A comparative genomics multitool for scientific discovery and conservation

*Zoonomia Consortium**

*A list of authors and their affiliations appears at the end of the paper

The Zoonomia Project is investigating the genomics of shared and specialized traits in eutherian mammals. We describe a whole-genome alignment of 240 species with unprecedented phylogenetic diversity, with over 80% of mammalian families represented, and new genome assemblies for 131 species. We find that regions of reduced genetic diversity are more abundant in species with high extinction risk, discern signals of evolutionary selection at unprecedented resolution, and describe insights enabled by individual reference genomes. By prioritizing phylogenetic diversity and making data available quickly, without restriction, the Zoonomia Project aims to support biological discovery, medical research, and biodiversity conservation.

The genomics revolution is enabling advances not just in medical research[1], but also in basic biology[2] and in biodiversity conservation, where genomic tools have helped apprehend poachers[3] and protect endangered populations[4]. Yet we still have only limited ability to predict which genomic variants lead to changes in organism-level phenotypes, such as increased disease risk — a task complicated by the sheer size of the human genome (~3 billion nucleotides)[5].

Comparative genomics can address this challenge by identifying nucleotide positions unchanged across millions of years of evolution[6], suggesting that changes negatively impact fitness and focusing the search for disease-causing variants. In 2011, the 29 Mammals Project[7] identified 12 base pair (bp) regions of evolutionary constraint, totalling 5.4% of the genome, by measuring sequence conservation in humans plus 28 other mammals. These regions proved to be more enriched for complex disease heritability than any other functional mark, including coding status[8]. By expanding the number of species, and making an alignment independent of any single reference genome, the Zoonomia Project was designed to detect evolutionary constraint in the eutherian lineage at unprecedented resolution, while providing new genomic resources for over 130 species.

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 with a representative genome from 49% to 82%, and include nine species that are the sole extant member of their family and seven that are critically endangered (**Figure 1**)[9]: 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 rhino (*Ceratotherium simum cottoni*) and black rhinoceros (*Diceros bicornis*).

We collaborated with 28 different institutions to collect samples, with nearly half (47%) provided by The Frozen Zoo® of San Diego Zoo Global (**Supplementary Table 1**). Since 1975, The Frozen Zoo® has stored renewable cell cultures for about 10,000 vertebrate animals representing over 1,100 taxa, including more than 200 species classified as vulnerable, endangered, critically endangered, or extinct[10]. For 36 target species we were unable to acquire a DNA sample of sufficient quality, even though our requirements were modest (see below), highlighting a major impediment to expanding the phylogenetic diversity of genomics.

We used two complementary approaches to generate genome assemblies (**Extended Data Table 1**). First, for 131 genomes, we generated shorter contiguity assemblies ("DISCOVAR assemblies") by performing a single lane of sequencing (2x250bp reads) on PCR-free libraries, and assembling with *DISCOVAR de novo*[11]. This method does not require intact cells and uses less than two micrograms of medium-quality DNA (most fragments >5 kilobases (kb)), which allowed us to include difficult-to-access species (**Extended Data Figure 1; Extended Data Figure 2**) while achieving contig (contiguous sequences constructed from overlapping short reads) lengths comparable to existing assemblies (median contig N50 of 46.8kb; compared to 47.9kb for Refseq genome assemblies).

For nine DISCOVAR genomes, and one pre-existing assembly, the lesser hedgehog tenrec (*Echinops telfairi)*, we increased contiguity 200-fold (median scaffold length increased from 90.5kb to 18.5Mb) through proximity-ligation, which uses chromatin interaction data to capture the physical relationships among genomic regions[12]. Unlike short-contiguity genomes, these assemblies capture structural changes like chromosomal rearrangements[13]. The upgraded assemblies increase the number of eutherian orders represented by a long-range assembly (contig N50 > 20kb, scaffold N50 > 10Mb) from 12 to 18 out of 19. We are working on upgrading the large treeshrew (*Tupaia tana*) assembly for the remaining order, Scandentia.

Comparative power of 240 species

The Zoonomia alignment includes 120 of our new assemblies and 121 existing assemblies representing a total of 240 species (the dataset includes assemblies for two different dogs) spanning ~110 million years of mammal evolution **(Supplementary Table 2)**. With a total evolutionary branch length of 16.6 substitutions per site, we expect just 191 positions in the human genome (0.000006%) to be identical across the aligned species due to chance (false positives) rather than evolutionary constraint (**Extended Data Table 2**). We applied this same calculation to The Exome Aggregation Consortium (ExAC), which analyzed exomes for 60,706 humans[14], 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 penetrance coding mutations, for which ExAC is optimally powered[15].

Biological insights from new assemblies

The scope and species diversity in the Zoonomia project supports evolutionary studies in many different lineages. Papers published to date, and the demonstrated utility of existing comparative genomics resources[16,17], illustrate the benefits of making new genome assemblies and alignments accessible to all researchers, without restrictions on use.

Speciation. Comparing our assembly for the endangered mantled howler monkey subspecies *Alouatta palliata mexicana* with the Guatemalan black howler monkey (*Alouatta pigra)*, which has a neighboring range, suggests different forms of selection shape reproductive isolation[18]. Initial divergence in allopatry was followed by positive selection on prezygotic isolating mechanisms, offering empirical support for a speciation process first outlined by Dobzhansky in 1935[19].

Protection from cancer. Using our assembly for the capybara (*Hydrochoerus hydrochaeris)*, a giant rodent, Herrera-Alvarez *et al.* identified positive selection on anti-cancer pathways, echoing earlier reports that another large mammal, the elephant, carries extra copies (retrogenes) of the tumor-suppressor gene *TP53*[20,21]. This offers a possible resolution to Peto's paradox — the observation that cancer in large mammals is rarer than expected — and could reveal new anti-cancer mechanisms.

Convergent evolution of venom. Casewell *et al.* used our assembly for the Hispaniolan solenodon (*Solenodon paradoxus*; **Extended Data Figure 2**), to investigate venom production, a trait in just a few eutherian lineages, including shrews and solenodons[22]. They identified paralogous copies of a kallikrein 1 serine protease (*KLK1*) that, together, encode solenodon venom, and showed that the *KLK1s* were independently co-opted in solenodons and shrews in an intriguing example of molecular convergence.

Informing biodiversity conservation strategies. Beichman *et al.* analyzed our giant otter (*Pteronura brasiliensis*) assembly and found low diversity and elevated burden of putatively deleterious genetic variants, consistent with the otter's recent population decline through overhunting and habitat loss[23]. Intriguingly, the giant otter had fewer putatively deleterious variants than either the southern or northern sea otter, suggesting highest potential for recovery if populations are protected.

Rapid assessment of species infection risk during COVID-19 pandemic. Using the Zoonomia alignment, and public genomic data from hundreds of other vertebrates, Damas *et al.* compared the structure of ACE2 (the SARS-CoV-2 receptor) and identified 47 mammals that have a high or very high likelihood of being virus reservoirs, intermediate hosts, or good model organisms to study COVID-19, and detected positive selection in the ACE2 receptor binding domain specific to bats[24].

Genetic diversity and extinction risk

We next asked whether a reference genome from a single individual can help identify populations with low genetic diversity to prioritize in biodiversity-conservation efforts. Diversity metrics reflect demographic history[25,26], and heterozygosity is lower in threatened species[27]. This

analysis was feasible because we used a single sequencing and assembly protocol for all DISCOVAR assemblies, minimizing 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 representing a different species (**Supplementary Table 3**). Four failed during analysis. For the remaining 126, we calculated two metrics: (1) the fraction of sites at which the sequenced individual is heterozygous ("overall heterozygosity"); and (2) the proportion of the genome residing in an extended region without any variation ("segments of homozygosity", or SoH). SoH is designed for short-contiguity assemblies, where scaffolds are potentially shorter than runs of homozygosity. Overall, heterozygosity and SoH are correlated (Pearson correlation r=-0.56; $p=1.8x10^{-9}$; N=98). However, while overall heterozygosity is correlated with contig N50 (Pearson correlation r_{het}=-0.39; $p_{het}=4x10^{-5}$; N_{het}=105), likely due to the difficulty of assembling more heterozygous genomes[28], SoH is not (Pearson correlation; r_{SoH}=0.09 p_{SoH}= 0.38; N_{SoH}=98). Overall heterozygosity and SoH are highly correlated with between the lower- and high-contiguity versions of the upgraded assemblies (Pearson correlation; r_{het}=0.999; $p_{het}=5x10^{-7}$; N_{het}=7; r_{SoH}=0.996; $p_{SoH}=1.4x10^{-6}$ N_{SoH}=7).

Genomic diversity varies significantly among species in different International Union of Conservation Nature (IUCN) conservation categories, as measured by overall heterozygosity (**Figure 2A**), and SoH (**Figure 2B**). SoH increases (p= 0.0235; R^2=0.055; N=94) with increasing levels of conservation concern; heterozygosity decreases (p=0.011; R^2=0.064; N=101). There is no significant difference between wild and captive populations in overall heterozygosity (**Figure 2C**) or SoH (**Figure 2D)**.

Unusual diversity values can suggest particular population demographics, although data from more than a single individual is needed to confirm these inferences. All seven critically endangered species have SoH higher than the median for species categorized as Least Concern (**Figure 2E**). The genomes with the lowest heterozygosity and highest SoH were the social tuco-tuco (*Ctenomys sociabilis;* het=0.00063*;*SoH=78.7%), sampled from small laboratory colony with just 12 founders[29], and the eastern mole (*Scalopus aquaticus;* het=0.0008; SoH=81.3%), supplied by a professional mole catcher and likely from a population bottlenecked by pest control measures.

The correlation between diversity metrics and IUCN is not explained by other species-level phenotypes. For Least Concern (N=75) species, we assessed 21 phenotypes cataloged in the Pantheria[30] database for correlation with heterozygosity or SoH. The most significant was between SoH and litter size, a trait also shown to predict extinction risk (p_{SoH}=0.02)[31], but none is significant after Bonferroni correction (**Extended Data Table 3)**.

Our inference that diversity trends lower in species at higher risk of extinction comes from a small fraction (2.6%) of threatened mammals[9]. Whether a direct correlation with extinction risk, or arising from association of diversity with species-level phenotypes such as litter size, it

suggests valuable information can be gleaned from sequencing just a single individual. Should this pattern prove robust across more species, diversity metrics from a single reference genome could help identify populations at-risk, even when few species-level phenotypes are documented, and prioritize species for population-level follow-up.

Resources for biodiversity conservation

For each genome assembly, we cataloged all high-confidence variant sites (broad.io/variants) to support the design of cost-effective, accurate genetic assays that are usable even when sample quality is low[32] and often preferable to designing expensive custom tools, relying on tools from related species, or sequencing random regions[33]. The reference genomes themselves support development of technologies such as using gene drive to control invasive species, or "de-extinction" through cloning and genetic engineering[34].

Our genomes have two notable limitations: We sequenced only a single individual, which is insufficient for studying population origins, population structure and recent demographic events[35,36], and the shorter contiguity of our assemblies prevented us from analyzing runs of homozygosity (RoH)[26]. This highlights a dilemma facing 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 released when we started the alignment) as part of a 600-way pan-amniote alignment using the Cactus alignment software (**Supplementary Table 2**)[37]. Rather than aligning to a single anchor genome, Cactus infers an ancestral genome for each pair of assemblies **(Figure 3A**). Consistent with our predictions, we have increased power to detect sequence constraint at individual bases relative to earlier studies. We detect 3.1% of bases in the human genome to be under purifying selection in the eutherian lineage (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 (**Figure 3B**), with improvement most notable in noncoding sequence (**Figure 3C**), and in increased resolution of individual features (**Figure 3D**). This represents a substantial proportion, but not all, of the 5 to 8% of the human genome suggested to be under purifying selection[7,38].

Next steps

Using our alignment of 240 mammalian genomes, we are pursuing four key analysis strategies. (1) ***Largest nuclear genome eutherian phylogeny:*** build a comprehensive phylogeny and timetree, including trees partitioned by functional annotations, mode of inheritance, and long-term recombination rates. (2) ***Detailed map of evolutionary constraint:*** identify highly conserved genomic regions, regions under accelerated evolution in particular lineages, and changes that likely impact phenotype, leveraging functional data from ENCODE[39], GTEx[40] and the Human Cell Atlas[41]. (3) ***Genotype–phenotype correlations***: investigate patterns of constraint

in human disease-associated regions, identify patterns of convergent adaptive evolution[2], and apply a forward genomics strategy to link functional elements to traits. (4) ***Evolution of genome structure*:** map syntenic regions between genomes, identify evolutionary breakpoints, and characterize the repeat landscape.

Conclusion

The Zoonomia Project has captured mammalian diversity at unprecedented scope, and is among the first of many projects underway to sequence, catalog, and characterize whole branches of Earth's eukaryotic biodiversity. Based on our experience, we propose the following principles for realizing the full value of large-scale comparative genomics:

(1) **Prioritizing sample collection**: We must support field researchers who collect samples and understand species ecology and behaviour, develop strategies for sample collection absent bulky laboratory equipment or cold chains, develop technology for using non-invasive sample types, and establish more repositories of renewable cell cultures[10].

(2) **Accessible, scalable tools for computational analysis**: Few research groups have access to 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 to make powerful computational pipelines accessible to all researchers.

(3) **Rapid data-sharing**: Data embargoes must not be permitted to delay analyses that directly benefit 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[42], Genome 10K[43], the Vertebrate Genomes Project, Bat 1K[44], Bird 10K, and DNA Zoo. As the number of genomes grows, so will the usefulness of comparative genomics in disease research and therapeutic development. Preserving, rather than merely recording, Earth's biodiversity 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.

Figure 1. The Zoonomia Projects brings the fraction of eutherian families represented by at least one assembly to 83%.

Phylogenetic tree of the mammalian families in the Zoonomia Project alignment, including both our new assemblies and all other high-quality mammalian genomes publicly available in Genbank when we started the alignment (March 2018; Supplementary Table 2). Tree topology is based on data from timetree.org[45]. Existing taxonomic classifications recognize a total of 127 extant eutherian mammalian families[46], including 43 families not previously represented in

Genbank (red boxes) and 41 families with additional representative genome assemblies (pink boxes). Of the remaining families, 21 had Genbank genome assemblies but no Zoonomia Project assembly (grey boxes) and 22 had no representative genome assembly (white boxes). Parenthetical numbers indicate the number of species with genome assemblies in a given family. Image credits: Fossa: Bertal/Wikimedia [CC BY-SA]; Arctic fox: Michael Haferkamp/Wikimedia [CC BY-SA]; (F) Hirola: JRProbert/Wikimedia [CC BY-SA]; Bumblebee bat: Sébastien J. Puechmaille [CC BY-SA]; Snowshoe hare: Denali National Park and Preserve/Wikimedia [Public domain]; Aye-aye: Tom Junek/Wikimedia [CC BY-SA]; Geoffroy's spider monkey: Patrick Gijsbers/Wikimedia [CC BY-SA]; Southern three-banded armadillo: Hedwig Storch/Wikimedia [CC BY-SA]; Giant Anteater: Graham Hughes/Wikimedia [CC BY-SA]; Brown-throated Sloth: Dick Culbert from Gibsons, B.C., Canada/Wikimedia [CC BY].

Figure 2. Genetic diversity varies across IUCN conservation categories.

(A) Heterozygosity declines and **(B)** SoH increases with level of concern for species conservation, as assessed by IUCN conservation categories. Horizontal gray lines indicate median. Comparing individuals sampled from wild and captive populations, we saw no statistically significant difference (independent samples t-test) in either **(C)** overall heterozygosity or **(D)** % segments of homozygosity, with similar means (horizontal gray lines) between birth population types. For A-D, total n=105 species, with n for each tested category indicated on the x axis. Statistical tests were two-sided. **(E)** Overall heterozygosity and SoH for all genomes analyzed (including those with high allelic balance ratio; n=124 species), with median SoH (17.1%, horizontal dashed line) and median overall heterozygosity (0.0026, vertical dashed line) for species categorized as Least Concern (dashed lines). Values for individuals from the seven critically endangered species are shown in red, with red text labels.

Figure 3. The Zoonomia alignment doubles the fraction of the human genome predicted to be under purifying selection at single base pair resolution.

(**A)** Cactus alignments are reference-genome-free, enabling detection of sequence absent from human (red) or other clades (purple), lineage-specific innovations (orange, green) and eutherian mammal-specific sequence (blue). (**B)** We compared phyloP predictions of conserved positions for a widely-used 100-vertebrate alignment (n=100 vertebrate species; grey) to the Zoonomia alignment (n=240 eutherian species; red). The cumulative portion of the genome expected to be covered by true vs. false positive calls is shown, starting from the highest confidence calls (solid line) and proceeding to calls with lower confidence (dashed lines), with a horizontal line indicating the point at which the confidence level drops below an expected FDR of 0.05 (two-sided). **(C)** A higher proportion of functionally annotated bases are detected as highly conserved (FDR<0.05) in the Zoonomia alignment (red) than the 100-vertebrate alignment (grey), most notably in noncoding regions. **(D)** At a putative androgen receptor binding site overlapping a ChIP-seq peak and a phastCons constrained element prediction, phyloP scores predict seven bases are under purifying selection in the Zoonomia alignment (red; FDR=0.05; two-sided),

peaking in positions with the most information content in the androgen receptor JASPAR[47] motif, compared to one (grey) in the 100-vertebrate alignment, with scores at FDR > 0.05 in light red and light grey.

Methods

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 either the Broad Institute (N=69) or Uppsala University (N=68). For samples received at Broad then sent to Uppsala, shipping approval was secured from the US Fish and Wildlife Service. IACUC approval was not required.

Sample quality control, library construction, and sequencing.

DNA integrity for each sample was visualized via agarose gel (at Broad) or Agilent tape station (at Uppsala). Samples passed QC if the bulk of DNA fragments were greater than 5kb. DNA concentration was then determined using Invitrogen Qubit dsDNA HS assay kit. For each of the samples that passed QC, 1-3µg of DNA was fragmented on the Covaris E220 Instrument using the 400bp standard program (10% Duty cycle, 140 PIP, 200 cycles per burst, 55s). Fragmented samples underwent SPRI double size selection (0.55X, 0.7Xf) followed by PCR-free Illumina library construction following the manufacturer's instructions (Kapa #KK8232) using PCR-free adapters from Illumina (#FC-121-3001). Final library fragment size distribution was determined on Agilent 2100 Bioanalyzer with High Sensitivity DNA Chips. Paired-end libraries were pooled, then sequenced on a single-lane of the Illumina HiSeq2500, set for Version 2 chemistry and 2x250bp reads. This yielded a total of mean 375M (standard deviation = 125M) reads per sample.

Assembly and validation

For each species, we applied DISCOVAR *de novo*[11] (discovardenovo-52488; ftp://ftp.broadinstitute.org/pub/crd/DiscovarDeNovo/) to assemble the 2x250bp 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 40.1x (s.d.+/- 14x). We assessed genome assembly, gene set, and transcriptome completeness using BUSCO, which provides quantitative measures based on gene content from near-universal single-copy orthologs[48]. BUSCO was run with default parameters, using the mammalian gene

model set (mammalia_odb9, n=4104), 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 downloaded from NCBI on April 10, 2019.

Genome upgrades. We selected genomes from each eutherian order without a preexisting long-contiguity assembly based on (1) whether the underlying assembly met the minimum quality threshold needed for HiRise upgrades; (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[49].

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 8 Dovetail-upgraded genomes. Four failed because they were either too fragmented to analyze (N=3) or due to undetermined errors (N=1). One assembly was excluded because it was a second individual from an already represented species.

Heterozygosity calls. We applied the standard GATK pipeline with genotype quality banding to identify the callable fraction of the genome[50,51]. First, we used *samtools* to subsample paired reads from the unmapped bam files[52]. After removing adapter sequences from the selected reads, we used BWA-MEM to map reads to the reference genome scaffolds of >10kb, removing duplicates using the PicardTools MarkDuplicates utility[53]. 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 callable for subsequent analyses. For the lists of high-confidence variant sites, we include only heterozygous positions after filtering at GQ>20, max DP<100, min DP>6, as described in the README file at 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 2x of the average sample read depth, then calculated global heterozygosity as the fraction of heterozygous calls over the whole callable genome.

Calling Segments of Homozygosity (SoH). We estimated the proportion of the genome within segments of homozygosity (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 62kb). We first split all scaffolds into windows with a maximum length of 50kb, with windows ranging from 20kb-50kb for scaffolds <50kb. 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 centered around the lower tail of the distribution and allowing the second to freely capture all the remaining heterozygosity variability[54,55]. 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 see no significant correlation between contig N50 and SoH (Pearson correlation=0.055, p=0.57, N=112).

Assessing the impact of % callable genome. We assessed whether the % of the genome that was callable (Supplementary Table 3) was likely to impact our analysis. The callable % was correlated with heterozygosity (r=-0.80, p<2.2e-16, N=130), and weakly with SoH (r=0.18, p=0.06, N=112). There is no significant difference in callable % among IUCN categories (p_{anova}=0.98; N=122) or between captive and wild populations ($p_{t\text{-}test}$=0.81; N=120).

Analyzing patterns of diversity. We excluded two genomes with exceptionally high heterozygosity (het > 0.02; > 5 standard deviations above the mean). Both were non-endangered, and thus removing them made our determination of lower heterozygosity in endangered species more conservative. Of the remaining 124, we excluded 19 genomes with allelic balance (ab) values more than one standard deviation above the mean (>0.36). Abnormally high ab can indicate sequencing biases with potential for artifacts 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 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 Cactus[37,56]. The topology used for the guide-tree of the alignment was taken from TimeTree[45]; 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[57] for those species which had corresponding entries in the 100-way. For species not present in the 100-way, distance matrix entries were more coarsely estimated using the distance estimated from Mash[58] 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 gene tree and species tree that could occur with either incomplete lineage-sorting or horizontal transfer. However, the guide tree has

minimal impact on the alignment, with little difference between alignments with different trees, even when using a tree that is purposely wrong[37]. Phenomena such as incomplete lineage sorting that affect a subset of species are unlikely to substantially impact the detection of purifying selection across the whole eutherian lineage described in Figure 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. Next, separate clade alignments were generated for each major clade in the alignment: Euarchonta, Glires, Laurasiatheria, and Afrotheria/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 the preprint introducing the progressive mode of Cactus[37]).

We created a neutral model for the conservation analysis using ancestral repeats detected by RepeatMasker[59] on the eutherian ancestral genome produced in the Cactus alignment (tRNA and "low complexity" repeats were removed). To fit the neutral model, we used phyloFit from the PHAST[60] 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. Since phyloFit does not support alignment columns containing duplicates, if a genome had more than one sequence in a single alignment block, they were replaced with a single entry representing the consensus base at each column.

We extracted initial conservation scores using PhyloP from the PHAST[60] 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 Hochberg[61]. 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 on a UCSC Assembly Hub[62] hosted at broad.io/genomes (the link may be loaded into the "Track Hubs" section of the browser) and at https: //alignment-output.s3.amazonaws.com/200m-v1.hal.

Acknowledgements

We thank the many individuals who provided samples and advice, including Chris Adenyo, Catherine Avila, Eric Baitchman, Richard Behringer, Adam Boyko, Matthew Breen, Kevin Campbell, Polly Campbell, Chris J. Conroy, Kimberly Cooper, Liliana M. Dávalos, Frederic Delsuc, Daniel Distel, Christopher Allan Emerling, Julie Fronczek, Neil Gemmel, Jeffrey Good, Kai He, Kris Helgen, Allyson Hindle, Hopi Hoekstra, Rodney Honeycutt, Pavel Hulva, Wiliam Israelsen, Boniface Kayang, Rosalind Kennerley, Marisa Korody, Danielle N. Lee, Edward Louis, Matt MacManes, Ann Misuraca, Anna Mitelberg, Phillip Morin , Alice Mouton , Miho Murayama, Michael Nachman, Asako Navarro, Rob Ogden, Bret Pasch, Sabastien Peuchmaille, TJ Robinson, Stephen Rossiter, Manuel Ruedi, Ashley Seifert, Steven Thomas, Samuel Turvey,

Goedele Verbeylen, and the late Dr. 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).

This project was funded by NIH NHGRI R01HG008742 (EKK, BB, DPG, RS, JTM, JJ, HN, BP, JA), Swedish Research Council Distinguished Professor Award (KLT, VDM, EM, JRSM), Knut and Alice Wallenberg Foundation (KLT, VDM, EM, JRSM), Uppsala University (KLT, VDM, EM, JRSM, JJ, JeA, LG), Broad Institute Next10 (LG), Gladstone Institutes (KSP), NIH NHGRI 5R01HG002939 (ArS, RH), NIH NHGRI 5U24HG010136 (ArS, RH), NIH NHGRI 5R01HG010485 (BP, MD), NIH NHGRI 2U41HG007234 (BP, MD, JA), NIH NIA 5PO1AG047200 (VNG), NIH NIA 1UH2AG064706 (VNG), BFU2017-86471-P MINECO/FEDER, UE (TM), Secretaria d'Universitats i Recerca and CERCA Programme del Departament d'Economia i Coneixement de la Generalitat de Catalunya GRC 2017 SGR 880 (TM), Howard Hughes International Early Career (TM), European Research Council Horizon 2020 #864203 (TM), Obra Social "La Caixa" (TM), BBSRC BBS/E/T/000PR9818, BBS/E/T/ 000PR9783 (WH, WN), BBSRC Core Strategic Programme Grant BB/P016774/1 (WH, WN, FD), Sir Henry Dale Fellowship 200517/Z/16/Z jointly funded by the Wellcome Trust and the Royal Society to NRC (NRC), FJCI-2016-29558 MICINN (DJ), Prince Albert II Foundation of Monaco and Canada, Global Genome Initiative, Smithsonian Institution (MN), European Research Council Research Grant ERC-2012-StG311000 (ECT), UK Medical Research Council MR/P026028/1 (WH, WN), National Science Foundation DEB-1457735 (MS), National Science Foundation DEB-1753760 (WJM), National Science Foundation IOS-2029774 (EKK, DPG), Robert and Rosabel Osborne Endowment (HL, JD), Swedish Research Council, FORMAS 221-2012-1531 (JRSM), NSF RoL: FELS: EAGER: DEB 1838283 (DR), Academy of Finland Finnish Center of Excellence #312041 (TK).

Consortium

Diane P. Genereux[1], Aitor Serres[2,3], Joel Armstrong[4], Jeremy Johnson[1], Voichita D. Marinescu[5], Eva Murén[5], David Juan[2,3], Gill Bejerano[6,7,8,9], Nicholas R. Casewell[10], Leona G. Chemnick[11], Joana Damas[12], Federica Di Palma[13,14], Mark Diekhans[4], Ian T. Fiddes[4], Manuel Garber[15], Vadim N. Gladyshev[1,16], Linda Goodman[1,17], Wilfried Haerty[14], Marlys L Houck[11], Robert Hubley[18], Teemu Kivioja[19,20,21], Klaus-Peter Koepfli[22], Lukas F. K. Kuderna[3], Eric S. Lander[1,23,24], Jennifer R. S. Meadows[5], William J. Murphy[25], Will Nash[14], Hyun Ji Noh[1], Martin Nweeia[26,27,28], Andreas R. Pfenning[29], Katherine S. Pollard[30,31,32], David Ray[33], Beth Shapiro[34,35], Arian Smit[18], Mark Springer[36], Cynthia C. Steiner[11], Ross Swofford[1], Jussi Taipale[19,21,37], Emma C. Teeling[38], Jason Turner-Maier[1], Jessica Alfoldi[1], Bruce Birren[1], Oliver A. Ryder[11], Harris Lewin[12,39], Benedict Paten[4], Tomas Marques-Bonet[2,3,40,41], Kerstin Lindblad-Toh[1,5], Elinor K. Karlsson[1,15,42]

(1) Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA; (2) CNAG-CRG, Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST),

Barcelona, Spain; (3) Institute of Evolutionary Biology (UPF-CSIC), PRBB, Barcelona, Catalonia, Spain; (4) Center for Biomolecular Science and Engineering, University of California Santa Cruz, Santa Cruz, CA, USA; (5) Science for Life Laboratory, Department of Medical Biochemistry and Microbiology, Uppsala University, Uppsala, Sweden; (6) Department of Biomedical Data Science, Stanford University, Stanford, CA, USA; (7) Department of Computer Science, Stanford University, Stanford, CA, USA; (8) Department of Developmental Biology, Stanford University, Stanford, CA, USA; (9) Department of Pediatrics, Stanford University, Stanford, CA, USA; (10) Liverpool School of Tropical Medicine, Liverpool, UK; (11) San Diego Zoo Institute for Conservation Research, San Diego, California, USA; (12) The UC Davis Genome Center, University of California, Davis, California, USA; (13) Department of Biological Sciences, University of East Anglia, Norwich, UK; (14) Earlham Institute, Norwich Research Park, Norwich, UK; (15) Program in Bioinformatics and Integrative Biology, University of Massachusetts Medical School, Worcester, Massachusetts, USA; (16) Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts, USA; (17) Stanford University, Stanford, California, USA; (18) Institute for Systems Biology, Seattle, Washington, USA; (19) Department of Biochemistry, University of Cambridge, Cambridge, United Kingdom; (20) Department of Medical Biochemistry and Biophysics, Karolinska Institutet, Stockholm, Sweden; (21) University of Helsinki, Helsinki, Finland; (22) Smithsonian Conservation Biology Institute, Center for Species Survival, National Zoological Park, Front Royal, Virginia and Washington, DC, USA; (23) Department of Biology, MIT, Cambridge, MA, USA; (24) Department of Systems Biology, Harvard Medical School, Boston, MA, USA; (25) Veterinary Integrative Biosciences, Texas A&M University, College Station, Texas, USA; (26) Marine Mammal Program, Smithsonian Institution, Washington, DC, USA; (27) Restorative Dentistry and Biomaterials Sciences, Harvard School of Dental Medicine, Boston, Massachusetts, USA; (28) School of Dental Medicine, Case Western Reserve University, Cleveland, Ohio, USA; (29) Carnegie Mellon University, School of Computer Science, Department of Computational Biology, Pittsburgh, Pennsylvania, USA; (30) Chan-Zuckerberg Biohub, San Francisco, California, USA; (31) Gladstone Institutes, San Francisco, California, USA; (32) University of California, San Francisco, California, USA; (33) Department of Biological Sciences, Texas Tech University, Lubbock, Texas, USA; (34) Ecology and Evolutionary Biology, University of California, Santa Cruz, California, USA; (35) Howard Hughes Medical Institute, Chevy Chase, Maryland, USA; (36) University of California, Riverside, California, USA; (37) Karolinska Institute, Solna, Sweden; (38) School of Biology and Environmental Science, University College Dublin, Dublin, Ireland; (39) Evolution and Ecology, University of California, Davis, California, USA; (40) Catalan Institution of Research and Advanced Studies (ICREA), Barcelona, Spain; (41) Institut Català de Paleontologia Miquel Crusafont, Universitat Autònoma de Barcelona, Edifici ICTA-ICP, Barcelona, Spain; (42) Program in Molecular Medicine, University of Massachusetts Medical School, Worcester, Massachusetts, USA

Author contributions

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

KLT, conceived the project, JJ, VDM, EM, NRC, LGC, JD, VNG, MLH, KPK, JRSM, WJM, MN, DR, RS, ECT, JeA, OAR, HL, KLT and EKK, contributed to the acquisition of the samples. JJ, VDM, EM, JD, LG, KPK, HN, CCS, RS, JTM, JeA, OAR, HL, KLT and EKK, contributed to the production of the genome assemblies. DPG, AS, JA, JJ, DJ, ITF, LFKK, HL, TM, KLT and EKK, contributed to the data analysis.. DPG, JJ, VDM, GB, FD, MD, ITF, MG, VNG, WH, RH, TK, ESL, JRSM, ARP, KSP, ArS, MS, JT, JeA, BB, OAR, HL, BP, TM, KLT and EKK contributed to the design and conduct of the project. DPG, ESL, WN, BS, OAR, KLT, and EKK, wrote the manuscript, with input from all other authors

Competing interests

LG 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.

Correspondence and requests for materials should be addressed to elinor@broadinstitute.org.

Reprints and permissions information is available at www.nature.com/reprints.

Data Availability

Details on each Zoonomia Project genome assembly, including NCBI Genbank[63] 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 http://broad.io/variants. Source data for Figure 2 is provided and in the Zoonomia github repository (DOI 10.5281/zenodo.3887432). The Cactus alignment is at https://alignment-output.s3.amazonaws.com/200m-v1.hal. A visualization of the alignments and PhyloP data is available by loading our assembly hub into the UCSC browser[62] 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.

Code Availability

The Discovar *de novo* assembly code is available at https://github.com/broadinstitute/discovar_de_novo/releases/tag/v52488 (DOI 10.5281/zenodo.3870889), the Cactus pipeline is available at https://github.com/ComparativeGenomicsToolkit/cactus (DOI 10.5281/zenodo.3873410) and

code for other analyses is available at https://github.com/broadinstitute/Zoonomia/ (DOI 10.5281/zenodo.3887432).

Extended Data Figure 1. Remarkable traits in non-human mammals.

Sequences from species with remarkable 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[64], achieving homeostasis to a degree elusive in treatment of human diabetes. **(B)** The North American beaver (*Castor canadensis)* avoids tooth decay by incorporating iron, rather than magnesium, into tooth enamel, yielding an orange hue[65]. (**C**) The thirteen-lined ground squirrel (*Ictidomys tridecemlineatus)* prepares for hibernation by rapidly increasing the thermogenic activity of brown fat[66], a process connected to improved glucose homeostasis and insulin sensitivity in humans[67–69]. (**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: (A) Merlin D. Tuttle/Science Source; (B) Stephen J. Krasemann/Science Source; (C) Allyson Hindle; (D) Sébastien J. Puechmaille [CC BY-SA]; (E) M. Watson/Science Source.

Extended Data Figure 2. Sample collection can be challenging, and sequencing methods must be selected to handle the sample quality. To enable 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 like the narwhal (*Monodon monoceros)*, simply accessing an individual in the wild can prove challenging. To sample DNA from the near-threatened narwhal, for example, Martin Nweeia and Inuit guide David Angnatsiak camped on an ice floe edge between Pond Inlet and Bylot Island, at the northeastern tip of Baffin Island. After a narwhal was harvested by Inuit hunters as part of an annual hunt, hours of flensing were necessary for collecting tissue samples. Shown, from left to right: Frank McCann, Hans Christian Schmidt, Frederick Eichmiller, Martin Nweeia, James Orr (facing backward), and Jack Orr (standing). **(b)** For endangered species like the Hispaniolan solenodon (*Solenodon paradoxus*), sample collection must be designed to minimize stress to the individual, limiting the amount of DNA that can be collected[22]. To collect DNA from the endangered solenodon without imposing stress on individuals in the wild, Nicholas Casewell turned to the world's only captive solenodons, housed off-exhibit at ZOODOM in the Dominican Republic. With help from Zoo veterinarians, Casewell collected a small amount of blood from

the solenodon's rugged tail. Narwhal photograph by Gretchen Freund and courtesy of Martin Nweeia. Solenodon photo courtesy of Lucy Emery.

Extended Data Table 1. The Zoonomia Project data includes 132 genome assemblies. These assemblies include 131 different species, with two narwhals (male and female), and 10 genomes upgraded to longer contiguity (including upgrade of an existing assembly for *Echinops 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; † upgraded to longer contiguity using existing assembly.

Extended Data Table 2: Power to detect constraint across data sets. 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 v3[70] 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 (human genome size). Branch length for Zoonomia was measured as substitutions/site in the phyloP analysis of the Cactus alignment.

Extended Data Table 3. Diversity statistics are not correlated with other species-level phenotypes. All phenotypes in the Pantheria database[30] for which at least 75% of the 75 "Least Concern" species had a value were included in the analysis. For continuous phenotypes, values were standardized to Z-scores prior to 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 were significant after Bonferroni correction for the number of traits considered (21).

References

1. Claussnitzer, M. *et al.* A brief history of human disease genetics. *Nature* **577**, 179–189 (2020).
2. Hiller, M. *et al.* A 'forward genomics' approach links genotype to phenotype using independent phenotypic losses among related species. *Cell Rep.* **2**, 817–823 (2012).
3. Wasser, S. K. *et al.* Genetic assignment of large seizures of elephant ivory reveals Africa's major poaching hotspots. *Science* **349**, 84–87 (2015).
4. Wright, B. *et al.* Development of a SNP-based assay for measuring genetic diversity in the Tasmanian devil insurance population. *BMC Genomics* **16**, 791 (2015).
5. Lappalainen, T., Scott, A. J., Brandt, M. & Hall, I. M. Genomic Analysis in the Age of Human Genome Sequencing. *Cell* **177**, 70–84 (2019).
6. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–315 (2014).
7. Lindblad-Toh, K. *et al.* A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478**, 476–482 (2011).
8. Finucane, H. K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).
9. IUCN. The IUCN Red List of Threatened Species. Version 2019-2. http://www.iucnredlist.org (2019).
10. Ryder, O. A. & Onuma, M. Viable Cell Culture Banking for Biodiversity Characterization and Conservation. *Annu Rev Anim Biosci* **6**, 83–98 (2018).
11. Weisenfeld, N. I. *et al.* Comprehensive variation discovery in single human genomes. *Nat. Genet.* **46**, 1350–1355 (2014).
12. Putnam, N. H. *et al.* Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. *Genome Res.* **26**, 342–350 (2016).
13. Kim, J. *et al.* Reconstruction and evolutionary history of eutherian chromosomes. *Proc. Natl. Acad. Sci. U. S. A.* **114**, E5379–E5388 (2017).

14. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
15. Balasubramanian, S. *et al.* Using ALoFT to determine the impact of putative loss-of-function variants in protein-coding genes. *Nat. Commun.* **8**, 382 (2017).
16. Meadows, J. R. S. & Lindblad-Toh, K. Dissecting evolution and disease using comparative vertebrate genomics. *Nat. Rev. Genet.* **18**, 624–636 (2017).
17. Cooper, G. M. & Shendure, J. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. *Nat. Rev. Genet.* **12**, 628–640 (2011).
18. Baiz, M. D., Tucker, P. K. & Cortés-Ortiz, L. Multiple forms of selection shape reproductive isolation in a primate hybrid zone. *Mol. Ecol.* (2018) doi:10.1111/mec.14966.
19. Dobzhansky, T. & Dobzhansky, T. G. *Genetics and the Origin of Species*. (Columbia University Press, 1937).
20. Abegglen, L. M. *et al.* Potential Mechanisms for Cancer Resistance in Elephants and Comparative Cellular Response to DNA Damage in Humans. *JAMA* **314**, 1850–1860 (2015).
21. Herrera-Álvarez, S., Karlsson, E., Ryder, O. A., Lindblad-Toh, K. & Crawford, A. J. How to make a rodent giant: Genomic basis and tradeoffs of gigantism in the capybara, the world's largest rodent. *bioRxiv* 424606 (2018) doi:10.1101/424606.
22. Casewell, N. R. *et al.* Solenodon genome reveals convergent evolution of venom in eulipotyphlan mammals. *Proc. Natl. Acad. Sci. U. S. A.* (2019) doi:10.1073/pnas.1906117.
23. Beichman, A. C. *et al.* Aquatic adaptation and depleted diversity: a deep dive into the genomes of the sea otter and giant otter. *Mol. Biol. Evol.* (2019) doi:10.1093/molbev/msz101.
24. Damas, J. *et al.* Broad Host Range of SARS-CoV-2 Predicted by Comparative and Structural Analysis of ACE2 in Vertebrates. *bioRxiv* 2020.04.16.045302 (2020) doi:10.1101/2020.04.16.045302.
25. Xue, Y. *et al.* Mountain gorilla genomes reveal the impact of long-term population decline and inbreeding. *Science* vol. 348 242–245 (2015).

26. 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).
27. Spielman, D., Brook, B. W. & Frankham, R. Most species are not driven to extinction before genetic factors impact them. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 15261–15264 (2004).
28. Vinson, J. P. *et al.* Assembly of polymorphic genomes: algorithms and application to Ciona savignyi. *Genome Res.* **15**, 1127–1135 (2005).
29. MacManes, M. D. & Lacey, E. A. The social brain: transcriptome assembly and characterization of the hippocampus from a social subterranean rodent, the colonial tuco-tuco (Ctenomys sociabilis). *PLoS One* **7**, e45524 (2012).
30. Jones, K. E. *et al.* PanTHERIA: a species-level database of life history, ecology, and geography of extant and recently extinct mammals: Ecological Archives E090-184. *Ecology* **90**, 2648–2648 (2009).
31. Cardillo, M. Biological determinants of extinction risk: why are smaller species less vulnerable? *Anim. Conserv.* **6**, 63–69 (2003).
32. Natesh, M. *et al.* Empowering conservation practice with efficient and economical genotyping from poor quality samples. *Methods Ecol. Evol.* **10**, 853–859 (2019).
33. Lowry, D. B. *et al.* Breaking RAD: an evaluation of the utility of restriction site-associated DNA sequencing for genome scans of adaptation. *Mol. Ecol. Resour.* **17**, 142–152 (2017).
34. Shapiro, B. Pathways to de-extinction: how close can we get to resurrection of an extinct species? *Funct. Ecol.* **31**, 996–1002 (2017).
35. Benazzo, A. *et al.* Survival and divergence in a small group: The extraordinary genomic history of the endangered Apennine brown bear stragglers. *Proc. Natl. Acad. Sci. U. S. A.* **114**, E9589–E9597 (2017).
36. Saremi, N. F. *et al.* Puma genomes from North and South America provide insights into the genomic consequences of inbreeding. *Nat. Commun.* **10**, 4769 (2019).
37. Armstrong, J. *et al.* Progressive alignment with Cactus: a multiple-genome aligner for the thousand-

genome era. (in submission).

38. Rands, C. M., Meader, S., Ponting, C. P. & Lunter, G. 8.2% of the Human genome is constrained: variation in rates of turnover across functional element classes in the human lineage. *PLoS Genet.* **10**, e1004525 (2014).

39. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).

40. GTEx Consortium *et al.* Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).

41. Regev, A. *et al.* The Human Cell Atlas. *Elife* **6**, (2017).

42. Lewin, H. A. *et al.* Earth BioGenome Project: Sequencing life for the future of life. *Proc. Natl. Acad. Sci. U. S. A.* **115**, 4325–4333 (2018).

43. Koepfli, K.-P., Paten, B., Genome 10K Community of Scientists & O'Brien, S. J. The Genome 10K Project: a way forward. *Annu Rev Anim Biosci* **3**, 57–111 (2015).

44. Teeling, E. C. *et al.* Bat Biology, Genomes, and the Bat1K Project: To Generate Chromosome-Level Genomes for All Living Bat Species. *Annu Rev Anim Biosci* **6**, 23–46 (2018).

45. Kumar, S., Stecher, G., Suleski, M. & Hedges, S. B. TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. *Mol. Biol. Evol.* **34**, 1812–1819 (2017).

46. *Mammal Species of the World. A Taxonomic and Geographic Reference (3rd ed)*. (Johns Hopkins University Press, 2005).

47. Vlieghe, D. *et al.* A new generation of JASPAR, the open-access repository for transcription factor binding site profiles. *Nucleic Acids Res.* **34**, D95–7 (2006).

Extended Data References

48. 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).
49. Farré, M. *et al.* A near-chromosome-scale genome assembly of the gemsbok (Oryx gazella): an iconic antelope of the Kalahari desert. *Gigascience* **8**, (2019).
50. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
51. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
52. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
53. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
54. Benaglia, T., Chauveau, D., Hunter, D. & Young, D. mixtools: An R package for analyzing finite mixture models. *J. Stat. Softw.* **32**, 1–29 (2009).
55. R Core Team. R: A Language and Environment for Statistical Computing. https://www.R-project.org/ (2019).
56. Paten, B. *et al.* Cactus: Algorithms for genome multiple sequence alignment. *Genome Res.* **21**, 1512–1528 (2011).
57. Haeussler, M. *et al.* The UCSC Genome Browser database: 2019 update. *Nucleic Acids Res.* **47**, D853–D858 (2019).
58. Ondov, B. D. *et al.* Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.* **17**, 132 (2016).
59. Smit, A. F. A. and Hubley, R. and Green, P. RepeatMasker Open-4.0. (2013-2015).
60. Hubisz, M. J., Pollard, K. S. & Siepel, A. PHAST and RPHAST: phylogenetic analysis with

space/time models. *Brief. Bioinform.* **12**, 41–51 (2011).

61. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)* vol. 57 289–300 (1995).

62. Nguyen, N. *et al.* Comparative assembly hubs: web-accessible browsers for comparative genomics. *Bioinformatics* **30**, 3293–3301 (2014).

63. Benson, D. A. *et al.* GenBank. *Nucleic Acids Res.* **41**, D36–42 (2013).

64. 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).

65. Gordon, L. M. *et al.* Dental materials. Amorphous intergranular phases control the properties of rodent tooth enamel. *Science* **347**, 746–750 (2015).

66. 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**, E284–99 (2014).

67. Stanford, K. I. *et al.* Brown adipose tissue regulates glucose homeostasis and insulin sensitivity. *J. Clin. Invest.* **123**, 215–223 (2013).

68. Chondronikola, M. *et al.* Brown adipose tissue improves whole-body glucose homeostasis and insulin sensitivity in humans. *Diabetes* **63**, 4089–4099 (2014).

69. 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).

70. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).