RESEARCH ARTICLE

Pervasive duplication of tumor suppressors in Afrotherians during the evolution of large bodies and reduced cancer risk

Juan M Vazquez[1†], Vincent J Lynch[2*]

[1]Department of Human Genetics, The University of Chicago, Chicago, United States; [2]Department of Biological Sciences, University at Buffalo, Buffalo, United States

Abstract The risk of developing cancer is correlated with body size and lifespan within species. Between species, however, there is no correlation between cancer and either body size or lifespan, indicating that large, long-lived species have evolved enhanced cancer protection mechanisms. Elephants and their relatives (Proboscideans) are a particularly interesting lineage for the exploration of mechanisms underlying the evolution of augmented cancer resistance because they evolved large bodies recently within a clade of smaller-bodied species (Afrotherians). Here, we explore the contribution of gene duplication to body size and cancer risk in Afrotherians. Unexpectedly, we found that tumor suppressor duplication was pervasive in Afrotherian genomes, rather than restricted to Proboscideans. Proboscideans, however, have duplicates in unique pathways that may underlie some aspects of their remarkable anti-cancer cell biology. These data suggest that duplication of tumor suppressor genes facilitated the evolution of increased body size by compensating for decreasing intrinsic cancer risk.

***For correspondence:** vjlynch@buffalo.edu

Present address: † Department of Integrative Biology, University of California – Berkeley, Berkeley, United States

Competing interests: The authors declare that no competing interests exist.

Funding: See page 17

Received: 19 November 2020
Accepted: 28 January 2021
Published: 29 January 2021

Reviewing editor: Antonis Rokas, Vanderbilt University, United States

Introduction

Among the constraints on the evolution of large bodies and long lifespans in animals is an increased risk of developing cancer. If all cells in all organisms have a similar risk of malignant transformation and equivalent cancer suppression mechanisms, then organisms with many cells should have a higher prevalence of cancer than organisms with fewer cells, particularly because large and small animals have similar cell sizes (***Savage et al., 2007***). Consistent with this expectation there is a strong positive correlation between body size and cancer incidence within species; for example, cancer incidence increases with increasing adult height in humans (***Million Women Study collaborators et al., 2011***; ***Nunney, 2018***) and with increasing body size in dogs, cats, and cattle (***Dobson, 2013***; ***Dorn et al., 1968***; ***Lucena et al., 2011***). There is no correlation, however, between body size and cancer risk between species; this lack of correlation is often referred to as 'Peto's Paradox' (***Caulin and Maley, 2011***; ***Leroi et al., 2003***; ***Peto et al., 1975***). Indeed, cancer prevalence is relatively stable at ~5% across species with diverse body sizes ranging from the minuscule 51 g grass mouse to the gargantuan 4800 kg African elephant (***Abegglen et al., 2015***; ***Boddy et al., 2020***; ***Tollis et al., 2020***). The ultimate resolution to Peto's Paradox is trivial, large-bodied and long-lived species evolved enhanced cancer protection mechanisms, but identifying and characterizing the mechanisms that underlie the evolution of augmented cancer protection has proven difficult (***Ashur-Fabian et al., 2004***; ***Seluanov et al., 2008***; ***Gorbunova et al., 2012***; ***Tian et al., 2013***; ***Sulak et al., 2016***).

eLife digest From the gigantic blue whale to the minuscule bumblebee bat, animals come in all shapes and sizes. Any species can develop cancer, but some are more at risk than others. In theory, if every cell has the same probability of becoming cancerous, then bigger animals should get cancer more often since they have more cells than smaller ones. Amongst the same species, this relationship is true: taller people and bigger dogs have a greater cancer risk than their smaller counterparts.

Yet this correlation does not hold when comparing between species: remarkably large creatures, like elephants and whales, are not more likely to have cancer than any other animal. But how have these gigantic animals evolved to be at lower risk for the disease?

To investigate, Vazquez and Lynch compared the cancer risk and the genetic information of a diverse group of closely related animals with different body sizes. This included elephants, woolly mammoths and mastodons as well as their small relatives, the manatees, armadillos, and marmot-sized hyraxes. Examining these species' genomes revealed that, during evolution, elephants had acquired extra copies of 'tumour suppressor genes' which can sense and repair the genetic and cellular damages that turn healthy cells into tumours. This allowed the species to evolve large bodies while lowering their risk of cancer.

Further studies could investigate whether other gigantic animals evolved similar ways to shield themselves from cancer; these could also examine precisely how having additional copies of cancer-protecting genes helps reduce cancer risk, potentially paving the way for new approaches to treat or prevent the disease.

One of the challenges for discovering how animals evolved enhanced cancer protection mechanisms is identifying lineages in which large-bodied species are nested within species with small body sizes. Afrotherian mammals are generally small-bodied, but also include the largest extant land mammals. For example, maximum adult weights are ~70 g in golden moles,~120 g in tenrecs,~170 g in elephant shrews,~3 kg in hyraxes, and ~60 kg in aardvarks (***Tacutu et al., 2013***). In contrast, while extant hyraxes are relatively small, the extinct Titanohyrax is estimated to have weighed ~1300 kg (***Schwartz et al., 1995***). The largest living Afrotheria are also dwarfed by the size of their recent extinct relatives: extant sea cows such as manatees are large bodied (~322–480 kg) but are relatively small compared to the extinct Stellar's sea cow which is estimated to have weighed ~8000–10,000 kg (***Scheffer, 1972***). Similarly African Savannah (4800 kg) and Asian elephants (3200 kg) are large, but are dwarfed by the truly gigantic extinct Proboscideans such as Deinotherium (~12,000 kg), *Mammut borsoni* (16,000 kg), and the straight-tusked elephant (~14,000 kg) (***Larramendi, 2015***). Remarkably, these large-bodied Afrotherian lineages are nested deeply within small-bodied species (***Figure 1***; ***O Leary et al., 2013a***; ***Springer et al., 2013***; ***O Leary et al., 2013b***; ***Puttick and Thomas, 2015***), indicating that gigantism independently evolved in hyraxes, sea cows, and elephants (Paenungulata). Thus, Paenungulates are an excellent model system in which to explore the mechanisms that underlie the evolution of large body sizes and augmented cancer resistance.

Box 1. Eutherian phylogenetic relationships.

Eutheria (eu- 'good' or 'right' and thēríon 'beast', hence 'true beasts') is one of three living (extant) mammalian lineages (Monotremes, Marsupials, and Eutherians) that diverged in the early–late Cretaceous. Eutheria was named in 1872 by Theodore Gill and refined by Thomas Henry Huxley in 1880. Living Eutherians are comprised of 18 orders, divided into two major clades (***Figure 1A***): Atlantogenata including the superorders Xenarthra (armadillos, anteaters, and sloths) and Afrotheria (Proboscidea, Sirenia, Hyracoidea, Tublidentata, Afroinsectivora, Cingulata, and Pilosa), and Boreoeutheria including the superorders Laurasiatheria (Insectivora, Artodactyla, Pholidota, and Carnovora) and Euarchontoglires (Lagomorpha, Rodentia, Scandentia, Dermoptera, and Primates). In our analyses, we have focused on identifying gene duplications in Afrotherian and Xenarthran genomes (***Figure 1B***), using the Xenarthrans Hoffmans two-toed sloth (*Choloepus hoffmanni*) and nine-banded armadillo (*Dasypus*

novemcinctus) as out-groups to the Afrotherians. This approach allows us to use phylogenetic methods to polarize gene duplication events and identify genes that duplicated in the Afrotherian stem-lineage.

Many mechanisms have been suggested to resolve Peto's paradox, including a decrease in the copy number of oncogenes, an increase in the copy number of tumor suppressor genes (**Caulin and Maley, 2011**; **Leroi et al., 2003**; **Nunney, 1999**), reduced metabolic rates, reduced retroviral activity and load (**Katzourakis et al., 2014**), and selection for 'cheater' tumors that parasitize the growth of other tumors (**Nagy et al., 2007**), greater sensitivity of cells to DNA damage (**Abegglen et al., 2015**; **Sulak et al., 2016**), enhanced recognition of neoantigens by T cells, among many others. Among the most parsimonious routes to enhanced cancer resistance may be through an increased copy number of tumor suppressors. For example, transgenic mice with additional copies of *TP53* have reduced cancer rates and extended lifespans (**García-Cao et al., 2002**), suggesting that changes in the copy number of tumor suppressors can affect cancer rates. Indeed, candidate genes studies have found that elephant genomes encode duplicate tumor suppressors such as *TP53* and *LIF* (**Abegglen et al., 2015**; **Sulak et al., 2016**; **Vazquez et al., 2018**) as well as other genes with putative tumor suppressive functions (**Caulin et al., 2015**; **Doherty and de Magalhães, 2016**). These studies, however, focused on a priori candidate genes; thus it is unclear whether duplication of tumor suppressor genes is a general phenomenon in the elephant lineage or reflects an ascertainment bias.

Here we trace the evolution of body mass, cancer risk, and gene copy number variation across Afrotherian genomes, including multiple living and extinct Proboscideans (**Figure 1**), to investigate whether duplications of tumor suppressors coincided with the evolution of large body sizes. Our estimates of the evolution of body mass across Afrotheria show that large body masses evolved in a stepwise manner, similar to previous studies (**O Leary et al., 2013a**; **Springer et al., 2013**; **O Leary et al., 2013b**; **Puttick and Thomas, 2015**) and coincident with dramatic reductions in intrinsic cancer risk. To explore whether duplication of tumor suppressors occurred coincident with the evolution of large body sizes, we used a genome-wide Reciprocal Best BLAT Hit (RBBH) strategy to identify gene duplications and used maximum likelihood to infer the lineages in which those duplications occurred. Unexpectedly, we found that duplication of tumor suppressor genes was common in Afrotherians, both large and small. Gene duplications in the Proboscidean lineage, however, were uniquely enriched in pathways that may explain some of the unique cancer protection mechanisms observed in elephant cells. These data suggest that duplication of tumor suppressor genes is pervasive in Afrotherians and preceded the evolution of species with exceptionally large body sizes.

Results

Step-wise evolution of body size in Afrotherians

Similar to previous studies of Afrotherian body size (**Puttick and Thomas, 2015**; **Elliot and Mooers, 2014**), we found that the body mass of the Afrotherian ancestor was inferred to be small (0.26 kg, 95% CI: 0.31–3.01 kg) and that substantial accelerations in the rate of body mass evolution occurred coincident with a 67.36× increase in body mass in the stem-lineage of Pseudoungulata (17.33 kg); a 1.45× increase in body mass in the stem-lineage of Paenungulata (25.08 kg); a 11.82× increase in body mass in the stem-lineage of Tehthytheria (296.56 kg); a 1.39× increase in body mass in the stem-lineage of Proboscidea (412.5 kg); and a 2.69× increase in body mass in the stem-lineage of Elephantimorpha (4114.39 kg), which is the last common ancestor of elephants and mastodons using the fossil record (**Figure 2A,B**). The ancestral Hyracoidea was inferred to be relatively small (2.86–118.18kg), and rate accelerations were coincident with independent body mass increases in large hyraxes such as *Titanohyrax andrewsi* (429.34 kg, 67.36× increase) (**Figure 2A,B**). While the body mass of the ancestral Sirenian was inferred to be large (61.7–955.51 kg), a rate acceleration occurred coincident with a 10.59× increase in body mass in Stellar's sea cow (**Figure 2A,B**). Rate accelerations also occurred coincident with dramatic reductions in body mass (36.6× decrease) in the stem-lineage of the dwarf elephants *Elephas (Palaeoloxodon) antiquus falconeri* and *Elephas cypriotes* (**Figure 2A,B**). These data indicate that gigantism in Afrotherians evolved step-wise, from small to

medium bodies in the Pseudoungulata stem-lineage, medium to large bodies in the Tehthytherian stem-lineage and extinct hyraxes, and from large to exceptionally large bodies independently in the Proboscidean stem-lineage and Stellar's sea cow (***Figure 2A,B***).

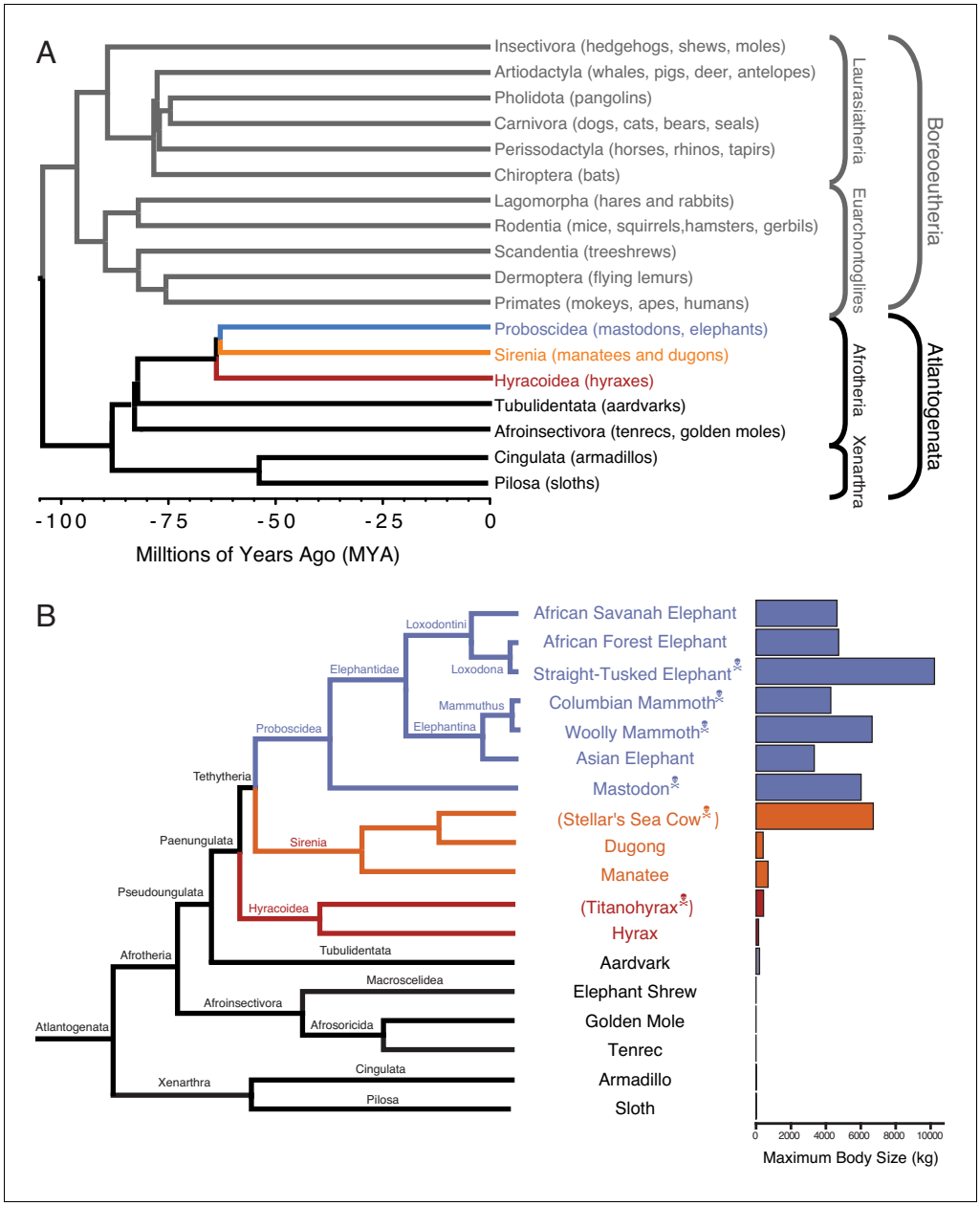

Figure 1. Large-bodied Afrotherians are nested within species with smaller body sizes (***Tacutu et al., 2013***; ***Puttick and Thomas, 2015***). (**A**) Phylogenetic relationships between Eutherian orders, examples of each order are given in parenthesis. Horizontal branch lengths are proportional to time since divergence between lineages (see scale, Millions of Ago [MYA]). The clades Atlantogenata and Boreoeutheria are indicated, the order Proboscidea is colored blue, Sirenia is colored orange, and Hyracoidea is colored red. (**B**) Phylogenetic relationships of extant and recently extinct Atlantogenatans with available genomes are shown along with clade names and maximum body sizes. Note that horizontal branch lengths are arbitrary, species indicated with skull and crossbones are extinct, and those in parentheses do not have genomes. The order Proboscidea is colored blue, Sirenia is colored orange, and Hyracoidea is colored red.

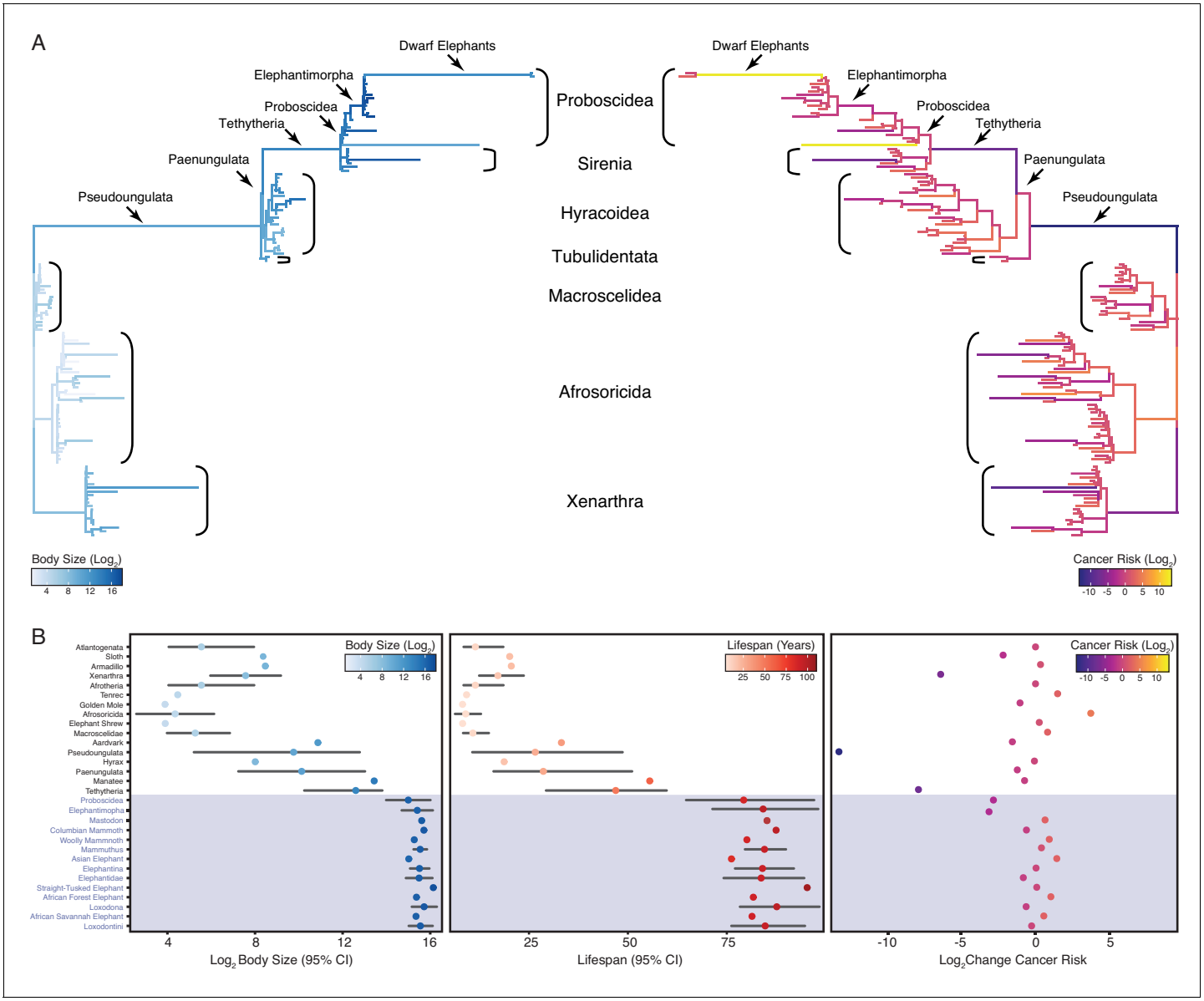

Figure 2. Convergent evolution of large-bodied, cancer resistant Afrotherians. (**A**) Atlantogenatan phylogeny, with branch lengths scaled by $\log_2$ change in body size (left) or $\log_2$ change in intrinsic cancer risk (right). Branches are colored according to ancestral state reconstruction of body mass or estimated intrinsic cancer risk. Clades and lineages leading to extant Proboscideans and dwarf elephants are labeled. (**B**) Extant and ancestral body size (left), lifespan (middle), and estimated intrinsic cancer risk reconstructions; data are shown as mean (dot) and 95% confidence interval (CI, whiskers).

Step-wise reduction of intrinsic cancer risk in large, long-lived Afrotherians

In order to account for a relatively stable cancer rate across species (***Abegglen et al., 2015***; ***Boddy et al., 2020***; ***Tollis et al., 2020***), intrinsic cancer risk must also evolve with changes in body size and lifespan across species. We used empirical body size and lifespan data from extant species and empirical body size and estimated lifespan data from extinct species to estimate intrinsic cancer risk (K) with the simplified multistage cancer risk model $K \approx Dt^6$, where D is the maximum body size and t is the maximum lifespan (***Peto et al., 1975***: ***Peto, 2015***; ***Armitage, 1985***; ***Armitage and Doll, 2004***). As expected, intrinsic cancer risk in Afrotheria also varies with changes in body size and longevity (***Figure 2A,B***), with a 6.41-$\log_2$ decreases in the stem-lineage of Xenarthra, followed by a 13.37-$\log_2$ decrease in Pseudoungulata, and a 1.49-$\log_2$ decrease in Aardvarks (***Figure 2A***). In contrast to the Paenungulate stem-lineage, there is a 7.84-$\log_2$ decrease in cancer risk in Tethytheria, a

0.67-log_2 decrease in Manatee, a 3.14-log_2 decrease in Elephantimorpha, and a 1.05-log_2 decrease in Proboscidea. Relatively minor decreases occurred within Proboscidea including a 0.83-log_2 decrease in Elephantidae and a 0.57-log_2 decrease in the American Mastodon. Within the Elephantidae, Elephantina and Loxodontini have a 0.06-log_2 decrease in cancer susceptibility, while susceptibility is relatively stable in Mammoths. The three extant Proboscideans, Asian Elephant, African Savana Elephant, and the African Forest Elephant, meanwhile, have similar decreases in body size, with slight increases in cancer susceptibility (***Figure 2A,B***).

Pervasive duplication of tumor suppressor genes in Afrotheria

Our hypothesis was that genes which duplicated coincident with the evolution of increased body mass (IBM) and reduced intrinsic cancer risk (RICR) would be uniquely enriched in tumor suppressor pathways compared to genes that duplicated in other lineages. Therefore, we identified duplicated genes in each Afrotherian lineage (***Table 1*** and ***Figure 3A***) and tested if they were enriched in Reactome pathways related to cancer biology (***Figure 3B***, ***Table 2***). No pathways related to cancer biology were enriched in either the Pseudoungulata (67.36-fold IBM, 13.37-log_2 RICR), but few genes were inferred to be duplicated in this lineage reducing power to detect enriched pathways. Consistent with our hypothesis, 18.18% of the pathways that were enriched in the Paenungulate stem-lineage (1.45-fold IBM, 1.17-log_2 RICR), 63% of the pathways that were enriched in the Tethytherian stem-lineage (11.82-fold IBM, 7.84-log_2 RICR), and 38.81% of the pathways that were enriched in the Proboscidean stem-lineage (1.06-fold IBM, 3.14-log_2 RICR) were related to tumor suppression (***Figure 3B***, ***Table 2***). Similarly, 21.28% and 38.00% of the pathways that were enriched in manatee (1.11-fold IBM, 0.89-log_2 RICR) and aardvark (67.36-fold IBM, 1.49-log_2 RICR), respectively, were related to tumor suppression. In contrast, only 2.86% of the pathways that were enriched in hyrax (1.6-fold IBM, 1.49-log_2 RICR) were related to tumor suppression (***Figure 3B***, ***Table 2***). Unexpectedly, however, lineages without major increases in body size or lifespan, or decreases in intrinsic cancer risk, were also enriched for tumor suppressor pathways. For example, 13.85%, 37.04%, and 22.00% of the pathways that were enriched in the stem-lineages of Afroinsectivoa and Afrosoricida, and in *E. telfairi*, respectively, were related to cancer biology (***Figure 3B***, ***Table 2***).

Our observation that gene duplicates in most lineages are enriched in cancer pathways suggest either that duplication of genes in cancer pathways is common in Afrotherians, or that there may be a systemic bias in the pathway enrichment analyses. For example, random gene sets may be generally enriched in pathway terms related to cancer biology. To explore this latter possibility, we generated 5000 randomly sampled gene sets of between 10 and 5000 genes, and tested for enriched Reactome pathways using ORA. We found that no cancer pathways were enriched (median hypergeometric p-value $\leq$0.05) among gene sets tested greater than 157 genes; however, in these smaller gene sets, 12–18% of enriched pathways were classified as cancer pathways. Without considering p-value thresholds, the percentage of enriched cancer pathways approaches ~15% (213/1381) in simulated sets. Thus, for larger gene sets, we used a simulated threshold of ~15% to determine if pathways related to cancer biology were enriched more than one would expect from sampling bias (***Table 2***). We directly compared our simulated and observed enrichment results by lineage and gene set size, and found that Afrosoricida, Cape golden mole, tenrec, Elephantidae, elephant shrew, Asian elephant, African Savannah elephant, African Forest elephant, Columbian mammoth, aardvark, Paenungulata, Proboscidea, Tethytheria, and manatee had enriched cancer pathway percentages above background with respect to their gene set sizes, that is expected enrichments based on random sampling of small gene sets (***Table 2***). Thus, we conclude that duplication of genes in cancer pathways is common in many Afrotherians but that the inference of enriched cancer pathway duplication is not different from background in some lineages, particularly in ancestral nodes with a small number of estimated duplicates.

Tumor suppressor pathways enriched exclusively within Proboscideans

While duplication of cancer associated genes is common in Afrotheria, the 157 genes that duplicated in the Proboscidean stem-lineage (***Figure 3A***) were uniquely enriched in 12 pathways related to cancer biology (***Figure 3B***). Among these uniquely enriched pathways (***Figure 3C***) were pathways related to the cell cycle, including 'G0 and Early G1', 'G2/M Checkpoints', and 'Phosphorylation of the APC/C', pathways related to DNA damage repair including 'Global Genome Nucleotide Excision

Table 1. Genomes used in this study.

Species	Common Name	Genomes	Highest Quality Genome	Reference(s)
Choloepus hoffmanni	Hoffmans two-toed sloth	choHof1, choHof2, choHof-C_hoffmanni-2.0.1_HiC	choHof-C_hoffmanni-2.0.1_HiC	**Dudchenko et al., 2017**
Chrysochloris asiatica	Cape golden mole	chrAsi1m	chrAsi1m	GCA_000296735.1
Dasypus novemcinctus	Nine-banded armadillo	dasNov3	dasNov3	GCA_000208655.2
Echinops telfairi	Lesser Hedgehog Tenrec	echTel2	echTel2	GCA_000313985.1
Elephantulus edwardii	Cape elephant shrew	eleEdw1m	eleEdw1m	GCA_000299155.1
Elephas maximus	Asian elephant	eleMaxD	eleMaxD	**Palkopoulou et al., 2018**
Loxodonta africana	African savanna elephant	loxAfr3, loxAfrC, loxAfr4	loxAfr4	ftp://ftp.broadinstitute.org/pub/assemblies/mammals/elephant/loxAfr4
Loxodonta cyclotis	African forest elephant	loxCycF	loxCycF	**Palkopoulou et al., 2018**
Mammut americanum	American mastodon	mamAmeI	mamAmeI	**Palkopoulou et al., 2018**
Mammuthus columbi	Columbian mammoth	mamColU	mamColU	**Palkopoulou et al., 2018**
Mammuthus primigenius	Woolly mammoth	mamPriV	mamPriV	**Palkopoulou et al., 2015**
Orycteropus afer	Aardvark	oryAfe1, oryAfe2	oryAfe2	**Dudchenko et al., 2017**
Palaeoloxodon antiquus	Straight tusked elephant	palAntN	palAntN	**Palkopoulou et al., 2018**
Procavia capensis	Rock hyrax	proCap1, proCap2, proCap-Pcap_2.0_HiC	proCap-Pcap_2.0_HiC	**Dudchenko et al., 2017**; **Lindblad-Toh et al., 2011**
Trichechus manatus latirostris	Manatee	triMan1, triManLat2	triManLat2	**Dudchenko et al., 2017**; **Foote et al., 2015**

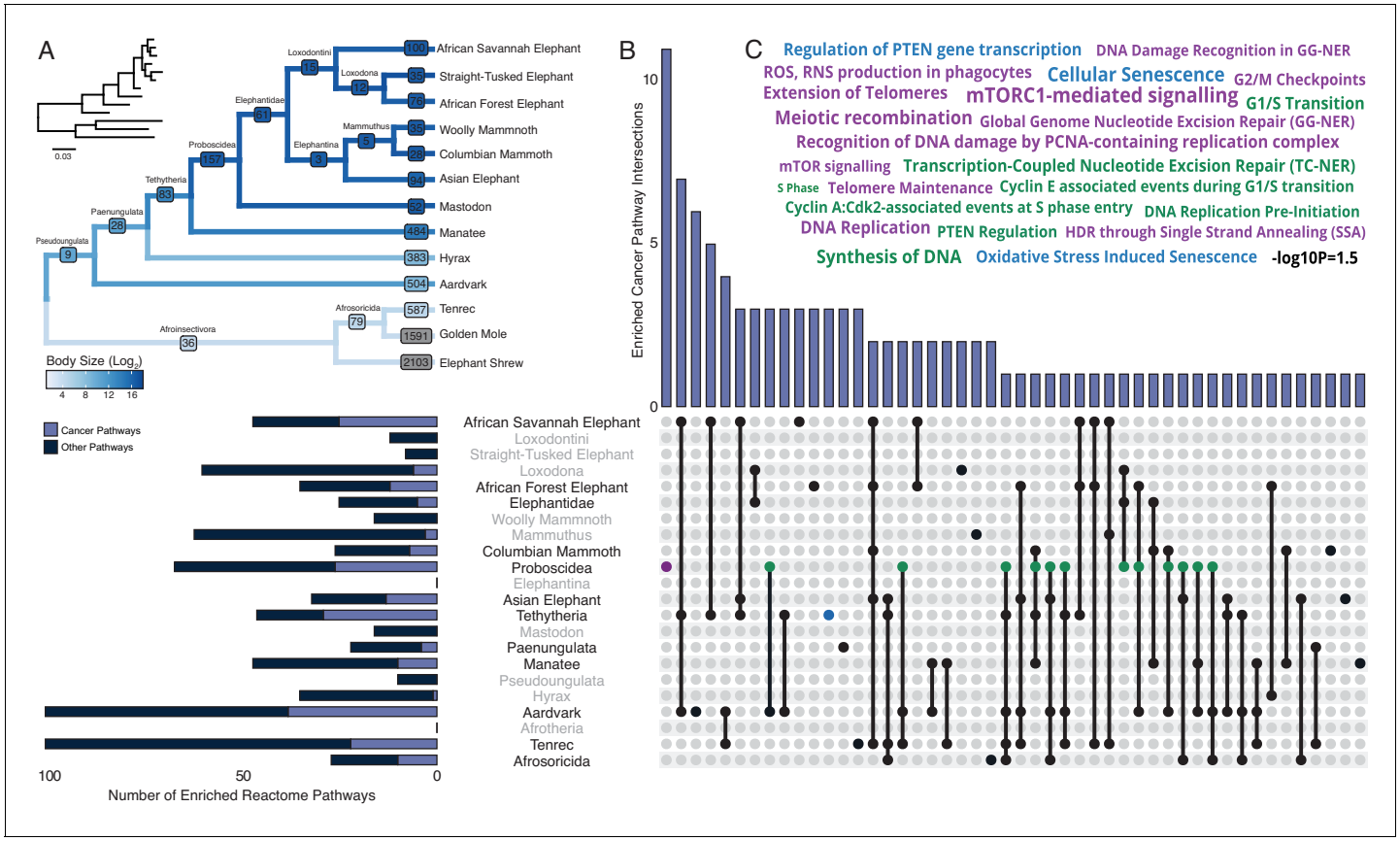

Figure 3. Pervasive duplication of tumor suppressors in Atlantogenata. (**A**) Afrotherian phylogeny indicating the number of genes duplicated in each lineage, inferred by maximum likelihood with Bayesian posterior probability (BPP) ≥0.80. Branches are colored according to $\log_2$ change in body size. Inset, phylogeny with branch lengths proportional to gene expression changes per gene. (**B**) Upset plot of cancer related Reactome pathways enriched in each Afrotherian lineage; lineages in which the cancer pathway enrichment percentage is less than background are shown in gray. Note that Upset plots are Euler diagrams showing intersections between sets; lines indicate intersections in pathway terms between lineages connected by that line (for example, the line connecting the points for Aardvark and Tenrec indicate pathway indications for those two lineages), and empty sets are not shown. (**C**) Wordcloud of pathways enriched exclusively in the Proboscidean stem-lineage (purple), shared between Proboscidea and Tethytheria (blue), or shared between Proboscidea and any other lineage (green).

The online version of this article includes the following figure supplement(s) for figure 3:

Figure supplement 1. Estimated Copy Number by Coverage (ECNC) consolidates fragmented genes while accounting for missing domains in homologs.

Figure supplement 2. Correlations between genome quality metrics and ECNC metrics.

Repair (GG-NER)', 'HDR through Single Strand Annealing (SSA)', 'Gap-filling DNA repair synthesis and ligation in GG-NER', 'Recognition of DNA damage by PCNA-containing replication complex', and 'DNA Damage Recognition in GG-NER', pathways related to telomere biology including 'Extension of Telomeres' and 'Telomere Maintenance', pathways related to the apoptosome including 'Activation of caspases through apoptosome-mediated cleavage', and pathways related to 'mTORC1-mediated signaling' and 'mTOR signaling', which play important roles in the biology of aging. Thus, duplication of genes with tumor suppressor functions is pervasive in Afrotherians, but genes in some pathways related to cancer biology and tumor suppression are uniquely duplicated in large-bodied (long-lived) Proboscideans (***Figure 4A,B***).

Among the genes uniquely duplicated within Proboscideans are *TP53*, *COX20*, *LAMTOR5*, *PRDX1*, *STK11*, *BRD7*, *MAD2L1*, *BUB3*, *UBE2D1*, *SOD1*, *LIF*, *MAPRE1*, *CNOT11*, *CASP9*, *CD14*, and *HMGB2* (***Figure 4C***). Two of these, *TP53* and *LIF*, have been previously described (***Abegglen et al., 2015***; ***Sulak et al., 2016***; ***Vazquez et al., 2018***). These genes are significantly enriched in pathways involved in apoptosis, cell cycle regulation, and both upstream and downstream pathways involving

Table 2. Summary of reactome pathways in Atlantogenata.

	Number of		Percentage		
	Genes	Pathways	Cancer pathways	Simulated cancer pathways	Cancer pathways greater than simulated?
Afroinsectivora	36	65	13.85%	15.42%	No
Afrosoricida	79	27	37.04%	15.42%	Yes
Chrysochloris asiatica	1591	100	27.00%	15.42%	Yes
Echinops telfairi	587	100	22.00%	15.42%	Yes
Elephantidae	61	25	20.00%	13.03%	Yes
Elephantulus edwardii	2103	100	22.00%	15.42%	Yes
Elephas maximus	94	32	40.63%	17.73%	Yes
Loxodona	12	60	10.00%	14.53%	No
Loxodonta africana	100	47	53.19%	15.42%	Yes
Loxodonta cyclotis	76	35	34.29%	16.11%	Yes
Loxodontini	15	12	0.00%	13.82%	No
Mammut americanum	52	16	0.00%	12.91%	No
Mammuthus	5	62	4.84%	15.29%	No
Mammuthus columbi	28	26	26.92%	12.88%	Yes
Mammuthus primigenius	35	16	0.00%	12.28%	No
Orycteropus afer	504	100	38.00%	15.42%	Yes
Paenungulata	28	22	18.18%	12.88%	Yes
Palaeoloxodon antiquus	35	8	0.00%	12.28%	No
Proboscidea	157	67	38.81%	9.52%	Yes
Procavia capensis	383	35	2.86%	15.42%	No
Pseudoungulata	9	10	0.00%	14.90%	No
Tethytheria	83	46	63.04%	18.52%	Yes
Trichechus manatus	484	47	21.28%	15.42%	Yes

TP53. The majority of these genes are expressed in African Elephant transcriptome data (***Figure 4D***), suggesting that they maintained functionality after duplication.

Coordinated duplication of TP53-related genes in Proboscidea

Prior studies found that the 'master' tumor suppressor *TP53* duplicated multiple times in elephants (***Abegglen et al., 2015***; ***Sulak et al., 2016***), motivating us to further study duplication of genes involved in *TP53*-related pathways in Proboscidea. We traced the evolution of genes in the TP53 pathway that appeared in one or more Reactome pathway enrichments for genes duplicated recently in the African Elephant, which has the most complete genome among Proboscideans and for which several RNA-Seq data sets are available. We found that the initial duplication of TP53 in Tethytheria, where body size expanded, was preceded by the duplication of *GTF2F1* and *STK11* in Paenungulata and was coincident with the duplication of *BRD7*. These three genes are involved in regulating the transcription of *TP53* (***Liang and Mills, 2013***; ***Launonen, 2005***; ***Drost et al., 2010***; ***Burrows et al., 2010***), and their duplication prior to that of *TP53* may have facilitated re-functionalization of *TP53* retroduplicates. Interestingly, *STK11* is also tumor suppressor that mediates tumor suppression via p21-induced senescence (***Launonen, 2005***). The other genes that are duplicated in the pathway are downstream of *TP53*; these genes duplicated either coincident with *TP53*, as in the case of *SIAH1*, or subsequently in Proboscidea, Elephantidae, or extant elephants (***Figure 4***). These genes are expressed in RNA-Seq data (***Figure 4D***), suggesting that they are functional.

While transcript abundance estimates inferred from RNA-Seq data can suggest that genes are functional, recent non-functional duplicates can still be transcribed. Therefore we inferred if each duplicate shown in ***Figure 4C/D*** encoded a putatively function protein by manually curation,

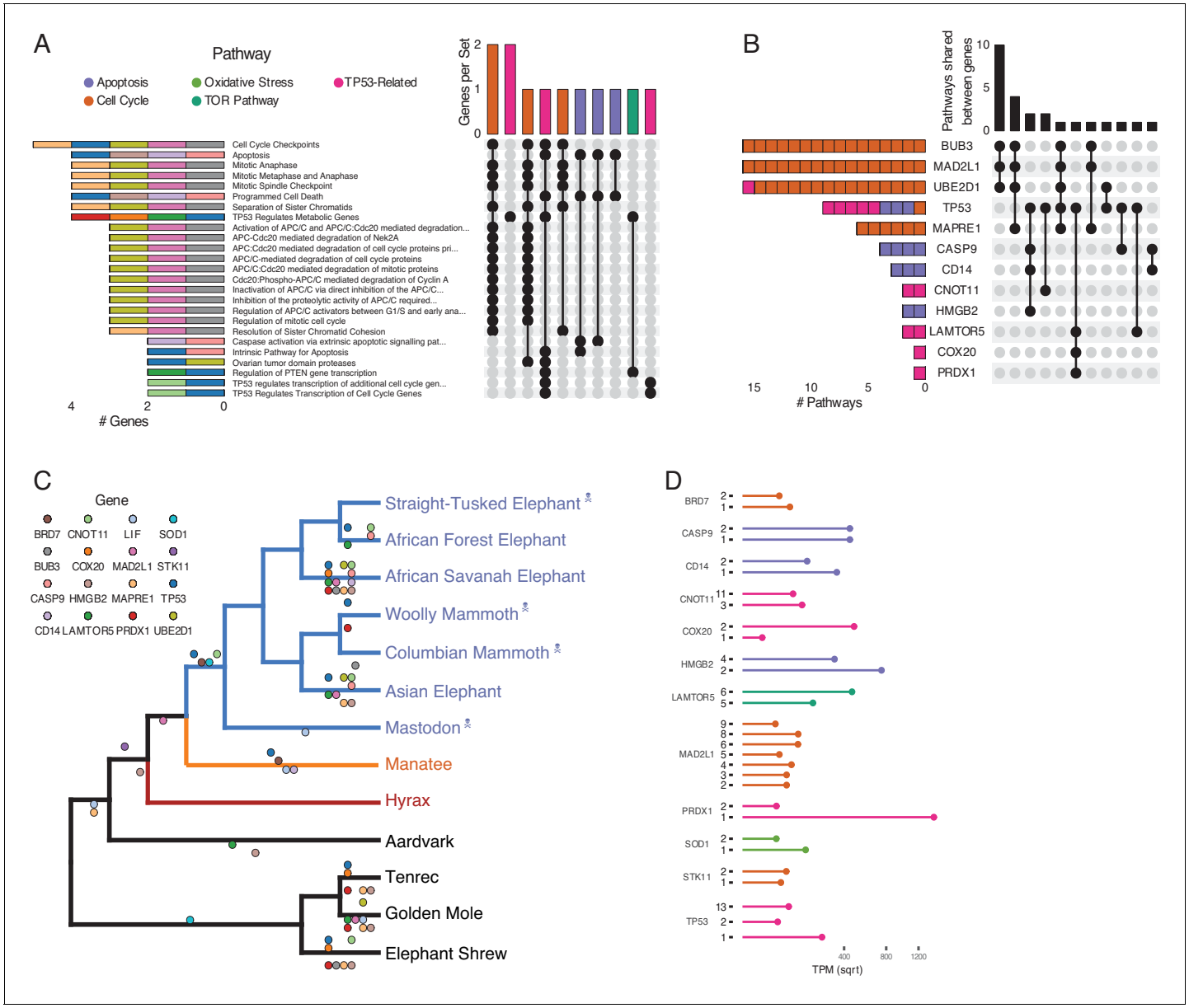

Figure 4. Duplications in the African savannah elephant (*Loxodonta africana*) are enriched for TP53-related and other tumor suppressor processes. (**A**) Upset plot of cancer-related Reactome pathways in African savannah elephant, highlighting shared genes in each set, and the pathway class represented by the combinations (see ***Figure 3*** for a description of Upset plots). (**B**) Inverted Upset plot from **A** showing the pathways shared by genes highlighted by WEBGESTALT in each pathway. (**C**) Cladogram of Afrotheria with sequenced genomes. Exemplar tumor suppressor duplicates are mapped onto lineages in which those genes are duplicated. Dots represent a duplication event of the color-coded genes. Note that we are unable to determine duplication status for some genes in Proboscideans because of assembly gaps in ancient genomes (indicated with skull and crossbones); these genes appear to be independently duplicated in extant species (African Forest, African Savanah, and Asian elephants) because they are missing from ancient genomes, biasing ancestral reconstructions of duplication status. (**D**) Gene expression levels of genes from panel **C** that have two or more expressed duplicates.

The online version of this article includes the following source data for figure 4:

Source data 1. Data set used for manual coding gene potential associated with ***Figure 4C,D***.

specifically to identify premature stop codons and overall sequence conservation. Most genes in ***Figure 4C/D***, such as *STK11*, *CD14*, *SOD1*, and *BRD7*, were well conserved and lacked premature stop codons. We also find that the *STK11*, *CD14*, and *BRD7* genes in the manatee were also well conserved, suggesting that extant manatees may also have enhanced tumor suppression and an

augmented stress response. However, some of the duplicate genes in the mantatee genome have premature stop codons suggesting they are not translated into functional proteins, including the additional copies of *MAPRE1*, *BUB3*, and *COX20* as well as at least one of the duplicate copies of *CNOT11*, *HMGB2*, *MAD2L1*, *LIF*, and *TP53*. For *TP53*, we have previously shown that duplicate copies of genes containing premature stop codons may still serve a functional role in regulating its progenitor's function. Thus, some of the genes with premature stop codons, such as duplicate *COX20* and *MAD2L1* which are expressed in RNA-Seq data, may encode functional lncRNA transcripts or truncated proteins. Some copies, including for *CASP9* and *PRDX1*, contained partial RBBH hits with no premature stop codons; however, they also lacked the totality of the coding sequence and thus may represent cases of pseudogenization, subfunctionalization, or neofunctionalization.

Discussion

Among the evolutionary, developmental, and life history constraints on the evolution of large bodies and long lifespans is an increased risk of developing cancer. While body size and lifespan are correlated with cancer risk within species, there is no correlation between species because large and long-lived organisms have evolved enhanced cancer suppression mechanisms. While this ultimate evolutionary explanation is straightforward (**Peto, 2015**), determining the mechanisms that underlie the evolution of enhanced cancer protection is challenging because many mechanisms with relatively small effects likely contribute to evolution of reduced cancer risk. Previous candidate gene studies in elephants have identified duplications of tumor suppressors such as *TP53* and *LIF*, among others, suggesting that an increased copy number of tumor suppressors may contribute to the evolution of large body sizes in the elephant lineage (**Abegglen et al., 2015**; **Sulak et al., 2016**; **Vazquez et al., 2018**; **Caulin et al., 2015**; **Doherty and de Magalhães, 2016**). Here we: (1) trace the evolution of body size and lifespan in Eutherian mammals, with particular reference to Afrotherians; (2) infer changes in cancer susceptibility across Afrotherian lineages; (3) use a genome-wide screen to identify gene duplications in Afrotherian genomes, including multiple living and extinct Proboscideans; and (4) show that while duplication of genes with tumor suppressor functions is pervasive in Afrotherian genomes, Proboscidean gene duplicates are enriched in unique pathways with tumor suppressor functions.

Correlated evolution of large bodies and reduced cancer risk

The hundred- to hundred-million-fold reductions in intrinsic cancer risk associated with the evolution of large body sizes in some Afrotherian lineages, in particular Elephantimorphs such as elephants and mastodons, suggests that these lineages must have also evolved remarkable mechanisms to suppress cancer. While our initial hypothesis was that large-bodied lineages would be uniquely enriched in duplicate tumor suppressor genes compared to other smaller-bodied lineages, we unexpectedly found that the duplication of genes in tumor suppressor pathways occurred at various points throughout the evolution of Afrotheria, regardless of body size. These data suggest that this abundance of tumor suppressors may have contributed to the evolution of large bodies and reduced cancer risk, but that these processes were not necessarily coincident. Interestingly, pervasive duplication of tumor suppressors may also have contributed to the repeated evolution of large bodies in hyraxes and sea cows, because at least some of the genetic changes that underlie the evolution of reduced cancer risk were common in this group. It remains to be determined whether our observation of pervasive duplication of tumor suppressors also occurs in other multicellular lineages. Using a similar reciprocal best BLAST/BLAT approach that focused on estimating copy number of known tumor suppressors in mammalian genomes, for example, **Caulin et al., 2015** found no correlation between copy number or tumor suppressors with either body mass or longevity, whereas **Tollis et al., 2020** found a correlation between copy number and longevity (but not body size) (**Tollis et al., 2020**; **Caulin et al., 2015**). These opposing conclusions may result from differences in the number of genes (81 vs 548) and genomes (8 vs 63) analyzed, highlighting the need for genome-wide analyses of many species that vary in body size and longevity.

There's no such thing as a free lunch: Trade-offs and constraints on tumor suppressor copy number

While we observed that duplication of genes in cancer related pathways – including genes with known tumor suppressor functions – is pervasive in Afrotheria, the number of duplicate tumor suppressor genes was relatively small, which may reflect a trade-off between the protective effects of increased tumor suppressor number on cancer risk and potentially deleterious consequences of increased tumor suppressor copy number. Overexpression of *TP53* in mice, for example, is protective against cancer but associated with progeria, premature reproductive senescence, and early death; however, transgenic mice with a duplication of the *TP53* locus that includes native regulatory elements are healthy and experience normal aging, while also demonstrating an enhanced response to cellular stress and lower rates of cancer (**García-Cao et al., 2002**; **Tyner et al., 2002**). These data suggest that duplication of tumor suppressors can contribute to augmented cancer resistance, if the duplication includes sufficient regulatory architecture to direct spatially and temporally appropriate gene expression. Thus, it is interesting that duplication of genes that regulate TP53 function, such as *STK11*, *SIAH1*, and *BRD7*, preceded the retroduplication *TP53* in the Proboscidean stem-lineage, which may have mitigated toxicity arising from dosage imbalances. Similar co-duplication events may have alleviated the negative pleiotropy of tumor suppressor gene duplications to enable their persistence and allow for subsequent co-option during the evolution of cancer resistance.

Caveats and limitations

Our genome-wide results suggest that duplication of tumor suppressors is pervasive in Afrotherians and may have enabled the evolution of larger body sizes in multiple lineages by lowering intrinsic cancer risk either prior to or coincident with increasing body size. However, our study has several inherent limitations. For example, we have shown that genome quality plays an important role in our ability to identify duplicate genes, and several species have poor quality genomes (and thus were excluded from further analyses). While several efforts have been established with the goal of generating high quality (chromosome length) reference genomes for mammals, such as DNAZoo, The Zoonomia Project, the Vertebrate Genomes Project, and Genome 10K, Atlantogenatans represent a minority of available genome projects. And while a few high quality Atlantogenatan genomes are available, they lack reference gene and transcriptome annotations, and genome browser graphical user interfaces that allow for easy access to genome data for the broader community, limiting their usefulness. Similarly, without comprehensive gene expression data we cannot be certain that duplicate genes are actually expressed, and thus functional. Our results on genome quality suggest several research priorities for these less well-studies species, including generating chromosome length reference genomes and genome annotations, and incorporating these species into existing genome browsers (such as UCSC Genome Browser).

We also assume that gene duplicates either maintain ancestral tumor suppressor functions and increase cancer resistance through dosage effects or provide redundancy to loss of function mutations thereby increasing robustness of tumor suppression. Many processes, such as developmental systems drift, neofunctionalization, and sub-functionalization, can cause divergence in gene functions and invalidate the assumption of conservation of gene function (**Rastogi and Liberles, 2005**; **Qian and Zhang, 2014**; **Stoltzfus, 1999**), leading to inaccurate inferences in gene and pathway functions which is a common problem in comparative genomic studies using pathway and gene ontologies to categorize gene function. In addition, we assume that most duplicate genes are functional but it is likely that some of the duplicates were identify are non-functional pseudogenes. Differentiating between functional and non-functional genes using comparative genomics can be challenging. For example, non-functional pseudogenes often accumulate non-synonymous amino acid substitutions and premature stop codons but these same changes can also occur in functional genes. For example, we have found that the elephant genome encodes *TP53* retogenes (*TP53RTGs*) all of which encode premature stop codons suggesting they are pseudogenes, but these *TP53TRGs* are expressed, encode functional separation of function mutants of the ancestral TP53 gene, and contribute to enhanced DNA damage sensitivity in elephant cells. Similarly, we have characterized duplicate *LIF* gene in elephants (*LIF6*) that lacks the start codon and exon 1 of the parent LIF gene. *LIF6* is expressed, encodes a functional protein with translation initiated at an alternative downstream start site, and also contributes to enhanced DNA damage sensitivity in elephant cells. In

addition, duplicate genes that lack coding potential, such as *PTENP1*, can also be expressed and while not translated function as LINC RNAs (in this case acting as a sponge for microRNAs that target the parent PTEN transcript). In each case classifying duplicates into putatively functional and non-functional categories based on sequence characteristic would misclassify *TP53RTG*s, *LIF6*, and *PTENP1*. Thus, sequence features of pseudogenes may maintain function, as a consequence of not excluding putative pseudogenes some of the genes we include in downstream analyses may be non-functional. Further experimental studies are needed to determine which duplicates are expressed and functional.

The focus of this study, motivated by our previous identification of *TP53* and *LIF* duplicates, was on the role gene duplication in general may have played in the resolution of Peto's paradox in large-bodied Afrotherians, particularly Proboscidea. Duplication of tumor suppressor genes, however, is unlikely to be the sole mechanism responsible for the evolution of large body sizes, long lifespans, and reduced cancer risk. The evolution of regulatory elements, coding genes, genes with non-canonical tumor suppressor functions, and immune cell recognition of cancerous cells are also likely important for reducing the risk of cancer.

Conclusions: All Afrotherians are equal, but some are more equal than others

While we found that duplication of tumor suppressor genes is common in Afrotheria, genes that duplicated in the Proboscidean stem-lineage (***Figure 3A,B***) were uniquely enriched in functions and pathways that may be related to the evolution of unique anti-cancer cellular phenotypes in the elephant lineage (***Figure 3C***). Elephant cells, for example, cannot be experimentally immortalized (***Fukuda et al., 2016***; ***Gomes et al., 2011***), rapidly repair DNA damage (***Sulak et al., 2016***; ***Hart and Setlow, 1974***; ***Francis et al., 1981***), are extremely resistant to oxidative stress (***Gomes et al., 2011***), and yet are also extremely sensitive to DNA damage (***Abegglen et al., 2015***; ***Sulak et al., 2016***; ***Vazquez et al., 2018***). Several pathways related to DNA damage repair, in particular nucleotide excision repair (NER), were uniquely enriched among genes that duplicated in the Proboscidean stem-lineage, suggesting a connection between duplication of genes involved in NER and rapid DNA damage repair (***Hart and Setlow, 1974***; ***Francis et al., 1981***). Similarly, we identified a duplicate *SOD1* gene in Proboscideans that may confer the resistance of elephant cells to oxidative stress (***Gomes et al., 2011***). Pathways related to the cell cycle were also enriched among genes that duplicated in Proboscideans, and cell cycle dynamics are different in elephants compared to other species; population doubling (PD) times for African and Asian elephant cells are 13–16 days, while PD times are 21–28 days in other Afrotherians (***Gomes et al., 2011***). Finally, the role of 'mTOR signaling' in the biology of aging is well known. Collectively these data suggest that gene duplications in Proboscideans may underlie some of their cellular phenotypes that contribute to cancer resistance.

Materials and methods

Ancestral body size reconstruction

We first assembled a time-calibrated supertree of Eutherian mammals by combining the time-calibrated molecular phylogeny of ***Bininda-Emonds et al., 2007***; ***Bininda-Emonds et al., 2008*** with the time-calibrated total evidence Afrotherian phylogeny from ***Puttick and Thomas, 2015***. While the ***Bininda-Emonds et al., 2007***; ***Bininda-Emonds et al., 2008*** phylogeny includes 1679 species, only 34 are Afrotherian, and no fossil data are included. The inclusion of fossil data from extinct species is essential to ensure that ancestral state reconstructions of body mass are not biased by only including extant species. This can lead to inaccurate reconstructions, for example, if lineages convergently evolved large body masses from a small-bodied ancestor. In contrast, the total evidence Afrotherian phylogeny of ***Puttick and Thomas, 2015*** includes 77 extant species and fossil data from 39 extinct species. Therefore, we replaced the Afrotherian clade in the ***Bininda-Emonds et al., 2008*** phylogeny with the Afrotherian phylogeny of ***Puttick and Thomas, 2015*** using Mesquite. Next, we jointly estimated rates of body mass evolution and reconstructed ancestral states using a generalization of the Brownian motion model that relaxes assumptions of neutrality and gradualism by considering increments to evolving characters to be drawn from a heavy-tailed stable distribution (the 'Stable Model')

implemented in StableTraits (**Elliot and Mooers, 2014**). The stable model allows for large jumps in traits and has previously been shown to outperform other models of body mass evolution, including standard Brownian motion models, Ornstein–Uhlenbeck models, early burst maximum likelihood models, and heterogeneous multi-rate models (**Elliot and Mooers, 2014**).

Reciprocal Best Hit BLAT

We developed a reciprocal best hit BLAT (RBHB) pipeline to identify putative homologs and estimate gene copy number across species. The Reciprocal Best Hit (RBH) search strategy is conceptually straightforward: (1) Given a gene of interest G_A in a query genome A, one searches a target genome B for all possible matches to G_A; (2) For each of these hits, one then performs the reciprocal search in the original query genome to identify the highest-scoring hit; (3) A hit in genome B is defined as a homolog of gene G_A if and only if the original gene G_A is the top reciprocal search hit in genome A. We selected BLAT (**Kent, 2002**) as our algorithm of choice, as this algorithm is sensitive to highly similar (>90% identity) sequences, thus identifying the highest-confidence homologs while minimizing many-to-one mapping problems when searching for multiple genes. RBH performs similar to other more complex methods of orthology prediction and is particularly good at identifying incomplete genes that may be fragmented in low quality/poorly assembled regions of the genome (**Altenhoff and Dessimoz, 2009**; **Salichos and Rokas, 2011**).

Effective copy number by coverage

In low-quality genomes, many genes are fragmented across multiple scaffolds, which results in BLA(S)T-like methods calling multiple hits when in reality there is only one gene. To compensate for this, we developed a novel statistic, Estimated Copy Number by Coverage (ECNC), which averages the number of times we hit each nucleotide of a query sequence in a target genome over the total number of nucleotides of the query sequence found overall in each target genome (**Figure 3—figure supplement 1**). This allows us to correct for genes that have been fragmented across incomplete genomes, while accounting for missing sequences from the human query in the target genome. Mathematically, this can be written as:

$$\mathrm{ECNC} = \frac{\sum_{n=1}^{l} C_n}{\sum_{n=1}^{l} \mathrm{bool}(C_n)} \quad (1)$$

where n is the given nucleotide in the query, l is the total length of the query, C_n is the number of instances that n is present within a reciprocal best hit, and bool (C_n) is 1 if C_n >1 C_n>0 or 0 if C_n =1 $C_n = 0$.

RecSearch pipeline

We created a custom Python pipeline for automating RBHB searches between a single reference genome and multiple target genomes using a list of query sequences from the reference genome. For the query sequences in our search, we used the hg38 UniProt proteome (**The UniProt Consortium, 2017**), which is a comprehensive set of protein sequences curated from a combination of predicted and validated protein sequences generated by the UniProt Consortium. Next, we excluded genes from downstream analyses for which assignment of homology was uncertain, including uncharacterized ORFs (991 genes), LOC (63 genes), HLA genes (402 genes), replication dependent histones (72 genes), odorant receptors (499 genes), ribosomal proteins (410 genes), zinc finger transcription factors (1983 genes), viral and repetitive-element-associated proteins (82 genes), and 'Uncharacterized', 'Putative', or 'Fragment' proteins (30,724 genes), leaving a final set of 37,582 query protein isoforms, corresponding to 18,011 genes. We then searched for all copies of 18,011 query genes in publicly available Afrotherian genomes (**Dobson, 2013**), including African savannah elephant (*Loxodonta africana*: loxAfr3, loxAfr4, loxAfrC), African forest elephant (*Loxodonta cyclotis*: loxCycF), Asian Elephant (*Elephas maximus*: eleMaxD), Woolly Mammoth (*Mammuthus primigenius*: mamPriV), Colombian mammoth (*Mammuthus columbi*: mamColU), American mastodon (*Mammut americanum*: mamAmeI), Rock Hyrax (*Procavia capensis*: proCap1, proCap2, proCap2*HiC*), West Indian Manatee (*Trichechus manatus latirostris:* triManLat1, triManLat1HiC), Aardvark (*Orycteropus afer*: oryAfe1, oryAfe1*HiC*), Lesser Hedgehog Tenrec (*Echinops telfairi:* echTel2), Nine-banded armadillo (*Dasypus novemcinctus:* dasNov3), Hoffman's two-toed sloth (*Choloepus hoffmannii*: choHof1,

choHof2, choHof2HiC), Cape golden mole (*Chrysochloris asiatica*: chrAsi1), and Cape elephant shrew (*Elephantulus edwardii*: eleEdw1) (**Dudchenko et al., 2017**; **Palkopoulou et al., 2015**; **Palkopoulou et al., 2018**; **Foote et al., 2015**).

A summary of gene duplications in each species is available in **Supplementary file 1**.

Duplication gene inclusion criteria

In order to condense transcript-level hits into single gene loci, and to resolve many-to-one genome mappings, we removed exons where transcripts from different genes overlapped, and merged overlapping transcripts of the same gene into a single gene locus call. The resulting gene-level copy number table was then combined with the maximum ECNC values observed for each gene in order to call gene duplications. We called a gene duplicated if its copy number was two or more, and if the maximum ECNC value of all the gene transcripts searched was 1.5 or greater; previous studies have shown that incomplete duplications can encode functional genes (**Sulak et al., 2016**; **Vazquez et al., 2018**), therefore partial gene duplications were included provided they passed additional inclusion criteria (see below). The ECNC cut-off of 1.5 was selected empirically, as this value minimized the number of false positives seen in a test set of genes and genomes. The results of our initial search are summarized in **Figure 3A**. Overall, we identified 13,880 genes across all species, or 77.1% of our starting query genes.

Genome quality assessment using CEGMA

In order to determine the effect of genome quality on our results, we used the gVolante webserver and CEGMA to assess the quality and completeness of the genome (**Nishimura et al., 2017**; **Parra et al., 2009**). CEGMA was run using the default settings for mammals ('Cut-off length for sequence statistics and composition'=1; 'CEGMA max intron length'=100,000; 'CEGMA gene flanks'=10,000, 'Selected reference gene set' = CVG). For each genome, we generated a correlation matrix using the aforementioned genome quality scores, and either the mean copy number or mean ECNC for all hits in the genome. We observed that the percentage of duplicated genes in non-Pseudoungulatan genomes was higher (12.94–23.66%) than Pseudoungulatan genomes (3.26–7.80%). Mean copy number, mean ECNC, and mean CN (the lesser of copy number and ECNC per gene) moderately or strongly correlated with genomic quality, such as LD50, the number of scaffolds, and contigs with a length above either 100K or 1M (**Figure 3—figure supplement 2**). The Afrosoricidians had the greatest correlation between poor genome quality and high gene duplication rates, including larger numbers of private duplications. The correlations between genome quality metric and number of gene duplications were particularly high for Cape golden mole (*Chrysochloris asiatica*: chrAsi1) and Cape elephant shrew (*Elephantulus edwardii*: eleEdw1); therefore we excluded these species from downstream pathway enrichment analyses.

Determining functionality of duplicated via gene expression

In order to ascertain the functional status of duplicated genes, we generated de novo transcriptomes using publicly available RNA-sequencing data for African savanna elephant, West Indian manatee, and nine-banded armadillo (**Supplementary file 2**). We mapped reads to the highest quality genome available for each species, and assembled transcripts using HISAT2 and StringTie (**Kim et al., 2015**; **Pertea et al., 2015**; **Pertea et al., 2016**). We found that many of our identified duplicates had transcripts mapping to them above a Transcripts Per Million (TPM) score of 2, suggesting that many of these duplications are functional. RNA-sequencing data was not available for Cape golden mole, Cape elephant shrew, rock hyrax, aardvark, or the lesser hedgehog tenrec.

Reconstruction of ancestral copy numbers

We encoded the copy number of each gene for each species as a discrete trait ranging from 0 (one gene copy) to 31 (for 32+ gene copies) and used IQ-TREE to select the best-fitting model of character evolution (**Minh et al., 2020**; **Hoang et al., 2018**; **Kalyaanamoorthy et al., 2017**; **Wang et al., 2018**; **Schrempf et al., 2019**), which was inferred to be a Jukes-Cantor type model for morphological data (MK) with equal character state frequencies (FQ) and rate heterogeneity across sites approximated by including a class of invariable sites (I) plus a discrete Gamma model with four rate categories (G4). Next we inferred gene duplication and loss events with the empirical Bayesian

ancestral state reconstruction (ASR) method implemented in IQ-TREE (**Minh et al., 2020**; **Hoang et al., 2018**; **Kalyaanamoorthy et al., 2017**; **Wang et al., 2018**; **Schrempf et al., 2019**), the best fitting model of character evolution (MK+FQ+GR+I) (**Soubrier et al., 2012**; **Yang et al., 1995**), and the unrooted species tree for Atlantogenata. We considered ancestral state reconstructions to be reliable if they had Bayesian Posterior Probability (BPP) $\geq$ 0.80; less reliable reconstructions were excluded from pathway analyses. We note that there may be 'ghost' duplication events, that is genes that duplicated in, for example, the Tethytherian stem-lineage that are maintained in the Stellar's sea cow genome and lost in the manatee genome. These genes will be reconstructed as a Proboscidean-specific duplication events because we cannot determine copy number in extinct species that lack genomes.

Pathway enrichment analysis

To determine if gene duplications were enriched in particular biological pathways, we used the WEB-based Gene SeT AnaLysis Toolkit (WebGestalt) (**Liao et al., 2019**) to perform Over-Representation Analysis (ORA) using the Reactome database (**Jassal et al., 2020**). Gene duplicates in each lineage were used as the foreground gene set, and the initial query set was used as the background gene set. WebGestalt uses a hypergeometric test for statistical significance of pathway over-representation, which we refined using two methods: a False Discovery Rate (FDR)-based approach and an empirical p-value approach (**Chen et al., 2013**). The Benjamini–Hochberg FDR multiple-testing correction was generated by WebGestalt. In order to correct p-values based on an empirical distribution, we modified the approach used by Chen et al. in Enrichr (**Chen et al., 2013**) to generate a 'combined score' for each pathway based on the hypergeometric p-value from WebGestalt, and a correction for expected rank for each pathway. In order to generate the table of expected ranks and variances for this approach, we randomly sampled foreground sets of 10–5000 genes from our background set 5000 times, and used WebGestalt ORA to obtain a list of enriched terms and P-values for each run; we then compiled a table of Reactome terms with their expected frequencies and standard deviation. These data were used to calculate a Z-score for terms in an ORA run, and the combined score was calculated using the formula $C = log(p) \cdot z$.

Estimating the evolution of cancer risk

The dramatic increase in body mass and lifespan in some *Afrotherian* lineages, and the relatively constant rate of cancer across species of diverse body sizes (**Abegglen et al., 2015**), indicates that those lineages must have also evolved reduced cancer risk. To infer the magnitude of these reductions we estimated differences in intrinsic cancer risk across extant and ancestral Afrotherians. Following **Peto, 2015**, we estimate the intrinsic cancer risk (*K*) as the product of risk associated with body mass and lifespan. In order to determine (*K*) across species and at ancestral nodes (see below), we first estimated ancestral lifespans at each node. We used Phylogenetic Generalized Least-Square Regression (PGLS) (**Felsenstein, 1985**; **Martins and Hansen, 1997**), using a Brownian covariance matrix as implemented in the R package *ape* (**Paradis and Schliep, 2019**), to calculate estimated ancestral lifespans across Atlantogenata using our estimates for body size at each node. In order to estimate the intrinsic cancer risk of a species, we first inferred lifespans at ancestral nodes using PGLS (**Supplementary file 3**) and the model. Next, we calculated K_1 at all nodes, and then estimated the fold-change in cancer susceptibility between ancestral and descendant nodes (**Figure 2**). Next, in order to calculate K_1 at all nodes, we used a simplified multistage cancer risk model for body size D and lifespan t: $K \approx Dt^6$ (**Peto et al., 1975**: **Peto, 2015**; **Armitage, 1985**; **Armitage and Doll, 2004**). The fold change in cancer risk between a node and its ancestor was then defined as $\log_2\left(\frac{K_2}{K_1}\right)$.

Data analysis

All data analysis was performed using Python version 3.8 and R version 4.0.2 (2020-06-22), and the complete reproducible manuscript, along with code and data generation pipeline, can be found on our GitHub page at https://github.com/docmanny/atlantogenataGeneDuplication (**Vazquez and Lynch, 2021**; copy archived at swh:1:rev:6bc68ac31ef148131480710e50b0b75d06077db2; **Paradis and Schliep, 2019**; **Paradis et al., 2020**; **R Development Core Team, 2019**; **Xie, 2020**; **Bolker and Robinson, 2020**; **Dowle and Srinivasan, 2019**; **Wickham et al., 2020a**; **Wickham, 2020**;

Harmon et al., 2020; *Yutani, 2020*; *Yu, 2020a*; *Campitelli, 2020*; *Wickham et al., 2020b*; *Yu, 2020b*; *Kassambara, 2020*; *Slowikowski, 2020*; *Xiao, 2018*; *Yu and Lam, 2020c*; *Zhu, 2019*; *Ooms, 2020*; *Bache and Wickham, 2014*; *Pinheiro and Bates, 2020*; *Sievert et al., 2020*; *Henry and Wickham, 2020*; *Wickham et al., 2018*; *Hlavac, 2018*; *Wickham, 2019a*; *Müller and Wickham, 2020*; *Wickham and Henry, 2020*; *Yu, 2020d*; *Wickham, 2019b*; *Yu, 2020e*; *Gehlenborg, 2019*; *Xie, 2016*; *Alfaro et al., 2009*; *Eastman et al., 2011*; *Slater et al., 2012*; *Harmon et al., 2008*; *Pennell et al., 2014*; *Wickham, 2016*; *Yu, 2020f*; *Yu et al., 2018*; *Yu et al., 2017*; *Sievert, 2020*; *Wickham et al., 2019*; *Wang et al., 2020*). All files necessary to reproduce the data in this manuscript are provided in *Source data 1*.

Manual verification of duplicate genes

We manually verified the coding potential of the 16 genes shown in *Figure 4* by first identifying the reciprocal best (DNA sequence) BLAT hits in the elephant and manatee genomes, which allowed us to determine conservation and presence of premature stop codons in the each open reading frame (ORF). We translated the ORF for each hit into amino acid sequences and grouped up hits for each gene into one FASTA file along with the UniProt protein sequences for the human, dog, cat, and cow orthologs. Using a pipeline hosted at NGPhlyogeny.fr (*Lemoine et al., 2019*), the homologs were aligned using MAFFT *Katoh and Standley, 2013*; the aligned sequences were cleaned using BMGE (*Criscuolo and Gribaldo, 2010*). Finally we used FastME (*Lefort et al., 2015*) to infer a gene tree for each duplicate. Alignments were then visually inspected for conservation and presence of premature stop codons.

Acknowledgements

We would like to thank Dr. Olga Dudchenko and Dr. Erez Aiden at Baylor College of Medicine for the Hi-C scaffolded *Procavia capensis*, *Trichechus manatus*, *Orycteropus afer*, and *Choloepus hoffmannii* genomes. We would also like to thank D.H. Vazquez for his indispensable support.

Additional information

Funding

Funder	Author
University of Chicago	Juan M Vazquez Vincent J Lynch

The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

Author contributions

Juan M Vazquez, Conceptualization, Resources, Data curation, Software, Formal analysis, Validation, Investigation, Visualization, Methodology, Writing - original draft, Writing - review and editing; Vincent J Lynch, Conceptualization, Data curation, Formal analysis, Supervision, Investigation, Visualization, Writing - original draft, Project administration, Writing - review and editing

Author ORCIDs

Juan M Vazquez https://orcid.org/0000-0001-8341-2390
Vincent J Lynch https://orcid.org/0000-0001-5311-3824

Decision letter and Author response

Decision letter https://doi.org/10.7554/eLife.65041.sa1
Author response https://doi.org/10.7554/eLife.65041.sa2

Additional files

Supplementary files

- Source data 1. All necessary data sets and scripts to reproduce results presented in this manuscript.
- Supplementary file 1. Summary of duplications in Atlantogenata.
- Supplementary file 2. RNA-Seq data sets used in this study, along with key biological and genome information.
- Supplementary file 3. Summary of PGLS model used to estimate lifespan.
- Transparent reporting form

Data availability

All data generated or analysed during this study are included in the manuscript and supporting files.

The following previously published datasets were used:

Author(s)	Year	Dataset title	Dataset URL	Database and Identifier
Di Palma F, Alfoldi J, Johnson J, Berlin A, Gnerre S, Jaffe D, MacCallum I, Young S, Walker BJ, Lindblad-Toh K	2013	Chrysochloris asiatica (Cape golden mole) genome	https://www.ncbi.nlm.nih.gov/assembly/GCF_000296735.1/	NCBI Assembly, ChrAsi1.0
Gnerre S, Heiman D, Young S, Fulton L, Delehaunty K, Minx P, Chinwalla A, Mardis E, Wilson R, Warren W	2018	The Genome Sequence of Choloepus hoffmanni (sloth)	https://www.dnazoo.org/assemblies/Choloepus_hoffmanni	NCBI Assembly, C_hoffmanni-2.0.1_HiC
Lindblad-Toh K, Chang JL, Gnerre S, Clamp M, Lander ES	2012	Dasypus novemcinctus (nine-banded armadillo) genome	https://www.ncbi.nlm.nih.gov/assembly/GCA_000208655.2/	NCBI Assembly, Dasnov3.0
Di Palma F, Alfoldi J, Johnson J, Berlin A, Gnerre S, Jaffe D, MacCallum I, Young S, Walker BJ, Lindblad-Toh K	2013	Echinops telfairi (small Madagascar hedgehog)	https://www.ncbi.nlm.nih.gov/assembly/GCA_000313985.1/	NCBI Assembly, EchTel2.0
Di Palma F, Alfoldi J, Johnson J, Berlin A, Gnerre S, Jaffe D, MacCallum I, Young S, Walker BJ, Lindblad-Toh K	2013	Elephantulus edwardii (Cape elephant shrew)	https://www.ncbi.nlm.nih.gov/assembly/GCA_000299155.1/	NCBI Assembly, EleEdw1.0
Palkopoulou E, Lipson M, Mallick S, Nielsen S, Rohland N, Baleka S, Karpinski E, Ivancevic AM, To TH, Kortschak RD, Raison JM, Qu Z, Chin TJ, Alt KW, Claesson S, Dalén L, MacPhee RDE, Meller H, Roca AL, Ryder OA, Heiman D, Young S, Breen M, Williams C, Aken BL, Ruffier M, Karlsson E, Johnson J, Di Palma F,	2018	A comprehensive genomic history of extinct and living elephants	https://www.ebi.ac.uk/ena/browser/view/PRJEB24361	NCBI Assembly, PRJEB24361

Alfoldi J, Adelson DL, Mailund T, Munch K, Lindblad-Toh K, Hofreiter M, Poinar H, Reich D				
Di Palma F, Alfoldi J, Johnson J, Berlin A, Gnerre S, Jaffe D, MacCallum I, Young S, Walker BJ, Lindblad-Toh K	2017	Aardvark (Orycteropus afer) genome	https://www.dnazoo.org/assemblies/Orycteropus_afer	NCBI Assembly, oryAfe2
Di Palma F, Alfoldi J, Johnson J, Berlin A, Gnerre S, Jaffe D, MacCallum I, Young S, Walker BJ, Lindblad-Toh K	2013	A comprehensive genomic history of extinct and living elephants	https://www.dnazoo.org/assemblies/Procavia_capensis	NCBI Assembly, proCap-Pcap_2.0_HiC
Andrew DF, Liu Y, Thomas GWC, Vinar T, Alfoldi J, Deng J, Dugan S	2015	West Indian manatee (Trichechus manatus) genome	https://www.dnazoo.org/assemblies/Trichechus_manatus	NCBI Assembly, triManLat2

References

Abegglen LM, Caulin AF, Chan A, Lee K, Robinson R, Campbell MS, Kiso WK, Schmitt DL, Waddell PJ, Bhaskara S, Jensen ST, Maley CC, Schiffman JD. 2015. Potential mechanisms for Cancer resistance in elephants and comparative cellular response to DNA damage in humans. *Jama* **314**:1850–1860. DOI: https://doi.org/10.1001/jama.2015.13134, PMID: 26447779

Alfaro ME, Santini F, Brock C, Alamillo H, Dornburg A, Rabosky DL, Carnevale G, Harmon LJ. 2009. Nine exceptional radiations plus high turnover explain species diversity in jawed vertebrates. *PNAS* **106**:13410–13414. DOI: https://doi.org/10.1073/pnas.0811087106, PMID: 19633192

Altenhoff AM, Dessimoz C. 2009. Phylogenetic and functional assessment of orthologs inference projects and methods. *PLOS Computational Biology* **5**:e1000262. DOI: https://doi.org/10.1371/journal.pcbi.1000262, PMID: 19148271

Armitage P. 1985. Multistage models of carcinogenesis. *Environmental Health Perspectives* **63**:195–201. DOI: https://doi.org/10.1289/ehp.8563195, PMID: 3908088

Armitage P, Doll R. 2004. The age distribution of Cancer and a multi-stage theory of carcinogenesis. *British Journal of Cancer* **91**:1983–1989. DOI: https://doi.org/10.1038/sj.bjc.6602297, PMID: 15599380

Ashur-Fabian O, Avivi A, Trakhtenbrot L, Adamsky K, Cohen M, Kajakaro G, Joel A, Amariglio N, Nevo E, Rechavi G. 2004. Evolution of p53 in hypoxia-stressed Spalax mimics human tumor mutation. *PNAS* **101**:12236–12241. DOI: https://doi.org/10.1073/pnas.0404998101, PMID: 15302922

Bache SM, Wickham H. 2014. *Magrittr: A Forward-Pipe Operator for R*. https://CRAN.R-project.org/package=magrittr

Bininda-Emonds OR, Cardillo M, Jones KE, MacPhee RD, Beck RM, Grenyer R, Price SA, Vos RA, Gittleman JL, Purvis A. 2007. The delayed rise of present-day mammals. *Nature* **446**:507–512. DOI: https://doi.org/10.1038/nature05634, PMID: 17392779

Bininda-Emonds ORP, Cardillo M, Jones KE, MacPhee RDE, Beck RMD, Grenyer R, Price SA, Vos RA, Gittleman JL, Purvis A. 2008. Erratum: the delayed rise of present-day mammals. *Nature* **456**:274. DOI: https://doi.org/10.1038/nature07347

Boddy AM, Abegglen LM, Pessier AP, Aktipis A, Schiffman JD, Maley CC, Witte C. 2020. Lifetime Cancer prevalence and life history traits in mammals. *Evolution, Medicine, and Public Health* **2020**:187–195. DOI: https://doi.org/10.1093/emph/eoaa015, PMID: 33209304

Bolker B, Robinson D. 2020. *Broom.mixed: Tidying Methods for Mixed Models*. https://CRAN.R-project.org/package=broom.mixed

Burrows AE, Smogorzewska A, Elledge SJ. 2010. Polybromo-associated BRG1-associated factor components BRD7 and BAF180 are critical regulators of p53 required for induction of replicative senescence. *PNAS* **107**:14280–14285. DOI: https://doi.org/10.1073/pnas.1009559107, PMID: 20660729

Campitelli E. 2020. *Ggnewscale: Multiple Fill and Colour Scales In' Ggplot2'*. https://CRAN.R-project.org/package=ggnewscale

Caulin AF, Graham TA, Wang LS, Maley CC. 2015. Solutions to Peto's paradox revealed by mathematical modelling and cross-species cancer gene analysis. *Philosophical Transactions of the Royal Society B: Biological Sciences* **370**:20140222. DOI: https://doi.org/10.1098/rstb.2014.0222, PMID: 26056366

Caulin AF, Maley CC. 2011. Peto's Paradox: evolution's prescription for cancer prevention. *Trends in Ecology & Evolution* **26**:175–182. DOI: https://doi.org/10.1016/j.tree.2011.01.002, PMID: 21296451

Chen EY, Tan CM, Kou Y, Duan Q, Wang Z, Meirelles GV, Clark NR, Ma'ayan A. 2013. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics* **14**:128. DOI: https://doi.org/10.1186/1471-2105-14-128, PMID: 23586463

Criscuolo A, Gribaldo S. 2010. BMGE (Block mapping and gathering with entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. *BMC Evolutionary Biology* **10**:210. DOI: https://doi.org/10.1186/1471-2148-10-210, PMID: 20626897

Dobson JM. 2013. Breed-Predispositions to Cancer in pedigree dogs. *ISRN Veterinary Science* **2013**:1–23. DOI: https://doi.org/10.1155/2013/941275

Doherty A, de Magalhães JP. 2016. Has gene duplication impacted the evolution of eutherian longevity? *Aging Cell* **15**:978–980. DOI: https://doi.org/10.1111/acel.12503, PMID: 27378378

Dorn CR, Taylor DO, Schneider R, Hibbard HH, Klauber MR. 1968. Survey of animal neoplasms in alameda and contra costa counties, California. II. Cancer morbidity in dogs and cats from alameda county. *Journal of the National Cancer Institute* **40**:307–318. PMID: 5694272

Dowle M, Srinivasan A. 2019. *Data.table: Extension of 'Data.frame*. https://CRAN.R-project.org/package=data.table

Drost J, Mantovani F, Tocco F, Elkon R, Comel A, Holstege H, Kerkhoven R, Jonkers J, Voorhoeve PM, Agami R, Del Sal G. 2010. BRD7 is a candidate tumour suppressor gene required for p53 function. *Nature Cell Biology* **12**:380–389. DOI: https://doi.org/10.1038/ncb2038, PMID: 20228809

Dudchenko O, Batra SS, Omer AD, Nyquist SK, Hoeger M, Durand NC, Shamim MS, Machol I, Lander ES, Aiden AP, Aiden EL. 2017. De novo assembly of the *aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science* **356**:92–95. DOI: https://doi.org/10.1126/science.aal3327, PMID: 28336562

Eastman JM, Alfaro ME, Joyce P, Hipp AL, Harmon LJ. 2011. A novel comparative method for identifying shifts in the rate of character evolution on trees. *Evolution* **65**:3578–3589. DOI: https://doi.org/10.1111/j.1558-5646.2011.01401.x

Elliot MG, Mooers AØ. 2014. Inferring ancestral states without assuming neutrality or gradualism using a stable model of continuous character evolution. *BMC Evolutionary Biology* **14**:226. DOI: https://doi.org/10.1186/s12862-014-0226-8, PMID: 25427971

Felsenstein J. 1985. Phylogenies and the comparative method. *The American Naturalist* **125**:1–15. DOI: https://doi.org/10.1086/284325

Foote AD, Liu Y, Thomas GW, Vinař T, Alföldi J, Deng J, Dugan S, van Elk CE, Hunter ME, Joshi V, Khan Z, Kovar C, Lee SL, Lindblad-Toh K, Mancia A, Nielsen R, Qin X, Qu J, Raney BJ, Vijay N, et al. 2015. Convergent evolution of the genomes of marine mammals. *Nature Genetics* **47**:272–275. DOI: https://doi.org/10.1038/ng.3198, PMID: 25621460

Francis AA, Lee WH, Regan JD. 1981. The relationship of DNA excision repair of ultraviolet-induced lesions to the maximum life span of mammals. *Mechanisms of Ageing and Development* **16**:181–189. DOI: https://doi.org/10.1016/0047-6374(81)90094-4, PMID: 7266079

Fukuda T, Iino Y, Onuma M, Gen B, Inoue-Murayama M, Kiyono T. 2016. Expression of human cell cycle regulators in the primary cell line of the african savannah elephant (loxodonta africana) increases proliferation until senescence, but does not induce immortalization. *In Vitro Cellular & Developmental Biology - Animal* **52**:20–26. DOI: https://doi.org/10.1007/s11626-015-9943-6, PMID: 26487427

García-Cao I, García-Cao M, Martín-Caballero J, Criado LM, Klatt P, Flores JM, Weill JC, Blasco MA, Serrano M. 2002. "Super p53" mice exhibit enhanced DNA damage response, are tumor resistant and age normally. *The EMBO Journal* **21**:6225–6235. DOI: https://doi.org/10.1093/emboj/cdf595, PMID: 12426394

Gehlenborg N. 2019. *UpSetR: A More Scalable Alternative to Venn and Euler Diagrams for Visualizing Intersecting Sets*. https://CRAN.R-project.org/package=UpSetR

Gomes NM, Ryder OA, Houck ML, Charter SJ, Walker W, Forsyth NR, Austad SN, Venditti C, Pagel M, Shay JW, Wright WE. 2011. Comparative biology of mammalian telomeres: hypotheses on ancestral states and the roles of telomeres in longevity determination. *Aging Cell* **10**:761–768. DOI: https://doi.org/10.1111/j.1474-9726.2011.00718.x, PMID: 21518243

Gorbunova V, Hine C, Tian X, Ablaeva J, Gudkov AV, Nevo E, Seluanov A. 2012. Cancer resistance in the blind mole rat is mediated by concerted necrotic cell death mechanism. *PNAS* **109**:19392–19396. DOI: https://doi.org/10.1073/pnas.1217211109, PMID: 23129611

Harmon LJ, Weir JT, Brock CD, Glor RE, Challenger W. 2008. GEIGER: investigating evolutionary radiations. *Bioinformatics* **24**:129–131. DOI: https://doi.org/10.1093/bioinformatics/btm538, PMID: 18006550

Harmon L, Pennell M, Brock C, Brown J, Challenger W, Eastman J, FitzJohn R, Glor R, Hunt G, Revell L, Slater G, Uyeda J, Weir J. 2020. *Geiger: Analysis of Evolutionary Diversification*. https://CRAN.R-project.org/package=geiger

Hart RW, Setlow RB. 1974. Correlation between deoxyribonucleic acid excision-repair and life-span in a number of mammalian species. *PNAS* **71**:2169–2173. DOI: https://doi.org/10.1073/pnas.71.6.2169, PMID: 4526202

Henry L, Wickham H. 2020. *Purrr: Functional Programming Tools*. https://CRAN.R-project.org/package=purrr

Hlavac M. 2018. *Stargazer: Well-Formatted Regression and Summary Statistics Tables*. https://CRAN.R-project.org/package=stargazer

Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. 2018. UFBoot2: improving the ultrafast bootstrap approximation. *Molecular Biology and Evolution* **35**:518–522. DOI: https://doi.org/10.1093/molbev/msx281, PMID: 29077904

Jassal B, Matthews L, Viteri G, Gong C, Lorente P, Fabregat A, Sidiropoulos K, Cook J, Gillespie M, Haw R, Loney F, May B, Milacic M, Rothfels K, Sevilla C, Shamovsky V, Shorser S, Varusai T, Weiser J, Wu G, et al. 2020. The reactome pathway knowledgebase. *Nucleic Acids Research* **48**:D498–D503. DOI: https://doi.org/10.1093/nar/gkz1031, PMID: 31691815

Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature Methods* **14**:587–589. DOI: https://doi.org/10.1038/nmeth.4285, PMID: 28481363

Kassambara A. 2020. *Ggpubr:' Ggplot2' Based Publication Ready Plots*. https://CRAN.R-project.org/package=ggpubr

Epub 2013 Jan 16 **Katoh K**, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution* **30**:772–780. DOI: https://doi.org/10.1093/molbev/mst010, PMID: 23329690

Katzourakis A, Magiorkinis G, Lim AG, Gupta S, Belshaw R, Gifford R. 2014. Larger mammalian body size leads to lower retroviral activity. *PLOS Pathogens* **10**:e1004214. DOI: https://doi.org/10.1371/journal.ppat.1004214, PMID: 25033295

Kent WJ. 2002. BLAT–the BLAST-like alignment tool. *Genome Research* **12**:656–664. DOI: https://doi.org/10.1101/gr.229202, PMID: 11932250

Kim D, Langmead B, Salzberg SL. 2015. HISAT: a fast spliced aligner with low memory requirements. *Nature Methods* **12**:357–360. DOI: https://doi.org/10.1038/nmeth.3317, PMID: 25751142

Larramendi A. 2015. Shoulder height, body mass, and shape of proboscideans. *Acta Palaeontologica Polonica* **61**:2014. DOI: https://doi.org/10.4202/app.00136.2014

Launonen V. 2005. Mutations in the human *LKB1/STK11* gene. *Human Mutation* **26**:291–297. DOI: https://doi.org/10.1002/humu.20222, PMID: 16110486

Lefort V, Desper R, Gascuel O. 2015. FastME 2.0: a comprehensive, accurate, and fast Distance-Based phylogeny inference program. *Molecular Biology and Evolution* **32**:2798–2800. DOI: https://doi.org/10.1093/molbev/msv150, PMID: 26130081

Lemoine F, Correia D, Lefort V, Doppelt-Azeroual O, Mareuil F, Cohen-Boulakia S, Gascuel O. 2019. Ngphylogeny.fr: new generation phylogenetic services for non-specialists. *Nucleic Acids Research* **47**:W260–W265. DOI: https://doi.org/10.1093/nar/gkz303, PMID: 31028399

Leroi AM, Koufopanou V, Burt A. 2003. Cancer selection. *Nature Reviews Cancer* **3**:226–231. DOI: https://doi.org/10.1038/nrc1016, PMID: 12612657

Liang J, Mills GB. 2013. AMPK: a contextual oncogene or tumor suppressor? *Cancer Research* **73**:2929–2935. DOI: https://doi.org/10.1158/0008-5472.CAN-12-3876, PMID: 23644529

Liao Y, Wang J, Jaehnig EJ, Shi Z, Zhang B. 2019. WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. *Nucleic Acids Research* **47**:W199–W205. DOI: https://doi.org/10.1093/nar/gkz401, PMID: 31114916

Lindblad-Toh K, Garber M, Zuk O, Lin MF, Parker BJ, Washietl S, Kheradpour P, Ernst J, Jordan G, Mauceli E, Ward LD, Lowe CB, Holloway AK, Clamp M, Gnerre S, Alföldi J, Beal K, Chang J, Clawson H, Cuff J, et al. 2011. A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478**:476–482. DOI: https://doi.org/10.1038/nature10530, PMID: 21993624

Lucena RB, Rissi DR, Kommers GD, Pierezan F, Oliveira-Filho JC, Macêdo JT, Flores MM, Barros CS. 2011. A retrospective study of 586 tumours in brazilian cattle. *Journal of Comparative Pathology* **145**:20–24. DOI: https://doi.org/10.1016/j.jcpa.2010.11.002, PMID: 21247583

Martins EP, Hansen TF. 1997. Phylogenies and the comparative method: a general approach to incorporating phylogenetic information into the analysis of interspecific data. *The American Naturalist* **149**:646–667. DOI: https://doi.org/10.1086/286013

Million Women Study collaborators, Green J, Cairns BJ, Casabonne D, Wright FL, Reeves G, Beral V. 2011. Height and Cancer incidence in the million women study: prospective cohort, and meta-analysis of prospective studies of height and total Cancer risk. *The Lancet Oncology* **12**:785–794. DOI: https://doi.org/10.1016/S1470-2045(11)70154-1, PMID: 21782509

Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, Lanfear R. 2020. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Molecular Biology and Evolution* **37**:1530–1534. DOI: https://doi.org/10.1093/molbev/msaa015, PMID: 32011700

Müller K, Wickham H. 2020. *Tibble: Simple Data Frames*. https://CRAN.R-project.org/package=tibble

Nagy JD, Victor EM, Cropper JH. 2007. Why don't all whales have Cancer? A novel hypothesis resolving Peto's paradox. *Integrative and Comparative Biology* **47**:317–328. DOI: https://doi.org/10.1093/icb/icm062, PMID: 21672841

Nishimura O, Hara Y, Kuraku S. 2017. gVolante for standardizing completeness assessment of genome and transcriptome assemblies. *Bioinformatics* **33**:3635–3637. DOI: https://doi.org/10.1093/bioinformatics/btx445, PMID: 29036533

Nunney L. 1999. Lineage selection and the evolution of multistage carcinogenesis. *Proceedings of the Royal Society of London. Series B: Biological Sciences* **266**:493–498. DOI: https://doi.org/10.1098/rspb.1999.0664

Nunney L. 2018. Size matters: height, cell number and a person's risk of cancer. *Proceedings of the Royal Society B: Biological Sciences* **285**:20181743. DOI: https://doi.org/10.1098/rspb.2018.1743

O Leary MA, Bloch JI, Flynn JJ, Gaudin TJ, Giallombardo A, Giannini NP, Goldberg SL, Kraatz BP, Luo ZX, Meng J, Ni X, Novacek MJ, Perini FA, Randall ZS, Rougier GW, Sargis EJ, Silcox MT, Simmons NB, Spaulding M, Velazco PM, et al. 2013a. The placental mammal ancestor and the post-K-Pg radiation of placentals. *Science* **339**:662–667. DOI: https://doi.org/10.1126/science.1229237, PMID: 23393258

O Leary MA, Bloch JJ, Flynn TJ, Gaudin A, Giallombardo NP, Giannini SL, Goldberg BPK, Luo Z-X, Meng J, Ni X, Novacek MJ, Perini FA, Randall Z, Rougier GW, Sargis EJ, Silcox MT, Simmons NB, Spaulding M, Velazco PM, Weksler M, et al. 2013b. Response to comment on "The placental mammal ancestor and the post-K-Pg radiation of placentals". *Science* **341**:613. DOI: https://doi.org/10.1126/science.1238162

Ooms J. 2020. *Magick: Advanced Graphics and Image-Processing in R*. https://CRAN.R-project.org/package=magick
Palkopoulou E, Mallick S, Skoglund P, Enk J, Rohland N, Li H, Omrak A, Vartanyan S, Poinar H, Götherström A, Reich D, Dalén L. 2015. Complete genomes reveal signatures of demographic and genetic declines in the woolly mammoth. *Current Biology* **25**:1395–1400. DOI: https://doi.org/10.1016/j.cub.2015.04.007, PMID: 25913407
Palkopoulou E, Lipson M, Mallick S, Nielsen S, Rohland N, Baleka S, Karpinski E, Ivancevic AM, To TH, Kortschak RD, Raison JM, Qu Z, Chin TJ, Alt KW, Claesson S, Dalén L, MacPhee RDE, Meller H, Roca AL, Ryder OA, et al. 2018. A comprehensive genomic history of extinct and living elephants. *PNAS* **115**:E2566–E2574. DOI: https://doi.org/10.1073/pnas.1720554115, PMID: 29483247
Paradis E, Blomberg S, Bolker B, Brown J, Claramunt S, Claude J, Cuong HS, Desper R, Didier G, Durand B, Dutheil J, Ewing R, Gascuel O, Guillerme T, Heibl C, Ives A, Jones B, Krah F, Lawson D, Lefort V, et al. 2020. *Ape: Analyses of Phylogenetics and Evolution*. https://CRAN.R-project.org/package=ape
Paradis E, Schliep K. 2019. Ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35**:526–528. DOI: https://doi.org/10.1093/bioinformatics/bty633, PMID: 30016406
Parra G, Bradnam K, Ning Z, Keane T, Korf I. 2009. Assessing the gene space in draft genomes. *Nucleic Acids Research* **37**:289–297. DOI: https://doi.org/10.1093/nar/gkn916, PMID: 19042974
Pennell MW, Eastman JM, Slater GJ, Brown JW, Uyeda JC, FitzJohn RG, Alfaro ME, Harmon LJ. 2014. Geiger v2.0: an expanded suite of methods for fitting macroevolutionary models to phylogenetic trees. *Bioinformatics* **30**:2216–2218. DOI: https://doi.org/10.1093/bioinformatics/btu181, PMID: 24728855
Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL. 2015. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology* **33**:290–295. DOI: https://doi.org/10.1038/nbt.3122, PMID: 25690850
Pertea M, Kim D, Pertea GM, Leek JT, Salzberg SL. 2016. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and ballgown. *Nature Protocols* **11**:1650–1667. DOI: https://doi.org/10.1038/nprot.2016.095, PMID: 27560171
Peto R, Roe FJ, Lee PN, Levy L, Clack J. 1975. Cancer and ageing in mice and men. *British Journal of Cancer* **32**:411–426. DOI: https://doi.org/10.1038/bjc.1975.242, PMID: 1212409
Peto R. 2015. Quantitative implications of the approximate irrelevance of mammalian body size and lifespan to lifelong Cancer risk. *Philosophical Transactions of the Royal Society B: Biological Sciences* **370**:20150198. DOI: https://doi.org/10.1098/rstb.2015.0198
Pinheiro J, Bates D. 2020. *R-Core, Nlme: Linear and Nonlinear Mixed Effects Models*. https://CRAN.R-project.org/package=nlme
Puttick MN, Thomas GH. 2015. Fossils and living taxa agree on patterns of body mass evolution: a case study with afrotheria. *Proceedings of the Royal Society B: Biological Sciences* **282**:20152023. DOI: https://doi.org/10.1098/rspb.2015.2023
Qian W, Zhang J. 2014. Genomic evidence for adaptation by gene duplication. *Genome Research* **24**:1356–1362. DOI: https://doi.org/10.1101/gr.172098.114, PMID: 24904045
R Development Core Team. 2019. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.https://www.R-project.org/
Rastogi S, Liberles DA. 2005. Subfunctionalization of duplicated genes as a transition state to neofunctionalization. *BMC Evolutionary Biology* **5**:28. DOI: https://doi.org/10.1186/1471-2148-5-28, PMID: 15831095
Salichos L, Rokas A. 2011. Evaluating ortholog prediction algorithms in a yeast model clade. *PLOS ONE* **6**:e18755. DOI: https://doi.org/10.1371/journal.pone.0018755, PMID: 21533202
Savage VM, Allen AP, Brown JH, Gillooly JF, Herman AB, Woodruff WH, West GB. 2007. Scaling of number, size, and metabolic rate of cells with body size in mammals. *PNAS* **104**:4718–4723. DOI: https://doi.org/10.1073/pnas.0611235104, PMID: 17360590
Scheffer VB. 1972. The weight of the steller sea cow. *Journal of Mammalogy* **53**:912–914. DOI: https://doi.org/10.2307/1379236
Schrempf D, Minh BQ, von Haeseler A, Kosiol C. 2019. Polymorphism-Aware species trees with advanced mutation models, bootstrap, and rate heterogeneity. *Molecular Biology and Evolution* **36**:1294–1301. DOI: https://doi.org/10.1093/molbev/msz043, PMID: 30825307
Schwartz GT, Rasmussen DT, Smith RJ. 1995. Body-Size diversity and community structure of fossil hyracoids. *Journal of Mammalogy* **76**:1088–1099. DOI: https://doi.org/10.2307/1382601
Seluanov A, Hine C, Bozzella M, Hall A, Sasahara TH, Ribeiro AA, Catania KC, Presgraves DC, Gorbunova V. 2008. Distinct tumor suppressor mechanisms evolve in rodent species that differ in size and lifespan. *Aging Cell* **7**:813–823. DOI: https://doi.org/10.1111/j.1474-9726.2008.00431.x, PMID: 18778411
Sievert C, Parmer C, Hocking T, Chamberlain S, Ram K, Corvellec M, Despouy P. 2020. *Plotly: Create Interactive Web Graphics Via' plotly.js'*. https://CRAN.R-project.org/package=plotly
Sievert C. 2020. *Interactive Web-Based Data Visualization with R, Plotly, and Shiny*. Chapman: Hall/CRC.
Slater GJ, Harmon LJ, Wegmann D, Joyce P, Revell LJ, Alfaro ME. 2012. Fitting models of continuous trait evolution to incompletely sampled comparative data using approximate bayesian computation. *Evolution* **66**:752–762. DOI: https://doi.org/10.1111/j.1558-5646.2011.01474.x
Slowikowski K. 2020. *Ggrepel: Automatically Position Non-Overlapping Text Labels With' Ggplot2'* . https://CRAN.R-project.org/package=ggrepel

Soubrier J, Steel M, Lee MS, Der Sarkissian C, Guindon S, Ho SY, Cooper A. 2012. The influence of rate heterogeneity among sites on the time dependence of molecular rates. *Molecular Biology and Evolution* **29**: 3345–3358. DOI: https://doi.org/10.1093/molbev/mss140, PMID: 22617951

Springer MS, Meredith RW, Teeling EC, Murphy WJ. 2013. Technical comment on "The placental mammal ancestor and the post-K-Pg radiation of placentals". *Science* **341**:613.2–61613. DOI: https://doi.org/10.1126/science.1238025, PMID: 23929967

Stoltzfus A. 1999. On the possibility of constructive neutral evolution. *Journal of Molecular Evolution* **49**:169–181. DOI: https://doi.org/10.1007/PL00006540, PMID: 10441669

Sulak M, Fong L, Mika K, Chigurupati S, Yon L, Mongan NP, Emes RD, Lynch VJ. 2016. *TP53* copy number expansion is associated with the evolution of increased body size and an enhanced DNA damage response in elephants. *eLife* **5**:e11994. DOI: https://doi.org/10.7554/eLife.11994, PMID: 27642012

Tacutu R, Craig T, Budovsky A, Wuttke D, Lehmann G, Taranukha D, Costa J, Fraifeld VE, de Magalhães JP. 2013. Human ageing genomic resources: integrated databases and tools for the biology and genetics of ageing. *Nucleic Acids Research* **41**:D1027–D1033. DOI: https://doi.org/10.1093/nar/gks1155, PMID: 23193293

The UniProt Consortium. 2017. UniProt: the universal protein knowledgebase. *Nucleic Acids Research* **45**:D158–D169. DOI: https://doi.org/10.1093/nar/gkw1099, PMID: 27899622

Tian X, Azpurua J, Hine C, Vaidya A, Myakishev-Rempel M, Ablaeva J, Mao Z, Nevo E, Gorbunova V, Seluanov A. 2013. High-molecular-mass hyaluronan mediates the Cancer resistance of the naked mole rat. *Nature* **499**:346–349. DOI: https://doi.org/10.1038/nature12234, PMID: 23783513

Tollis M, Ferris E, Campbell MS, Harris VK, Rupp SM, Harrison TM, Kiso WK, Schmitt DL, Garner MM, Aktipis CA, Maley CC, Boddy AM, Yandell M, Gregg C, Schiffman JD, Abegglen LM. 2020. Elephant genomes reveal insights into differences in disease defense mechanisms between species. *bioRxiv*. DOI: https://doi.org/10.1101/2020.05.29.124396

Tyner SD, Venkatachalam S, Choi J, Jones S, Ghebranious N, Igelmann H, Lu X, Soron G, Cooper B, Brayton C, Park SH, Thompson T, Karsenty G, Bradley A, Donehower LA. 2002. p53 mutant mice that display early ageing-associated phenotypes. *Nature* **415**:45–53. DOI: https://doi.org/10.1038/415045a, PMID: 11780111

Vazquez JM, Sulak M, Chigurupati S, Lynch VJ. 2018. A zombie LIF gene in elephants is upregulated by TP53 to induce apoptosis in response to DNA damage. *Cell Reports* **24**:1765–1776. DOI: https://doi.org/10.1016/j.celrep.2018.07.042, PMID: 30110634

Vazquez JM, Lynch VJ. 2021. Pervasive duplication of tumor suppressors in Afrotherians during the evolution of large bodies and reduced cancer risk. *Software Heritage*. swh:1:rev:6bc68ac31ef148131480710e50b0b75d06077db2. https://archive.softwareheritage.org/swh:1:dir:ae8a48052c5d5e4e3757e45dc5d632d07c443e3b;origin=https://github.com/docmanny/atlantogenataGeneDuplication;visit=swh:1:snp:d095eb620de426a079259bc6c9f6e8ebd57daf06;anchor=swh:1:rev:6bc68ac31ef148131480710e50b0b75d06077db2/

Wang HC, Minh BQ, Susko E, Roger AJ. 2018. Modeling site heterogeneity with posterior mean site frequency profiles accelerates accurate phylogenomic estimation. *Systematic Biology* **67**:216–235. DOI: https://doi.org/10.1093/sysbio/syx068, PMID: 28950365

Wang LG, Lam TT, Xu S, Dai Z, Zhou L, Feng T, Guo P, Dunn CW, Jones BR, Bradley T, Zhu H, Guan Y, Jiang Y, Yu G. 2020. Treeio: an R package for phylogenetic tree input and output with richly annotated and associated data. *Molecular Biology and Evolution* **37**:599–603. DOI: https://doi.org/10.1093/molbev/msz240, PMID: 31633786

Wickham H. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. New York: Springer-Verlag.

Wickham H, Hester J, Francois R. 2018. *Readr: Read Rectangular Text Data*. https://CRAN.R-project.org/package=readr

Wickham H, Averick M, Bryan J, Chang W, McGowan L, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen T, Miller E, Bache S, Müller K, Ooms J, Robinson D, Seidel D, Spinu V, Takahashi K, et al. 2019. Welcome to the tidyverse. *Journal of Open Source Software* **4**:1686. DOI: https://doi.org/10.21105/joss.01686

Wickham H. 2019a. *Stringr: Simple, Consistent Wrappers for Common String Operations*. https://CRAN.R-project.org/package=stringr

Wickham H. 2019b. *Tidyverse: Easily Install and Load The' Tidyverse*. https://CRAN.R-project.org/package=tidyverse

Wickham H. 2020. *Forcats: Tools for Working with Categorical Variables (Factors)*. https://CRAN.R-project.org/package=forcats

Wickham H, François R, Henry L, Müller K. 2020a. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr

Wickham H, Chang W, Henry L, Pedersen TL, Takahashi K, Wilke C, Woo K, Yutani H, Dunnington D. 2020b. *Ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics*. https://CRAN.R-project.org/package=ggplot2

Wickham H, Henry L. 2020. *Tidyr: Tidy Messy Data*. https://CRAN.R-project.org/package=tidyr

Xiao N. 2018. *Ggsci: Scientific Journal and Sci-Fi Themed Color Palettes For' Ggplot2'*. https://CRAN.R-project.org/package=ggsci

Xie Y. 2016. *Bookdown: Authoring Books and Technical Documents with R Markdown*. Florida: Hall/CRC.

Xie Y. 2020. *Bookdown: Authoring Books and Technical Documents with R Markdown*. https://CRAN.R-project.org/package=bookdown

Yang Z, Kumar S, Nei M. 1995. A new method of inference of ancestral nucleotide and amino acid sequences. *Genetics* **141**:1641–1650. DOI: https://doi.org/10.1093/genetics/141.4.1641, PMID: 8601501

Yu G, Smith DK, Zhu H, Guan Y. 2017. Ggtree : an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data . *Methods in Ecology and Evolution* **8**:28–36. DOI: https://doi.org/10.1111/2041-210X.12628

Yu G, Lam TT, Zhu H, Guan Y. 2018. Two methods for mapping and visualizing associated data on phylogeny using *Ggtree*. *Molecular Biology and Evolution* **35**:3041–3043. DOI: https://doi.org/10.1093/molbev/msy194, PMID: 30351396

Yu G. 2020a. *Ggimage: Use Image In' Ggplot2'*. https://CRAN.R-project.org/package=ggimage

Yu G. 2020b. *Ggplotify: Convert Plot To' Grob' Or' Ggplot' Object*. https://CRAN.R-project.org/package=ggplotify

Yu G. 2020d. *Tidytree: A Tidy Tool for Phylogenetic Tree Data Manipulation*. https://yulab-smu.github.io/treedata-book/

Yu G. 2020e. *Treeio: Base Classes and Functions for Phylogenetic Tree Input and Output*. https://rdrr.io/bioc/treeio/

Yu G. 2020f. Using ggtree to Visualize Data on Tree-Like Structures. *Current Protocols in Bioinformatics* **69**:e96. DOI: https://doi.org/10.1002/cpbi.96

Yu G, Lam TT. 2020c. *Ggtree: An R Package for Visualization of Tree and Annotation Data*. https://yulab-smu.github.io/treedata-book/

Yutani H. 2020. *Gghighlight: Highlight Lines and Points In' Ggplot2'*. https://CRAN.R-project.org/package=gghighlight

Zhu H. 2019. *KableExtra: Construct Complex Table With' Kable' and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra

Appendix 1

Summary of duplicate gene annotations associated with ***Figure 4C,D***.

BRD7:

Three manatee copies, but tml_BRD7_3 has PMSs. Overall high sequence similarity.

BUB3:

#manatee=#elephant.

PMS in loxafr #2–3, PMS in triman #2–3.

High AA conservation even in pseudogenes.

Casp9:

Two extra elephant copies are identical in AA sequence but hits only encompass a 47-AA hit that is highly conserved with matching domain in CASP9.

CD14:

#Manatee = #Elephant, all four copies are highly conserved.

CNOT11:

9/11 elephant copies do contain PMS; however, many of these are still mostly full length and highly conserved. The ones with an early PMS are less well conserved.

COX20:

Second elephant copy has two PMSs, but is still well conserved.

HMGB2:

2/4 elephant copies have PMSs, all copies highly conserved. Manatee 2/3 copies have PMS, lower conservation.

LAMTOR5:

Only one elephant copy has a PMS; three very highly conserved copies, others with conserved domains.

LIF:

Manatee's have 4 of 13 copies previously reported in elephants. See ***Vazquez et al., 2018***.

TP53:

Elephant: 9/19 with PMSs. Low N' conservation, but high conservation on C' end. See ***Sulak et al., 2016***.

STK11:

Very highly conserved, with divergence in the second copies in the elephant and manatee.

SOD1:

Highly conserved elephant copies. Manatee copies are only partial hits, with moderate conservation.

PRDX1:

Extremely strong conservation for main copies. Second elephant copy is a partial hit and has some divergence from the main elephant copy.

MAPRE1:

2/3 elephant copies have early PMSs, but with subsequent ATGs. Very high conservation in main and 1/2 duplicate elephant copies, only partial hits on third copy. Very high conversation between manatee copies.

MAD2L1:

7/9 elephant copies have PMSs. 2/6 manatee copies have PMSs. However, all copies are very highly conserved.