generalized Least Squares (PGLS) model between IUCN full categories (Least Concern - LC, Near Threatened - NT, Vulnerable - VU, Endangered - EN, Critically Endangered - CR) and H was significant. The phylogenetic signal (λ) in the model was significant ($\lambda = 0.887$; 95% CI = 0.584–0.972), implying there was phylogenetic non-independence among data which has been accounted for in the model. The secondary PGLS model of H against IUCN binary categories of Threatened and Non-Threatened (LC + NT + VU = Non-Threatened; EN + CR = Threatened) was also significant with $\lambda = 0.867$. The PGLS models also revealed a significant relationship between the ROH burden (as measured by $F_{100\text{kb}}$) and IUCN full categories and between the ROH burden and binary IUCN categories. None of the PGLS models revealed significant associations between D or $F_{1\text{Mb}}$ and IUCN category (full or binary). Among significant models, the $F_{100\text{kb}}$ with IUCN binary categories model was best, only slightly better than the model with IUCN full categories, followed by H (Supplementary Table S3). The results among models of “population trend” in place of IUCN categorization were similar to the IUCN category models in general. The H as an individual factor and a whole model with $F_{100\text{kb}}$ were significant, but not $F_{100\text{kb}}$ as an individual factor (Supplementary Figure S17). Technical Dimension 1 was the significant variable of this case. Models with D and $F_{1\text{Mb}}$ were non-significant with also non-significant independent variables of interest. Among models with the geographic range as an independent variable of interest, $F_{1\text{Mb}}$ showed significance with geographic range as an individual factor (Supplementary Figure S18).

Our analyses indicate that H and $F_{100\text{kb}}$ were the best conservation metrics of GD. Thus, either H or $F_{100\text{kb}}$ were used as independent variables in the PGLS for phylogenetic ordinal regression against IUCN categorization. Models with H as the independent variable were significant when IUCN binary category was the dependent variable, in contrast to the case of the IUCN full category as the dependent variable. Individual heterozygosity was also significant, but $F_{100\text{kb}}$ was never a significant predictor of IUCN category (whether full or binary). Results of the machine learning classifier models are presented in Supplementary Table S4. In general, H was identified as a better predictor of IUCN categories than $F_{100\text{kb}}$ across models.

Discussion

Genome resequencing data offer remarkably high information content per individual (e.g., estimates of GD such as mean H or $F_{100\text{kb}}$). This means that sampling only a few individuals can provide key insights into population biology. The relationships among GD, N_e , and fitness have been thoroughly reviewed and summarized by previous studies^{30–32}. These and other studies indicate that GD, as measured by H or related measures, is a critical component not only of contemporary fitness but also of

future evolutionary potential. The idea that GD can serve as an indicator of future evolutionary potential should not be overlooked considering the global environmental challenges facing natural populations today.

A reduction in GD, with its concomitant loss of fitness and increased probability of extinction^{9,33}, is expected to result from demographic events like population bottlenecks, population subdivision, and founder events that reduce population sizes. Neutral GD is determined by the product of the generational mutation rate and the effective population size (N_e), and thus GD is determined in part by the census size of the population^{32,34}. Moreover, and not surprisingly, population census size is positively correlated with geographic range size. According to conservation theory, small, threatened populations tend to have lower GD than large, broadly distributed populations which are typically not threatened³⁵.

Our analyses of empirical data bear out those theoretical predictions (Fig. 1). We analyzed population genomic data from 83 species belonging to 11 Orders of mammals representing the various IUCN conservation categories. For each species, we calculated GD metrics and tested for significant associations between these metrics and various biological parameters, such as geographic distribution or body size, that might impact diversity. The overarching goal of the research was to determine the relationship between population-level GD metrics and IUCN conservation categories while simultaneously identifying key intrinsic drivers of mammalian GD, which we address first.

Description of mammalian genomic diversity

Our results are consistent with a long history of empirical genetic studies dating to the 1960's when protein electrophoresis was first used to measure GD in natural populations of mammals. For example, Fig. 1 indicates that the three species with the highest H values are *O. virginianus* (white-tailed deer), *Peromyscus maniculatus* (deer mouse, including 2 subspecies), and *Myotis lucifugus* (little brown bat). Nevo et al. (1984) compiled an allozyme dataset of GD metrics, including H , from 1111 species of animals and plants including from 184 species of mammals. Their dataset was comprised of GD estimates from only a few dozen allozyme markers per species, and they examined only a few of the same species that we did. However, there are some remarkable similarities between Nevo et al. (1984) and our current study. Nevo et al. (1984) only 12 species of mammals (not including humans and domestic cat) that had values of $H \geq 0.09$. Among them were *O. virginianus*, *P. maniculatus*, and two species of bats of the genus *Myotis*. The fact that the three species with the highest H in our dataset are either the same species or a congener of high GD species reported by Nevo et al. (1984) using such a different analytical approach is reassuring. It bolsters our confidence that evolutionary genetics theory is buttressed by existing, publicly-available genomic datasets that can be readily exploited by conservationists.

Taxonomic Order is the taxonomic level in which member species share a broad suite of morphological, physiological, genetic, and ecological characteristics; species of different Orders can easily be

distinguished by many conservationists. If we just consider the 4 most speciose Orders, Rodentia had the highest mean value of $H = 0.00520$ and Carnivora had the lowest mean value $H = 0.00088$. This is not unexpected given that small herbivores generally have much larger population sizes and nucleotide substitution rates than do carnivores³⁶. Conversely, Carnivores had the highest mean $F1Mb = 0.06209$ and rodents have the second lowest mean $F1Mb = 0.02441$. Again, this is consistent with their population biology in which rodents are expected to have higher effective mutation rates and larger population sizes than carnivores, where there is generally far more opportunity for inbreeding in isolated populations. Primates have relatively high inbreeding with $F1Mb = 0.05569$. This is perhaps a reflection of a high degree of social structuring, small census population sizes, and slower rates of molecular evolution in primates³⁶.

Genomic diversity and Red List status

The major finding of this study is that key population GD metrics are predictive of IUCN conservation categories that presumably reflect extinction threat status. This supports the idea that GD is indirectly reflected by the current Red List assessment methodology. Our results also indicate that Threatened species or populations have reduced GD compared to those with Non-Threatened status. We found that H (and its correlates) was the best conservation metric, followed by $F100kb$ (a measure of autozygosity that is reflective of inbreeding). Two individual Red List criteria, “population trend” and “geographic range”, also reflect GD. Species with “Stable” population trends had significantly higher H than do “Decreasing” or “Increasing” species. Geographic range was inversely proportional to longer fraction of ROH (Supplementary Figure S18), another reasonable result in that habitat contraction can result in elevated levels of inbreeding relative to random mating³⁷.

Since H and $F100kb$ were the best predictors of Red List designation, we plotted their global distributions (Supplementary Figures S19 and S20) to illustrate world-wide patterns of GD. Mammalian populations in Asia and Africa, where the human footprint is the oldest, generally had higher levels of inbreeding than did other continents whereas North America seemed to have relatively healthier distributions of mammals with regard to their GD. Taken as a whole, the worldwide GD distribution calls for more active conservation efforts and research in Asia and the Global South.

The correlation between GD and Red List status has been tested before^{10,11,38,39} but mostly with mitochondrial or microsatellite marker data. There has been no scientific consensus on whether the Red List indirectly captures GD. Recently, Schmidt et al. (2023) performed a meta-analysis of studies that used different markers and corroborated Willoughby et al. (2015), who found that GD is modestly predictive of Red List status. Our results are consistent with this interpretation. Several authors^{10,11,40} have suggested using the loss of GD rather than snapshot values of GD in conservation assessments. In the next section, we extend this line of reasoning by detailing an approach for including GD as an explicit criterion in future conservation assessments.

An explicit genetic criterion for conservation assessments

Over thirty years ago, Mace and Lande (1991) originally suggested an assessment criterion based on N_e in Version 1.0 of the Red List Categories and Criteria, but the most recent iteration of these Criteria (Version 3.1) still do not embrace N_e despite recent pleas to include genetic considerations in status determinations (e.g., ^{10,11,42}). We suggest that an additional criterion that explicitly considers GD metrics and thresholds would help further inform conservation assessments, especially for species that might otherwise be deemed Data Deficient.

Our proposal for an explicit new GD criterion for status assessments is based on the mean loss of heterozygosity over time⁴³. We chose H not only because the concept of heterozygosity is well understood by most biologists, but because our results indicate that it was the best indicator as well as the best predictor of existing IUCN categories. Furthermore, H has a solid theoretical foundation based on Crow and Kimura's equation:

$$H_T = H_0 \left(1 - \frac{1}{2N_e}\right)^T$$

where N_e = effective population size, H_0 = observed heterozygosity, H_T = heterozygosity at time T , and T = the number of generations in 100 years (e.g., T is 100 for most insects or annual plants, T is 50 for antelope with 2-year generation times, and T is 5 for whales with 20-year generation times). Our proposed GD criterion is illustrated in Fig. 3 and, in principle, could be readily applied by any conservation organization that conducts status assessments given that the model parameters can be estimated from publicly available resources¹⁰. For example, H_0 can be estimated from population genomic datasets and generation time is generally known from life history studies. N_e can either be estimated indirectly from census population size (N_c) where N_e is crudely estimated from N_c ⁴⁴, or directly from population genomic data. For example, contemporary N_e can be estimated using the linkage-disequilibrium-based method (e.g., GONE⁴⁵) or with a coalescence-based method (e.g., Stairway Plot 2⁴⁶) so long as practitioners recognize that genomes do not immediately register demographic changes (i.e., there is a lag time^{47,48}).

We suggest that GD can be used to assign threat categories (e.g., CR or VU) when a population is expected to lose a given proportion of its H in 100 years⁴⁹⁻⁵² as follows:

CR: if H_T is 90% or less of H_0 (i.e., a 10% or more loss of heterozygosity in 100 years)

EN: if H_T is 90–95% of H_0 (a 5–10% loss of heterozygosity in 100 years)

VU: if H_T is 95–97.5% or less of H_0 (a 2.5-5% loss of heterozygosity in 100 years) OR $N_e < 1000$

NT: if H_T is more than 97.5% of H_0 AND $1000 \leq N_e < 5000$

LC if H_T is more than 97.5% of H_0 AND $N_e \geq 5000$

We tested this new GD criterion using maximum population size estimates for species with Red List information, or by employing Stairway Plot 2 to estimate N_e for “Data Deficient” species (i.e., those without Red List information). The results are presented in Fig. 4 and Supplementary Table S5. Compared to the official IUCN Red List categories, the “GD categories” that we derived from the GD criterion described above were generally more conservative, likely because we used the maximum population size estimates available. The percentage loss of heterozygosity in 100 years (“Het_loss_” in Supplementary Table S5) was less than 2.5% in many cases thanks to large census population sizes and/or long generation intervals. The two “Data-Deficient” species were assigned as “LC” or “NT” based on large estimates of N_e . However, while all the “LC” and “NT” species of the Red List remained in “Non-Threatened” categories, some “EN” and “CR” species according to the Red List were elevated or remained as “CR” when evaluated using only our new GD criterion, perhaps foreshadowing genomic manifestations of the extinction vortex⁵.

Our analyses show that the five conservation criteria currently used by IUCN (census population size, demographic trajectory, geographic range size, a combined index of population size and geographic range size, and associated quantitative analyses) indirectly capture heterozygosity, a key element of GD. However, many species on IUCN’s Red List are “Data Deficient” because parameters like census population size or demographic trajectory are extremely difficult to estimate. We think that GD could become valuable as a sixth criterion for conservation assessments, in large part because GD can be more easily and inexpensively evaluated than census size or demographic trajectory and can be estimated directly (by anyone) from public databases that are expanding rapidly.

Regardless of whether the scientific community adopts our specific GD criterion, we think conservationists would do well to explicitly assess GD metrics as part of a comprehensive evaluation of each species. We expect other genomic assessments, such as genetic load or genomic offset, could ultimately be incorporated into a more comprehensive GD criterion at some point in the future, but heterozygosity estimates for many species can be generated today as conservationists struggle with the ongoing biodiversity crises. Our study outlines the theoretical and empirical justification for a new GD criterion, a bioinformatic pipeline for estimating GD from publicly-available population genomic data, an analytical framework, and explicit recommendations for use by conservation authorities. We have illustrated our ideas using mammalian data, but they are applicable to most branches of the tree of life.

Materials and Methods

Overall workflow of this study is shown in Supplementary Figure S21. We evaluated five population genomic metrics that each has a strong theoretical justification for being conservation-informative: 1) mean nucleotide diversity (π); 2) Watterson’s theta (θ_W); 3) mean observed genome-wide heterozygosity (H); 4) Tajima’s D (D); and 5) the extent of autozygosity as measured by runs of homozygosity (ROH). The first, π , is the basic genetic diversity index that conveys the average number of nucleotide differences

per site between all pairs of sequences in a population⁵³. The second, θ_W , represents the number of mutations in a population under the infinite site assumption⁵⁴. Third, H measures the proportion of heterozygous sites considered in a given sample⁵⁵. At the population level, mean H is averaged across all individual estimates. Fourth, D is computed as the difference between π and θ_W , divided by its variance under mutation-drift equilibrium⁵⁶. Tajima's D can be used to identify signatures of selection on individual loci, but demographic trends can also be detected when it is measured across the genome: $D < 0$ indicates population growth after a bottleneck, $D = 0$ indicates population stability, and $D > 0$ indicates a sudden population decline. Lastly, ROHs describe the proportion of contiguous homozygous regions along the genome and can be used to directly estimate both the extent and timing of inbreeding (and indirectly, the level of inbreeding depression due to associated reductions in fitness⁵⁷). We calculated two ROH estimators, namely $F_{\text{ROH}>100\text{kb}}$ (the fraction of ROH longer than 100 kb, $F_{100\text{kb}}$; representing the cumulative inbreeding level) and $F_{\text{ROH}>1\text{Mb}}$ (the fraction of ROH longer than 1 Mb, $F_{1\text{Mb}}$; representing the recent inbreeding level). We could not estimate $F_{\text{ROH}>1\text{Mb}}$ in some species because of low contiguity of their reference genome and so could not determine if the absence of detection of long ROHs was due to their true (biological) absence or if absence of detection was due to technical factors such as being scattered across contigs.

Data collection

We collected four types of data from public databases: 1) reference genomes; 2) population-level whole genome resequencing (WGR) reads for the inference of GD metrics; 3) IUCN Red List information; and 4) ecological characteristics (trophic level, body mass, and habitat breadth) for statistical tests of association with the GD metrics. We used subspecies or regional population level data whenever available, because conservation status can vary among demographically independent populations within the same species.

We searched all the available reference genomes of mammalian species (as of 2021) from NCBI and collected assembly identifiers (e.g., accession number and assembly name) required for our bioinformatic pipeline. We also collected additional information on the assembly level (i.e., contig, scaffold, or chromosome), contig N50, scaffold N50, and assembly size for downstream analyses (Supplementary Dataset S1). Species with a reference genome were further searched and population-level WGR data were accessed via NCBI's Sequence Read Archive (SRA). We use population-level data when: 1) "WGS" type data was available; 2) the data were comprised of paired-end reads; 3) the data were sequenced with Illumina technologies, such as Genome Analyzer, HiSeq, Novaseq, or Nextseq platforms; and lastly 4) a minimum of 2 different individuals from the same wild population were available. We followed the data author's population designation and limited the maximum number of individuals to 25 for computational tractability. We recorded the types of sequencing chemistry (i.e., 2-channel or 4-channel) and the number of individuals for downstream use (Supplementary Dataset S1, <https://github.com/AnnaBrunicheOlsen/theta>).

For each species evaluated, we used the IUCN Red List to record conservation category (i.e., CR - Critically Endangered, EN - Endangered, VU - Vulnerable, NT - Near Threatened, LC - Least Concern), population trend (i.e., decreasing, stable, or increasing), and extent of geographic range. We imported the shape file of species geographic range to ArcGIS Pro 2.9.0⁵⁸ and calculated the total habitat ranges except for “Extinct” and “Possibly Extinct” species. The shape files were clipped to match with “subspecies” or “subpopulation” of the collection site whenever apparent and applicable based on the associated metadata.

Bioinformatic analysis

We downloaded each species’ reference genome assembly, sorted them by length using BBMap (<https://sourceforge.net/projects/bbmap/>) and indexed each using samtools⁵⁹. If mitochondrial sequences were labeled in the assembly, they were culled. Short scaffolds less than 100 kb were also removed, then the resultant assembly was indexed again. Repeat files were downloaded from the assembly file if readily available or were created by running RepeatMasker⁶⁰ with “rush” option using the mammalian repeat database.

For each species, WGR SRA files (fastq format) were downloaded using the sra-toolkit. We employed fastqc⁶¹ to check the raw quality of downloaded fastq files and TrimGalore (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) to cull adapter sequences (using “very stringent” setting), low quality ends (less than 20nt) or reads of short length (less than 30nt), and read pairs of short length (less than 30nt). Sequence quality was checked again after filtering and samples where < 80% of reads passed quality filters were removed from downstream analyses.

Quality-filtered SRA reads from each species were mapped onto the respective preprocessed reference assembly using bwa-mem⁶² after creating a reference genome dictionary using Picard tools (<http://broadinstitute.github.io/picard/>). To improve read mapping quality, we locally realigned reads using GATK’s “RealignerTargetCreator” and “IndelRealigner” tools⁶³. We used samtools to estimate summary statistics (mapping rate, depth and breadth of coverage) from the resultant bam files and species data of low quality (< 80% mapping rate, depth, or breadth) were removed. We estimated mappability using genmap⁶⁴ with 100-bp k-mer setting allowing two mismatches. Sites with low mappability (< 1) were not considered. Non-repeat regions were identified from the length-filtered reference genome using bedtools complement⁶⁵. Intersecting regions among the non-repeat regions, regions of mappability = 1, and scaffolds longer than 100kb were identified using bedtools.

We estimated GD metrics using ANGSD⁶⁶ and bcftools. We applied conservative filters in ANGSD, including removing low quality reads and ambiguously mapped reads. We estimated genotype likelihoods with GATK and maximum likelihood estimates of the folded site frequency spectrum were obtained using the realSFS tool. We estimated π , θ_W , and D applying a sliding window approach with non-overlapping 50 kb windows. Individual genome-wide H was estimated using a similar process and averaged to provide a population-level mean H estimate for each species. To estimate the ROH burden, a

bcf file was generated using ANGSD from bam files. Subsequently, bcftools/roh⁶⁷ was employed to identify individual ROHs applying the hidden Markov model. The fraction of ROHs in individual genomes were averaged to obtain a population-level estimate per species using an in-house python script.

Statistical analysis

Descriptive statistics (mean and standard deviation) and the distribution of GD metrics were summarized and plotted by both IUCN categories and taxonomic Orders. Before the main analyses described below, we partitioned the full dataset into two. The first data partition, the “IUCN dataset” (Supplementary Dataset S2), included all the species having their own categorical Red List assessment but excluded those listed as “Data-Deficient”. The second data partition, the “EcoEvo dataset” (Supplementary Dataset S3), included all the species having data from the COMBINE database⁶⁸.

For the main comparison between GD and IUCN categorization, uncorrelated GD metrics were first identified using Pearson’s correlation test. Uncorrelated GD metrics were then individually tested against IUCN categories (including binary-transformed categories where Threatened = CR + EN + VU versus Non-Threatened = NT + LC) to determine if there was a significant difference between mean GD values across IUCN categories. We considered technical factors as well (e.g., sequence read depth) and controlled for them in the statistical tests (Technical Dimensions; Dim.1–Dim.4) as described in A1. To account for phylogenetic signal (λ) we ran Phylogenetic Generalized Least Squares models (PGLS; GD ~ IUCN category + Dim.1 + Dim.2 + Dim.3 + Dim.4 - 1) using the R package ‘caper’⁶⁹ with the maximum likelihood method. The mammalian phylogenetic tree used in the models was derived from VertLife⁷⁰ with sampling 1,000 trees from the “Mammals birth-death node-dated completed trees” distribution⁷⁰. The ‘averageTree’ function with default option in R package ‘phytools’⁷¹ was applied to obtain a consensus tree from the 1,000 trees, then rooted with *Sarcophilus harrisii* as an outgroup⁷² using the ‘root’ function in R package ‘ape’⁷³. Additional tips for each subspecies were manually added to the tree as a sister taxon of its closest relative using the ‘AddTip’ function in R package ‘TreeTools’⁷⁴. Effect sizes of significant independent variables of interest were reported as partial omega-squared using R package ‘sjstats’⁷⁵ and model comparisons were assessed using Akaike Information Criterion (AIC) values. Two Red List assessment criteria, “population trend” and “geographic range”, were compared in place of the IUCN category with the same procedure above.

We used ordinal regression tests to examine the explanatory power of GD metrics that were significantly correlated with IUCN categorization after accounting for phylogenetic non-independence. Each model consisted of IUCN full categories or of the binary categories (i.e., Threatened vs. Non-Threatened) as a dependent variable and one of the significant GD metrics as an independent variable. IUCN categories were treated as pseudo-continuous following⁷⁶.

We considered several machine learning (random forest, k-nearest neighbors, and support vector machine) classifier models using the ‘scikit-learn’ package⁷⁷ in Python 3.9. These machine learning algorithms more efficiently capture potential non-linear relationships, such as those between GD and

IUCN categories, and are applicable to small datasets⁷⁸. All GD metrics identified as significant in linear models were included as predictors, whereas IUCN binary categories were included as responses. For the random forest and linear support vector machine classifiers, we used predictors together and compared their feature importance. For k-nearest neighbors and non-linear support vector machine classifier, we used predictors individually and compared their model accuracy. Predictors were standardized in k-nearest neighbors and support vector machine models. We used 70% of the data for model training and 30% for model testing. After hyperparameter tuning by a wide range of randomized grid searches and/or a finer parameter grid search, the final random forest classifier was set as described in Supplementary Table S3.

We tested associations between IUCN categories and a) population trend and b) geographic range estimates, two criteria currently used to help determine IUCN status. We did so to provide perspective on the signal (or lack thereof) contained in GD metrics. The two GD metrics which best predicted IUCN status were used to plot the global distribution of GD values on a global map. We collected geo-coordinates of the WGR data from their original sources or by using the GeoNames database (<https://www.geonames.org/>) when a coordinate was unavailable.

To strengthen our conservation-oriented analyses by accounting for potential confounding factors, the distribution of all the GD metrics, conservation criteria and eco-evolutionary factors across species was displayed on a Multi Factor Analysis (MFA) plot using R packages 'FactoMineR'⁷⁹ with the first two dimensions. We also exploited the data to address evolutionary questions. We compared GD against key eco-evolutionary factors that could drive levels of standing GD, including trophic level, habitat breadth, and body mass. See Appendix A2 for details.

Declarations

Data and code availability

Genomic data collected from the NCBI (species, population, accession numbers for reference assembly, BioProject numbers for WGR, etc.) are reported in Supplementary Table S6. Data for statistical analyses were shared as Supplementary Dataset S1-S3. Bash, Python, and R scripts used in the study were uploaded to the GitHub repository: <https://github.com/AnnaBrunicheOlsen/theta>.

Acknowledgments

We thank Dr. Esteban Fernandez-Juricic for his careful advice on statistical approaches. We also thank Dr. Avril Harder for thorough advice on ROH calculations and Jeeyung Kim for advice on the machine learning analyses procedure. This work was supported in part by U.S. Department of Agriculture Hatch project 1025651 to JRW. ABO was supported by a Carlsberg Foundation Reintegration Fellowship (grant no. CF19-0427) and JAD was supported in part by the USDA's National Institute for Food and Agriculture.

Author Contributions

JAD originally envisioned the project; JRW, ANB, ABO, and JWB helped refine project design; JYJ led project execution; JYJ, EJH, AJM, GFL, and SJ collected the data, JYJ conducted most statistical analyses, and all authors helped refine the original writing of JYJ and JAD.

Competing Interest Statement

We declare no competing interests, but JAD is a member of IUCN's Species Survival Commission (SSC) North America and of the IUCN SSC Conservation Genetics Specialist Group.

References

1. Schlaepfer, D. R., Braschler, B., Rusterholz, H.-P. & Baur, B. Genetic effects of anthropogenic habitat fragmentation on remnant animal and plant populations: a meta-analysis. *Ecosphere* **9**, e02488 (2018).
2. Almeida-Rocha, J. M., Soares, L. A. S. S., Andrade, E. R., Gaiotto, F. A. & Cazetta, E. The impact of anthropogenic disturbances on the genetic diversity of terrestrial species: A global meta-analysis. *Mol Ecol* **29**, 4812–4822 (2020).
3. Ceballos, G. *et al.* Accelerated modern human-induced species losses: Entering the sixth mass extinction. *Sci Adv* **1**, e1400253 (2015).
4. Van Der Valk, T. *et al.* Significant loss of mitochondrial diversity within the last century due to extinction of peripheral populations in eastern gorillas. *Sci Rep* **8**, 1–10 (2018).
5. Gilpin, M. E. & Soulé, M. E. Minimum viable populations: processes of extinction. in *Conservation Biology: The Science of Scarcity and Diversity* (ed. Soulé, M. E.) 19–34 (Sinauer Associates, 1986).
6. Blomqvist, D., Pauliny, A., Larsson, M. & Flodin, L.-Å. Trapped in the extinction vortex? Strong genetic effects in a declining vertebrate population. *BMC Evol Biol* **10**, 1–9 (2010).
7. DeWoody, J. A., Harder, A. M., Mathur, S. & Willoughby, J. R. The long-standing significance of genetic diversity in conservation. *Mol Ecol* **30**, 4147–4154 (2021).
8. England, P. R. *et al.* Effects of intense versus diffuse population bottlenecks on microsatellite genetic diversity and evolutionary potential. *Conserv Genet* **4**, 595–604 (2003).
9. Frankham, R. Genetics and extinction. *Biol Conserv* **126**, 131–140 (2005).
10. Willoughby, J. R. *et al.* The reduction of genetic diversity in threatened vertebrates and new recommendations regarding IUCN conservation rankings. *Biol Conserv* **191**, 495–503 (2015).
11. Garner, B. A., Hoban, S. & Luikart, G. IUCN Red List and the value of integrating genetics. *Conserv Genet* **21**, 795–801 (2020).
12. Laikre, L. *et al.* Post-2020 goals overlook genetic diversity. *Science* **367**, 1083–1085 (2020).
13. van Oosterhout, C. Mutation load is the spectre of species conservation. *Nat Ecol Evol* **4**, 1004–1006 (2020).
14. Reed, D. H. & Frankham, R. Correlation between fitness and genetic diversity. *Conserv Biol* **17**, 230–237 (2003).

15. Kardos, M. *et al.* The crucial role of genome-wide genetic variation in conservation. *Proc Natl Acad Sci* **118**, e2104642118 (2021).
16. Brüniche-Olsen, A. *et al.* The inference of gray whale (*Eschrichtius robustus*) historical population attributes from whole-genome sequences. *BMC Evol Biol* **18**, 87 (2018).
17. Khan, A. *et al.* Genomic evidence for inbreeding depression and purging of deleterious genetic variation in Indian tigers. *Proc Natl Acad Sci* **118**, e2023018118 (2021).
18. Rosel, P. E., Wilcox, L. A., Yamada, T. K. & Mullin, K. D. A new species of baleen whale (Balaenoptera) from the Gulf of Mexico, with a review of its geographic distribution. *Mar Mamm Sci* **37**, 577–610 (2021).
19. Petit-Marty, N., Vázquez-Luis, M. & Hendriks, I. E. Use of the nucleotide diversity in COI mitochondrial gene as an early diagnostic of conservation status of animal species. *Conserv Lett* **14**, e12756 (2021).
20. Genereux, D. P. *et al.* A comparative genomics multitool for scientific discovery and conservation. *Nature* **587**, 240–245 (2020).
21. Wilder, A. P. *et al.* The contribution of historical processes to contemporary extinction risk in placental mammals. *Science* **380**, eabn5856 (2023).
22. Brüniche-Olsen, A., Kellner, K. F., Belant, J. L. & DeWoody, J. A. Life-history traits and habitat availability shape genomic diversity in birds: implications for conservation. *Proc R Soc B* **288**, 20211441 (2021).
23. Nei, M. & Roychoudhury, A. K. Sampling variances of heterozygosity and genetic distance. *Genetics* **76**, 379–390 (1974).
24. Gorman, G. C. & Renzi, J. Genetic distance and heterozygosity estimates in electrophoretic studies: Effects of sample size. *Copeia* **1979**, 242–249 (1979).
25. Bromham, L., Rambaut, A. & Harvey, P. H. Determinants of rate variation in mammalian DNA sequence evolution. *J Mol Evol* **43**, 610–621 (1996).
26. Romiguier, J. *et al.* Comparative population genomics in animals uncovers the determinants of genetic diversity. *Nature* **515**, 261–263 (2014).
27. Ellegren, H. & Galtier, N. Determinants of genetic diversity. *Nat Rev Genet* **17**, 422–433 (2016).
28. Anderson, S. J., Côté, S. D., Richard, J. H. & Shafer, A. B. A. Genomic architecture of phenotypic extremes in a wild cervid. *BMC Genom* **23**, 126 (2022).
29. McDonald, J. S. & Miller, K. v. *A history of white-tailed deer restocking in the United States, 1878 to 2004*. (Quality Deer Management Association, 2004).
30. Nevo, E., Beiles, A. & Ben-Shlomo, R. The evolutionary significance of genetic diversity: ecological, demographic and life history correlates. in *Evolutionary Dynamics of Genetic Diversity: Proceedings of a Symposium held in Manchester, England, March 29–30, 1983* (ed. Mani, G. S.) 132–213 (Springer-Verlag, 1984).
31. Mitton, J. B. Molecular approaches to population biology. *Annu Rev Ecol Syst* **25**, 45–69 (1994).

32. James, J. & Eyre-Walker, A. Mitochondrial DNA sequence diversity in mammals: a correlation between the effective and census population sizes. *Genome Biol Evol* **12**, 2441–2449 (2020).
33. Flight, P. A. Phylogenetic comparative methods strengthen evidence for reduced genetic diversity among endangered tetrapods. *Conserv Biol* **24**, 1307–1315 (2010).
34. Leffler, E. M. *et al.* Revisiting an old riddle: what determines genetic diversity levels within species? *PLoS Biol* **10**, e1001388 (2012).
35. Frankham, R. Relationship of genetic variation to population size in wildlife. *Conserv Biol* **10**, 1500–1508 (1996).
36. Zhang, L. *et al.* Maintenance of genome sequence integrity in long- and short-lived rodent species. *Sci Adv* **7**, eabj3284 (2021).
37. Nonaka, E. *et al.* Scaling up the effects of inbreeding depression from individuals to metapopulations. *J Anim Ecol* **88**, 1202–1214 (2019).
38. Nabholz, B., Mauffrey, J.-F., Bazin, E., Galtier, N. & Glemin, S. Determination of mitochondrial genetic diversity in mammals. *Genetics* **178**, 351–361 (2008).
39. Brüniche-Olsen, A., Kellner, K. F., Belant, J. L. & DeWoody, J. A. Life-history traits and habitat availability shape genomic diversity in birds: implications for conservation. *Proc R Soc B* **288**, 20211441 (2021).
40. Schmidt, C., Hoban, S., Hunter, M., Paz-Vinas, I. & Garroway, C. J. Genetic diversity and IUCN Red List status. *Conserv Biol* **37**, e14064 (2023).
41. Mace, G. M. & Lande, R. Assessing extinction threats: toward a reevaluation of IUCN threatened species categories. *Conserv Biol* **5**, 148–157 (1991).
42. Laikre, L. Genetic diversity is overlooked in international conservation policy implementation. *Conserv Genet* **11**, 349–354 (2010).
43. Crow, J. F. & Kimura, M. *Introduction to Population Genetics Theory*. (Harper & Row Publishers, 1970).
44. Palstra, F. P. & Ruzzante, D. E. Genetic estimates of contemporary effective population size: what can they tell us about the importance of genetic stochasticity for wild population persistence? *Mol Ecol* **17**, 3428–3447 (2008).
45. Santiago, E. *et al.* Recent demographic history inferred by high-resolution analysis of linkage disequilibrium. *Mol Biol Evol* **37**, 3642–3653 (2020).
46. Liu, X. & Fu, Y.-X. Stairway Plot 2: demographic history inference with folded SNP frequency spectra. *Genome Biol* **21**, 280 (2020).
47. Waples, R. S. Conservation genetics of Pacific salmon. III. Estimating effective population size. *J Hered* **81**, 277–289 (1990).
48. Patton, A. H. *et al.* Contemporary demographic reconstruction methods are robust to genome assembly quality: A case study in Tasmanian devils. *Mol Biol Evol* **36**, 2906–2921 (2019).
49. Lande, R. Genetics and demography in biological conservation. *Science* **241**, 1455–1460 (1988).

50. Lynch, M. & Lande, R. The critical effective size for a genetically secure population. *Anim Conserv* **1**, 70–72 (1998).
51. Frankham, R., Bradshaw, C. J. A. & Brook, B. W. Genetics in conservation management: revised recommendations for the 50/500 rules, Red List criteria and population viability analyses. *Biol Conserv* **170**, 56–63 (2014).
52. Allendorf, F. W. & Ryman, N. The role of genetics in population viability analysis. in *Population viability analysis* (eds. Beissinger, S. R. & McCullough, D. R.) (University of Chicago Press, 2002).
53. Nei, M. & Li, W.-H. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc Natl Acad Sci* **76**, 5269–5273 (1979).
54. Watterson, G. A. On the number of segregating sites in genetical models without recombination. *Theor Popul Biol* **7**, 256–276 (1975).
55. Nei, M. Estimation of average heterozygosity and genetic distance from a small number of individuals. *Genetics* **89**, 583–590 (1978).
56. Tajima, F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**, (1989).
57. Ceballos, F. C., Joshi, P. K., Clark, D. W., Ramsay, M. & Wilson, J. F. Runs of homozygosity: windows into population history and trait architecture. *Nat Rev Genet* **19**, 220–234 (2018).
58. Esri. ArcGIS Pro. (2021).
59. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
60. Smit, A., Hubley, R. & Green, P. RepeatMasker Open-4.0. 2013-2015. <http://www.repeatmasker.org> (2015).
61. Andrews, S. FastQC: A Quality Control Tool for High Throughput Sequence Data. <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/> (2010).
62. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
63. van der Auwera, G. & O’Connor, B. *Genomics in the Cloud: Using Docker, GATK, and WDL in Terra*. (O’Reilly Media, 2020).
64. Pockrandt, C., Alzamel, M., Iliopoulos, C. S. & Reinert, K. GenMap: ultra-fast computation of genome mappability. *Bioinformatics* **36**, 3687–3692 (2020).
65. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
66. Korneliussen, T. S., Albrechtsen, A. & Nielsen, R. ANGSD: analysis of next generation sequencing data. *BMC Bioinform* **15**, 1–13 (2014).
67. Narasimhan, V. *et al.* BCFtools/RoH: a hidden Markov model approach for detecting autozygosity from next-generation sequencing data. *Bioinformatics* **32**, 1749–1751 (2016).

68. Soria, C. D., Pacifici, M., Di Marco, M., Stephen, S. M. & Rondinini, C. COMBINE: a coalesced mammal database of intrinsic and extrinsic traits. *Ecology* **102**, e03344 (2021).
69. Orme, D. *et al.* caper: Comparative Analyses of Phylogenetics and Evolution in R. (2018).
70. Upham, N. S., Esselstyn, J. A. & Jetz, W. Inferring the mammal tree: species-level sets of phylogenies for questions in ecology, evolution, and conservation. *PLoS Biol* **17**, e3000494 (2019).
71. Revell, L. J. phytools: an R package for phylogenetic comparative biology (and other things). *Methods Ecol Evol* 217–223 (2012).
72. Damas, J. *et al.* Evolution of the ancestral mammalian karyotype and syntenic regions. *Proc Natl Acad Sci* **119**, e2209139119 (2022).
73. Paradis, E. & Schliep, K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35**, 526–528 (2019).
74. Smith, M. TreeTools: create, modify and analyse phylogenetic trees. (2019).
75. Lüdtke, D. sjstats: Statistical Functions for Regression Models. (2022).
76. Graber, S. Phylogenetic comparative methods for discrete responses in evolutionary biology. *Thesis*, (2013).
77. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *J Mach Learn Res* **12**, 2825–2830 (2011).
78. Osisanwo, F. Y. *et al.* Supervised machine learning algorithms: classification and comparison. *IJCTT* **48**, 128–138 (2017).
79. Lê, S., Josse, J. & Husson, F. FactoMineR: An R Package for Multivariate Analysis. *J Stat Softw* **25**, 1–18 (2008).

Figures

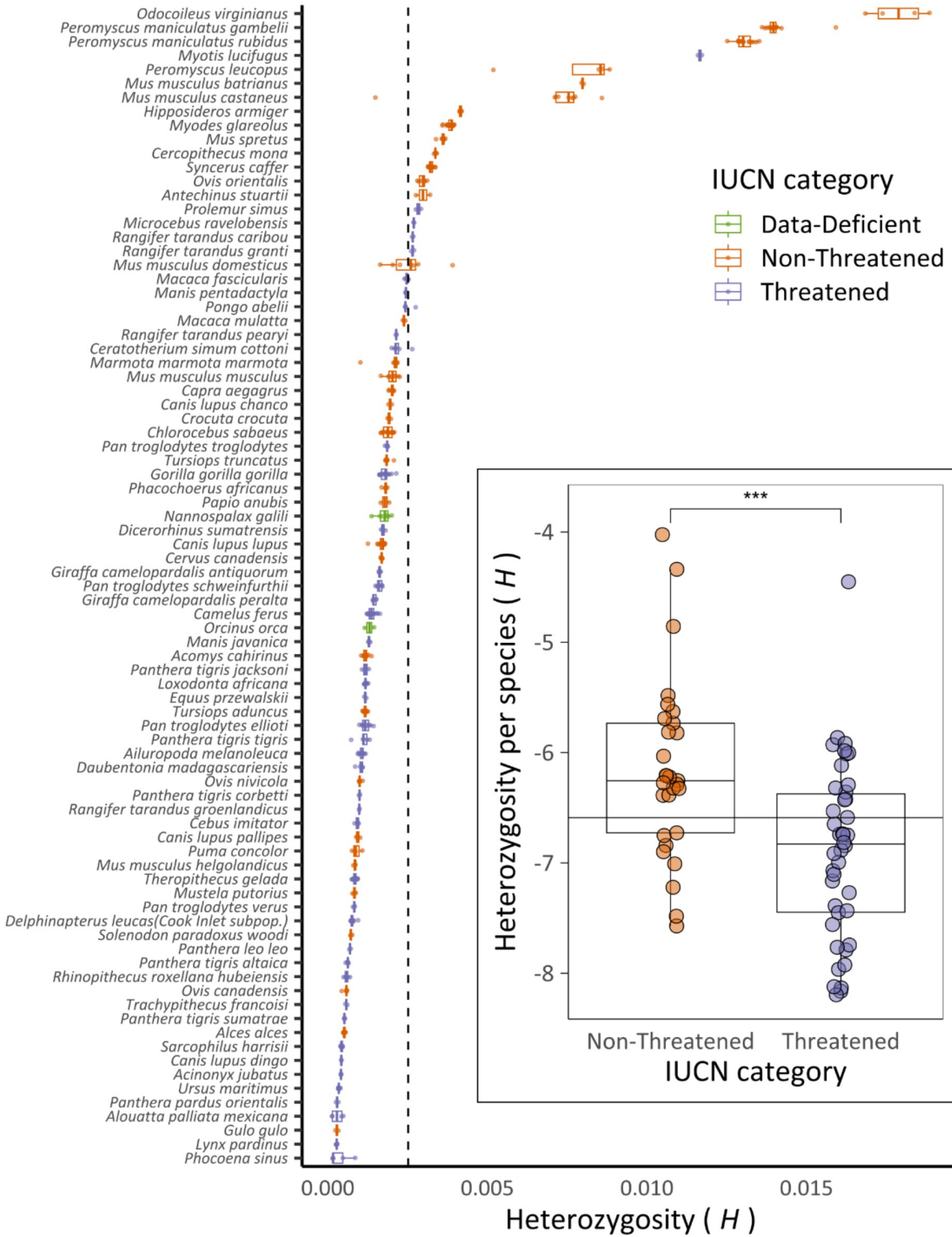


Figure 1

A box plot of heterozygosity by species. Species are arranged by descending median value of heterozygosity and colored by IUCN Threatened/Non-Threatened categories, plus “Data-Deficient”. Dashed line indicates the overall mean value. It should be noted, however, that *Odocoileus virginianus* WGR data is pooled-sequencing which should contribute to the high GD value, and is included in this study since the NCBI SRA does not separately categorize “pooled-seq” from “WGS” data type. Species

names according to NCBI are shown on the y-axis. The inset shows a box plot of log-transformed observed heterozygosity against IUCN categories. Non-Threatened category is compared to Threatened category using a Wilcoxon test and the significance is shown (***: $p < 0.001$). Dashed line indicates the overall mean value.

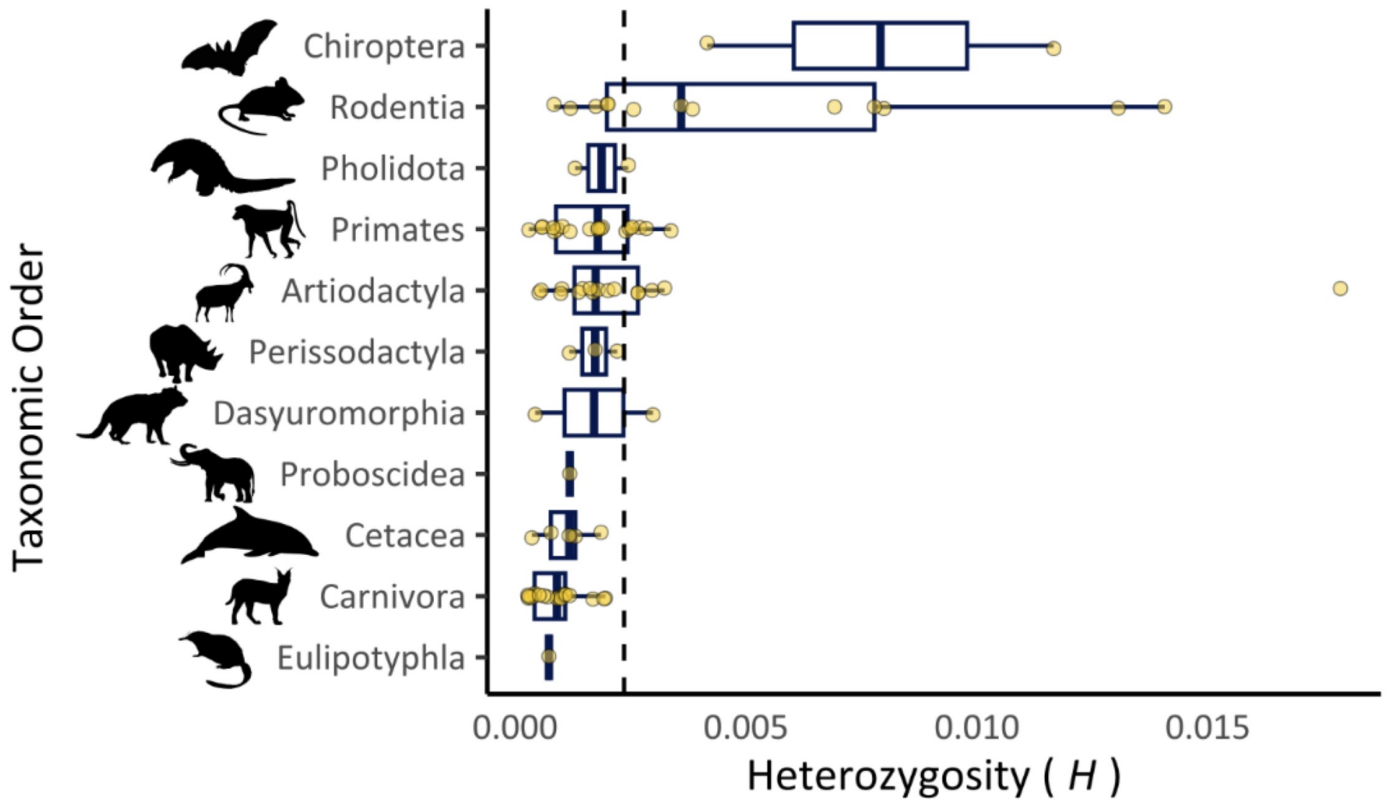


Figure 2

A box plot of observed heterozygosity by taxonomic Order. Taxonomic Orders are arranged by descending median value of heterozygosity. Dashed line indicates the overall mean value. Silhouette images of animals are adapted from PhyloPic (<https://www.phylopic.org/>).

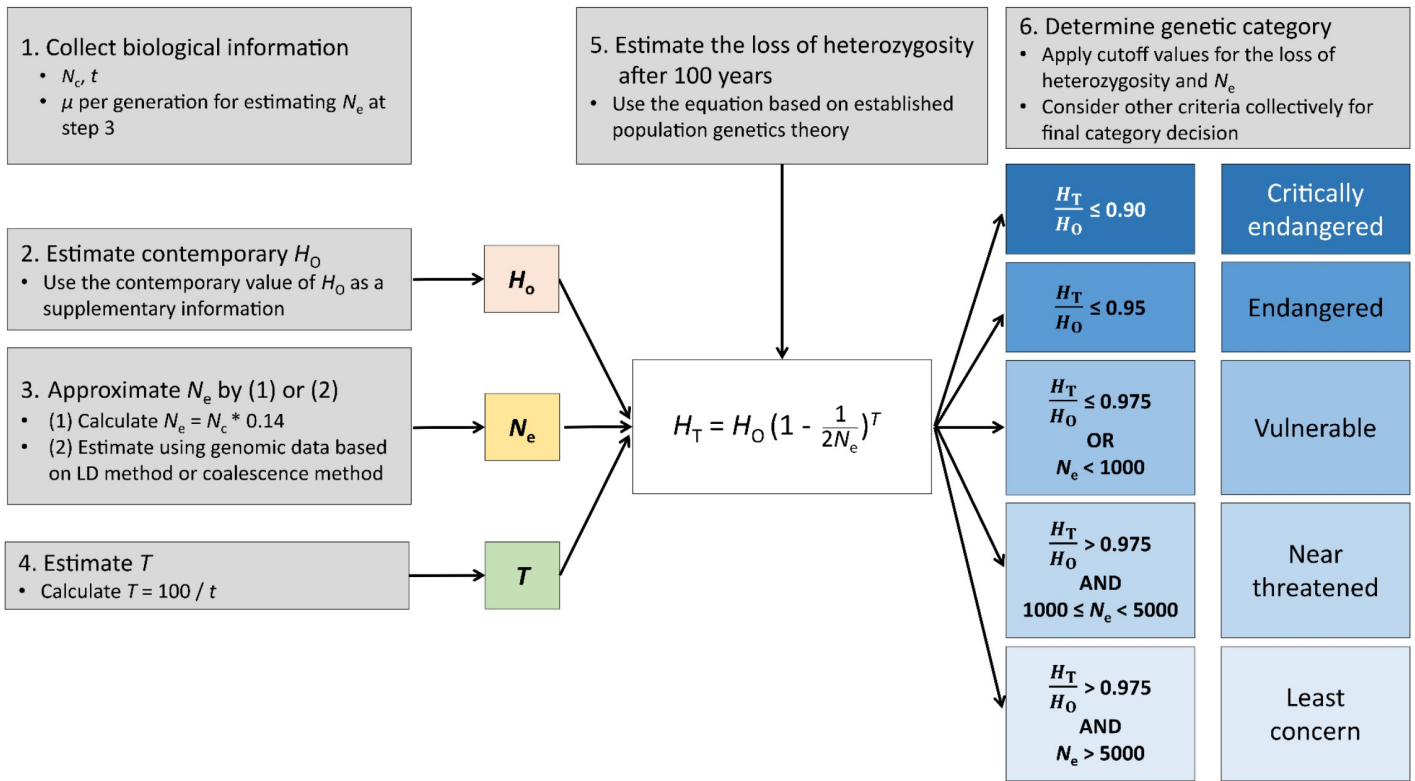
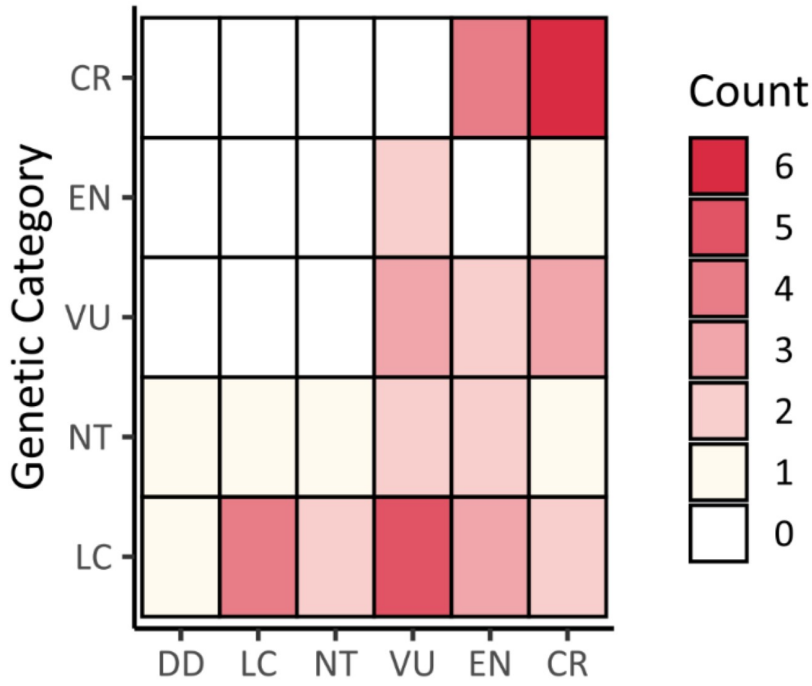


Figure 3

A practical outline of how genetic or genomic diversity could be explicitly used to help determine formal conservation status. Using reliable scientific knowledge, the loss of heterozygosity after 100 years can be predicted and used to determine conservation status according to cutoff thresholds. Abbreviations: N_c = census population size, t = generation time, μ = mutation rates, H_0 = observed heterozygosity, N_e = effective population size, LD = Linkage-disequilibrium, T = the number of generations in 100 years, H_T = reduced heterozygosity after T generations.

(a)



(b)

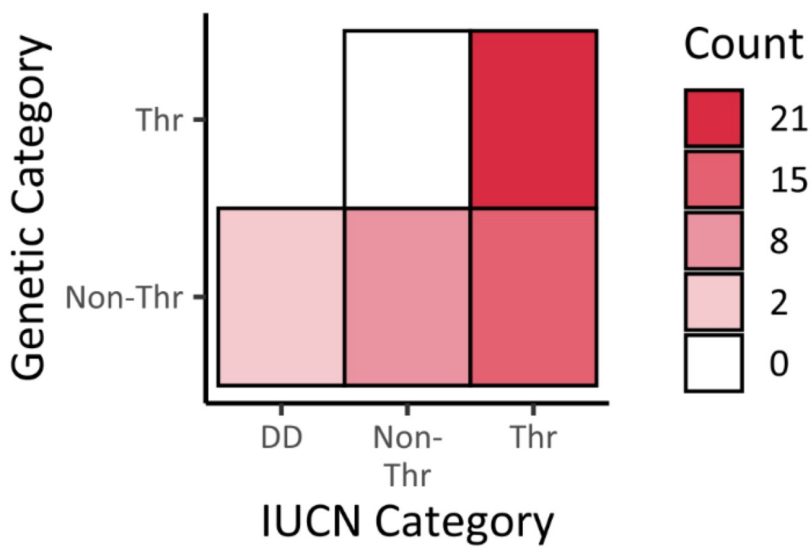


Figure 4

The comparison between original, official IUCN categories and hypothetical categories based solely on the genetic criterion described in this study, whether full (a) or binary (b) plus “Data-Deficient”. Abbreviations: LC - Least Concern, NT - Near Threatened, VU - Vulnerable, EN - Endangered, CR - Critically Endangered, DD - Data Deficient, Thr - Threatened.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [ThetaNatSustsubmissionSI.docx](#)
- [SupplementaryTableS5.xlsx](#)
- [SupplementaryTableS6.xlsx](#)
- [SupplementaryDatasetS1.xlsx](#)
- [SupplementaryDatasetS2.xlsx](#)
- [SupplementaryDatasetS3.xlsx](#)