THE UNIVERSITY OF CHICAGO

THE ROLE OF GENE DUPLICATION IN MEDIATING PETO'S PARADOX IN *AFROTHERIA* AND *CHIROPTERA*

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE BIOLOGICAL SCIENCES
AND THE PRITZKER SCHOOL OF MEDICINE
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF HUMAN GENETICS

BY
JUAN MANUEL VÁZQUEZ DÍAZ

CHICAGO, ILLINOIS
AUGUST 2020

To all who have walked alongside me,

to those who paved the path before me,

to those who had to stop and say goodbye,

and to those who will walk with me until the end.

"¡Ah desgraciado si el dolor te abate,
si el cansancio tus miembros entumece!
Haz como el árbol seco: reverdece
y como el germen enterrado: late."
- José de Diego, "En la Brecha"

Table of Contents

List of Figures

List of Tables

ACKNOWLEDGMENTS

In the hierarchy of acknowledgement, it is crucial that I acknowledge the environment and materials that molded the road I traveled on to this day. This work was made by a Puerto Rican scientist, born and raised in a colonial situation, whose fundamental right to suffrage is withheld by the laws of the United States of America so long as he calls himself a resident of Puerto Rico [161, 120, 111]. The trials and tribulations leading to the completion of this dissertation went beyond the standard fare for biologists, as Hispanic and other underrepresented groups in science suffer greatly increased attrition rates at every stage of their academic trajectories [185, 11, 95, 49]. Furthermore, this work, and all future works, are read in the context of a system where the impact and merit of work from underrepresented groups are undervalued and discounted relative to their peers [132, 70]. Despite recent efforts to enact change, the paucity of diversity is readily and conspicuously apparent at the highest levels of academia [57, 30]. It is the hope of the author that this thesis, and his eventual presence among tenured faculty, will provide a stake in the proverbial mountain to help the next generation of underrepresented scientists ascent to the heights of academia.

I would like to thank my friends, family and loved ones for paving the road I have taken: my mother, Dra. Libia A. Diaz Rivera, who passed away early this year, whose love, sacrifice, and dedication made me who I am; my uncle, Jesus M. Diaz Rivera, for being the father I never had; my cousin, Yanira M. Diaz Loyola, who has been a sister to me; my aunt, Karen Loyola Peralta, who was always there when I needed a second mom; my extended family for their support; my dear wife, Harshita Mira Venkatesh; my spiritual blood brother Rene Sagardia; Coral Rosario and Francisco Vargas for being family when I needed them; and Adele Mouakad and Gian Toyos for inspiring me to become a scientist, and giving me my first chance.

I would like to acknowledge the history of cancer in my family, and my own torrid affair with the concept of aging, in inspiring this research.

ABSTRACT

Cancer is a disease intrinsic to multicellularity. Within a species, body size and lifespan are strongly correlated with cancer risk; between species, however, this correlation no longer holds. This phenomena, known as Peto's Paradox, requires that species evolve cancer suppression mechanisms alongside increases in size and lifespan. Previous studies have identified instances of tumor suppressor duplications in large, long-lived species, suggesting a greater role for gene duplication in resolving Peto's Paradox. Thus, in this thesis, I identified all protein-coding gene duplications in available genomes to determine if tumor suppressor pathways were enriched among duplicated genes in large, long-lived species. Then, I selected two hits in large, long-lived species to characterize in primary fibroblasts, and determine their effects on cell cycle and cell death in response to stress: *LIF* in the African Elephant (*Loxodonta africana*) and *TP53* in the Little Brown Bat (*Myotis lucifugus*).

To determine if tumor suppressors gene duplications are more common in large-bodied *Atlantogenatans*, I used a Reciprocal Best-Hit BLAT strategy to obtain copy numbers of all protein-coding genes in *Atlantogenatan* genomes. From an initial set of 18,011 protein-coding genes, I identified a median of 13,880 genes in *Atlantogenatan* genomes, of which a median of 940 genes are duplicated. Just as body size fluctuates throughout *Atlantogenata*, tumor suppressor genes also duplicated throughout the phylogenetic tree; furthermore, many of them remain transcriptionally active in extant elephants. Together, the data suggest that the duplication of tumor suppressor genes facilitated the evolution of increased body size in *Atlantogenata*.

The resurrection and re-functionalization of a *LIF* pseudogene (*LIF6*) with pro-apoptotic functions in elephants and their extinct relatives (*Proboscideans*) may have played a role in resolving Peto's Paradox. *LIF6* is transcriptionally up-regulated by *TP53* in response to DNA damage, and translocates to the mitochondria where it induces apoptosis. Phylogenetic analyses of living and extinct *Proboscidean LIF6* genes indicates its *TP53* response element

evolved coincident with the evolution of large body sizes in the *Proboscidean* stem-lineage. These results suggest that re-functionalizing of a pro-apoptotic *LIF* pseudogene may have been permissive (though not sufficient) for the evolution of large body sizes in *Proboscideans*.

In the long-lived bat, *Myotis lucifugus*, I describe a duplication of the *TP53-WRAP53* locus which may play a role in shaping its unique stress response. While pseudogene copies of *TP53* are common in Myotis bats, *M. lucifugus* has a unique, syntenic duplication of *TP53-WRAP53* that has conserved both regulatory and transcriptional functionality. Relative to 4 other closely related bat species (*M. evotis*, *M. thysanodes*, *M. yumanensis*, and *E. fuscus*), the *M. lucifugus* demonstrates a unique resistance to DNA damage and generalized oxidative stress, resembling the phenotype of a *TP53-WRAP53* locus duplication in a previously-described transgenic mouse model.

Overall, these results suggest that gene duplication plays an important role in Peto's Paradox. While tumor suppressor duplications may facilitate the evolution of increased lifespans and body sizes in the short term, my work suggests the need for a polygenic or omnigenic model for Peto's Paradox in order to comprehensively lay this question to rest.

CHAPTER 1

INTRODUCTION

1.1 Two paradigms, one paradox

The relationship between cancer, body size, and lifespan in mice and men has been known for quite some time. It is known that differences in body size between members of the same species lead to proportional differences in cell counts within their body. Thus, if any cell in the body has the potential to become cancerous, then taller individuals with more cells should have a proportionally higher risk of cancer; unsurprisingly, this holds true not only in humans, but in other species such as dogs and mice.

Similarly, the time-dependent nature of mutagenesis and oncogenesis should lead to a positive relationship between cancer risk and age. As time passes, cells acquire and accumulate mutations which eventually lead to oncogenesis; furthermore, there may be other biological processes associated with age, such as decreased immunosurveillance, that can allow tumors to establish and thrive in the body. The increased incidence rate of cancer in older populations relative to younger populations has been well-established, not only in humans, but in many other species as well.

The fact that the relationships between cancer and body size and lifespan is present in multiple species suggests that this relationship is a fundamental biological fact, as opposed to a species-specific curiosity. Extrapolating these within-species studies of cancer epidemiology to comparisons between species, however, lead to the discovery of an equally fundamental contradiction that holds the promise of a new world of insight into the biology of cancer avoidance and treatment, cancer risk does not corelate with either body size or lifespan across species, an observation that has become known as "Peto's Paradox".

Numerous mechanisms have been proposed to resolve Peto's paradox, including reduced copy number of oncogenes, an increase in the copy number of tumor suppressor genes

[21, 99, 125], reduced metabolic rates leading to decreased free radical production, reduced retroviral activity and load [88], increased immune surveillance, and selection for 'cheater' tumors that parasitize the growth of other tumors [122], among many others. Gene duplication has long been recognized to play an important role in the generation of evolutionarily relevant phenotypic variation but thus far been understudied as a particularly parsimonious resolution to the evolution of Peto's Paradox. By sequencing new genomes, many studies have examined positive selection and conservation of tumor suppressor genes in large, long-lived species to elucidate which genes are involved in mediating Peto's Paradox [190, 158, 135, 40, 116, 44, 93, 43, 100, 54, 192, 89]. However, many studies have described individual cases of tumor suppressor gene duplications[180, 170, 1], which suggests that sequence evolution in 1:1 orthologous genes may not fully resolve Peto's Paradox; they have also used methods that mask recent gene duplications.

In this work, I explore more thoroughly the possibility that gene duplication played a role in the resolution of Peto's Paradox in lineages with a high theoretical risk of cancer, such as Elephants, Whales, and Bats. To do so, I first investigate the overall pattern of gene duplication for all human protein-coding homologs in other genomes to determine if tumor suppressor gene duplications are especially enriched among the pool of genes which have duplicated in lineages with exceptional body sizes or lifespans. Then, I functionally characterize two such duplicated genes - LIF and TP53 - to determine if the duplicate copies conserve functionality *in vitro* using a primary cell culture model that accurately reflects the biology of the whole organism.

1.2 How one becomes too many: the Multistage Model of Carcinogenesis

The crux of Peto's Paradox and this work lies in the understanding of how tumors form and develop. While many theories of carcinogenesis have been postulated, one of the simplest

and most powerful models is the Multistage Model of Carcinogenesis [7, 6, 140]. The model describes cancer as a multistage process, where cells progress through a number of states until reaching a rate-limiting "precancerous" stage; at this point, the next state change will create a cancerous cell that begins to propagate and divide uncontrollably. The transitions from a normal cell to a cancer cell are, functionally, mutations and other disorders that are well-described by various hallmarks of cancer [61, 62].

Each state change is a time-dependent process, as such, the cancer risk of a single cell is proportional to the time that the organism has been alive. This age-dependence of cancer is familiar to humans: Figure 1.1A displays the positive correlation between age and cancer incidence rate per 100,000 individuals based on data sourced from the Surveillance, Epidemiology, and End Results Program by the National Cancer Institute's Division of Cancer Control and Population Sciences [172]. Perhaps unsurprisingly, tissues that are exposed to chronic sources of stress, such as digestive tissues (acidity, replication stress) [18, 162] and lungs (oxidative stress, carcinogen exposure) have a stronger time-relationship than tissues such as bones and joints, which replicate slowly and experience more limited stress, and have smaller populations of epithelial cells which are particularly prone to cancerous transformation [112, 119]. However, importantly, even these tissues see a correlation with age, indicating that time affects the cancer risk of all cells.

An individual's overall cancer risk is the sum of the cancer risk over all the cells in their body; as such, individuals with a greater number of cells are at greater risk of cancer than individuals with a lower number of cells. Height serves as a useful proxy for cell number (independent of estimates such as body mass index which is correlated with other health problems), as actively dividing cells? size remains invariant both within and between species. Various studies, including the five population-level studies examined by Nunney (2018) [97, 56, 81, 186, 171, 127], have shown that both height and BMI correlate with various cancers; Figure 1.1B reproduces the summary data from Nunney (2018) [127] for both sexes,

for various cancers. The overall mean hazard ratio per 10cm height increase for all cancers and sexes was 1.11 (95% CI 1.09-1.12), indicating that there is a significant effect of body size on cancer risk.

Aside from humans, the multistage model of carcinogenesis is supported by data from other species. Larger dog lbreeds, for example, are at a greater risk of cancer than short-lived species [34]; furthermore, for both dogs and cats, cancer incidence rates per 100,000 have been shown to increase significantly over the lifetime of the species [38]. Meanwhile, in cattle, age has been shown to be significantly correlated with the incidence rate of various neoplasia [104]. Neoplasia has additionally been reported for various other species of mammals, birds, and dogs in the literature, although data correlating these with lifespan and body size within the species are not readily available [41, 4]. Thus, body size and lifespan are significant risk factors for cancer in not just humans, but in many other species as well.

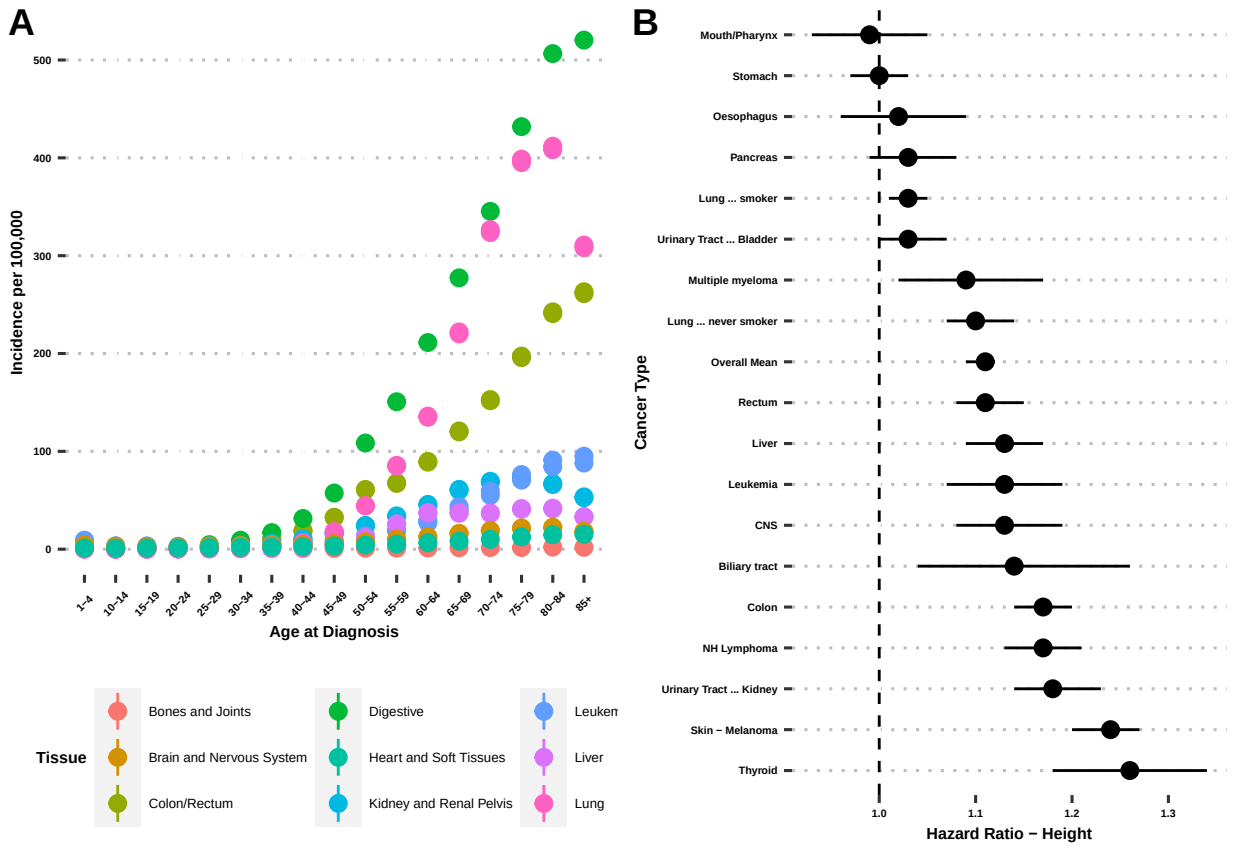

Figure 1.1: The relationship between body size, lifespan, and cancer risk within Humans. A) SEER21 Data demonstrating the positive relationship between the age of the population and the cancer incidence rate for cancers in all tissue types. B) Figure reproduced from Nunney (2018) summarizing the hazard ratio per 10cm of height for various types of cancers, collected from 5 international studies [127].

1.3 Cancer rates between species: Peto's Paradox

Among the many morphological differences between species, body size and lifespan are among the most starkly apparent (Figure 1.2A). Between distantly related species, these differences can be enormous: consider the classic paradigm of the mouse and the elephant, for example. These differences in size have been shown to be due to an increase in cell count, rather than cell size [154]. As such, one would initially assume that the within-species paradigm of "more

cells means more cancer" would also translate into higher cancer rates in elephants versus mice.

Further compounding the relationship between size and cancer risk is the positive correlation between body size and lifespan. Not only do large species have more cells than smaller species, they also live comparatively longer: for example, the African Elephant lives nearly 21x longer than the House Mouse [140, 173] (Figure 1.2A). And so, given the paradigm of lifespan and cancer risk within species, one would expect that these large, long-lived species would be even more at risk of cancer than their smaller, shorter-lived cousins.

However, while cancer, size, and lifespan are correlated within species, they are not correlated between species. In Figure 1.2B, data reproduced from Abegglen et al (2015) [1] and colored by phylogenetic Order demonstrates that across mammals of all sizes and lifespans, we see no correlation between these demographic traits and the species? cancer risk.

The observation that species' cancer risks hold no correlation with either body size or lifespan was observed by various groups around the same time, but was coined "Peto's Paradox" after the publications by the statistical epidemiologist Sir Richard Peto [7, 6, 140]. In stark contrast with the correlation of body size and/or lifespan with cancer risk between members of the same species, across a variety of species, when one compares these species and others' cancer risks with their average body sizes and maximum lifespans, the correlations you see at the species-level disappear entirely. This paradox has been best studied in mammals [1, 141], but has also been observed in other vertebrate clades such as birds [121], and I have conducted initial studies in long-lived reptiles such as turtles and tortoises.

The existence of Peto's Paradox suggests that the evolution of increased body sizes and longevity must coincide with the evolution of enhanced cancer resistance in these species. Furthermore, depending on the timeframe of evolution, the co-evolution of these traits must track each other closely. The ideal study design for Peto's Paradox would be to examine clades

where closely-related species show especially high variability in size and lifespan. Two clades fit this paradigm well: the clade *Atlantogenata*, which is home to elephants and hyraxes; and the clade *Chiroptera*, which contains 20-25% of extant mammalian species, and encompasses a wide variety of all sizes and lifespans. Thus, these clades provide a robust starting point for addressing the question of how Peto's Paradox has been resolved by Evolution (Figure 1.3).

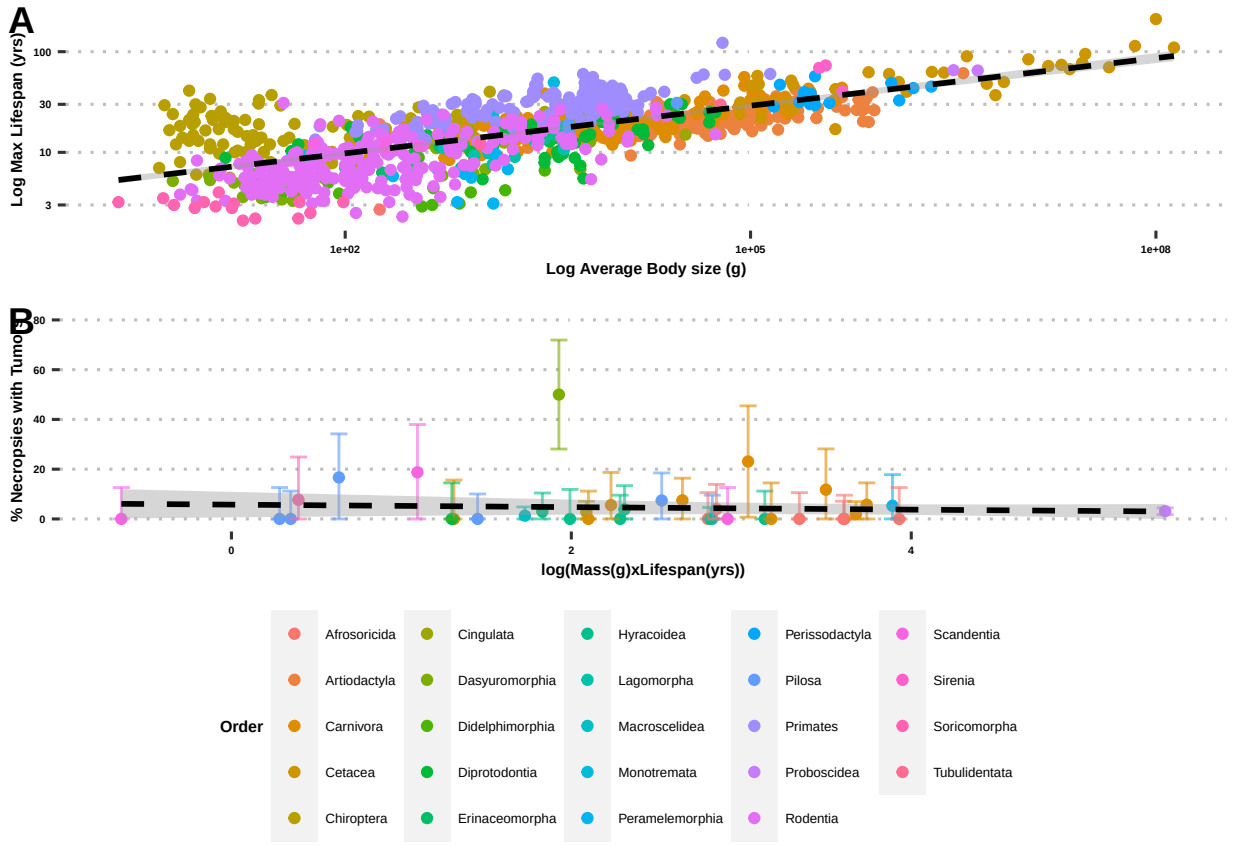

Figure 1.2: Peto's Paradox describes the lack of expected correlation between body size, lifespan, and cancer risk between species. A) Body size and lifespan for a plethora of mammalian species gathered from Anage [173]. B) No correlation between body size and lifespan across mammalian species; data collected by [1].

1.4 Clade selection and study design

For this thesis, I focus my attention on two clades: *Atlantogenata* (Figure 1.3B), and *Chiroptera* (Figure 1.3C). These clades have a large or long-lived species nested deeply in a clade of smaller, short-lived species, which indicates a recent expansion in size and/or lifespan. Combined with the available genomes for these clades, I will be able to search for gene duplication events along the tree, and determine where any tumor suppressor genes duplicated in the lineage leading to the main species of interest: the African Bush Elephant (Loxodonta africana, 65 years, 4800 kg) in *Atlantogenata*; and the Little Brown Bat (Myotis luficufugus, 34 years, 10 g); both of these species have relatively high-quality genomes. Furthermore, for these two clades, I have primary dermal fibroblasts as well as many outgroup species, allowing us to follow up on my genomic results with functional validation.

In Chapter 2, I describe how I developed a pipeline for identifying gene homologs in highly-fragmented genomes using a Reciprocal Best-Hit BLAT search method, and how I applied it to publically-available genomes in *Atlantogenata*, including living and extinct elephants, to determine the relationship between the evolution of body size, and gene duplication pathway enrichments. I found that both body size increases as well as tumor suppressor duplications are prevalent throughout *Atlantogenata*, and that many of these duplicates are conserved and show transcriptional activity in extant elephants. In Chapter 3, I demonstrate the functional characterization of one of these hits in the African Elephant: the retrogene LIF6. Although LIF has undergone various segmental duplications in the common ancestor of elephants, manatees, and hyraxes, one copy - LIF6 - was resurrected into functional gene by the creation of an upstream TP53 binding site, and induces apoptosis in response to DNA damage. Finally, in Chapter 4, I describe a syntenic duplication of the TP53-WRAP53 locus in the Little Brown Bat, *Myotis lucifugus*, which has retained both regulatory and transcriptional functionality. While a causal role has not established between this duplication and stress response in *Myotis lucifugus*, the patterns of stress response shown by this species

relative to other Myotis species is similar a previous mouse model of TP53-WRAP53 locus duplication described in the literature. Ultimately while no single mechanism can explain the evolution of cancer resistant species, gene duplication appear to play a major role in mediating the cancer resistance of the large, long-lived species included in this study.

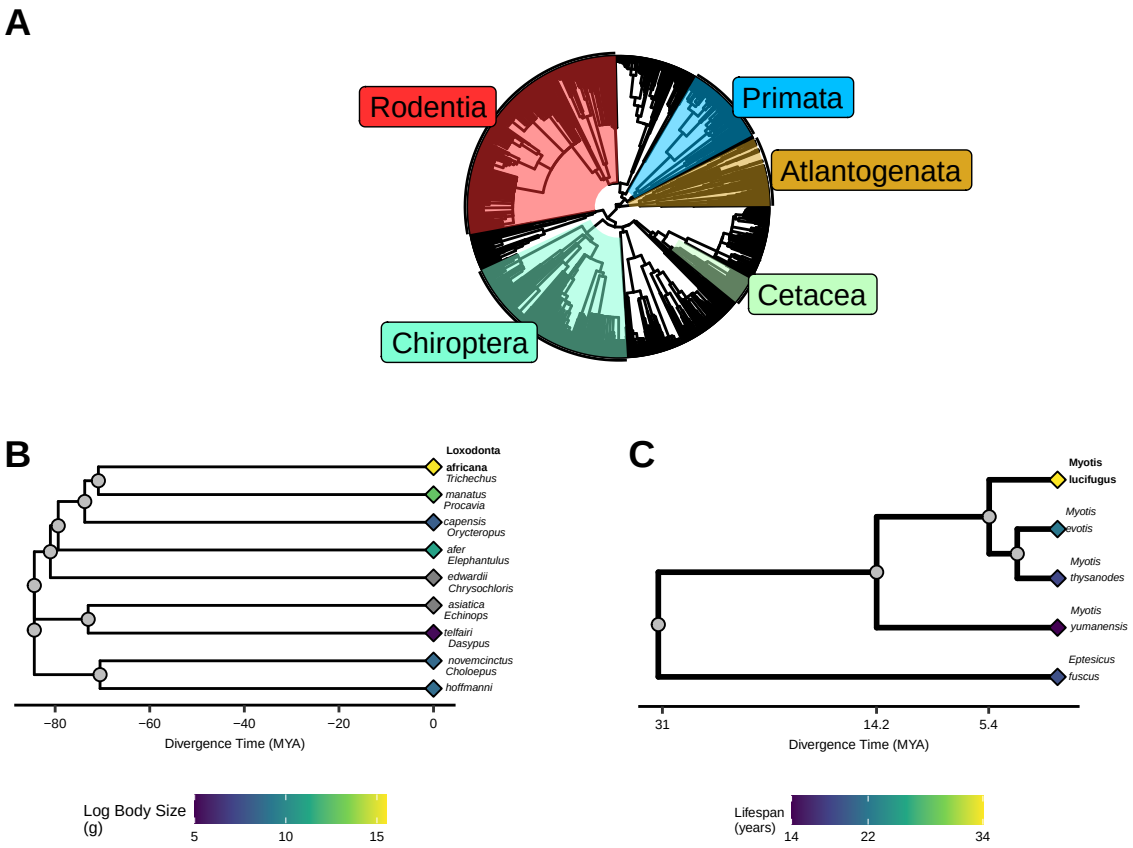

Figure 1.3: *Atlantogenata* and *Chiroptera* in their phylogenetic context. A) A time-calibrated species tree for *Eutheria*, with major clades highlighted [14]. B) *Atlantogenatan* species with publicly-available genomes, used in this thesis. Tip points are colored based on the log body size of the species, where the African Elephant *Loxodonta africana* is the largest species, deeply nested within much smaller species. C) *Chiropteran* species related to the long-lived Little Brown Bat *Myotis lucifugus*, for which primary fibroblasts are available for *in vitro* functional characterizations. Tip points are colored based on maximum reported lifespans; note that *Myotis lucifugus* is a long-lived species nested within a clade of much-shorter-lived species.

CHAPTER 2

PERVASIVE DUPLICATION OF TUMOR SUPPRESSOR GENES PRECEDED PARALLEL EVOLUTION OF LARGE BODIED *ATLANTOGENATANS*

2.1 Introduction

Among the major constraints on the evolution of large body sizes (and long life-spans) 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, organism 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 [154]. 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 [56, 127] and in dogs [34, 38]. 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' [21, 99, 140]. The ultimate resolution to Peto's Paradox is obvious: large bodied and long-lived species evolved enhanced cancer protection mechanisms. However, identifiying the specific genetic, molecular, and cellular mechanisms that underlie the evolution of augmented cancer protection has been difficult [c.f. 8, 160, 55, 176, 170].

Figure 2.1: *Atlantogenatans* with sequenced genomes, body sizes, and known lifespans [173, 145].

Among 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 ˜70g in golden moles, ˜120g in tenrecs, ˜170g in elephant shrews, ˜3kg in hyraxes, and 60kg in aardvarks [173]. While extant hyraxes are relatively small, the extinct Titanohyrax is estimated to have weighted up to ˜1300kg [157]. The largest members of *Afrotheria*, too, are dwarfed by the size of their recent ancestors: extant cows manatees are large bodied (˜322-480kg) but are relatively small compared to the extinct Stellar's sea cow which is estimated to have weight 8000-10000kg [155]. Similarly African (4,800kg) and Asian elephants (3,200kg) are the largest living elephant species, but are dwarfed by the truly gigantic extinct *Proboscideans* such as *Deinotherium* (˜132,000kg), *Mammut borsoni* (110,000kg), and the Asian straight-tusked elephant (˜220,000kg), the largest known land mammal [96]. Remarkably, these large-bodied *Afrotherian* lineages are nested within small bodied species (**Figure 2.1**) [129, 166, 130, 145], 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.

Although many mechanisms can potentially resolve Peto's paradox, among the most parsimonious routes to enhanced cancer resistance is through an increased copy number of tumor suppressors. Indeed, candidate genes studies have found that the elephant genome encodes duplicate such as *TP53* and *LIF* [1, 170, 180] as well as other genes with putative tumor suppressive functions [20, 35]. As these studies focus on *a priori* gene sets, however, it remains unknown whether this is a general, genome-wide trend in *Afrotherian* genomes; and whether such a general trend is associated with the recent increases in body size – and therefore expected cancer risk – in these species.

Here, we trace the evolution of body mass and gene copy number variation in across *Afrotherian* genomes in order to investigate whether duplication of tumors suppresor is common in large, long-lived *Proboscideans*. Our estimates of the evolution of body mass, similarly to previous studies [129, 166, 130, 145], show that large body masses evolved in a step-wise manner, with major increases in body mass in the *Pseudoungulata* (17kg), *Paenungulata* (25kg), *Tethytheria* (296kg), and *Proboscidea* (4,100kg) stem-lineages. To explore whether duplication of tumor suppressor genes 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 all *Afrotherians*, both large and small. These data suggest that duplication of tumor suppressor genes is pervasive in *Afrotherians* and proceeded the evolution of species with very large body sizes.

2.2 Methods

2.2.1 Ancestral Body Size Reconstruction

We built a time-calibrated supertree of *Eutherian* mammals by combining the time-calibrated molecular phylogeny of Bininda-Emonds *et al.* (2008) [14] with the time-calibrated total evidence *Afrotherian* phylogeny from Puttick and Thomas (2015) [145]. While the Bininda-Emonds *et al.* [14] phylogeny includes 1,679 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) [145] includes 77 extant species and fossil data

from 39 extinct species. Therefore we replaced the *Afrotherian* clade in the Bininda-Emonds *et al.* [14] phylogeny with the *Afrotherian* phylogeny of Puttick and Thomas [145] 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 [42]. The stable model allows for occasional large jumps in traits and has previously been shown to out-perform other models of body mass evolution, including standard Brownian motion models, Ornstein–Uhlenbeck models, early burst maximum likelihood models, and heterogeneous multi-rate models [42].

2.2.2 Identification of Duplicate Genes

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 [90] as our algorithm of choice, as this algorithm is sensitive to highly simliar (>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/poor assembled regions of the genome [5, 151].

Effective Copy Number By Coverage: In lower-quality genomes, many genes are

fragmented across multiple scaffolds, which results in BLAT calling multiple hits when in reality there is only one gene. To compensate for this, we came up with a novel statistic, Estimated Copy Number by Coverage (ECNC), which averages the number of times we see each nucleotides 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 S.2.1**). This allows us to correct for genes that have been fragmented across incomplete genomes, while also taking into account missing sequences from the human query in the target genome. Mathematically, this can be written as:

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

where n is a 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 $C_n > 0$ or 0 if $C_n = 0$.

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 Proteome provided by UniProt [Accession #UP000005640; 24], which is a comprehensive set of protein sequences curated from a combination of predicted and validated protein sequences generated by the UniProt Consortium. In order to refine our search, we omitted protein sequences originating from long, noncoding RNA loci (e.g. LINC genes); poorly-studied genes from predicted open reading frames (C-ORFs); and sequences with highly repetitive sequences such as zinc fingers, protocadherins, and transposon-containing genes, as these were prone to high levels of false positive hits.

After filtering out problematic protein queries (see below), we then used our pipeline to search for all copies of our `n.GenesSearched` query genes in publicly available *Afrotherian* genomes (S.2.2), 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, triManLat1_HiC), 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, choHof2_HiC), Cape golden mole (*Chrysochloris asiatica*: chrAsi1), and Cape elephant shrew (*Elephantulus edwardii*: eleEdw1).

Query gene inclusion criteria: To assemble our query list, we first removed all unnamed genes from UP000005640. 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 any protein described as either "Uncharacterized," "Putative," or "Fragment" by UniProt in UP000005640 (30724 genes), leaving us with a final set of 37582 query protein isoforms, corresponding to 18011 genes.

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 [170, 180], 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 2.4. Overall, we identified 13880 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 [123, 136]. CEGMA was run using the default settings for mammals ("Cut-off length for sequence statistics and composition" = 1;"CEGMA max intron length" = 100000; "CEGMA gene flanks" = 10000, "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.

2.2.3 Evidence for Functionality of Gene Duplicates

To validate and filter out duplicate gene calls, we intersected our results with either gene prediction or transcriptomic evidence as a proxy for functionality.

Transcriptome Assembly: For the African Savana Elephant, Asian Elephant, West Indian Manatee, and Nine-Banded Armadillo, we generated *de novo* transcriptomes using publically-available RNA-sequencing data from NCBI SRA (S.2.1). We mapped reads to all genomes available for each species, and assembled transcripts using HISAT2 and StringTie, respectively [92, 139, 138]. RNA-sequencing data was not available for Cape Golden Mole, Cape Elephant Shrew, Rock Hyrax, Aardvark, or the Lesser Hedgehog Tenrec.

Gene Prediction: We obtained tracks for genes predicted using GenScan for all the genomes available via UCSC Genome Browser: African savannah elephant (loxAfr3), Rock Hyrax (proCap1), West Indian Manatee (triManLat1), Aardvark (oryAfe1), Lesser Hedgehog

Tenrec (echTel2), Nine-banded armadillo (dasNov3), Hoffman's Two-Toed Sloth (choHof1), Cape golden mole (chrAsi1), and Cape Elephant Shrew (eleEdw1); gene prediction tracks for higher-quality assemblies were not available.

Evidenced Duplicate Criteria: We intersected our records of duplicate hits identified in each genome with the gene prediction tracks and/or transcriptome assemblies using `bedtools`. When multiple lines of evidence for functionality were present for a genome, we used the union of all intersections as the final output for evidenced duplicates. When analyzing the highest-quality assemblies available for each species, if a species had neither gene prediction tracks nor RNA-seq data for the highest-quality genome available, we conservatively included all hits for the genome in the final set of evidenced duplicates.

2.2.4 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 [118, 68, 82, 183, 156], 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 [118, 68, 82, 183, 156], the best fitting model of character evolution (MK+FQ+GR+I) [165, 187], and the unrooted species tree for *Atlantogenata*. We considered ancestral state reconstructions to be reliable if they had Bayeisan Posterior Probability (BPP) ?0.80; less reliable reconstructions were excluded for pathway analyses.

2.2.5 Pathway Enrichment Analysis

To determine if gene duplications were enriched in particular biological pathways, we used the WEB-based Gene SeT AnaLysis Toolkit (WEBGESTALT)[102] to perform overrepresentation analysis (ORA); pathway databases included Reactome [80], Wikipathways, [163], and KEGG [83]. Gene duplicates in each lineage were used as the foreground gene set, and the initial query set was used as the background gene set. Statistical significance of enriched terms was assessed with a hypergeometric test, and controlled for multiple testing using the Benjamini-Hochberg false discovery rate (FDR); each analysis was run at FDR=0.1, FDR=0.2, FDR=0.3, and FDR=0.5. In order to determine an empirical false positive rate for term enrichment, we randomly sampled 1000-10,000 genes from our background set 1000 times, and ran the aforementioned analyses to see which terms were likely to randomly appear; there were no terms which appeared at FDR ≤ 0.3.

2.2.6 Estimating the Evolution of Cancer Risk

The dramatic increase in body mass and lifespan in some *Afrotherian* lineages implies 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 [141] we estimate the intrinsic cancer risk as the product of risk associated with body mass and lifespan. In order to determine the intrinsic cancer risk (K) across species and at ancestral nodes (see below), we first needed to estimate ancestral lifespans at each node. We used Phylogenetic Generalized Least-Square Regression (PGLS) [46, 113] implemented in the R package `ape` 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 and the model $ln(lifespan) = \beta_1 corBrowninan + \beta_2 ln(size) + \epsilon$. Next, we calculated K_1 at all nodes, and then estimated the fold change in cancer susceptibility between ancestral

and descendant nodes (Figure 2.3, Table 2.2). 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: $K \approx Dt^6$ [6, 7, 141, 140]. The fold change in cancer risk between a node and its ancestor was then defined as $\log_2(\frac{K_2}{K_1})$.

2.3 Results

2.3.1 Step-wise evolution of body size in Afrotherians

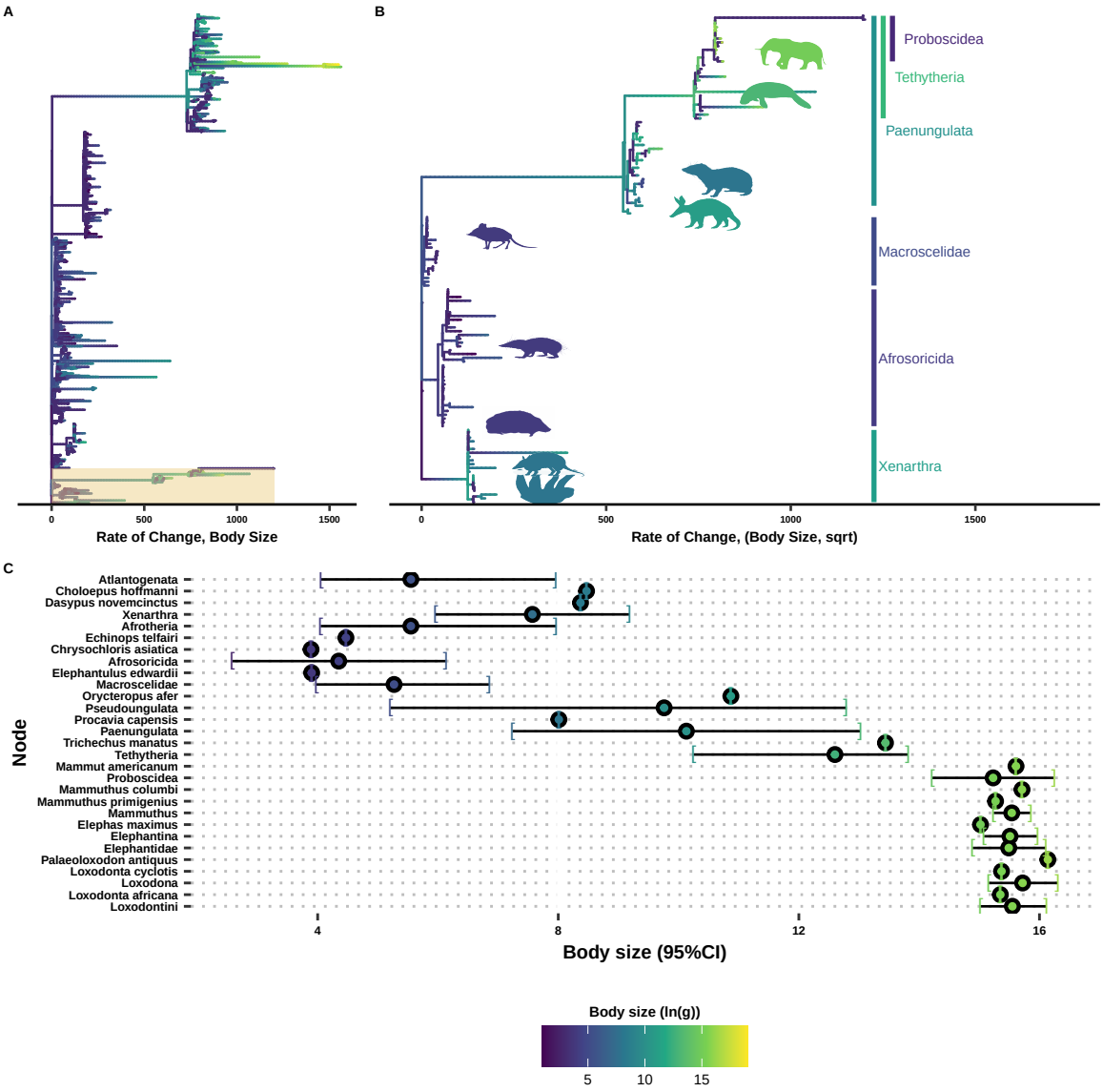

Figure 2.2: Body sizes rapidly and frequently expand in *Eutherians*, especially in *Atlantogenata*.

(continued on next page)

Similar to previous studies of *Afrotherian* body size [21,27], we found that the body mass of the *Afrotherian* ancestor was inferred to be small (0.26kg, 95% CI: 0.31-3.01kg) and that

Figure 2.2: Body sizes rapidly and frequently expand in *Eutherians*, especially in *Atlantogenata*. **A)** Tree of *Eutherian* species, colored by ln(Body Size) and with branch lengths set to the rate of change in body sizes, normalized by the square root of the root branch. *Atlantogenata* is highlighted at the bottom. **B)** Zoom-in of (**A**) on *Atlantogenata*. Silhuetes for the African Elephant, West Indian Manatee, Cape Elephant Shrew, Lesser Hedgehog Tenrec, Cape Golden Mole, Nine-Banded Armadillo, and Hoffman's Two-Toed Sloth are colored by their extant body sizes, while clade labels are colored based on the common ancestor's estimated body size. **C)** Confidence interval plot for representative species and ancestral nodes.

substantial accelerations in the rate of body mass evolution occurred coincident with a 67.36x increase in body mass in the stem-lineage of *Pseudoungulata* (17.33kg), a 1.45x increase in body mass in the stem-lineage of *Paenungulata* (25.08kg), a 11.82x increase in body mass in the stem-lineage of *Tehthytheria* (296.56kg), and a 2.69x increase in body mass in the stem-lineage of *Proboscidea* (4114.39kg) (Figure 2.2, Table 2.1). The ancestral *Hyracoidea* was inferred to be relatively small (2.86kg-118.18kg), and rate accelerations were coincident with independent body mass increases in large hyraxes such as *Titanohyrax andrewsi* (429.34kg, 67.36x increase). While the body mass of the ancestral *Sirenian* was inferred to be large (61.7kg-955.51kg), a rate acceleration occurred coincident with a 10.59x increase in body mass in Stellar's sea cow. Rate accelerations also occurred coincident with 36.6x decrease in body mass in the stem-lineage of the dwarf elephants *Elephas (Palaeoloxodon) antiquus falconeri* and *Elephas cypriotes*. These data suggest 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 2.2, Table 2.1).

Table 2.1: Body Size and Confidence Intervals in *Atlantogenata* estimated using StableTraits.

Node	Size (log(g))	95% CI (Low)	95% CI (High)	Rate (sqrt)
Cryptochloris wintoni	3.13	3.13	3.13	5.78
Amblysomus marleyi	3.53	3.53	3.53	3.79
Elephantulus revoili	3.48	3.48	3.48	1.10
Titanohyrax andrewsi	12.97	12.97	12.97	0.07
Titanohyrax ultimus	14.08	14.08	14.08	34.61
Megalohyrax sp nov	12.52	12.52	12.52	7.21
Elephas maximus asurus	15.66	15.66	15.66	0.34
Protenrec tricuspis	1.14	1.14	1.14	69.75
Microgale parvula	1.16	1.16	1.16	33.46
Microgale pusilla	1.25	1.25	1.25	34.31
Geogale aurita	1.90	1.90	1.90	40.07
Microgale longicaudata	2.09	2.09	2.09	0.77
Microgale brevicaudata	2.19	2.19	2.19	0.60
Microgale jobihely	2.30	2.30	2.30	1.07
Microgale principula	2.32	2.32	2.32	0.17
Dilambdogale gheerbranti	2.38	2.38	2.38	2.21
Microgale taiva	2.47	2.47	2.47	0.13
Microgale cowani	2.62	2.62	2.62	0.57
Eremitalpa granti	3.14	3.14	3.14	9.65
Calcochloris obtusirostris	3.27	3.27	3.27	13.38
Neamblysomus julianae	3.33	3.33	3.33	5.72

Table 2.1: Body Size and Confidence Intervals in *Atlantogenata* estimated using StableTraits. *(continued)*

Node	Size (log(g))	95% CI (Low)	95% CI (High)	Rate (sqrt)
Chlorotalpa duthieae	3.38	3.38	3.38	0.32
Chlorotalpa sclateri	3.54	3.54	3.54	0.09
Macroscelides proboscideus	3.64	3.64	3.64	14.17
Chrysochloris stuhlmanni	3.74	3.74	3.74	0.33
Oryzorictes hova	3.79	3.79	3.79	22.77
Elephantulus myurus	3.81	3.81	3.81	0.95
Elephantulus brachyrhynchus	3.81	3.81	3.81	0.93
Elephantulus rozeti	3.81	3.81	3.81	10.51
Elephantulus fuscus	3.82	3.82	3.82	0.68
Elephantulus intufi	3.82	3.82	3.82	1.15
Microgale talazaci	3.88	3.88	3.88	61.40
Chrysochloris asiatica	3.89	3.89	3.89	3.34
Elephantulus edwardii	3.90	3.90	3.90	0.24
Carpitalpa arendsi	3.94	3.94	3.94	0.45
Amblysomus corriae	3.94	3.94	3.94	0.98
Amblysomus hottentotus	3.98	3.98	3.98	0.02
Elephantulus fuscipes	4.04	4.04	4.04	1.93
Elephantulus rufescens	4.05	4.05	4.05	0.12
Neamblysomus gunningi	4.09	4.09	4.09	3.26
Elephantulus rupestris	4.12	4.12	4.12	0.32
Amblysomus septentrionalis	4.23	4.23	4.23	0.52

Table 2.1: Body Size and Confidence Intervals in *Atlantogenata* estimated using StableTraits. *(continued)*

Node	Size (log(g))	95% CI (Low)	95% CI (High)	Rate (sqrt)
Chambius kasserinensis	4.27	4.27	4.27	11.84
Amblysomus robustus	4.33	4.33	4.33	1.38
Micropotamogale lamottei	4.36	4.36	4.36	2.82
Echinops telfairi	4.47	4.47	4.47	7.75
Limnogale mergulus	4.52	4.52	4.52	121.95
Hemicentetes semispinosus	4.75	4.75	4.75	4.68
Chrysospalax villosus	4.77	4.77	4.77	0.13
Petrodromus tetradactylus	5.29	5.29	5.29	24.61
Herodotius pattersoni	5.50	5.50	5.50	11.64
Setifer setosus	5.61	5.61	5.61	12.52
Rhynchocyon cirnei	5.86	5.86	5.86	3.30
Metoldobotes sp nov	5.93	5.93	5.93	15.94
Chrysospalax trevelyani	6.13	6.13	6.13	62.84
Rhynchocyon petersi	6.15	6.15	6.15	2.13
Rhynchocyon chrysopygus	6.28	6.28	6.28	0.40
Potamogale velox	6.49	6.49	6.49	103.04
Rhynchocyon udzungwensis	6.57	6.57	6.57	4.33
Tenrec ecaudatus	6.75	6.75	6.75	79.50
Dasypus sabanicola	7.05	7.05	7.05	12.18
Tolypeutes matacus	7.11	7.11	7.11	15.96
Dasypus septemcinctus	7.30	7.30	7.30	4.44

Table 2.1: Body Size and Confidence Intervals in *Atlantogenata* estimated using StableTraits. *(continued)*

Node	Size (log(g))	95% CI (Low)	95% CI (High)	Rate (sqrt)
Zaedyus pichiy	7.31	7.31	7.31	5.54
Dasypus hybridus	7.31	7.31	7.31	4.05
Chaetophractus villosus	7.61	7.61	7.61	0.42
Chaetophractus nationi	7.67	7.67	7.67	0.09
Heterohyrax brucei	7.78	7.78	7.78	1.64
Cabassous centralis	7.92	7.92	7.92	0.25
Seggeurius amourensis	7.98	7.98	7.98	2.82
Procavia capensis	8.01	8.01	8.01	0.00
Microhyrax lavocati	8.13	8.13	8.13	0.73
Bradypus tridactylus	8.23	8.23	8.23	0.48
Bradypus torquatus	8.27	8.27	8.27	0.03
Dasypus novemcinctus	8.37	8.37	8.37	14.73
Euphractus sexcinctus	8.43	8.43	8.43	14.99
Choloepus hoffmanni	8.47	8.47	8.47	0.32
Bradypus variegatus	8.49	8.49	8.49	0.51
Tamandua tetradactyla	8.52	8.52	8.52	10.44
Cyclopes didactylus	8.53	8.53	8.53	2.15
Choloepus didactylus	8.71	8.71	8.71	0.64
Thyrohyrax meyeri	8.78	8.78	8.78	3.55
Saghatherium bowni	9.13	9.13	9.13	15.85
Dasypus kappleri	9.23	9.23	9.23	74.13

Table 2.1: Body Size and Confidence Intervals in *Atlantogenata* estimated using StableTraits. *(continued)*

Node	Size (log(g))	95% CI (Low)	95% CI (High)	Rate (sqrt)
Thyrohyrax domorictus	9.30	9.30	9.30	1.15
Dimaitherium patnaiki	9.57	9.57	9.57	18.23
Phosphatherium escuilliei	9.62	9.62	9.62	326.23
Saghatherium antiquum	9.73	9.73	9.73	2.90
Thyrohyrax litholagus	10.01	10.01	10.01	28.58
Myrmecophaga tridactyla	10.26	10.26	10.26	41.03
Myorycteropus africanus	10.27	10.27	10.27	0.57
Selenohyrax chatrathi	10.73	10.73	10.73	14.99
Priodontes maximus	10.82	10.82	10.82	268.43
Orycteropus afer	10.87	10.87	10.87	6.59
Antilohyrax pectidens	10.93	10.93	10.93	13.69
Bunohyrax fajumensis	11.32	11.32	11.32	1.45
Afrohyrax championi	11.32	11.32	11.32	0.19
Geniohyus mirus	11.33	11.33	11.33	5.44
Prorastomus sirenoides	11.49	11.49	11.49	13.61
Elephas antiquus falconeri	11.51	11.51	11.51	6.12
Pachyhyrax crassidentatus	11.81	11.81	11.81	2.29
Megalohyrax eocaenus	11.95	11.95	11.95	0.24
Elephas cypriotes	12.21	12.21	12.21	1.90
Bunohyrax major	12.36	12.36	12.36	11.39
Titanohyrax angustidens	12.48	12.48	12.48	0.04

Table 2.1: Body Size and Confidence Intervals in *Atlantogenata* estimated using StableTraits. *(continued)*

Node	Size (log(g))	95% CI (Low)	95% CI (High)	Rate (sqrt)
Daouitherium rebouli	12.80	12.80	12.80	0.74
Arcanotherium savagei	12.89	12.89	12.89	7.29
Dugong dugon	12.92	12.92	12.92	5.85
Trichechus senegalensis	13.03	13.03	13.03	0.57
Trichechus inunguis	13.08	13.08	13.08	0.69
Protosiren smithae	13.20	13.20	13.20	33.69
Numidotherium koholense	13.23	13.23	13.23	2.29
Omanitherium dhofarensis	13.35	13.35	13.35	0.03
Trichechus manatus	13.44	13.44	13.44	1.39
Moeritherium spp	13.82	13.82	13.82	5.71
Phiomia spp	13.89	13.89	13.89	3.64
Elephas maximus	15.02	15.02	15.02	5.81
Barytherium spp	15.20	15.20	15.20	73.58
Mammuthus primigenius	15.27	15.27	15.27	2.17
Mammut borsoni	16.49	16.49	16.49	15.33
Mammuthus trogontherii	16.38	16.38	16.38	16.00
Loxodonta africana	15.35	15.35	15.35	1.28
Loxodonta cyclotis	15.37	15.37	15.37	3.72
Palaeoloxodon antiquus	16.14	16.14	16.14	0.01
Palaeoloxodon namadicus	16.81	16.81	16.81	12.81
Mammut americanum	15.61	15.61	15.61	0.95

Table 2.1: Body Size and Confidence Intervals in *Atlantogenata* estimated using StableTraits. *(continued)*

Node	Size (log(g))	95% CI (Low)	95% CI (High)	Rate (sqrt)
Mammuthus columbi	15.71	15.71	15.71	0.91
Hydrodamalis gigas	15.72	15.72	15.72	172.52
Atlantogenata	5.55	4.06	7.95	0.03
Afrotheria	5.55	4.05	7.96	0.00
Afrosoricida	4.35	2.58	6.13	44.49
Macroscelidae	5.27	3.98	6.85	2.49
Pseudoungulata	9.76	5.21	12.78	545.83
Paenungulata	10.13	7.24	13.02	4.42
Tethytheria	12.60	10.25	13.81	187.47
Proboscidea	15.23	14.22	16.24	30.28
Elephantidae	15.49	14.89	16.10	2.21
Elephantina	15.51	15.08	15.96	0.01
Mammuthus	15.54	15.24	15.85	0.47
Loxodontini	15.55	15.02	16.11	0.11
Loxodona	15.72	15.16	16.30	0.86
Xenarthra	7.57	5.96	9.18	124.94

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

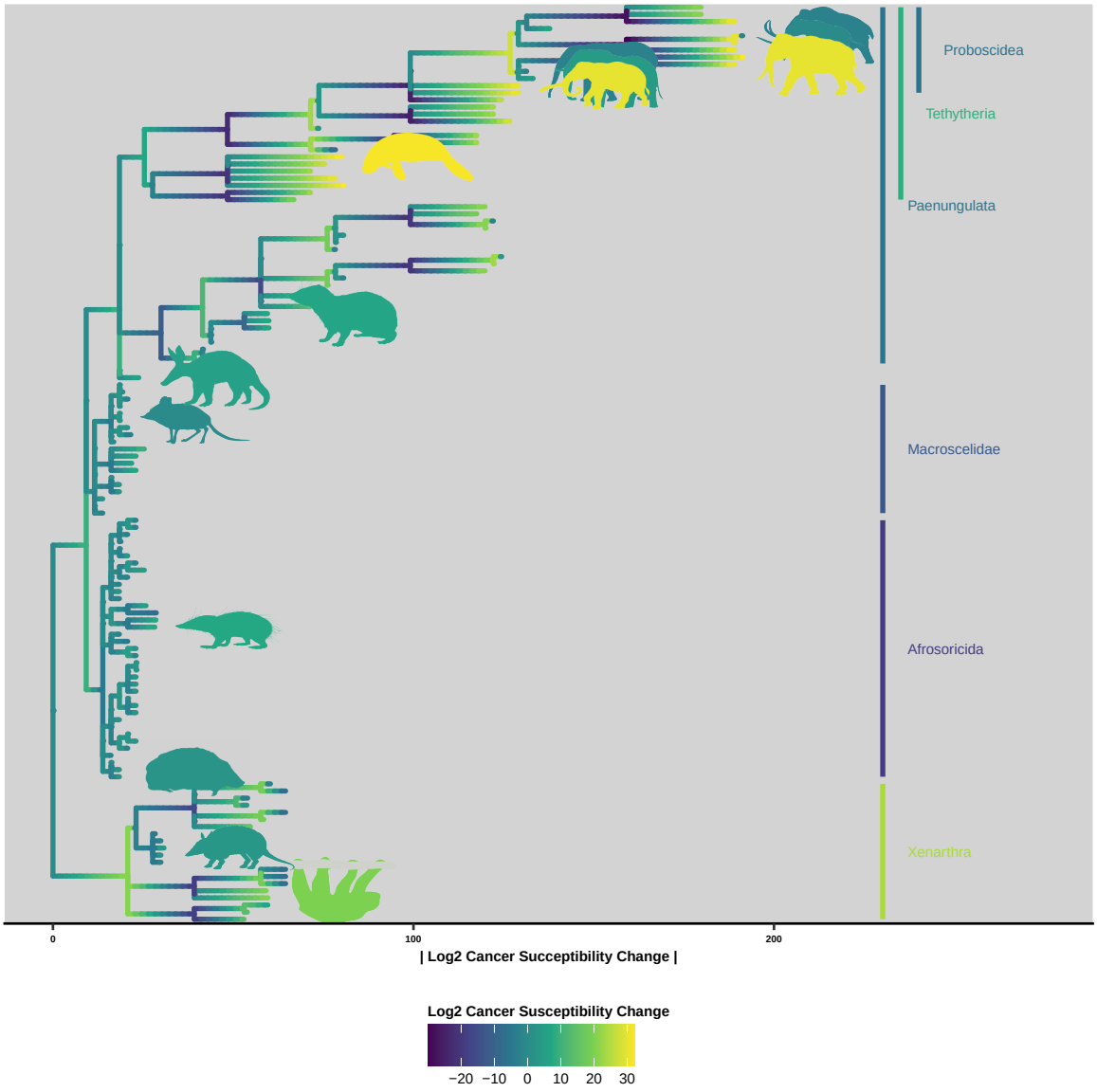

Figure 2.3: Cancer susceptibility across *Atlantogenata.* Branch lengths are set to the magnitude of change in cancer susceptibility; colors indicate the magnitude and direction of the change.

As expected, intrinsic cancer susceptibility in *Afrotheria* also varies with changes in body size and longevity (Figure 2.3, Table 2.2), with an initial 9.22-fold and 20.75-fold increase in the stem-lineage of *Afrotheria* and *Xenarthra*, respectively, followed by a 9.22-fold increases in Pseudoungulata and a 5.38-fold increase in Aardvarks (Figure 2.3, Table 2.2). In contrast to the *Paenungulate* stem-lineage, there is a 6.92-fold increase in cancer risk in Tethytheria, a 8.62-fold increase in Manatee, and dramatic increases within *Proboscidea* including a 27.66-fold increase in Elephantidae and a 29.97-fold in the American Mastodon. Within the Elephantidae, Elephantina and Loxodontini have a 2.31-fold increase 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 both size and cancer susceptibility (Figure 2.3, Table 2.2).

Table 2.2: Estimated Cancer Susceptibility for nodes in *Atlantogenata*

Node	Est. Lifespan	K1	K2	Change in K	log2 Change
Loxodontini	34.38	1.47e+16	2.97e+15	4.94e+00	2.31
Loxodonta africana	65.00	2.47e+17	1.47e+16	1.68e+01	4.07
Loxodona	34.38	1.47e+16	1.47e+16	1.00e+00	0.00
Loxodonta cyclotis	31.12	2.97e+15	1.47e+16	2.02e-01	-2.31
Palaeoloxodon antiquus	34.38	1.47e+16	1.47e+16	1.00e+00	0.00
Elephantidae	31.12	2.97e+15	1.40e+07	2.13e+08	27.66
Elephantina	34.38	1.47e+16	2.97e+15	4.94e+00	2.31
Elephas maximus	65.50	2.58e+17	1.47e+16	1.76e+01	4.14
Mammuthus	34.38	1.47e+16	1.47e+16	1.00e+00	0.00
Mammuthus primigenius	31.12	2.97e+15	1.47e+16	2.02e-01	-2.31
Mammuthus columbi	34.38	1.47e+16	1.47e+16	1.00e+00	0.00
Proboscidea	9.41	1.40e+07	1.21e+14	1.15e-07	-23.05
Mammut americanum	34.38	1.47e+16	1.40e+07	1.05e+09	29.97
Tethytheria	25.49	1.21e+14	1.01e+12	1.21e+02	6.92
Trichechus manatus	69.00	4.77e+16	1.21e+14	3.93e+02	8.62
Paenungulata	18.91	1.01e+12	1.01e+12	1.00e+00	0.00
Procavia capensis	14.80	3.13e+10	1.01e+12	3.11e-02	-5.01
Pseudoungulata	18.91	1.01e+12	1.69e+09	5.97e+02	9.22
Orycteropus afer	29.80	4.19e+13	1.01e+12	4.17e+01	5.38
Elephantulus edwardii	10.40	6.90e+07	1.69e+09	4.09e-02	-4.61
Afrosoricida	10.40	6.90e+07	1.69e+09	4.09e-02	-4.61
Chrysochloris asiatica	10.40	6.90e+07	6.90e+07	1.00e+00	0.00
Echinops telfairi	19.00	2.57e+09	6.90e+07	3.72e+01	5.22
Afrotheria	12.69	1.69e+09	2.83e+06	5.97e+02	9.22
Xenarthra	20.89	4.97e+12	2.83e+06	1.76e+06	20.75
Dasypus novemcinctus	22.30	3.67e+11	4.97e+12	7.37e-02	-3.76
Choloepus hoffmanni	41.00	1.42e+13	4.97e+12	2.85e+00	1.51
Atlantogenata	8.52	2.83e+06	2.83e+06	1.00e+00	0.00
Afroinsectivora	12.69	1.69e+09	1.69e+09	1.00e+00	0.00

2.3.3 *Identification and evolutionary history of gene duplications*

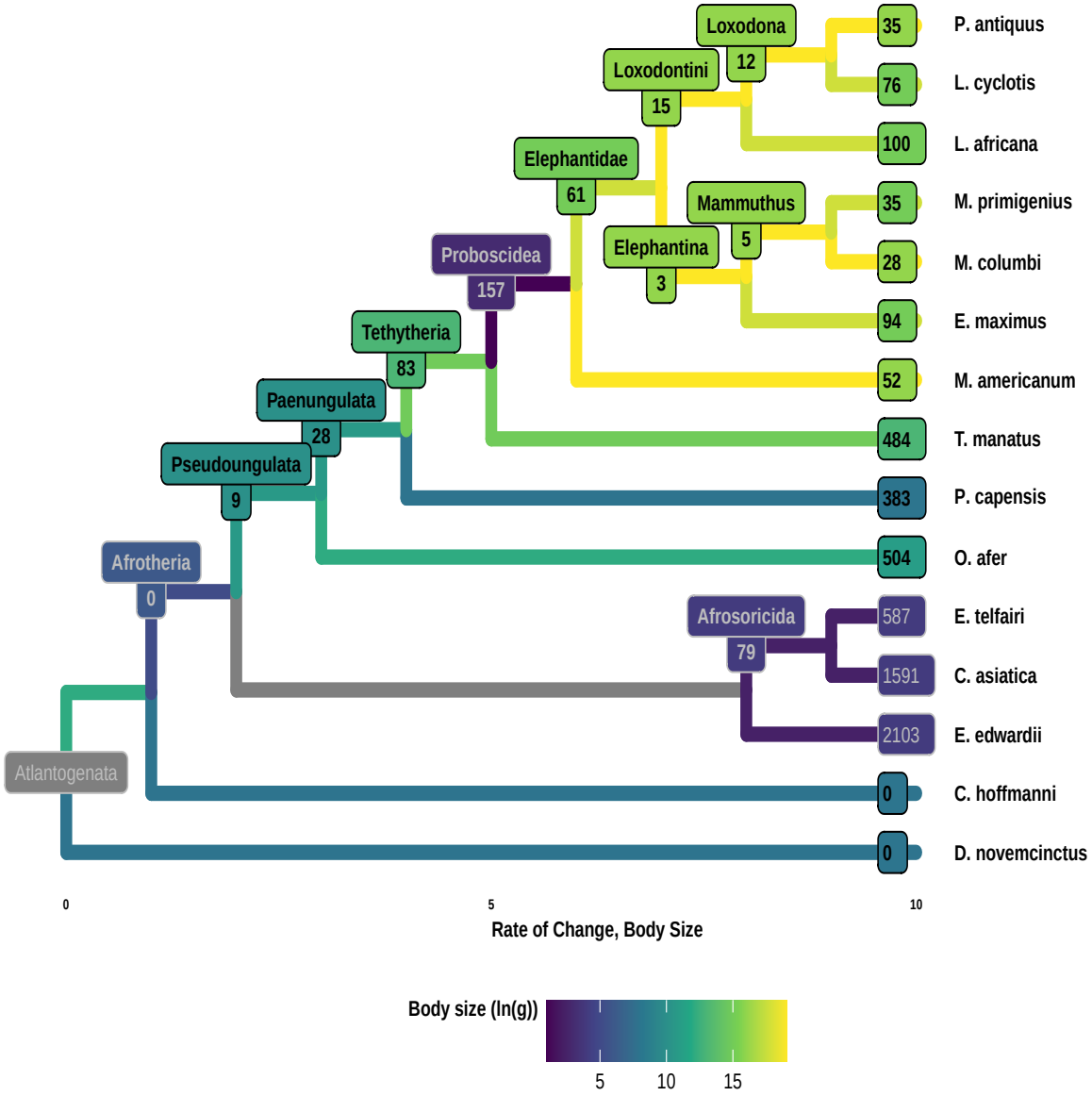

Figure 2.4: Gene duplications occur readily throughout *Atlantogenata*. Shown here is a tree of *Atlantogenatan* species with genomes, with the number of genes that underwent an increase in copy number overlayed at each node.

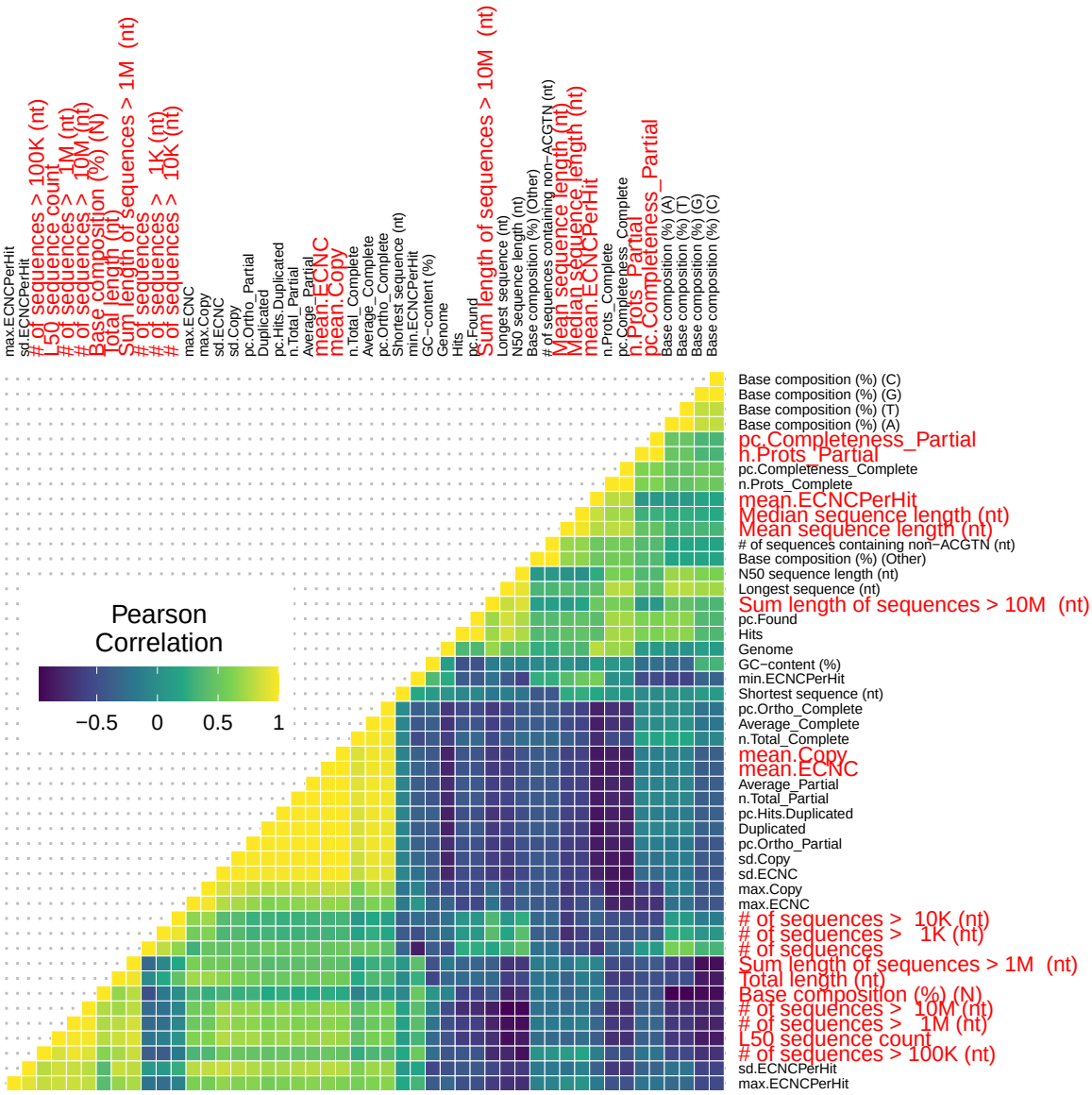

Figure 2.5: Correlations between genome quality metrics and ECNC metrics. Gene copy number metrics, and the genome quality metrics most strongly associated with them, are highlighted in red.

We found that gene duplications were common in *Atlantogenatan* genomes (Figure 2.4, Table 2.3), identifying an average of 76.2% genes, with an average of 10.53% duplicated; this is in-line with other studies describing the rates of gene duplications over time [106]. We observed that the percentage of duplicated genes in non-*Pseudoungulatan* genomes was

Table 2.3: Summary of duplications in *Atlantogenata*

Species	Common Name	Size (g)	#Hits	#Duplicated	% Genes Found	% Hits Duplicated	Mean ECNC/Hit
Choloepus hoffmanni	Hoffmans Two-Toed Sloth	4.3e+03	14082	3204	78.19%	22.75%	0.98
Chrysochloris asiatica	Cape Golden Mole	49	13547	2716	75.22%	20.05%	0.99
Dasypus novemcinctus	Nine-Banded Armadillo	4.8e+03	13819	2605	76.73%	18.85%	0.98
Echinops telfairi	Lesser Hedgehog Tenrec	87	12903	1670	71.64%	12.94%	0.99
Elephantulus edwardii	Cape Elephant Shrew	49	12884	3048	71.53%	23.66%	0.99
Elephas maximus	Asian Elephant	3.3e+06	14073	907	78.14%	6.44%	1.00
Loxodonta africana	African Savanna Elephant	4.6e+06	14051	940	78.01%	6.69%	1.00
Loxodonta cyclotis	African Forest Elephant	4.7e+06	14065	900	78.09%	6.40%	1.00
Mammut americanum	American Mastodon	6e+06	13840	737	76.84%	5.33%	1.00
Mammuthus columbi	Columbian Mammoth	6.6e+06	13059	426	72.51%	3.26%	1.00
Mammuthus primigenius	Woolly Mammoth	4.3e+06	13935	723	77.37%	5.19%	1.00
Orycteropus afer	Aardvark	5.3e+04	13880	1083	77.06%	7.80%	0.99
Palaeoloxodon antiquus	Straight Tusked Elephant	1e+07	13969	745	77.56%	5.33%	1.00
Procavia capensis	Rock Hyrax	3e+03	13672	788	75.91%	5.76%	1.00
Trichechus manatus	Manatee	6.9e+05	14092	1046	78.24%	7.42%	1.00

significantly higher: while *Pseudoungulatan* genomes had duplicates percentages ranging anywhere from 3.26% to 7.80%, outgroup species' duplication rates ranged from 12.94% to 23.66%. To explore whether genome quality may adversely effect our inferences, we used CEGMA and the gVolante server [136, 123] to assess the correlation between genome quality and copy number estimates. As shown in Figure 2.5, mean Copy Number, mean ECNC, and mean CN (the lesser of Copy Number and ECNC per gene) moderately or strongly correlate with genomic quality, such as LD50, the number of scaffolds, and contigs with a length above either 100K or 1M.

Among the genes that increased in copy number in the elephant lineage are *TP53* and *LIF*, as previously described; however, we find that these two genes represent a fraction of the 940 genes that are duplicated in the African Elephant overall, which accumulated over various steps through their evolution. While the extinct elephantids have acceptable genome quality metrics according to CEGMA, they are nonetheless missing a significant number of sequences; this may contribute to the low number of duplicated genes that occured in internal nodes. The number of duplicates that occur at each branch is also proportional to the density of the sampling of the clade overall, as would be expected. In branches, such as *Afrosoricida*, where the number of species is relatively minuscule compared to the size of the clade, we see many significantly larger numbers of duplications private to these species.

2.3.4 Duplications that occurred recently in Probodiscea are enriched for tumor suppressor pathways

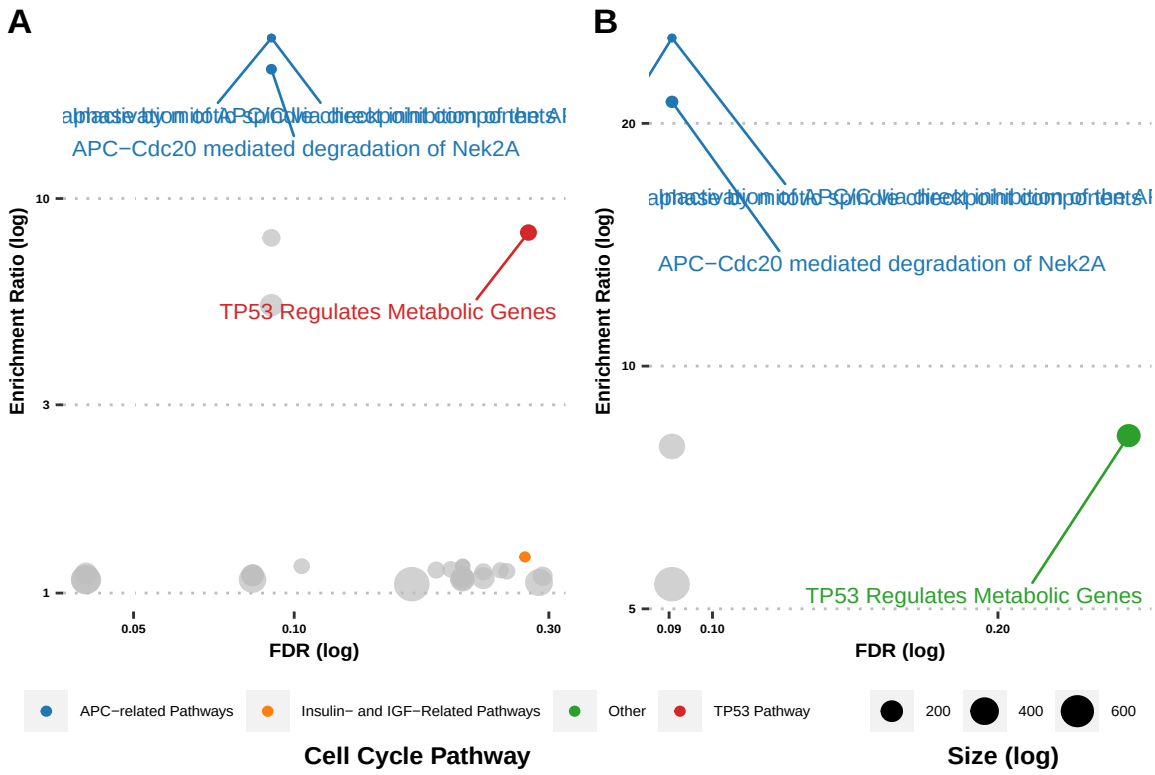

Figure 2.6: Overrepresentation Analysis of Duplicated Genes in *Atlantogenata* using Reactome Pathways.

Our initial hypothesis was that genes which duplicated in lineages that experienced a growth in size would be enriched for membership in tumor suppressor pathways. Thus, we used WebGestalt and its Overrepresentation Analysis (ORA) functionality to determine what pathways were enriched in our duplicated gene sets in each branch relative to our initial query set. For our database, we used Reactome for our primary analysis, but additionally used the KEGG, Panther, Wikipathways, and Wikipathways_cancer databases using WebGestalt ORA. Going through the tree, at no FDR ≤ 0.5 is there any significant pathway representation for genes that increased in the branches leading to *Afrosoricida*, *Pseudoungulata*, *Elephantina*,

Loxodontini, *Elephas maximus*, *Mammuthus*, *Loxodonta cyclotis*, *Palaeoloxodon antiquus*, *Loxodona*, *Mammuthus columbi*, *Mammuthus primigenius*, *Procavia capensis*, *Elephantidae*, or *Proboscidea*; furthermore, there are no significant pathway enrichments at FDR<0.5 for genes whose copy number did not change between branches (copy-number-stable) for *Afrotheria*. Note that because *Xenarthra* was selected as the outgroup, it is not possible to polarize the changes in their gene copy numbers along the tree.

For the other branches, the number of pathways that came up as significantly enriched at each FDR is shown in Table 2.4. For the species with high duplication rates and lower-quality, highly-fragmented genomes, such as with *Chrysochloris asiatica* (20.05% duplicated hits) and *Elephantulus edwardii* (23.66% duplicated hits), it is unsurprising that there is a proportionally large number of pathway enrichments. In the case of these two species, their many pathway enrichments also span an incredible range of processes at every level of biology; this, in combination with the high number of copy numbers identified for these genes, further suggests a need for improvement and refinement in these genomes. In the cell cycle pathways called as significant in the genomes of these two species plus *Orycteropus afer* and *Echinops telfairi*, the duplicated genes included in these sets are from the same gene families, such as the APC subunit family; the proteosome subunit families; and the protein phosphatase 2 family, among others. It is highly possible that these results reflect true expansions of these gene families, especially in the higher-quality *Orycteropus afer* genome; however, it is also possible that it simply reflects artifactual duplications, and so require further study.

Table 2.4: Number of pathways overrepresented among duplicated genes at different FDRs.

Ancestor	Node	Pathways at FDR≤ 0.1	Pathways at FDR≤ 0.2	Pathways at FDR≤ 0.3	Pathways at FDR≤ 0.5
Afroinsectivora	Elephantulus edwardii	252	37	30	87
Afrosoricida	Chrysochloris asiatica	90	48	43	105
Afrosoricida	Echinops telfairi	0	2	0	31
Afrotheria	Afroinsectivora	0	0	0	33
Loxodontini	Loxodonta africana	6	0	1	0
Paenungulata	Tethytheria	0	0	0	2
Proboscidea	Mammut americanum	0	0	0	6
Pseudoungulata	Orycteropus afer	27	67	29	67
Tethytheria	Proboscidea	0	3	0	3
Tethytheria	Trichechus manatus	4	0	0	2

The pathway enrichments for genes whose copy number either did not change, or whose copy number increased, between *Loxodona* and the African Elephant are shown in Figure 2.6. Among the few enriched pathways in the African Elephant, we see that that two tumor suppression pathways - APC Complex-related pathways, and "TP53 Regulates Metabolic Genes"- appear not only in the case of stable genes, but also in the set of newly duplicated genes. The other pathways we see in the set of recently-duplicated genes include "Functionalization of Compounds" and its daughter pathway "Xenobiotics". Genes in these pathways serve to add functional groups to lipophylic compounds which would otherwise not be reactive in the cell, and are types of metabolic pathways. In the stable set we see enrichment of pathways such as "Neuronal Systems" and "Axon Guidance," which fit in well with what is known about elephant biology and evolution [53]. Overall in elephants, we see enrichments within duplicated genes for pathways involved in what makes an elephant an elephant - including tumor suppressor pathways.

2.3.5 Concerted duplication of TP53 and TP53-related genes towards Probodiscea

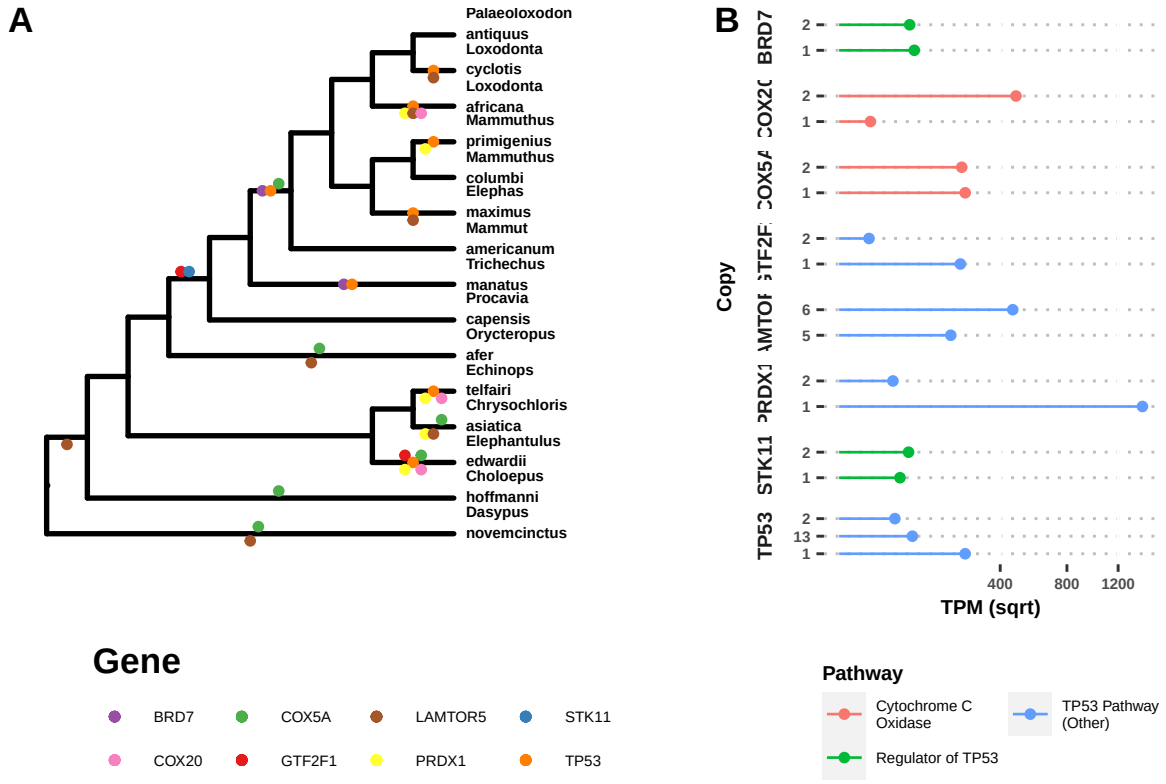

Figure 2.7: TP53-related genes are also duplicated and functional in *Loxodonta africana*. **A)** Cladogram of *Atlantogenata* highlighting along each branch when duplications for each gene occured. **B)** Gene expression data collected from pulically-available RNA-seq data for each duplicate in **A)**.

Prior studies looking at the duplication of TP53 in the African Elephant motivated further study for the enrichment of genes involved in TP53-related metabolic pathway. We traced the evolution of all genes involved in this TP53-duplicated pathway that were duplicated in the African Elephant, and used publicly-available RNA-seq data to see which genes are actively expresses in living elephants. Excitingly, 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 coincided by the duplication of BRD7. These two genes are involved in regulating the transcription of TP53, and their duplication prior to that of TP53 may have facilitated its retroduplication. Interestingly enough, STK11 is a tumor suppressor gene in its own right, and plays additional roles in mediating tumor suppression via p21-induced senescence [98].

The other genes that are duplicated in the pathway are all downstream of TP53; these genes duplicated either alongside TP53 in the case of SIAH1, or subsequently in *Probodiscea*, *Elephantidae*, and in modern elephants. (Figure 2.7). These genes are all expressed in RNA-seq data, suggesting that they encode functional genes in modern elephants (Figure 2.7).

2.4 Discussion

With our results, we have demonstrated that gigantism in *Atlantogenata* was not limited to extant elephants, but rather occurred at various points in the evolution of the clade; however, the hundred-fold to hundred-million-fold increases in cancer risk that is associated with these increases however poses an innate challenge in the evolution and persistence of this trait. The environmental selective pressures on body size have long been the fascination of evolutionary biologists, and the influence of climate, predation, geography, and ecological niche on body size have been well established. [25]. Indeed, there is a general trend, known as Cope's Rule, for body sizes of species to increase over time [71, 76, 79]. However, for all the research on body size that has been done thus far, the mechanisms that enable a release from the negative pressure on body size exerted on cancer risk has proven more elusive [71, 76, 79].

Furthermore, we show that tumor suppressor duplications are enriched not only in large, extant species, but also in large common ancestors, and that these duplications evolved throughout the tree, rather than in concert. The ancestral body sizes of many of the subclades in *Atlantogenata* were estimated to be large, and the estimated cancer risk increases - even

for the small clades like *Afroinsectivora* may explain why these TSG gene duplications occurred early on, and persisted even in species that we did not expect to have high risks of cancer. However, some of our results also provide interesting evidence for a paradigm of TSG duplications being a pervasive phenomenon, rather than a specific mechanism that occurs after the evolution of gigantism. At the common ancestor of *Proboscidea*, which was small relative to both its ancestors and its descendants, we see the emergence of various TSG duplications, which may have enabled the stratospheric increase in body size of modern elephants. We also identified many TSG duplication events in smaller species and lineages such as in *Chrysochloris asiatica* and *Elephantulus edwardii*, although these may have been the result of low-quality genomes.

The impacts and takeaways of this study are limited by the quality and quantity of *Atlantogenatan* genomes that were available, and our available knowledge of the lifespans and cancer risk of the extant species. For many of our species, studies in captivity are limited, and the species are relatively understudied from a longitudinal perspective, such as with the Cape Elephant Shrew and Cape Golden Mole. Furthermore, while there is recent interest in resequencing and improving the quality of these assemblies, at the time of this writing there is still quite a ways to go in order to have genomes of a sufficiently rigorous quality to make stronger inferences about gene copy number expansions and contractions (which were not considered in this study for this reason).

The lack of a stronger signal from tumor suppressor duplications is likely a result of the strong effect size on both cancer risk and organismal toxicity that a TSG duplication would provide. The duplication of a tumor suppressor in many cases is associated with mild toxicity, although it greatly varies given the context and TSG in question; however, a single TSG duplication can also provide significant protection against cancer. For example, the overexpression of TP53 in mice, while protective of cancer, is associated with progeria and early death; however, if an additional copy of TP53 is introduced with its regulatory

elements intact, the mice are healthy and experience normaging, while also demonstrating an enhanced response to cellular stress and lower rates of cancer. [178, 52]. In light of this, it is fascinating that our results in the elephant lineage suggest that the duplication of TP53 regulators preceded the retroduplication and expansion of TP53, as this likely would have lowered the toxicity of the initial duplication and thus enabled it to occur. Given a sufficient selective pressure on increasing body size, it stands to reason that events like this could alleviate the negative pleiotropy of TSG duplications sufficiently to enable their persistence and allow for subsequent refinement over evolutionary time.

By combining a phylogenetic study on body size in addition to a survey of copy number across nearly all protein-coding genes, we provide a comprehensive look at the question of the role of cancer risk and body size in *Atlantogenata* that may provide broader insight to other mammalian species. Our study was initially motivated by the identification of functional duplicates of tumor suppressors, such as TP53 and LIF in elephants [1, 170, 180]. Further in support of our results, a larger candidate gene study by Caulin et al. [20] characterized the copy number of 830 known tumor-suppressor genes across 36 mammals and identified 382 putative duplicates, including duplicates in species with large body sizes and long life-spans. However, while candidate gene studies are useful, by their very design they are biased in determining larger patterns of evolution of traits. Without addressing these questions with a genome-wide approach, any and all insights will be inevitably limited to a fraction of the whole story.

Our results suggest that the pervasive duplication of tumor suppressors may help enable the evolution of larger body sizes by lowering the cancer risk of species, either prior to or in lockstep with increasing body size. however, this is unlikely to be the only genetic mechanism at play in this scenario. In genome-wide studies of unusually large or long-lived species such as the bowhead whale [89], Myotid bats [158, 190], naked mole rat [93], and blind mole rat [43], there were cases of overrepresentation of TSGs among duplicate genes that were

outshadowed by the identification of strong signatures of positive selection at TSGs. While the evolution of regulatory and coding elements of both TSGs and other non-canonical tumor suppressor genes have been shown to be important for mediating the cancer risk of long-lived species, there has been no attention given to the possibility of TSG duplications providing a relaxation of possible negative pleitropy that could result from these traits. It has been well-established in the literature that genes duplication events allow for evolutionary drift in one of the copies, which may result in neofunctionalization or specialization of the two copies [147, 146, 167]. While this is beyond the scope of our study, a promising future direction of this work would include an evolutionary analysis of duplicated genes relative to each other to see if this has already occurred between the pairs of genes we have identified.

2.5 Supplementary Figures

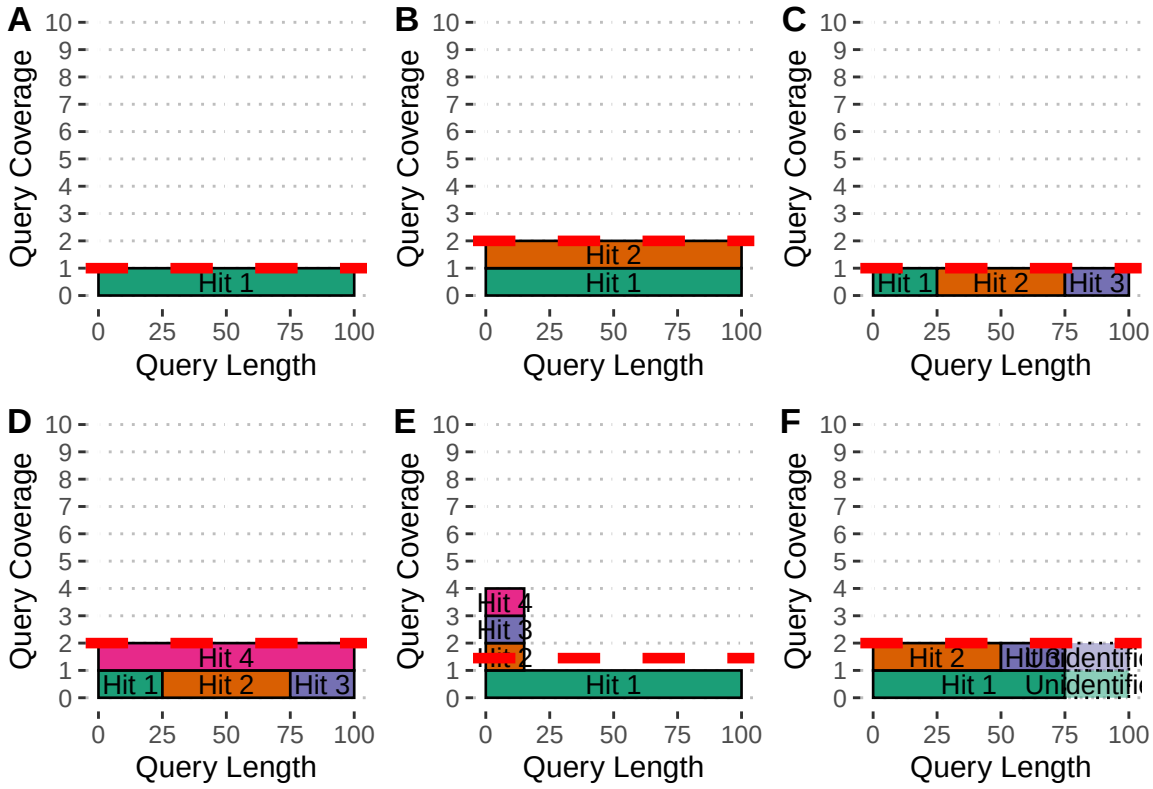

Figure S.2.1: Estimated Copy Number by Coverage (ECNC) consolidates fragmented genes while accounting for missing domains in homologs. **A)** A single, contiguous gene homolog in a target genome with 100% query length coverage has an ECNC of 1.0. **B)** Two contiguous gene homologs, each with 100% query length coverage have an ECNC of 2.0. **C)** A single gene homolog, split across multiple scaffolds and contigs in a fragmented target genome; BLAT identifies each fragment as a single hit. Per nucleotide of query sequence, there is only one corresponding nucleotide over all the hits, thus the ECNC is 1.0. **D)** Two gene homologs, one fragmented and one contiguous. 100% of nucleotides in the query sequence are represented between all hits; however, every nucleotide in the query has two matching nucleotides in the target genome, thus the ECNC is 2.0. (Continued on next page)

Figure S.2.1: (Continued from prior page) **E)** One true gene homolog in the target genome, plus multiple hits of a conserved domain that span 20% of the query sequence. While 100% of the query sequence is represented in total, 20% of the nucleotides have 4 hits. Thus, the ECNC for this gene is 1.45. **F)** Two real gene homologs; one hit is contiguous, one hit is fragmented in two, and the tail end of both sequences was not identified by BLAT due to sequence divergence. Only 75% of the query sequence was covered in total between the hits, but for that 75%, each nucleotide has two hits. As such, ECNC is equal to 2.0 for this gene.

2.6 Supplementary Tables

Table S.2.1: NCBI SRA datasets used in this study, along with key biological and genome information.

Organism	Common Name	Genome	SRA Acc.	Tissues
Dasypus novemcinctus	Nine-banded armadillo	dasNov3	SRR494779, SRR494767, SRR494780, SRR494770, SRR309130, SRR494771, SRR4043756, SRR494776, SRR494778, SRR4043762, SRR4043755, SRR6206923, SRR4043761, SRR4043760, SRR6206913, SRR4043763, SRR494772, SRR494781, SRR494774, SRR494777, SRR494775, SRR4043754, SRR1289524, SRR4043758, SRR6206903, SRR1289523, SRR4043759, SRR3222425, SRR494768, SRR494769, SRR6206908, SRR4043757, SRR494766, SRR6206918, SRR494773	Kidney, Spleen, Cerebellum W/ Brainstem, Rt. Quadricep, Mid-Stage Pregnant Endometrium, Cervix, Lung, Liver, Skeletal Muscle, Ascending Colon, Pregnant Armadillo Endometrium, Heart, Placenta
Loxodonta africana	African savanna elephant	loxAfr3, loxAfrC, loxAfr4	SRR6307198, SRR1041765, SRR6307199, SRR6307201, SRR6307196, SRR6307202, SRR6307200, SRR6307195, SRR975188, SRR6307194, SRR6307204, SRR3222430, SRR6307205, SRR975189, SRR6307197, SRR6307203	Blood, Fibroblast, Placenta
Trichechus manatus latirostris	Manatee	triMan1, triMan-Lat2	SRR4228542, SRR4228545, SRR4228544, SRR4228539, SRR4228541, SRR4228538, SRR4228546, SRR4228537, SRR4228540, SRR4228543, SRR4228547	Buffy Coat

Table S.2.2: Genomes used in this study.

Species	Common Name	Genomes	Highest Quality Genome	Citation
Choloepus hoffmanni	Hoffmans two-toed sloth	choHof1, choHof2, choHof-C_hoffmanni-2.0.1_HiC	choHof-C_hoffmanni-2.0.1_HiC	39
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	133
Loxodonta africana	African savanna elephant	loxAfr3, loxAfrC, loxAfr4	loxAfr4	133, Broad/loxAfr4.0[1]
Loxodonta cyclotis	African forest elephant	loxCycF	loxCycF	133
Mammut americanum	American mastodon	mamAmeI	mamAmeI	133
Mammuthus columbi	Columbian mammoth	mamColU	mamColU	133
Mammuthus primigenius	Woolly mammoth	mamPriV	mamPriV	134
Orycteropus afer	Aardvark	oryAfe1, OryAfe1.0_HiC	OryAfe1.0_HiC	39
Palaeoloxodon antiquus	Straight tusked elephant	palAntN	palAntN	133
Procavia capensis	Rock hyrax	proCap1, proCap2, proCap-Pcap_2.0_HiC	proCap-Pcap_2.0_HiC	39, 103
Trichechus manatus latirostris	Manatee	triMan1, TriMan-Lat1.0_HiC	TriManLat1.0_HiC	39, 47

1. Available at ftp://ftp.broadinstitute.org/pub/assemblies/ mammals/elephant/loxAfr4

CHAPTER 3

A ZOMBIE LIF GENE IN ELEPHANTS IS UP-REGULATED BY TP53 TO INDUCE APOPTOSIS IN RESPONSE TO DNA DAMAGE

3.1 Introduction

The risk of developing cancer places severe constraints on the evolution of large body sizes and long lifespans in animals. If all cells have a similar risk of malignant transformation and equivalent cancer suppression mechanisms, organism with many cells should have a higher risk of developing cancer than organisms with fewer cells. Similarly organisms with long lifespans have more time to accumulate cancer-causing mutations than organisms with shorter lifespans and therefore should also be at an increased risk of developing cancer, a risk that is compounded in large-bodied, long-lived organisms [17, 21, 36, 141, 140]. Consistent with these expectations, there is a strong positive correlation between body size and cancer incidence *within* species. Larger dog breeds, for example, have higher rates of cancer than smaller breeds [34] and human cancer incidence increases with increasing adult height for numerous cancer types [56]. In stark contrast, there are no correlations between body size or lifespan and cancer risk *between* species [1]; this lack of correlation is often referred to as 'Peto's Paradox' [21, 99, 140].

While the ultimate resolution to Peto's paradox is that large bodied and/or long-lived species evolved enhanced cancer protection mechanisms, identifying and characterizing those mechanisms is essential for elucidating how enhanced cancer resistance and thus large bodies and long lifespans evolved. Numerous and diverse mechanisms have been proposed to resolve Peto's paradox [21, 31, 88, 99, 108, 122, 125, 175], but discovering those mechanisms has been challenging because the ideal study system is one in which a large, long-lived species is deeply nested within a clade of smaller, short-lived species – all of which have sequenced genomes.

Unfortunately, few lineages fit this pattern. Furthermore while comparative genomics can identify genetic changes that are phylogenetically associated the evolution of enhanced cancer protection, determining which of those genetic changes are causally related to cancer biology through traditional reverse and forward genetics approaches are not realistic for large species such as whales and elephants. Thus we must use other methods to demonstrate causality.

Among the most parsimonious mechanisms to resolve Peto's paradox are a reduced number of oncogenes and/or an increased number of tumor suppressor genes [21, 99, 125], but even these relatively simple scenarios are complicated by transcriptional complexity and context dependence. The multifunctional interleukin-6 class cytokine *leukemia inhibitory factor* (*LIF*), for example, can function as either a tumor suppressor or an oncogene depending on the context. Classically *LIF* functions as an extracellular cytokine by binding the *LIF* receptor (*LIFR*) complex, which activates downstream *PI3K/AKT*, *JAK/STAT3*, and *TGF*β signaling pathways. The *LIF* gene encodes at least three transcripts, *LIF-D*, *LIF-M*, and *LIF-T*, which contain alternative first exons spliced to common second and third exons [59, 67, 148, 181]. Remarkably while the *LIF-D* and *LIF-M* isoforms are secreted proteins that interact with the *LIF* receptor [148, 181], the *LIF-T* isoform lacks the propeptide sequence and is an exclusively intracellular protein [59, 181] that induces caspase-dependent apoptosis through an unknown mechanism [60].

Here we show that the genomes of *Paenungulates* (elephant, hyrax, and manatee) contain numerous duplicate *LIF* pseudogenes, at least one (*LIF6*) of which is expressed in elephant cells and is up-regulated by *TP53* in response to DNA damage. *LIF6* encodes a separation of function isoform structurally similar to *LIF-T* that induces apoptosis when overexpressed in multiple cell types and is required for the elephant-specific enhanced cell death in response to DNA-damage. These results suggest that the origin of a zombie *LIF* gene (a reanimated pseudogene that kills cells when expressed) may have contributed to the evolution of enhanced cancer resistance in the elephant lineage and thus the evolution large body sizes and long

lifespans.

3.2 Methods

3.2.1 Identification of LIF genes in Mammalian genomes

We used BLAT to search for *LIF* genes in 53 *Sarcopterygian* genomes using the human *LIF* protein sequences as an initial query. After identifying the canonical *LIF* gene from each species, we used the nucleotide sequences corresponding to this *LIF* CDS as the query sequence for additional BLAT searches within that species genome. To further confirm the orthology of each *LIF* gene we used a reciprocal best BLAT [90] approach, sequentially using the putative CDS of each *LIF* gene as a query against the human genome; in each case the query gene was identified as *LIF*. Finally we used the putative amino acid sequence of the *LIF* protein as a query sequence in a BLAT search.

We thus used BLAT to characterize the *LIF* copy number in Human (*Homo sapiens*; GRCh37/hg19), Chimp (*Pan troglodytes*; CSAC 2.1.4/panTro4), Gorilla (*Gorilla gorilla gorilla*; gorGor3.1/gorGor3), Orangutan (*Pongo pygmaeus abelii*; WUGSC 2.0.2/ponAbe2), Gibbon (*Nomascus leucogenys*; GGSC Nleu3.0/nomLeu3), Rhesus (*Macaca mulatta*; BGI CR_1.0/rheMac3), Baboon (*Papio hamadryas*; Baylor Pham_1.0/papHam1), Marmoset (*Callithrix jacchus*; WUGSC 3.2/calJac3), Squirrel monkey (*Saimiri boliviensis*; Broad/saiBol1), Tarsier (*Tarsius syrichta*; Tarsius_syrichta2.0.1/tarSyr2), Bushbaby (*Otolemur garnettii*; Broad/otoGar3), Mouse lemur (*Microcebus murinus*; Broad/micMur1), Chinese tree shrew (*Tupaia chinensis*; TupChi_1.0/tupChi1), Squirrel (*Spermophilus tridecemlineatus*; Broad/speTri2), Mouse (*Mus musculus*; GRCm38/mm10), Rat (*Rattus norvegicus*; RGSC 5.0/rn5), Naked mole-rat (*Heterocephalus glaber*; Broad HetGla_female_1.0/hetGla2), Guinea pig (*Cavia porcellus*; Broad/cavPor3), Rabbit (*Oryctolagus cuniculus*; Broad/oryCun2), Pika (*Ochotona princeps*; OchPri3.0/ochPri3), Kangaroo rat (*Dipodomys ordii*; Broad/dipOrd1),

Chinese hamster (*Cricetulus griseus*; C_griseus_v1.0/criGri1), Pig (*Sus scrofa*; SGSC Sscrofa10.2/susScr3), Alpaca (*Vicugna pacos*; Vicugna_pacos-2.0.1/vicPac2), Dolphin (*Tursiops truncatus*; Baylor Ttru_1.4/turTru2), Cow (*Bos taurus*; Baylor Btau_4.6.1/bosTau7), Sheep (*Ovis aries*; ISGC Oar_v3.1/oviAri3), Horse (*Equus caballus*; Broad/equCab2), White rhinoceros (*Ceratotherium simum*; CerSimSim1.0/cerSim1), Cat (*Felis catus*; ICGSC Felis_catus 6.2/felCat5), Dog (*Canis lupus familiaris*; Broad CanFam3.1/canFam3), Ferret (*Mustela putorius furo*; MusPutFur1.0/musFur1), Panda (*Ailuropoda melanoleuca*; BGI-Shenzhen 1.0/ailMel1), Megabat (*Pteropus vampyrus*; Broad/pteVam1), Microbat (*Myotis lucifugus*; Broad Institute Myoluc2.0/myoLuc2), Hedgehog (*Erinaceus europaeus*; EriEur2.0/eriEur2), Shrew (*Sorex araneus*; Broad/sorAra2), Minke whale (*Balaenoptera acutorostrata scammoni*; balAcu1), Bowhead Whale (*Balaena mysticetus*; v1.0), Rock hyrax (*Procavia capensis*; Broad/proCap1), Sloth (*Choloepus hoffmanni*; Broad/choHof1), Elephant (*Loxodonta africana*; Broad/loxAfr3), Cape elephant shrew (*Elephantulus edwardii*; EleEdw1.0/eleEdw1), Manatee (*Trichechus manatus latirostris*; Broad v1.0/triMan1), Tenrec (*Echinops telfairi*; Broad/echTel2), Aardvark (*Orycteropus afer afer*; OryAfe1.0/oryAfe1), Armadillo (*Dasypus novemcinctus*; Baylor/dasNov3), Opossum (*Monodelphis domestica*; Broad/monDom5), Tasmanian devil (*Sarcophilus harrisii*; WTSI Devil_ref v7.0/sarHar1), Wallaby (*Macropus eugenii*; TWGS Meug_1.1/macEug2), and Platypus (*Ornithorhynchus anatinus*; WUGSC 5.0.1/ornAna1).

3.2.2 Phylogenetic analyses and gene tree reconciliation of Paenungulate LIF genes

The phylogeny of *LIF* genes was estimated using an alignment of the *LIF* loci from the African elephant, hyrax, manatee, tenrec, and armadillo genomes and BEAST (v1.8.3) [150]. We used the HKY85 substitution, which was chosen as the best model using HyPhy, empirical nucleotide frequencies (+F), a proportion of invariable sites estimated from the

data (+I), four gamma distributed rate categories (+G), an uncorrelated random local clock to model substitution rate variation across lineages, a Yule speciation tree prior, uniform priors for the GTR substitution parameters, gamma shape parameter, proportion of invariant sites parameter, and nucleotide frequency parameter. We used an Unweighted Pair Group Arithmetic Mean (UPGMA) starting tree. The analysis was run for 10 million generations and sampled every 1000 generations with a burn-in of 1000 sampled trees; convergence was assessed using Tracer, which indicated convergence was reached rapidly (within 100,000 generations). We used Notung v2.6 [22] to reconcile the gene and species trees.

3.2.3 Gene expression data (Analyses of RNA-Seq data and RT-PCR)

To determine if duplicate *LIF* genes were basally transcribed, we assembled and quantified elephant *LIF* transcripts with HISAT2 and StringTie [92, 139, 138] using deep 100bp paired-end RNA-Seq data (over 138 million reads) we previously generated from Asian elephant dermal fibroblasts [170], as well as more shallow (approx. 30 million reads) singe-end sequencing from African elephant dermal fibroblasts [27] and placenta [170], and Asian elephant peripheral blood mononuclear cells (PBMCs) [149]. HISAT2 and StringTie were run on the Galaxy web-based platform (https://usegalaxy.org) [2] using default settings, and without a guide GTF/GFF file.

We determined if *LIF* transcription was induced by DNA damage and p53 activation in African elephant Primary fibroblasts (San Diego Frozen Zoo) using RT-PCR and primers designed to amplify elephant duplicate *LIF* genes, including LIF1-F: 5'-GCACAGAGAAGGACAAGCTG-3', LIF1-R: 5'-CACGTGGTACTTGTTGCACA-3', LIF6-F: 5'-CAGCTAGACTTCGTGGCAAC-3', LIF6-R: 5'-AGCTCAGTGATGACCTGCTT-3', LIF3-R: 5'-TCTTTGGCTGAGGTGTAGGG-3', LIF4-F: 5'-GGCACGGAAAAGGACAAGTT-3', LIF4-R: 5'-GCCGTGCGTACTTTATCAGG-3', LIF5-F: 5'-CTCCACAGCAAGCTCAAGTC-3', LIF5-R: 5'-GGGGATGAGCTGTGTGTACT-3'. We also used primers to elephant *BAX* to determine if it

was up-regulated by TP53: *BAX*-F: 5'-CATCCAGGATCGAGCAAAGC-3', *BAX*-R: 5'-CCACAGCTGCAATCATCCTC-3'. African elephant Primary fibroblasts were grown to 80% confluency in T-75 culture flasks at 37°C/5% CO2 in a culture medium consisting of FGM/EMEM (1:1) supplemented with insulin, FGF, 6% FBS and Gentamicin/Amphotericin B (FGM-2, singlequots, Clonetics/Lonza). At 80% confluency, cells were harvested and seeded into 6-well culture plates at 10,000 cells/well. Once cells recovered to 80% confluency they were treated with either vehicle control, 50μM Doxorubicin, or 50μM Nutlin-3a.

Total RNA was extracted using the RNAeasy Plus Mini kit (Qiagen), then DNase treated (Turbo DNA-free kit, Ambion) and reverse-transcribed using an olgio-dT primer for cDNA synthesis (Maxima H Minus First Strand cDNA Synthesis kit, Thermo Scientific). Control RT reactions were otherwise processed identically, except for the omission of reverse transcriptase from the reaction mixture. RT products were PCR-amplified for 45 cycles of 94°/20 seconds, 56°/30 seconds, 72°/30 seconds using a BioRad CFX96 Real Time qPCR detection system and SYBR Green master mix (QuantiTect, Qiagen). PCR products were electrophoresed on 3% agarose gels for 1 hour at 100 volts, stained with SYBR safe, and imaged in a digital gel box (ChemiDoc MP, BioRad) to visualize relative amplicon sizes.

3.2.4 Statistical methods

We used a Wilcox or T-test test implanted in R for all statistical comparisons, with at least four biological replicates. The specific statistical test used and number replicates for each experiment are indicated in figure legends.

3.2.5 Luciferase assay and cell culture

We used the JASPAR database of transcription factor binding site (TFBS) motifs [115] to computationally predict putative TFBSs within a 3kb window around Atlantogenatan *LIF* genes and identified matches for the *TP53* motif (MA0106.3), including a match

(sequence: CACATGTCCTGGCAACCT, score: 8.22, relative score: 0.82) 1kb upstream of the African elephant *LIF6* start codon. To test if the putative p53 binding site upstream of elephant *LIF6* was a functional p53 response element, we synthesized (GeneScript) and cloned the –1100bp to +30bp region of the African elephant *LIF6* gene (loxAfr3_dna range=scaffold_68:4294134-4295330 strand=+ repeatMasking=none) and a mutant lacking the CACATGTCCTGGCAACCT sequence into the pGL3-Basic[minP] luciferase reporter vector.

African elephant primary fibroblasts (San Diego Frozen Zoo) were grown to 80% confluency in T-75 culture flasks at 37°C/5% CO2 in a culture medium consisting of FGM/EMEM (1:1) supplemented with insulin, FGF, 6% FBS and Gentamicin/Amphotericin B (FGM-2, single-quots, Clonetics/Lonza). At 80% confluency, 104 cells were harvested and seeded into 96-well white culture plates. 24 hours later cells were transfected using Lipofectamine LTX and either 100g of the pGL3-Basic[minP], pGL3-Basic[minP] –1100bp to +30bp, pGL3-Basic[minP] -1100bp–+30bp Δp53TFBS luciferase reporter vectors and 1ng of the pGL4.74 [hRluc/TK] Renilla control reporter vector according the standard protocol with 0.5 ul/well of Lipofectamine LTX Reagent and 0.1ul/well of PLUS Reagent. 24 hours after transfection cells were treated with either vehicle control, 50μM Doxorubicin, or 50μM Nutlin-3a. Luciferase expression was assayed 48 hours after drug treatment, using the Dual-Luciferase Reporter Assay System (Promega) in a GloMax-Multi+ Reader (Promega). For all experiments luciferase expression was standardized to Renilla expression to control for differences transfection efficiency across samples; Luc./Renilla data is standardized to (Luc./Renilla) expression in untreated control cells. Each luciferase experiment was replicated three independent times, with 8-16 biological replicates per treatment and control group.

3.2.6 ChIP-qPCR and cell culture

African elephant primary fibroblasts were grown to 80% confluency in T-75 culture flasks at 37°C/5% CO2 in a culture medium consisting of FGM/EMEM (1:1) supplemented with insulin, FGF, 6% FBS and Gentamicin/Amphotericin B (FGM-2, singlequots, Clonetics/Lonza). 104 cells were seeded into each well of 6-well plate and grown to 80% confluency. Cells were then treated with either a negative control siRNA or equimolar amounts of a combination of three siRNAs that specifically target the canonical *TP53* transcript using Lipofectamine LTX according to the suggested standard protocol. The next day, cells were treated with either water, DMSO, 50μM Doxorubicin, or 50μM Nutlin-3a in three biological replicates for each condition. After 18 hrs of incubation with each drug, wells were washed three times with ice cold PBS and PBS replaced with fresh media, and chromatin cross linked with 1% fresh formaldehyde for 10 minutes. We used The MAGnify Chromatin Immunoprecipitation System (ThermoFischer #492024) to perform chromatin immunoprecipitation according to the suggested protocol. However rather than shearing chromatin by sonication, we used the ChIP-It Express Enzymatic Shearing Kit (Active Motif # 53009) according to the suggested protocol. Specific modifications to the MAGnify Chromatin Immunoprecipitation System included using 3ug of the polyclonal *TP53* antibody (FL-393, lot #DO215, Santa Cruz Biotechnology).

We used qPCR to assay for enrichment of *TP53* binding from the ChIP-Seq using the forward primer 5'-TGGTTTCCAGGAGTCTTGCT-3' and the reverse primer 5'-CATCCCCTC CTTCCTCTGTC-3'. 100ng of ChIP DNA was used per PCR reaction, which was amplified for 45 cycles of 94°/20 s, 56°/30 s, 72°/30 s using a BioRad CFX96 Real Time qPCR detection system and SYBR Green master mix (QuantiTect, Qiagen). Data are shown as fold increase in *TP53* ChIP signal relative to the background rabbit IgG ChIP signal and standardized to the control water for DOX or DMSO for nutlin-3a treatments.

3.2.7 ApoTox-Glo Viability/Cytotoxicity/Apoptosis experiments

T75 culture flasks were seeded with 200,000 African Elephant primary fibroblasts, and grown to 80% confluency at 37°C/5% CO2 in a culture medium consisting of FGM/EMEM (1:1) supplemented with insulin, FGF, 6% FBS and Gentamicin/Amphotericin B (FGM-2, singlequots, Clonetics/Lonza). 5000 cells were seeded into each well of two opaque-bottomed 96-well plates. In each plate, half of the columns in the plate were transfected with pcDNA3.1/LIF6/eGFP (GenScript) using Lipofectamine LTX (Thermo Scientific 15338100); the other half were mock transfected with the same protocol without any DNA. In the plate designated for the 18hr timepoint, each column was treated with either: 50μM (-)-Nutlin-3 (Cayman 18585); 20μM Z-VAD-FMK (Cayman 14463); 2uM Cyclosporin A (Cayman 12088); 50μM Doxorubicin (Fisher BP251610); DMSO (Fisher BP231100); or DPBS (Gibco 14190136). For the 24hr timepoint, the same schema for treatment was used, but with half-doses. Each treatment contained eight biological replicates for each condition. After 18 hrs of incubation with each drug, cell viability, cytotoxicity, and Caspase-3/7 activity were measured using the ApoTox-Glo Triplex Assay (Promega) in a GloMax-Multi+ Reader (Promega). Z-VAD-FMK readings were normalized to the PBS-treated, mock-transfected cells; all others were normalized to the DMSO-treated, mock-transfected cells.

T75 culture flasks were seeded with 250,000 wild-type (ATCC CRL-2907) and *Bak*/*Bax* double knockout (ATCC CRL-2913) mouse embryonic fibroblasts (MEFs), or Chinese hamster ovary cells (CHO-K1, Thermo R75807) and allowed to grow to 80% confluency at 37°C/5% CO2 in a culture medium consisting of high-glucose DMEM (Gibco) supplemented with GlutaMax (Gibco), Sodium pyruvate (Gibco), 10% FBS (Gibco), and Penicillin-Streptomycin (Gibco). 3000 cells were seeded into each well of an opaque, bottomed 96-well plate. Half of the columns in the plate were transfected with pcDNA3.1/LIF6/eGFP (GenScript) using Lipofectamine LTX (ThermoFisher Scientific 15338100); the other half were mock transfected with the same protocol without any DNA. 6 hours post-transfection, the transfection reagents

and media from each well was replaced: for the 24-hour timepoint, drug-supplemented media was placed within the wells; for the 48-hour timepoint, untreated media was placed in the wells, and then replaced with treatment media 24-hours later. Each column was treated with either: 50μM (-)-Nutlin-3 (Cayman 18585); 20μM Z-VAD-FMK (Cayman 14463); 2uM Cyclosporin A (Cayman 12088); 50μM Doxorubicin (Fisher BP251610); DMSO (Fisher BP231100); or DPBS (Gibco 14190136). Each treatment contained eight biological replicates for each condition. After 18 hrs of incubation with each drug, cell viability, cytotoxicity, and Caspase-3/7 activity were measured using the ApoTox-Glo Triplex Assay (Promega) in a GloMax-Multi+ Reader (Promega). Z-VAD-FMK readings were normalized to the PBS-treated, mock-transfected cells; all others were normalized to the DMSO-treated, mock-transfected cells.

For knockdown experiments T75 culture flasks were seeded with 200,000 African Elephant primary fibroblasts, and grown to 80% confluency at 37°C/5% CO2 in a culture medium consisting of FGM/EMEM (1:1) supplemented with insulin, FGF, 6% FBS and Gentamicin/Amphotericin B (FGM-2, singlequots, Clonetics/Lonza). 5000 cells were seeded into each well of two opaque-bottomed 96-well plates. In each plate, pairs of rows were transfected with either Silencer™ Select Negative Control No. 1 siRNA (Thermo 4390843), P53 siRNA (Dharmacon) [170], and either with or without pcDNA3.1/LIF6/eGFP (GenScript) using Lipofectamine LTX (Thermo Scientific 15338100). In the plate designated for the 18hr timepoint, each column was treated with either: 50μM Doxorubicin (Fisher BP251610); or an equivalent dilution of Ethanol (Fisher BP2818100). For the 24hr timepoint, the same schema for treatment was used, but with half-doses. Each treatment contained eight biological replicates for each condition. After 18 hrs of incubation with each drug, cell viability, cytotoxicity, and Caspase-3/7 activity were measured using the ApoTox-Glo Triplex Assay (Promega) in a GloMax-Multi+ Reader (Promega). All data were normalized to the ethanol-treated scrambled siRNA control samples. siRNAs were designed to specifically-target the elephant *LIF6* gene. Sequences of the three LIF6-specific siRNAs used are as follows: 1)

5'-GAAUAUACCUGGAGGAAUGUU-3', 2) 5'-GGAAGGAGGCCAUGAUGAAUU-3', 3) 5'-CACAAUAAGACUAGGAUAUUU-3' (Dharmacon). We also validated efficiency of the knockdown via qRT-PCR using the primer sets described earlier, which specifically the *LIF6* gene, and confirmed the combination of all three *LIF6* siRNAs was 88

To determine if *LIF6* was sufficient to induce apoptosis we synthesized and cloned (GeneScript) the African elephant *LIF6* gene into the pcDNA3.1+C-DYK expression vector, which adds at DYK epitope tag immediately C-terminal to the *LIF6* protein. We transiently transfected Chinese hamster ovary (CHO) cells or MEFs with LIF6_pcDNA3.1+C-DYK expression vector using Lipofectamine LTX according to manufacturer protocol and as described above, and assayed cell viability, cytotoxicity, and the induction of apoptosis using an ApoTox-Glo triplex assay. Mitochondrion membrane potential was assayed in CHO cells using the fluorometric Mitochondrion Membrane Potential Kit (Sigma MAK147) 48 hours after transfection.

3.2.8 Evolutionary analyses of LIF genes

We used a Bayesian approach to date *LIF* duplication events implemented in BEAST (v1.8.3) [150], including all identified African elephant, hyrax, and manatee *LIF* duplicates, as well as cannonical *LIF* genes from armadillo, sloth, aardvark, golden mole, and *LIF6* genes from Asian elephant, woolly and Columbian mammoth, straight-tusked elephant, and American Mastodon [134]. We used the GTR substitution, which was chosen as the best model using HyPhy, empirical nucleotide frequencies (+F), a proportion of invariable sites estimated from the data (+I), four gamma distributed rate categories (+G) with the shape parameter estimated from the data, an uncorrelated random local clock to model substitution rate variation across lineages, a Yule speciation tree prior, uniform priors for the GTR substitution parameters, gamma shape parameter, proportion of invariant sites parameter, and nucleotide frequency parameter. We used an Unweighted Pair Group Arithmetic Mean (UPGMA)

starting tree. The analysis was run for 10 million generations and sampled every 1000 generations with a burn-in of 1000 sampled trees; convergence was assessed using Tracer, which indicated convergence was reached rapidly (within 100,000 generations).

To constrain nodes we used normal priors with estimated confidence intervals, the root node was constrained to be 105 MYA, the root of *Xenarthra* was constrained to be 66 MYA, the root of *Afrosoricida* was constrained to be 70 MYA, the root of *Afrosoricida-Macroselidea* divergence constrained to be 75 MYA, the *Elephantidea* root was constrained to be 7.5 MYA, the *Afrotheria* root was constrained to be 83 MYA, the *Paeungulata* root was constrained to be 68 MYA, and the *Proboscidea* root was constrained to be 16 MYA. Divergence dates were obtained from www.timetree.org using the 'Expert Result' divergence dates.

We used the RELAX method to [184] test if duplicate *LIF* genes experienced a relaxation of the intensity of selection using the DataMonkey web server [33]. The alignment included all duplicate *LIF* genes identified in the African elephant, hyrax, and manatee genomes, as well as cannonical *LIF* genes from armadillo, sloth, aardvark, golden mole, and *LIF6* genes from Asian elephant, woolly and Columbian mammoth, straight-tusked elephant, and American Mastodon. Alignment confidence was assessed using GUIDANCE2 [159] with the MAFFT [87] algorithm and 100 bootstrap replicates.

3.3 Results

3.3.1 Repeated segmental duplications increased LIF copy number in Paenungulates

We characterized *LIF* copy number in 53 mammalian genomes, including large, long-lived mammals such as the African elephant (*Loxodonta africana*), Bowhead (*Balaena mysticetus*) and Minke (*Balaenoptera acutorostrata scammoni*) whales, as well as small, long-lived mammals such bats and the naked mole rat. We found that most Mammalian genomes

encoded a single *LIF* gene, however, the manatee (*Trichechus manatus*), rock hyrax (*Procavia capensis*), and African elephant genomes contained 7-11 additional copies of *LIF* (Figure 3.1). None of the duplicate *LIF* genes includes the 5'-UTR, coding exon 1, or a paired low complexity (CGAG)n/CT-rich repeat common to the canonical *LIF* genes in elephant, hyrax, manatee, tenrec, and armadillo (Figure 3.2A). Most of the duplicates include complex transposable element insertions composed of tandem tRNA-Asn-AAC/AFROSINE and AFROSINE3/tRNA-RTE/MIRc elements within introns one and two (Figure 3.2A). Fine mapping of the duplicate ends by reciprocal best BLAT indicates that there is no region of homology upstream of the tRNA-Asn-AAC/AFROSINE elements for duplicates that include exon 2, whereas duplicate *LIF* genes that lack exon 2 have 150-300bp regions of homology just upstream of the paired AFROSINE3/tRNA-RTE/MIRc elements in intron 2. The *LIF* encoding loci in the hyrax and manatee genomes have not been assembled into large-scale scaffolds, but the African elephant *LIF* loci are located within a 3.5Mb block of chromosome 25 (loxAfr4).

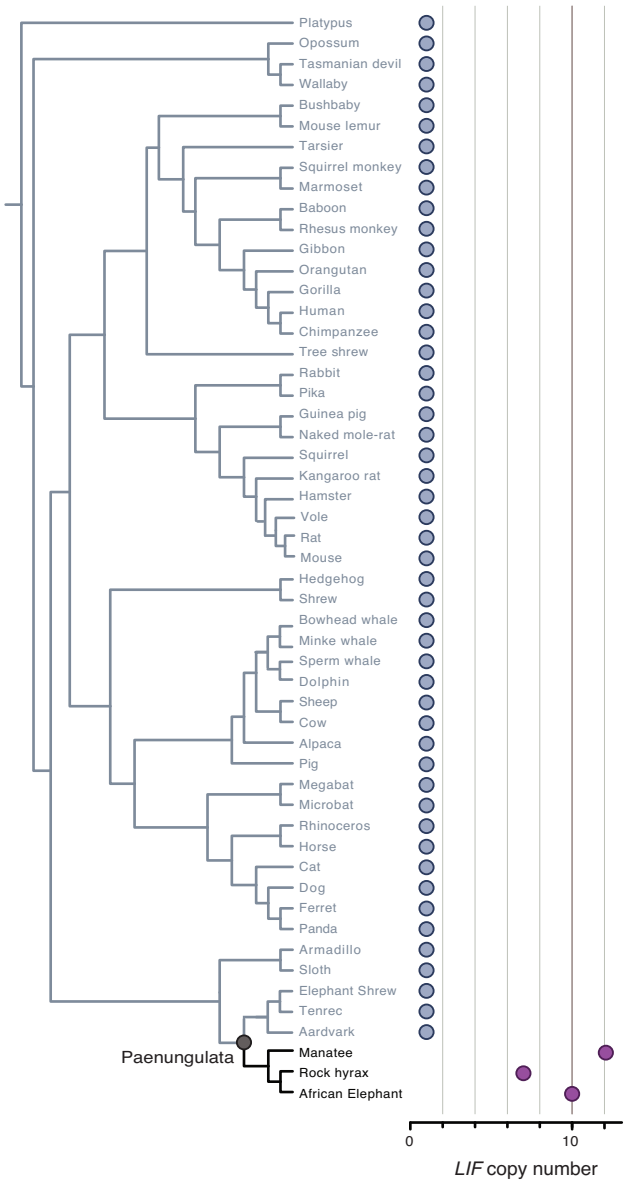

Figure 3.1: Expansion of *LIF* copy number in Paenungulata.*LIF* copy number in mammalian genomes. Clade names are shown for lineages in which the genome encodes more than one *LIF* gene or pseudogene.

LIF duplicates may result from independent duplication events in the elephant, hyrax, and manatee lineages, ancestral duplications that occurred in the Paenungulate stem-lineage followed by lineage-specific duplication and loss events, or some combination of these processes.

We used Bayesian phylogenetic methods to reconstruct the *LIF* gene tree and gene tree reconciliation to reconstruct the pattern of *LIF* duplication and loss events in Paenungulates. Consistent with a combination of ancestral and lineage-specific duplications, our phylogenetic analyses of Paenungulate *LIF* genes identified well-supported clades containing loci from multiple species as well as clades containing loci from only a single species (Figure 3.2B). The reconciled tree identified 17 duplication and 14 loss events (Figure 3.2C). These data indicate that the additional *LIF* genes result from repeated rounds of segmental duplication, perhaps mediated by recombination between repeat elements.

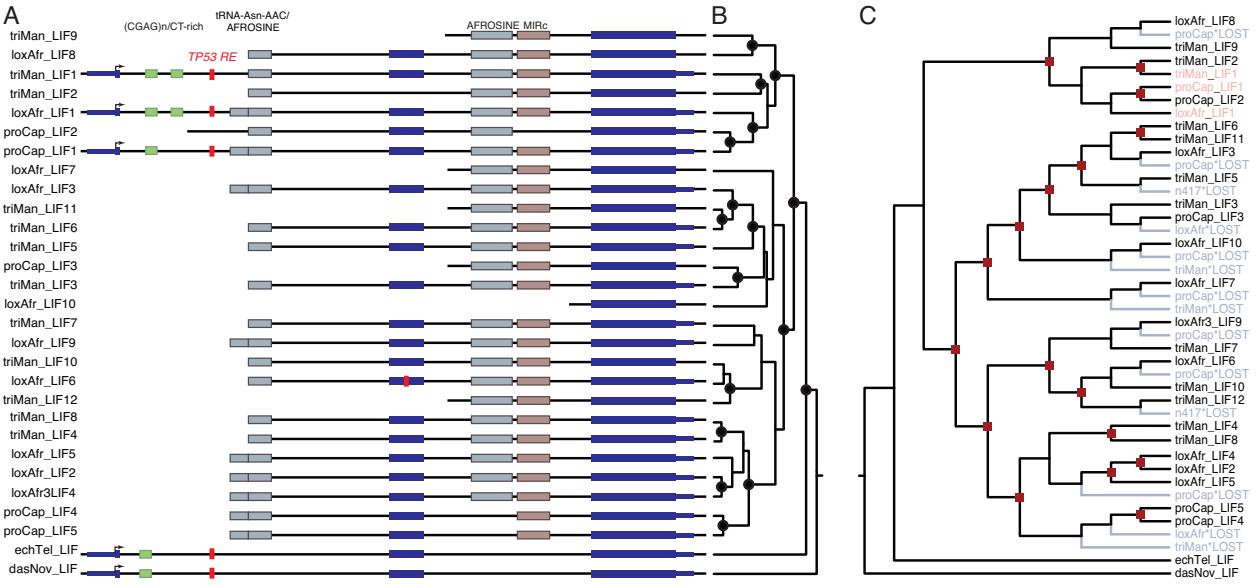

Figure 3.2: *LIF* copy number increased through segmental duplications. **A)** Organization of the *LIF* loci in African elephant (loxAfr), hyrax (ProCap), and manatee (triMan), tenrec (echTel), and armadillo (dasNov) genomes. The location of homologous transposable elements around *LIF* genes and *TP53* transcription factor binding sites are shown. **B)** *LIF* gene tree, nodes with Bayesian Posterior Probabilities (BPP) ¿ 0.9 are indicated with black circles. **C)** Reconciled *LIF* gene trees African elephant (loxAfr), hyrax (ProCap), and manatee (triMan). Duplication events are indicated with red squares, gene loss events are indicated with in blue and noted with '*LOST'. Cannonical *LIF* genes (*LIF1*) are shown in red.

3.3.2 Duplicate LIF genes are structurally similar to the LIF-T

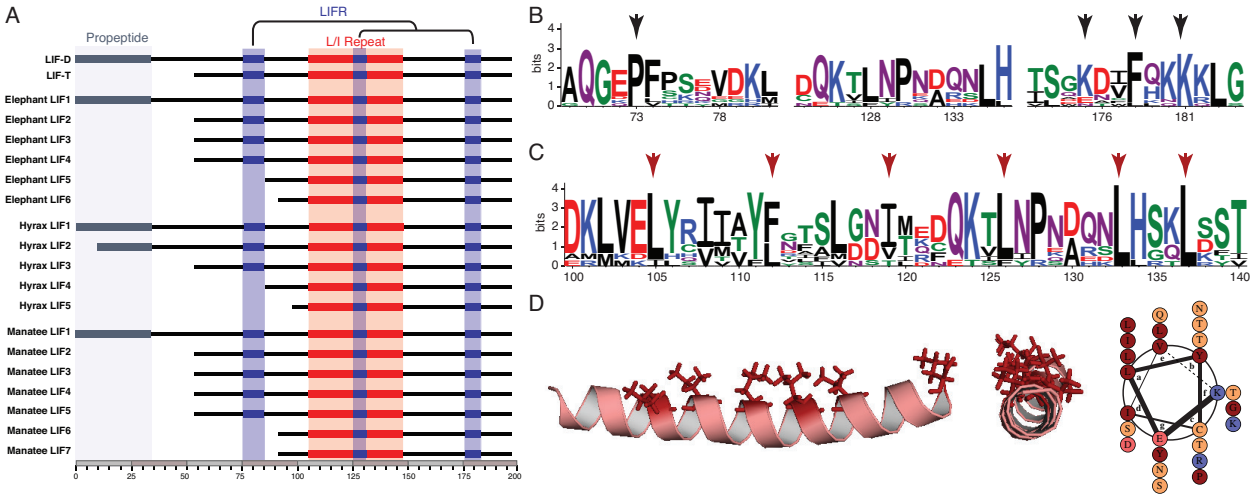

Figure 3.3: Structure of duplicate *LIF* genes with coding potential. **A)** Domain structure of the *LIF-D* and *LIF-T* isoforms and of duplicate elephant, hyrax, and manatee *LIF* duplicates with coding potential. Locations of the propeptide, interactions sites with the *LIF* receptor (*LIFR*), and L/I repeat are shown. **B)** Sequence logo showing conservation of *LIF* receptor (*LIFR*) interaction sites in duplicate *LIF* proteins. Residues in *LIF* that make physical contacts with *LIFR* are indicated with black arrows. Amino acids are colored according to physicochemical properties. Column height indicates overall conservation at that site (4, most conserved). **C)** Sequence logo showing conservation of the leucine/isoleucine repeat region in duplicate *LIF* proteins. Leucine/isoleucine residues required for pro-apoptotic functions of *LIF-T* are indicated with red arrows. Amino acids are colored according to physicochemical properties. Column height indicates overall conservation at that site (4, most conserved). **D)** Leucine/isoleucine residues in the African elephant *LIF6* form an amphipathic alpha helix. Structural model of the *LIF6* protein (left, center), and helical wheel representation of the *LIF6* amphipathic alpha helix.

Barring transcription initiation from cryptic upstream sites encoding in frame start codons, all duplicate *LIF* genes encode N-terminally truncated variants that are missing exon 1, lack

the propeptide sequence, and are similar in primary structures to *LIF-T* (Figure 3.3A). While some duplicates lack the N-terminal LIFR interaction site (Figure 3.3A), all include the leucine/isoleucine repeat required for inducing apoptosis (Figure 3.3A) [60]. Crucial residues that mediate the interaction between *LIF* and LIFR (Figure 3.3B) [75, 77] are relatively well conserved in duplicate *LIF* proteins, as are specific leucine/isoleucine residues that are required for the pro-apoptotic functions of *LIF-T* (Figure 3.3C) [60]. Haines et al. (2000) [60] suggested that the leucine/isoleucine residues of *LIF-T* are located on a single face of helix B, and may form an amphipathic α-helix. Similar to LIF-T, leucine/isoleucine residues of duplicate *LIF* proteins are located on a single face of helix B (Figure 3.3D). These data suggest that at least some of the structural features that mediate *LIF* functions, in particular the pro-apoptotic function(s) of LIF-T, are conserved in duplicate LIFs.

3.3.3 Elephant LIF6 is up-regulated by TP53 in response to DNA damage

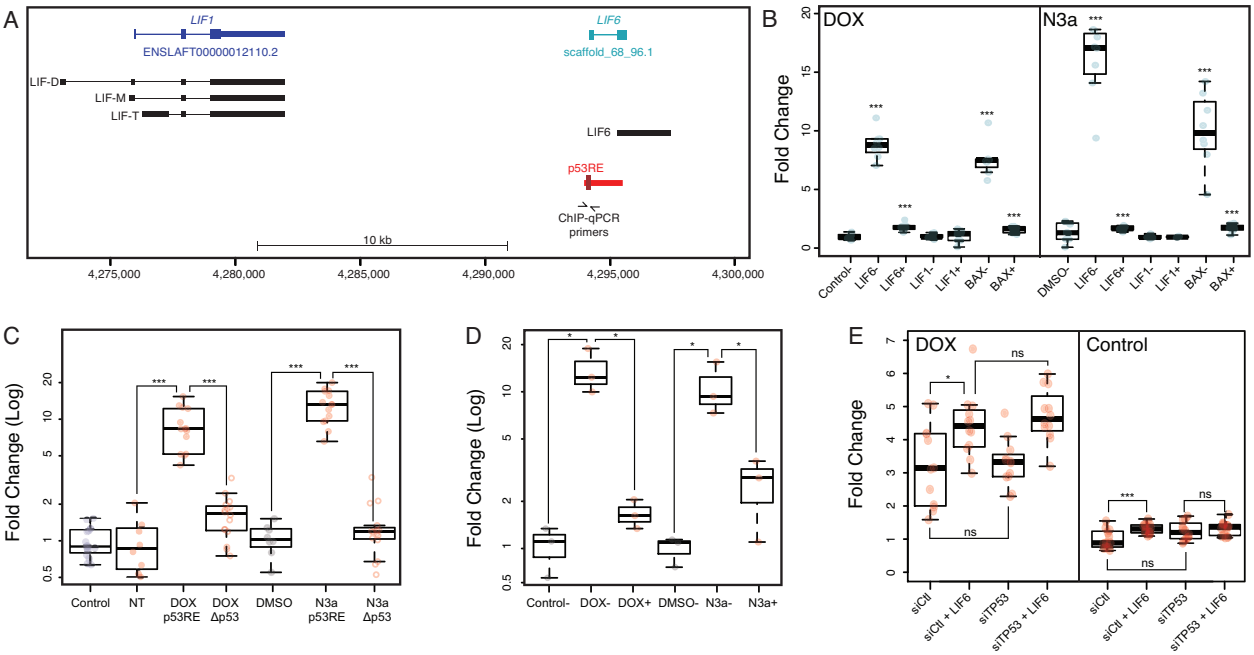

Figure 3.4: African elephant *LIF6* is transcriptionally up-regualted by *TP53* in response to DNA damage. (Continued on next page)

Figure 3.4: (Continued from previous page) **A)** Structure of the African elephant *LIF/LIF6* locus (loxAfr3). The ENSEMBL *LIF* and geneID gene models are shown in blue and cyan. Transcripts assembled by StringTie (option 'do not use GFF/GTF') are shown in black. The region upstream of *LIF6* used in transcription factor binding site prediction and luciferase assays is shown in red; the location of the putative p53 binding-site is shown in dark red. **(B)** Quantitative real-time PCR (qPCR) showing that *LIF6* is up-regulated in African elephant fibroblasts treated with doxorubicin (DOX) or nutlin-3a (N3a) and either a negative control siRNA (-) or an siRNA to knockdown *TP53* expression (+); *TP53* knockdown prevents *LIF6* up-regulation in response to DOX or N3a. Data shown as fold-change relative to control (water) or DMSO (a carrier for nutlin-3a). N=8, *** Wilcox test P<0.001. **C)** Dual luciferase reporter assay indicates that the *LIF6* upstream region (p53RE) activates luciferase expression in African elephant fibroblasts treated in response to doxorubicin (DOX) or nutlin-3a treatment (N3a), and is significantly attenuated by deletion of the putative *TP53* binding site (Δp53). Data shown as fold-change relative to controls (water for DOX, DMSO for N3a). NT, no DOX or nutlin-3a treatment. N=8, *** Wilcox test P<0.001. **D)** ChIP-qPCR indicates that the putative *TP53* binding site is bound by *TP53* in response to in response to doxorubicin (DOX-) or nutlin-3a treatment (N3a-), and is significantly attenuated by siRNA mediated *TP53* knockdown (DOX+ or N3a-). Data shown as fold-change relative to carrier controls (water or DMSO) and standardized to IgG control. N=3, * unequal variance T-test P$\leq$ 0.06. **E)** Knockdown of *TP53* inhibits DOX induced apoptosis in elephant African elephant fibroblasts. Fibroblasts were transiently transfected with either an negative control siRNA (siCtl) or three siRNAs targeting TP53, and either a empty vector control or a *LIF6* expression vector. Apoptosis was assayed using an ApoTox-Glo 18 hours after treatment with DOX or control media. N=8, **** Wilcox test P<0.05, *** Wilcox test P<0.001.

If expansion of the *LIF* gene repertoire plays a role in the evolution of enhanced cancer resistance, then one or more of the *LIF* genes should be transcribed. To determine if duplicate

LIF genes were transcribed, we assembled and quantified elephant *LIF* transcripts with HISAT2 and StringTie [92, 139, 138] using deep 100bp paired-end RNA-Seq data (¿138 million reads) we previously generated from Asian elephant dermal fibroblasts [170], as well as more shallow (30 million reads) singe-end sequencing from Asian elephant peripheral blood mononuclear cells (PBMCs) [149], African elephant dermal fibroblasts [27] and placenta [170]. We identified transcripts corresponding to the LIF-D, LIF-M, and *LIF-T* isoforms of the canonical *LIF1* gene, and one transcript of a duplicate *LIF* gene (*LIF6*) in Asian elephant dermal fibroblasts (Figure 3.4A). The *LIF6* transcript initiates just downstream of canonical exon 2 and expression was extremely low (0.33 transcripts per million), as might be expected for a pro-apoptotic gene. No other RNA-Seq dataset identified duplicate *LIF* transcripts.

Previous studies have shown that *TP53* regulates basal and inducible transcription of *LIF* in response to DNA damage through a binding site located in *LIF* intron 1 [12, 73], suggesting that duplicate *LIF* genes may be regulated by TP53. Therefore we computationally predicted *TP53* binding sites within a 3kb window around Atlantogenatan *LIF* genes and identified binding site motifs in the first intron of African elephant, hyrax, manatee, tenrec, and armadillo *LIF1* genes whereas the only duplicate *LIF* gene with a putative *TP53* binding site was elephant *LIF6*; note that the putative *TP53* binding sites around *LIF1* and *LIF6* are not homologous (Figure S.3.1). Next we treated African elephant primary dermal fibroblasts with the DNA damaging agent doxorubicin (DOX) or the MDM2 antagonist nutlin-3a and quantified the transcription of canonical *LIF1*, duplicate *LIF* genes, and the *TP53* target gene *Bax* by qRT-PCR. DOX treatment induced *LIF6* expression 8.18-fold (Wilcox test, P=1.54×10-6) and nutlin-3a induced *LIF6* expression 16.06-fold (Wilcox test, P=1.00×10-4), which was almost completely attenuated by siRNA mediated *TP53* knockdown (Figure S.3.2 and Figure 3.4B). Treatment with DOX (Wilcox test, P=1.55×10-4) or nutlin-3a (Wilcox test, P=1.55×10-4) also up-regulated the *TP53* target gene *BAX* (Figure 3.4B), which again was almost blocked by knockdown of *TP53* (Figure 3.4B). In contrast neither treatment

up-regulated *LIF1* (Figure 3.4B) and we observed no expression of the other duplicate *LIF* genes in African elephant fibroblasts or any *LIF* duplicate in hyrax fibroblasts treated with DOX or nutlin-3a. These data suggest that while *LIF6* encodes a transcribed gene in elephants, transcription of the other *LIF* duplicates is either induced by different signals or they are pseudogenes.

To test if the putative *TP53* binding site upstream of elephant *LIF6* was a functional *TP53* response element, we cloned the –1100bp to +30bp region of the African elephant *LIF6* gene into the pGL3-Basic[minP] luciferase reporter vector and tested its regulatory ability in dual luciferase reporter assays. We found that the African elephant *LIF6* upstream region had no effect on basal luciferase expression in transiently transfected African elephant fibroblasts (Wilcox test, P=0.53). In contrast, both DOX (Wilcox test, P=1.37×10-8) and nutlin-3a (Wilcox test, P=1.37×10-8) strongly increased luciferase expression (Figure 3.4C), which was almost completely abrogated by deletion of the putative *TP53* binding-site in DOX (Wilcox test, P=1.37×10-8) and N3a (Wilcox test, P=1.37×10-9) treated cells (Figure 3.4C). Next we performed ChIP-qPCR to determine if the *TP53* binding-site upstream of *LIF6* is bound by *TP53* in African elephant fibroblasts treated with DOX or nutlin-3a using a rabbit polyclonal *TP53* antibody (FL-393) that we previously demonstrated recognizes elephant *TP53* [170]. DOX treatment increased *TP53* binding 14.26-fold (unequal variance t-test, P=0.039) and nutlin-3a increased *TP53* binding 10.75-fold (unequal variance t-test, P=0.058) relative to ChIP-qPCR with normal mouse IgG control antibody. This increased binding was almost completely attenuated by siRNA mediated *TP53* knockdown (Figure 3.4D).

Finally, we transiently transfected elephant fibroblasts with either a negative control siRNA or siRNAs targeting *TP53* and a *LIF6* expression vector and assayed cell viability, cytotoxicity, and apoptosis using an ApoTox-Glo assay 18 hours after treatment with DOX or control media. We found that *LIF6* expression with negative control siRNAs augmented

the induction of apoptosis by DOX (Wilcox test, P=0.033; Figure 3.4E and Figure S.3.3). Knockdown of *TP53* did not inhibit the induction of apoptosis (Wilcox test, P=0.033; Figure 3.4E and Figure S.3.3), suggesting *TP53* knockdown was insufficient to alter the induction of apoptosis; note that while siRNA mediated knockdown significantly reduced *TP53* transcript levels (Figure S.3.2), we were unable to validate knockdown of the *TP53* protein because the FL-393 antibody that recognizes elephant *TP53* is no longer available. Interestingly, however, *LIF6* transfection induced apoptosis in elephant fibroblasts treated with control media and negative control siRNAs (Wilcox test, P=0.008), suggesting that *LIF6* can induce apoptosis in the absence of DNA damage similar to *LIF-T* (Figure 3.4E and Figure S.3.3). Thus, we conclude that elephant *LIF6* is transcriptionally up-regulated by *TP53* in response to DNA damage and may have pro-apoptotic functions.

3.3.4 Elephant LIF6 contributes to the augmented DNA-damage response in elephants

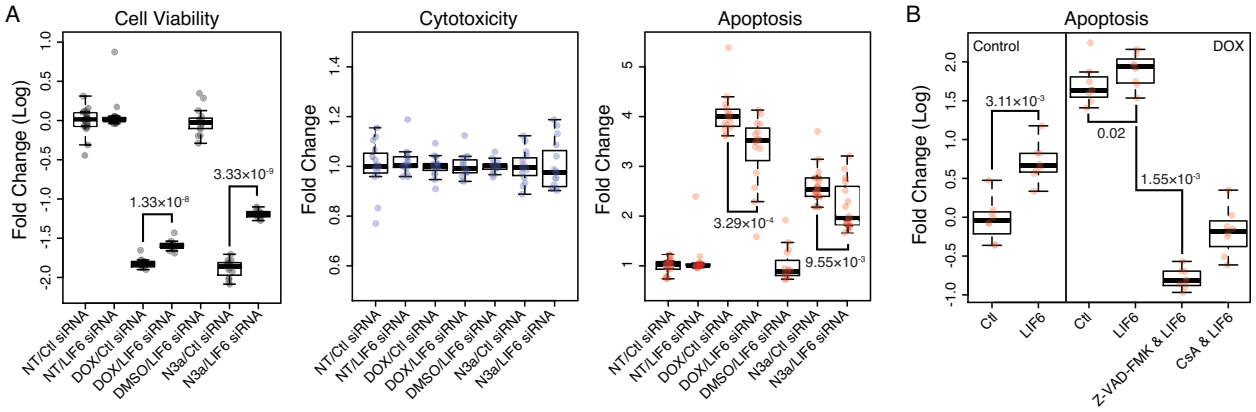

Figure 3.5: African elephant *LIF6* contributes to the augmented DNA damage response in elephants. **A)** African elephant fibroblasts were treated with either doxorubicin (DOX) or nutlin-3a (N3a), or an equimolar mixture of 3 siRNAs targeting *LIF6* and doxorubicin (DOX/*LIF6* siRNA) or nutlin-3a treatment (N3a/*LIF6* siRNA). Cell viability, cytoxicity, and the induction of apoptosis was assayed using an ApoTox-Glo assay 24 hours after treatment. NT, no treatment. Ctl siRNA, negative control siRNA. DMSO, carrier for nutlin-3a. N=16, Wilcox test. **B)** African elephant fibroblasts were transiently transfected with either an empty expression vector (Ctl) or a *LIF6* encoding expression vector (*LIF6*), and treated with either DOX, the caspase inhibitor Z-VAD-FMK, or the cyclosporine A (CsA) which inhibits opening of the opening of the mitochondrial permeability transition pore. N=8, Wilcox test

We have previously shown that elephant cells evolved to be extremely sensitive to genotoxic stress and induce apoptosis at lower levels of DNA damage than their closest living relatives, including the African Rock hyrax (*Procavia capensis capensis*), East African aardvark (*Orycteropus afer lademanni*), and Southern Three-banded armadillo (*Tolypeutes matacus*) [170]. To test the contribution of *LIF6* to this derived sensitivity, we designed a set of three siRNAs that specifically target *LIF6* and reduce *LIF6* transcript abundance 88% (Figure

S.3.2). Next, we treated African elephant dermal fibroblasts with DOX or nutlin-3a and either *LIF6* targeting siRNAs or a control siRNA and assayed cell viability, cytotoxicity, and apoptosis using an ApoTox-Glo assay 24 hours after treatment. Both DOX (Wilcox test, P=3.33×10-9) and nutlin-3a (Wilcox test, P=3.33×10-9) reduced cell viability 85%, which was attenuated 5-15% by *LIF6* knockdown in DOX (Wilcox test, P=1.33×10-8) or nutlin-3a (Wilcox test, P=3.33×10-9) treated cells (Figure 3.5A). While neither DOX nor nutlin-3a induced cytotoxicity (Figure 3.5A), both DOX (4.05-fold, Wilcox test, P=3.33×10-9) and nutlin-3a (2.64-fold, Wilcox test, P=3.33×10-9) induced apoptosis (Figure 3.5A).

To determine if *LIF6* expression was sufficient to induce apoptosis, we transiently transfected a *LIF6* expression vector in to African elephant dermal fibroblasts and assayed cell viability, cytotoxicity, and apoptosis using the ApoTox-Glo assay 24 hours after transfection. We again found that *LIF6* overexpression induced apoptosis in the absence of either DNA damage by DOX or *TP53* activation by nutlin-3a treatment (Wilcox test, P=3.11×10-4), and augmented apoptosis induced with DOX (Wilcox test, P=0.02). Induction of apoptosis by *LIF6* was almost completely blocked by co-treatment with the irreversible broad-spectrum caspase inhibitor Z-VAD-FMK (Wilcox test, P=1.55×10-4) but not cyclosporine A (Wilcox test, P=0.23), which inhibits opening of the opening of the mitochondrial permeability transition pore (Figure 3.5B and Figure S.3.4). These data suggest that *LIF6* contributes to the enhanced apoptotic response that evolved in the elephant lineage, likely through a mechanism that induces caspase-dependent apoptosis.

3.3.5 Elephant LIF6 induces mitochondrial dysfunction and caspase-dependent apoptosis

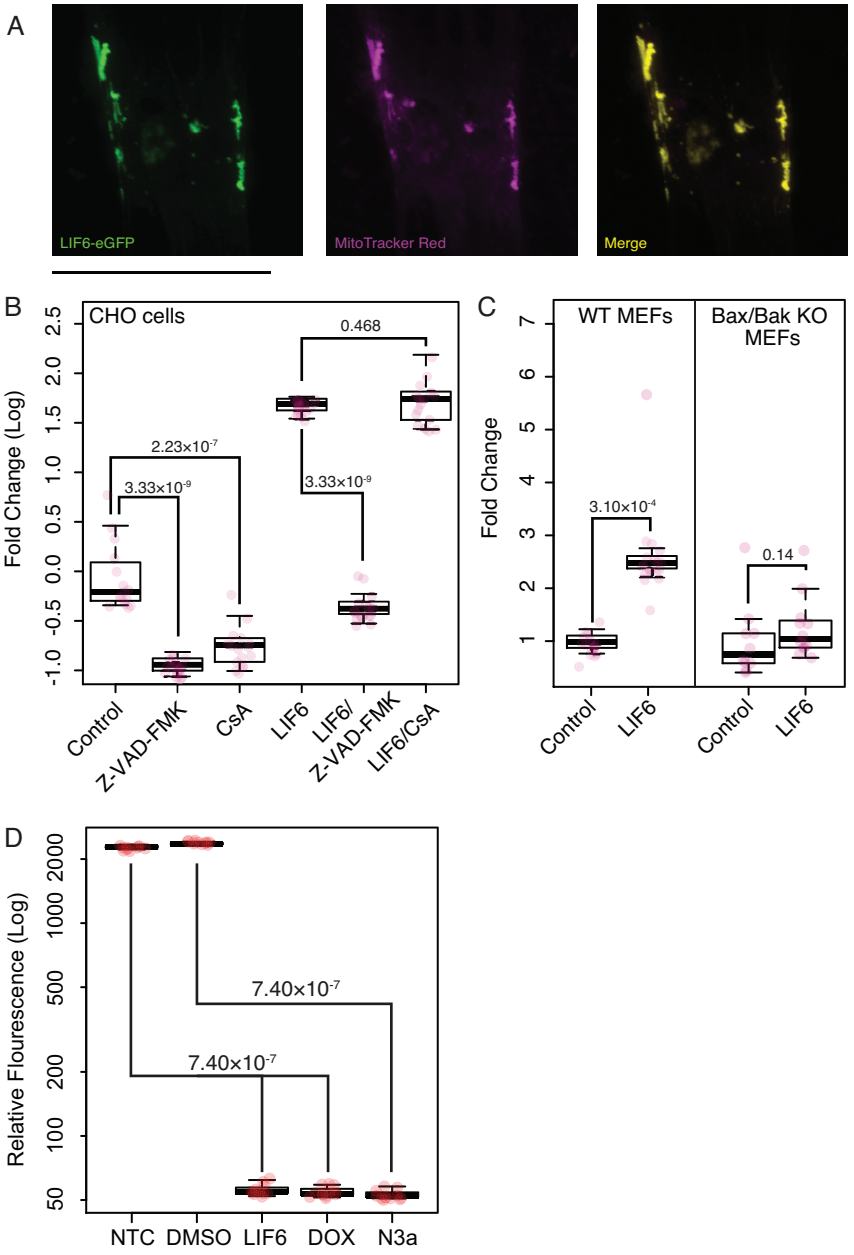

Figure 3.6: African elephant *LIF6* is mitochondrial localized and induces caspase dependent apoptosis. (Continued on next page)

Figure 3.6: (Continued from previous page) **A)** African elephant fibroblasts were transiently transfected with an expression vector encoding a eGFP tagged *LIF6* gene and mitochondria stained with MitoTracker Red CM-H2XRos. A single representative cell is shown. **B)** Chinese hamster ovary (CHO) cells (which do not express *LIFR*) were transiently transfected with an expression vector encoding the African elephant *LIF6* gene and assayed for the induction of apoptosis with an ApoTox-Glo assay 24 hours after transfection. Induction of apoptosis by *LIF6* was inhibited by co-treatment with the irreversible broad-spectrum caspase inhibitor Z-VAD-FMK but not cyclosporine-A (CsA). Treatment of CHO cells with Z-VAD-FMK or CsA alone reduced apoptosis. N=16, Wilcox test. **C)** Overexpression of *LIF6* in *Bax/Bak* double knockout mouse embryonic fibroblasts does not induce apoptosis, not augmented nutlin-3a induced apoptosis. N=8, Wilcox test. **D)** Overexpression of *LIF6* in CHO cells induces loss of mitochondrial membrane potential 48 hours after transfection. N=8, Wilcox test.

To infer the mechanism(s) by which *LIF6* contributes to the induction of apoptosis, we first determined the sub-cellular localization of a LIF6¬¬–eGFP fusion protein in African elephant dermal fibroblasts. Unlike LIF-T, which has diffuse cytoplasmic and nuclear localization [60], LIF6¬¬–eGFP was located in discrete foci that co-localized with MitoTracker Red CM-H2XRos stained mitochondria (Figure 3.6A). Mitochondria are critical mediators of cell death, with distinct pathways and molecular effectors underlying death through either apoptosis [86, 174] or necrosis [174, 179]. During apoptosis, for example, the *Bcl-2* family members *Bax/Bak* form large pores in the outer mitochondrial membrane that allow cytochrome c to be released into the cytosol thereby activating the caspase cascade [86, 174]. In contrast, during necrosis, *Bax/Bak* in the outer membrane interact with cyclophilin D (*CypD*) and the inner membrane complex leading to the opening of the mitochondrial permeability transition pore (MPTP), swelling, and eventual rupture [174, 179].

To test if *LIF6* induced apoptosis was specific to elephant cells and independent of *LIF* receptor (LIFR) mediated signaling, we transiently transfected Chinese hamster (*Cricetulus griseus*) ovary (CHO) cells, which do not express *LIFR* [131], with the *LIF6* expression vector and assayed the induction of apoptosis with the ApoTox-Glo assay. Overexpression of *LIF6* induced apoptosis 5.38-fold (Wilcox test, P=3.33×10-9) 24 hours after transfection, consistent with a pro-apoptotic function independent of *LIFR* (Figure 3.6B). Induction of apoptosis by

LIF6, however, was almost completely blocked by co-treatment with Z-VAD-FMK (Figure 3.6B) but not cyclosporine A (CsA) (Figure 3.6B). To test if *LIF6* induced apoptosis is dependent upon *Bax* and *Bak*, we overexpressed *LIF6* in *Bax*/*Bak* knockout mouse embryonic fibroblasts (MEFs) but did not observe an induction of apoptosis (Wilcox test, P=0.14; Figure 3.6C and Figure S.3.5). In contrast *LIF6* overexpression induced apoptosis in wild-type MEFs (Wilcox test, P=0. 3.10×10-4; Figure 3.6C and Figure S.3.5). During apoptosis, collapse of the mitochondrial membrane potential (MMP) coincides with the opening of the mitochondrial transition pores, leading to the release of proapoptotic factors into the cytosol. Consistent with this mechanism, we found that *LIF6* overexpression, treatment with DOX, or with nutlin-3a induced loss of MMP in CHO cells 48 hours after transfection (Wilcox test, P=7.40×10-7; Figure 3.6D). Thus *LIF6* is sufficient to induce mitochondrial dysfunction and apoptosis mediated through *Bax*/*Bak* and independent of MPTP opening.

3.3.6 Elephant LIF6 is a refunctionalized pseudogene

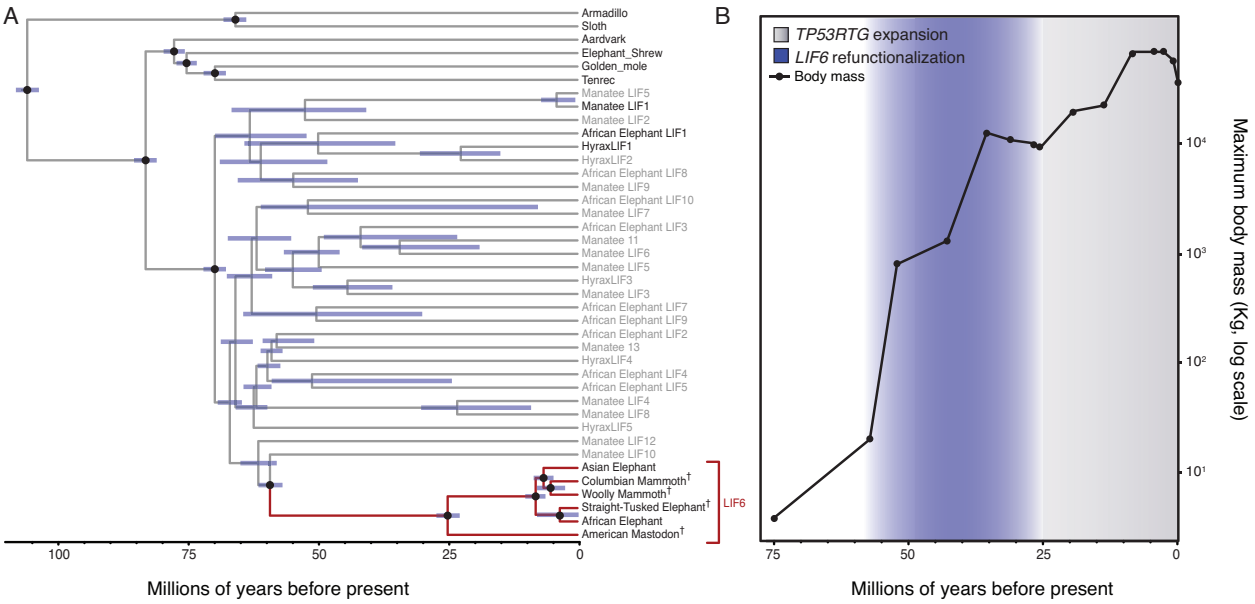

Figure 3.7: *LIF6* is a re-functionalized pseudogene. **A)** Time calibrated Bayesian phylogeny of Atlantogenatan *LIF* genes. The *Proboscidean LIF6* clade is highlighted in red, canonical *LIF* genes in black, *LIF* duplicates in grey. The 95% highest posterior density (HPD) of estimated divergence dates are shown as blue bars. Nodes used to calibrate divergence dates are shown with black circles. **B)** Proboscidean *LIF6* re-functionalized during the evolution of large body sizes in the *Proboscidean* lineage.

We reasoned that most duplicate *LIF* genes are (likely) pseudogenes because elephant *LIF6* is deeply nested within the duplicate *LIF* clade, is the only expressed duplicate, and is the only duplicate with a *TP53* response element, suggesting elephant *LIF6* re-evolved into a functional gene from a pseudogene ancestor. To test this hypothesis and reconstruct the evolutionary history of the *LIF6* gene in the *Proboscideans* with greater phylogenetic resolution, we annotated the *LIF6* locus in the genomes of Elephantids including the African Savannah elephant (*Loxodonta africana*), African Forest elephant (*Loxodonta cyclotis*), Asian elephant (*Elephas maximus*), woolly mammoth (*Mammuthus primigenius*), Columbian mammoth

(*Mammuthus columbi*), and straight-tusked elephant (*Palaeoloxodon antiquus*), as well as the American Mastodon (*Mammut americanum*), an extinct Mammutid. We found that the genomes of each extinct *Proboscidean* contained a *LIF6* gene with coding potential similar to the African and Asian elephant *LIF6* genes as well as the *TP53* binding-site, indicating that *LIF6* evolved to be a *TP53* target gene in the stem-lineage of *Proboscideans*.

While functional genes evolve under selective constraints that reduce their d_n/d_s (ω) ratio to below one, pseudogenes are generally free of such constraints and experience a relaxation in the intensity of purifying selection and an elevation in their d_n/d_s ratio. Therefore, we used a random effects branch-site model (RELAX) to test for relaxed selection on duplicate *LIF* genes compared to canonical *LIF* genes. The RELAX method fits a codon model with three ω rate classes to the phylogeny (null model), then tests for relaxed/intensified selection along lineages by incorporating a selection intensity parameter (K) to the inferred ω values; relaxed selection (both positive and negative) intensity is inferred when K$\leq$ 1 and increased selection intensity is inferred when K¿1. As expected for pseudogenes, *LIF* duplicates (other than *Proboscidean LIF6* genes) had significant evidence for a relaxation in the intensity of selection (K=0.36, LRT=42.19, P=8.26×10-11) as did the *Proboscidean LIF6* stem-lineage (K=0.00, LRT=3.84, P=0.05). In contrast, *Proboscidean LIF6* genes had significant evidence for selection intensification (K=50, LRT=4.46, P=0.03). We also found that the branch-site unrestricted statistical test for episodic diversification (BUSTED), which can detect gene-wide (not site-specific) positive selection on at least one site and on at least one branch, inferred a class of strongly constrained sites in (ω=0.00, 23.7%), a class of moderately constrained sites (ω=0.64, 75.85%), and a few sites that may have experienced positive selection in *Proboscidean LIF6* genes (ω=10000.00, 0.41%; LRT=48.81, P$\leq$ 0.001). These data are consistent with the reacquisition of constraints after refunctionalization.

Finally we inferred a Bayesian time-calibrated phylogeny of *Atlantogenatan LIF* genes, including *LIF6* from African and Asian elephant, woolly and Columbian mammoth, straight-

tusked elephant, and American Mastodon, to place upper and lower bounds on when the *Proboscidean LIF6* gene may have refunctionalized (Figure 3.7A). We found that estimated divergence date of the *Proboscideans LIF6* lineage was 59 MYA (95% HPD: 61-57 MYA) whereas the divergence of *Proboscideans* was 26 MYA (95% HPD: 23.28 MYA). These data indicate that the *Proboscidean LIF6* gene refunctionalized during the evolutionary origin of large body sizes in this lineage, although precisely when within this time interval is unclear (Figure 3.7B). Thus *LIF6* was reanimated sometime before the demands of maintaining a larger body existed in the *Proboscidean* lineage, suggesting *LIF6* is permissive for the origin of large bodies but is not sufficient.

3.4 Discussion

A comprehensive analyses of genetic changes associated with the resolution of Peto's paradox in the elephant lineage has yet to be performed, but candidate gene studies have identified functional duplicates of the master tumor suppressor *TP53* as well as putative duplicates of other tumor suppressor genes [1, 20, 170]. Caulin et al, for example, characterized the copy number of 830 tumor-suppressor genes [65] across 36 mammals and identified 382 putative duplicates, including five copies of *LIF* in African elephants, seven in hyrax, and three in tenrec. Here we show that an incomplete duplication of the *LIF* gene in the *Paenungulate* stem-lineage generated a duplicate missing the proximal promoter and exon 1, generating a gene with similar structure to the *LIF-T* isoform [59], which functions as an intra-cellular pro-apoptotic protein independently from the LIFR-mediated signaling. Additional duplications of this original duplicate increased *LIF* copy number in *Paenungulates*, however, most *LIF* duplicates lack regulatory elements, are not expressed in elephant or hyrax fibroblasts (manatee cells or tissues are unavailable), and, with the exception of elephant LIF6, are likely pseudogenes.

While we are unable to do the kinds of reverse and forward genetic experiments that

traditionally establish causal associations between genotypes and phenotypes, we were able to use primary African elephant and hyrax dermal fibroblasts to functionally characterize *LIF* duplicates. We found, for example, that the elephant *LIF6* gene is transcribed at very low levels under basal conditions, but is up-regulated by *TP53* in response to DNA damage. One of the constraints on the refunctionalization of pseudogenes is that they must evolve new cis-regulatory elements to direct their expression, but random DNA sequences can evolve into promoters with only a few substitutions suggesting de novo origination of regulatory elements may be common [189]. There should be strong selection against the origin of constitutively active enhancers/promoters for pro-apoptotic pseudogenes, however, because their expression will be toxic. These results imply refunctionalizing *LIF* pseudogenes may impose a potential evolutionary cost. One of the ways to avoid that cost is through the gain of inducible regulatory elements that appropriately respond to specific stimuli, such as a *TP53* signaling. Indeed our phylogenetic analysis indicates that a *TP53* response element up-stream of *LIF6* evolved before the divergence of mastodons and the modern elephant lineage, suggesting that *LIF6* refunctionalized in the stem-lineage of *Proboscideans* coincident with the origin of large body sizes and thus may have been permissive for the large bodies.

The precise mechanisms by which mitochondrial dysfunction leads to apoptosis are uncertain, however, during early stages of apoptosis the pro-death *Bcl-2* family members *Bax* and *Bak* hetero- and homo-oligomerize within the mitochondrial outer membrane leading to permeabilization (MOMP) and the release of pro-apoptotic protein such as cytochrome c [86, 85]. In contrast, during necrosis the collapse of the MMP and the opening of the mitochondrial permeability transition pore (MPTP) leads to mitochondrial swelling, rupture, and cell death [105]. Our observations that cyclosporine A (CsA) did not inhibit *LIF6* induced apoptosis, and that *LIF6* overexpression did not induce apoptosis in *Bax*/*Bak* null MEFs suggests that *LIF6* functions in a manner analogous to the pro-apoptotic *Bcl-2* family members by inducing the opening of the outer mitochondrial membrane pore. Furthermore

our observation that *LIF6* overexpression incudes apoptosis in elephant dermal fibroblasts, Chinese hamster ovary cells, and mouse embryonic fibroblasts indicates the *LIF6* mechanism of action is neither of cell-type nor species specific. The molecular mechanisms by which *LIF6* induces apoptosis, however, are unclear and the focus of continued studies.

3.5 Supplemental Figures

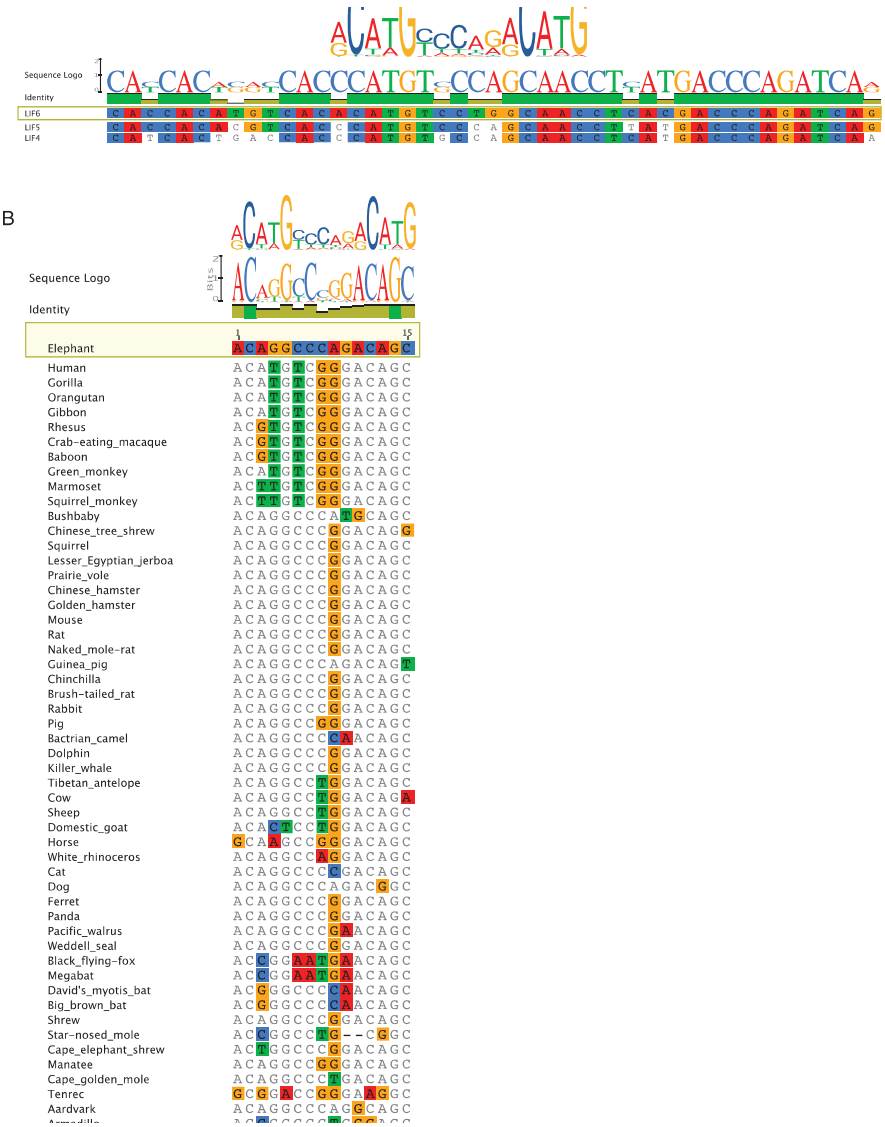

Figure S.3.1: Similarity of the *LIF6* and *LIF1 TP53* binding sites. (Continued on next page)

Figure S.3.1: (Continued from previous page) to the *TP53* binding motif (JASPAR MA0106.2), related to Figure 3.2. **A)** Alignment of the *LIF4, LIF5*, and *LIF6 TP53* binding sites. Bases are colored according to identity to LIF6, identical nucleotides are indicated with green columns above the alignment. A sequence logo is displayed on top. The experimentally validated *TP53* binding motif is aligned on top of the putative *LIF4, LIF5*, and *LIF6 TP53* binding sites. Note 3-4 nucleotide differences between *LIF6* and *LIF4* and *LIF5*. **B)** Sequence logo of the *LIF1* intron 1 *TP53* binding site from 53 Eutherian mammals. The JASPAR TP53 motif (MA0106.2) is shown aligned and above a sequence logo of the TP53 motif from 53 mammals. Sequences from each of the 53 mammals is show below, with differences from the elephant *LIF1* intron 1 TP53 binding site shown in color.

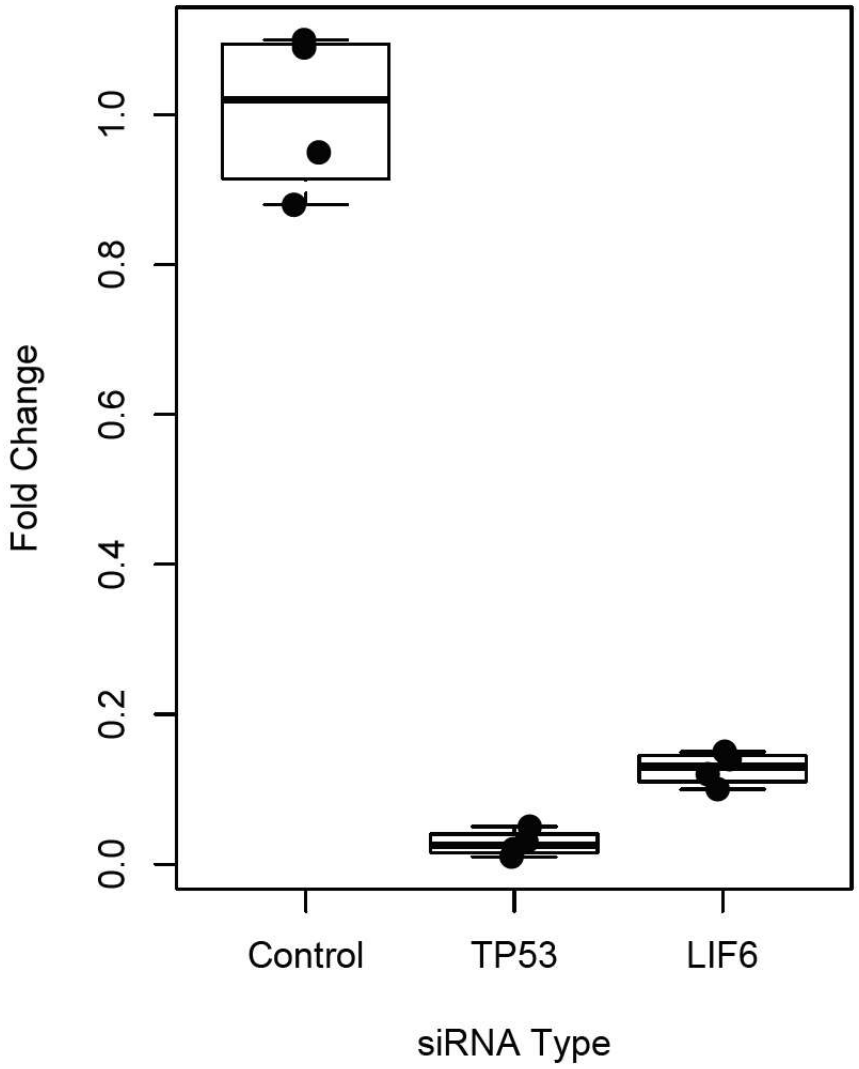

Figure S.3.2: Efficacy of siRNAs targeting *TP53* and *LIF6* transcripts, related to Figure 3.4. Fold change in *TP53* and *LIF6* transcript abundance upon siRNA mediated knockdown compared to negative control siRNAs. N=4, Wilcox test P=0.028 for TP53 knockdown and P=0.029 for *LIF6* knockdown.

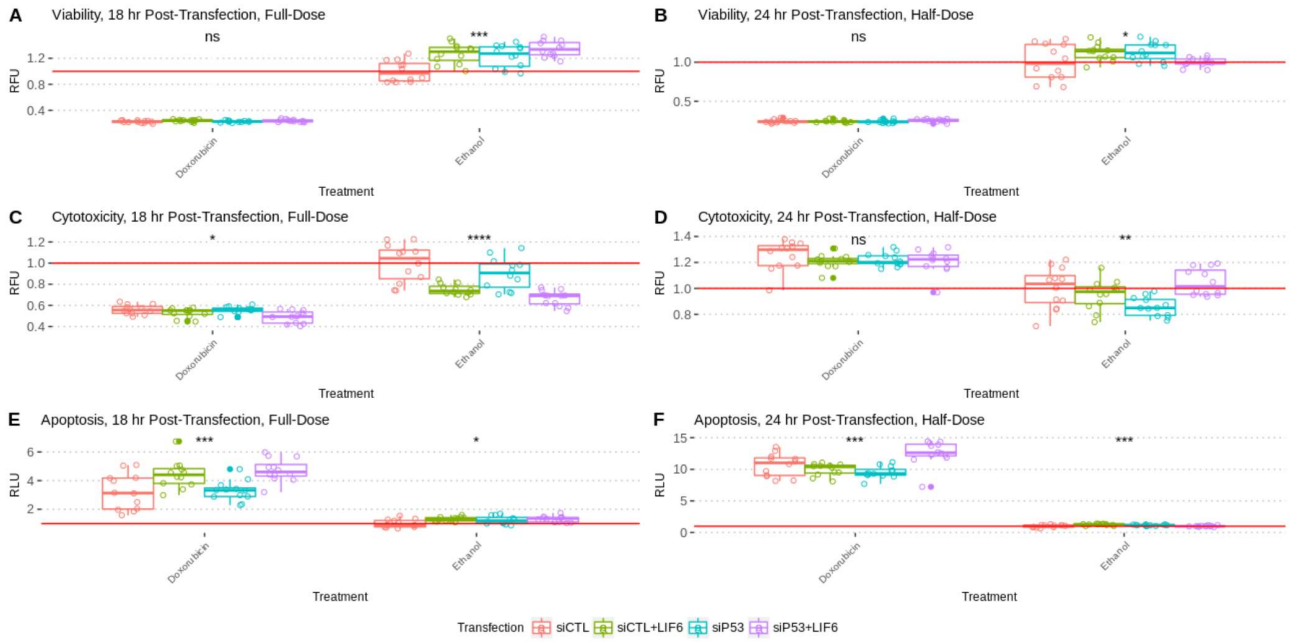

Figure S.3.3: ApoTox-Glo results for elephant cells treated with LIF6 and siRNA to knockdown TP53, related to Figure 3.4E. (Continued on next page)

Figure S.3.3: (Continued from previous page) Apoptosis (**A,B**), Cytotoxicity (**C,D**), and Viability (**E,F**) rates in African Elephant primary fibroblasts transfected with either scrambled control siRNA (siCTL) or anti-P53 siRNA (siP53); and with or without LIF6. After 6 hours of transfection, cells were treated with either 50-μM of Doxorubicin and tested 12 hours later at 18hr post transfection (**A,C,E**); or were treated with 25-μM Doxorubicin and tested 18 hours later at 24hr post-transfection (**B,D,F**). Co-transfecting siCTL with LIF6 results replicates the previously-seen apoptosis effect at 18 and 24 hours; at 24-hours, knocking down P53 rescues the apoptosis phenotype.

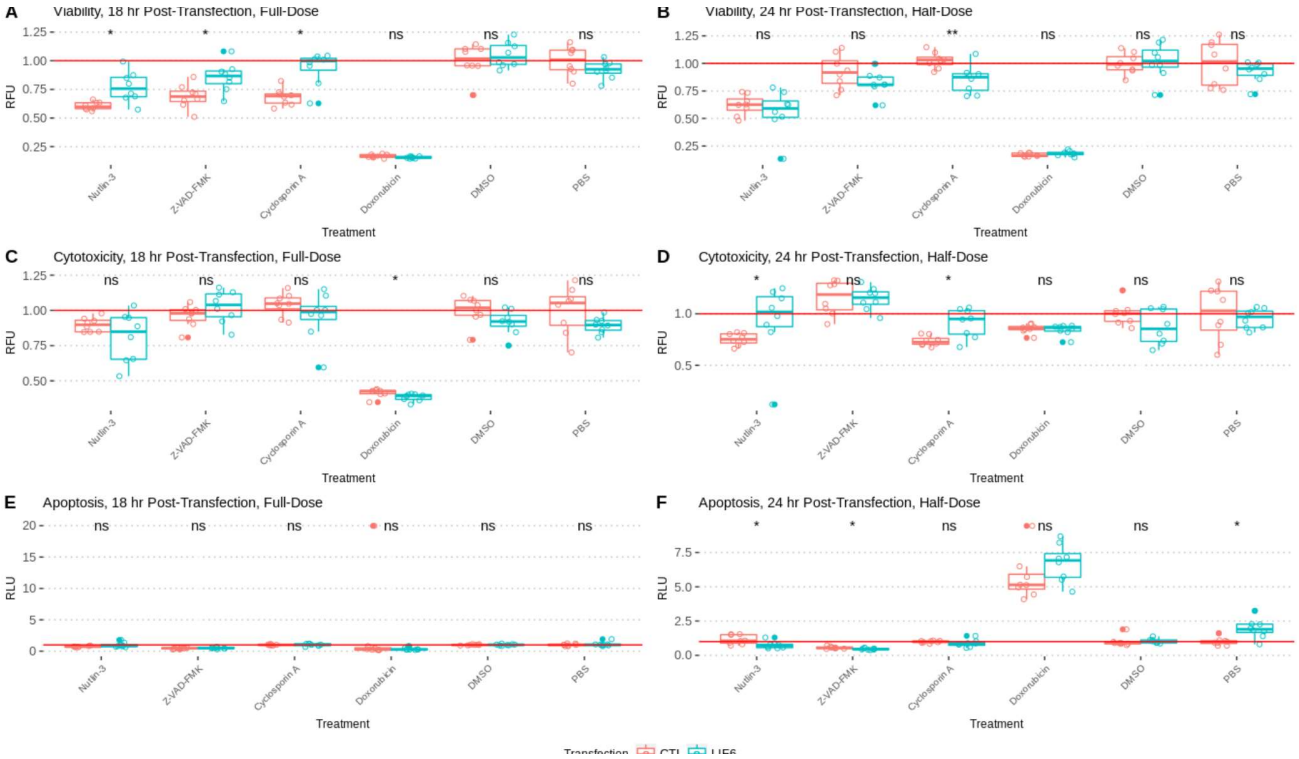

Figure S.3.4: ApoTox-Glo results for elephant cells transfected with LIF6, related to Figure 3.5B. (Continued on next page)

Figure S.3.4: (Continued from previous page) Apoptosis (**A,B**), Cytotoxicity (**C,D**), and Viability (**E,F**) rates in African Elephant primary fibroblasts transfected with *LIF6*, assayed the ApoToxGlo Triplex Assay. (**A,C,E**) Cells were treated 6 hours post-transfection with either 50-μM Nutlin-3, 20-μM Z-VAD-FMK, 2-μM Cyclosporin A, or 50-μM Doxorubicin, and were assayed 12 hours later, at 18 hours post-transfection. (**B, D, F**) Cells were treated as in **A**, **C**, and **E**, with half-doses of treatments, and tested 18 hours later at 24hr post-transfection. Apoptosis rates are markedly increased in cells transfected with LIF6 at 24 hours, which is inhibited by Z-VAD-FMK. Nutlin-3, which disrupts P53-MDM2 binding and thus activates P53, results in an increase in cytotoxicity, yet a decrease of apoptosis, in LIF6(+) cells compared to the mock-transfected control and to the PBS-treated LIF6(+) cells.

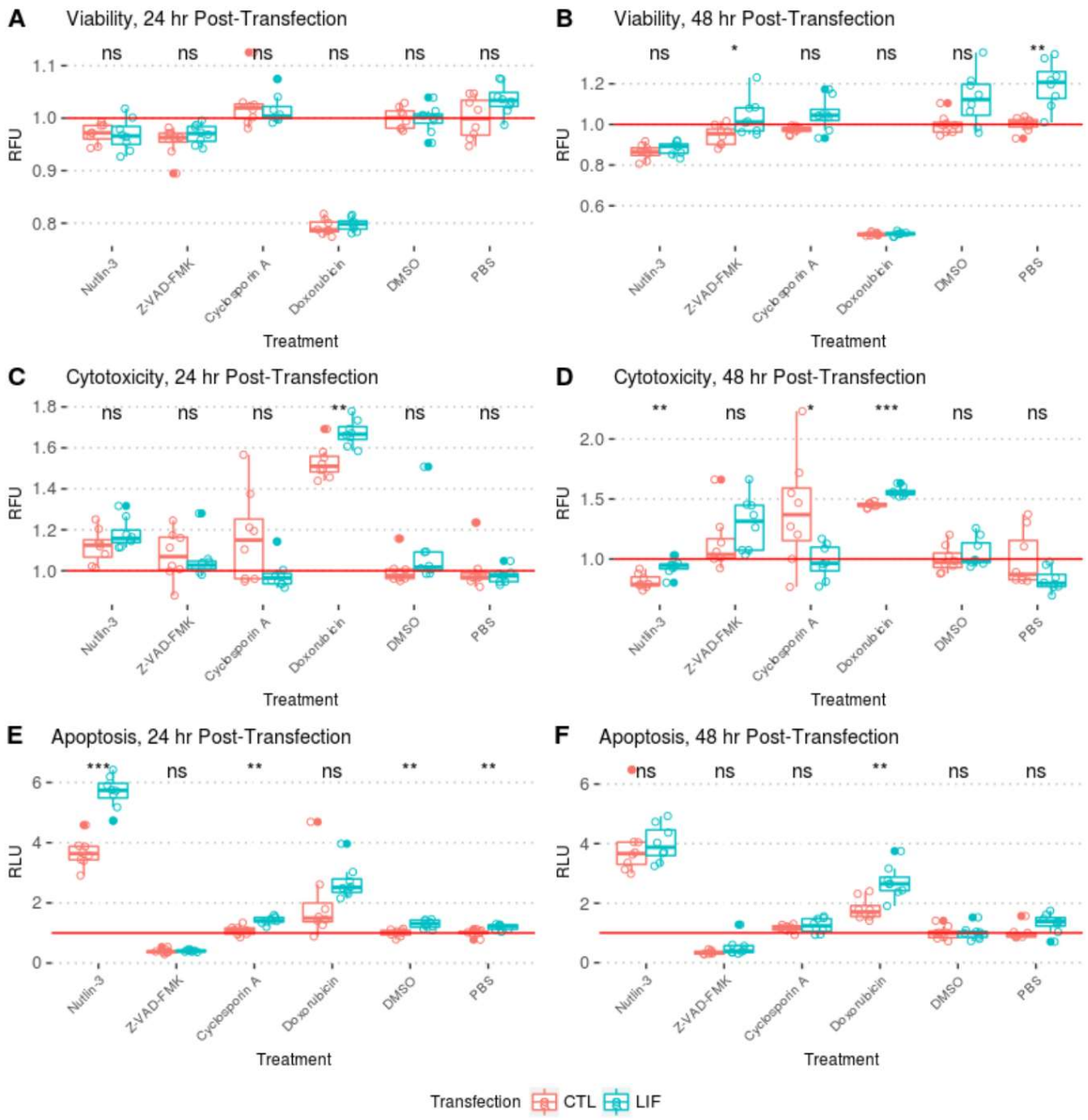

Figure S.3.5: ApoTox-Glo results for mouse embryonic fibroblasts (MEFs) transfected with LIF6, related to Figure 3.6C. (Continued on next page)

Figure S.3.5: (Continued from previous page) Cells were transfected with LIF6 for either 24-hours (**A-C**) or 48-hours (**D-F**), and were treated for 18 hours with either Nutlin-3, Z-VAD-FMK, Cyclosporin A, or Doxorubicin. The Viability (**A, D**), Cytotoxicity (**B, E**), and Apoptosis (**C, F**) rates in these cells were then measured using the ApoToxGlo Triplex Assay. Apoptosis rates are elevated for WT-MEF cells transfected with LIF6, but the effect is ablated when cells are treated with Z-VAD-FMK, a pan-caspase inhibitor; this ablation is not observed when treating cells with Cyclosporin A, an inhibitor of necrosis, indicating that the mechanism of LIF6-induced apoptosis is caspase-dependent. Treatment with Nutlin-3 - which increases P53 activity by disrupting binding between P53-MDM2 - intensifies apoptosis in LIF6 cells more than it does in untransfected WT-MEFs, suggesting a P53-dependent mechanism for caspase induction.

CHAPTER 4

A FULL-LOCUS DUPLICATION OF TP53 ENHANCES THE STRESS RESPONSE OF THE LITTLE BROWN BAT, *MYOTIS LUCIFUGUS*

4.1 Introduction

Bats are an exceptional clade that accounts for 20% of all extant mammalian species. [26] In addition to being the only volant clade of mammal, bats posess many unique adaptations, including echolocation and a high basal metabolism. [26, 16, 69, 169]. Bats are also phenotypically diverse, and come in a variety of sizes and lifespans (**Figure 4.1A**). For example, Kitti's hog-nosed bat (*Craseonycteris thonglongyai*) is the smallest species of bat, and weighs a maximum of 2.0 g [66]; on the other hand, the largest bat, the Giant golden-crowned flying fox (*Acerodon jubatus*) can weigh over 1 kg [64]. Similarly, the maximum lifespan of the shortest-lived bat, the Velvety free-tailed bat (*Molossus molossus*), is 5.6 years [51], while Brandt's bat (*Myotis brandtii*) can live over 41 years. [143] This dramatic diversity in life history traits is especially interesting given that the common ancestor of bats was only 58.9 MYA, and so occured in a small amount of evolutionary time. [3]

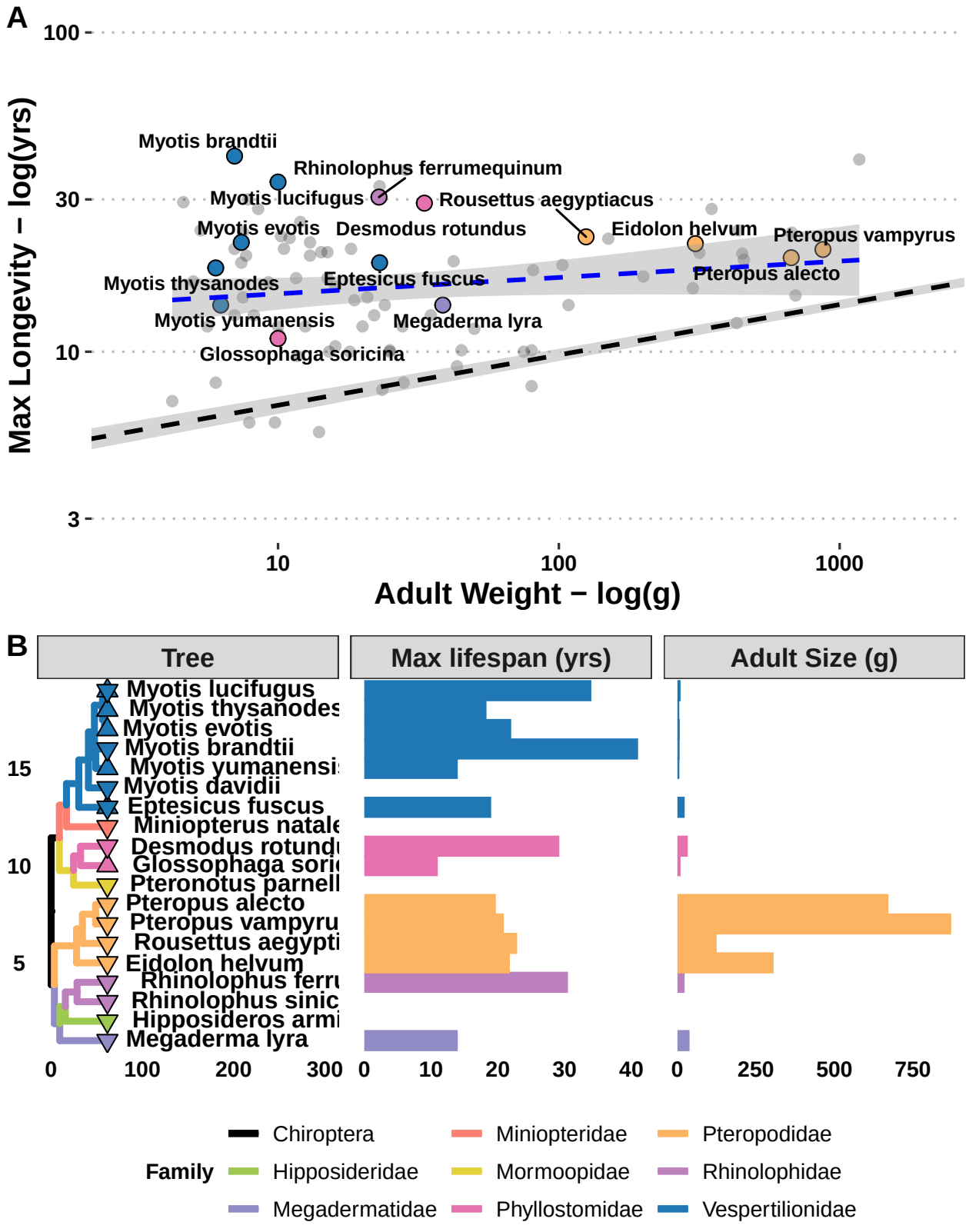

Figure 4.1: Bats come in a variety of sizes and lifespans. (continued on next page)

One of the constraints on the evolution of long life spans is an increased risk of developing cancer. Variations in either size or lifespan can affect an individual's cancer risk. For example, taller humans and larger dogs have an increased risk of various cancers compared to shorter, smaller individuals [127, 15]; similarly, older humans and dogs are at higher risk of cancer incidence than younger individuals [172, 15]. These observations fit the multistage model of carcinogenesis, which postulates that all cells have an intrinsic risk of becoming cancerous, and that this risk increases as a function of time. [6, 7, 126]; however the correlations between size, lifespan, and cancer risk disappear when comparing different species. [1, 141, 126] This phenomena is known as Peto's Paradox, and bats appear to be no exception: there are remarkably few reports of cancer in bats, which combined with their long lifespan, suggests that overall cancer rates must be low. [4, 128, 13]

The evolutionary history and phylogeny of bats make them a perfect clade for studying Peto's Paradox. In order to escape Peto's Paradox, large and long-lived bats must have evolved enhanced cancer suppression mechanisms relative to smaller, shorter-lived species. Some bat families, such as *Myotis*, have members with similar body sizes, but where a few individual species have independently evolved much longer lifespans; similarly, we also see families such as *Pteropus* where the lifespans of member species is similar, but there is a great diversity in size (**Figure 4.1B**). The number of genetic changes involved in increasing size or lifespan – and the cancer resistance mechanisms that coevolved with them – are likely minimal due to the short divergence time between these bat species. As such, these represent ideal clades for studying the genetic causes of Peto's Paradox.

Figure 4.1: (Continued from previous page) **A)** The maximum lifespans and adult body sizes of bat species in HAGR. [173] Species of bats with either published genomes or primary cell lines are highlighted. The correlation between size and lifespan of mammals is represented by the black dotted line, while the best-fit line of size and lifespan in bats is shown by the blue dotted line. **B)** A time-calibrated phylogeny of selected bat species, with maximum body sizes and lifespans. Species with primary cell lines are marked with a star, while species with published genomes are marked with a downwards triangle. [3]

One possible mechanism for resolving Peto's Paradox is through the duplication of pre-existing tumor suppressor genes. Indeed, some tumor suppressor gene duplicates have been identified - but not functionally validated - in various species of bats. [158] In other large and long-lived species, tumor suppressor gene duplications have previously been described and shown to play a functional role in primary cells from these species [170, 180]. Sulak (2016) [170] demonstrated that the additional copies of TP53 in elephants were playing a functional role in amplifying their DNA damage response; however, 8 TP53 duplications were also described in the "microbat", *Myotis lucifugus*, which were not present in the "megabat" *Pteropus vampiris*. This suggests that additional copies of TP53 may have contributed to the evolution of augmented cancer resistance and long lifespans in the microbat lineage.

Here, we investigate whether any of these extra copies of TP53 in *Myotis lucifugus* were transcribed and played a functional role in their cells. Starting with 15 publically-available bat genomes, we used a Reciprocal Best-Hit BLAT approach to identify if any additional copies of TP53 were present in these other genomes, or if all 8 copies were unique to TP53. We then investigated whether or not these copies were transcribed in these bats using publically available RNA-seq datasets; we also quantified expression in primary cell lines using RT-qPCR. Finally, we investigated the functional role of the additional copies of TP53 by comparing the DNA damage response of cells from various closely related bat species. Here, we describe and characterize a duplication of the TP53-WRAP53 locus unique to *Myotis lucifugus*. The duplicate encodes a fully-functional copy of TP53, which sensitizes their cells to DNA damage.

4.2 Methods

4.2.1 Bat Primary Fibroblasts

All primary bat fibroblasts were kindly provided by William Kohler and Richard Miller. Bat cell media consists of a high-glucose DMEM base (Gibco 10566-016) supplemented with

GlutaMax (Gibco 35050-061), 10% FBS (Gibco 26140-079), 1% Sodium pyruvate (Gibco 11360-070), and 1% Pen-Strep (Gibco 15140-122). All bat cells are propagated using the same protocol; for experiments, cells for each species were passaged in parallel using reagents from the same lot.

4.2.2 Cell Culture

A plate of cells was rinsed with one volume of DPBS (Gibco 14190-250); cells were then incubated in 0.25% Trypsin-EDTA (Gibco 25200-072) for 5-7 minutes at 37°C at 5% CO_2. After incubation, the cell suspension was transfered to a 15-mL conical tube (Thermo 339650) with an equal volume of media to stop trypsinization. Cells were then pelleted at $500g$ for 5 minutes, then resuspended in 1 mL media. The concentration of live cells in suspension was determined using a TC10 Automated Cell Counter (Bio-Rad); the concentration of cells in suspension was then adjusted to 10^6 cells/mL by adding additional media. For propagation, $5x10^5$ cells were plated in a T75 flask (Thermo 156499); for ApoTox Glo and other experiments, $1x10^6$ cells were plated in a T175 flask (Thermo 159910).

4.2.3 Transfection of Bat Cells

Initial attempts to transfect cells using lipofection resulted in low transfection efficiencies (data not shown). Cells were transfected using the Amaxa Basic Nucleofector? Kit for Primary Mammalian Fibroblasts (Lonza VPI1002) and a Nucleofector-2b device (Lonza AAB-1001). Transfections were done as per the instruction manual using the U-12 program, $1x10^6$ cells, and either 5 μg of DNA or 5 pmol of siRNA. All transfections were done in duplicate or triplicate. After transfection, replicate transfections were pooled and cells were plated in T25 flasks; media was exchanged after 12 hours, and cells were harvested 24-48 hours after transfection for further experiments.

4.2.4 Identification of TP53 Copy Number in 15 Bat Genomes via Reciprocal Best-Hit BLAT

The Reciprocal Best-Hit BLAT was done locally. The BLAT component programs `gfServer` and `gfClient`, (version 0.351, [90]), the genomes of 15 species of bat (**Table 4.1**), and the human genome (hg38, [58]) were downloaded. To prepare the databases for gfServer, the genomes were soft-masked using available RepeatMasker tracks, and converted into twoBit files using the UCSC tool `faToTwoBit`. `gfServer` was used to host the genomes locally in memory for searching using the UCSC reccomended settings. To search the genomes, the protein sequence of human TP53 (Uniprot P04637-1) [24] was used as the initial query sequence. The genomes were queried using `gfClient`; the nucleotide sequence corresponding to the top hit in each species were then used as the query for a second search within each species. To perform the reciprocal search, all hits of the second search using the native TP53 was used as the query for a search against hg38. Every hit that returned the human TP53 locus as the top reciprocal hit was noted as a copy of TP53.

4.2.5 Treatment and RNA Extraction

All samples were generated in parallel using 6-well plates. For each bat species, 500,000 cells were plated per well and allowed to grow for 24 hours. Cells were then treated for 4 hours with either etoposide (Cayman 12092); paraquat (Sigma 36541); tunicamycin (Cayman 11445); hydrogen peroxide (Sigma); or DMSO as a control. After incubation, the cells were rinsed once with DPBS, then lysed in-place using 350 μL of Buffer RLT Plus per well. RNA was extracted using the RNEasy Plus Mini Kit (Qiagen 74134) by following the standard protocol. Concentration of RNA was measured using a NanoDrop 2000 spectrophotometer (Thermo ND-2000).

4.2.6 RT-qPCR of TP53 response in response to stress

cDNA from all samples was generated using the QuantiTect Reverse Transcription Kit from Qiagen (Qiagen 205311), including the DNA removal step. Primers specific to either GAPDH, TP53.1, or TP53.2 were designed using Primer-BLAST, and were validated by sequencing (**Table 4.3**). [188] The QuantiTect SYBR Green PCR Kit (Qiagen 204141) was used for all qPCR reactions. Reactions were run for 100 cycles; TP53-knockdown samples were run for 300 cycles.

4.2.7 Sample Prep, Library Preparation and RNA Sequencing

Cells from *Eptesicus fuscus* and two individuals of *Myotis lucifugus* were transfected with either TP53 siRNA or a scrambled siRNA control in triplicate. Cells were allowed to recover for 36 hours, and then were replated in 6-well plates. Cells were then treated with for 4 hours with either etoposide (Cayman 12092); paraquat (Sigma 36541); tunicamycin (Cayman 11445); hydrogen peroxide (Sigma); or with control media. RNA was extracted as described, and then quantified using an Agilent 2100 BioAnalyzer with the RNA 6000 Nano Kit (Agilent 5067-1511). Due to a combination of poor quality and quantity, Peroxide samples and matching controls were recollected in a second batch.

4.2.8 RNA-seq Analysis

The SRA accession numbers of the RNA-seq datasets used in our analysis are noted in **Table 4.2**. Reads from each SRA record were mapped to their respective genome using `HISAT`. [92] Mapped reads were then assembled into initial putative transcripts using `StringTie`, and were merged into a guide GTF file using `StringTie --merge`. [139]. A final set of transcripts for each SRA record was made by re-running `StringTie` using the guide GTF.

4.2.9 Dual Luciferase Assays for Promoter Activity

The promoter region for both copies of TP53 were identified via Reciprocal Best-Hit BLAT using the human TP53 promoter as a starting point ([173], HAGRID 0006), followed by narrowing down the region of homology via Reciprocal Best-Hit BLAT between the two promoters. The final promoters sequences for TP53-1 and TP53-1 are referred to as pTP53-1 and pTP53-2, respectively. These sequences were synthesized and cloned (GenScript) into the pGL4.14 and pGL4.26 empty vectors, (EV) creating the following vectors: pGL4.14/EV; pGL4.14/pTP53-1; pGL4.14/pTP53-2; pGL4.26/EV; pGL4.26/pTP53-1; and pGL4.26/pTP53-2.

For the dual luciferase assay, *Myotis lucifugus* cells were transfected with 1:10 mixtures of a Renilla luciferase vector and each of the experimental Firefly luciferase vectors. 24 hours post-transfection, cells were harvested via trypsinization and replated in two 96-well plates (Corning 353296) at a density of 5000 cells/well. After a further 24 hours, the media was aspirated from the plates; half the wells were then filled with media supplemented with 20 μM of either Etoposide or Nutlin, while the other half were filled with DMSO-treated control media. The plates were incubated for 6 hours, and were then rinsed once with PBS. 100 μL of 1x Passive Lysis Buffer (Promega) was then added to each well, and the plates were shaken at room temperature for 1 hour. Readings were taken using a 96-well luminometer with dual-injectors for Renilla and Firefire luciferase reagents (Promega).

4.2.10 Kinetic measurements of Apoptosis and Necrosis Rates

In order to determine the rates of apoptosis for our cell lines, we utilized the RealTime-Glo assay from Promega. The RealTime-Glo assay uses an Annexin-V-based luciferase probe to detect early-stage apoptosis in the mitochondria, while simultaneously measuring cellular permiability using a DNA-sensistive fluorophore. As such, it can detect either early-stage apoptosis (luminsescence, but no fluorescence), late-stage apoptosis (luminescence and fluorescence), or necrosis and other cell death pathways (fluorescence, but no luminescence).

Cells were plated at an initial concentration of 5000 cells per well. 24 hours after plating, the media in the plate was asperated column-by-column and replaced with 50 μL of appropriate treatment media, plus 50 μL of freshly-made 2x RealTime Glo Detection Reagent. Plates were simultaneously incubated at standard conditions and imaged every 15 minutes for 36 hours using a Cytation 5 multi-mode plate reader (BioTek).

4.2.11 Quantification of Viability, Cytotoxicity, and Apoptosis in Response to Stress using ApoToxGlo

In each assay, 6 cell lines representing either 5 species (with two *M. lucifugus* individuals), or 6 individuals of *M. lucifugus*, were tested at 3 distinct timepoints, with two cell lines per 96-well plate. Cells were plated at an initial concentration of 5000 cells per well. 24 hours after plating, the media was asperated column-by-column and replaced with 100 μL of appropriate treatment media. Plates were then assayed using the ApoToxGlo kit (Promega) in a 96-well luminometer (Promega).

4.3 Results

4.3.1 Myotis lucifugus *has a unique, functional duplication of the TP53 locus*

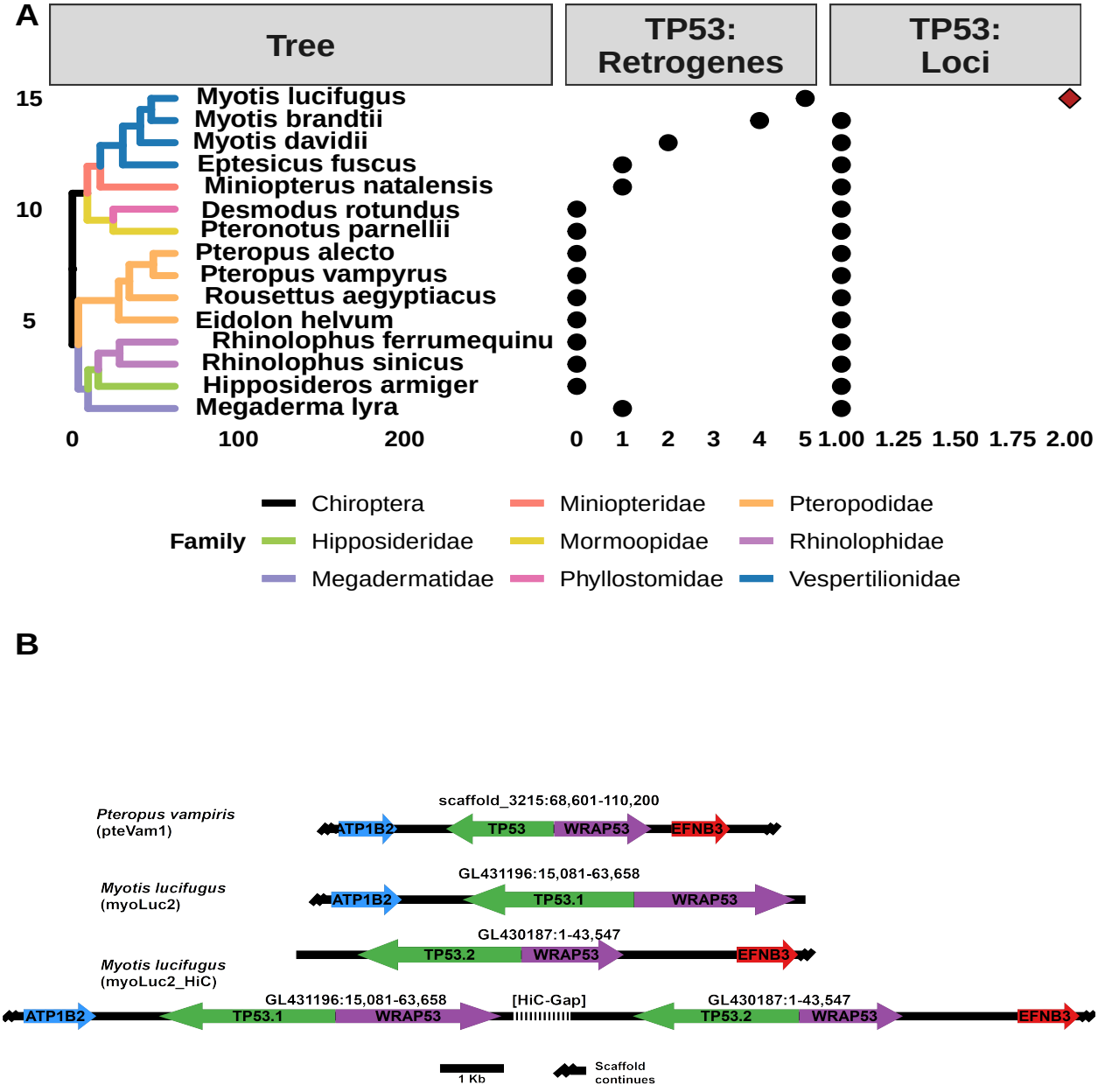

Figure 4.2: *Myotis lucifugus* has a unique, second copy of TP53. (Continued on next page)

Figure 4.2: (Continued from previous page) **A)** Copy numbers of TP53 across 15 bat genomes. Here, the TP53 locus is defined as the TP53 gene with conserved exon-intron structure, promoter region, and the adjacent WRAP53. **B)** A cartoon of the TP53 locus in the Megabat genome, *Pteropus vampyris* (pteVam1), the *M. lucifugus* genome (myoLuc2), and in the Hi-C-scaffolded myoLuc2 genome (myoLuc2-HiC). The duplication event occured within the boundaries of the genes flanking TP53; these flanking genes are not duplicated in the genome, suggesting that the duplication was only of TP53-WRAP53.

In order to determine the copy number of TP53 throughout *Chiroptera*, we used BLAT to search for all possible homologues of TP53 in the published genomes of 15 species of bat: Common vampire bat (*Desmondus rotundus*, desRot2), Straw-colored fruit bat (*Eidolon helvum*, eidHel1), Big brown bat (*Eptesicus fuscus*, eptFus1), Great roundleaf bat (*Hipposideros armiger*, hipArm1), Greater false vampire bat (*Megaderma lyra*, megLyr1), Natal long-fingered bat (*Miniopterus natalensis*, Eckalbar20161), Little brown bat (*Myotis lucifugus*, myoLuc2), Davids myotis (*Myotis davidii*, myoDav1), Black flying fox (*Pteropus alecto*, pteAle1), Brandts bat (*Myotis brandtii*, myoBra1), Large flying fox (*Pteropus vampiris*, pteVam1/pteVam2), Parnells mustached bat (*Pteronotus parnellii*, ptePar1), Greater horseshoe bat (*Rhinolophus ferrumequinum*, rhiFer1), Egyptian fruit bat (*Rousettus aegyptiacus*, rouAeg2), Chinese rufous horseshoe bat (*Rhinolophus sinicus*, rhiSin1) [116, 135, 37, 40, 103, 190, 158, 137]. Using a reciprocal best-hit BLAT approach, we validated all forward hits for TP53 in other genomes by searching the human genome using the putative TP53 hit: a forward hit was identified as TP53 if and only if human TP53 was the top hit of the reciprocal search. The results of these searches are detailed in **Figure 4.2A**. While we found that various species of bats within the superfamily *Vespertilionoidea* have multiple pseudogene copies of TP53, these copies are highly degraded and are not expressed in publically available RNA-seq datasets. Within the genome of *M. lucifugus*, however, we identified a second, full-length duplicate of the TP53-WRAP53 locus. These two loci are found on two separate scaffolds of the draft myoLuc2.0 genome. In other genomes, the TP53-WRAP53 locus is flanked by the genes ATP1B2 and EFNB3 on the 5' and 3' ends, respectively. In the myoLuc2 genome, the copy

of TP53 on scaffold GL431196 (TP53.1) is flanked by ATP1B2 at the 5' end, and the end of the scaffold at the 3' end; the other copy lies at the start of scaffold GL430187, and has EFNB3 on its 3' end (**Figure 4.2B**). This suggests that the two copies of TP53 originated via a syntenic duplication event between the two flanking genes. Confirming this suspicion, in a Hi-C scaffolded version of the myoLuc2 genome that was recently generated, we see that the two copies are indeed syntenic and located back-to-back as shown in Figure 4.2B. [39]

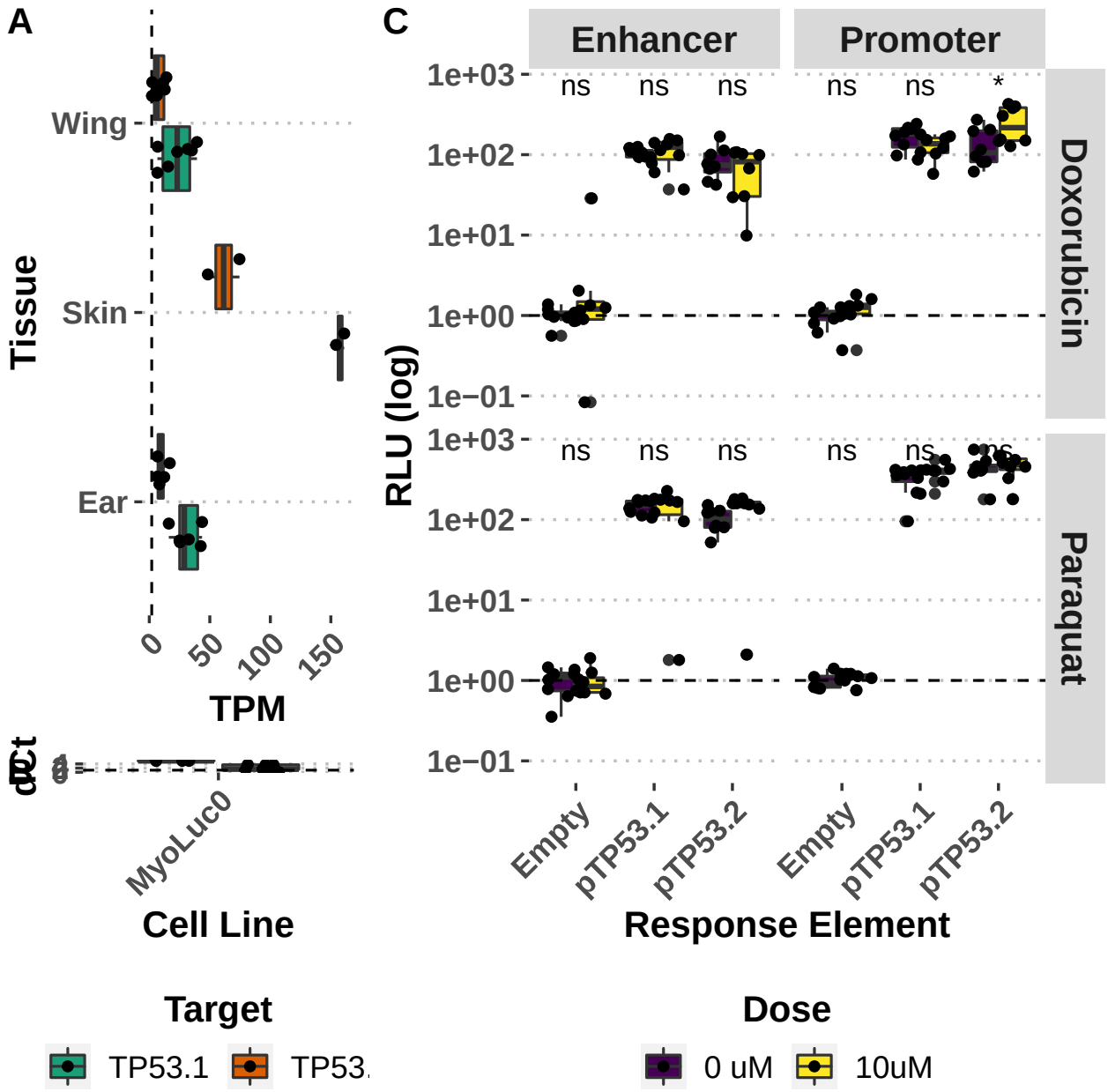

Figure 4.3: The two full-length copies of TP53 in *Myotis lucifugus* are expressed and driven by functional promoters. (Continued on next page)

To see if any of the additional copies of TP53 were expressed in any species, we assembled *de novo* transcriptomes from public RNA-seq datasets for each species using the "Tuxedo" suite of RNA-seq tools, including HISAT2 and StringTie [92, 139, 138]. Of all 7 copies of TP53 in *M. lucifugus*, only the two full-length copies, TP53.1 and TP53.2, showed any

Figure 4.3: (Continued from previous page) **A)** Transcripts per Million (TPM) abundances of transcripts uniquely mapping to either TP53.1 or TP53.2. No transcripts were mapped to any of the TP53 pseudogenes. **B)** RT-qPCR of the primary *Myotis lucifugus* cell line, myoLuc0, showing expression of TP53.1 and TP53.2 in untreated cells. **C)** The two promoters show strong activity in a dual luciferase assay. Values are normalized to co-transfected *Renilla* luciferase and to the empty vector controls. Expression remains high even after treatment with either Doxorubicin (DNA damage) or Paraquat (mitochondrial oxidative stress).

evidence of transcriptional activity (**Figure 4.3A**); none of the other copies in the myoLuc2 genome showed any transcriptional activty. Similarly, for all other bat species, only the canonical copy of TP53 showed evidence of transcription, suggesting that none of the TP53 pseudogenes identified in *Chiroptera* are functional. We additionally confirmed that TP53.1 and TP53.2 are expressed in one of our primary cell lines from *Myotis lucifugus,* myoLuc0* (**Figure 4.3B**).

As both TP53.1 and TP53.2 have preserved the TP53 promotor site, we hypothesized that the regulatory activity at these sites may be driving expression of both transcripts. The two promoter regions contain binding sites for TATA-Binding Protein, NFkB, and TP53 according to JASPAR and CONSITE [Supplementary Methods 91, 153]. To test whether either of the two promoter sequences have *in vitro* promoter or enhancer function, we cloned the two promoter sequences into two vectors, pGL4.14 and pGL4.26. Each vector has a firefly luciferase. In pGL4.14, expression of the luciferase is dependant on the promoter potential of the inserted DNA, while pGL4.26 has a minimal promoter, and therefore tests if the DNA has enhancer activity. As shown in **Figure 4.3C**, both promoters have exceedingly high promoter and enhancer activity, with 100-fold increases in luciferase expression relative to their respective empty vectors.

The activity of the promoters was not significantly influenced by the addition of either Doxorubicin (a DNA damaging agent), or Paraquat (which induces oxidative stress). While unexpected, this may have been due to the exceedingly high levels of activity of these two promoters, or due to a lack of turnover of the luciferase at the protein level. Nonetheless, the

results demonstrate that both copies of TP53 are expressed, both in the whole-organism and in primary cell culture; and that they possess functional promoter sequences that drive their expression.

4.3.2 Myotis lucifugus *is more sensitive to various sources of stress than other bat species*

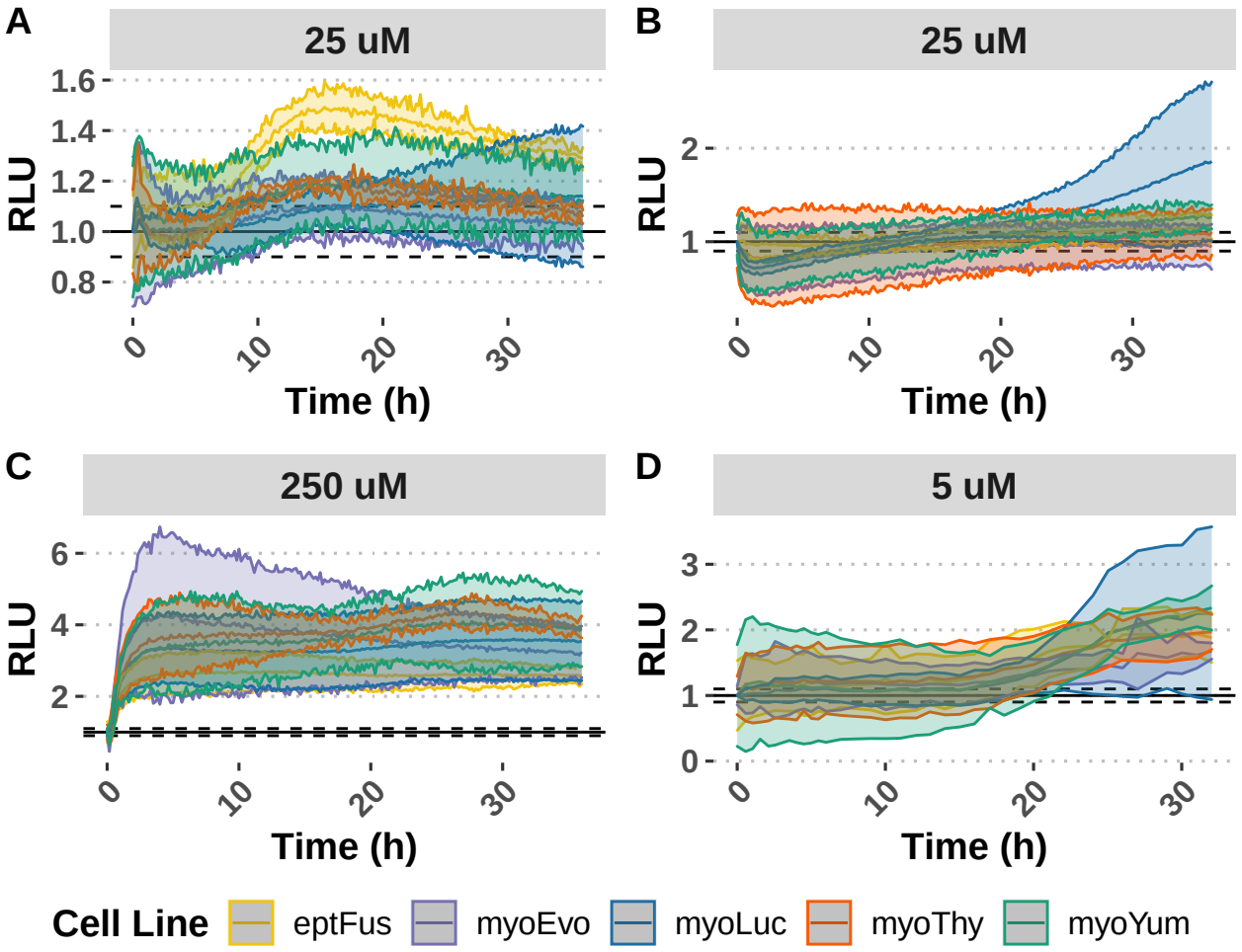

Figure 4.4: Kinetic rates of apoptosis from DNA damage (Etoposide, **A**), oxidative stress (Paraquat, **B**; Hydrogen Peroxide, **C**), and unfolded protein response (Tunicamycin, **D**) were assessed in 5 different bat species. The kinetics of apoptosis in response to each treament was measured over 36 hours.

One possible outcome of having multiple copies of TP53 is an increased sensitivity to external sources of stress. In order to quantify the stress response of *Myotis lucifugus*, we obtained primary skin fibroblasts derived from 11 individuals, and tested their sensitivity to various sources of stress compared to 5 other closely related bat species: *Myotis evotis*, *Myotis thysanodes*, *Myotis yumanensis*, and *Eptesicus fuscus*. We treated these cells with various doses of Etoposide (DNA double-strand breaks); Hydrogen Peroxide (general oxidative stress); Paraquat (mitochrondrial-specific oxidative stress); and Tunicamycin (unfolded protein response). We then assayed both the kinetics and the dynamic range of apoptosis, necrosis, or cell cycle arrest in these cells using two assays: the RealTime-Glo assay, and the ApoTox-Glo assay. We hypothesized that *M. lucifugus*, with its multiple copies of TP53, would either react faster TP53-depended forms of stress; or react to the stressors with either increased magintude of response, or a different approach to resolving the stress.

The kinetics of apoptosis, as measured using the RealTime-Glo assay in Figures 4, do not vary between species, although the magnitude of the shift does differ slightly, and are similar between species. 6 hours after treatment with either Etoposide or Paraquat, we see a sharp increase in the amount of DNA damage in the cells of various bat species. At the doses of Etoposide and Paraquat where we see the greatest differences in the apoptosis rates between the different species, we do not see any differences in the amount of DNA damage between the species. This suggests that the cause of cellular death between the difference species is not the damage itself, but rather how the cells respond to it. Further supporting this idea is the observation that both TP53.1 and TP53.2 increase in expression at the same timepoint during treatment (**Figure 4.2B**).

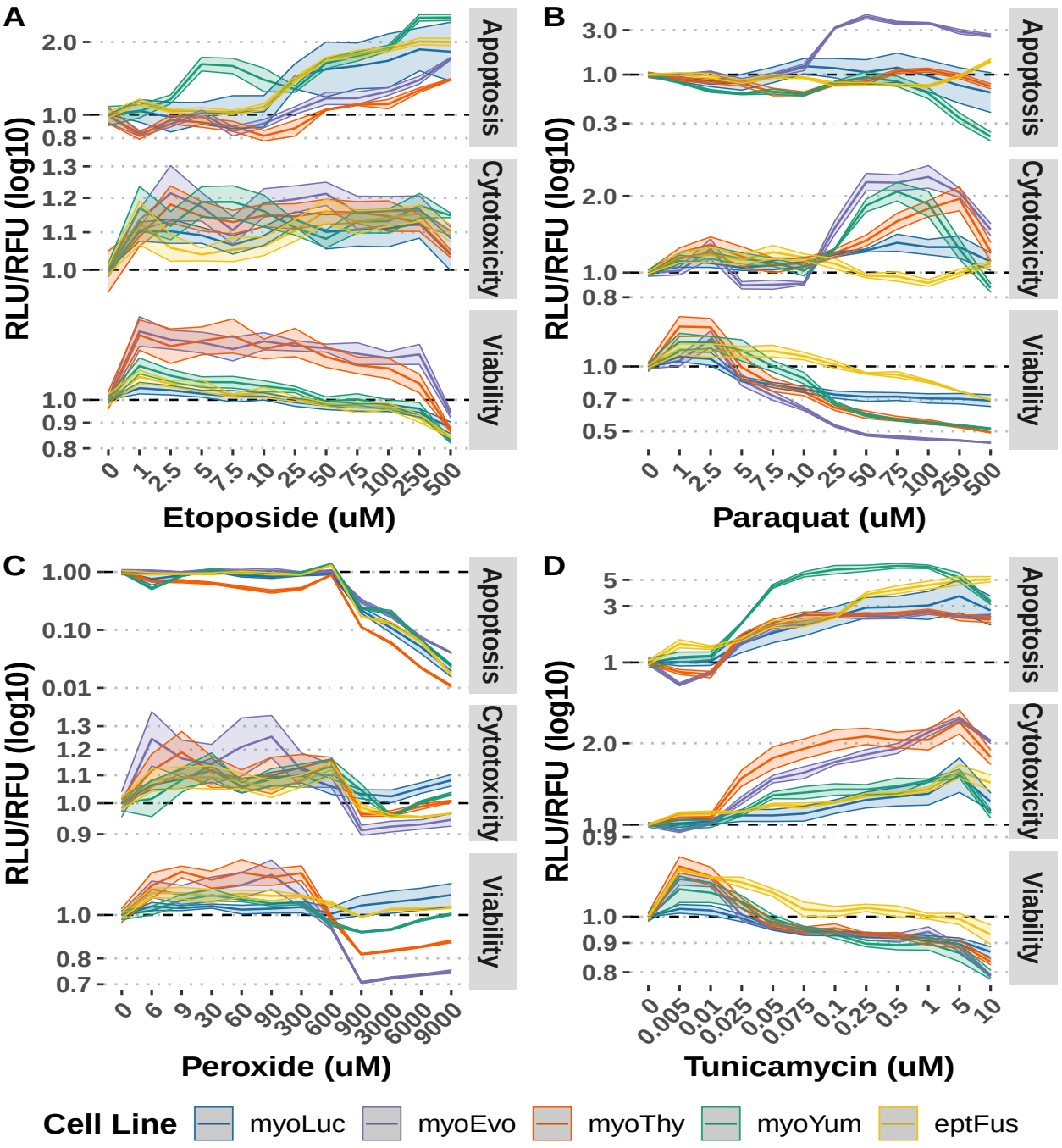

Figure 4.5: Dose-response curves for cell viability, cytotoxicity, and apoptosis from DNA damage (Etoposide, **A**), oxidative stress (Paraquat, **B**; Hydrogen Peroxide, **C**), and unfolded protein response (Tunicamycin, **D**) were assessed in 5 different bat species. Measurements were taken at 24 hours for all stresses, except for Paraquat (**B**), which was measured after 48 hours.

In order to test the hypothesis that *Myotis lucifugus* experiences a distinct stress response pattern than its sister species, we used the ApoTox Glo assay to quantify dose-dependent levels of viability, cytotoxicity, and cell death in response to our 4 stressors (Figure 4.5). At low doses of Etoposide ($\leq 10\mu M$) (Figure 4.5A), *Myotis lucifugus* - unlike other *Myotis* species - does not have a significant reaction in terms of apoptosis, and only shows a slight increase in cell cycle arrest (signfied by an elevated viability and cytotoxicity). However, the increase in cytotoxicity is less than the other *Myotis* species, suggesting that other species are experiencing great non-apoptotic cell death than *M. lucifugus*. Similarly, in response to Hydrogen Peroxide (Figure 4.5C), we see that *Myotis lucifugus* has a flat signal across the dosage range, with only cell cycle arrest signals apparent. *Myotis lucifugus* is more sensitive to Paraquat treatment (Figure 4.5B) compared to other species, except for *Myotis evotis*. Unlike Hydrogen Peroxide, paraquat is oxidized in the mitochrondria, and induces mitochondrial-origin redox stress; as such, the difference between the two suggests that while *Myotis lucifugus* is more readily able to respond to cytosolic redox stress and oxidative DNA/RNA damage than other species, within the mitochondria, it is more prone to the induction of apoptosis. Together, these results suggest that *Myotis lucifugus* has higher tolerance to stress than other bat species.

4.4 Discussion

The resistance of bats to various forms of stress has long been a focus of the literature [84, 144, 29, 152, 45, 16]. Due to the high metabolic demands of flight, bats are near-constantly subjected to conditions of temperature and oxidative stress that would be dangerous to other organisms. As such, many of these studies focused on comparisons to land-dwelling mammals; however, fewer studies have been performed to compare and characterize the stress tolerance between bat species, let alone closely-related ones.

With the advent of next-gen sequencing technologies came various studies looking at how

genetic variation between bats, and how signatures of positive selection were enriched near longevity-associated genes in long-lived bats; indeed, TP53 had previously been identified as a gene undergoing recent positive selection in long-lived bats. [158] Additionally, some studies had previously identified putative gene duplications in long-lived bats such as *Myotis lucifugus* [158].

With the comprehensive inter-species stress tolerance comparisons in this study, we have shown that *Myotis lucifugus* has a unique pattern of stress tolerance consistent with its increased longevity, and one where its duplication of TP53 may play a functional role. Lamentably, there were various problems in optimizing the knockdown via siRNA of TP53 in cells of *Myotis lucifugus*, and due to the extenuating circumstances created by the global COVID-19 pandemic, further work must be postponed. However, given the work that has been done in mice regarding TP53, we can hypothesize what the consequences of this duplication may be on the stress response patterns in *Myotis lucifugus*, and what the consequences of a knockdown would be.

While the possibility of performing *in vivo* studies of TP53 in *M. lucifugus* are remote due to various logistical and ethical concerns, we can draw on insights from previous work in mice. Initial generations of "super-TP53" mice were made by stably transfecting mice embryos with vectors overexpressing TP53; these mice, although tumor resistant, also suffered from premature aging due to chronic overactivation of TP53. This chronic overactivation of TP53 would lead to elevated rates of cell death and senescence, which would take their toll on the mice's health [178]. However, later generations of "super-P53" mice were made with BACs containing the 130 kb TP53-WRAP53 neighborhood. [52] These mice possessed the same tumor resistance phenotype, and enhanced levels of senescence and apoptosis in response to ionizing radiation - however, the mice were otherwise indistinguishable from healthy, normal mice, and no longer suffered from progeria. The presence of endogenous *cis* regulators in the BAC were likely responsible for keeping TP53 expression in check outside of moments of

acute stress. In light of these mouse models, we can expect that the TP53 duplication in *Myotis lucifugus* would reflect the latter model - and in fact, the stress response patterns that we have observed in this study strongly correlate with the increased-damage surveillance of the Garcia-Cao Super-P53 mice. As such, one would expect that a knockdown of TP53 in the Little Brown Bat would result in a decrease in apoptosis and senescence in response to DNA damage and oxidative stress, respectively, bringing these down to the value of the outgroup species.

Not only have the two copies of the TP53-WRAP53 loci conserved the same levels of regulatory activity as measured by dual luciferase output, but because they represent a syntenic duplication in the same topologically-associated domain. This would allow cells to not only conserve the *cis* regulation of both copies, but also enable a conservation of *trans* regulation in the 3D space of the genome. The existance of a natural example of a "super-P53" mammal opens up many such questions about the evolutionary cost and adaptations that come with such a development, such as whether there are negative pleitropic effects of this arrangement, and if there have been other compensatory genetic changes to accomodate this arrangement, such as the duplication or upregulation of additional TP53 regulatory proteins. Additionally, as WRAP53 is known to regulate TP53 expression at both the transcriptional and translational levels, more work must be done to determine what are the functional consequences of having an additional copy of WRAP53 in the genome.

4.5 TABLES

Table 4.1: Bat genomes used in this study.

Genome	Common Name	Species	Ref
desRot2	Common vampire bat	Desmondus rotundus	116
eidHel1	Straw-colored fruit bat	Eidolon helvum	135
eptFus1	Big brown bat	Eptesicus fuscus	Broad Institute
hipArm1	Great roundleaf bat	Hipposideros armiger	37
megLyr1	Greater false vampire bat	Megaderma lyra	135
minNat1	Natal long-fingered bat	Miniopterus natalensis	40
myoLuc2	Little brown bat	Myotis lucifugus	103
myoDav1	Davids myotis	Myotis davidii	190
pteAle1	Black flying fox	Pteropus alecto	190
myoBra1	Brandts bat	Myotis brandtii	158
pteVam1, pteVam2	Large flying fox	Pteropus vampiris	103
ptePar1	Parnells mustached bat	Pteronotus parnellii	135
rhiFer1	Greater horseshoe bat	Rhinolophus ferrumequinum	135
rouAeg2	Egyptian fruit bat	Rousettus aegyptiacus	137
rhiSin1	Chinese rufous horseshoe bat	Rhinolophus sinicus	37

Table 4.2: *Myotis lucifugus* SRAs used in this study.

Run	Sex	Tissue	Library Selection	Type	Size (MB)
SRR1270869	unknown	Ear	RANDOM	PAIRED	1173
SRR1270919	unknown	Ear	RANDOM	PAIRED	1540
SRR1270921	unknown	Ear	RANDOM	PAIRED	1610
SRR1270922	unknown	Ear	RANDOM	PAIRED	1532
SRR1270923	unknown	Ear	RANDOM	PAIRED	1680
SRR1916841	Female	Wing	PolyA	PAIRED	1565
SRR1869462	Male	Wing	PolyA	PAIRED	1822
SRR1916834	Female	Wing	PolyA	PAIRED	1615
SRR1916836	Female	Wing	PolyA	PAIRED	1655
SRR1916839	Male	Wing	PolyA	PAIRED	1629
SRR5676383	Male	Wing	cDNA	PAIRED	13345
SRR5676382	Male	Wing	cDNA	PAIRED	9648
SRR1013468	Male	unknown	PCR	PAIRED	2510
SRR4249979	unknown	Skin Fibroblast	cDNA	PAIRED	1173
SRR4249988	unknown	Skin Fibroblast	cDNA	PAIRED	1358

Table 4.3: Primer sequences used in this study.

ID	Target	Sequence
myoLuc2TP53.1-F1	TP53.1	GGGAAGGGACAGAGGATGAC
myoLuc2TP53.1-R1	TP53.1	TGACAATGATCTGAATCCTGAGG
myoLuc2TP53.2-F2.1	TP53.2	CAAAGAAGCCAGCGATGAA
myoLuc2TP53.2-R2.1	TP53.2	AAAGGTGCCGGTATTTTGCT
myoLuc2_GAPDH-F	GAPDH	TGACCCCTTCATTGACCTCAAC
myoLuc2_GAPDH-R	GAPDH	TGACTGTGCCCTTGAACTTG

CHAPTER 5

DISCUSSION & CONCLUSION

While the biology underlying the relationship between cancer, body size, and lifespan within species has been known since the 1950‘s, the mechanisms that disentangle these correlations between species are much less understood [7, 6, 140, 61, 62, 125, 177, 20, 9]. In order to understand Peto”s Paradox, as well as the evolution of longevity and body size, many groups have looked phenotypically *in vivo* and *in vitro* for traits, such as stress response, which are associated with longevity and body size across species [10, 55, 160, 63]. More recently, with the advent of genomic era, various groups have also begun to explore the genetic differences between long- and short-lived species from evolutionary and candidate gene perspectives [190, 158, 135, 40, 116, 44, 93, 43, 100, 54, 192, 89]. This work has examined an underappreciated angle by which Peto's Paradox can be resolved, which is through the duplication of tumor suppressor genes.

I have shown how long-lived species have duplicate many genes ” including tumor suppresors ” along their lineage (Chapter 2), and have demonstrated that many of these duplicate genes, such as LIF (Chapter 3) and TP53 (Chapter 4), have retained or regained their function as non-canonical and canonical tumor suppressors in cellula using functional genomics and primary cell cultures from the species in question. Given the number of genes identified, a full functional characterization of each one would be beyond the efforts of a single thesis; however, the filtering criteria and the expression of these duplicate genes *in vivo* suggest that they are conserved and functional, at either the RNA or protein level, and thus may be contributing to the resolution of Peto's Paradox in their host species.

Among the most interesting results from Chapter 2 was the discovery that tumor suppressor genes are frequently duplicated, and that these duplications occurred throughout *Atlantogenata*. This tracks with other studies observing that genes duplicate readily at a rate of 1% per gene per million years [106]. As such, it is possible that the observation that tumor

suppressor duplications precede body size increases represents more than a casual coincidence. However, given the limited number of genomes I have, relative to the large number of species that exist currently in this clade, I lack the power to confidently determine if body size and tumor suppression count is likely causal, rather than suggestively coincidental.

The results from Chapters 3 and 4 demonstrate *in vitro* the functional impact of the duplicates I identified in Chapter 2, and demonstrate how they likely act to suppress cancer risk of large, long lived species in vivo. In the case of LIF6, a pseudogene which duplicated in the common ancestors of elephants and manatees is resurrected by the evolution of a novel TP53 binding site upstream, and kills cells upon expression. This likely acts in concert with the other TP53-related duplications I identified in Chapter 1 to remove damaged cells in response to stress, thus conserving a living pool of undamaged cells that are likely non-cancerous. Meanwhile, the unique apoptotic and senescence response of cells from the Little Brown Bat, *Myotis lucifugus*, is consistent with the predicted effects having a second, full-length copy of the TP53 locus, which enables enhanced cell damage detection without premature aging in similar mouse models. While it appears that the specific genes that undergo duplication in each species are private, overall I showed that genes in similar tumor suppressor pathways are duplicated in large, long-lived species, which supports the thesis that gene duplication has provided an indispensable mechanism for resolving Peto's Paradox.

5.1 Limitations of approaches & impact on outcomes

In addition to the chapter-specific caveats that are discussed at length in each section, there are larger methodological and biological factors that frame these results. My studies, by focusing solely on protein-coding genes, only encompass a small fraction of a much greater fraction of the genome. Outside the domain of protein-coding genes, there is a world of "known-unknowns" in the genome that, while well-explored in model organisms of aging, remain woefully understudied in other long-lived species. And beyond the nucleotides of the

genome, there are layers upon layers of biology where more mechanisms underlying Peto's Paradox may hide.

Relationship between in vitro and in vivo effects of tumor suppressor genes

By using primary cell cultures to validate tumor suppressor duplicates *in vitro* rather than in vivo, there is always the threat of overlooking or overexaggerating the real biology that occurs within an organisms. Primary cell culture models are preferred over immortalized cell culture samples, as these cells have undergone various cytological and genetic changes, and may no longer reflect the original biology of the donor organism after so much time. However, even the youngest primary tissue samples will lack their original environmental context when removed from the donor. Decades of oncological work have demonstrated the impact that tumor environment plays on cancer cells, which can either restrict and limit the proliferation of cancer cells, or even promote their growth and metastasis [61, 62]. As such, it is difficult to ascertain any non-cell-autonomous cancer suppression mechanisms using a two-dimensional monoculture of primary cells from one or more individuals.

A strength of my *in vitro* study designs relative to prior studies is the use of closely-related species to address my questions in a phylogenetically-sound manner. Other comparative studies comparing cellular responses to stress frequently compare single representatives from each major class in *Eutheria*, which provides a very low resolution for studying the true association between traits such as cell stress, and the evolution of longevity and body size [10, 63, 84, 107]. There are a few studies comparing cancer resistance mechanisms at the transcriptomic level *in vivo*; however, these studies not only suffer from the same aforementioned evolutionary challenges in their design, but also from additional environmental and technical confounding factors due to their use third-party-generated data generated at different times and locations [109, 50]. True *in vivo* functional assays of tumor suppressors as described in this work in endangered and threatened species such as elephants and bats would not only be deeply unethical, but technically intractable.

Obviously, *in vivo* studies of tumor suppressor genes in species such as elephants, whales, and bats, are not possible, however, many aspects of cancer are cellular phenomena thus I can explore the biology of cancer using robust *in vitro* primary cell culture system that accounts for not only inter-individual variation, but also for tissue type and other technical batch effects. This depends greatly on one"s ability to find willing collaborators and ethical sources for fresh, primary tissue. For my work on Elephants (Chapter 3), I was forced to use an increased number of sample-level replicates in order to compensate for a lack of multiple individuals; however, for my work in Chapter 4 with the Little brown bat *Myotis lucifugus*, I was able to obtain samples from multiple individuals, and thus correct for inter-individual variability when looking at their stress response. However, due to the logistical and ethical challenges in collecting tissue samples, the cells used in this work are all skin fibroblasts; there is active debate in the literature as to whether or not these cells are the most relevant cell type for stress response and cancer resistance studies. For the purposes of this work, however, so long as the cell types and tissues of origin from each species are properly matched, any cell line that regularly exhibits neoplasia in the population that express the genes of interest would be suitable. As skin is the largest organ system in any mammal, and it"s neoplasia and carcinomas has one of the strongest correlations to body size of any tissue in humans [127, 56], I also find that it is especially relevant to my current question.

Exclusion of non-protein-coding genetic elements and their impact on Peto's Paradox

In my design, I intentionally and explicitly limit my initial search in Chapter 1 " and thus throughout this work - to only protein-coding genes. As such, the role of the noncoding genome on Peto"s Paradox has yet to be discussed, including gene regulatory elements and non-coding RNAs (ncRNAs). While I do explore changes in the regulatory landscape surrounding both LIF6 (Chapter 3) and TP53 (Chapter 4), I do not explore more expansively how sequence and coding conservation associates with longevity, body size, and other traits

relevant to Peto"s Paradox, as these are already adequately discussed in the literature[190, 158, 135, 40, 116, 44, 93, 43, 100, 54, 192, 89, 94]. On the other hand, ncRNAs are grossly understudied in the context of Peto's Paradox, likely due to costs of sequencing and the poor quality of many non-model-organism genomes. Nonetheless, at least one study has shown that siRNAs expressed in the blood of a long-lived species of bat likely regulate genes in oncogenic pathways such as inflammation [74].

The definition of "Tumor Suppressor" and polygenic effects in cancer resistance

While my study presupposes the existence and categorization of genes into three basic categories - tumor suppressor (anti-oncogenic), oncogenic, and other - the lines between these categories are frequently blurred and unclear. Many genes, such as APOBEC3B [182, @ Hashemi2018], AMPK [101], and p63 [117], are either a tumor suppressor or an oncogene depending on the context of their expression, the mutations they acquire, and even based on the stage of cancer where they become dysfunctional. It is theoretically possible that by duplicating these genes, the anti-oncogenic and the pro-oncogenic functions of these genes could be split between the two copies in an instance of sub-functionalization [167, 48, 147, 146]; or that redundant copies of tumor suppressors could abrogate any oncogenic effects of mutations in other copies [124, 32, 110, 23, 191].

Furthermore, our knowledge of how genes and gene networks contribute to cancer risk is still in its infancy, and it is possible that of the duplicated genes that I identified, many of them do play critical roles in suppressing cancer in large, long lived species, but that this biological function of theirs was overlooked or unknown. As large-scale genomics studies of cancer have become more cost-effective and computationally feasible, more and more data has come out showing the incredible genetic diversity within tumors. Endeavors such as the Catalogue of Somatic Mutations in Cancer (COSMIC) Cancer Gene Census [164] and the Cancer Genome Atlas [19] have begun to tease apart which genes and mutations are either

causal to or casualties from cancer. However, other analyses have shown that networks of mutations, rather than individual gene drivers, can also lead to dysregulation and oncogenesis in cells [78, 72, 28]. A systems biology approach to Peto's Paradox, where knowledge of the networks of cancer suppressing and cancer promoting pathways is leveraged to identify genes and regulatory elements that have evolved in large, long-lived species, is a tantalizing prospect; while this thesis lays some of the groundwork towards such a project, there is still much work to be done in establishing basic biology and genomic tools in these species before the question can be pursued.

Terra incognita of Peto's Paradox: epigenetic contributions to cancer resistance

Beyond the realm of genetic sequences and gene expression, there are many other ways that evolution can optimize cancer resistance and thus resolve Peto's Paradox. Of these, genomic stability has shown to be tightly associated with cancer risk, both within model organisms, as well as in large, long-lived species such as the Naked Mole Rat. [142, 93, 114, 168]. Among the genes that I identified in Chapter 1 as duplicated, I did find some genes that are involved in chromosome reorganization; additionally, there are other large, long-lived animals such as the Bowhead Whale which have duplications of proteins like the histone deacetylase SIRT7 [89]. While I did not find a significant enrichment of genes involved in these pathways, it is possible a small number of duplicated genes here play an outsized role, or that other forms of evolutionary adaptation in these processes are involved in mediating Peto's Paradox through genomic stability.

While it is clear that large, long-lived species must resolve their cancer risk during evolution in order to achieve their large sizes and lifespans, there are many ways that this can occur. In this work, I establish that gene duplication may contribute significantly to the ablation of cancer risk during the evolution of longevity and body size. In *Atlantogenata*, as body sizes expand, so do the copy numbers of tumor suppressor genes, suggesting that the two events

are tightly intertwined. In extant elephants, these tumor suppressors remain functional, and as such, are likely still functional, including a series of duplications both upstream and downstream of TP53, which was already known to be duplicated. I then characterized an unexpected, resurrected retrogene of LIF, called LIF6, which induces apoptosis in elephant cells in response to stress. And finally, I describe a syntenic TP53 duplication in the Little Brown Bat *Myotis lucifugus* which has preserved both regulatory potential and expression patterns similar to the canonical copy; the stress profile of the Little Brown Bat relative to other bat species matches what one would expect given what is known about a TP53 locus duplication in mouse models, but a causal role has yet to be established in vitro. While gene duplication alone cannot fully explain how large, long-lived species overcome their increased cancer risk, it does represent a major contributor to the mosaic of mechanisms at play in resolving Peto's Paradox.

References

[1] Lisa M Abegglen, Aleah F Caulin, Ashley Chan, Kristy Lee, Rosann Robinson, Michael S Campbell, Wendy K Kiso, Dennis L Schmitt, Peter J Waddell, Srividya Bhaskara, Shane T Jensen, Carlo C Maley, and Joshua D Schiffman. Potential Mechanisms for Cancer Resistance in Elephants and Comparative Cellular Response to DNA Damage in Humans. *JAMA*, 314(17):1850–1860, 2015.

[2] Enis Afgan, Dannon Baker, Marius van den Beek, Daniel Blankenberg, Dave Bouvier, Martin Čech, John Chilton, Dave Clements, Nate Coraor, Carl Eberhard, Björn Grüning, Aysam Guerler, Jennifer Hillman-Jackson, Greg Von Kuster, Eric Rasche, Nicola Soranzo, Nitesh Turaga, James Taylor, Anton Nekrutenko, and Jeremy Goecks. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic Acids Research*, 44(W1):W3–W10, July 2016.

[3] Ingi Agnarsson, Carlos M Zambrana-Torrelio, Nadia Paola Flores-Saldana, and Laura J May-Collado. A time-calibrated species-level phylogeny of bats (chiroptera, mammalia). *PLoS currents*, 3:RRN1212, 11 2011.

[4] Thales A. F. Albuquerque, Luisa Drummond do Val, Aoife Doherty, and João Pedro de Magalhães. From humans to hydra: patterns of cancer across the tree of life. *Biological Reviews*, (3), 2018.

[5] Adrian M Altenhoff and Christophe Dessimoz. Phylogenetic and functional assessment of orthologs inference projects and methods. *PLoS computational biology*, 5(1):e1000262, 2009.

[6] P Armitage. Multistage models of carcinogenesis. *Environmental health perspectives*, 63:195–201, 1985.

[7] P Armitage and R Doll. The Age Distribution of Cancer and a Multi-stage Theory of Carcinogenesis. *British Journal of Cancer*, 8(1):1–12, 1954.

[8] O Ashur-Fabian, A Avivi, L Trakhtenbrot, K Adamsky, M Cohen, G Kajakaro, A Joel, N Amariglio, E Nevo, and G Rechavi. Evolution of p53 in hypoxia-stressed Spalax mimics human tumor mutation. *Proceedings of the National Academy of Sciences*, 101(33):12236–12241, 2004.

[9] S N Austad and K E Fischer. Mammalian aging, metabolism, and ecology: evidence from the bats and marsupials. *Journal of gerontology*, 46(2):B47–53, 1991.

[10] S.N. Austad. Methusaleh's Zoo: How Nature provides us with Clues for Extending Human Health Span. *Journal of Comparative Pathology*, 142:S10–S21, 2010.

[11] Donald A. Barr, Maria Elena Gonzalez, and Stanley F. Wanat. The leaky pipeline: Factors associated with early decline in interest in premedical studies among underrepresented minority undergraduate students. *Academic Medicine*, 83(5), 2008.

[12] Euan W Baxter and Jo Milner. p53 Regulates LIF expression in human medulloblastoma cells. *Journal of neuro-oncology*, 97(3):373–382, May 2010.

[13] MARK BECK, J. BECK, and EDWIN B. HOWA. BILE DUCT ADENOCARCINOMA IN A PALLID BAT (ANTROZEOUS PALLIDUS). *Journal of Wildlife Diseases*, 18:365–367, 1982.

[14] Olaf R P Bininda-Emonds, Marcel Cardillo, Kate E Jones, Ross D E MacPhee, Robin M D Beck, Richard Grenyer, Samantha A Price, Rutger A Vos, John L Gittleman, and Andy Purvis. Erratum: The delayed rise of present-day mammals. *Nature*, 456(7219):274–274, 2008.

[15] R T Bronson. Variation in age at death of dogs of different sexes and breeds. *American journal of veterinary research*, 43:2057–9, 1982.

[16] Anja K Brunet-Rossinni. Reduced free-radical production and extreme longevity in the little brown bat (Myotis lucifugus) versus two non-flying mammals. *Mechanisms of Ageing and Development*, 125:11–20, 2004.

[17] J Cairns. Mutation selection and the natural history of cancer. *Science of Aging Knowledge Environment*, 2006.

[18] Lauren Peirce Carcas. Gastric cancer review. *Journal of carcinogenesis*, 13(1):14, 2014.

[19] Jian Carrot-Zhang, Nyasha Chambwe, Jeffrey S. Damrauer, Theo A. Knijnenburg, A. Gordon Robertson, Christina Yau, Wanding Zhou, Ashton C. Berger, Kuan-lin Huang, Justin Y. Newberg, R. Jay Mashl, Alessandro Romanel, Rosalyn W. Sayaman, Francesca Demichelis, Ina Felau, Garrett M. Frampton, Seunghun Han, Katherine A. Hoadley, Anab Kemal, Peter W. Laird, Alexander J. Lazar, Xiuning Le, Ninad Oak, Hui Shen, Christopher K. Wong, Jean C. Zenklusen, Elad Ziv, Cancer Genome Atlas Analysis Network, Francois Aguet, Li Ding, John A. Demchok, Michael K.A. Mensah, Samantha Caesar-Johnson, Roy Tarnuzzer, Zhining Wang, Liming Yang, Jessica Alfoldi, Konrad J. Karczewski, Daniel G. MacArthur, Matthew Meyerson, Christopher Benz, Joshua M. Stuart, Andrew D. Cherniack, and Rameen Beroukhim. Comprehensive Analysis of Genetic Ancestry and Its Molecular Correlates in Cancer. *Cancer Cell*, 37(5):639–654.e6, 2020.

[20] Aleah F Caulin, Trevor A Graham, Li-San Wang, and Carlo C Maley. Solutions to Peto's paradox revealed by mathematical modelling and cross-species cancer gene analysis. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 370(1673):20140222, 2015.

[21] Aleah F Caulin and Carlo C Maley. Peto's Paradox: evolution's prescription for cancer prevention. *Trends in ecology & evolution*, 26(4):175–82, 2011.

[22] K Chen, D Durand, and M Farach-Colton. NOTUNG: a program for dating gene duplications and optimizing gene family trees. *J. Comput. Biol.*, 7(3-4):429–447, 2000.

[23] Wei-Hua Chen, Xing-Ming Zhao, Vera van Noort, and Peer Bork. Human monogenic disease genes have frequently functionally redundant paralogs. *PLoS computational biology*, 9(5):e1003073, 2013.

[24] The UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Research*, 45(D1):D158–D169, 2017.

[25] Natalie Cooper and Andy Purvis. Body Size Evolution in Mammals: Complexity in Tempo and Mode. *The American Naturalist*, 175(6):727–738, 2010.

[26] G B Corbet and J E Hill. Wilson, D. E., and D. M. Reeder (eds.). 1993. MAMMAL SPECIES OF THE WORLD: A TAXONOMIC AND GEOGRAPHIC REFERENCE, 2nd Edition. Smithsonian Institution Press, Washington, D.C., xviii + 1206 pp. ISBN 1-56098-217-9. Price (hardcover), $75.00 ($60.00 to members of The American Society of Mammalogists). *Journal of Mammalogy*, 75:239–243, 1994.

[27] Diego Cortez, Ray Marin, Deborah Toledo-Flores, Laure Froidevaux, Angelica Liechti, Paul D Waters, Frank Grutzner, and Henrik Kaessmann. Origins and functional evolution of Y chromosomes across mammals. *Nature*, 508(7497):488–493, April 2014.

[28] Pau Creixell, Jüri Reimand, Syed Haider, Guanming Wu, Tatsuhiro Shibata, Miguel Vazquez, Ville Mustonen, Abel Gonzalez-Perez, John Pearson, Chris Sander, Benjamin J Raphael, Debora S Marks, B F Francis Ouellette, Alfonso Valencia, Gary D Bader, Paul C Boutros, Joshua M Stuart, Rune Linding, Nuria Lopez-Bigas, and Lincoln D Stein. Pathway and network analysis of cancer genomes. *Nature methods*, 12(7):615–21, 2015.

[29] Anna Csiszar, Andrej Podlutsky, Natalia Podlutskaya, William E. Sonntag, Steven Z. Merlin, Eva E. R. Philipp, Kristian Doyle, Antonio Davila, Fabio A. Recchia, Praveen Ballabh, John T. Pinto, and Zoltan Ungvari. Testing the Oxidative Stress Hypothesis of Aging in Primate Fibroblasts: Is There a Correlation Between Species Longevity and Cellular ROS Production? *The Journals of Gerontology: Series A*, 67:841–852, 2012.

[30] Sandra Daley, Deborah L. Wingard, and Vivian Reznik. Improving the retention of underrepresented minority faculty in academic medicine. *Journal of the National Medical Association*, 98(17019910):1435–1440, September 2006.

[31] Chi V Dang. A metabolic perspective of Peto's paradox and cancer. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 370(1673), July 2015.

[32] E Jedediah Dean, Jerel C Davis, Ronald W Davis, and Dmitri A Petrov. Pervasive and Persistent Redundancy among Duplicated Genes in Yeast. *PLoS Genetics*, 4(7):e1000113, 2008.

[33] Wayne Delport, Art F Y Poon, Simon D W Frost, and Sergei L Kosakovsky Pond. Datamonkey 2010: a suite of phylogenetic analysis tools for evolutionary biology. *Bioinformatics*, 26(19):2455–2457, October 2010.

[34] Jane M Dobson. Breed-predispositions to cancer in pedigree dogs. *ISRN veterinary science*, 2013:941275, 2013.

[35] Aoife Doherty and João de Magalhães. Has gene duplication impacted the evolution of Eutherian longevity? *Aging Cell*, 15(5):978–980, 2016.

[36] R Doll. The age distribution of cancer: implications for models of carcinogenesis. *Journal of the Royal Statistical Society Series A (...*, 1971.

[37] Dong Dong, Ming Lei, Panyu Hua, Yi-Hsuan Pan, Shuo Mu, Guantao Zheng, Erli Pang, Kui Lin, and Shuyi Zhang. The genomes of two bat species with long constant frequency echolocation calls. *Molecular Biology and Evolution*, 34(1):20–34, 2016.

[38] C. R. Dorn, D. O. N. Taylor, R. Schneider, H. H. Hibbard, and M. R. Klauber. Survey of Animal Neoplasms in Alameda and Contra Costa Counties, California. II. Cancer Morbidity in Dogs and Cats From Alameda County¡xref ref-type="fn" rid="FN2"¿2¡/xref¿. *JNCI: Journal of the National Cancer Institute*, 40(2):307–318, 1968.

[39] Olga Dudchenko, Sanjit S Batra, Arina D Omer, Sarah K Nyquist, Marie Hoeger, Neva C Durand, Muhammad S Shamim, Ido Machol, Eric S Lander, Aviva Presser Aiden, and Erez Lieberman Aiden. De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. *Science*, 356(6333):92–95, 2017.

[40] Walter L Eckalbar, Stephen A Schlebusch, Mandy K Mason, Zoe Gill, Ash V Parker, Betty M Booker, Sierra Nishizaki, Christiane Muswamba-Nday, Elizabeth Terhune, Kimberly A Nevonen, Nadja Makki, Tara Friedrich, Julia E VanderMeer, Katherine S Pollard, Lucia Carbone, Jeff D Wall, Nicola Illing, and Nadav Ahituv. Transcriptomic and epigenomic characterization of the developing bat wing. *Nature Genetics*, 48(5):528–536, 2016.

[41] Marc Effron, Lynn Griner, and Kurt Benirschke. Nature and Rate of Neoplasia Found in Captive Wild Mammals, Birds, and Reptiles at Necropsy. *JNCI: Journal of the National Cancer Institute*, 59(1):185–198, 1977.

[42] Michael G Elliot and Arne Ø Mooers. Inferring ancestral states without assuming neutrality or gradualism using a stable model of continuous character evolution. *BMC evolutionary biology*, 14(1):226, 2014.

[43] Xiaodong Fang, Eviatar Nevo, Lijuan Han, Erez Y Levanon, Jing Zhao, Aaron Avivi, Denis Larkin, Xuanting Jiang, Sergey Feranchuk, Yabing Zhu, Alla Fishman, Yue Feng, Noa Sher, Zhiqiang Xiong, Thomas Hankeln, Zhiyong Huang, Vera Gorbunova, Lu Zhang, Wei Zhao, Derek E Wildman, Yingqi Xiong, Andrei Gudkov, Qiumei Zheng, Gideon Rechavi, Sanyang Liu, Lily Bazak, Jie Chen, Binyamin A Knisbacher, Yao Lu, Imad Shams, Krzysztof Gajda, Marta Farré, Jaebum Kim, Harris A Lewin, Jian Ma, Mark Band, Anne Bicker, Angela Kranz, Tobias Mattheus, Hanno Schmidt, Andrei

Seluanov, Jorge Azpurua, Michael R McGowen, Eshel Ben Jacob, Kexin Li, Shaoliang Peng, Xiaoqian Zhu, Xiangke Liao, Shuaicheng Li, Anders Krogh, Xin Zhou, Leonid Brodsky, and Jun Wang. Genome-wide adaptive complexes to underground stresses in blind mole rats Spalax. *Nature Communications*, 5(1):3966, 2014.

[44] Xiaodong Fang, Inge Seim, Zhiyong Huang, Maxim V Gerashchenko, Zhiqiang Xiong, Anton A Turanov, Yabing Zhu, Alexei V Lobanov, Dingding Fan, Sun Hee Yim, Xiaoming Yao, Siming Ma, Lan Yang, Sang-Goo Lee, Eun Bae Kim, Roderick T Bronson, Radim ?umbera, Rochelle Buffenstein, Xin Zhou, Anders Krogh, Thomas J Park, Guojie Zhang, Jun Wang, and Vadim N Gladyshev. Adaptations to a Subterranean Environment and Longevity Revealed by the Analysis of Mole Rat Genomes. 8(5), 1900.

[45] Martin E. Feder and Gretchen E. Hofmann. HEAT-SHOCK PROTEINS, MOLECULAR CHAPERONES, AND THE STRESS RESPONSE: Evolutionary and Ecological Physiology. *Annual Review of Physiology*, 61:243–282, 1999.

[46] Joseph Felsenstein. Phylogenies and the Comparative Method. *The American Naturalist*, 125(1):1–15, 1985.

[47] Andrew D. Foote, Yue Liu, Gregg W. C. Thomas, Tomáš Vinař, Jessica Alföldi, Jixin Deng, Shannon Dugan, Cornelis E. van Elk, Margaret E. Hunter, Vandita Joshi, Ziad Khan, Christie Kovar, Sandra L. Lee, Kerstin Lindblad-Toh, Annalaura Mancia, Rasmus Nielsen, Xiang Qin, Jiaxin Qu, Brian J. Raney, Nagarjun Vijay, Jochen B. W. Wolf, Matthew W. Hahn, Donna M. Muzny, Kim C. Worley, M. Thomas P. Gilbert, and Richard A. Gibbs. Convergent evolution of the genomes of marine mammals. *Nature Genetics*, 47(3):272–275, 2015.

[48] A Force, M Lynch, F B Pickett, A Amores, Y L Yan, and J Postlethwait. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics*, 151(4):1531–45, 1999.

[49] Santo Fortunato, Carl T Bergstrom, Katy Börner, James A Evans, Dirk Helbing, Staša Milojević, Alexander M Petersen, Filippo Radicchi, Roberta Sinatra, Brian Uzzi, et al. Science of science. *Science*, 359(6379):eaao0185, 2018.

[50] Alexey A. Fushan, Anton A. Turanov, Sang?Goo Lee, Eun Bae Kim, Alexei V. Lobanov, Sun Hee Yim, Rochelle Buffenstein, Sang?Rae Lee, Kyu?Tae Chang, Hwanseok Rhee, Jong?So Kim, Kap?Seok Yang, and Vadim N. Gladyshev. Gene expression defines natural changes in mammalian lifespan. *Aging Cell*, 14(3):352–365, 2015.

[51] Yann Gager, Olivier Gimenez, M. Teague O'Mara, and Dina K. N. Dechmann. Group size, survival and surprisingly short lifespan in socially foraging bats. *BMC Ecology*, 16:2, 2016.

[52] Isabel García-Cao, Marta García-Cao, Juan Martín-Caballero, Luis M. Criado, Peter Klatt, Juana M. Flores, Jean Claude Weill, María A. Blasco, and Manuel Serrano.

'Super p53' mice exhibit enhanced DNA damage response, are tumor resistant and age normally. *The EMBO Journal*, 21(22):6225–6235, 2002.

[53] M Goodman, K N Sterner, M Islam, M Uddin, C C Sherwood, P R Hof, Z C Hou, L Lipovich, H Jia, L I Grossman, and D E Wildman. Phylogenomic analyses reveal convergent patterns of adaptive evolution in elephant and human ancestries. *Proceedings of the National Academy of Sciences*, 106(49):20824–20829, 2009.

[54] V Gorbunova, A Seluanov, ZD Zhang, VN Gladyshev, and J Vijg. Comparative genetics of longevity and cancer: insights from long-lived rodents. 15(8).

[55] Vera Gorbunova, Christopher Hine, Xiao Tian, Julia Ablaeva, Andrei V Gudkov, Eviatar Nevo, and Andrei Seluanov. Cancer resistance in the blind mole rat is mediated by concerted necrotic cell death mechanism. *Proceedings of the National Academy of Sciences of the United States of America*, 109(47):19392–6, 2012.

[56] Jane Green, Benjamin J Cairns, Delphine Casabonne, F Lucy Wright, Gillian Reeves, Valerie Beral, and for the Million Women Study collaborators. 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(8):785–794, 2011.

[57] James P. Guevara, Emem Adanga, Elorm Avakame, and Margo Brooks Carthon. Minority Faculty Development Programs and Underrepresented Minority Faculty Representation at US Medical Schools. *JAMA*, 310(21):2297–2304, 12 2013.

[58] Yan Guo, Yulin Dai, Hui Yu, Shilin Zhao, David C. Samuels, and Yu Shyr. Improvements and impacts of GRCh38 human reference on high throughput sequencing data analysis. *Genomics*, 109, 2017.

[59] B P Haines, R B Voyle, T A Pelton, R Forrest, and P D Rathjen. Complex conserved organization of the mammalian leukemia inhibitory factor gene: regulated expression of intracellular and extracellular cytokines. *Journal of immunology (Baltimore, Md. : 1950)*, 162(8):4637–4646, April 1999.

[60] B P Haines, R B Voyle, and P D Rathjen. Intracellular and extracellular leukemia inhibitory factor proteins have different cellular activities that are mediated by distinct protein motifs. *Molecular biology of the cell*, 11(4):1369–1383, April 2000.

[61] Douglas Hanahan and Robert A. Weinberg. The hallmarks of cancer. *Cell*, 100(1):57–70, January 2000.

[62] Douglas Hanahan and Robert A Weinberg. Hallmarks of Cancer: The Next Generation. 144(5).

[63] James M. Harper, Adam B. Salmon, Scott F. Leiser, Andrzej T. Galecki, and Richard A. Miller. Skin-derived fibroblasts from long-lived species are resistant to some, but not

all, lethal stresses and to the mitochondrial inhibitor rotenone. *Aging Cell*, 6(1):1–13, 2007.

[64] Lawrence R Heaney and Nina R Ingle. A key to the bats of the Philippine Islands / Nina R. Ingle, Lawrence R. Heaney. *Publication (USA)*, 1992.

[65] Maureen E Higgins, Martine Claremont, John E Major, Chris Sander, and Alex E Lash. CancerGenes: a gene selection resource for cancer genome projects. *Nucleic Acids Research*, 35(Database issue):D721–6, January 2007.

[66] JE Hill and SE Smith. Craseonycteris thonglongyai. *Mammalian species*, 1981.

[67] Toru Hisaka, Alexis Desmoulière, Jean-Luc Taupin, Sophie Daburon, Véronique Neaud, Nathalie Senant, Jean-Frédéric Blanc, Jean-François Moreau, and Jean Rosenbaum. Expression of leukemia inhibitory factor (LIF) and its receptor gp190 in human liver and in cultured human liver myofibroblasts. Cloning of new isoforms of LIF mRNA. *Comparative hepatology*, 3(1):10, November 2004.

[68] Diep Thi Hoang, Olga Chernomor, Arndt von Haeseler, Bui Quang Minh, and Le Sy Vinh. UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Molecular Biology and Evolution*, 35(2):518–522, 2017.

[69] Raymond J. Hock. The Metabolic Rates and Body Temperatures of Bats. *Biological Bulletin*, 101(3):289–299, 1951.

[70] Bas Hofstra, Vivek V. Kulkarni, Sebastian Munoz-Najar Galvez, Bryan He, Dan Jurafsky, and Daniel A. McFarland. The diversity–innovation paradox in science. *Proceedings of the National Academy of Sciences*, 117(17):9284–9291, 2020.

[71] D HONE and M BENTON. The evolution of large size: how does Cope's Rule work? *Trends in Ecology & Evolution*, 20(1):4–6, 2005.

[72] Jorrit J Hornberg, Frank J Bruggeman, Hans V Westerhoff, and Jan Lankelma. Cancer: A Systems Biology disease. *Biosystems*, 83(2-3):81–90, 2006.

[73] Wenwei Hu, Zhaohui Feng, Angelika K Teresky, and Arnold J Levine. p53 regulates maternal reproduction through LIF. *Nature*, 450(7170):721–724, November 2007.

[74] Zixia Huang, David Jebb, and Emma C Teeling. Blood miRNomes and transcriptomes reveal novel longevity mechanisms in the long-lived bat, Myotis myotis. *BMC Genomics*, 17(1):906, 2016.

[75] K R Hudson, A B Vernallis, and J K Heath. Characterization of the receptor binding sites of human leukemia inhibitory factor and creation of antagonists. *Journal of Biological Chemistry*, 271(20):11971–11978, May 1996.

[76] G Hunt and K Roy. Climate change, body size evolution, and Cope's Rule in deep-sea ostracodes. *Proceedings of the National Academy of Sciences*, 103(5):1347–1352, 2006.

[77] Trevor Huyton, Jian-Guo Zhang, Cindy S Luo, Mei-Zhen Lou, Douglas J Hilton, Nicos A Nicola, and Thomas P J Garrett. An unusual cytokine:Ig-domain interaction revealed in the crystal structure of leukemia inhibitory factor (LIF) in complex with the LIF receptor. *Proceedings of the National Academy of Sciences of the United States of America*, 104(31):12737–12742, July 2007.

[78] Jaime Iranzo, Iñigo Martincorena, and Eugene V Koonin. Cancer-mutation network and the number and specificity of driver mutations. *Proceedings of the National Academy of Sciences*, 115(26):E6010–E6019, 2018.

[79] David Jablonski. Body-size evolution in Cretaceous molluscs and the status of Cope's rule. *Nature*, 385(6613):250–252, 1997.

[80] Bijay Jassal, Lisa Matthews, Guilherme Viteri, Chuqiao Gong, Pascual Lorente, Antonio Fabregat, Konstantinos Sidiropoulos, Justin Cook, Marc Gillespie, Robin Haw, Fred Loney, Bruce May, Marija Milacic, Karen Rothfels, Cristoffer Sevilla, Veronica Shamovsky, Solomon Shorser, Thawfeek Varusai, Joel Weiser, Guanming Wu, Lincoln Stein, Henning Hermjakob, and Peter D'Eustachio. The reactome pathway knowledgebase. *Nucleic acids research*, 48(D1):D498–D503, 2020.

[81] Geoffrey C Kabat, Mimi Y Kim, Albert R Hollenbeck, and Thomas E Rohan. Attained height, sex, and risk of cancer at different anatomic sites in the NIH-AARP diet and health study. *Cancer causes & control : CCC*, 25(12):1697–706, 2014.

[82] Subha Kalyaanamoorthy, Bui Quang Minh, Thomas K F Wong, Arndt von Haeseler, and Lars S Jermiin. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature Methods*, 14(6):587–589, 2017.

[83] Minoru Kanehisa and Susumu Goto. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, 28(1):27–30, 2000.

[84] Pankaj Kapahi, Michael E Boulton, and Thomas B.L Kirkwood. Positive correlation between mammalian life span and cellular resistance to stress. *Free Radical Biology and Medicine*, 26(5-6):495–500, 1999.

[85] Jason Karch, Onur Kanisicak, Matthew J Brody, Michelle A Sargent, Demetria M Michael, and Jeffery D Molkentin. Necroptosis Interfaces with MOMP and the MPTP in Mediating Cell Death. *PloS one*, 10(6):e0130520, 2015.

[86] Jason Karch, Jennifer Q Kwong, Adam R Burr, Michelle A Sargent, John W Elrod, Pablo M Peixoto, Sonia Martinez-Caballero, Hanna Osinska, Emily H-Y Cheng, Jeffrey Robbins, Kathleen W Kinnally, and Jeffery D Molkentin. Bax and Bak function as the outer membrane component of the mitochondrial permeability pore in regulating necrotic cell death in mice. *eLife*, 2:e00772, August 2013.

[87] Kazutaka Katoh, Kei-ichi Kuma, Hiroyuki Toh, and Takashi Miyata. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Research*, 33(2):511–518, 2005.

[88] Aris Katzourakis, Gkikas Magiorkinis, Aaron G Lim, Sunetra Gupta, Robert Belshaw, and Robert Gifford. Larger Mammalian Body Size Leads to Lower Retroviral Activity. *PLoS Pathogens*, 10(7):e1004214, 2014.

[89] Michael Keane, Jeremy Semeiks, Andrew E Webb, Yang I Li, Víctor Quesada, Thomas Craig, Lone Madsen, Sipko van Dam, David Brawand, Patrícia I Marques, Pawel Michalak, Lin Kang, Jong Bhak, Hyung-Soon Yim, Nick V Grishin, Nynne Nielsen, Mads Heide-Jørgensen, Elias M Oziolor, Cole W Matson, George M Church, Gary W Stuart, John C Patton, Craig J George, Robert Suydam, Knud Larsen, Carlos López-Otín, Mary J O?Connell, John W Bickham, Bo Thomsen, and João de Magalhães. Insights into the Evolution of Longevity from the Bowhead Whale Genome. *Cell Reports*, 10(1):112–122, 2015.

[90] James W Kent. Blat: The blast-like alignment tool. *Genome Research*, 12(4):656–664, 2002.

[91] Aziz Khan, Oriol Fornes, Arnaud Stigliani, Marius Gheorghe, Jaime A. Castro-Mondragon, Robin van der Lee, Adrien Bessy, Jeanne Chèneby, Shubhada R. Kulkarni, Ge Tan, Damir Baranasic, David J. Arenillas, Albin Sandelin, Klaas Vandepoele, Boris Lenhard, Benoît Ballester, Wyeth W. Wasserman, François Parcy, and Anthony Mathelier. JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Research*, pages gkx1126–, 2018.

[92] Daehwan Kim, Ben Langmead, and Steven L Salzberg. HISAT: a fast spliced aligner with low memory requirements. *Nature Methods*, 12(4):357–360, 2015.

[93] Eun Bae Kim, Xiaodong Fang, Alexey A Fushan, Zhiyong Huang, Alexei V Lobanov, Lijuan Han, Stefano M Marino, Xiaoqing Sun, Anton A Turanov, Pengcheng Yang, Sun Hee Yim, Xiang Zhao, Marina V Kasaikina, Nina Stoletzki, Chunfang Peng, Paz Polak, Zhiqiang Xiong, Adam Kiezun, Yabing Zhu, Yuanxin Chen, Gregory V Kryukov, Qiang Zhang, Leonid Peshkin, Lan Yang, Roderick T Bronson, Rochelle Buffenstein, Bo Wang, Changlei Han, Qiye Li, Li Chen, Wei Zhao, Shamil R Sunyaev, Thomas J Park, Guojie Zhang, Jun Wang, and Vadim N Gladyshev. Genome sequencing reveals insights into physiology and longevity of the naked mole rat. *Nature*, 479(7372):223–7, 2011.

[94] Amanda Kowalczyk, Raghavendran Partha, Nathan L Clark, and Maria Chikina. Pan-mammalian analysis of molecular constraints underlying extended lifespan. *eLife*, 9:e51089, 2020.

[95] Tonisha B. Lane. Beyond academic and social integration: Understanding the impact of a stem enrichment program on the retention and degree attainment of underrepresented students. *CBE—Life Sciences Education*, 15(3):ar39, 2016. PMID: 27543638.

[96] Asier Larramendi. Shoulder Height, Body Mass, and Shape of Proboscideans. *Acta Palaeontologica Polonica*, 61(3), 2015.

[97] Béatrice Lauby-Secretan, Chiara Scoccianti, Dana Loomis, Yann Grosse, Franca Bianchini, Kurt Straif, and International Agency for Research on Cancer Handbook Working Group. Body Fatness and Cancer: Viewpoint of the IARC Working Group. *New England Journal of Medicine*, 375(8):794–798, 2016.

[98] Virpi Launonen. Mutations in the human LKB1/STK11 gene. *Human Mutation*, 26(4):291–297, 2005.

[99] Armand M Leroi, Vassiliki Koufopanou, and Austin Burt. Cancer selection. *Nature Reviews Cancer*, 3(3):226–231, 2003.

[100] Kaitlyn N. Lewis, Ilya Soifer, Eugene Melamud, Margaret Roy, R. Scott McIsaac, Matthew Hibbs, and Rochelle Buffenstein. Unraveling the message: insights into comparative genomics of the naked mole-rat. *Mammalian Genome*, 27(7-8):259–278, 2016.

[101] Jiyong Liang and Gordon B Mills. AMPK: a contextual oncogene or tumor suppressor? *Cancer research*, 73(10):2929–35, 2013.

[102] Yuxing Liao, Jing Wang, Eric J Jaehnig, Zhiao Shi, and Bing Zhang. WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. *Nucleic Acids Research*, 47(W1):W199–W205, 2019.

[103] Kerstin Lindblad-Toh, Manuel Garber, Or Zuk, Michael F Lin, Brian J Parker, Stefan Washietl, Pouya Kheradpour, Jason Ernst, Gregory Jordan, Evan Mauceli, Lucas D Ward, Craig B Lowe, Alisha K Holloway, Michele Clamp, Sante Gnerre, Jessica Alföldi, Kathryn Beal, Jean Chang, Hiram Clawson, James Cuff, Federica Di Palma, Stephen Fitzgerald, Paul Flicek, Mitchell Guttman, Melissa J Hubisz, David B Jaffe, Irwin Jungreis, W James Kent, Dennis Kostka, Marcia Lara, Andre L Martins, Tim Massingham, Ida Moltke, Brian J Raney, Matthew D Rasmussen, Jim Robinson, Alexander Stark, Albert J Vilella, Jiayu Wen, Xiaohui Xie, Michael C Zody, Team, Broad Institute Sequencing Platform and Whole Genome Assembly, Jen Baldwin, Toby Bloom, Chee Whye Chin, Dave Heiman, Robert Nicol, Chad Nusbaum, Sarah Young, Jane Wilkinson, Kim C Worley, Christie L Kovar, Donna M Muzny, Richard A Gibbs, Baylor College of Medicine Human Genome Sequencing Center Sequencing Team, Andrew Cree, Huyen H Dihn, Gerald Fowler, Shalili Jhangiani, Vandita Joshi, Sandra Lee, Lora R Lewis, Lynne V Nazareth, Geoffrey Okwuonu, Jireh Santibanez, Wesley C Warren, Elaine R Mardis, George M Weinstock, Richard K Wilson, Genome Institute at Washington University, Kim Delehaunty, David Dooling, Catrina Fronik, Lucinda Fulton, Bob Fulton, Tina Graves, Patrick Minx, Erica Sodergren, Ewan Birney, Elliott H Margulies, Javier Herrero, Eric D Green, David Haussler, Adam Siepel, Nick Goldman, Katherine S Pollard, Jakob S Pedersen, Eric S Lander, and Manolis Kellis. A high-resolution map of human evolutionary constraint using 29 mammals. *Nature*, 478(7370):476–482, 2011.

[104] R B Lucena, D R Rissi, G D Kommers, F Pierezan, J C Oliveira-Filho, J T S A Macêdo, M M Flores, and C S L Barros. A Retrospective Study of 586 Tumours in Brazilian Cattle. *Journal of Comparative Pathology*, 145(1):20–24, 2011.

[105] J D Ly, D R Grubb, and A Lawen. The mitochondrial membrane potential (deltapsi(m)) in apoptosis; an update. *Apoptosis : an international journal on programmed cell death*, 8(2):115–128, March 2003.

[106] Michael Lynch and John S. Conery. The Evolutionary Fate and Consequences of Duplicate Genes. *Science*, 290(5494):1151–1155, 2000.

[107] Carlos López-Otín, Maria A. Blasco, Linda Partridge, Manuel Serrano, and Guido Kroemer. The Hallmarks of Aging. *Cell*, 153(6):1194–1217, 2013.

[108] Sebastian Maciak and Pawel Michalak. Cell size and cancer: a new solution to Peto's paradox? *Evolutionary applications*, 8(1):2–8, January 2015.

[109] Sheila L MacRae, Matthew McKnight Croken, R B Calder, Alexander Aliper, Brandon Milholland, Ryan R White, Alexander Zhavoronkov, Vadim N Gladyshev, Andrei Seluanov, Vera Gorbunova, Zhengdong D Zhang, and Jan Vijg. DNA repair in species with extreme lifespan differences. *Aging*, 7(12):1171–1182, 2015.

[110] Giulia Malaguti, Param Priya Singh, and Hervé Isambert. On the retention of gene duplicates prone to dominant deleterious mutations. *Theoretical Population Biology*, 93:38–51, 2014.

[111] Pedro A. Malavet. Puerto rico: Cultural nation, american colony. *Mich. J. Race & L.*, 6(1):1–106, 2000.

[112] Stavros C Manolagas. Birth and Death of Bone Cells: Basic Regulatory Mechanisms and Implications for the Pathogenesis and Treatment of Osteoporosis*. *Endocrine Reviews*, 21(2):115–137, 2000.

[113] Emilia P Martins and Thomas F Hansen. Phylogenies and the Comparative Method: A General Approach to Incorporating Phylogenetic Information into the Analysis of Interspecific Data. *The American Naturalist*, 149(4):646–667, 1997.

[114] Alexander Y Maslov and Jan Vijg. Genome instability, cancer and aging. *Biochimica et biophysica acta*, 1790(10):963–9, 2009.

[115] A Mathelier, O Fornes, D J Arenillas, and C Chen. JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic acids ...*, 2015.

[116] M. Lisandra Zepeda Mendoza, Zijun Xiong, Marina Escalera-Zamudio, Anne Kathrine Runge, Julien Thézé, Daniel Streicker, Hannah K. Frank, Elizabeth Loza-Rubio, Shengmao Liu, Oliver A. Ryder, Jose Alfredo Samaniego Castruita, Aris Katzourakis,

George Pacheco, Blanca Taboada, Ulrike Löber, Oliver G. Pybus, Yang Li, Edith Rojas-Anaya, Kristine Bohmann, Aldo Carmona Baez, Carlos F. Arias, Shiping Liu, Alex D. Greenwood, Mads F. Bertelsen, Nicole E. White, Michael Bunce, Guojie Zhang, Thomas Sicheritz-Pontén, and M. P. Thomas Gilbert. Hologenomic adaptations underlying the evolution of sanguivory in the common vampire bat. *Nature Ecology & Evolution*, 2(4):659–668, 2018.

[117] Alea A Mills. p63: oncogene or tumor suppressor? *Current Opinion in Genetics & Development*, 16(1):38–44, 2006.

[118] Bui Quang Minh, Heiko A Schmidt, Olga Chernomor, Dominik Schrempf, Michael D Woodhams, Arndt von Haeseler, and Robert Lanfear. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution*, 37(5):1530–1534, 2020.

[119] Amirhossein Misaghi, Amanda Goldin, Moayd Awad, and Anna A Kulidjian. Osteosarcoma: a comprehensive review. *SICOT-J*, 4:12, 2018.

[120] José Trías Monge. *Puerto Rico: The trials of the oldest colony in the world*. Yale University Press, 1999.

[121] AP Møller, J Erritzøe, and JJ Soler. Life history, immunity, Peto's paradox and tumours in birds. *Journal of Evolutionary Biology*, 2017.

[122] John D. Nagy, Erin M. Victor, and Jenese H. Cropper. Why don't all whales have cancer? A novel hypothesis resolving Peto's paradox. *Integrative and Comparative Biology*, 47(2):317–328, 2007.

[123] Osamu Nishimura, Yuichiro Hara, and Shigehiro Kuraku. gVolante for standardizing completeness assessment of genome and transcriptome assemblies. *Bioinformatics*, 33(22):3635–3637, 2017.

[124] Martin A Nowak, Maarten C Boerlijst, Jonathan Cooke, and John Maynard Smith. Evolution of genetic redundancy. *Nature*, 388(6638):167–171, 1997.

[125] Leonard Nunney. Lineage selection and the evolution of multistage carcinogenesis. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 266(1418):493–498, 1999.

[126] Leonard Nunney. Commentary: The multistage model of carcinogenesis, Peto's paradox and evolution. *International Journal of Epidemiology*, 45:649–653, 2016.

[127] Leonard Nunney. Size matters: height, cell number and a person's risk of cancer. *Proc. R. Soc. B*, 285(1889):20181743, 2018.

[128] June E. Olds, Eric R. Burrough, Amanda J. Fales-Williams, Aaron Lehmkuhl, Darin Madson, Abby J. Patterson, and Michael J. Yaeger. RETROSPECTIVE EVALUATION

OF CASES OF NEOPLASIA IN A CAPTIVE POPULATION OF EGYPTIAN FRUIT BATS (ROUSETTUS AEGYPTIACUS). *Journal of Zoo and Wildlife Medicine*, 46:325–332, 2015.

[129] Maureen A O'Leary, Jonathan I Bloch, John J Flynn, Timothy J Gaudin, Andres Giallombardo, Norberto P Giannini, Suzann L Goldberg, Brian P Kraatz, Zhe-Xi Luo, Jin Meng, Xijun Ni, Michael J Novacek, Fernando A Perini, Zachary S Randall, Guillermo W Rougier, Eric J Sargis, Mary T Silcox, Nancy B Simmons, Michelle Spaulding, Paúl M Velazco, Marcelo Weksler, John R Wible, and Andrea L Cirranello. The placental mammal ancestor and the post-K-Pg radiation of placentals. *Science (New York, N.Y.)*, 339(6120):662–7, 2013.

[130] Maureen A O'Leary, Jonathan I Bloch, John J Flynn, Timothy J Gaudin, Andres Giallombardo, Norberto P Giannini, Suzann L Goldberg, Brian P Kraatz, Zhe-Xi Luo, Jin Meng, Xijun Ni, Michael J Novacek, Fernando A Perini, Zachary Randall, Guillermo W Rougier, Eric J Sargis, Mary T Silcox, Nancy B Simmons, Michelle Spaulding, Paúl M Velazco, Marcelo Weksler, John R Wible, and Andrea L Cirranello. Response to comment on "The placental mammal ancestor and the post-K-Pg radiation of placentals". *Science (New York, N.Y.)*, 341(6146):613, 2013.

[131] Camila A Orellana, Esteban Marcellin, Robin W Palfreyman, Trent P Munro, Peter P Gray, and Lars K Nielsen. RNA-Seq Highlights High Clonal Variation in Monoclonal Antibody Producing CHO Cells. *Biotechnology journal*, 13(3):e1700231, March 2018.

[132] Scott E Page. *The diversity bonus: How great teams pay off in the knowledge economy*. Princeton University Press, 2019.

[133] Eleftheria Palkopoulou, Mark Lipson, Swapan Mallick, Svend Nielsen, Nadin Rohland, Sina Baleka, Emil Karpinski, Atma M. Ivancevic, Thu-Hien To, R. Daniel Kortschak, Joy M. Raison, Zhipeng Qu, Tat-Jun Chin, Kurt W. Alt, Stefan Claesson, Love Dalén, Ross D. E. MacPhee, Harald Meller, Alfred L. Roca, Oliver A. Ryder, David Heiman, Sarah Young, Matthew Breen, Christina Williams, Bronwen L. Aken, Magali Ruffier, Elinor Karlsson, Jeremy Johnson, Federica Di Palma, Jessica Alfoldi, David L. Adelson, Thomas Mailund, Kasper Munch, Kerstin Lindblad-Toh, Michael Hofreiter, Hendrik Poinar, and David Reich. A comprehensive genomic history of extinct and living elephants. *Proceedings of the National Academy of Sciences*, 115(11):E2566–E2574, 2018.

[134] Eleftheria Palkopoulou, Swapan Mallick, Pontus Skoglund, Jacob Enk, Nadin Rohland, Heng Li, Ayça Omrak, Sergey Vartanyan, Hendrik Poinar, Anders Götherström, David Reich, and Love Dalén. Complete genomes reveal signatures of demographic and genetic declines in the woolly mammoth. *Current biology : CB*, 25(10):1395–1400, May 2015.

[135] Joe Parker, Georgia Tsagkogeorga, James A. Cotton, Yuan Liu, Paolo Provero, Elia Stupka, and Stephen J. Rossiter. Genome-wide signatures of convergent evolution in echolocating mammals. *Nature*, 502(7470):228–231, 2013.

[136] Genis Parra, Keith Bradnam, Zemin Ning, Thomas Keane, and Ian Korf. Assessing the gene space in draft genomes. *Nucleic Acids Research*, 37(1):289–297, 2008.

[137] Stephanie S. Pavlovich, Sean P. Lovett, Galina Koroleva, Jonathan C. Guito, Catherine E. Arnold, Elyse R. Nagle, Kirsten Kulcsar, Albert Lee, Françoise Thibaud-Nissen, Adam J. Hume, Elke Mühlberger, Luke S. Uebelhoer, Jonathan S. Towner, Raul Rabadan, Mariano Sanchez-Lockhart, Thomas B. Kepler, and Gustavo Palacios. The Egyptian Rousette Genome Reveals Unexpected Features of Bat Antiviral Immunity. *Cell*, 173:1098–1110.e18, 2018.

[138] Mihaela Pertea, Daehwan Kim, Geo M Pertea, Jeffrey T Leek, and Steven L Salzberg. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nature Protocols*, 11(9):1650–1667, 2016.

[139] Mihaela Pertea, Geo M Pertea, Corina M Antonescu, Tsung-Cheng Chang, Joshua T Mendell, and Steven L Salzberg. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology*, 33(3):290–295, 2015.

[140] R Peto, FJ Roe, PN Lee, L Levy, and J Clack. Cancer and ageing in mice and men. *British Journal of Cancer*, 32(4):411–426, 1975.

[141] Richard Peto. Quantitative implications of the approximate irrelevance of mammalian body size and lifespan to lifelong cancer risk. *Phil. Trans. R. Soc. B*, 370(1673):20150198, 2015.

[142] I. O. Petruseva, A. N. Evdokimov, and O. I. Lavrik. Genome stability maintenance in naked mole-rat. *Acta naturae*, 9:31–41, Oct-Dec 2017.

[143] Andrej J. Podlutsky, Alexander M. Khritankov, Nikolai D. Ovodov, and Steven N. Austad. A New Field Record for Bat Longevity. *The Journals of Gerontology: Series A*, 60:1366–1368, 2005.

[144] Harrison Pride, Zhen Yu, Bharath Sunchu, Jillian Mochnick, Alexander Coles, Yiqiang Zhang, Rochelle Buffenstein, Peter J. Hornsby, Steven N. Austad, and Viviana I. Pérez. Long-lived species have improved proteostasis compared to phylogenetically-related shorter-lived species. *Biochemical and Biophysical Research Communications*, 457:669–675, 2015.

[145] Mark N Puttick and Gavin H Thomas. Fossils and living taxa agree on patterns of body mass evolution: a case study with Afrotheria. *Proceedings. Biological sciences / The Royal Society*, 282(1821):20152023, 2015.

[146] Wenfeng Qian and Jianzhi Zhang. Genomic evidence for adaptation by gene duplication. *Genome research*, 24(8):1356–62, 2014.

[147] Shruti Rastogi and David A Liberles. Subfunctionalization of duplicated genes as a transition state to neofunctionalization. *BMC Evolutionary Biology*, 5(1):28, 2005.

[148] P D Rathjen, S Toth, A Willis, J K Heath, and A G Smith. Differentiation inhibiting activity is produced in matrix-associated and diffusible forms that are generated by alternate promoter usage. *Cell*, 62(6):1105–1114, September 1990.

[149] Puli Chandramouli Reddy, Ishani Sinha, Ashwin Kelkar, Farhat Habib, Saurabh J Pradhan, Raman Sukumar, and Sanjeev Galande. Comparative sequence analyses of genome and transcriptome reveal novel transcripts and variants in the Asian elephant Elephas maximus. *Journal of biosciences*, 40(5):891–907, December 2015.

[150] Nadin Rohland, David Reich, Swapan Mallick, Matthias Meyer, Richard E Green, Nicholas J Georgiadis, Alfred L Roca, and Michael Hofreiter. Genomic DNA sequences from mastodon and woolly mammoth reveal deep speciation of forest and savanna elephants. *PLoS Biology*, 8(12):e1000564, 2010.

[151] Leonidas Salichos and Antonis Rokas. Evaluating ortholog prediction algorithms in a yeast model clade. *PloS one*, 6(4):e18755, 2011.

[152] Adam B. Salmon, Shanique Leonard, Venkata Masamsetti, Anson Pierce, Andrej J. Podlutsky, Natalia Podlutskaya, Arlan Richardson, Steven N. Austad, and Asish R. Chaudhuri. The long lifespan of two bat species is correlated with resistance to protein oxidation and enhanced protein homeostasis. *The FASEB Journal*, 23:2317–2326, 2009.

[153] Albin Sandelin, Wyeth W. Wasserman, and Boris Lenhard. ConSite: web-based prediction of regulatory elements using cross-species comparison. *Nucleic Acids Research*, 32:W249–W252, 2004.

[154] Van M. Savage, Andrew P. Allen, James H. Brown, James F. Gillooly, Alexander B. Herman, William H. Woodruff, and Geoffrey B. West. Scaling of number, size, and metabolic rate of cells with body size in mammals. *Proceedings of the National Academy of Sciences*, 104(11):4718–4723, 2007.

[155] V B Scheffer. The Weight of the Steller Sea Cow. *Journal of Mammalogy*, 53(4):912–914, 1972.

[156] Dominik Schrempf, Bui Quang Minh, Arndt von Haeseler, and Carolin Kosiol. Polymorphism-Aware Species Trees with Advanced Mutation Models, Bootstrap, and Rate Heterogeneity. *Molecular Biology and Evolution*, 36(6):1294–1301, 2019.

[157] G T Schwartz, D T Rasmussen, and R J Smith. Body-Size Diversity and Community Structure of Fossil Hyracoids. *Journal of Mammalogy*, 76(4):1088–1099, 1995.

[158] Inge Seim, Xiaodong Fang, Zhiqiang Xiong, Alexey V. Lobanov, Zhiyong Huang, Siming Ma, Yue Feng, Anton A. Turanov, Yabing Zhu, Tobias L. Lenz, Maxim V. Gerashchenko, Dingding Fan, Sun Hee Yim, Xiaoming Yao, Daniel Jordan, Yingqi Xiong, Yong Ma, Andrey N. Lyapunov, Guanxing Chen, Oksana I. Kulakova, Yudong Sun, Sang-Goo Lee, Roderick T. Bronson, Alexey A. Moskalev, Shamil R. Sunyaev, Guojie Zhang, Anders Krogh, Jun Wang, and Vadim N. Gladyshev. Genome analysis reveals insights into

physiology and longevity of the Brandt's bat Myotis brandtii. *Nature Communications*, 4(1):2212, 2013.

[159] Itamar Sela, Haim Ashkenazy, Kazutaka Katoh, and Tal Pupko. GUIDANCE2: accurate detection of unreliable alignment regions accounting for the uncertainty of multiple parameters. *Nucleic Acids Research*, 43(W1):W7–14, July 2015.

[160] Andrei Seluanov, Christopher Hine, Michael Bozzella, Amelia Hall, Tais H C Sasahara, Antonio A C M Ribeiro, Kenneth C Catania, Daven C Presgraves, and Vera Gorbunova. Distinct tumor suppressor mechanisms evolve in rodent species that differ in size and lifespan. *Aging cell*, 7(6):813–23, 2008.

[161] Dorian A. Shaw. The status of puerto rico revisited: Does the current u.s.-puerto rico relationship uphold international law note. *Fordham Int'l L.J.*, 17(4):1006–1061, 1993.

[162] Robert Sitarz, Ma?gorzata Skierucha, Jerzy Mielko, Johan Offerhaus, Ryszard Maciejewski, and Wojciech Polkowski. Gastric cancer: epidemiology, prevention, classification, and treatment. *Cancer Management and Research*, Volume 10:239–248, 2018.

[163] Denise N Slenter, Martina Kutmon, Kristina Hanspers, Anders Riutta, Jacob Windsor, Nuno Nunes, Jonathan Mélius, Elisa Cirillo, Susan L Coort, Daniela Digles, Friederike Ehrhart, Pieter Giesbertz, Marianthi Kalafati, Marvin Martens, Ryan Miller, Kozo Nishida, Linda Rieswijk, Andra Waagmeester, Lars M T Eijssen, Chris T Evelo, Alexander R Pico, and Egon L Willighagen. WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. *Nucleic Acids Research*, 46(D1):D661–D667, 2017.

[164] Zbyslaw Sondka, Sally Bamford, Charlotte G Cole, Sari A Ward, Ian Dunham, and Simon A Forbes. The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. *Nature Reviews Cancer*, 18(11):696–705, 2018.

[165] Julien Soubrier, Mike Steel, Michael S Y Lee, Clio Der Sarkissian, Stéphane Guindon, Simon Y W Ho, and Alan Cooper. The Influence of Rate Heterogeneity among Sites on the Time Dependence of Molecular Rates. *Molecular Biology and Evolution*, 29(11):3345–3358, 2012.

[166] Mark S Springer, Robert W Meredith, Emma C Teeling, and William J Murphy. Technical comment on "The placental mammal ancestor and the post-K-Pg radiation of placentals". *Science (New York, N.Y.)*, 341(6146):613, 2013.

[167] Arlin Stoltzfus. On the Possibility of Constructive Neutral Evolution. *Journal of Molecular Evolution*, 49(2):169–181, 1999.

[168] Zuzana Storchova and David Pellman. From polyploidy to aneuploidy, genome instability and cancer. *Nature Reviews Molecular Cell Biology*, 5(1):45–54, 2004.

[169] Eugene H Studier and Michael J O'Farrell. Biology of Myotis thysanodes and M. lucifugus (Chiroptera: Vespertilionidae)—III. Metabolism, heart rate, breathing rate, evaporative water loss and general energetics. *Comparative Biochemistry and Physiology Part A: Physiology*, 54:423–432, 1976.

[170] Michael Sulak, Lindsey Fong, Katelyn Mika, Sravanthi Chigurupati, Lisa Yon, Nigel P Mongan, Richard D Emes, and Vincent J Lynch. TP53 copy number expansion is associated with the evolution of increased body size and an enhanced DNA damage response in elephants. *eLife*, 5:e11994, 2016.

[171] J Sung, Y M Song, D A Lawlor, G D Smith, and S Ebrahim. Height and Site-specific Cancer Risk: A Cohort Study of a Korean Adult Population. *American Journal of Epidemiology*, 170(1):53–64, 2009.

[172] Surveillance, Epidemiology, and End Results (SEER) Program. SEER*Stat Database: Incidence - SEER Research Data, 9 Registries, Nov 2019 Sub (1975-2017) - Linked To County Attributes - Time Dependent (1990-2017) Income/Rurality, 1969-2017 Counties, National Cancer Institute, DCCPS, Surveillance Research Program.

[173] Robi Tacutu, Thomas Craig, Arie Budovsky, Daniel Wuttke, Gilad Lehmann, Dmitri Taranukha, Joana Costa, Vadim E Fraifeld, and João de Magalhães. Human Ageing Genomic Resources: Integrated databases and tools for the biology and genetics of ageing. *Nucleic Acids Research*, 41(D1):D1027–D1033, 2013.

[174] Stephen W G Tait and Douglas R Green. Mitochondria and cell death: outer membrane permeabilization and beyond. *Nat. Rev. Mol. Cell Biol.*, 11(9):621–632, September 2010.

[175] Kazuhiro Takemoto, Masato Ii, and Satoshi S Nishizuka. Importance of metabolic rate to the relationship between the number of genes in a functional category and body size in Peto's paradox for cancer. *Royal Society open science*, 3(9):160267, September 2016.

[176] Xiao Tian, Jorge Azpurua, Christopher Hine, Amita Vaidya, Max Myakishev-Rempel, Julia Ablaeva, Zhiyong Mao, Eviatar Nevo, Vera Gorbunova, and Andrei Seluanov. High molecular weight hyaluronan mediates the cancer resistance of the naked mole-rat. 499(7458), 2013.

[177] Marc Tollis, Joshua D Schiffman, and Amy M Boddy. Evolution of cancer suppression as revealed by mammalian comparative genomics. *Current Opinion in Genetics & Development*, 42:40–47, 2017.

[178] Stuart D. Tyner, Sundaresan Venkatachalam, Jene Choi, Stephen Jones, Nader Ghebranious, Herbert Igelmann, Xiongbin Lu, Gabrielle Soron, Benjamin Cooper, Cory Brayton, Sang Hee Park, Timothy Thompson, Gerard Karsenty, Allan Bradley, and Lawrence A. Donehower. p53 mutant mice that display early ageing-associated phenotypes. *Nature*, 415(6867):45, 2002.

[179] Angelina V Vaseva, Natalie D Marchenko, Kyungmin Ji, Stella E Tsirka, Sonja Holzmann, and Ute M Moll. p53 opens the mitochondrial permeability transition pore to trigger necrosis. *Cell*, 149(7):1536–1548, June 2012.

[180] Juan Manuel Vazquez, Michael Sulak, Sravanthi Chigurupati, and Vincent J. Lynch. A Zombie LIF Gene in Elephants Is Upregulated by TP53 to Induce Apoptosis in Response to DNA Damage. *Cell Reports*, 24(7):1765–1776, 2018.

[181] R B Voyle, B P Haines, M F Pera, R Forrest, and P D Rathjen. Human germ cell tumor cell lines express novel leukemia inhibitory factor transcripts encoding differentially localized proteins. *Experimental cell research*, 249(2):199–211, June 1999.

[182] Duowei Wang, Xianjing Li, Jiani Li, Yuan Lu, Sen Zhao, Xinying Tang, Xin Chen, Jiaying Li, Yan Zheng, Shuran Li, Rui Sun, Ming Yan, Decai Yu, Guangwen Cao, and Yong Yang. APOBEC3B interaction with PRC2 modulates microenvironment to promote HCC progression. *Gut*, 68(10):1846–1857, 2019.

[183] Huai-Chun Wang, Bui Quang Minh, Edward Susko, and Andrew J Roger. Modeling Site Heterogeneity with Posterior Mean Site Frequency Profiles Accelerates Accurate Phylogenomic Estimation. *Systematic Biology*, 67(2):216–235, 2017.

[184] Joel O Wertheim, Ben Murrell, Martin D Smith, Sergei L Kosakovsky Pond, and Konrad Scheffler. RELAX: Detecting Relaxed Selection in a Phylogenetic Framework. *Molecular biology and evolution*, 32(3):820–832, March 2015.

[185] Potter Wickware. Along the leaky pipeline. *Nature*, 390(6656):202–203, 1997.

[186] Sara Wirén, Christel Häggström, Hanno Ulmer, Jonas Manjer, Tone Bjørge, Gabriele Nagel, Dorthe Johansen, Göran Hallmans, Anders Engeland, Hans Concin, Håkan Jonsson, Randi Selmer, Steinar Tretli, Tanja Stocks, and Pär Stattin. Pooled cohort study on height and risk of cancer and cancer death. *Cancer Causes & Control*, 25(2):151–159, 2014.

[187] Z Yang, S Kumar, and M Nei. A new method of inference of ancestral nucleotide and amino acid sequences. *Genetics*, 141(4):1641–50, 1995.

[188] Jian Ye, George Coulouris, Irena Zaretskaya, Ioana Cutcutache, Steve Rozen, and Thomas L Madden. Primer-BLAST: A tool to design target-specific primers for polymerase chain reaction. *BMC Bioinformatics*, 13:134, 2012.

[189] A H Yona, E J Alm, and J Gore. Random Sequences Rapidly Evolve Into De Novo Promoters. *bioRxiv*, 2017.

[190] Guojie Zhang, Christopher Cowled, Zhengli Shi, Zhiyong Huang, Kimberly A. Bishop-Lilly, Xiaodong Fang, James W. Wynne, Zhiqiang Xiong, Michelle L. Baker, Wei Zhao, Mary Tachedjian, Yabing Zhu, Peng Zhou, Xuanting Jiang, Justin Ng, Lan Yang, Lijun Wu, Jin Xiao, Yue Feng, Yuanxin Chen, Xiaoqing Sun, Yong Zhang, Glenn A. Marsh,

Gary Crameri, Christopher C. Broder, Kenneth G. Frey, Lin-Fa Wang, and Jun Wang. Comparative Analysis of Bat Genomes Provides Insight into the Evolution of Flight and Immunity. *Science*, 339(6118):456–460, 2012.

[191] Jianzhi Zhang. Advances in Experimental Medicine and Biology. *Advances in experimental medicine and biology*, 751:279–300, 2012.

[192] Úlfur Árnason, Fritjof Lammers, Vikas Kumar, Maria A. Nilsson, and Axel Janke. Whole-genome sequencing of the blue whale and other rorquals finds signatures for introgressive gene flow. *Science Advances*, 4(4):eaap9873, 2018.