The Zoonomia Project is investigating the genomics of shared and specialized traits in eutherian mammals. Here we provide genome assemblies for 131 species, of which all but 9 are previously uncharacterized, and describe a whole-genome alignment of 240 species of considerable phylogenetic diversity, comprising representatives from more than 80% of mammalian families. We find that regions of reduced genetic diversity are more abundant in species at a high risk of extinction, discern signals of evolutionary selection at high resolution and provide insights from individual reference genomes. By prioritizing phylogenetic diversity and making data available quickly and without restriction, the Zoonomia Project aims to support biological discovery, medical research and the conservation of biodiversity.

Designing a comparative-genomics multitool

When selecting species, we sought to maximize evolutionary branch length, to include at least one species from each eutherian family, and to prioritize species of medical, biological or biodiversity conservation interest. Our assemblies increase the percentage of eutherian families that all but 9 are previously uncharacterized, and include 9 species that are the sole extant member of their family and 7 species that are critically endangered* (Fig. 1): the Mexican howler monkey (Alouatta palliata mexicana), hirola (Beatragus hunteri), Russian saiga (Saiga tatarica tatarica), social tuco-tuco (Ctenomys sociabilis), indri (Indri indri), northern white rhinoceros (Ceratotherium simum cottoni) and black rhinoceros (Diceros bicornis).

Comparative power of 240 species

The Zoonomia alignment includes 120 newly generated assemblies and 121 existing assemblies, representing a total of 240 species (the
dataset includes assemblies for two different dogs) and spanning about 110 million years of mammalian evolution (Supplementary Table 2). With a total evolutionary branch length of 16.6 substitutions per site, we expect only 191 positions in the human genome (0.000006%) to be identical across the aligned species owing to chance (false positives) rather than evolutionary constraint (Extended Data Table 2). We applied this same calculation to data from The Exome Aggregation Consortium (ExAC) — who analysed exomes for 60,706 humans — and estimated that 88% of positions would be expected to have no variation. This illustrates the potential for relatively small cross-species datasets to inform human genetic studies — even for diseases driven by high-penetration coding mutations, for which ExAC data are optimally powered.

Biological insights from additional assemblies

The scope and species diversity in the Zoonomia Project supports evolutionary studies in many lineages. Previously published papers (discussed in the subsections below), and the demonstrated utility of existing comparative genomics resources 

 dataset includes assemblies for two different dogs) and spanning about 110 million years of mammalian evolution (Supplementary Table 2). With a total evolutionary branch length of 16.6 substitutions per site, we expect only 191 positions in the human genome (0.000006%) to be identical across the aligned species owing to chance (false positives) rather than evolutionary constraint (Extended Data Table 2). We applied this same calculation to data from The Exome Aggregation Consortium (ExAC) — who analysed exomes for 60,706 humans — and estimated that 88% of positions would be expected to have no variation. This illustrates the potential for relatively small cross-species datasets to inform human genetic studies — even for diseases driven by high-penetration coding mutations, for which ExAC data are optimally powered.

Biological insights from additional assemblies

The scope and species diversity in the Zoonomia Project supports evolutionary studies in many lineages. Previously published papers (discussed in the subsections below), and the demonstrated utility of existing comparative genomics resources 

 dataset includes assemblies for two different dogs) and spanning about 110 million years of mammalian evolution (Supplementary Table 2). With a total evolutionary branch length of 16.6 substitutions per site, we expect only 191 positions in the human genome (0.000006%) to be identical across the aligned species owing to chance (false positives) rather than evolutionary constraint (Extended Data Table 2). We applied this same calculation to data from The Exome Aggregation Consortium (ExAC) — who analysed exomes for 60,706 humans — and estimated that 88% of positions would be expected to have no variation. This illustrates the potential for relatively small cross-species datasets to inform human genetic studies — even for diseases driven by high-penetration coding mutations, for which ExAC data are optimally powered. 

Biological insights from additional assemblies

The scope and species diversity in the Zoonomia Project supports evolutionary studies in many lineages. Previously published papers (discussed in the subsections below), and the demonstrated utility of existing comparative genomics resources 

 dataset includes assemblies for two different dogs) and spanning about 110 million years of mammalian evolution (Supplementary Table 2). With a total evolutionary branch length of 16.6 substitutions per site, we expect only 191 positions in the human genome (0.000006%) to be identical across the aligned species owing to chance (false positives) rather than evolutionary constraint (Extended Data Table 2). We applied this same calculation to data from The Exome Aggregation Consortium (ExAC) — who analysed exomes for 60,706 humans — and estimated that 88% of positions would be expected to have no variation. This illustrates the potential for relatively small cross-species datasets to inform human genetic studies — even for diseases driven by high-penetration coding mutations, for which ExAC data are optimally powered.

Biological insights from additional assemblies

The scope and species diversity in the Zoonomia Project supports evolutionary studies in many lineages. Previously published papers (discussed in the subsections below), and the demonstrated utility of existing comparative genomics resources 

 dataset includes assemblies for two different dogs) and spanning about 110 million years of mammalian evolution (Supplementary Table 2). With a total evolutionary branch length of 16.6 substitutions per site, we expect only 191 positions in the human genome (0.000006%) to be identical across the aligned species owing to chance (false positives) rather than evolutionary constraint (Extended Data Table 2). We applied this same calculation to data from The Exome Aggregation Consortium (ExAC) — who analysed exomes for 60,706 humans — and estimated that 88% of positions would be expected to have no variation. This illustrates the potential for relatively small cross-species datasets to inform human genetic studies — even for diseases driven by high-penetration coding mutations, for which ExAC data are optimally powered.

Biological insights from additional assemblies

The scope and species diversity in the Zoonomia Project supports evolutionary studies in many lineages. Previously published papers (discussed in the subsections below), and the demonstrated utility of existing comparative genomics resources 

 dataset includes assemblies for two different dogs) and spanning about 110 million years of mammalian evolution (Supplementary Table 2). With a total evolutionary branch length of 16.6 substitutions per site, we expect only 191 positions in the human genome (0.000006%) to be identical across the aligned species owing to chance (false positives) rather than evolutionary constraint (Extended Data Table 2). We applied this same calculation to data from The Exome Aggregation Consortium (ExAC) — who analysed exomes for 60,706 humans — and estimated that 88% of positions would be expected to have no variation. This illustrates the potential for relatively small cross-species datasets to inform human genetic studies — even for diseases driven by high-penetration coding mutations, for which ExAC data are optimally powered.

Biological insights from additional assemblies

The scope and species diversity in the Zoonomia Project supports evolutionary studies in many lineages. Previously published papers (discussed in the subsections below), and the demonstrated utility of existing comparative genomics resources 

 dataset includes assemblies for two different dogs) and spanning about 110 million years of mammalian evolution (Supplementary Table 2). With a total evolutionary branch length of 16.6 substitutions per site, we expect only 191 positions in the human genome (0.000006%) to be identical across the aligned species owing to chance (false positives) rather than evolutionary constraint (Extended Data Table 2). We applied this same calculation to data from The Exome Aggregation Consortium (ExAC) — who analysed exomes for 60,706 humans — and estimated that 88% of positions would be expected to have no variation. This illustrates the potential for relatively small cross-species datasets to inform human genetic studies — even for diseases driven by high-penetration coding mutations, for which ExAC data are optimally powered.
resolution to Peto’s paradox—the observation that cancer in large mammals is rarer than expected—and could reveal anti-cancer mechanisms.

**Convergent evolution of venom**

A previous publication has used our assembly for the Hispaniolan solenodon (Solenodon paradoxus) (Extended Data Fig. 2) to investigate venom production—a trait that is found in only a few eutherian lineages, including shrews and solenodons. They identified paralogous copies of a kallikrein 1 serine protease gene (KLK1) that together encode solenodon venom, and showed that the KLK1 gene was independently co-opted for venom production in both solenodons and shrews, in an example of molecular convergence.

**Informing biodiversity conservation strategies**

A previous analysis of our giant otter (Pteronura brasiliensis) assembly found low diversity and an elevated burden of putatively deleterious genetic variants, consistent with the recent population decline of this species through overhunting and habitat loss. The giant otter had fewer putatively deleterious variants than either the southern or northern sea otter (Enhydra lutris nereis and E. lutris kenyonii, respectively), which suggests that it has highest potential for recovery among these species if populations are protected.

**Rapid assessment of species infection risk**

Using the Zoonomia alignment and public genomic data from hundreds of other vertebrates, a previous publication compared the structure of ACE2—the receptor for severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), the causative agent of coronavirus disease 2019 (COVID-19)—and identified 47 mammals that have a high or very high likelihood of being virus reservoirs, intermediate hosts or good model organisms for the study of COVID-19, and detected positive selection in the ACE2 receptor-binding domain that is specific to bats.

**Genetic diversity and extinction risk**

We next asked whether a reference genome from a single individual can help to identify populations with low genetic diversity to prioritize in efforts to conserve biodiversity. Diversity metrics reflect demographic history, and heterozygosity is lower in threatened species. This analysis was feasible because we used a single sequencing and assembly protocol for all DISCOVAR assemblies, which minimized variation in accuracy, completeness and contiguity due to the sequencing technology and the assembly process that would otherwise confound species comparisons.

We estimated genetic diversity for 130 of our DISCOVAR assemblies, each of which represented a different species (Supplementary Table 3). Four of these estimates failed during analysis. For the remaining 126 DISCOVAR assemblies, we calculated 2 metrics: (1) the fraction of sites at which the sequenced individual is heterozygous (overall heterozygosity); and (2) the proportion of the genome that resides in an extended region without any variation (segments of homozygosity (SoH)). The SoH measurement is designed for short-contiguity assemblies, in which scaffolds are potentially shorter than runs of homozygosity. Overall, heterozygosity and SoH values are correlated (Pearson correlation $r = -0.39, P = 1.8 \times 10^{-4}, n = 98$. Although overall heterozygosity is correlated with contig N50 values (Pearson correlation $r_{cont} = -0.39$,
Genomic diversity varies significantly among species in different IUCN conservation categories, as measured by overall heterozygosity (Fig. 2a) and SoH values (Fig. 2b). SoH values increase ($P = 0.024$, $R^2 = 0.055, n = 94$) with increasing levels of conservation concern, whereas heterozygosity decreases ($P = 0.011, R^2 = 0.064, n = 101$). There is no significant difference between wild and captive populations in overall heterozygosity (Fig. 2c) or SoH values (Fig. 2d).

Unusual diversity values can suggest particular population demographics, although data from more than a single individual are needed to confirm these inferences. All seven critically endangered species have SoH values that are higher than the median for species categorized as of least concern (Fig. 2e). The genomes with the lowest heterozygosity and highest SoH values were the social tuco-tuco (heterozygosity = 0.00063 and SoH = 78.7%), which was sampled from a small laboratory colony with only 12 founders, and the eastern mole (Scalopus aquaticus) (heterozygosity = 0.0008 and SoH = 81.3%), which was supplied by a professional mole catcher and was probably from a population that had experienced a bottleneck owing to pest control measures.

The correlation between diversity metrics and IUCN category is not explained by other species-level phenotypes. For species of least concern ($n = 75$), we assessed 21 phenotypes that are catalogued in the PanTHERIA database for correlation with heterozygosity or SoH values. The most significant was between SoH value and litter size, a trait that has previously been shown to predict extinction risk ($P = 0.02$), but none is significant after Bonferroni correction (Extended Data Table 3).

Our inference that diversity trends lower in species at a higher risk of extinction comes from a small fraction (2.6%) of threatened mammals. Whether this is a direct correlation with extinction risk or arises from an association between diversity and species-level phenotypes such as litter size, it suggests that valuable information can be gleaned from sequencing only a single individual. Should this pattern prove robust across more species, diversity metrics from a single reference genome could help to identify populations that are at risk—even when few species-level phenotypes are documented—and to prioritize species for follow-up at the population level.

Resources for biodiversity conservation

For each genome assembly, we catalogued all high-confidence variant sites (http://broad.io/variants) to support the design of cost-effective and accurate genetic assays that are usable even when the sample quality is low; such assays are often preferable to designing expensive custom tools, relying on tools from related species or sequencing random regions. The reference genomes themselves support the development of technologies such as using gene drives to control invasive species or pursuing ‘de-extinction’ through cloning and genetic engineering.

Our genomes have two notable limitations. We sequenced only a single individual for each species, which is insufficient for studying population origins, population structure and recent demographic events, and the shorter contiguity of our assemblies prevented us from analysing runs of homozygosity. This highlights a dilemma that faces all large-scale genomics initiatives: determining when the value of sequencing additional individuals exceeds the value of improving the reference genome itself.

Whole-genome alignment

We aligned the genomes of 240 species (our assemblies and other mammalian genomes that were released when we started the alignment) as part of a 600-way pan-ammnoite alignment using the Cactus alignment software (Supplementary Table 2). Rather than aligning to a single anchor genome, Cactus infers an ancestral genome for each pair of assemblies (Fig. 3a). Consistent with our predictions, we have increased power to detect sequence constraint at individual bases relative to previous studies. We detect 3.1% of bases in the human genome to be under purifying selection in the eutherian lineage (false-discovery rate (FDR) < 5%), without using windowing or other means to integrate contextual information across neighbouring bases. This is more than double the number from the largest previous 100-vertebrate alignment (Fig. 3b), with improvements being most notable in the non-coding sequence (Fig. 3c) and in the increased resolution of individual features (Fig. 3d). This represents a substantial proportion—but not all—of the 5 to 8% of the human genome that has previously been suggested to be under purifying selection.

Next steps

Using our alignment of 240 mammalian genomes, we are pursuing four key strategies of analysis. First, we aim to provide the largest eutherian
phylogeny based on nuclear genomes by building a comprehensive phylogeny and time tree, including trees partitioned by functional annotations, mode of inheritance and long-term recombination rates. Second, we will produce a detailed map of evolutionary constraint, identifying highly conserved genomic regions, regions under accelerated evolution in particular lineages and changes that probably affect phenotype, leveraging functional data from ENCODE\textsuperscript{38}, GTEx\textsuperscript{41} and the Human Cell Atlas\textsuperscript{32}. Third, we will use genotype–phenotype correlations to investigate patterns of constraint in regions associated with disease in humans, identify patterns of convergent adaptive evolution\textsuperscript{7} and apply a forward genomics strategy to link functional elements to traits. Finally, we will explore the evolution of genome structure by mapping syntenic regions between genomes, identifying evolutionary breakpoints and characterizing the repeat landscape.

Conclusion

The Zoonomia Project has captured mammalian diversity at a high resolution, and is among the first of many projects that are underway to sequence, catalogue and characterize whole branches of the eukaryotic biodiversity of the Earth. On the basis of our experience, we propose the following principles for realizing the full value of large-scale comparative genomics.

First, we should prioritize sample collection. We must support field researchers who collect samples and understand species ecology and behaviour, develop strategies for sample collection that do not rely on bulky laboratory equipment or cold chains, develop technology for using non-invasive types of sampling and establish more repositories of renewable cell cultures\textsuperscript{10}.

Second, we need accessible and scalable tools for computational analysis. Few research groups have access to the computational resources necessary for work with massive genomic datasets. We must address the shortage of skilled computational scientists, and design software and data-storage systems that make powerful computational pipelines accessible to all researchers.

Finally, we should promote rapid data-sharing. Data embargoes must not be permitted to delay analyses that directly benefit the conservation of endangered species, human health or progress in basic science. Genomic data should be shared as quickly as possible and without restrictions on use.

Numerous large-scale genome-sequencing efforts are now underway, including the Earth BioGenome Project\textsuperscript{35}, Genome 10K\textsuperscript{46}, the Vertebrate Genomes Project, Bat IK\textsuperscript{15}, Bird 10K\textsuperscript{46} and DNA Zoo. As the number of genomes grows, so too will the usefulness of comparative genomics in disease research and the development of therapeutic strategies. Preserving, rather than merely recording, the biodiversity of the Earth must be a priority. Through global scientific collaborations, and by making genomic resources available and accessible to all research communities, we can ensure that the legacy of genomics is not a digital archive of lost species.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2876-6.
45. Teeling, E. C. et al. Bat biology, genomes, and the BatIK project: to generate chromosome-level genomes for all living bat species. *Annu. Rev. Anim. Biosci.* **6**, 23–46 (2018).
46. Feng, S. et al. Dense sampling of bird diversity increases power of comparative genomics. *Nature* https://doi.org/10.1038/s41586-020-2873-9 (2020).
47. Kumar, S., Stecher, G., Suleski, M. & Hedges, S. B. TimeTree: a resource for timelines, timeframes, and divergence times. *Mol. Biol. Evol.* **34**, 1812–1819 (2017).
48. Wilson, D. F. & Reeder, D. M. (eds) *Mammal Species of the World*. A Taxonomic and Geographic Reference 3rd edn (Johns Hopkins Univ. Press, 2005).
49. Vliegh, D. et al. A new generation of JASPAR, the open-access repository for transcription factor binding site profiles. *Nucleic Acids Res.* **34**, D95–D97 (2006).

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2020

Zoonomia Consortium

Diane P. Genereux1, Aitor Serres1, Joel Armstrong1, Jeremy Johnson1, Voichita D. Marinescu2, Eva Murén2, David Juan3, Gill Bejerano4,5,6, Nicholas R. Casewell9, Leona G. Chemnick10, Martin Nweeia24,25,26, Andreas R. Pfenning27, Katherine S. Pollard28,29,30, David A. Ray31, Eric S. Lander1,21,22, Jennifer R. S. Meadows4, William J. Murphy23, Will Nash13, Hyun Ji Noh1, Beth Shapiro32,33, Arian F. A. Smit17, Mark S. Springer34, Cynthia C. Steiner10, Ross Swofford1, Diane P. Genereux1, Aitor Serres1, Joel Armstrong1, Jeremy Johnson1, Voichita D. Marinescu2, Eva Murén2, David Juan3, Gill Bejerano4,5,6, Nicholas R. Casewell9, Leona G. Chemnick10, Martin Nweeia24,25,26, Andreas R. Pfenning27, Katherine S. Pollard28,29,30, David A. Ray31, Eric S. Lander1,21,22, Jennifer R. S. Meadows4, William J. Murphy23, Will Nash13, Hyun Ji Noh1, Beth Shapiro32,33, Arian F. A. Smit17, Mark S. Springer34, Cynthia C. Steiner10, Ross Swofford1, Eric S. Lander1,21,22, Jennifer R. S. Meadows4, William J. Murphy23, Will Nash13, Hyun Ji Noh1, Beth Shapiro32,33, Arian F. A. Smit17, Mark S. Springer34, Cynthia C. Steiner10, Ross Swofford1.

1Broad Institute of MIT and Harvard, Cambridge, Cambridge, MA, USA. 2Institute of Evolutionary Biology (UPF-CSIC), PRBB, Barcelona, Spain. 3UC Santa Cruz Genomics Institute, University of California Santa Cruz, Santa Cruz, CA, USA. 4Science for Life Laboratory, Department of Medical Biochemistry and Microbiology, Uppsala University, Uppsala, Sweden. 5Department of Biomedical Data Science, Stanford University, Stanford, CA, USA. 6Department of Computer Science, Stanford University, Stanford, CA, USA. 7Department of Developmental Biology, Stanford University, Stanford, CA, USA. 8Department of Pediatrics, Stanford University, Stanford, CA, USA. 9Centre for Snakebite Research and Interventions, Liverpool School of Tropical Medicine, Liverpool, UK. 10San Diego Zoo Global, Beckman Center for Conservation Research, San Diego, CA, USA. 11The UC Davis Genome Center, University of California, Davis, CA, USA. 12Department of Biological Sciences, University of East Anglia, Norwich, UK. 13Earlam Institute, Norwich, UK. 14Program in Bioinformatics and Integrative Biology, University of Massachusetts Medical School, Worcester, MA, USA. 15Bingham and Women’s Hospital, Harvard Medical School, Boston, MA, USA. 16Fauna Bio Incorporated, Emeryville, CA, USA. 17Institute for Systems Biology, Seattle, WA, USA. 18Department of Biochemistry, University of Cambridge, Cambridge, UK. 19Applied Tumor Genomics Research Program, University of Helsinki, Helsinki, Finland. 20Smithsonian-Mason School of Conservation, Front Royal, VA, USA. 21Department of Biology, MIT, Cambridge, MA, USA. 22Department of Systems Biology, Harvard Medical School, Boston, MA, USA. 23Veterinary Integrative Biosciences, Texas A&M University, College Station, TX, USA. 24Marine Mammal Program, Smithsonian Institution, Washington, DC, USA. 25Restorative Dentistry and Biomaterials Sciences, Harvard School of Dental Medicine, Boston, MA, USA. 26School of Dental Medicine, Case Western Reserve University, Cleveland, OH, USA. 27Carnegie Mellon University, School of Computer Science, Department of Computational Biology, Pittsburgh, PA, USA. 28Chan Zuckerberg Biohub, San Francisco, CA, USA. 29Glade Institute, San Francisco, CA, USA. 30Department of Epidemiology and Biostatistics, Institute for Computational Health Sciences and Institute for Human Genetics, University of California, San Francisco, CA, USA. 31Department of Biomedical Sciences, Texas Tech University, Lubbock, TX, USA. 32Department of Ecology and Evolutionary Biology, University of California, Santa Cruz, Santa Cruz, CA, USA. 33Howard Hughes Medical Institute, University of California Santa Cruz, Santa Cruz, CA, USA. 34Department of Evolution, Ecology and Organismal Biology, University of California, Riverside, Riverside, CA, USA. 35Department of Medical Biochemistry and Biophysics, Karolinska Institutet, Stockholm, Sweden. 36School of Biology and Environmental Science, University College Dublin, Dublin, Ireland. 37Department of Evolution, Behavior, and Ecology, Division of Biology, University of California, San Diego, La Jolla, CA, USA. 38Department of Evolution and Ecology, University of California, Davis, CA, USA. 39Catalan Institution of Research and Advanced Studies (ICREA), Barcelona, Spain. 40Institut Català de Paleontologia Miquel Crusafont, Institut de Paleontologia de Catalunya, Barcelona, Spain. 41CNAG-CRG, Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), Barcelona, Spain. 42Program in Molecular Medicine, University of Massachusetts Medical School, Worcester, MA, USA.

These authors contributed equally: Kerstin Lindblad-Toh, Elinor K. Karlsson.

✉ e-mail: elinor@broadinstitute.org
Analysis

Methods

The number of samples (species) required to detect evolutionary conservation at a single base was estimated by applying a Poisson model of the distribution of substitution counts in the genome.

Species selection, sample shipping and regulatory approvals
Species were selected to maximize branch length across the eutherian mammal phylogeny, and to capture genomes of species from previously unrepresented eutherian families. Of 172 species initially selected for inclusion, we obtained sufficiently high-quality DNA samples for genome sequencing for 137. DNA samples from collaborating institutions were shipped to the Broad Institute (n = 69) or Uppsala University (n = 68). For samples received at the Broad Institute that were then sent to Uppsala, shipping approval was secured from the US Fish and Wildlife Service. Institutional Animal Care and Use Committee approval was not required.

Sample quality control, library construction and sequencing
DNA integrity for each sample was visualized via agarose gel (at the Broad Institute) or Agilent tape station (at Uppsala University). Samples passed quality control if the bulk of DNA fragments were greater than 5 kb. DNA concentration was then determined using Invitrogen Qubit dsDNA HS assay kit. For each of the samples that passed quality control, 1–3 μg of DNA was fragmented on the Covaris E220 Instrument using the 400-bp standard programme (10% duty cycle, 140 PIP, 200 cycles per burst, 35 s). Fragmented samples underwent SPRI double-size selection (0.55×, 0.7×) followed by PCR-free Illumina library construction following the manufacturer’s instructions (Kapa no. KK8232) using PCR-free adapters from Illumina (no. FC-121-3001). Final library fragment size distribution was determined on Agilent 2100 Bioanalyzer with High Sensitivity DNA Chips. Paired-end libraries were pooled, and then sequenced on a single lane of the Illumina HiSeq2500, set for Version 2 chemistry and 2×200 bp reads. This yielded a total of mean 375 million (s.d. = 125 million) reads per sample.

Assembly and validation
For each species, we applied DISCOVAR de novo (discovardenovo-52488) (ftp://ftp.broadinstitute.org/pub/crd/DiscovarDeNovo/) to assemble the 2×250 bp read group, using the following command: DiscovarDeNovo READS = [READFILE] OUT_DIR = [SPECIES_ID]/[SPECIES_ID].discovar_files NUM_THREADS = 24 MAX_MEM_GB = 200G.

Coverage for each genome was automatically calculated by DISCOVAR, with a mean coverage of 40.1× (s.d. = 14×). We assessed genome assembly, gene set and transcriptome completeness using Benchmarking Universal Single-Copy Orthologs (BUSCO), which provides quantitative measures on the basis of gene content from near-universal single-copy orthologs. BUSCO was run with default parameters, using the mammalian gene model set (mammalia_odb9, n = 4,104), using the following command: python ./BUSCO.py -i [input fasta] -o [output file] -l ./mammalia_odb9/ -m genome -c 1 -sp. human.

Median contig N50 for existing RefSeq assemblies was calculated using the assembly statistics for the most recent release of 118 eutherian mammals with RefSeq assembly accession numbers. Assemblies were all classified as either reference genome or representative genome. Assembly statistics were downloaded from the NCBI on 10 April 2019.

Genome upgrades. We selected genomes from each eutherian order without a pre-existing long-contiguity assembly on the basis of (1) whether the underlying assembly met the minimum quality threshold needed for HiRise upgrades; and (2) whether a second sample of sufficient quality could be obtained from that individual. All upgrades were done with Dovetail Chicago libraries and assembled with HiRise 2.1, as previously described.

Estimating heterozygosity
Selection of assemblies for heterozygosity analysis. Heterozygosity statistics were calculated for all but four of our short read assemblies (n = 126) as well as eight Dovetail-upgraded genomes. Four failed because they were either too fragmented to analyse (n = 3) or because of undetermined errors (n = 1). One assembly was excluded because it was a second individual from a species that was already represented.

Heterozygosity calls. We applied the standard GATK pipeline with genotype quality banding to identify the callable fraction of the genome. First, we used samtools to subsample paired reads from the unmapped .bam files. After removing adapter sequences from the selected reads, we used BWA-MEM to map reads to the reference genome scaffolds of >10 kb, removing duplicates using the Picard Tools MarkDuplicates utility. We then called heterozygous sites using standard GATK-Haplotypecaller specifications, and with additional gVCF banding at 0, 10, 20, 30, 40, 50 and 99 qualities. We used the fraction of the genome with genotype quality >15 for subsequent analyses. For the lists of high-confidence variant sites, we include only heterozygous positions after filtering at GQ >20, maximum DP <100, minimum DP >6, as described in the README file at http://broad.io/variants.

Inferring overall heterozygosity. To avoid confounding by sex chromosomes or complex regions, we excluded all scaffolds with less than 0.5 or greater than 2× of the average sample read depth, then calculated global heterozygosity as the fraction of heterozygous calls over the whole callable genome.

Calling SoH. We estimated the proportion of the genome within SoH using a metric designed for genomes with scaffold N50 shorter than the expected maximum length of runs of homozygosity (our median scaffold N50 is 62 kb). We first split all scaffolds into windows with a maximum length of 50 kb, with windows ranging from 20 to 50 kb for scaffolds <50 kb. For each window, we calculated the average number of heterozygous sites per bp. We discriminated windows with extremely low heterozygosity by using the Python 3.5.2 pomegranate package to fit a two-component Gaussian mixture model to the joint distribution of window heterozygosity, forcing the first component to be centred around the lower tail of the distribution and allowing the second to freely capture all the remaining heterozygosity variability. As heterozygosity cannot be negative, and normal distributions near zero can cross into negative values, we used the normal cumulative distribution function to correct the posterior distribution by the negative excess—effectively fitting a truncated normal to the first component. The final SoH value was calculated using the posterior maximum likelihood classification between both components. We saw no significant correlation between contig N50 and SoH (Pearson correlation = 0.055, P = 0.57, n = 112).

Assessing the effect of the percentage of callable genome. We assessed whether the percentage of the genome that was callable (Supplementary Table 3) was likely to affect our analysis. The callable percentage was correlated with heterozygosity (r = −0.80, P = 2.2 × 10−16, n = 130), and weakly with SoH values (r = 0.18, P = 0.06, n = 112). There is no significant difference in callable percentage among IUCN categories (analysis of variance P = 0.98, n = 122) or between captive and wild populations (t-test P = 0.81, n = 120).

Analysing patterns of diversity. We excluded two genomes with exceptionally high heterozygosity (heterozygosity >0.02; >5 s.d. above the mean). Both were of non-endangered species, and thus removing them made our determination of lower heterozygosity in endangered species more conservative. Of the remaining 124 genomes, we excluded 19 with allelic balance values that were more than one s.d. above the
mean (>0.36). Abnormally high allelic balance can indicate sequencing biases with potential for artefacts in estimates of heterozygosity and/or SoH. Our final dataset contains heterozygosity values for 105 genomes and SoH values for 98 genomes (Supplementary Table 3). For seven genomes, we were unable to estimate SoH because the two components of the Gaussian mixture model overlapped completely. To ask about a possible directional relationship between level of IUCN concern and overall heterozygosity or SoH, we applied regression using the IUCN category as an ordinal predictor. We also asked about the relationship of diversity metrics to a set of species-level phenotypes for which correlations were previously reported (Extended Data Table 3).

Alignment
The alignment was generated using the progressive mode of Cactus57,58. The topology used for the guide tree of the alignment was taken from TimeTree57; the branch lengths of the guide tree were generated by a least-squares fit from a distance matrix. The distance matrix was based on the UCSC 100-way phyloP fourfold-degenerate site tree 59 for those species that had corresponding entries in the 100-way tree. For species not present in the 100-way tree, distance matrix entries were more coarsely estimated using the distance estimated from Mash60 to the closest relative included in the 100-way data.

Cactus does not attempt to fully resolve the gene tree when multiple duplications take place along a single branch, as there is an implicit restriction in Cactus that a duplication event be represented as multiple regions in the child species aligned to a single region in the parent species. This precludes representing discordance between the gene tree and species tree that could occur with incomplete lineage-sorting or horizontal transfer. However, the guide tree has a minimal effect on the alignment, with little difference between alignments with different trees—even when using a tree that is purposely wrong61. Phenomena such as incomplete lineage sorting that affect a subset of species are unlikely to substantially affect the detection of purifying selection across the whole euthaniser lineage described in Fig. 3.

The alignment was generated in several steps, on account of its large scale. First, a backbone alignment of several long contiguity assemblies was generated, using the genomes of two non-placental mammals (Tasmanian devil (Sarcophilus harrisii) and platypus (Ornithorhynchus anatinus)), to inform the reconstruction of the placental root. Next, separate clade alignments were generated for each major clade in the alignment: Euarchonta, Glires, Laurasiatheria, Afrotheria and Xenarthra. The roots of these clade alignments were then aligned to the corresponding ancestral genomes from the backbone, stitching these alignments together to create the final alignment. The process of aligning a genome to an existing ancestor is complex and further described in an accompanying Article that introduces the progressive mode of Cactus62.

We created a neutral model for the conservation analysis using ancestral repeats detected by RepeatMasker63 on the euthaniser ancestral genome produced in the Cactus alignment (tRNA and low-complexity repeats were removed). To fit the neutral model, we used phyloPfit from the PHAST4 package, using the REV (generalized reversible) model and EM optimization method. The training input was a MAF exported on columns from the set of ancestral repeats mentioned above. Because phyloPfit does not support alignment columns that contain duplicates, if a genome had more than one sequence in a single alignment block, these were replaced with a single entry representing the consensus base at each column.

We extracted initial conservation scores using phyloP from the PHAST4 package on a MAF exported using human as a reference. We converted the phyloP scores (which represent log-scaled P values of acceleration or conservation) into P values, then into q values using the FDR-correction of Benjamini and Hochberg64. Any column with a resulting q value less than 0.05 was deemed significantly conserved or accelerated.

The alignment—as well as conservation annotations—are available at https://cglgenomics.ucsc.edu/data/cactus/.

Reporting summary
Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability
The project website is http://zoonomiaproject.org/. Details of each Zoonomi genome assembly—including NCBI GenBank65 accession numbers—are provided in Supplementary Table 1. Sequence data and genome assemblies are available at https://www.ncbi.nlm.nih.gov/. Variant lists for each species are provided at http://broad.io/variants. Further source data for Fig. 2 are provided in the Zoonomi GitHub repository (https://doi.org/10.5281/zenodo.3878432). The Cactus alignment is provided at https://cglgenomics.ucsc.edu/data/cactus/. A visualization of the alignments and phyloP data is available by loading our assembly hub into the UCSC browser66 by copying the hub link https://comparative-genomics-hubs.s3-us-west-2.amazonaws.com/200m_hub.txt into the Track Hubs page. There are no restrictions on use. Source data are provided with this paper.

Code availability
The DISCOVAR de novo assembly code is available at https://github.com/broadinstitute/discover_de_novo/releases/tag/v52488 (https://doi.org/10.5281/zenodo.3870899), the Cactus pipeline is available at https://github.com/ComparativeGenomicsToolkit/cactus (https://doi.org/10.5281/zenodo.3873410) and code for other analyses is available at https://github.com/broadinstitute/Zoonomi/ (https://doi.org/10.5281/zenodo.3874332).

50. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31, 3210–3212 (2015).
51. Fané, M. et al. A near-chromosome-scale genome assembly of the gemsbok (Oryx gazella): an iconic antelope of the Kalahari desert. Gigascience 8, givy162 (2019).
52. McKenna, A. et al. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20, 1297-1303 (2010).
53. DePristo, M. A. et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. 43, 491–498 (2011).
54. Li, H. et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079 (2009).
55. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754–1760 (2009).
56. Benaglia, T., Chauveau, D., Hunter, D. & Young, D. mixtools: an R package for analyzing finite mixture models. J. Stat. Softw. 32, 1–29 (2009).
57. R Core Team. R: A Language and Environment for Statistical Computing. https://www.R-project.org/ (2019).
58. Paten, B. et al. Cactus: algorithms for genome multiple sequence alignment. Genome Res. 21, 1512–1528 (2011).
59. Ondov, B. D. et al. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol. 17, 132 (2016).
60. Smit, A. F. A., Hubley, R. & Green, P. RepeatMasker Open-4.0. 4.0. http://www.repeatmasker.org/ (2013–2015).
61. Hubisz, M. J., Poland, K. S. & Siepel, A. PHAST and RPHAST: phylogenetic analysis with space/time models. Brief. Bioinform. 12, 41–51 (2011).
62. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. B 57, 289–300 (1995).
63. Benson, D. A. et al. GenBank. Nucleic Acids Res. 41, D36–D42 (2013).
64. Nguyen, N. et al. Comparative assembly hubs: web-accessible browsers for comparative genomics. Bioinformatics 30, 3293–3301 (2014).
65. Karcewski, K. J. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature 581, 434–443 (2020).
66. Pinheiro, E., Tadeu, V. A., Migliorini, R. H. & Kettelhut, I. C. Effect of fasting on carbohydrate metabolism in frugivorous bats (Artibeus lituratus and Artibeus jamaicensis). Comp. Biochem. Physiol. B Biochem. Mol. Biol. 143, 279–284 (2006).
67. Gordon, L. M. et al. Amorphous intergranular phases control the properties of rodent tooth enamel. Science 347, 746–750 (2015).
68. Hindle, A. G. & Martin, S. L. Intrinsic circannual regulation of brown adipose tissue form and function in tune with hibernation. Am. J. Physiol. Endocrinol. Metab. 306, E284–E299 (2014).
69. Stanford, K. I. et al. Brown adipose tissue regulates glucose homeostasis and insulin sensitivity. J. Clin. Invest. 123, 215–223 (2013).
70. Chondronikola, M. et al. Brown adipose tissue improves whole-body glucose homeostasis and insulin sensitivity in humans. Diabetes 63, 4089–4099 (2014).
Analysis

71. Saito, M. et al. High incidence of metabolically active brown adipose tissue in healthy adult humans: effects of cold exposure and adiposity. Diabetes 58, 1526–1531 (2009).

Acknowledgements We thank the many individuals who provided samples and advice, including C. Adenyo, C. Avila, E. Baichman, R. Behringer, A. Boyko, M. Breen, K. Campbell, P. Campbell, C. J. Conroy, K. Cooper, L. M. Dávalos, F. Delsuc, D. Distel, C. A. Emerling, J. Franzek, N. Gemmel, J. Good, K. He, K. Helgen, A. Hindle, H. Hoekstra, R. Honeycutt, F. Huval, W. Israelsen, B. Kayang, R. Kennerley, M. Korody, D. N. Lee, E. Louis, M. MacManes, A. Misuraca, A. Mitelberg, P. Morin, A. Mouton, M. Murayama, M. Nachman, A. Navarro, R. Ogden, B. Pasch, S. Puechmaillal, T. J. Robinson, S. Rossiter, M. Ruedi, A. Seifert, S. Thomas, S. Turey, G. Verbeylen and the late R. J. Baker. We also thank the Broad Institute Genomics Platform and SNP & SEQ Technology Platform (part of the National Genomics Infrastructure (NGI) Sweden and Science for Life Laboratory) and Swedish National Infrastructure for Computing (SNIC) at Uppmax. This project was funded by NIH NHGRI R01HG008742 (E.K.K., B.B., D.P.G., R.S., J.T.-M., J.J., H.J.N., B.P. and J. Armstrong), Swedish Research Council Distinguished Professor Award (K.L.-T., V.D.M., E.M. and J.R.S.M.), Swedish Research Council grant 2018-05973 (K.L.-T.), Knut and Alice Wallenberg Foundation (K.L.-T., V.D.M., E.M. and J.R.S.M.), Uppsala University (K.L.-T., V.D.M., E.M., J.R.S.M., J.J., J. Alfoldi and L.G.), Broad Institute NextXf (L.G.), Gladstone Institutes (K.S.P.), NIH NHGRI SR01HG002939 (A.F.A.S. and R.H.), NIH NHGRI SU24HG010136 (A.F.A.S. and R.H.), NIH NHGRI SR01HG010485 (B.P. and M.D.), NIH NHGRI 2U41HG007234 (B.P., M.D. and J. Armstrong), NIH NIA SP01AG047200 (V.N.G.), NIH NIA 1U1H2AG064706 (V.N.G.), BFU2017-86471-P MINECO/FEDER, UE (T.M.-B.), Secretaria d’Universitats i Recerca and CERCA Programme del Departament d’Economia i Coneixement de la Generalitat de Catalunya GRC 2017 SGR 880 (T.M.-B.), Howard Hughes International Early Career (T.M.-B.), European Research Council Horizon 2020 no. 864203 (T.M.-B.), Obra Social ‘La Caixa’ (T.M.-B.), BBSRC BBS/E/100086418, BBS/E/100086783 (W.H. and W.N.), BBSRC Core Strategic Programme Grant BB/P016774/T (W.H., W.N. and F.D.), Sir Henry Dale Fellowship 200517/Z/16/Z jointly funded by the Wellcome Trust and the Royal Society (N.R.C.), FICJ-2016-29558 MICINN (D.J.), Prince Albert II Foundation of Monaco and Canada, Global Genome Initiative, Smithsonian Institution (M.N.), European Research Council Research Grant ERC-2012-283110000 (E.C.T.), Irish Research Council Laureate Award (E.C.T.), UK Medical Research Council MR/P026028/1 (W.H. and W.N.), National Science Foundation DEB-1457735 (M.S.S.), National Science Foundation DEB-1753760 (W.J.M.), National Science Foundation IOS-2029778 (E.K.K. and D.P.G.), Robert and Rosabel Osborne Endowment (H.A.L. and J.D.), Swedish Research Council, FORMAS 221-2012-1531 (J.R.S.M.), NSF Rol.: FELS. EAGER. DEB 1838283 (D.A.R.) and Academy of Finland grant to Center of Excellence in Tumor Genetics Research no. 310242 (T.K. and J.T.).

Author contributions K.L.-T. conceived the project. J.J., V.D.M., E.M., N.R.C., L.G.C., J.D., V.N.G., M.L.H., K.-P.K., J.R.S.M., W.J.M., M.N., D.A.R., R.S., E.C.T., J. Alfoldi, O.A.R., H.A.L., K.L.-T. and E.K.K. contributed to the acquisition of the samples. J.J., V.D.M., E.M., J.D., L.G., K.-P.K., H.I.N., C.C.S., R.S., J.T.-M., J. Alfoldi, O.A.R., H.A.L., K.L.-T. and E.K.K. contributed to the production of the genome assemblies. D.P.G., A.S., J. Armstrong, J.J., D.J., I.T.F., L.F.K.K., H.A.L., T.M.-B., K.L.-T. and E.K.K. contributed to the data analysis. D.P.G., J.J., V.D.M., G.B., F.D.P., M.D., I.T.F., M.G., V.N.G., W.H., R.H., T.K., E.S.L., J.R.S.M., A.R.P., K.S.P., A.F.A.S., M.S.S., J.T., J. Alfoldi, B.B., O.A.R., H.A.L., B.P., T.M.-B., K.L.-T. and E.K.K. contributed to the design and conduct of the project. D.P.G., E.S.L., W.N., B.S., O.A.R., K.L.-T. and E.K.K. wrote the manuscript, with input from all other authors.

Competing interests L.G. is a co-founder of, equity owner in and chief technical officer at Fauna Bio Incorporated.

Additional information Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2876-6.

Correspondence and requests for materials should be addressed to E.K.K.

Peer review information Nature thanks Chris Ponting, Steven Salzberg, Guojie Zhang and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at http://www.nature.com/reprints.
Extended Data Fig. 1 | Notable traits in non-human mammals. Sequences from species with notable phenotypes can inform human medicine, basic biology and biodiversity conservation, but sample collection can be challenging. a, The Jamaican fruit bat (Artibeus jamaicensis) maintains constant blood glucose across intervals of fruit-eating and fasting, achieving homeostasis to a degree that is unknown in the treatment of human diabetes. b, The North American beaver (Castor canadensis) avoids tooth decay by incorporating iron rather than magnesium into tooth enamel, which yields an orange hue. c, The thirteen-lined ground squirrel (Ictidomys tridecemlineatus) prepares for hibernation by rapidly increasing the thermogenic activity of brown fat, a process that—in humans—is connected to improved glucose homeostasis and insulin sensitivity. d, The tiny bumblebee bat (Craseonycteris thonglongyai) is among the smallest of mammals, making it a sparse source of DNA. e, The remote habitat of the very rare Amazon River dolphin (Inia geoffrensis) precludes collection of the high-molecular weight DNA. Image sources: Merlin D. Tuttle/Science Source (a); Stephen J. Krasemann/Science Source (b); Allyson Hindle (c); Sébastien J. Puechmaille (CC BY-SA) (d); M. Watson/Science Source (e).
Extended Data Fig. 2 | Sample collection can be challenging, and sequencing methods must be selected to handle the sample quality. To enable the inclusion of species from across the eutherian tree (including from the 50% of mammalian families not represented in existing genome databases), the Zoonomia Project needed sequencing and assembly methods that produce reliable data from DNA collected in remote locations, sometimes in only modest quantities and often without benefit of cold chains for transport. a, For the marine species such as the narwhal (*Monodon monoceros*), simply accessing an individual in the wild can prove challenging. For example, to sample DNA from the near-threatened narwhal, M.N. and Inuit guide D. Angnatsiak camped on the edge of an ice floe between Pond Inlet and Bylot Island, at the northeastern tip of Baffin Island. After a narwhal was collected by Inuit hunters as part of an annual hunt, hours of flensing were necessary for the collection of tissue samples. From left to right, F. McCann, H. C. Schmidt, F. Eichmiller, M.N., J. Orr (facing backward) and J. Orr (standing). b, For endangered species such as the Hispaniolan solenodon (*S. paradoxus*), sample collection must be designed to minimize stress to the individual, limiting the amount of DNA that can be collected22. To collect DNA from the endangered solenodon without imposing stress on individuals in the wild, N.R.C. turned to the world’s only captive solenodons, which are housed off-exhibit at ZOODOM in the Dominican Republic. With help from veterinarians at the zoo, N.R.C. collected a small amount of blood from the rugged tail of the solenodon. Narwhal photograph by G. Freund, and courtesy of M.N. Solenodon photograph courtesy of L. Emery.
| n | Common Name | Species | Family |
|---|-------------|---------|--------|
| 1 | 1 Hispaniolan solenodon | Solenodon paradoxus (EN) | Solenodontidae |
| 2 | 1 Indochinese shrew | Crocidura indochinensis | Soricidae |
| 3 | 1 Eastern mole | Scapanus aquaticus | Talpidae |
| 4 | 1 Great shrew-like mole | Uroslitus gracilis | Talpidae |
| 5 | 1 Arctic fox | Vulpes lagopus | Canidae |
| 6 | 1 Domestic dog (village dog) | Canis lupus familiaris | Canidae |
| 7 | 1 Fossa | Cryptoprocta ferox (VU) | Eupleridae |
| 8 | 1 Black-footed cat | Felis nigripes (VU) | Felidae |
| 9 | 1 Jaguar | Panthera onca (NT) | Felidae |
| 10 | 1 Dwarf mongoose | Helogale parvula | Hirundinidae |
| 11 | 1 Meerkat | Suricata suricatta | Herpestidae |
| 12 | 1 S Abridged mongoose | Mungos mungo | Herpestidae |
| 13 | 1 Striped hyena | Hyaena hyaena (NT) | Hyaenidae |
| 14 | 1 Western spotted skunk | Spilogale gracilis | Mephitidae |
| 15 | 1 Giant otter | Pteronura brasiliensis (EN) | Mustelidae |
| 16 | 1 Honey badger | Mellivora capensis | Mustelidae |
| 17 | 1 California sea lion | Zalophus californianus | Otariidae |
| 18 | 1 Northern elephant seal | Mirounga angustirostris | Phocidae |
| 19 | 1 Asian palm civet | Paradoxurus hermaphroditus | Viverridae |
| 20 | 1 Tree pargolin | Manis tricuspis* (VU) | Manidae |
| 21 | 1 Black rhinoceros | Diceros bicornis (CR) | Rhinocerotidae |
| 22 | 1 Northern white rhino | Ceratotherium simum (CR) | Rhinocerotidae |
| 23 | 1 Malayan tapir | Tapirus indicus (EN) | Tapiridae |
| 24 | 1 South American tapir | Tapirus terrestris (NT) | Tapiridae |
| 25 | 1 Pronghorn | Antilocapra americana* | Antilocapridae |
| 26 | 1 North Pacific right whale | Eubalaena japonica (EN) | Balaenidae |
| 27 | 1 Hilde | Steatorrhinus hunteri (CR) | Dasyuridae |
| 28 | 1 Nillig tahr | Hamitragus hylotis (EN) | Bovidae |
| 29 | 1 Peninsular bighorn sheep | Ovis canadensis (EN) | Bovidae |
| 30 | 1 Russian salga | Saiga tatarica tatarica (CR) | Cervidae |
| 31 | 1 Siberian reindeer | Rangifer tarandus (CR) | Cervidae |
| 32 | 1 Grey whale | Eschrichtius robustus | Eschrichtiuidae |
| 33 | 1 Hippopotamus | Hippopotamus amphibius* (VU) | Hippopotamlidae |
| 34 | 1 Amazon river dolphin | Inia geoffrensis (DD) | Iniidae |
| 35 | 1 Pygmy sperm whale | Kogoa breviceps (DD) | Kogodidae |
| 36 | 1 Narwhal | Monodon monoceros | Monodontidae |
| 37 | 1 Narwhal | Monodon monoceros | Monodontidae |
| 38 | 1 Siberian musk deer | Moschus moschiferus* (VU) | Moschidae |
| 39 | 1 Harp seal | Phoca groenlandica | Phocidae |
| 40 | 1 Indus river dolphin | Platanista gangetica (EN) | Platanistidae |
| 41 | 1 La plata dolphin | Pontoporia blainvillii (EN) | Iniidae |
| 42 | 1 Chacoan peccary | Catagonus wagneri* (EN) | Tayassuidae |
| 43 | 1 Java slender loris | Tragulus javanicus* (VU) | Tragulidae |
| 44 | 1 Cuvier's beaked whale | Ziphus cavirostris | Ziphiidae |
| 45 | 1 Sowerby's beaked whale | Mesoplodon bidens (DD) | Ziphiidae |
| 46 | Chiropteridae (bats) | Chiropteridae (bats) | Chiroptera |
| 47 | 1 Cantor's leaf-nosed bat | Hipposideros griseus | Hipposideridae |
| 48 | 1 Great false vampire bat | Megaderma lyra | Megadermatidae |
| 49 | 1 Mexican free-tailed bat | Tadarida brasiliensis | Molossidae |
| 50 | 1 Ghost-faced bat | Mormoops megalophylla | Molossidae |
| 51 | 1 Greater bulldog bat | Noctilio leporinus | Phyllostomidae |
| 52 | 1 California leaf-nosed bat | Macrotus californicus | Phyllostomidae |
| 53 | 1 Hairy big-eared bat | Myronectes hirtula | Phyllostomidae |
| 54 | 1 Jamaican fruit-eating bat | Artibeus jamaicensis | Phyllostomidae |
| 55 | 1 Seba's short-tailed shrew | Carollia perspicillata | Soricidae |
| 56 | 1 Stripe-headed round-eared bat | Tonia sauridra | Vespertilionidae |
| 57 | 1 Tailed bat | Anoura caudifera | Vespertilionidae |
| 58 | 1 Egyptian fruit bat | Rousettus aegyptiacus | Pteropodidae |
| 59 | 1 Long-tongued fruit bat | Macrograculus sorbinus | Pteropodidae |
| 60 | 1 Greater horseshoe bat | Rhinolophus ferrumequinum | Rhinolophidae |
| 61 | 1 Ashy gray-tube-nosed bat | Myotis fulvus | Myotisidae |
| 62 | 1 Common bent-wing bat | Miniopterus schreibersi (NT) | Miniopterusidae |
| 63 | 1 Common pipistrelle | Pipistrellus pipistrellus | Vespertilionidae |
| 64 | 1 Eastern red bat | Lasiurus borealis | Vespertilionidae |
| 65 | 1 Egyptian slit-faced bat | Nyctereutes helalensis | Vespertilionidae |
| 66 | 1 Greater mouse-eared bat | Myotis mysticellus | Vespertilionidae |
| 67 | 1 Pallid bat | Antrozous pallidus | Vespertilionidae |

| Extended Data Table 1 | The Zoonomia Project data includes 132 genome assemblies |

These assemblies include 131 different species, with 2 narwhals (male and female), and 10 genomes upgraded to longer contiguity (including upgrade of an existing assembly for E. telfairi).

Species of concern on the IUCN Red List are indicated as near-threatened (NT), vulnerable (V), endangered (EN) or critically endangered (CR).

*Upgraded to longer contiguity (using existing assembly).

**Upgraded to longer contiguity using existing assembly.
The expected number of variants conserved by chance (false positives) was estimated for four genomic resources (the 29 Mammals Project\(^7\) dataset, the human-only ExAC\(^14\) and gnomAD v.3\(^6\) datasets, and the Zoonomia Project dataset) by applying a Poisson model of the distribution of substitution counts in the genome. Branch length for gnomAD was estimated by dividing 526,001,545 single-nucleotide variants by 3.088 gigabases (size of the human genome). Branch length for Zoonomia was measured as the number of substitutions per site in the phyloP analysis of the Cactus alignment.

| Dataset                | Number of samples | Branch length | Expected fraction with no substitutions | Expected number of false positives |
|------------------------|-------------------|---------------|----------------------------------------|----------------------------------|
| 29 Mammals Project     | 29                | 4.9           | 7.5x10^{-3}                            | 22,995,049                       |
| ExAC (33 megabases, exome only) | 60,706          | 0.12          | 0.89                                   | 29,268,374                       |
| gnomAD v3              | 71,702            | 0.17          | 0.84                                   | 2,604,359,690                    |
| Zoonomia Project       | 240               | 16.6          | 6.2 x10^{-8}                          | 191                              |
Extended Data Table 3 | Diversity statistics are not correlated with other species-level phenotypes

| Test      | Phenotype                        | heterozygosity N | p     | segments of homozygosity N | p     | Description                                                                                                                                                                                                 |
|-----------|----------------------------------|------------------|-------|---------------------------|-------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| **ANOVA** | 12-1 Habitat Breadth             | 58               | 0.277 | 55                        | 0.418 | Number of habitat layers used by non-captive populations; Categories: above ground dwelling, aquatic, fossorial and ground dwelling                                                                           |
| LM        | 15-1 Litter Size                 | 64               | 0.094 | 59                        | 0.018 | Number of offspring born per litter per female                                                                                                                                                           |
| LM        | 26-1 GR Area km2                 | 64               | 0.258 | 60                        | 0.171 | Calculated using total extent of a species range with a global equal-area projection                                                                                                                    |
| LM        | 26-2 GR MaxLat dd                | 64               | 0.473 | 60                        | 0.423 | Maximum latitudinal extent of each species range                                                                                                                                                         |
| LM        | 26-3 GR MinLat dd                | 64               | 0.850 | 60                        | 0.773 | Minimum latitudinal extent of each species range                                                                                                                                                         |
| LM        | 26-4 GR MidRangeLat dd           | 64               | 0.038 | 60                        | 0.179 | Median latitudinal extent of each species range                                                                                                                                                           |
| LM        | 26-5 GR MaxLong dd               | 64               | 0.655 | 60                        | 0.694 | Maximum longitudinal extent of each species range                                                                                                                                                         |
| LM        | 26-6 GR MinLong dd               | 64               | 0.632 | 60                        | 0.516 | Minimum longitudinal extent of each species range                                                                                                                                                         |
| LM        | 26-7 GR MidRangeLong dd          | 64               | 0.624 | 60                        | 0.579 | Median longitudinal extent of each species range                                                                                                                                                         |
| LM        | 27-1 HuPopDen Min n-km2          | 64               | 0.567 | 60                        | 0.268 | Minimum human population density                                                                                                                                                                          |
| LM        | 27-2 HuPopDen Mean n-km2         | 64               | 0.342 | 60                        | 0.330 | Mean human population density                                                                                                                                                                             |
| LM        | 27-3 HuPopDen 5p n-km2           | 64               | 0.727 | 60                        | 0.488 | 5th percentile human population density                                                                                                                                                                  |
| LM        | 27-4 HuPopDen Change             | 64               | 0.372 | 60                        | 0.107 | Mean rate of increase in human population density                                                                                                                                                         |
| LM        | 28-1 Precip Mean mm              | 64               | 0.092 | 60                        | 0.433 | Mean monthly precipitation                                                                                                                                                                                 |
| LM        | 28-2 Temp Mean 01degC            | 64               | 0.098 | 60                        | 0.063 | Mean monthly temperature (0.1°C)                                                                                                                                                                          |
| LM        | 30-1 AET Mean mm                 | 64               | 0.101 | 60                        | 0.608 | Mean monthly AET (Actual Evapotranspiration Rate) from 1920 to 1980 (mm)                                                                                                                                   |
| LM        | 30-2 PET Mean mm                 | 64               | 0.078 | 60                        | 0.154 | Mean monthly PET (Potential Evapotranspiration Rate) from 1920 to 1980                                                                                                                                    |
| LM        | 5-1 AdultBodyMass g              | 66               | 0.228 | 61                        | 0.823 | Mass of adult (or age unspecified) live or freshly-killed specimens (excluding pregnant females)                                                                                                           |
| **ANOVA** | 6-1 Diet Breadth                 | 59               | 0.657 | 55                        | 0.531 | Number of dietary categories; for non-captive or non-provisioned populations; Categories: vertebrate, invertebrate, fruit, flowers/nectar/pollen, leaves/branches/bark, seeds, grass and roots/tubers |
| **ANOVA** | 6-2 Trophic Level                | 59               | 0.966 | 55                        | 0.894 | Trophic level of each species for non-captive or non-provisioned populations; Categories: (1) herbivore; (2) omnivore, or (3) carnivore                                                                      |
| LM        | 9-1 GestationLen d               | 55               | 0.074 | 52                        | 0.331 | Length of time of non-inactive fetal growth                                                                                                                                                              |
| **ANOVA** | Family                           | 35               | 0.088 | 33                        | 0.421 | Families with more than 1 representative species categorized as Least Concern, including: Canidae (2), Caviidae (2), Cebeidae (2), Cercopithecidae (3), Cricetidae (2), Dipodidae (2), Herpestidae (3), Phylllostomidae (6), Pitheciidae (2), Procaviidae (2), Pteropodidae (2), Talpidae (2), Vesperillionidae (5) |
| **ANOVA** | Order                            | 62               | 0.108 | 56                        | 0.619 | Orders with 4 or more species categorized as Least Concern, including Carnivora (9); Cetartiodactyla (5); Chiroptera (18); Primates (7); Rodentia (23).                                                          |

All phenotypes in the PanTHERIA database were for which at least 75% of the 75 species of least concern had a value were included in the analysis. For continuous phenotypes, values were standardized to $Z$-scores before analysis (latitude was calculated as an absolute value) and correlation measured by fitting a linear model using the core R function lm. For categorical phenotypes with more than two categories, group means were compared using the core R function aov to fit an analysis of variance model. None was significant after Bonferroni correction for the number of traits considered (21).
Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

☑ n/a

- The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement (Confirmed)
- A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The statistical test(s) used AND whether they are one- or two-sided
- Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- For null hypothesis testing, the test statistic (e.g. F, t, r) with confidence intervals, effect sizes, degrees of freedom and P value noted
  - Give P values as exact values whenever suitable.
- For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- Estimates of effect sizes (e.g. Cohen’s d, Pearson’s r), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection

- No software was used

Data analysis

- DISCOVAR de novo (discovardenovo-52488); BUSCO 2.0; HiRise 2.1; RStudio 1.2; R version 3.6.1; Samtools 1.8; BWA 0.7.17-r1188; GATK 3.6; Picard-Tools 2.21.3; Python 3.5.2 pomegranate package; ordPens package for R; Cactus (https://www.biorxiv.org/content/10.1101/730531v3.full); v1.5 PHAST; Custom python scripts for implementing SoH and heteroygosity analyses as described in methods.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:
- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Details on each Zoonomia Project genome assembly, including NCBI Genbank accession numbers, are in Supplementary Table 1. Sequence data and genome assemblies are available at https://www.ncbi.nlm.nih.gov/. Variant lists for each species are at broad.io/variants. Raw data for figure 3 is in Supplementary Table 2. There are no restrictions on data availability.
Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

- Life sciences
- Behavioural & social sciences
- Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

| Sample size | Sample size was determined based on evolutionary branch length calculations, as described in the manuscript. |
| Data exclusions | Diversity analysis: We excluded 2 genomes with high heterozygosity (> 6 standard deviations above the mean) and 17 genomes with allelic balance values more than one standard deviation above the mean. Exclusion criteria were established prior to analysis and described in Methods. |
| Replication | No replication. Study design required just one individual from each species be sequenced. |
| Randomization | Not relevant. This study did not involve experimental groups. |
| Blinding | Not relevant. This study did not involve experimental groups. |

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

| Materials & experimental systems | Methods |
| n/a | Involved in the study |
| ☒ | Antibodies |
| ☒ | Eukaryotic cell lines |
| ☒ | Palaeontology |
| ☒ | Animals and other organisms |
| ☒ | Human research participants |
| ☒ | Clinical data |
| n/a | Involved in the study |
| ☒ | ChIP-seq |
| ☒ | Flow cytometry |
| ☒ | MRI-based neuroimaging |