Loss of critical developmental and human disease-causing genes in 58 mammals

Yatish Turakhia^{1,†}, Heidi I. Chen^{2,†}, Amir Marcovitz^{2,†} & Gill Bejerano^{2,3,4,5,*}

¹Department of Electrical Engineering, Stanford University, Stanford, California 94305, USA

²Department of Developmental Biology, Stanford University, Stanford, California 94305, USA

³Department of Computer Science, Stanford University, Stanford, California 94305, USA

⁴Department of Biomedical Data Science, Stanford University, Stanford, California 94305, USA

⁵Department of Pediatrics, Stanford University, Stanford, California 94305, USA

*To whom correspondence should be addressed:

Beckman Center B-300

279 Campus Drive West (MC 5329)

Stanford CA 94305-5329

+1 (650) 725 6792

bejerano@stanford.edu

[†]equal contribution

Gene losses provide an insightful route for studying the morphological and physiological adaptations of species, but their discovery is challenging. Existing genome annotation tools and protein databases focus on annotating intact genes and do not attempt to distinguish nonfunctional genes from genes missing annotation due to sequencing and assembly artifacts. Previous attempts to annotate gene losses have required significant manual curation, which hampers their scalability for the ever-increasing deluge of newly sequenced genomes. Using extreme sequence erosion (deletion and non-synonymous substitution) as an unambiguous signature of loss, we developed an automated approach for detecting high-confidence protein-coding gene loss events across a species tree. Our approach relies solely on gene annotation in a single reference genome, raw assemblies for the remaining species to analyze,

and the associated phylogenetic tree for all organisms involved. Using the hg38 human assembly as a reference, we discovered over 500 unique human genes affected by such high-confidence erosion events in different clades across 58 mammals. While most of these events likely have benign consequences, we also found dozens of clade-specific gene losses that result in early lethality in outgroup mammals or are associated with severe congenital diseases in humans. Our discoveries yield intriguing potential for translational medical genetics and for evolutionary biology, and our approach is readily applicable to large-scale genome sequencing efforts across the tree of life.

Introduction

The placental mammal radiation exhibits tremendous phenotypic diversity [1]. These myriad forms and functions arose from modification of inherited traits from one generation to the next. An extreme case of this “descent with modification” involves disappearance (either sudden or gradual) of ancestral phenotypes, caused by one or more function-inactivating mutations. If it fixes in the species, this functional inactivation will be accompanied by relaxation of purifying selection on all genomic regions encoding the trait, allowing erosion of the sequence. When a gene loses its functional role this way, the result can be a so-called “gene loss” – a nonfunctional, dead gene [2].

Previous studies have identified a few intriguing examples of gene losses that are

associated with phenotypic traits. For example, the enzyme gene *GULO* is inactivated in independent mammalian lineages that consequently lost the ability to synthesize Vitamin C, and must instead have a constant supply of it in their diet [3, 4]; multiple visual system genes are no longer functional in subterranean mammals which live primarily in darkness [5, 6]; taste receptor and other genes are lost in fully aquatic cetaceans [7–9]; renal transporter genes *URAT1*, *GLUT9* and *OAT1* are dead in fruit-eating bats, which could have facilitated their frugivorous diet [10]; the immune genes *MX1* and *MX2* are eroded in toothed whales, possibly making them more susceptible to certain viral pathogens [11]; and the loss of *PON1* in several marine mammals may confer increased vulnerability to agricultural pesticide pollution [12]. Systematic annotation of gene losses across species may therefore not only reveal fascinating evolutionary events and genotype-phenotype relationships, but could also point us to “natural knockout” models for human pathologies that reveal compensating molecular pathways in species missing otherwise-indispensable genes [2, 3, 13].

Manual annotation of pseudogenized genes, such as by ENCODE-HAVANA [14], while accurate, has been applied to only a handful of species (human, mouse, rat and zebrafish) and is not scalable for hundreds to thousands of newly sequenced species [15]. Automating the discovery of gene losses is scalable but challenging. Recently, Sharma et al. [10] developed an automated method using whole-genome multiple-sequence alignment to identify genes with inactivating mutations. However, their overly inclusive approach predicted hundreds of gene losses per species across 62 mammals, of which only 21 total were reported, presumably because manual inspection found many loss predictions to

be spurious. Similar technique was used to further semi-manually annotate 85 gene losses in cetaceans [16]. Other computational approaches for gene annotation, like the Ensembl pipeline [17], combine *ab initio* predictions with protein sequence evidence [18] and with gene models inferred from RNA-seq data. However, they focus only on annotating *intact* genes and do not attempt to distinguish decaying genes from those that only confoundingly appear so, for instance, as a result of sequencing gaps and alignment errors. By design, these tools are thus limited in their ability to identify gene loss events.

We developed an improved, conservative approach for identifying gene loss events and applied it to 58 mammalian genomes, using human assembly hg38 as the reference. We hypothesized that the most accurate strategy for declaring an ancestral gene function to be lost is to demonstrate that it is highly sequence-eroded compared to other genes in the same genome. While it is impossible to guarantee that no gene transcript is expressed from the orthologous locus (especially if its promoter remains intact), the erosion of gene sequence implies that the gene does not maintain its original function. To that end, we used pairwise whole-genome alignments to locate the orthologs of human protein-coding genes in each of the 58 query species by applying a novel synteny-aware mapping procedure. We employed a Mahalanobis distance-based model to identify the most highly mutated orthologs in each genome in a way that normalizes for the baseline extent of sequence divergence between this species and the reference (human). We considered these per-genome predictions to be *likely eroded* and functionally dead orthologs in the species represented by the assembly. Lastly, we applied a stringent phylogenetic filter, requiring that all final candidate erosions were supported by observations from at least two closely

related species and affected ancestral genes – in order to minimize confounders due to assembly artefacts, private (individual-specific) mutations, or reference genome biases. Though very conservative, our approach uncovered hundreds of *high-confidence ortholog erosion* (*hiconfErosion*) events, affecting over 50 mammals, including some very surprising gene losses with intriguing evolutionary and biomedical implications. Importantly, our method was designed to allow easy addition of genomes to the analysis, such as the those from countless newly sequenced mammals, birds [19], vertebrates [20], insects [21] and more.

Results

Pre-existing gene annotation methods fail to identify and determine the functional state for thousands of orthologs

Of the 59 mammalian genomes (including the human reference) that we used in our study, we found Ensembl gene predictions (release 86) for only 34 genomes. The number of protein-coding gene annotations varied widely (Figure 1A), from 11,765 annotations in alpaca to 22,285 annotations in human. When plotting the number of Ensembl predicted genes against the assembly N50, it is clear that the annotation completeness is artifactually dependent on assembly contiguity or lack thereof (Figure 1A). For instance, absence of the critical chordate development gene *PAX6* in the dog annotation set is likely due to a

sequencing gap in the genome assembly (Figure 1B). Similarly, while *TP53* and multiple adjacent genes are absent from the alpaca annotation set, its genome includes a region with high sequence similarity to the human *TP53* ortholog (Figure 1C), implying a potentially missed prediction. We have observed that pseudogene annotation pipelines are also incomplete – for example, manual curation of mouse pseudogenes by ENCODE-HAVANA has been on-going for several years, but the list is still not available online (<https://www.encodegenes.org/>). Based on these observations, we devised an improved approach for identifying a high-confidence, conservative subset of gene loss events across the 58 placental mammals, using human as the reference.

Mapping gene orthologs across mammalian genomes

To identify *ab initio* orthologs of human genes across the 58 query species, we used whole-genome pairwise alignments (i.e. Jim Kent’s BLASTZ-based *chains* [22]) to map genes from human coordinate space to each of the other genomes. Chains identify conserved sequences not only in coding regions (covering 2% of the human genome) but also in non-coding, gene-regulatory regions (covering 5-10% of the human genome) in which the genes are interspersed. As such, chains facilitate accurately determining genomic orthology between species. Using these 58 pairwise alignment chains anchored to the human reference, we determined the single locus in the aligning (query) genome that exhibited the greatest sequence similarity and surrounding synteny with each human protein-coding gene. Specifically, each successfully mapped gene was (i) positioned within a chain that

had alignment and gene-in-syteny scores sufficiently higher than the next-highest-scoring (likely paralogous) chain and (ii) did not overlap in genomic space with any other gene mappings proposed in the previous step (see Methods, Figure 2A,B). In total, our mapping procedure identified 11,155 - 17,690 orthologs per species (median: 16,782; Figure 3A) across the 58 non-human mammals considered.

Importantly, this chain-based mapping approach succeeded for previously discovered gene losses, even accurately deciphering pseudogenized orthologs with extensive sequence erosion compared to the ancestral gene. For example, much of the exonic portions of *SERPINA10* are absent in the dog genome and do not yield a functional gene product [23]. As such, Ensembl does not offer a corresponding gene annotation in the dog genome assembly (canFam3), which would leave ambiguity about the existence and function this ortholog in dogs if not for experimental evidence. Our approach, however, precisely located the remnants of *SERPINA10* in the canFam3 assembly and allowed for assessing the semantic status (reading frame integrity and occurrence of substitution, insertion, and deletion) of the remaining chain-mapped exons (Figure 2C).

Per-species identification of likely eroded orthologs using a Mahalanobis distance outlier model

Because functional inactivation will be accompanied by relaxation of purifying selection on all genomic regions encoding the trait, we used the extent of coding sequence erosion to

declare the loss of gene function. In particular, to account for mutation rate heterogeneity across different branches of the mammalian tree, we restricted our results to orthologs that appear extremely eroded (using outlier detection) in sequence *relative* to all other mapped genes in the same given species. Unlike Sharma et al. [10], who used single-event features, such as in-frame stop codons and frameshift indels, we considered two different aggregate features of sequence erosion – i.e. the fraction of exonic nucleotides causing amino acid substitutions and the fraction out-right deleted with respect to the sequence of the human ortholog. This way, we avoided making spurious outlier calls caused by some mutations, such as stop codons near the end of a gene or multiple frameshift indels that restore the original frame, which do not necessarily lead to loss of the ortholog’s function. Additionally, before declaring outliers, we masked out (i.e. did not consider) alignment gaps that could have resulted from sequencing gaps in the target species (see Methods).

For outlier detection, we used a Mahalanobis distance (MD) [24, 25] based model (see Methods) to mark a small fraction of the orthologs (typically 1-2% or 120-284 genes per species, median of 192; Figure 3A) as outliers for their extreme fraction both of non-synonymous substitutions and of deleted bases. We considered these outlier orthologs to be likely eroded in sequence and function in the species represented by the given genome assembly. For instance, by performing these methods on the 17,836 genes confidently mapped to the mm10 mouse assembly (Figure 3B), we distinguished previously discovered gene losses for *Cetp* [26] and *Eys* [27] (plus other novel predictions) as outliers ($MD > 15$), compared to uneroded, relatively sequence-conserved orthologs ($MD \leq 5$), including genes *Tp53* [28], *Notch1* [29], and *Sox11* [30] (Figure 3B,C) with indispensable

function in mice.

Phylogenetic filtering to identify high-confidence ortholog erosions in multiple related species

Two caveats must be considered for the per-species gene erosion predictions: Firstly, not all genes in the human reference set were necessarily present in functional form in the common ancestor shared by human and the a species under consideration. Sequences in other species that have some resemblance (such as paralogs) to a functional human gene, and that appear eroded, but were never present/functional in their common ancestor with human, could be confounded for gene erosion. Secondly, the erosion could be real but private to the single sequenced individual (i.e. not representative of the species at-large).

To address these caveats, we applied a number of strict phylogenetic filters to restrict the final candidate list to the highest-confidence predictions (Figure 3D). First, we required a final candidate to have an uneroded, sequence-conserved ortholog ($MD < 5$) in at least one other mammal in addition to human. With the assumption of parsimony, this first step ensures that the candidate ortholog was present, of functional importance, and under purifying selection in the common ancestor of human and any other extant lineages in which this ortholog is conserved in sequence. Second, even more importantly, we also required that each final candidate was considered likely eroded in *two or more* phylogenetically neighboring species (such as rat/mouse or manatee/elephant/cape elephant shrew) sharing

a common ancestor with human in which the gene was declared conserved. Doing so minimized false positives from gene losses that are real but private to a sequenced individual, as well as those resulting from assembly, alignment, and mapping artifacts. We designated the final candidates as *high-confidence ortholog erosions*, or *hiconfErosions*.

In all, we identified 412 genes that were eroded in one or more multi-species groups, yielding 2,192 and 539 unique ortholog-species and ortholog-clade pairs, respectively (Figure 3A). Our hiconfErosion predictions include many previously known functional losses of *Cetp* [26], *Eys* [27], and *Tcn1* [31] in mouse, *ZPI* in bovine genomes [32], *CRYGB* in subterranean mammals [5], *GSDMA* in whales [10], and *RHBG* in fruit bats [10]. The number of hiconfErosions per species ranged from 5 in aardvark to 88 in brush-tailed rat and chinchilla (median of 41, Figure 4A).

To estimate the false positive rate (FPR) for these results, we tabulated how many predicted hiconfErosions for mouse and rat were contradicted by proteomic or transcriptomic evidence from UniProt [18] (see Methods). We found 6/68 (8.82%) and 2/68 (2.94%) hiconfErosion predictions in mouse and rat conflicting with Uniprot evidence, respectively. We conservatively re-scaled the conflicts to account for potentially missing UniProt entries but still found the FPR estimate to be less than 10% in both species (Figure 4B, see Methods). In contrast, if we applied our phylogenetic filtering directly to the orthologs missing Ensembl annotations for mouse and rat, the FPR for both species would be above 30% (Figure 4B), highlighting the importance of our orthologous chain mapping and outlier detection approach.

Selection and variation in mammalian genes exhibiting multi-species ortholog erosion

Next, we asked how many hiconfErosions involve genes whose knockout mouse models do not survive beyond early development and, surprisingly, found ten such genes (Table 1, see Methods). We also found 27 hiconfErosions involve HGMD [33] genes implicated in severe human congenital disorders (Table 2, see Methods). Using the Residual Variation Intolerance Score (RVIS) [34], which is a gene-based measure of intolerance to common genetic variation (with minor allele frequencies greater than 5%), we found another 13 genes affected by hiconfErosions that have a relatively low RVIS score percentile ($< 20\%$, Table 3, see Methods). We also found that our final hiconfErosion predictions (), containing 412 unique human orthologs, is depleted of genes causing lethality when inactivated in developing mice ($p\text{-value} < 2.2e-16$; $\text{fold} = 8.43$; Figure 5A), of known human HGMD disease-causing genes ($p\text{-value} < 1.3e-11$, $\text{fold} = 3.17$; Figure 5B), and of those intolerant to functional variation in humans ($p\text{-value} < 1e-6$, Figure 5C).

Erosion of *GCKR* in microbats may facilitate their high-fat insectivorous diet

Microbats and megabats, the two primary suborders of bats, have vastly different diets. Microbats have extremely fat-rich, insect-based diets [35], while megabats have fruit-based diets depleted of fat [36]. We found that *GCKR* (Glucokinase regulatory protein; OMIM:

600842), variants of which are associated with elevated high-density lipoprotein (HDL) cholesterol in human [37], is specifically eroded in microbats (see Figure 6 and Table 2). This potentially provides a genotypic basis for an observation by Widmaier et al. [38] who found that HDL-cholesterol levels in a lactating microbat species (*Tadarida brasiliensis mexicana*) were 10-fold higher in comparison to three species of megabats. Specifically, they proposed that these extraordinarily high HDL levels, possibly facilitated by *GCKR* erosion, could explain why microbats are protected from atherosclerosis and other common diseases associated with high-fat intake.

Erosion of *PKD1L1* and *MMP21* evolved in artiodactyla (even-toed ungulates) but causes prenatal lethality in mice and congenital heart defects in humans

Pkd1l1 (polycystin kidney disease protein 1-like 1) null mutations in mice result in a high rate of embryonic lethality [39] (Table 1) characterized by misregulated nodal signaling [40] and anomalous Left-Right (L-R) symmetry patterning [41, 42] (Figure 7A). Indeed, humans with homozygous loss of function in *PKD1L1* (OMIM: 609721) present with laterality defects ranging from *Situs Inversus Totalis* to heterotaxy and with congenital heart disorders, all of which can be deleterious early in life [43]. As such, it was remarkable to find *PKD1L1* erosion, with deletion of as much as 98% of coding bases in some species, in a clade of 9 even-toed ungulates (including cetaceans, see Figure 7B). Congruent with this finding, an Ensembl annotation for *PKD1L1* does not exist for the cow, sheep,

dolphin, alpaca and pig genomes, and we detected no orthology between human *PKDILI* to any other gene in these genomes. In contrast, we found high sequence homology for the human protein sequence of *PKDILI* in the genomes of outgroup lineages such as horse and elephant, suggesting that the gene is intact in those species, and lost in artiodactyla (even-toed ungulates).

Interestingly, we also discovered a related hiconfErosion in the same group of even-toed ungulates (including cetaceans) affecting the gene *MMP21* (Matrix Metalloproteinase 21; OMIM: 608416; Figure 7C) that encodes components of Nodal signaling [43] and is also implicated in heterotaxy [44] and in congenital heart disorders [45, 46] (Figure 7A). Interestingly, Double Outlet Right Ventricle – a cardiac disorder associated with mutation in mouse *Mmp21* [47] and with mutation in human *PKDILI* [43] – has significantly higher incidence in cattle with congenital heart disorder relative to human and other animals with this condition [48] (presumably because hiconfErosion of *both* *PKDILI* and *MMP21* naturally occurs in artiodactyla but not in other mammals).

Erosion of endocrine regulator *GPRC6A* evolved in odontoceti (toothed whales) but causes testicular insufficiency in humans and mice

GPRC6A (OMIM:613572) is a G protein-coupled receptor that plays an important role in metabolic and endocrine regulation [49]. Highly expressed in Leydig cells of the testis (in addition to many other tissues), *GPRC6A* is activated by testosterone, osteocalcin, basic

amino acids, and various cations [50, 51]. In mouse models, *Gprc6a* ablation results in dramatic decrease of testosterone levels, smaller testis size, and reduced sperm count [52]. Similarly in humans, loss-of-function mutations in *GPRC6A* are associated with testicular insufficiency (subfertility, altered sperm parameters, low circulating testosterone levels, and high circulating luteinizing hormone levels). A number of studies also found that *GPRC6A* is associated with prostate cancer progression and could serve as a potential drug target for this malignancy [51, 53].

GPRC6A is a highly conserved gene across mammals and even has orthologs in species as distant as zebrafish (Figure 8), suggesting it may have arisen early in vertebrate evolution. It is highly remarkable, then, that we observed hiconfErosion of *GPRC6A* in dolphin and killer whale – where over 50% of coding bases are deleted in both species (including the entirety of exons 5 and 6), and the remaining regions contain several frameshift indels and nonsense mutations (Figure 8). The gene is otherwise well-conserved in bovidae (domestic goat, sheep, Tibetan antelope, and cow), suggesting it was functional in the cetartiodactyla ancestor (superorder of odontoceti and bovidae) but was inactivated and then allowed to erode in the evolutionary past of toothed whales. *GPRC6A* erosion in toothed whales is especially striking when considering that toothed whales exhibit 7-25 times greater testis-to-body mass ratio relative to other mammals [54].

Discussion

In this work, we devised a novel approach to automate the survey of genomes for instances of coding gene sequence erosion. Knowledge of gene losses has far-reaching applications in studying the morphological and physiological adaptations of different species [10], in discovering alternative molecular pathways for critical genes [2], in medical genetics [13], and even in animal conservation programs [12]. We designed our approach for low false positive rates in order to be practically useful for the experimentalists tackling these research efforts. In particular, our conservative approach addresses a number of challenges that hindered previous automated methods.

Pre-existing genome annotation methods [17] were designed to identify intact genes and were not tuned to distinguish sequence erosion due to a bonafide gene loss from confounds due to alignment, assembly, and other technical artifacts. In contrast, our mapping procedure using whole-genome pairwise alignments is syntenic-aware, is robust to sequencing gaps and other artifacts, and produces strictly one-to-one ortholog assignments that are not confounded by gene duplication events (paralogs) across species. For each genome, we used the mapped gene annotations to scan for signatures of gene erosion (large deletions and non-synonymous substitutions) [10] while also carefully accounting for sequencing gaps and other assembly artifacts. To enrich for high-confidence gene erosion predictions, we used a Mahalanobis distance-based metric to refine the candidates to just those orthologs exhibiting extreme amounts of deletion and non-synonymous substitution relative to other genes in that species. We designated these orthologs as being likely

eroded in sequence and function in the species represented by each examined assembly. To even further minimize erroneous predictions, we distilled out just those candidates that are considered likely eroded in at least two closely related species but sequence-conserved in at least one other outgroup besides human. Despite the phylogenetic filter reducing sensitivity for detecting gene losses, we discovered 2,192 orthologs (hiconfErosions) across 52 species that are severely eroded. Many of these hiconfErosions are novel candidates with evolutionary and disease significance (Table 1 and Table 2). Evaluated against UniProt expression and proteomic observations in mouse and rat (see Methods), this final candidate set has an estimated false prediction rate of less than 10%.

While a majority of the hiconfErosions were predicted to have benign consequences, dozens involved loci associated with lethal developmental outcomes in mouse knockout models (Table 1), human disease (Table 2), and/or intolerance to variation in human populations (Table 3). The evolutionary mechanisms contributing to these functionally significant gene losses were likely diverse. Some may have resulted from regressive evolution (i.e., degeneration of formerly useful anatomical and physiological functions), such as the degraded *OPTC* vision gene [55] in nocturnal microbats or the *CRYGB* erosion in subterranean cape golden moles. Other losses may actually have contributed to species adaptation, such as erosion of the epithelium gene *GSDMA* [56, 57] in cetaceans which have skin adaptations to life in the water [10].

Species with hiconfErosion that tolerate functional changes known to cause disease in humans provide intriguing potential for medical genetics [13]. Some of these “knockout experiments of nature” may tolerate and even derive selective advantage from gene loss(es)

because of factors specific to their habitat niche. Others, however, may have evolved compensatory suppressor mutations to genetically negate otherwise deleterious consequences. It is these latter cases that could be harnessed for treating human congenital disorders [13]. For instance, despite the erosion of *GPRC6A* in toothed whales, these magnificent creatures have relatively massive testes for their size – the categorically opposite phenotype of when this gene is inactivated in mice and humans. We posit that further examination of odontoceti could reveal complementary testis development pathways that might inform therapy for male hypogonadism. More broadly, extension of this gene loss survey to additional or newly sequenced species across the metazoan tree of life will likely reveal more such evolutionary curiosities with great clinical potential.

To that end, we emphasize the ease with which genomes can be added to analyses using our approach. Each additional genome would independently be aligned to the reference assembly – without the high computational cost of realigning all genomes, as is necessary for methods requiring multiple-sequence alignments. Thus, our approach is easily scalable for the ever-increasing number of new species being sequenced [15, 58].

Methods

Genome assemblies, reference genome and protein-coding gene annotations

We used genome assemblies of 59 mammalian species (listed in Supplementary Table 1) in this study, with human genome assembly GRCh38/hg38 as reference. From Ensembl biomart release 86, we downloaded lists of gene identifiers and counted how many unique ‘protein_coding’ gene identifiers are found. Of the 59 mammalian species, 34 species had gene annotation in Ensembl. For our reference gene set, we started with 22,072 unique human protein-coding genes annotated in Ensembl and excluded gene annotations from un-placed and un-localized scaffolds, leaving 19,729 protein-coding genes. We then obtained coordinates for segmental duplications (genomic fragments larger than 1kb with more than 90% sequence identity to other genomic fragments) from the UCSC genome browser and excluded from our analysis genes with 10% or more overlap with any segmentally duplicated fragment. We also restricted the reference gene set to genes having “complete” ENSEMBL transcripts. These filters generated our reference gene set of 17,860 unique human protein-coding genes for mapping across mammalian genomes. We collected assembly N50 measures for each genome from NCBI: <https://www.ncbi.nlm.nih.gov/assembly>.

Mapping orthologs across mammalian genomes

We used Jim Kent’s BLASTZ-based, whole-genome pairwise alignment chains [22] to map – when possible with high confidence – each reference gene to a single orthologous position in each of the 58 query species (Figure 2A). First, we assigned every base in the canonical transcript of the reference gene to the highest-scoring chain (in terms of UCSC chain alignment scores) that contains the base in its alignment (Figure 2B1-B2). The chain to which most number of bases get assigned was picked as the *best chain* (C_b) for that gene (Figure 2B3). Since chains are preferentially ordered based on the alignment score before assigning bases (Figure 2B1), the best chain was also the chain with highest synteny for those aligning bases, and was treated as the most likely ortholog for that gene. Next, this procedure was repeated after removing the *best chain* (C_b) from consideration to pick another chain as the *second-best chain* (C_{sb}) for the same gene (Figure 2B4). This chain C_{sb} was treated as containing a likely paralog of the gene. To ensure that the orthologous chain was easily differentiable from paralogs, we required the *second-best chain* to have an alignment score at least 20 times lower than that of the best chain. To also ensure high synteny, we required the number of bases in the aligning blocks of the best chain be at least 20 times greater than the number of bases in the gene itself, i.e. $gene-in-syteny \geq 20$, where $gene-in-syteny = \text{length of } C_b / \text{length of gene}$. We also required unique mapping of coordinates between reference and query genomes, such that if two or more reference genes were mapped to the same location, all overlapping mappings were discarded.

In each species, for every gene to which an orthologous chain was assigned, we used

the alignments derived from the chain to determine the orthologous amino acid sequence for every transcript of that gene. Because exon boundaries sometimes shift for evolutionary or alignment reasons, we excluded the first and last two amino acid positions in each exon for all downstream analyses [59]. Gaps in chain alignments found in regions containing an assembly gap in the query species were masked out to avoid confusing them with true deletions in the query genome.

Outlier detection model

To identify the genes orthologs that have a surprisingly high number of non-synonymous coding substitutions and deletions in each of the 58 query species, we used an outlier detection approach. We first used the derived amino acid sequence of each transcript for every gene in a query species, as described previously, to compute a feature vector $\vec{x} = (x_1, x_2)$, where x_1 is the fraction of the transcript gene deleted in the query species and x_2 is the fraction of non-synonymous coding substitutions in the remaining intact portion of the gene transcript compared to the orthologous gene transcript sequence in the reference species (human).

We then used $\vec{x}$ across all genes in the query species to compute the mean vector, $\vec{\mu}$, and a covariance matrix, S . Mahalanobis distance (MD) [24, 25] measures how “far” a given gene with feature vector, $\vec{x}$, is from the distribution of all genes in the query species, as follows:

$$MD = \sqrt{(\vec{x} - \vec{\mu})^T S^{-1} (\vec{x} - \vec{\mu})} \quad (1)$$

Since most genes exhibit relatively few non-synonymous substitutions and deletions (as placental mammals are relatively closely related, see Figure 3B), a very large value of MD implies high extensive erosion of the gene. Conversely, a very small value of MD implies that the gene is well- conserved. Since the coding sequences are expected to diverge more with increasing phylogenetic distance between reference and query species, we computed a different distribution ($\vec{\mu}$ and S) for each query species. If all transcripts of an ortholog had MD greater than an absolute threshold ($MD > 15$), we marked the gene an outlier (✓) since it is likely sequence-eroded and nonfunctional in the species represented by the query genome. If all transcripts of an ortholog exhibited $MD \leq 5$, we regarded the gene sequence-conserved and functional (✓) in the species represented by the query genome. All other orthologs that were not categorized into either of the former two groups were marked with undetermined (?).

Phylogenetic filtering

We used the phylogenetic tree of the 59 mammalian species (including human; Supplementary Figure 1) to identify which of the outlier orthologs would be considered to have undergone hiconfErosion (Figure 3D illustrates this procedure). For each gene, we recorded which species exhibit a sequence-conserved ortholog (based on the results of the

outlier detection model described above; also see Figure 3B). Since the gene under consideration is already annotated to be functionally present in hg38 by Ensembl, we inferred by parsimony that the last common ancestor of human and of all species containing the conserved gene likely had a functional copy of this gene with a sequence similar to that of its extant functional orthologs. Finally, we search all subtrees starting from the earliest ancestral species marked in this fashion with gene conservation to identify specific groups where hiconfErosion has occurred. A group in which hiconfErosion has occurred is one in which there are at least two species with the gene marked as likely eroded (all transcripts with $MD > 15$) and no leaf-node species with the gene marked as sequence-conserved ($MD \leq 5$).

Curation of Uniprot data and details of FPR (False Positive Rate) estimation in mouse and rat

To measure the rate of false positives in hiconfErosion predictions for mouse and rat, we obtained experimentally curated coding gene models downloaded as protein sequences with ‘protein level’ or ‘transcript level’ evidence from Uniprot [18] (release-2017_01; ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/complete).

We found 16,385 and 7,591 gene models in mouse and rat, respectively. We then mapped those amino acid sequences to their respective genomes (mouse sequences to mm10; rat sequences to rn6) using protein BLAT [60] with default settings (minimum mapping score

30; minimum sequence identity 25; maximum intron size 750,000). After mapping these protein sequences, we computed what proportion of orthologs predicted to be affected by hiconfErosion is contradicted by empirical evidence. Specifically, for a given species and its genome, any predicted hiconfErosion-affected orthologs that deviate by more than 20% in sequence length, share less than 90% identity, or disagree with the locus of the relevant Uniprot sequence were considered false predictions. Lastly, since we had 17,860 tested proteins in the reference gene set, we conservatively scaled up the number of conflicts in mouse and rat by a factor of 1.09 ($=17,860/16,385$) and 2.35 ($=17,860/7,591$), respectively, to estimate our False Positive Rate (FPR) while accounting for potentially missing Uniprot entries.

Curation of lethal loss of function genes in mouse models

We used gene function annotations from the MGI Phenotype Ontology (Mouse Genome Informatics, <http://www.informatics.jax.org/>) to derive a subset of genes whose disruption leads to severe developmental perturbations and lethal consequences. The MGI Phenotype Ontology catalogs spontaneous, induced, and genetically-engineered mouse mutations and their associated phenotypes. Ontology data (containing 8,949 phenotypic terms; v6.04) were lifted over from mouse to human, resulting in 609,253 (canonicalized) gene-phenotype associations. We selected all 50 unique phenotypic terms with the string “lethal” in their description (e.g., MP:0008762 *embryonic lethality*; MP:0008569 *lethality at weaning*; MP:0006206 *embryonic lethality between somite formation and em-*

bryo turning), identifying a total of 3,617 genes associated with causing early lethality when disrupted.

Analysis of hiconfErosion genes implicated in HGMD

We obtained an access to a professional version (2016.2) of the Human Gene Mutation Database [33] containing 165,939 entries of known disease-causing gene mutations (referenced to GRCh38 assembly) and their respective phenotypes. We then obtained the list of 4,014 monogenic disease genes (i.e. genes with at least one “DM” mutation) from this database, of which 3,711 genes were also present in our reference gene set.

Analysis of genic variation intolerability with RVIS score

Residual Variation Intolerance Score (RVIS) is a unified genome-wide metric that quantifies the extent to which genes tolerate coding mutation load (i.e., the higher the score, the more a gene can tolerate common genetic variation without negative consequences). We downloaded the precomputed percentiles of RVIS scores (inferred based on Exome Aggregation Consortium variants with Minor Allele Frequency of 0.05%) available for 17,276 genes (http://genic-intolerance.org/data/RVIS_Unpublished_ExAC_May2015.txt). We recorded the hiconfErosion genes with low ($< 20\%$) RVIS score percentiles, which are relatively intolerant to functional variation in humans.

Statistical significance of depletion

To estimate statistical significance of depletion [61] for the hiconfErosion group of genes relative to categories such as *mouse lethal genes* or *HGMD disease genes*, we used the Fisher exact test to compute a one-sided p-value (using R):

$$p_{depletion} = \frac{\binom{\Pi}{L_{\Pi}} \binom{N - \Pi}{L - L_{\Pi}}}{\binom{N}{L}} \quad (2)$$

In the above equation, N and L denote the total number of genes and the hiconfErosion subset of genes in the analysis, respectively. Similarly, Π and L_{Π} denotes the subset of N and L associated with a gene category such as HGMD disease genes or mouse lethal genes.

We also compute a ‘fold change’ for depletion as $1/Fold_{enrichment}$:

$$Fold_{depletion} = \frac{1}{Fold_{enrichment}} = \frac{\Pi * L}{N * L_{\Pi}} \quad (3)$$

We calculate the two-tailed p-value using R to test how different were the mean RVIS scores of the 412 genes implicated in hiconfErosions compared to the reference gene set.

References

1. O’Leary, M. A. & Kaufman, S. MorphoBank: phylophenomics in the “cloud”. *Cladistics* **27**, 529–537 (2011).
2. Albalat, R. & Cañestro, C. Evolution by gene loss. *Nature Reviews Genetics* **17**, 379–391 (2016).
3. Hiller, M. *et al.* A “forward genomics” approach links genotype to phenotype using independent phenotypic losses among related species. *Cell Reports* **2**, 817–823 (2012).
4. Marcovitz, A., Jia, R. & Bejerano, G. “Reverse Genomics” Predicts Function of Human Conserved Noncoding Elements. *Molecular Biology and Evolution* **33**, 1358–1369 (2016).
5. Partha, R. *et al.* Subterranean mammals show convergent regression in ocular genes and enhancers, along with adaptation to tunneling. *eLife* **6**, e25884 (2017).
6. Emerling, C. A. & Springer, M. S. Eyes underground: regression of visual protein networks in subterranean mammals. *Molecular Phylogenetics and Evolution* **78**, 260–270 (2014).
7. Jiang, P. *et al.* Major taste loss in carnivorous mammals. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 4956–4961 (2012).
8. McGowen, M. R., Gatesy, J. & Wildman, D. E. Molecular evolution tracks macroevolutionary transitions in Cetacea. *Trends in Ecology & Evolution* **29**, 336–346 (2014).

9. Strasser, B., Mlitz, V., Fischer, H., Tschachler, E. & Eckhart, L. Comparative genomics reveals conservation of filaggrin and loss of caspase-14 in dolphins. *Experimental Dermatology* **24**, 365–369 (2015).
10. Sharma, V. *et al.* A genomics approach reveals insights into the importance of gene losses for mammalian adaptations. *Nature Communications* **9**, 1215 (2018).
11. Braun, B. A., Marcovitz, A., Camp, J. G., Jia, R. & Bejerano, G. Mx1 and Mx2 key antiviral proteins are surprisingly lost in toothed whales. *Proceedings of the National Academy of Sciences of the United States of America* **112**, 8036–8040 (2015).
12. Meyer, W. K. *et al.* Ancient convergent losses of Paraoxonase 1 yield potential risks for modern marine mammals. *Science* **361**, 591–594 (2018).
13. Emerling, C. A., Widjaja, A. D., Nguyen, N. N. & Springer, M. S. Their loss is our gain: regressive evolution in vertebrates provides genomic models for uncovering human disease loci. *Journal of Medical Genetics* **54**, 787–794 (2017).
14. Harrow, J. *et al.* GENCODE: producing a reference annotation for ENCODE. *Genome biology* **7**, S4 (2006).
15. Of Scientists, G. I. C. Genome 10K: a proposal to obtain whole-genome sequence for 10 000 vertebrate species. *Journal of Heredity* **100**, 659–674 (2009).
16. Huelsmann, M. *et al.* Genes lost during the transition from land to water in cetaceans highlight genomic changes associated with aquatic adaptations. *Science Advances* **5**, eaaw6671 (2019).
17. Zerbino, D. R. *et al.* Ensembl 2018. *Nucleic Acids Research* **46**, D754–D761 (2018).

18. Bateman, A. *et al.* UniProt: the universal protein knowledgebase. *Nucleic Acids Research* **45**, D158–D169 (2017).
19. Zhang, G. *et al.* Comparative genomics reveals insights into avian genome evolution and adaptation. *Science* **346**, 1311–1320 (2014).
20. Venkatesh, B. Evolution and diversity of fish genomes. *Current Opinion in Genetics & Development* **13**, 588–592 (2003).
21. Clark, A. G. *et al.* Evolution of genes and genomes on the Drosophila phylogeny. *Nature* **450**, 203–218 (2007).
22. Kent, W. J., Baertsch, R., Hinrichs, A., Miller, W. & Haussler, D. Evolution's cauldron: Duplication, deletion, and rearrangement in the mouse and human genomes. *Proceedings of the National Academy of Sciences* **100**, 11484–11489 (2003).
23. Derrien, T. *et al.* Revisiting the missing protein-coding gene catalog of the domestic dog. *BMC genomics* **10**, 62 (2009).
24. Mahalanobis, P. C. *On the generalized distance in statistics* in (1936).
25. De Maesschalck, R., Jouan-Rimbaud, D. & Massart, D. L. The mahalanobis distance. *Chemometrics and intelligent laboratory systems* **50**, 1–18 (2000).
26. Hogarth, C. A., Roy, A. & Ebert, D. L. Genomic evidence for the absence of a functional cholesteryl ester transfer protein gene in mice and rats. *Comparative Biochemistry and Physiology. Part B, Biochemistry & Molecular Biology* **135**, 219–229 (2003).

27. Abd El-Aziz, M. M. *et al.* EYS, encoding an ortholog of *Drosophila* spacemaker, is mutated in autosomal recessive retinitis pigmentosa. *Nature Genetics* **40**, 1285–1287 (2008).
28. Bowling, S. *et al.* P53 and mTOR signalling determine fitness selection through cell competition during early mouse embryonic development. *Nature Communications* **9**, 1763 (2018).
29. Su, R.-W., Strug, M. R., Jeong, J.-W., Miele, L. & Fazleabas, A. T. Aberrant activation of canonical Notch1 signaling in the mouse uterus decreases progesterone receptor by hypermethylation and leads to infertility. *Proceedings of the National Academy of Sciences of the United States of America* **113**, 2300–2305 (2016).
30. Hargrave, M. *et al.* Expression of the Sox11 gene in mouse embryos suggests roles in neuronal maturation and epithelio-mesenchymal induction. *Developmental Dynamics: An Official Publication of the American Association of Anatomists* **210**, 79–86 (1997).
31. Greibe, E., Fedosov, S. & Nexø, E. The cobalamin-binding protein in zebrafish is an intermediate between the three cobalamin-binding proteins in human. *PLoS One* **7**, e35660 (2012).
32. Goudet, G., Mugnier, S., Callebaut, I. & Monget, P. Phylogenetic analysis and identification of pseudogenes reveal a progressive loss of zona pellucida genes during evolution of vertebrates. *Biology of reproduction* **78**, 796–806 (2008).

33. Stenson, P. D. *et al.* The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. *Human Genetics* **136**, 665–677 (2017).
34. Petrovski, S. *et al.* The Intolerance of Regulatory Sequence to Genetic Variation Predicts Gene Dosage Sensitivity. *PLOS Genetics* **11**, e1005492 (2015).
35. Phillips, C. J. *et al.* Dietary and flight energetic adaptations in a salivary gland transcriptome of an insectivorous bat. *PloS one* **9**, e83512 (2014).
36. Voigt, C. C., Zubaid, A., Kunz, T. H. & Kingston, T. Sources of assimilated proteins in Old and New World phytophagous bats. *Biotropica* **43**, 108–113 (2011).
37. Singaraja, R. R. *et al.* Identification of four novel genes contributing to familial elevated plasma HDL cholesterol in humans. *Journal of lipid research* **55**, 1693–1701 (2014).
38. Widmaier, E. *et al.* High plasma cholesterol, but low triglycerides and plaque-free arteries, in Mexican free-tailed bats. *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology* **271**, R1101–R1106 (1996).
39. Vogel, P. *et al.* Situs inversus in *Dpcd/Poll^{-/-}*, *Nme7^{-/-}*, and *Pkd111^{-/-}* mice. *Veterinary Pathology* **47**, 120–131 (2010).
40. Field, S. *et al.* *Pkd111* establishes left-right asymmetry and physically interacts with *Pkd2*. *Development (Cambridge, England)* **138**, 1131–1142 (2011).
41. Grimes, D. T. *et al.* Genetic Analysis Reveals a Hierarchy of Interactions between Polycystin-Encoding Genes and Genes Controlling Cilia Function during Left-Right Determination. *PLoS genetics* **12**, e1006070 (2016).

42. Kamura, K. *et al.* Pkd111 complexes with Pkd2 on motile cilia and functions to establish the left-right axis. *Development (Cambridge, England)* **138**, 1121–1129 (2011).
43. Vetrini, F. *et al.* Bi-allelic Mutations in PKD1L1 Are Associated with Laterality Defects in Humans. *The American Journal of Human Genetics* **99**, 886–893 (2016).
44. Perles, Z. *et al.* A human laterality disorder caused by a homozygous deleterious mutation in MMP21. *Journal of Medical Genetics* **52**, 840–847 (2015).
45. Akawi, N. *et al.* Discovery of four recessive developmental disorders using probabilistic genotype and phenotype matching among 4,125 families. *Nature Genetics* **47**, 1363–1369 (2015).
46. Li, Y. *et al.* Global genetic analysis in mice unveils central role for cilia in congenital heart disease. *Nature* **521**, 520–524 (2015).
47. Eppig, J. T. *et al.* Mouse Genome Informatics (MGI): Resources for Mining Mouse Genetic, Genomic, and Biological Data in Support of Primary and Translational Research. *Methods in Molecular Biology (Clifton, N.J.)* **1488**, 47–73 (2017).
48. Ho, S. Y. *et al.* *Congenital Heart Malformations in Mammals: An Illustrated Text* (World Scientific, 2000).
49. Clemmensen, C., Smajilovic, S., Wellendorph, P. & BrÄuner-Osborne, H. The GPCR, class C, group 6, subtype A (GPRC6A) receptor: from cloning to physiological function. *British Journal of Pharmacology* **171**, 1129–1141 (2014).
50. Pi, M. & Quarles, L. D. Multiligand Specificity and Wide Tissue Expression of GPRC6A Reveals New Endocrine Networks. *Endocrinology* **153**, 2062–2069 (2012).

51. Pi, M. & Quarles, L. D. GPRC6A regulates prostate cancer progression. *The Prostate* **72**, 399–409 (2012).
52. Pi, M. *et al.* GPRC6A Null Mice Exhibit Osteopenia, Feminization and Metabolic Syndrome. *PLoS ONE* **3**. doi:10.1371/journal.pone.0003858 (2008).
53. Ye, R., Pi, M., Cox, J. V., Nishimoto, S. K. & Quarles, L. D. CRISPR/Cas9 targeting of GPRC6A suppresses prostate cancer tumorigenesis in a human xenograft model. *Journal of Experimental & Clinical Cancer Research : CR* **36**. doi:10.1186/s13046-017-0561-x (2017).
54. Kenagy, G. J. & Trombulak, S. C. Size and Function of Mammalian Testes in Relation to Body Size. *Journal of Mammalogy* **67**, 1–22 (1986).
55. Acharya, M. *et al.* Evaluation of the OPTC gene in primary open angle glaucoma: functional significance of a silent change. *BMC molecular biology* **8**, 21 (2007).
56. Saeki, N., Kuwahara, Y., Sasaki, H., Satoh, H. & Shiroishi, T. Gasdermin (Gsdm) localizing to mouse Chromosome 11 is predominantly expressed in upper gastrointestinal tract but significantly suppressed in human gastric cancer cells. *Mammalian Genome: Official Journal of the International Mammalian Genome Society* **11**, 718–724 (2000).
57. Saeki, N. *et al.* GASDERMIN, suppressed frequently in gastric cancer, is a target of LMO1 in TGF-beta-dependent apoptotic signalling. *Oncogene* **26**, 6488–6498 (2007).
58. Lewin, H. A. *et al.* Earth BioGenome Project: Sequencing life for the future of life. *Proceedings of the National Academy of Sciences* **115**, 4325–4333 (2018).

59. Marcovitz, A. *et al.* A functional enrichment test for molecular convergent evolution finds a clear protein-coding signal in echolocating bats and whales. *Proceedings of the National Academy of Sciences*. doi:10.1073/pnas.1818532116 (2019).
60. Kent, W. J. BLAT The BLAST-Like Alignment Tool. *Genome Research* **12**, 656–664 (2002).
61. Rivals, I., Personnaz, L., Taing, L. & Potier, M.-C. Enrichment or depletion of a GO category within a class of genes: which test? *Bioinformatics* **23**, 401–407 (2006).

Acknowledgements We thank the members of the Bejerano Laboratory, particularly M. J. Berger, K. Jagadeesh, J. Birgmeier, B. Yoo and others for technical advice and helpful discussions. We thank Hiram Clawson and the UCSC Genome Browser team for providing us the mammalian alignment chains and multiple sequence alignments and technical advice. We thank D. Cooper and P. Stenson for Human Gene Mutation Database (HGMD) access.

Competing Interests The authors declare that they have no competing financial interests.

Tables and Figures

Gene	Mouse knockout phenotype	Species with gene hiconfErosion
CA5B	preweaning lethality, incomplete penetrance	killer whale, dolphin
CCDC94	embryonic lethality	black flying-fox, megabat, microbat, big brown bat David's myotis bat
DMBT1	embryonic lethality	^(I) naked mole-rat, brush-tailed rat, guinea pig, chinchilla ^(II) goat, sheep, Tibetan antelope, cow
HAP1	lethality during fetal growth through weaning	naked mole-rat, brush-tailed rat, guinea pig, chinchilla
MESP1	embryonic lethality	naked mole-rat, brush-tailed rat, guinea pig, chinchilla
MLKL	embryonic lethality	cat, dog, ferret, panda, pacific walrus, Weddell seal, horse, white rhinoceros
PKD1L1	lethality throughout fetal growth and development	alpaca, Bactrian camel, killer whale, dolphin, goat, sheep, Tibetan antelope, cow
RESP18	prenatal lethality	microbat, David's myotis bat
SERPINA10	lethality, incomplete penetrance	^(I) cat, dog, ferret, panda, pacific walrus, Weddell seal ^(II) rabbit, pika
ZNF565	embryonic lethality	pacific walrus, Weddell seal

^(I), ^(II) Independent hiconfErosion loss events: ^(I) species belonging to group I ^(II) species belonging to group II

Table 1

Genes implicated in mammalian hiconfErosions that result in lethality when inactivated in mouse models.

Gene	Associated human monogenic disease	Species with gene hiconfErosion
<i>A2ML1</i>	<ul style="list-style-type: none"> Noonan syndrome Otitis media (susceptibility to) 	naked mole-rat, brush-tailed rat, guinea pig, chinchilla
<i>CETP</i>	<ul style="list-style-type: none"> Higher/Lower HDL cholesterol level Hyperalphalipoproteinaemia Cholesterol ester transfer protein deficiency 	^(I) Bactrian camel, alpaca ^(II) goat, sheep, Tibetan antelope, cow
<i>CRYBB3</i>	<ul style="list-style-type: none"> Cataract 	microbat, big brown bat, David's myotis bat
<i>CRYGB</i>	<ul style="list-style-type: none"> Congenital cataract 	cape golden mole, tenrec
<i>EYS</i>	<ul style="list-style-type: none"> Retinitis pigmentosa Retinal dystrophy Peripheral dystrophy Usher syndrome 	mouse, rat, Chinese hamster, golden hamster, prairie vole
<i>FCN2</i>	<ul style="list-style-type: none"> FCN2 deficiency 	goat, sheep, Tibetan antelope, cow
<i>GCKR</i>	<ul style="list-style-type: none"> Elevated HDL-cholesterol Hypertriglyceridaemia 	microbat, big brown bat, David's myotis bat
<i>HMGCS2</i>	<ul style="list-style-type: none"> Mitochondrial HMG-CoA synthase deficiency 	black flying-fox, megabat
<i>HSD3B2</i>	<ul style="list-style-type: none"> Adrenal hyperplasia Hypospadias, non-syndromic Idiopathic hypospadias Pseudohermaphroditism 	goat, sheep, Tibetan antelope, cow
<i>KLKB1</i>	<ul style="list-style-type: none"> Prekallikrein deficiency 	dolphin, killer whale
<i>KRT3</i>	<ul style="list-style-type: none"> Corneal dystrophy, Meesmann 	mouse, rat
<i>MESP1</i>	<ul style="list-style-type: none"> Tetralogy of Fallot Ventricular septal defect 	naked mole-rat, brush-tailed rat, guinea pig, chinchilla
<i>MMP21</i>	<ul style="list-style-type: none"> Heterotaxy 	pig, dolphin, killer whale, Bactrian camel, alpaca, goat, sheep, Tibetan antelope, cow
<i>MYD88</i>	<ul style="list-style-type: none"> MYD88 deficiency 	brush-tailed rat, chinchilla
<i>OPTC</i>	<ul style="list-style-type: none"> Glaucoma, primary open angle 	cape golden mole, tenrec, microbat, big brown bat, David's myotis bat
<i>PCSK9</i>	<ul style="list-style-type: none"> Hypercholesterolaemia Low/High LDL cholesterol 	^(I) cat, dog, ferret, panda, pacific walrus, Weddell seal ^(II) goat, sheep, Tibetan antelope, cow
<i>PLA2G2A</i>	<ul style="list-style-type: none"> Colorectal cancer 	dolphin, killer whale
<i>RIPPLY1</i>	<ul style="list-style-type: none"> Intellectual disability 	black flying-fox, megabat
<i>SFXN4</i>	<ul style="list-style-type: none"> Colorectal cancer Mitochondriopathy and macrocytic anaemia 	^(I) Chinese hamster, golden hamster, prairie vole ^(II) naked mole-rat, brush-tailed rat, guinea pig, chinchilla
<i>STAP1</i>	<ul style="list-style-type: none"> Hypercholesterolaemia 	dolphin, killer whale
<i>TCN1</i>	<ul style="list-style-type: none"> Transcobalamin I deficiency 	mouse, rat
<i>TNFRSF6B</i>	<ul style="list-style-type: none"> Systemic lupus erythematosus 	mouse, rat, Chinese hamster, golden hamster, prairie vole
<i>UMOD</i>	<ul style="list-style-type: none"> Chronic kidney disease Congenital anomalies of the kidney and urinary tract FJHN/MCKD syndrome Glomerulocystic kidney disease Uromodulin-associated kidney disease Hyperuricaemic nephropathy 	microbat, big brown bat, David's myotis bat, black flying-fox, megabat
<i>XAF1</i>	<ul style="list-style-type: none"> Multiple sessile serrated adenoma 	cat, dog, ferret, panda, pacific walrus, Weddell seal
<i>ZNF543</i>	<ul style="list-style-type: none"> IgA nephropathy 	cat, dog, ferret, panda, pacific walrus, Weddell seal
<i>ZNF589</i>	<ul style="list-style-type: none"> Intellectual disability 	ferret, panda, pacific walrus, Weddell seal
<i>ZP1</i>	<ul style="list-style-type: none"> Infertility 	pig, dolphin, killer whale, Bactrian camel, alpaca, goat, sheep, Tibetan antelope, cow

^{(I), (II)} Independent hiconfErosion loss events: ^(I) species belonging to group I ^(II) species belonging to group II

Table 2

Genes implicated in mammalian hiconfErosions that harbor monogenic disease variants in human patients (disease mutations in HGMD [33]).

Gene	RVIS Score Percentile	Species with gene hiconfErosion
<i>DDX60</i>	4.36%	Tibetan antelope, Sheep, Domestic goat, Cow
<i>ADAM30</i>	8.36%	Guinea Pig, Brush-tailed rat, Naked mole-rat, Chinchilla
<i>CYP27C1</i>	10.99%	David's myotis (bat), Microbat, Big brown bat
<i>PCDHA7</i>	11.82%	Panda, Weddell seal, Ferret, Pacific walrus
<i>UBC</i>	12.52%	^(I) Hedgehog, Star-nosed mole, Shrew ^(II) Naked mole-rat, Guinea Pig, Chinchilla, Brush-tailed rat ^(III) Ferret, Panda, Pacific walrus, Weddell seal ^(IV) Alpaca, Bactrian camel
<i>UMOD</i>	13.17%	Black flying-fox, Megabat, David's myotis (bat), Microbat, Big brown bat
<i>SFXN4</i>	13.78%	^(I) Chinese Hamster, Golden hamster, Prairie vole ^(II) Naked mole-rat, Guinea Pig, Chinchilla, Brush-tailed rat
<i>ZYG11B</i>	14.17%	Rhesus macaque, Crab-eating macaque
<i>MCF2L2</i>	15.17%	Golden hamster, Mouse, Rat, Lesser Egyptian jerboa, Prairie vole, Chinese Hamster
<i>TRMT2B</i>	16.72%	Guinea Pig, Chinchilla, Naked mole-rat, Brush-tailed rat
<i>SLC6A16</i>	17.45%	Black flying-fox, Megabat
<i>DCBLD1</i>	18.09%	Lesser Egyptian jerboa, Prairie vole, Golden hamster, Chinese Hamster, Mouse, Rat
<i>MCHR2</i>	18.15%	Naked mole-rat, Guinea Pig, Chinchilla, Brush-tailed rat

^(I), ^(II), ^(III), ^(IV) Independent hiconfErosion loss events:
^(I) species belonging to group I ^(II) species belonging to group II ^(III) species belonging to group III ^(IV) species belonging to group IV

Table 3

Genes implicated in mammalian hiconfErosions that have low RVIS (Residual Variation Intolerance Score) score percentile. The RVIS scores were inferred based on Exome Aggregation Consortium variants with Minor Allele Frequency of 0.05%.

Figure 1

Incomplete coding gene annotations in genomic databases. A) Of the 59 mammalian genomes used in this study, 34 have Ensembl (release 86) gene annotations. The number of annotations per genome appears to artifactually increase with assembly quality (as measured by N50 contig length), with a correlation coefficient of 0.67, saturating at some point beyond N50 $\approx 5e+5$ for highly contiguous assemblies (such as human and mouse). B) Top: *PAX6* Ensembl gene model absence (gray highlight) in the dog genome assembly due to a sequencing gap. The genome browser view shows the minus strand. Bottom: The orthologous position of *PAX6* (and flanking genes) in the human genome. C) Top: *TP53* Ensembl gene model absence (gray highlight) in the alpaca genome within a conserved region. The genome browser view shows the minus strand. Bottom: The orthologous position of *TP53* (and flanking genes) in the human genome.

Figure 3

A pipeline to call hiconfErosions in genes across mammalian lineages. A) Using pairwise chained alignments, we determined an orthologous location for human coding genes in each genome. Then we computed the amount of coding deletion and amino acid substitution to find, for each each genome, the outlier genes among all the genes that were mapped (i.e., likely eroding genes). B) Scatter plot showing the distribution of the mouse (mm10) genes along the two features (% coding deletion and % coding substitution) for calculating Mahalanobis Distance (MD). The histogram along each axis shows the distribution of points when projected on the respective axis. (Per-species) likely eroded genes (in red, such as known losses of *Cetp* and *Eys*) have much higher values of % coding deletions and % coding substitutions compared to most other genes (in black). C) Histogram of the MD values for mouse genes along with the threshold values (dashed vertical lines) for conserved ($\checkmark$, $MD \leq 5$), undetermined ($?$), and likely eroded ($\dagger$, $MD > 15$) genes. D) An example illustration of applying phylogenetic filtering. We call hiconfErosion ($\times$, e.g. in mouse and squirrel) if a gene is marked likely eroded ($MD > 15$) in more than one species in a clade that is flanked by at least one group in which the gene is conserved ($\checkmark$). An undetermined gene in an individual species ($?$, e.g. in rat) is inferred as eroded if it resides in a clade with hiconfErosion. The likely eroded ($\dagger$) gene in cow, however, is not considered a hiconfErosion since the finding is not supported by another adjacent species.

Figure 4

Distribution of hiconfErosions across species and estimation of False Positive Rate (FPR). A) Number of hiconfErosions predicted in different species. B) If hiconfErosions are directly inferred by applying the phylogenetic filtering (Figure 3D) to missing gene annotations in Ensembl (instead of eroded genes from Figure 3C), the estimated FPR (red bars) in mouse and rat is significantly higher than in our set of predicted hiconfErosions.

Figure 5

Selection and variation in mammalian genes affected by hiconfErosion. A) hiconfErosion genes are depleted from genes whose inactivation in mouse models leads to lethal consequences in early development. B) hiconfErosion genes are depleted from genes with known Mendelian disease mutations (HGMD genes). C) We compare the genome-wide distribution of Residual Variation Intolerance Score (RVIS) percentiles (derived from ExAC) for common variation at MAF of 0.05 to the RVIS percentiles of genes affected by hiconfErosions. By and large, hiconfErosion genes are more tolerant of common functional variation in humans, suggesting that selection over coding genes is similar across mammals. Several genes at the lower tail of the RVIS scores are highlighted (see Table 3 for details).

Figure 6

***GCKR* hiconfErosion in microbats.** A genome browser view of *GCKR* (Glucokinase regulatory protein) on the human genome (GRCh38) and a mammalian multiple alignment highlight coding lesions in microbat. Positions of deleted genomic fragments (black) and non-synonymous substitutions (green) are highlighted for David's myotis and big brown bat (both of *microchiroptera*).

Figure 8

***GPRC6A* hiconfErosion in toothed whales.** A genome browser view of *GPRC6A* (on the human genome, GRCh38) and a mammalian multiple alignment highlight coding lesions in dolphin and killer whale. Positions of deleted fragments (black) and non-synonymous substitutions (green) are shown for both species. *GPRC6A* is a key hormonal and metabolism regulation protein, and the knockout of *GPRC6A* in male mice results in over 6-fold decrease of testosterone levels and reduced size and weight of the testis [52].

SI Tables and Figures

bioRxiv preprint doi: <https://doi.org/10.1101/819169>; this version posted October 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

	Scientific Name	Common Name	UCSC Assembly ID
1	<i>Ailuropoda melanoleuca</i>	Panda	ailMel1
2	<i>Bos taurus</i>	Cow	bosTau8
3	<i>Callithrix jacchus</i>	Marmoset	calJac3
4	<i>Camelus ferus</i>	Bactrian camel	camFer1
5	<i>Canis lupus familiaris</i>	Dog	canFam3
6	<i>Capra aegagrus hircus</i>	Domestic goat	capHir1
7	<i>Cavia porcellus</i>	Guinea Pig	cavPor3
8	<i>Ceratotherium simum simum</i>	White Rhino	cerSim1
9	<i>Chinchilla lanigera</i>	Chinchilla	chiLan1
10	<i>Chlorocebus sabaesus</i>	Green monkey	chlSab2
11	<i>Chrysochloris asiatica</i>	Cape golden mole	chrAsi1
12	<i>Condylura cristata</i>	Star-nosed mole	conCri1
13	<i>Cricetulus griseus</i>	Chinese Hamster	criGri1
14	<i>Dasypus novemcinctus</i>	Armadillo	dasNov3
15	<i>Echinops telfairi</i>	Tenrec	echTel2
16	<i>Elephantulus zayadai</i>	Cape elephant shrew	eleEdw1
17	<i>Eptesicus fuscus</i>	Big brown bat	eptFus1
18	<i>Equus caballus</i>	Horse	equCab2
19	<i>Erinaceus europaeus</i>	Hedgehog	eriEur2
20	<i>Felis catus</i>	Cat	felCat8
21	<i>Heterocephalus glaber</i>	Naked mole-rat	hetGla2
22	<i>Homo sapiens</i>	Human	hg38
23	<i>Jaculus jaculus</i>	Lesser Egyptian jerboa	jacJac1
24	<i>Leptonychotes weddellii</i>	Weddell seal	lepWed1
25	<i>Loxodonta africana</i>	African Elephant	loxAfr3
26	<i>Macaca fascicularis</i>	Crab-eating macaque	macFas5
27	<i>Macaca mulatta</i>	Rhesus macaque	rheMac8
28	<i>Mesocricetus auratus</i>	Golden hamster	mesAur1
29	<i>Microtus ochrogaster</i>	Prairie vole	micOch1
30	<i>Mus musculus</i>	Mouse	mm10
31	<i>Mustela putorius furo</i>	Ferret	musFur1
32	<i>Myotis davidii</i>	David’s myotis (bat)	myoDav1
33	<i>Myotis lucifugus</i>	Microbat	myoLuc2
34	<i>Nomascus leucogenys</i>	Gibbon	nomLeu3
35	<i>Ochotona princeps</i>	Pika	ochPri3
36	<i>Octodon degus</i>	Brush-tailed rat	octDeg1
37	<i>Odobenus rosmarus divergens</i>	Pacific walrus	odoRosDiv1
38	<i>Orcinus orca</i>	Killer whale	orcOrc1
39	<i>Orycteropus afer afer</i>	Aardvark	oryAfe1
40	<i>Oryctolagus cuniculus</i>	Rabbit	oryCun2
41	<i>Otolemur garnettii</i>	Bushbaby	otoGar3
42	<i>Ovis aries</i>	Sheep	oviAri3
43	<i>Pan troglodytes</i>	Chimp	panTro5
44	<i>Pantholops hodgsonii</i>	Tibetan antelope	panHod1
45	<i>Papio Anubis</i>	Baboon	papAnu2
46	<i>Pongo pygmaeus abelii</i>	Orangutan	ponAbe2
47	<i>Pteropus alecto</i>	Black flying-fox	pteAle1
48	<i>Pteropus vampyrus</i>	Megabat	pteVam1
49	<i>Rattus norvegicus</i>	Rat	rn6
50	<i>Saimiri boliviensis boliviensis</i>	Squirrel monkey	saiBol1
51	<i>Sorex araneus</i>	Shrew	sorAra2
52	<i>Spermophilus tridecemlineatus</i>	Squirrel	speTri2
53	<i>Sus scrofa</i>	Pig	susScr3
54	<i>Tarsius syrichta</i>	Tarsier	tarSyr2
55	<i>Trichechus manatus</i>	Manatee	triMan1
56	<i>Tupaia belangeri</i>	Tree shrew	tupBel1
57	<i>Tupaia belangeri chinensis</i>	Chinese tree shrew	tupChi1
58	<i>Tursiops truncatus</i>	Dolphin	turTru2
59	<i>Vicugna pacos</i>	Alpaca	vicPac2

Supplementary Table 1

The 59 mammalian genome assemblies used in this study. Scientific name, common name, and the assembly name of the 59 mam-
malian species (including human) used in this study.

Supplementary Figure 1

Phylogenetic tree of the 59 mammalian species. The phylogenetic tree of the 59 mammalian species (including human) used in this study. Our pipeline (Figure 3) requires knowledge of only the topology of the tree as shown in this figure and not of the branch lengths.