Life History Affects Cancer Gene Copy Numbers in Mammalian Genomes

by

Aika Kunigunda Schneider-Utaka

A Thesis Presented in Partial Fulfillment
of the Requirements for the Degree
Master of Science

Approved October 2019 by the
Graduate Supervisory Committee:

Carlo Maley, Chair
Melissa Wilson
Marc Tollis

ARIZONA STATE UNIVERSITY

December 2019

ABSTRACT

Cancer is a disease which can affect all animals across the tree of life. Certain species have undergone natural selection to reduce or prevent cancer. Mechanisms to block cancer may include, among others, a species possessing additional paralogues of tumor suppressor genes, or decreasing the number of oncogenes within their genome. To understand cancer prevention patterns across species, I developed a bioinformatic pipeline to identify copies of 545 known tumor suppressor genes and oncogenes across 63 species of mammals. I used phylogenetic regressions to test for associations between cancer gene copy numbers and a species' life history. I found a significant association between cancer gene copies and species' longevity quotient. Additional paralogues of tumor suppressor genes and oncogenes is not solely dependent on body size, but rather the balance between body size and longevity. Additionally, there is a significance association between life history traits and genes that are both germline and somatic tumor suppressor genes. The bioinformatic pipeline identified large tumor suppressor gene and oncogene copy numbers in the naked mole rat (*Heterocephalus glaber*), armadillo (*Dasypus novemcinctus*), and the two-fingered sloth (*Choloepus hoffmanni*). These results suggest that increased paralogues of tumor suppressor genes and oncogenes are these species' modes of cancer resistance.

ACKNOWLEDGMENTS

This research was made possible from the support of my committee members, Dr. Carlo C. Maley, Dr. Melissa A. Wilson, and Dr. Marc Tollis. I would like to especially thank Dr. Marc Tollis for mentoring me and directing my projects for 3.5 years. Even though Dr. Tollis relocated to NAU, I appreciate the time he took out of his schedule to meet with me weekly.

This research was also made possible by the Arizona State University Bioinformatics Core Lab, and the resources allocated from Arizona State University School of Life Sciences, and the Biodesign Institute at Arizona State University.

TABLE OF CONTENTS

Page

CHAPTER Page

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Cancer Across Species

Cancer is a disease that has a mortality rate of 11% to 25% in humans (Ferlay et al., 2015) and affects 1.7 million US residents yearly (Siegel et al., 2019). While many risk factors are involved with cancer, some include height (Lahmann et al. 2016) and age (White et al. 2014). For instance, every 10cm increase in height above average has been associated with a 14-18% lifetime risk of melanoma (Lahmann et al. 2016). In the United States, lifetime cancer diagnosis risk is 41% (White et al. 2014). Also, more than 50% of cancers become diagnosed in patients that are 65 and older (White et al. 2014). Similar patterns are also observed in dogs (Paoloni et al., 2007). For example, large breeds of dogs have a 61% higher chance of getting bone sarcoma than smaller dogs (Tjalma R. A. 1966). Due to these patterns, large body size and increased cell divisions may elevate cancer risk. Every cell division has the consequence of obtaining a harmful, cancer-initiating mutation. Therefore, larger animals theoretically should have more cancer (Tollis, Boddy, et al., 2017). However, Richard Peto noticed that there was not a correlation between body size and lifespan across species (Peto et al., 1975). Mice and humans have comparable cancer rates, yet humans have 1,000 times more cells and have longer lifespans (Peto et al., 1975). Therefore, within a species, cancer rates correlate with body size.

Life History Theory

Natural selection has acted on species to prevent cancer. Studied compare cancer resistance mechanisms in mammals with fast life-history and those with slow life

histories. Characterization of a fast-life history include smaller body masses, shorter lifespans, and high reproduction rates (Kraus et al., 2005). In short-lived animals, reproduction has a higher distribution of energy than somatic maintenance. These animals reproduce quickly and most likely die from other causes besides cancer (Boddy et al. 2015). Larger body masses, long lifespans, and fewer offspring describe a slow life-history (Kraus et al., 2005). Species under selection for slow life histories, are under selection, in part, to maintain their soma over long periods, which likely involves preventing cancer in those cells (Boddy et al. 2015).

Anti-Aging Mechanisms

Bats have extended lifespans compared to mammals with comparable body sizes. Species in the *Myotis* genus have the most pronounced longevity in Chiroptera. The genus contains 13 species that have a lifespan that exceeds 20 years (Foley et al., 2018). The species, *Myotis brandtii*, is the longest-lived bat, with a maximum longevity of 41 years and a body mass of 7 grams (Seim et al., 2013). Telomere maintenance contributes to a bat's extended lifespan. In most mammals, telomeres shorten with repetitive cell division, which eventually limits the total number of times a cell may divide, and therefore limits the replenishment of stem cells that maintain tissues. However, 21 telomere maintenance undergo positive selection in *Myotis*. These genes are responsible for telomere lengthening and DNA repair (Foley et al., 2018).

In addition to bats, naked mole rats (*Heterocephalus glaber)* have low rates of cancer, with a lifespan of 30 years (Buffenstein, R., 2005). Their longevity is approximately 8 times longer than the lifespan of an average mouse (Lewis et al., 2012). They live in a controlled underground environment with little light and constant

temperatures. Out of a thousand studies, reports found six cases of neoplasia and cancer. These cases occurred due to a higher exposure of temperatures and more light than average (Seluanov et al., 2018). Similarly, to Chiropterans, the naked mole rat exhibits positive selection for telomeric lengthening genes such as TOP2A, which could contribute to their prolonged lifespan (Tollis, Schiffman, et al., 2017). Besides telomere preservation, naked mole rats have insignificant senescence; observations of age-related disease occurred only when they reach maximum longevity (Buffenstein, R., 2008). Their lack of senescence may be due to the expression of a senescence regulating gene, TP53. The naked mole rat's expression of TP53 is 50% higher than the levels in other rodents such as the mouse (*Mus musculus*) (Lewis et al., 2012). The selection pressures on telomere maintenance and senescence suggest that the naked mole rate would be an ideal model organism for cancer resistance research.

Mechanisms for Cancer Resistance

Large, long-lived mammals should have higher cancer rates due to the larger number of cell divisions required to generate and maintain their bodies. However, species such as the African Savanna elephant, *Loxodonta africana*, only has a cancer mortality rate of 5% (Abegglen et al., 2016). This species has a body mass of $4.5 \bullet 10_6$ grams (100 times greater than humans), and a maximum longevity of 80 years (Jones et al., 2009). The low cancer mortality rates may be caused, in part, by the elephant's 20 copies (40 alleles) of the tumor suppressor gene, TP53 (Abegglen et al., 2016; Sulak et al., 2016). TP53's responsibility to maintain the fidelity of the genome gave it the name, "guardian of the genome" (Caulin et al., 2015). It is responsible for apoptosis, senesce and cell-cycle arrest. High expression levels of TP53 could regulate the cell cycle and induce

apoptosis due to the presence of damaged DNA. Elephants have a higher apoptotic response to DNA damage than humans (Abegglen et al., 2016).

Cetaceans are the largest placental mammals, with the bowhead whale (*Balaena mysticetus)* containing 1000 times more cells than humans. Bowhead whales can live past 200 years and have no reports of cancer (though, to be fair, there is no data on cancer rates in bowhead whales, one way or the other). Unlike the African Savanna elephant, the bowhead whale does not have duplications of TP53. However, genomic analysis on the bowhead whale has revealed that there is positive selection on the DNA repair gene, ERCC1. Excess copies of the gene may reduce mutation rates and reduce the need for genes that control the cell cycle (Seluanov et al., 2018). Positive selection has also been found on anti-aging genes such as APTX, ERCC3, FGFR1, FOXO3, NIG, and SOCS2 (Keane et al., 2015). These genes may indirectly aid the whale in suppressing cancer. Similar to the bowhead whale, the humpback whale (*Megaptera novaeangliae*) demonstrates a positive selection on cancer genes. These genes include ATR, BCORL1, PICALM, PRDM2, and TPR (Tollis et al., 2019). Additionally, the species has duplications in growth and apoptosis genes. These gene include NOX5, PRMT2, and SLC25A6 (Tollis et al., 2019). The positive selection and duplications on these genes may be responsible for the cetacean's gigantism and their (inferred) low cancer risk. The only report of cancer in cetaceans comes from the beluga whales (*Delphinapterus leucas*) of the St. Lawrence estuary, which was highly polluted. In that population, belugas had an 18% cancer rate due to do with their exposure to pollution (Martineau et al., 2002).

Numerous studies have sought the genomic mechanisms underlying cancer resistance in mammals (Sulak et al.,2016, Keane et al., 2015, Tollis et al., 2019, Vicens et

al., 2018). Two genomic mice models were conducted to see the impact of TP53 in small mammals. In the first model, mice had overexpressed isoforms of *p44.* These mice underwent extreme aging but were cancer-resistant (Reinhardt et al., 2012). In the second model, mice were engineered to have two additional alleles of TP53. These mice were cancer-resistant, but did not undergo extreme aging (Reinhardt et al., 2012). Taken with the evidence from the elephant genome, this is evidence that gene duplications may provide a powerful mechanism for the evolution of new traits, including cancer suppression. However, a systematic study to determine if tumor suppressor gene duplications are associated with life-history traits such as body mass or lifespan across mammals has not been done.

Natural section may have selected many ways to prevent cancer. An increase in tumor suppressor genes may allow a species to repair DNA more efficiently or increase the rate of apoptosis with the presence of damaged DNA. The number of oncogenic paralogues could be reduced to avoid possible cancer driving mutations. Of course, oncogenes (technically, proto-oncogenes) have important functions in normal cells, often as part of the regulation of cell proliferation. Similarly, to the naked mole rat, the habitat in which a species lives may affect their cancer susceptibility. Those that live in a sun-exposed environment may have increased mechanisms to avoid skin cancer. These traits may allow a species to demonstrate cancer resistant mechanisms. Therefore, the study had three goals: 1) to identify gene copies in tumor suppressor genes and oncogenes across mammals with the use of a bioinformatic pipeline; 2) to determine if there is a correlation between life history data with gene duplications; and 3) to identify specific animals or genes that should be further studied to understand mechanisms of cancer

resistance. The latter goal has the potential to inform future human therapies. I seek to reach those goals by testing for evidence of natural selection in genes involved with cancer resistance, across species.

CHAPTER 2

METHODS AND MATERIALS

Collection of Tumor Suppressor Genes and Oncogenes

The tumor suppressor genes and oncogenes analyzed were acquired from COSMIC: Catalogue of Somatic Mutations in Cancer (Tate et al., 2018). In December 2018, I retrieved 548 gene symbols from a curated list. In addition to the gene symbols, the gene names, cancer association, and classification between tumor suppressor genes were recorded. Cataloged tumor suppressor genes were then further classified as genes that act as tumor suppressors in the germline, because inactivation of them leads to heritable cancer syndromes, or act as tumor suppressors in somatic cells because their inactivation in somatic cells increases the probability that those cells will evolve into cancers. Some genes act as tumor suppressors in both the germline and the soma. COSMIC had data on 242 tumor suppressor genes. This was further subdivided into 43 genes that act as tumor suppressor genes in both the soma and germline, 35 germline tumor suppressor genes, and 143 somatic tumor suppressor genes. COSMIC identified 72 genes that can act as both tumor suppressors and oncogenes depending on the types of mutations they acquire. The database also had information of 240 oncogenes.

Determination of Gatekeeper and Caretaker Genes

Each tumor suppressor gene was given the classification of being a gatekeeper gene or a caretaker gene. Caulin et al., (2015) provided a list of 59 gatekeeper and caretaker genes. The remainder of the genes, including genes that were considered tumor suppressor genes and oncogenes, were classified using the gene summaries from GeneCards (Stelzer et al., 2016). Genes that are gatekeepers are responsible for the

control of cell checkpoints and proliferation. Caretaker genes are responsible for DNA repair and inhibiting DNA damage (Caulin et al., 2015). The remainder of the genes, including genes that were considered tumor suppressor genes and oncogenes, were classified using the gene summaries from GeneCards (Stelzer et al., 2016).

Collection of Human Protein Sequences

I gathered human protein sequences from Ensembl (Hunt et al., 2018) based on the gene symbols collected from COSMIC using the tool, BioMart (Hunt et al., 2018). The database used in BioMart was Ensembl Genes 98, with the Human genes (GRCh38.p13) dataset. The gene names from COSMIC were used in the external reference identification parameter. The output results were peptide sequences. If there were multiple peptide sequences available for one gene, the longest sequence was collected.

Collection of Mammal Genomes

The whole genomes of 63 mammals were collected from NCBI's Genome Browser (O'Leary et al., 2016) and The Bowhead Whale Genome Resource (Keane et al., 2015). The assembly names, assembly level, genome length, genome coverage, and sequencing method can be found in Table 1. The genomes collected represent animals from five Superorders. The study looked at genomes of 49 Laurasiatherians, 8 Euarchontoglires, 3 Afrotherians, 2 Xenarthrans and 1 species of. A phylogeny of the mammals can be found in Figure 1.

Table 1

Genome Assemblies

Species Name	NCBI Assembly ID	Assembly Level	Genome Length	Genome Coverage	Sequencing Method
Vicugna pacos Alpaca	Vi_pacos_V1.0	Scaffolds	2092.95 Mb	72.5x	Illumina HiSeq2000
Bison bison American Bison	Bison_UMD1.0	Scaffolds	2828.03 Mb	60x	454; Illumina HiSeq
Arctocephalus gazella Antarctic Fur Seal	ArcGazv1.4	Scaffolds	2313.49 Mb	200x	Illumina HiSeq
Balaenoptera bonaerensis Antarctic Minke Whale	ASM97880v1	Scaffolds	2234.64 Mb	60x	Illumina HiSeq2000
Camelus dromedarius Arabian Camel	PRJNA234474_Ca_dromedarius_V1.0	Scaffolds	2084.54 Mb	46.43x	Illumina HiSeq2000
Dasypus novemcinctus Nine-Banded Armadillo	Dasnov3.0	Scaffolds	3631.52 Mb	6x	Sanger
Camelus bactrianus Bactrian Camel	Ca_bactrianus_MBC_1.0	Scaffolds	1780.72 Mb	79.2x	Illumina HiSeq2000
Delphinapterus leucas Beluga Whale	ASM228892v3	Scaffolds	2362.78 Mb	117x	Illumina HiSeqX
Ursus americanus American Black Bear	ASM334442v1	Scaffolds	2588.39 Mb	100x	Illumina; PacBio
Balaena mysticetus Bowhead Whale	NA	Scaffolds	NA	150	Illumina HiSeq
Ursus arctos horribilis Brown Bear	ASM358476v1	Scaffolds	2328.66 Mb	50x	Illumina HiSeq
Pan troglodytes Chimpanzee	Clint_PTRv2	Chromosome	3050.4 Mb	124x	Illumina HiSeq
Tursiops truncatus	Ttru_1.4	Scaffolds	2477.89 Mb	2.5x	Sanger; 454 FLX; Illumina HighSeq

Species Name	NCBI Assembly ID	Assembly Level	Genome Length	Genome Coverage	Sequencing Method
Common Bottlenose Dolphin					
Bos taurus Cattle	ARS-UCD1.2	Chromosome	2715.85 Mb	80x	PacBio; Illumina NextSeq 500; Illumina HiSeq; Illumina GAII
Canis lupus Dog	CanFam3.1	Chromosome	2407.29 Mb	7x	Sanger
Equus asinus Donkey	ASM30337v1	Scaffolds	2356.05 Mb	61x	Illumina
Loxodonta africana African Savanna Elephant	Loxafr3.0	Scaffolds	3196.74 Mb	7x	Sanger
Neophocaena asiaeorientalis Yangtze Finless Porpoise	Neophocaena_asiaeorientalis_V1	Scaffolds	2284.63 Mb	106x	Illumina HiSeq2000
Giraffa tippelskirchi Giraffe	ASM165123v1	Scaffolds	2705.07 Mb	37x	Illumina HiSeq
Gorilla gorilla Western Lowland Gorilla	GorGor4	Chromosome	3063.36 Mb	80x	Illumina HiSeq
Eschrichtius robustus Grey Whale	ASM218922v1	Scaffolds	2849.45 Mb	11x	Illumina HiSeq
Cavia porcellus Domestic Guinea Pig	Cavpor3.0	Scaffolds	2849.45 Mb	6.8x	Sanger
Phocoena phocoena Harbor Porpoise	ASM307100v1	Scaffolds	2571.07 Mb	87x	Illumina NextSeq 500
Monachus schauinslandi Hawaiian Monk Seal	ASM220157v1	Chromosome X and scaffolds	2400.93 Mb	61x	Illumina HiSeq

Species Name	NCBI Assembly ID	Assembly Level	Genome Length	Genome Coverage	Sequencing Method
Hippopotamus amphibius Hippopotamus	ASM299558 v1	Scaffolds	2579.62 Mb	55x	HiSeq200
Equus caballus Horse	EquCab3.0	Chromosomes	2474.93 Mb	88x	Sanger; Illumina HiSeq; PacBio
Homo sapiens Human	GRCh38.p13	Chromosomes	2987.97 Mb	NA	Sanger
Megaptera novaeangliae Humpback Whale	megNov1	Scaffolds	2265.79 Mb	102x	Illumina HiSeq
Bos indicus x Bos taurus Hybrid Cattle	UOA_Angus_1	Scaffolds	2630.86 Mb	136x	PacBio RSII; PacBio Sequel; Illumina
Tursiops aduncus Indo-Pacific Bottlenose Dolphin	ASM322739 v1	Scaffolds	2503.93 Mb	180x	Illumina HiSeq
Sousa chinensis Indo-Pacific Humpbacked Dolphin	S_chinensis_fine_genome_map	Scaffolds	2338.99 Mb	107.6x	Illumina HiSeq
Pteropus vampyrus Large Flying Fox	Pvam_2.0	Scaffolds	2198.28 Mb	188x	Illumina
Myotis lucifugus Little Brown Bat	Myoluc2.0	Scaffolds	2034.58 Mb	7x	Sanger
Trichechus manatus Florida Manatee	TriManLat1.0	Scaffolds	3103.81 Mb	150x	Illumina HiSeq
Balaenoptera acutorostrata Minke Whale	BalAcu1.0	Scaffolds	2431.69 Mb	92x	Illumina HiSeq 2000
Mus musculus House Mouse	GRCm38.p6	Chromosome	2689.66 Mb	NA	Sanger

Species Name	NCBI Assembly ID	Assembly Level	Genome Length	Genome Coverage	Sequencing Method
Heterocephalus glaber Naked Mole-Rat	HetGla_female_1.0	Scaffolds	2631.08 Mb	90x	Illumina HiSeq
Monodon monoceros Narwhal	NGI_Narwhal_1	Scaffolds	2414.06 Mb	42x	10X Genomics; Dovetail Chicago; Dovetail
Okapia johnstoni Okapi	ASM166083 v1	Scaffolds	2878.13 Mb	30x	Illumina HiSeq
Monodelphis domestica Gray Short-Tailed Opossum	MonDom5	Scaffolds	3598.44 Mb	6.8x	Sanger
Orcinus orca Killer Whale	Oorc_1.1	Scaffolds	2372.92 Mb	200x	Illumina HiSeq
Lagenorhynchus obliquidens Pacific White-Sided Dolphin	ASM367639 v1	Scaffolds	2334.47 Mb	35.68x	Illumina HiSeq
Ailuropoda melanoleuca Giant Panda	ASM200744 v1	Scaffolds	2363.89 Mb	70x	Illumina HiSeq
Sus scrofa Pig	Sscrofa11.1	Scaffolds	2459.03 Mb	65x	PacBio
Ursus maritimus Polar Bear	UrsMar_1.0	Scaffolds	2301.38 Mb	101x	Illumina Genome Analyzer II
Equus przewalskii Przewalski's Horse	Burgud	Scaffolds	2395.95 Mb	85.63	Illumina HiSeq
Rattus norvegicus Norway Rat	Rnor_6.0	Chromosome	2743.3 Mb	3x	Sanger; SOLiD; PacBio
Macaca mulatta Rhesus Monkey	rheMacS_1.0	Chromosome	2971.33 Mb	100x	PacBio Sequel
Procavia capensis Cape Rock Hyrax	Pcap_2.0	Scaffolds	3749.9 Mb	107x	Illumina; Sanger dideoxy sequencing

Species Name	NCBI Assembly ID	Assembly Level	Genome Length	Genome Coverage	Sequencing Method
Mesoplodon bidens Sowerby's Beaked Whale	MesBid_v1_BIUU	Scaffolds	2797.69 Mb	32.4x	Illumina HiSeq
Physeter catodon Sperm Whale	ASM283717v2	Chromosome and scaffolds	2512.14 Mb	248x	BGISEQ-500
Panthera tigris Amur Tiger	PanTig1.0	Scaffolds	2391.08 Mb	99x	Illumina HiSeq 2000
Choloepus hoffmanni Hoffmann's Two-Fingered Sloth	C_hoffmanni_2.0.1	Scaffolds	3286.01 Mb	65x	Illumina
Odobenus rosmarus Pacific Walrus	Oros_1.0	Scaffolds	2400.15 Mb	200x	Illumina
Bubalus bubalis Water Buffalo	Bubbub1.0	Scaffolds	2836.17	119x	Illumina HiSeq 2000
Leptonychotes weddellii Weddell Seal	LepWed1.0	Scaffolds	3156.9 Mb	82x	Illumina HiSeq
Ceratotherium simum Southern White Rhinoceros	CerSimSim1.0	Scaffolds	2666.62 Mb	91x	Illumina HiSeq
Camelus ferus Wild Bactrian Camel	CB1	Scaffolds	2009.19 Mb	30x	Illumina GAIIx; 454 GS-FLX Titanium; SOLid 3
Bos mutus Wild Yak	BosGru_v2.0	Scaffolds	2703.27 Mb	130x	Illumina HiSeq; Illumina GA
Lipotes vexillifer Yangtze River Dolphin	Lipotes_vexillifer_v1	Scaffolds	2429.21 Mb	115x	Illumina HiSeq 2000
Bos indicus Zebu cattle	ASM293397v1	Chromosome	2707.15 Mb	100x	454; IonTorrent; Illumina NextSeq; Illumina MiSeq

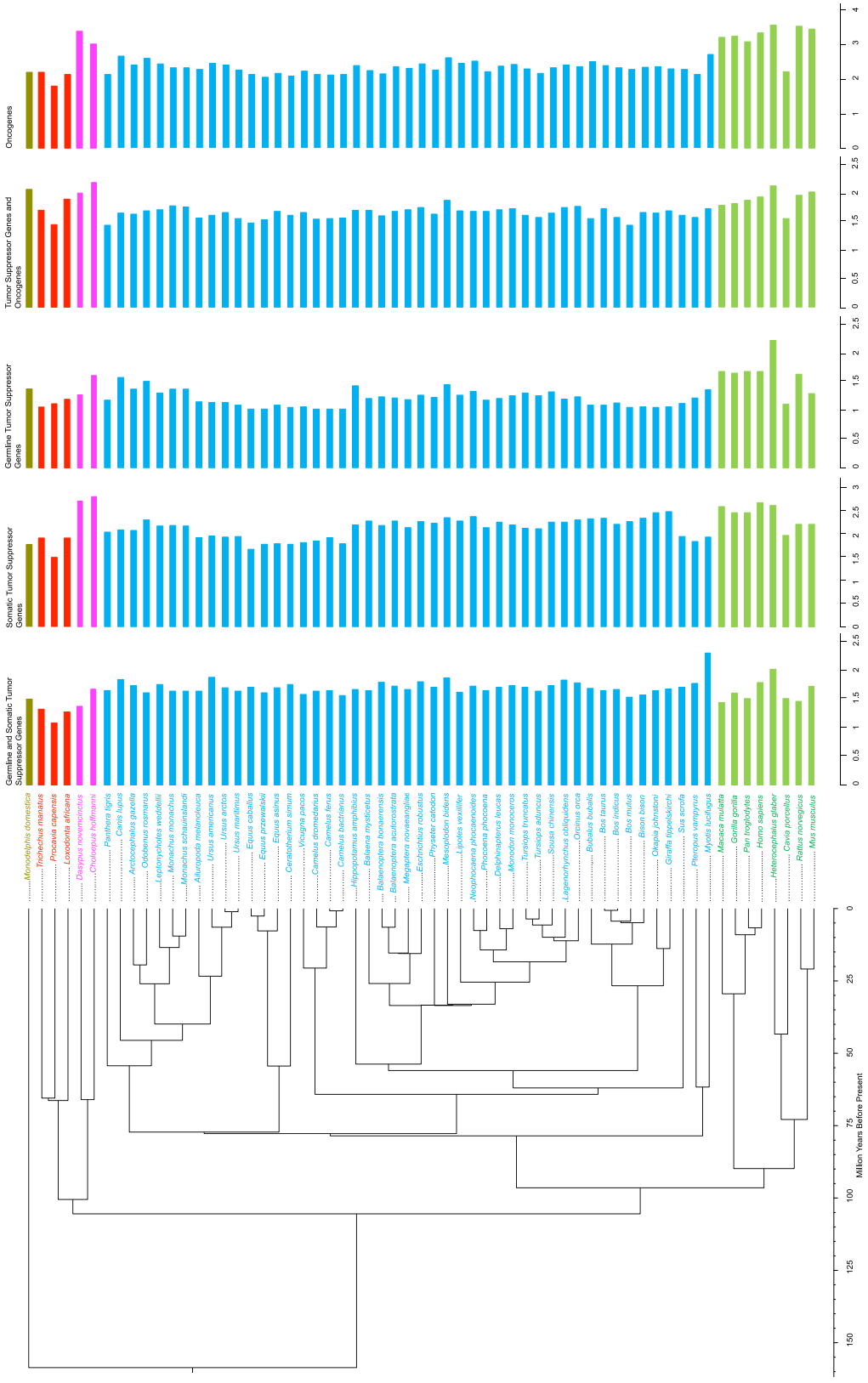

Figure 1. Time calibrated phylogenetic tree with corresponding gene duplications.

Collection of Life History Data and Ecology Data

The study observed the relationship between the proportion of tumor suppressor genes and oncogenes to each animal's life history and ecology. Pantheria was the source for the values for body mass and lifespan in captivity, as well as the superorder and order of each animal (Jones et al., 2009). The smallest animal studied, the little brown bat (*Myotis lucifugus*), has a body mass of 7.15g. The largest animal studied, the bowhead whale (*Balaena mysticetus*), has a body mass of 8.0•107g (Jones et al., 2009). Out of the 63 species studied, 16 species were smaller than humans, and 46 species were larger than humans. Additionally, the mean basal metabolic rate (BMR ml O2/hr) was collected for 25 species. Sayres et al. (2011) provided a preliminary list of BMR rates for 32 mammals, 19 of which were in my study. The BMR of 18 mammals was gathered from Sieg et al. (2009), the BMR of 7 mammals was collected from Jones et al. (2009), and one mammal's BMR was collected from Gumal et al. (1998). Also, the animal's biome (tropical or non-tropical), habitat (aquatic or terrestrial) and diet (carnivore, omnivore or herbivore) were obtained from Animal Diversity Web (Dewey et al., 2010). The identification of hemochorial, endotheliochorial and epitheliochorial placentation for 31 mammals came from Comparative Placentation (Benirschke 2007; Benirschke 2008; Benirschke 2010; Benirschke 2011) and in the paper by Mossman H. (1987). The life history data for all 63 mammals is in Appendix A.

Longevity Quotient Calculations

The Amniote Life History Database contained information regarding log-body mass, average longevity and maximum longevity for 2,320 eutherians, and 229 marsupials (Wolfram Research 2016). I used this information to calculate the longevity

quotient. The longevity quotient was calculated using the formula presented in Foley et al. (2018). The application of a linear regression exhibited the relationship between the log maximum longevity to the log body mass for the non-flying eutherians. The relationship created the line, y=0.2718x + 0.1396, and a R2 value of 0.5058 (Schneider-Utaka 2018). The equation was used to predict the expected longevity. Calculating the difference between the observed longevity over the expected longevity provided the values for longevity quotient in each mammal (Foley et al., 2018; Schneider-Utaka 2018). The longevity quotients for 62 mammals studied are found in Figure 2. The lifespan for Sowerby's Beaked Whale, (*Mesoplodon bidens*) is unknown. Therefore, its longevity quotient was not calculated. The longevity quotient for each mammal can be found in Appendix A.

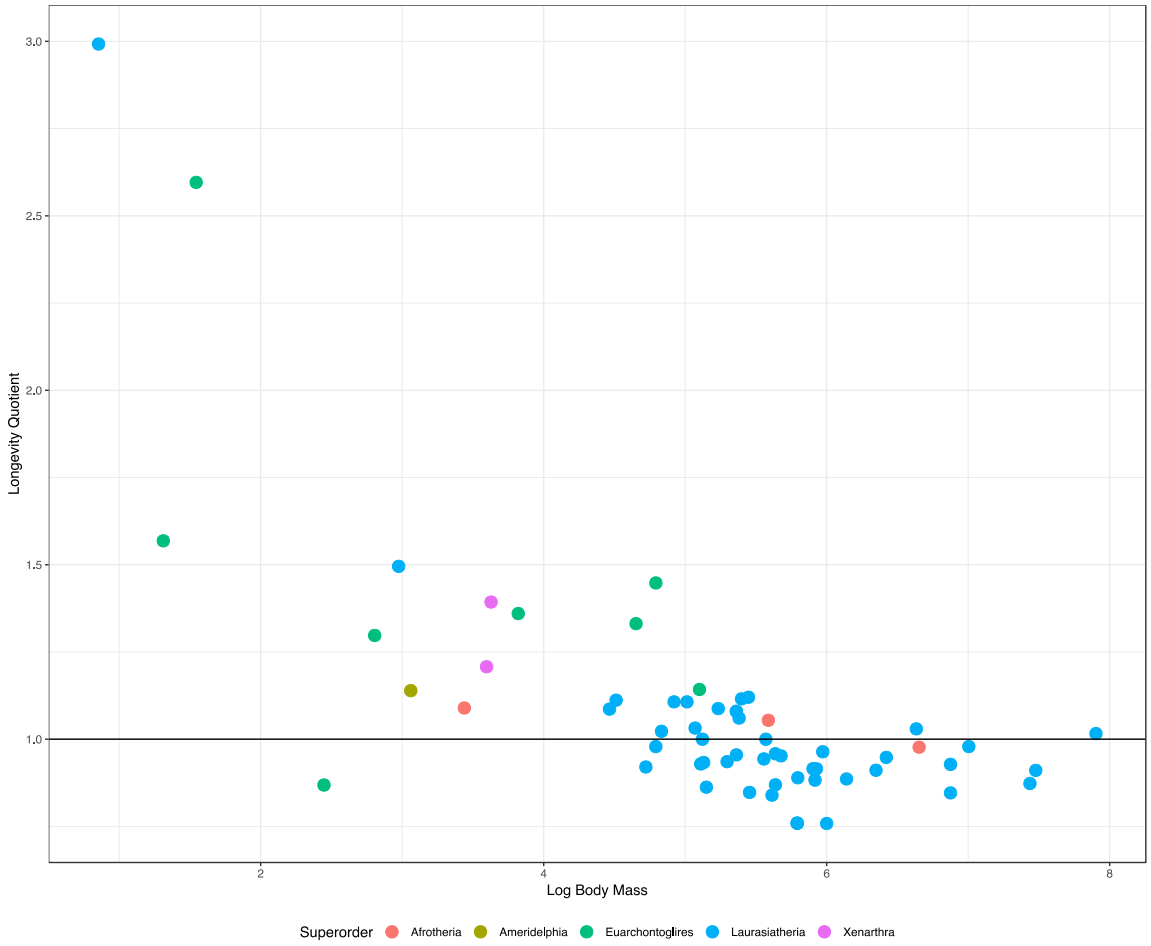

Figure 2. Longevity quotient across placental mammals. The species are color coordinated by mammalian superorder.

Bioinformatic Pipeline

A bioinformatic pipeline was written to identify gene duplications in placental mammals. The pipeline identified orthologues and paralogues of 548 tumor suppressor genes and oncogenes across 63 mammals. The pipeline required an input of human protein-coding peptide sequences, and whole-genome sequences, both in FASTA format. An altered version of web BLAT ran with a minscore of 55, and a minidentity of 60,65, and 70.

The pipeline used the BLAT parameters to search the specified genome for a copy that most closely matched the human protein sequence. Once it collected the top-scoring sequence and position, BLAT was rerun using the highest sequence to search the genome for candidate paralogues. To validate the orthologues and paralogues, the sequences ran against the RefSeq human protein database (taxis:9606) with BLASTX (Boratyn et al., 2012). The final results translated protein sequences into amino acid sequences.

The pipeline ran for all genomes and genes under a minidentity of 65% and 70%. The genes, PRDM1 and QKI resulted in no copies for all studied mammalian genomes, even with humans. To identify gaps or missing genes from the results, the pipeline was run with a minidentity of 60% and run using web BLAT parameters. The UCSC Genome Browser uses a minscore of 20, and a minidentity of 0 (Kent et al., 2012; Kent 2012). The two tests continued to result in errors with the genes; orthologues and paralogues were not found.

Manual BLAT

Once the pipeline had been run with multiple parameters, two genes consistently came up with zero copies. Since there were human sequences available, there was a pipeline error. Therefore, to obtain results for PRDM1 and QKI, the genes were manually collected using web BLAT on Ensembl and the UCSC Genome Browser. Ensembl had access to 29 mammalian genomes that were present in the pipeline (Hunt et al., 2018). These animals included: *Ailuropoda melanoleuca, Bison bison, Bos indicus, Bos indicus x Bos Taurus, Bos mutus, Bos Taurus, Canis lupus familiaris, Cavia porcellus, Choloepus hoffmanni, Dasypus novemcinctus, Equus asinus asinus, Equus caballus, Gorilla gorilla gorilla, Heterocephalus glaber, Homo sapiens, Loxodonta africana,*

Monodelphis domestica, *Mus musculus*, *Myotis lucifugus*, *Pan troglodytes*, *Panthera tigris*, *Procavia capensis*, *Pteropus vampyrus*, *Rattus norvegicus*, *Sus scrofa*, *Tursiops truncates*, *Ursus americanus*, *Ursus maritimus and Vicugna pacos.* The UCSC Genome Browser had access to an additional four mammalian genomes (Kent WJ, Sugnet 2002; Bhagwat et al., 2012). These animals include: *Balaenoptera acutorostrata*, *Ceratotherium simum*, *Macaca mulatta and Trichechus manatus latirostris.* PRDM1 and QKI had a maximum of 33 species which contained orthologues and paralogues. Their sequences, and the number of duplications were added to the combined results. The 30 additional animals received a NA for copy number for both genes.

Combination of Pipeline Results

The pipeline results were combined to determine the maximum number of gene copies found in the tumor suppressor genes and oncogenes. The minidentity of 70% provided results for 546 genes. The results from the 65% minidentity contributed to the gene copies for 56 genes. Although there were higher gene copy numbers found with the 65% protein identity, false positives were identified. A noticeable error occurred in humans with TP53. Humans have one copy of TP53 (Abegglen et al., 2015). However, the pipeline identified two copies. Therefore, the pipeline was rerun and validated with a 70% protein identity. The genes identified using the 65% identity were manually verified using BLASTX before adding them to the results. More false positives were identified using the parameter. Lastly, two genes, PRDM1 and QKI's manual web BLAT results were added for 33/63 mammals. The combined results are can be found in Appendix A.

Normalization of Gene Copies

Normalization occurred for the results for tumor suppressor genes, oncogenes, caretaker genes, and gatekeeper genes from the total number of duplications. The results were normalized by determining the relationship between the number of duplications by the number of genes with at least one copy. This equation determines if the animals have more copies of genes than expected.

Collection of phastCons Scores

Conservation scores for all 545 genes were obtained from the UCSC Table Browser (Karolchik et al., 2004). The clade studied was mammals, primarily focusing on the human genome. The assembly used was December 2013 (GRCh38/hg38). The group studied was comparative genomics with a conservative track. The table used to obtain the results was collected from Cons 100 Verts (phastCons100way). The gene positions were gathered from Ensembl. These gene positions match the sequences using BioMart from Ensembl. The phastCons scores had a 10,000,000-line filter applied. The phastCons scores for each gene position were averaged together to determine a gene's level of conservation.

Orthologues Identification

OrthoDB was used to determine if orthologues were present in the most distant ancestor of my animals, *Monodelphis domestica* (Kriventseva et al. 2018). If the gene was not present in *Monodelphis domestica*, its presence was investigated in an Afrotherian (*Loxodonta africana*) and a Xenarthran (*Dasypus novemcinctus*).

PGLS ANOVA Tests

The number of tumor suppressor genes and oncogenes were tested for their significance against the mammals' life history and their ecology. I used R (version 1.1.456) to compute Phylogenetic Generalized Least Squares (PGLS) regressions (Fritz et al., 2009) with the caper package (version 2.14) (Orme et al., 2018). The phylogenetic supertree (Figure 1) used was computed with the mammals in the Amniote Database (Wolfram Research 2016). The species, *Mesoplodon bidens* (Sowerby's Beaked Whale) and the *Bos indicus x Bos taurus* (Hybrid cattle), were unavailable in the Amniote Database and were excluded from the phylogenetic tree.

A total of 106 regressions were applied to the pipeline results. The results for all regressions can be found in Appendix B. For all the regressions, normalized values for tumor suppressor genes and oncogenes were used. To adjust for the probability of false positives over multiple-comparisons, I conducted both a Bonferroni correction and a false-discovery rate test. The conservative method, the Bonferroni correction, concluded that a P-value below 0.00047 is significant. The false-discovery rate was calculated with a P-value of 0.05 (McDonald 2014). The results for significance can be found in Table 2.

Table 2

Phylogenetic Regressions with Correction Tests

Test Preformed	**Lambda**	**R_2**	**P-Value**	**Bonferroni Significance (P<0.00047)**	**False Discovery Rate Significance**	**False Discovery Rate P-Value**
Total TSGs Normalized vs Total Oncogenes Normalized	0.981	0.6811	2.95 E-14	Significant	Significant	3.07983 E-12
Gatekeeper Genes Normalized vs Oncogenes Normalized	1	0.6726	5.81 E-14	Significant	Significant	3.07983 E-12
Total Oncogenes Normalized vs Superorder	0	0.7418	1.46 E-13	Significant	Significant	5.141 E-12
Somatic/ Germline TSGs Normalized vs Longevity Quotient	0.71	0.5379	6.26 E-10	Significant	Significant	1.6589 E-08
Total Oncogenes Normalized vs Order	0	0.7821	9.72 E-10	Significant	Significant	2.06128 E-08
Caretaker Genes Normalized vs Oncogenes Normalized	0.973	0.4663	1.77 E-08	Significant	Significant	3.13053 E-07

Test Preformed	Lambda	R_2	P-Value	Bonferroni Significance (P<0.00047)	False Discovery Rate Significance	False Discovery Rate P-Value
TSGs and Oncogenes Normalized vs Superorder	0	0.5604	3.91 E-08	Significant	Significant	5.92086 E-07
Caretaker Genes Normalized vs Gatekeeper Genes Normalized	0.866	0.3994	3.88 E-07	Significant	Significant	4.81711 E-06
TSGs and Oncogenes Normalized vs Order	0	0.6972	4.09 E-07	Significant	Significant	4.81711 E-06
Somatic TSGs Normalized vs Order	0.761	0.6844	8.59 E-07	Significant	Significant	9.10858 E-06
Somatic/ Germline TSGs Normalized vs Order	0	0.6534	4.51 E-06	Significant	Significant	4.31067E-05
Somatic/ Germline TSGs Normalized vs Superorder	0	0.4584	4.88 E-06	Significant	Significant	4.31067 E-05
Germline TSGs Normalized vs Order	0	0.6135	2.99 E-05	Significant	Significant	0.0002439

Test Preformed	Lambda	R_2	P-Value	Bonferroni Significance (P<0.00047)	False Discovery Rate Significance	False Discovery Rate P-Value
Total TSGs Normalized vs Order	0.699	0.5913	7.70 E-05	Significant	Significant	0.0005829
Total TSGs Normalized vs Superorder	0.943	0.3497	0.0003	Not Significant	Significant	0.0021695
Somatic/ Germline TSGs Normalized vs Log-body mass	0.813	0.1785	0.0016	Not significant	Significant	0.010759
Caretaker Genes Normalized vs Longevity Quotient	0.817	0.168	0.0025	Not significant	Significant	0.0158875
Germline TSGs Normalized vs Longevity Quotient	0.985	0.1614	0.0032	Not significant	Significant	0.018603
Total Oncogenes Normalized vs Longevity Quotient	0.992	0.1561	0.0037	Not significant	Significant	0.0208932
TSGs/Oncogenes Normalized vs Diet	1	0.1737	0.0085	Not significant	Significant	0.0450023

Collection of Neoplasia Rates

The Northwest ZooPath database contained the neoplasia rates for 36 of my mammals (Garner et al. 2019). The total records, neoplasia records, neoplasia rate, average age, and average age with neoplasia can be found in Table 3. Species with fewer than 25 total records were excluded from phylogenetic regressions. Finalized tests used 27/36 mammals in Table 3.

Table 3

Neoplasia Rates

Scientific Name	Total Records	Neoplasia Records	Neoplasia Rate	Average Age (months)	Average Age Neoplasia (months)
Bison bison	142	14	0.1	83.54	249.29
Bos indicus	6	0	0	117.67	NA
Bos taurus	52	10	0.19	105.7	109.39
Bubalus bubalis	10	6	0.6	168	168
Camelus dromedarius	150	44	0.29	172.21	179.87
Giraffa tippelskirchi	34	2	0.06	87.12	300
Okapia johnstoni	58	6	0.1	135.42	188.37
Sus scrofa	112	62	0.55	110.49	137.81
Ailuropoda melanoleuca	12	8	0.67	87.27	115.73
Canis lupus	917	445	0.49	103.77	118.85
Odobenus rosmarus	26	4	0.15	242.77	336
Ursus americanus	44	10	0.23	149.45	266.4
Ursus arctos	160	82	0.51	260.49	272.74
Ursus maritimus	246	66	0.27	249.4	266.13
Delphinapterus leucas	2	0	0	156	NA
Orcinus orca	8	0	0	588	NA
Phocoena phocoena	4	0	0	30	NA
Tursiops truncatus	2698	30	0.01	259.09	336
Dasypus novemcinctus	76	26	0.34	182.07	255.39
Monodelphis domestica	18	4	0.22	42.67	48
Procavia capensis	138	20	0.14	61.26	79.2

Scientific Name	Total Records	Neoplasia Records	Neoplasia Rate	Average Age (months)	Average Age Neoplasia (months)
Camelus bactrianus	28	10	0.36	186	184.8
Vicugna pacos	1552	406	0.26	67.21	96.93
Ceratotherium simum	94	12	0.13	286.98	458
Equus asinus	183	6	0.03	71.05	188.23
Choloepus hoffmanni	32	12	0.38	189.75	260
Gorilla gorilla	288	52	0.18	333.31	409.92
Macaca mulatta	174	28	0.16	130.74	210.15
Pan troglodytes	444	128	0.29	334.4	423.34
Loxodonta africana	164	10	0.06	368.49	429.6
Cavia porcellus	3619	1260	0.35	24.1	47.7
Heterocephalus glaber	114	8	0.07	80.42	81
Mus musculus	1199	216	0.18	6.74	12.08
Rattus norvegicus	3736	1602	0.43	17.14	28.48
Trichechus manatus	20	4	0.2	177.69	120.47
Hippopotamus amphibius	12	2	0.17	622.16	718.37

Housekeeping Genes Analysis

The relationship between tumor suppressor genes and oncogenes with life history traits as compared to the relationship between housekeeping genes with life history traits. Dorus et al. (2014) provided a list of 94 housekeeping genes. 7 genes were also found in COSMIC and were removed from the list of housekeeping genes. The protein sequences for the 87 genes were obtained using BioMart on Ensembl (Hunt et al., 2018). The bioinformatic pipeline ran with a 70% mindidentity to determine if there were orthologues and paralogues in my 63 mammals. The copy numbers for the genes can be found in Appendix C. Normalization of gene copy numbers occurred, and underwent PGLS regressions using the phylogenetic tree in Figure 1.

CHAPTER 3

DATA ANALYSIS AND DISCUSSION

Collection of Gene Copies

Using the available tumor suppressor genes and oncogenes from COSMIC (Tate et al., 2018), gene duplications for 63 mammals were identified with the use of a bioinformatic pipeline (Methods: Bioinformatic Pipeline). The pipeline found at least one orthologue in 546/548 genes. The pipeline failed to collect results for PRDM1 and QKI. The genes ran in the pipeline with a protein identity of 60%, 65%, 70%, and the web BLAT parameters. Each time, there were no sequences collected. Sequences for the genes were manually obtained for 33 mammals using BLAT from Ensembl and UCSC genome browser (Hunt et al., 2018; Kent et al., 2002; Bhagwat et al., 2012).

In addition to the failure to identify copies within a genome, the pipeline also failed to validate the copies for three genes: HMGA2, HNRNPA2B1, and MUC4. The gene HMGA2 is a part of the structure of the enhanceosome. Humans have one copy of the gene, with limited expression in 27 different tissues (HMGA2 2019). However, the pipeline identified 2051 copies in human. The pipeline also found 1276 copies in chimpanzee, 9819 copies in gorilla, and 1282 copies in rhesus. A maximum of 2 copies of HMGA2 was found for all others. The gene, HNRNPA2B1 is responsible for pre-mRNA processing and transport (HNRNPA2B1 2019). The pipeline identified 114 copies of the gene in the dog, 95 copies in walrus, 84 copies in mouse and 83 copies in rat. Lastly, the gene, MUC4 is a mucus protein that has been linked to cell differentiation (MUC4 2019). Copies for the gene were only identified in humans and chimpanzees. The pipeline identified 128 copies in human, and 23 copies in chimpanzee. Random

sequences for the three genes were manually implemented in BLASTX to distinguish if these copies were false positives (Boratyn et al., 2012). Each gene contained false positives and were removed from the data set. Additional genes with a minimum of one copy were manually checked with BLASTX to determine if the pipeline results were accurate. The pipeline's overall failure to identify genes correctly was 0.91%.

Gene duplications may also be the result of the mammal's genome coverage (Table 1). Due to sequencing and the presence of scaffolds, some genomes may possess copies of the gene are un-identifiable using the scaffold data in the pipeline. In addition, due to the genome assembly method, the pipeline may have identified some gene copies which may not be present if assembled differently. *Dasypus novemcinctus, Tursiops truncates, Canis lupus familiaris, Loxodonta africana, Cavia porcellus, Homo sapiens, Myotis lucifugus, Mus musculus, Monodelphis domestica,* and *Rattus norvegicus* used Sanger sequencing for their genome assemblies. Since Sanger sequencing produces longer fragments, the duplications identified may be more accurate for these species.

The pipeline determined that 476 genes or 87.34% of tested sequences had at least a 1: many relationship (Table 4). Also, 293 genes had at least three paralogues (53.76%), 176 genes had at least four paralogues (36.33%), 124 genes had at least five paralogues (24.95%) and 45 genes had at least paralogues (8.26%). Validation of paralogues occurred with the use of BLASTX back to the human genome to eliminate false positives (Boratyn et al., 2012). Caulin et al., conducted a similar test with 830 tumor suppressor genes in 36 mammals using Ensembl (2015). They identified 383 genes with a 1:many relationship (46.14%) compared to the pipeline's 87.34% (Caulin et al., 2015). This study contains at fewer genes, yet it also covers oncogenes. According to the

breakdown of gene duplications from Table 4, consistently, there were more gene duplications found in oncogenes than tumor suppressor genes. In addition, the parameters of the pipeline may have identified more gene paralogues. The Ensembl annotation pipeline has the parameter to include only transcribed evidence when outputting their results (Hunt et al., 2018). The bioinformatic pipeline does not include that criteria. Therefore, the pipeline found genomic sequences that have not been previously annotated or published in the past. Also, the results from the pipeline notes that duplications of sequences are common across placental mammals.

Table 4

Breakdown of Gene Duplications

Gene Duplications	Gene Type	Number of Genes	Percentage to Total Genes
Two or more copies	Total TSGs	201	87.34%
	Somatic and Germline TSGs	33	
	Somatic TSGs	141	
	Germline TSGs	27	
	TSGs and Oncogenes	62	
	Oncogenes	213	
Three or more copies	Total TSGs	109	53.76%
	Somatic and Germline TSGs	16	
	Somatic TSGs	81	
	Germline TSGs	12	
	TSGs and Oncogenes	42	
	Oncogenes	142	
Four or more copies	Total TSGs	75	36.33%
	Somatic and Germline TSGs	12	
	Somatic TSGs	58	
	Germline TSGs	5	
	TSGs and Oncogenes	27	
	Oncogenes	96	
Five or more copies	Total TSGs	51	24.95%
	Somatic and Germline TSGs	10	
	Somatic TSGs	37	
	Germline TSGs	4	
	TSGs and Oncogenes	16	
	Oncogenes	69	
Ten or more copies	Total TSGs	19	8.26%
	Somatic and Germline TSGs	4	
	Somatic TSGs	12	
	Germline TSGs	3	
	TSGs and Oncogenes	4	
	Oncogenes	22	

Breakdown of Highest and Lowest Gene Duplications

The genes were categorized into caretaker genes, gatekeeper genes, total tumor suppressor genes, tumor suppressor genes that are both germline and somatic, germline

tumor suppressor genes, somatic tumor suppressor genes, oncogenes, and genes that are both tumor suppressor and oncogenes. The number of gene duplications were normalized to determine if animals have more copies of genes than expected. The normalized data for all the animals can be found in Appendix A.

The normalized data were separated into the categories to determine which animals had the most duplications, and which animals had the least (Table 5). The animals with the most normalized total tumor suppressor genes were *Heterocephalus glaber* (2.39), *Choloepus hoffmanni* (2.33), and *Homo sapiens* (2.3). *Procavia capensis* had the lowest duplications of tumor suppressor genes (1.34). Under a normalization method that calculated the number of tumor suppressor gene copies over the total number of tumor suppressor genes studied, again, *Heterocephalus glaber* has the most paralogues. The superorder, Euachontoglires, had the highest normalized copy numbers for caretaker genes. *Heterocephalus glaber* (2.03), *Homo sapiens* (1.61) and *Rattus norvegicus* (1.57) had the most copy numbers, and *Procavia capensis* (1.15) had the least. Similarly, to total tumor suppressor genes, *Choloepus hoffmanni* (2.54), *Heterocephalus glaber* (2.42) and *Homo sapiens* (2.41) had the greatest number of genes, and *Procavia capensis* (1.43) had the fewest. Rodentia had the highest copy numbers for oncogenes; *Heterocephalus glaber* (3.41), *Rattus norvegicus* (3.37) and *Mus musculus* (3.29) had three times more oncogenes than expected. Using the alternative normalization method, *Heterocephalus glaber* consistently had the highest number of oncogenes. The species, *Procavia capensis* (1.71) had the lowest normalized copy numbers for oncogenes.

Table 5

Normalized Gene Duplications for Non-Conserved and Conserved Genes

Gene Type	**Normalized Gene Copies**			**Normalized Non-Conserved Gene Copies (<0.3)**			**Normalized Conserved Gene Copies (>0.3)**		
Caretaker Genes	Max	Naked Mole Rat	2.03	Max	Naked Mole Rat	1.95	Max	Human	3.25
		Human	1.61		Two-Fingered Sloth	1.57		Naked Mole Rat/ Mouse	3
		Rat	1.57		Human	1.51		Chimp/ Guinea Pig	2.75
	Min	Rock Hyrax	1.15	Min	Rock Hyrax	1.12	Min	Two-Fingered Sloth	
		American Bison	1.17		American Bison	1.13		White Rhino/ Yangtze River Dolphin/ Horse/ Large Flying Fox	1.25
		Giant Panda	1.18		Indo-Pacific Bottlenose Dolphin	1.34			
Gatekeeper Genes	Max	Two-Fingered Sloth	2.54	Max	Two-Fingered Sloth	2.63	Max	Rat	1.74
		Naked Mole Rat	2.42		Naked Mole Rat/ Armadillo	2.49		Mouse	1.70
		Human	2.39					Beluga Whale/ Orca Whale	1.68
	Min	Rock Hyrax	1.43	Min	Rock Hyrax	1.45	Min	Polar Bear	1.15
		Horse	1.57		Horse	1.59		Wild Yak	1.18
		Bactrian Camel	1.63		Bactrian Camel	1.66		Tiger	1.19
Total Tumor Suppressor Genes	Max	Naked Mole Rat	2.39	Max	Naked Mole Rat	2.41	Max	Mouse	2.25
		Two-Fingered Sloth	2.33		Two-Fingered Sloth	2.39		Naked Mole Rat	2.13
		Human	2.3		Human	2.33		Orca Whale	2.07
	Min	Rock Hyrax	1.34	Min	Rock Hyrax	1.35	Min	Rock Hyrax	1.2
		Horse	1.48		Horse	1.48		Sperm Whale	1.27
		Przewalski Horse	1.54		Bactrian Camel	1.53		Common Bottlenose Dolphin	1.21

Gene Type	Normalized Gene Copies			Normalized Non-Conserved Gene Copies (<0.3)			Normalized Conserved Gene Copies (>0.3)		
Oncogenes	Max	Naked Mole Rat	3.41	Max	Naked Mole Rat	2.99	Max	Rat	8.47
		Rat	3.37		Human	2.89		Mouse	7.93
		Mouse	3.29		Armadillo	2.83		Naked Mole Rat	6.17
	Min	Rock Hyrax	1.71	Min	Rock Hyrax	1.56	Min	Rock Hyrax	2.84
		Przewalski Horse	1.94		Indo-Pacific Bottlenose Dolphin	1.78		Przewalski Horse	2.86
		White Rhino	1.98		Przewalski Horse	1.81		Horse	2.93
Germline and Somatic Tumor Suppressor Genes	Max	Naked Mole Rat	2.02	Max	Naked Mole Rat	2.13	Max	Dog/ Harbor Porpoise	2.5
		Little Brown Bat	1.83		Little Brown Bat	1.92		Finless Porpoise/ Beluga Whale/ Donkey/ Hybrid Cattle/ Pacific White Sided Dolphin/ Humpback/ Opossum/ Mouse/ Orca Whale/ Pig	2
		Human	1.79		Human	1.85			
	Min	Rock Hyrax	1.09	Min	Rock Hyrax	1.09	Min	American Bison/ Antarctic Minke Whale/ Bactrian Camel/ Tiger/ Wild Bactrian Camel/ Wild Yak/ Cow/ Guinea Pig/ Two-Fingered Sloth/	1
		Wild Yak	1.22		Bactrian Camel/ Harbor Porpoise	1.26			
		Bactrian Camel	1.24		Wild Yak	1.27			

Gene Type	Normalized Gene Copies			Normalized Non-Conserved Gene Copies (<0.3)			Normalized Conserved Gene Copies (>0.3)		
								Armadillo/ Elephant/ Weddell Seal/ Giraffe/ Grey Whale/ Naked Mole Rat/ Okapi/ Yangtze River Dolphin/ Little Brown Bat/ Narwhal/ Rock Hyrax/ Hippo/ Przewalski Horse/ Sperm Whale/ Indo-Pacific Bottlenose Dolphin/ Giant Panda/ Common Bottlenose Dolphin/ Alpaca/ Walrus	
Germline Tumor Suppressor Genes	Max	Naked Mole Rat	2.23	Max	Naked Mole Rat	2.19			
		Human/ Rhesus/ Chimp	1.69		Human/ Chimp/ Rhesus/ Gorilla/	1.69			
		Gorilla	1.66						
	Min	Przewalski Horse/ Horse	1.03	Min	Przewalski Horse/ Horse/ Wild Bactrian Camel	1.03			
		Arabian Camel/ Wild Bactrian Camel	1.031						
		Bactrian Camel							

Gene Type	Normalized Gene Copies			Normalized Non-Conserved Gene Copies (<0.3)			Normalized Conserved Gene Copies (>0.3)		
Somatic Tumor Suppressor Genes	Max	Two Fingered Sloth	2.69	Max	Two Fingered Sloth	2.81	Max	Naked Mole Rat/ Mouse	2.29
		Armadillo	2.61		Armadillo	2.68		Orca/ Rhesus	2.08
		Human	2.57		Human	2.61			
	Min	Rock Hyrax	1.44	Min	Rock Hyrax	1.45	Min	Rock Hyrax	1.23
		Horse	1.61		Horse	1.61		Sperm Whale	1.31
		Opossum	1.70		Opossum	1.70		Common Bottlenose Dolphin/ Large Flying Fox	1.36
Tumor Suppressor Genes and Oncogenes	Max	Two Fingered Sloth	2.18	Max	Two Fingered Sloth	2.34	Max	Rat	2
		Naked Mole Rat	2.12		Naked Mole Rat	2.2		Naked Mole Rat/ Human/ Mouse	1.67
		Opossum	2.07		Armadillo	2.19			
	Min	Wild Yak	1.43	Min	Rock Hyrax	1.48	Min	Polar Bear/ Armadillo/ Two Fingered Sloth/ Tiger	1
		Tiger	1.438		Wild Yak	1.5		Manatee/ Wild Yak	1.08
		Rock Hyrax	1.44		Tiger	1.52		Horse	1.09

Within the categories of tumor suppressor genes, the species *Homo sapiens*, *Heterocephalus glaber* and *Choloepus hoffmanni* have the highest number of paralogues. Since the reference genome used came from *Homo sapiens*, there are expectations for increased gene copies in the species. *Heterocephalus glaber* has little to no cancer rates. In addition to living in a controlled environment, *Heterocephalus glaber's* low cancer rates may be the result of additional tumor suppressor genes. Unlike the *Heterocephalus glaber*, cancer in the *Choloepus hoffmanni* is relatively unknown. The species may have

low cancer rates due to their high copy numbers of tumor suppressor genes in addition to their low metabolic rates.

Within oncogenes, rodents have the highest normalized copy numbers. Along the Rodentia lineage, the species have gained 1773 genes and lost 378 genes (Demuth et al., 2006). Between the separation of *Rattus norvegicus* and *Mus musculus*, *Rattus norvegicus* has gained an additional 235 genes, and *Mus musculus* has gained an additional 843 genes (Demuth et al., 2006). Also, both *Rattus norvegicus* and *Mus musculus* have fast-life histories. They reproduce quickly and are easily susceptible to cancer. The higher oncogene copy numbers may be the cause of their high cancer rates.

Consistently, *Procavia capensis* has the lowest normalized values for all gene categories besides genes that have both tumor suppressor and oncogenic properties. This may be due to the sequencing method of the genome (107x with Illumina). In addition, the low duplication values may be due to the animal's most recent common ancestor distance. The TMRCA between *Homo sapiens* and *Procavia capensis* is 102 million years (Kumar et al., 2017). Therefore, gene duplications may not have been identified. With the use of another reference genome (preferably from the superorder Afrotheria), an increase in copy numbers may occur.

Balance Between Tumor Suppressor Genes and Oncogenes

Phylogenetic regressions were applied between life history data and genetic categories to identify patterns of cancer resistance. A total of 106 phylogenetic regressions (PGLS) were applied to the pipeline results. Significant tests had a P-value < 0.05; 26 tests showed significance. Due to the large number of tests completed, I applied a Bonferroni significance of $P < 0.00047$, and a corrected false discovery rate to each

regression. The Bonferroni correction supported the significance of 14 tests, and the false discovery rate supported 20 tests.

The most significant test compared the relationship between total tumor suppressor genes and total oncogenes. This has a high phylogenetic signal of 0.981, an R_2 of 0.6811, and a P-value of 2.95•10-14. This illustrates that the number of tumor suppressor genes and oncogenes are positively correlated. The relationship between gatekeeper genes and oncogenes (R_2 of 0.6726 and P-value of 5.81•10-14) and caretaker genes with oncogenes (R_2 of 0.4663 and P-value of 1.77•10-08) was also highly significant (Figure 3). Caulin et al., (2015) completed similar tests, which validated the positive association between gatekeeper genes and proto-oncogenes with an R_2 of 0.85, and a P-value < 0.001. Caulin et al., (2015) did not find significance between the caretaker genes and proto-oncogenes. However, their smaller sample size may contribute to the lack of significance. Higher copy numbers of tumor suppressor genes drives copy numbers of oncogenes. Due to the high number of paralogues, these species may have more complex regulatory networks compared to species with lower cancer gene copy numbers. Additionally, species with large duplications of oncogenes may have more tumor suppressor gene paralogues to protect itself from cancer-initiating mutations.

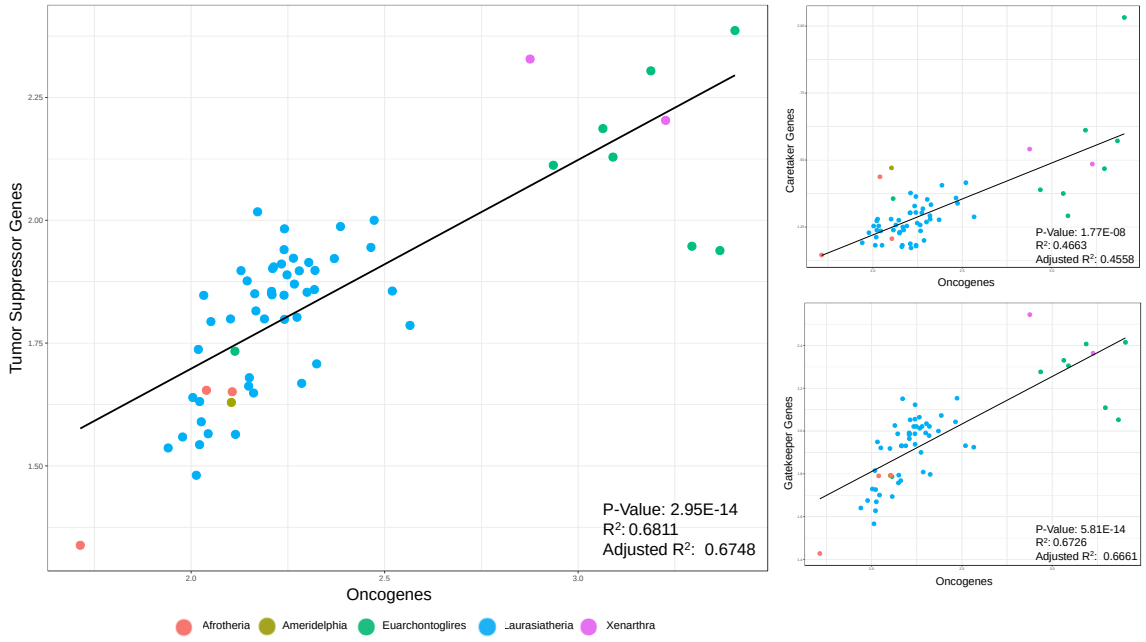

Figure 3. The relationship between tumor suppressor genes and oncogenes. Tumor suppressor genes are further categorized in caretaker genes and gatekeeper genes. The mammals are color coordinated by mammalian superorder.

Application of Phylogenetic Regressions on Life History Data

Regressions were applied to longevity quotient to determine if the combination of body mass and lifespan was significant. A highly significant test compared the relationship between the longevity quotient and tumor suppressor genes with both germline and somatic properties. The relationship demonstrates an R_2 of 0.5379 and a P-value of 6.26•10-10. Significance was also calculated and supported with the false discovery rate between longevity quotient with caretaker genes, germline tumor suppressor genes, and oncogenes. However, under a Bonferroni correction test, they were insignificant.

Longevity quotient and caretaker genes had an R_2 of 0.168, and a P-value of 0.0025. Longevity quotient and germline tumor suppressor genes had a strong phylogenetic signal of 0.985, an R_2 of 0.1614, and a P-value of 0.003159. Longevity quotient with oncogenes also has a strong phylogenetic signal of 0.992, an R_2 of 0.1561, and a P-value of 0.003745. The relationship between longevity quotient with tumor suppressor genes that are both germline and somatic, germline tumor suppressor genes and genes that have both tumor suppressor and oncogenic properties can be found in Figure 4. Under a Bonferroni correction test and a false discovery rate, longevity quotient is not correlated with genes that are both tumor suppressor genes and oncogenes, total tumor suppressor genes, gatekeeper genes, and somatic tumor suppressor genes. Longevity quotient can predict copy numbers for genes (with the most confidence in tumor suppressor genes with both germline and somatic properties) in mammals.

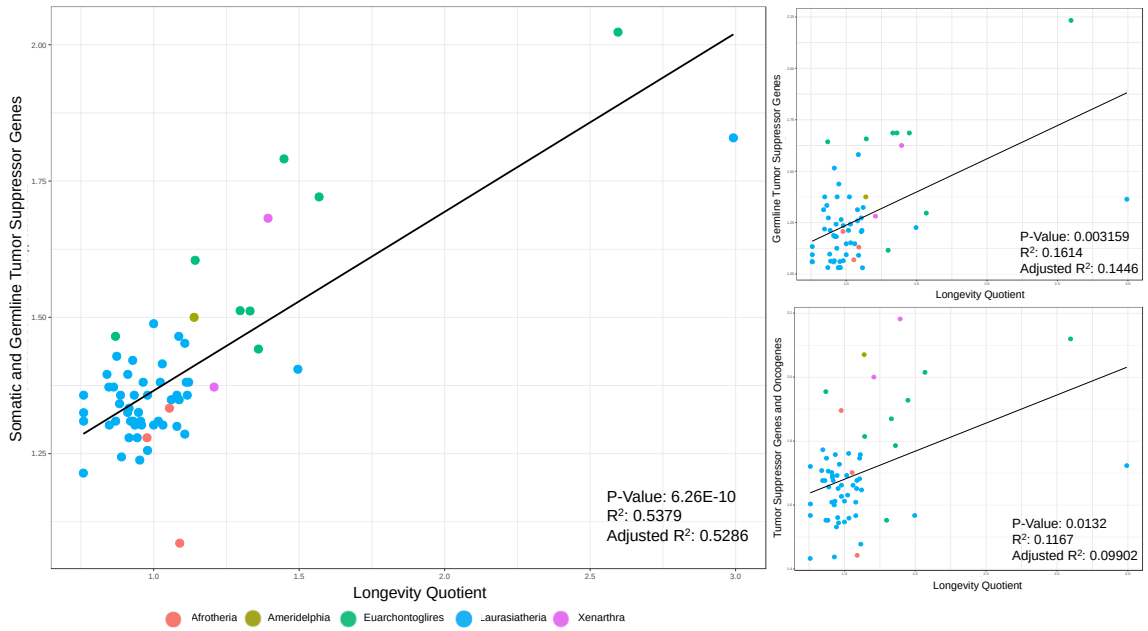

Figure 4. The relationship between longevity quotient and tumor suppressor genes. The mammals are color coordinated by mammalian superorder.

Predictions suggest that large body masses should more tumor suppressor genes since they undergo more cell divisions. To determine if there is a relationship between body mass and tumor suppressor genes, a phylogenetic regression tested log-body mass against tumor suppressor genes that are both germline and somatic. This regression resulted in an R_2 of 0.1785 and a P-value of 0.0016. The log-body mass was also tested against total oncogenes. The relationship had an R_2 of 0.07697 and a P-value of 0.0443. Under a Bonferroni correction test and a false discovery rate test, the results were not significant. However, both tests exhibit a negative correlation between body mass and copy numbers (Figure 5). Generally, when the log-body mass was tested using phylogenetic regressions, there was no positive correlation between mass and copy numbers. Caulin et al., (2015) found similar patterns; where they could not determine a correlation between tumor suppressor genes and body mass. Thus, cancer prevention is

not driven by body mass.

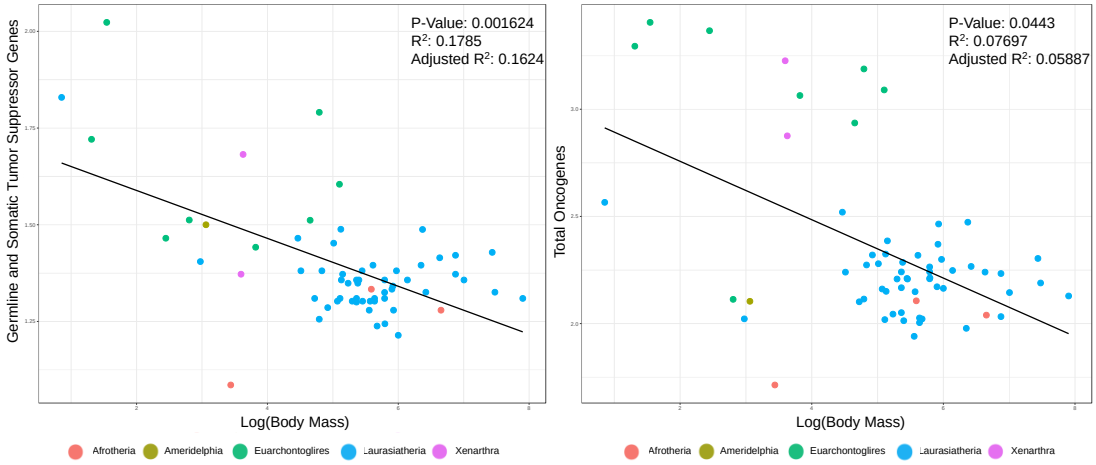

Figure 5. The relationship between log-body mass and gene categories. The mammals are color coordinated by mammalian superorder.

Cetaceans are the largest placental mammals. The study investigated 18 species of cetaceans with a body mass between 32,500g (*Neophocaena asiaeorientalis*) and 8.0•107 (*Balaena mysticetus*). Due to their large body mass, it was expected that these mammals would have large duplications in tumor suppressor genes such as African elephant (*Loxodonta africana*) with TP53. However, large duplications generally did not appear in these species. The pipeline identified high copy numbers for the germline somatic suppressor gene, TPM3. Cetaceans had a range between 7 and 12 paralogues of the gene. The humpback whale (*Megaptera novaeangliae*) has 7 duplications of the gene (Tollis et al., 2019). The bioinformatic pipeline retrieved 7 paralogues of TPM3. However, all species studied had similar copy numbers for TPM3. However, the number of duplications in these genes was relatively average compared to other species. Due to the lack of significant gene duplications within cetaceans, their large body masses and longevity are most likely the result of non-cancer related mechanisms.

Due to the presence of multiple orthologues and paralogues in tumor suppressor genes and oncogenes, I identified the relationship between genome size and cancer gene copy numbers. The genome sizes ranged from 1780.72 Mb (*Camelus bactrianus*) to 3631.52 Mb (*Dasypus novemcinctus*). The length of each genome is found in Table 1. Genomes with longer lengths may have more tumor suppressor genes and oncogenes. However, under a phylogenetic regression, the relationship between cancer genes and genome length was insignificant (Figure 6). The association between tumor suppressor genes and genome length had a P-value of 0.5076 and an R_2 of 0.009815. The connection between oncogenes and genome length had a P-value of 0.608 and an R_2 of 0.005896. These results demonstrate that genome length cannot predict the number of cancer gene paralogues.

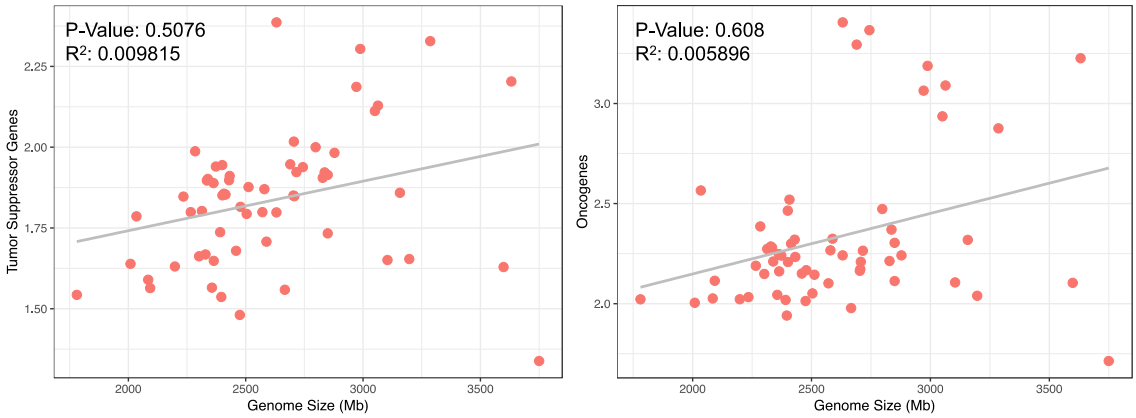

Figure 6. The relationship between genome size and gene categories. Genome size is compared with tumor suppressor genes and oncogenes.

A species' metabolism creates reactive oxygen species (ROS), which may damage DNA (Dang C.V. 2012). Therefore, animals with higher metabolic rates should have mechanisms to reduce the amount of ROS damage due to natural selection. I collected the basal metabolic rates for 25 species to observe if species with high BMR also have a

higher proportion of tumor suppressor genes. As shown in Figure 7, there was no correlation between BMR with tumor suppressor genes and oncogenes. Using a phylogenetic regression (PGLS test), the relationship between metabolic rate and tumor suppressor genes resulted in a P-value of 0.8311, and an R_2 value of 0.002113. Similarly, the relationship between metabolic rate and oncogenes had a P-value of 0.5635 and an R_2 value of 0.01539.

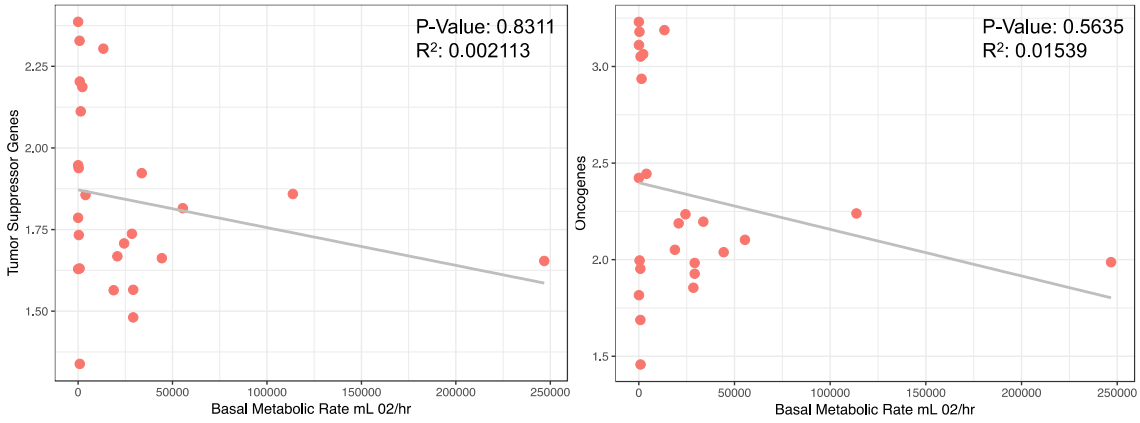

Figure 7. The relationship between basal metabolic rate and gene categories. The BMR was collected for 25 species and compared with total normalized tumor suppressor genes and oncogenes.

The animals studied covered every placental mammal superorder. The relationship between superorders and oncogenes, tumor suppressor genes and oncogenes, tumor suppressor genes that are germline and somatic and total tumor suppressor genes were highly significant (Figure 8). A phylogenetic regression between superorders and tumor suppressor genes and oncogenes resulted in R_2 of 0.7418 and P-value of 1.46•10-13. A phylogenetic regression between superorders and oncogenes resulted in R_2 of 0.5604 and P-value of 3.19•10-8. A regression between superorders and tumor

suppressor genes that are both germline and somatic resulted in R_2 of 0.4584 and P-value of 4.88•10-6. The relationship between superorders and total tumor suppressor genes resulted in R_2 of 0.3497 and a P-value of 0.000307. For the above tests, both Euarchontoglires and Xenarthrans have the highest significance and copy numbers. Phylogenetic regressions with animal order had high significance, due to small numbers of animals in each order, further testing is required to confirm the results.

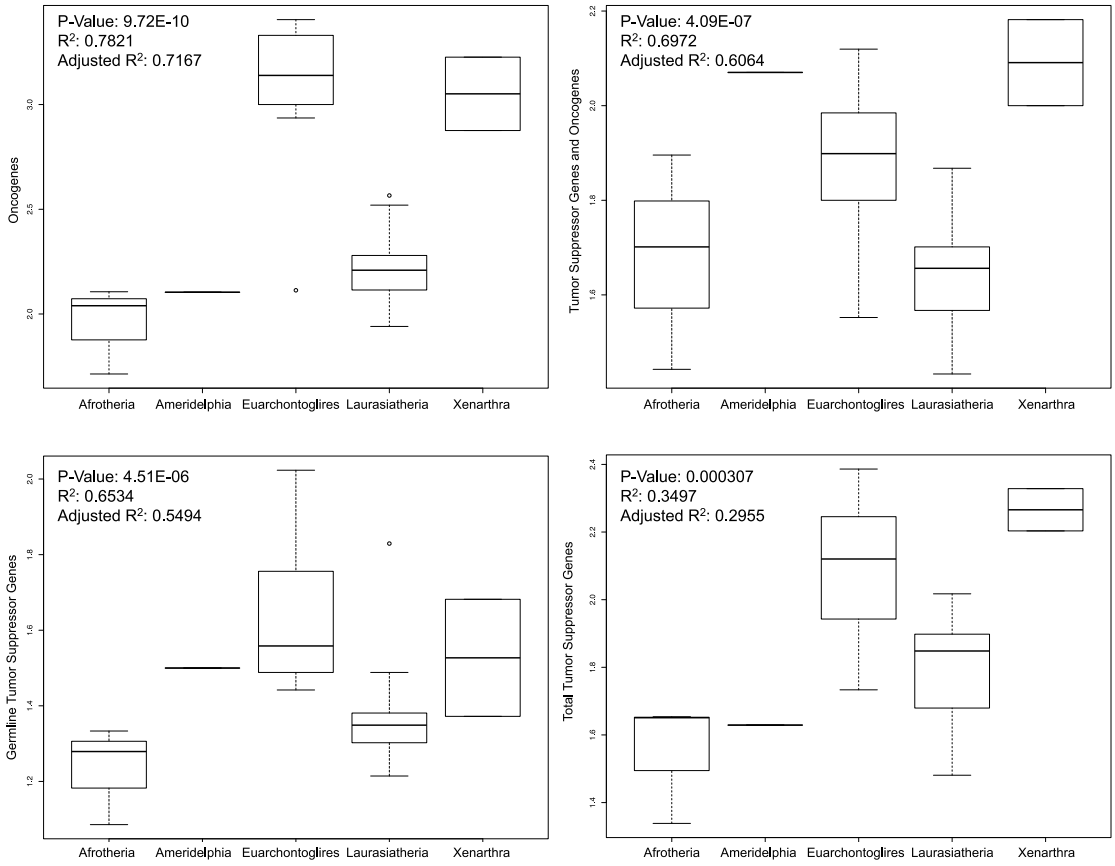

Figure 8. The relationship between gene categories and mammalian superorder.

Life history traits were analyzed in two superorders: Euarchontoglires and Laurasiatherians using ANOVA phylogenetic regressions (Orme et al., 2018). Due to the small sample sizes in the other superorders, they were not tested. A statistically

significant, positive correlation was calculated for total tumor suppressor genes and somatic tumor suppressor genes with log-body mass x lifespan in Euarchontoglires (Figure 9). Total tumor suppressor genes and log-body mass x lifespan had an R_2 of 0.6855 and a P-value of 0.02145. Somatic tumor suppressor genes and log-body mass x lifespan had an R_2 of 0.6193 and a P-value of 0.03547. However, with the application of a Bonferroni correct test and a false discovery rate test, both tests become insignificant. Phylogenetic regressions were also applied in Laurasiatherians; all tests were insignificant. Patterns are observed within superorders, however, life history traits are only significant across the mammals.

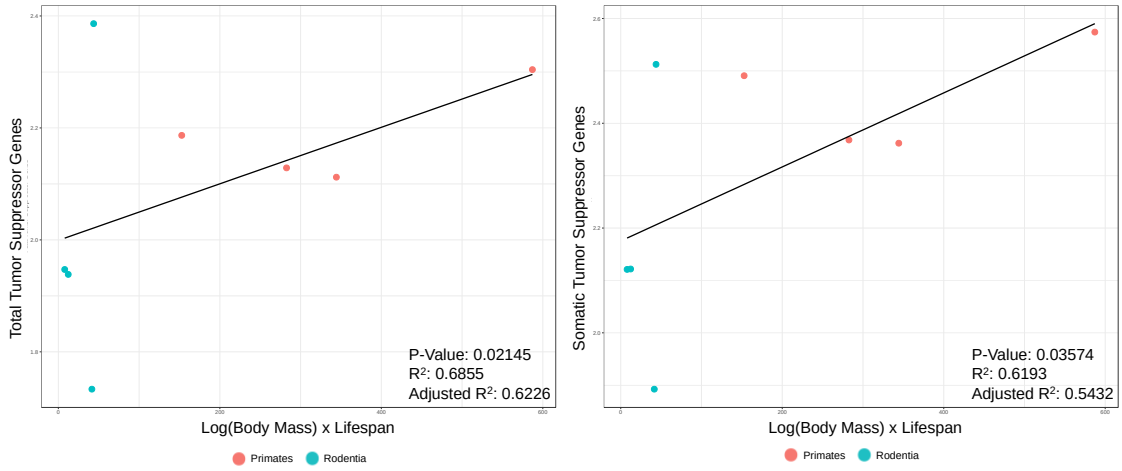

Figure 9. Log-body mass x lifespan in Eauchontoglires. The mammals are color coordinated by mammalian order.

A total of 86 tests between life history traits and copy numbers were insignificant with phylogenetic regressions. The majority of these tests compared the relationship between cancer gene copy numbers and an animal's ecological background. Only one test between diet and genes that were both tumor suppressor and oncogenes was significant (R_2 of 0.1737 and P-value of 0.008491). However, under a Bonferroni correction, the test

becomes insignificant. Significance was found for ecology data without a phylogenetic regression. This was apparent placentation. Researchers speculate that species with hemochorial placentas (invasive placentas) will have higher numbers of tumor suppressor genes and oncogenes. Under a non-phylogenetic model, germline tumor suppressor genes have an R^2 of 0.2649 and a P-value of 0.5301. However, when phylogeny is applied, the R^2 is 0.04951 and the P-value is 0.01346. With a non-phylogenetic model, oncogenes have an R^2 of 0.3761 and a P-value of 0.011354. With phylogeny, the R^2 is 0.1633 and the P-value is 0.1076. An animal's biome, diet, and placentation cannot predict their tumor suppressor gene and oncogene copy numbers.

Neoplasia Rates

Species with low neoplasia rates should have a higher proportion of tumor suppressor genes to protect the genome. Therefore, I investigated if there was a correlation between tumor suppressor genes and oncogenes with neoplasia rates from 27 mammals from the Northwest ZooPath data (Table 1) (Garner et al. 2019). With the use of a phylogenetic model, there was no correlation between cancer gene copy numbers and neoplasia rate (Figure 10). Between tumor suppressor genes and neoplasia rate, there was a P-value of 0.9571 and an R^2 of 0.0001344. Similarly, the P-value between oncogenes and neoplasia rate was 0.4884 and an R^2 of 0.02248. According to the results, duplications of cancer genes cannot predict neoplasia rates.

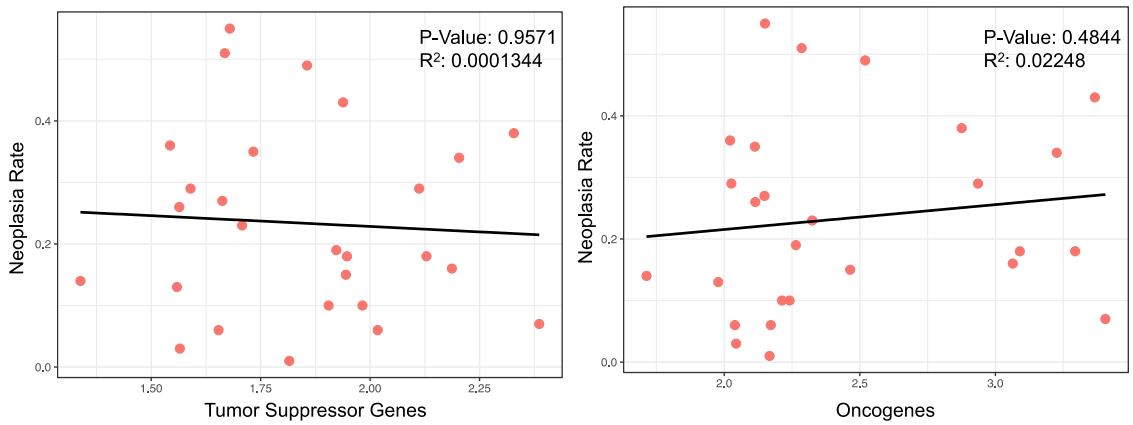

Figure 10. Neoplasia rates compared to tumor suppressor genes and oncogenes. The neoplasia rates are from the Northwest ZooPath database (Garner et al. 2019).

Potential Bias Towards Humans

The human genome is the best-annotated genome. It was used as a reference to find orthologues and paralogues of tumor suppressor genes and oncogenes in mammals. Therefore, there is a potential bias towards humans in my data. This is possible, however, there are non-human primates that also have a high quantity of tumor suppressor genes and oncogenes (Table 6). These animals include the *Heterocephalus glaber* (a rodent), *Choloepus hoffmanni* and *Dasypus novemcinctus* (both Xenarthrans). Both species of Xenarthrans have a 102 MYA time divergence from humans (TMRCA). Therefore, the pipeline was able to identify these desired genes in genomes that are very distant from humans. Additionally, I conducted a linear regression test to see if there was a correlation between tumor suppressor genes and oncogene copy numbers with the time divergence from humans (Figure 11). There was no correlation between cancer genes and TMRCA (TSG P-value of 0.6311 and oncogene P-value of 0.6191). Therefore, cancer copy numbers are not dependent on the distance from humans.

Table 6

Normalized Gene Duplications with Life History Traits

Scientific Name	Body Mass (g)	Max Longevity (y)	Total TSGs	Total Oncogenes	TMRCA (MYA)
Homo sapiens	62035	122.5	366	480	0
Pan troglodytes	44983.5	74	442	498	6.4
Gorilla gorilla	126215.495	55.4	420	507	8.6
Macaca mulatta	6614	40	399	437	28.81
Cavia porcellus	639.1	14.8	372	464	89
Heterocephalus glaber	35	28.3	520	713	89
Mus musculus	20.5	6	358	465	89
Rattus norvegicus	280	5	442	508	89
Ailuropoda melanoleuca	117500	36.8	403	523	94
Arctocephalus gazella	67979.43	30.6	444	479	94
Balaena mysticetus	8.00E+07	211	397	512	94
Balaenoptera acutorostrata	7.50E+06	50	509	687	94
Balaenoptera bonaerensis	7.50E+06	50	423	492	94
Bison bison	624577.07	33.5	448	514	94
Bos indicus	618642.42	20	438	572	94
Bos indicus x Bos taurus	618642.42	20	371	464	94
Bos mutus	1.00E+06	22	387	465	94
Bos taurus	618642.42	20	436	504	94
Bubalus bubalis	827250.485	34.9	468	493	94
Camelus bactrianus	475000	40	513	720	94
Camelus dromedarius	434000	40	446	523	94
Camelus ferus	434000	28.4	390	467	94
Canis lupus	29190.755	29.5	421	475	94
Ceratotherium simum	2.23E+06	50	435	499	94
Delphinapterus leucas	1.38E+06	40	432	510	94
Equus asinus	171249.245	50	348	451	94
Equus caballus	250000	62	553	746	94
Equus przewalskii	360000	36	421	497	94

Scientific Name	Body Mass (g)	Max Longevity (y)	Total TSGs	Total Oncogenes	TMRCA (MYA)
Eschrichtius robustus	2.73E+07	77	419	502	94
Giraffa tippelskirchi	800000	39.5	400	441	94
Hippopotamus amphibius	2.64E+06	61.2	469	544	94
Lagenorhynchus obliquidens	103000	46	445	504	94
Leptonychotes weddellii	410833.335	25	380	457	94
Lipotes vexillifer	83500	24	409	567	94
Megaptera novaeangliae	3.00E+07	95	383	476	94
Mesoplodon bidens	2.35E+06	NA	436	499	94
Monachus monachus	284940.665	23.7	451	507	94
Monachus schauinslandi	197758.36	30	442	728	94
Monodon monoceros	938126.44	50	556	756	94
Myotis lucifugus	7.15	34	430	522	94
Neophocaena asiaeorientalis	32500	33	454	502	94
Neophocaena phocaenoides	141150	53	347	425	94
Odobenus rosmarus	846498.125	40	454	513	94
Okapia johnstoni	230001.14	33.5	442	522	94
Orcinus orca	4.30E+06	100	389	482	94
Panthera tigris	128800	26.3	393	486	94
Phocoena phocoena	52730.93	20.4	389	477	94
Physeter catodon	1.01E+07	100	358	425	94
Pteropus vampyrus	944.685	20.9	440	744	94
Sousa chinensis	279999.99	24.95	527	717	94
Sus scrofa	135000	27	265	341	94
Tursiops aduncus	230000	53	452	549	94
Tursiops truncatus	230000	53	441	489	94
Ursus americanus	132405	34	403	434	94
Ursus arctos	240500	50	305	394	94
Ursus maritimus	371703.81	45	457	557	94
Vicugna pacos	62000	25.8	444	538	94
Choloepus hoffmanni	4250	37	435	524	102

Scientific Name	Body Mass (g)	Max Longevity (y)	Total TSGs	Total Oncogenes	TMRCA (MYA)
Dasypus novemcinctus	3949.01	22.33333333	371	449	102
Loxodonta africana	4.50E+06	80	377	453	102
Procavia capensis	2750	14.80833333	433	487	102
Trichechus manatus	387500	56	446	522	102
Monodelphis domestica	1149.875	7	414	486	160

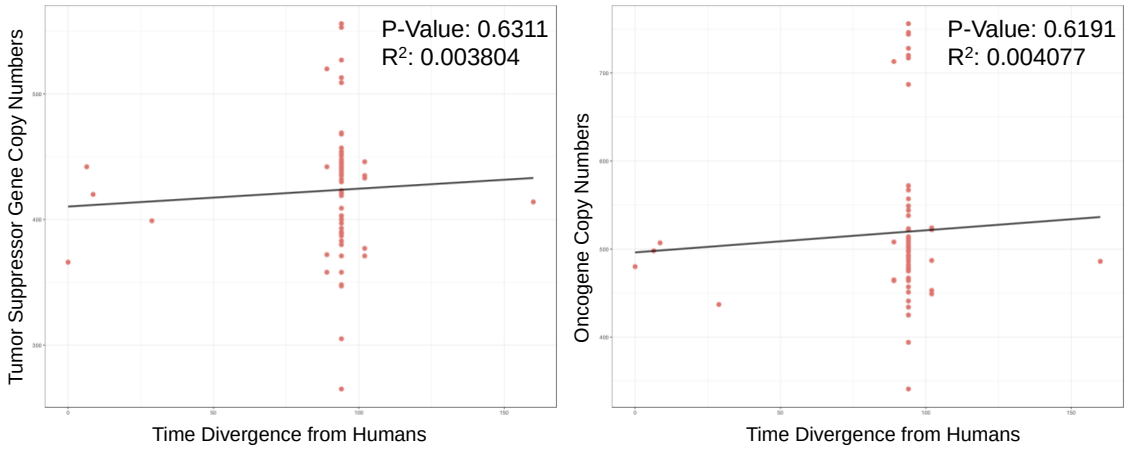

Figure 11. The relationship between human time divergence and cancer genes. The number of tumor suppressor genes and oncogenes are represented as total copy numbers rather than normalized copy numbers.

Additionally, to determine if my results had a bias towards humans, phastCons scores were collected for the 545 genes. PhastCons scores use a phylogenetic Markov model to predict conservation. Siepel et al., research (2005) established a baseline conservation score of 0.3. Genes with a conservation score lower than 0.3 are less conserved. The function of these genes in other mammals may not be the same as it is in humans. Genes with a conservation score higher than 0.3 are highly conserved. These genes are more likely to have orthologues and paralogues for other mammals, and the gene has been favored by natural selection.

The phastCons scores for the 545 genes ranged from a conservation score of 0.00985529 to a conservation score of 0.89897887. The mean phastCons score was 0.17011576. There were 487 genes with a conservation score below 0.3, and 58 genes with a conservation score above 0.3. The majority of the highly conserved genes were oncogenes the distribution of phastCons scores can be found in Figure 12.

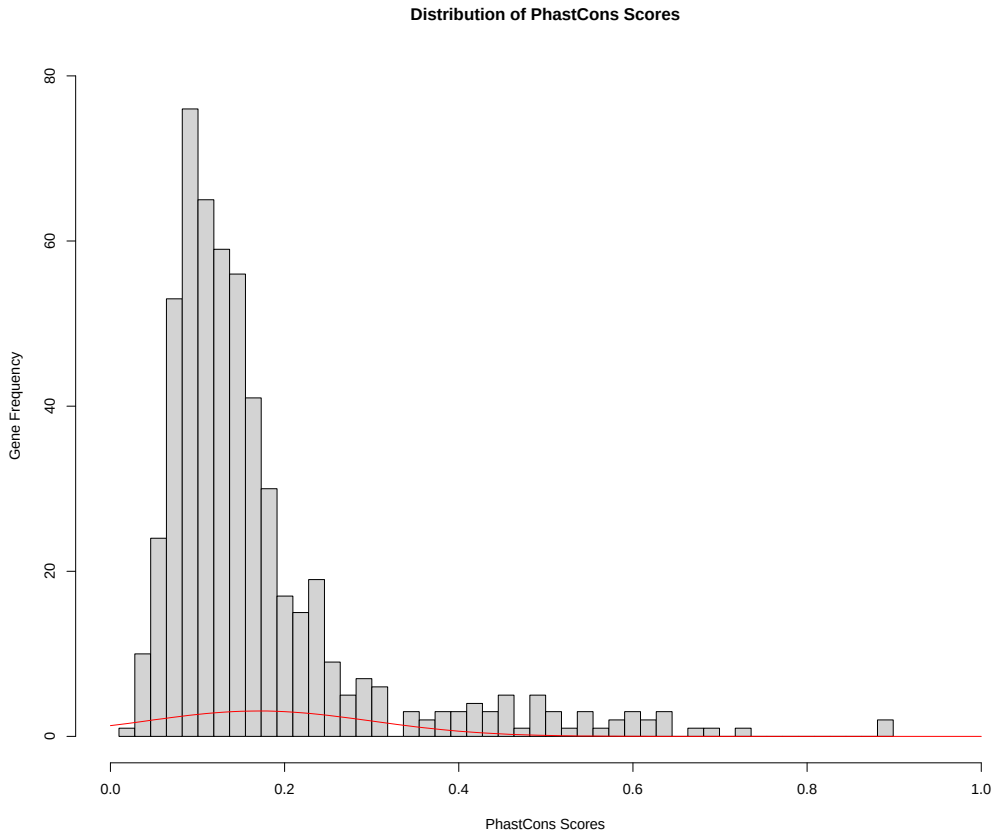

Figure 12. The distribution of phastCons scores.

The low conservation scores favor a bias of tumor suppressor genes and oncogenes towards Homo sapiens. However, *Homo sapiens* have experienced gene gains over the primate lineage. Primates have experienced a net loss of 162 genes. Within the *Homo sapiens* lineage, they have a net gain of 441 genes (Demuth et al., 2006). This net gain is too large to be by chance. Therefore, natural selection has acted to increase

paralogues in *Homo sapiens* (Richard et al., 2008). Other primates such as *Pan troglodytes* do not share the same pattern large numbers of paralogues as *Homo sapiens* do. *Pan troglodytes* have experienced a net loss of 865 genes (Demuth et al., 2006.). Richard et al. (2008), identified 32 Mb of DNA present in *Homo sapiens* that are missing in the genomes of *Pan troglodytes*. These reasons may explain the high tumor suppressor gene and oncogene paralogues in *Homo sapiens*.

The phastCons scores for *Homo sapiens* were compared to the number of cancer gene copy numbers observed from the pipeline (Figure 13). I determined that conserved genes have higher copy numbers per cancer gene (P-value of 0.005157 and R_2 of 0.01434) using a linear regression. Therefore, there is a weak effect between copy numbers and the conservation of tumor suppressor genes and oncogenes.

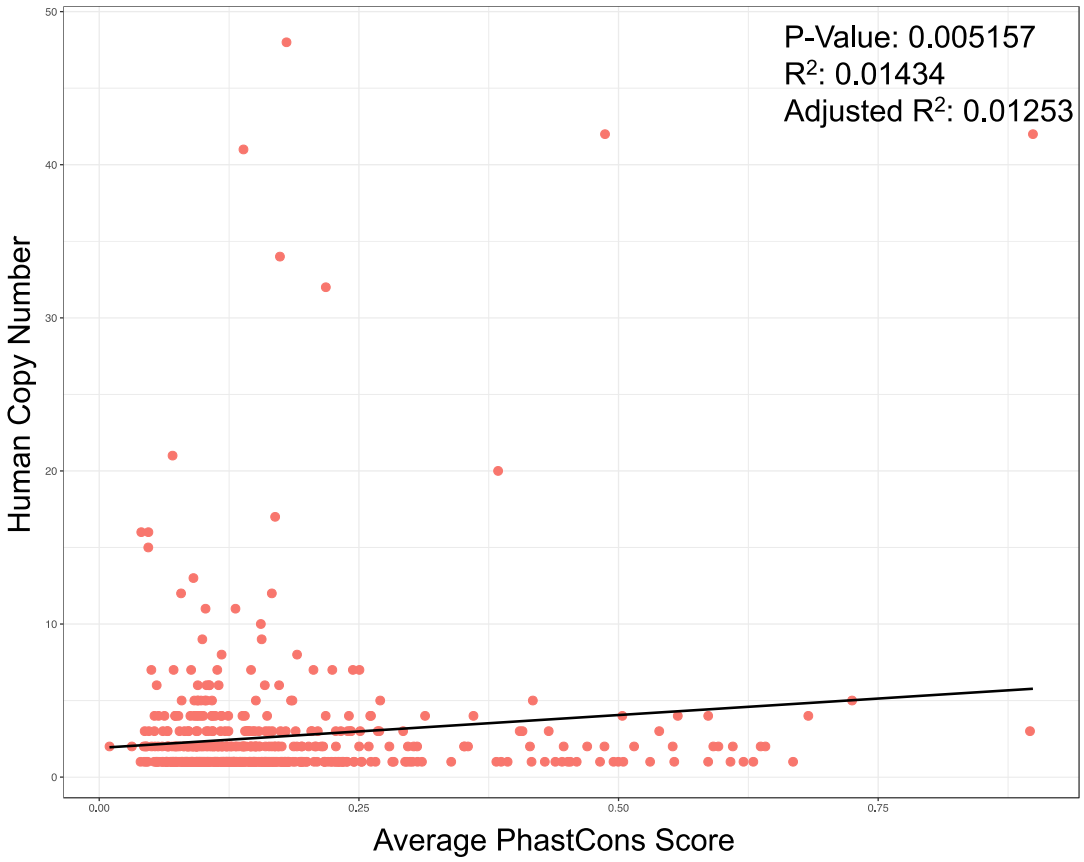

Figure 13. The relationship between phastCons scores and cancer genes in humans. The human copy numbers are not normalized.

Due to the lack of conservation for human cancer genes, I identified if there should be orthologues present for all 545 genes using OrthoDB (Kriventseva et al. 2018). Orthologues were observed in the most distant ancestor studied, *Monodelphis domestica*. For all the genes that the pipeline identified at least one cancer gene copy in *Monodelphis domestica*, OrthoDB also acknowledged the presence of orthologues. If OrthoDB did not have orthologues for *Monodelphis domestica*, the genes were observed in *Loxodonta africana* and *Dasypus novemcinctus*. The presence of these genes aligned with the pipeline results. Therefore, the orthologues and paralogues identified in the pipeline

should be found in the placental mammalian genomes even with the lack of gene conservation.

The results for the normalized gene copies for low conservation (<0.3) mimicked the results without conservation taken into account (Table 5). Genes with low-conservation scores in total tumor suppressor genes were *Heterocephalus glaber* (2.41), *Choloepus hoffmanni* (2.39), and *Homo sapiens* (2.33). *Procavia capensis* had the lowest duplications of tumor suppressor genes (1.35). These results mimic those of normalized data without specification of conservation. Species with the highest copy numbers for caretaker genes were *Heterocephalus glaber* (1.95), *Choloepus hoffmanni* (1.57) and *Homo sapiens* (1.61). *Procavia capensis* (1.12) had the least copy numbers for caretaker genes. Large duplications within gatekeeper genes were identified in *Choloepus hoffmanni* (2.63), *Heterocephalus glaber* (2.492) and *Dasypus novemcinctus* (2.49). The species with the lowest number of paralogues in gatekeeper genes were *Procavia capensis* (1.45). The animals with the largest number of oncogenes are unlike the animals without conservation taken into account. *Heterocephalus glaber* (2.99), *Homo sapiens* (2.89) and *Dasypus novemcinctus* (2.83) have the most oncogenes. *Procavia capensis* (1.56) possess the lowest number of oncogenes, which is comparable to copy numbers in all oncogenes.

Genes with high conservation were mainly oncogenes. The species with the highest number of total tumor suppressor genes were *Mus musculus* (2.25), *Heterocephalus glaber* (2.13) and Orcinus orca (2.07). Similarly, to the total normalized data and the non-conserved genes, *Procavia capensis* (1.2) had the lowest number of tumor suppressor copies. The species Homo sapiens (3.25), *Heterocephalus glaber* (3.0)

and *Mus musculus* (3.0) have the highest copy numbers for caretaker genes. Duplications were not observed in caretaker genes for *Choloepus hoffmanni*. The species with the highest normalized gatekeeper genes were *Rattus norvegicus* (1.74), *Mus musculus* (1.7), *Orcinus orca* (1.68) and *Delphinapterus leucas* (1.68). The species, *Ursus maritimus* (1.15), had the least amount of gene duplications. The species with the highest number of oncogenes are all from the Rodentia order. *Rattus norvegicus* (8.47) has the most duplications, followed by *Mus musculus* (7.93) and *Heterocephalus glaber* (6.17). Again, *Procavia capensis* (2.84) has the least number of paralogues. From the population of 58 conserved genes, zero genes were classified as germline tumor suppressor genes. Therefore, the highest and lowest normalized copy numbers are not calculated for this category (Table 5). Normalized values are higher in the conserved genes due to the presence of genes with 10 or more paralogues. The conserved genes make up 10.6% of all the genes studied. Therefore, the general population of tumor suppressor genes and oncogenes are not conserved across mammals.

Housekeeping Genes

Housekeeping genes are expressed in all tissues and are responsible for cell preservation (Eisenberg et al. 2013). Some housekeeping genes are also considered tumor suppressor genes. Due to their role in cell maintenance, animals with low cancer rates should have more copies of housekeeping genes. A phylogenetic regression demonstrated that housekeeping genes are not driven by longevity (P-value of 0.3025 and an R_2 of 0.02125). This pattern is apparent in Figure 14. There was also no relationship between housekeeping genes and body mass (P-value of 0.1287 and an R_2 of 0.04555). The

patterns observed in cancer genes are different from other sets of genes, including housekeeping genes.

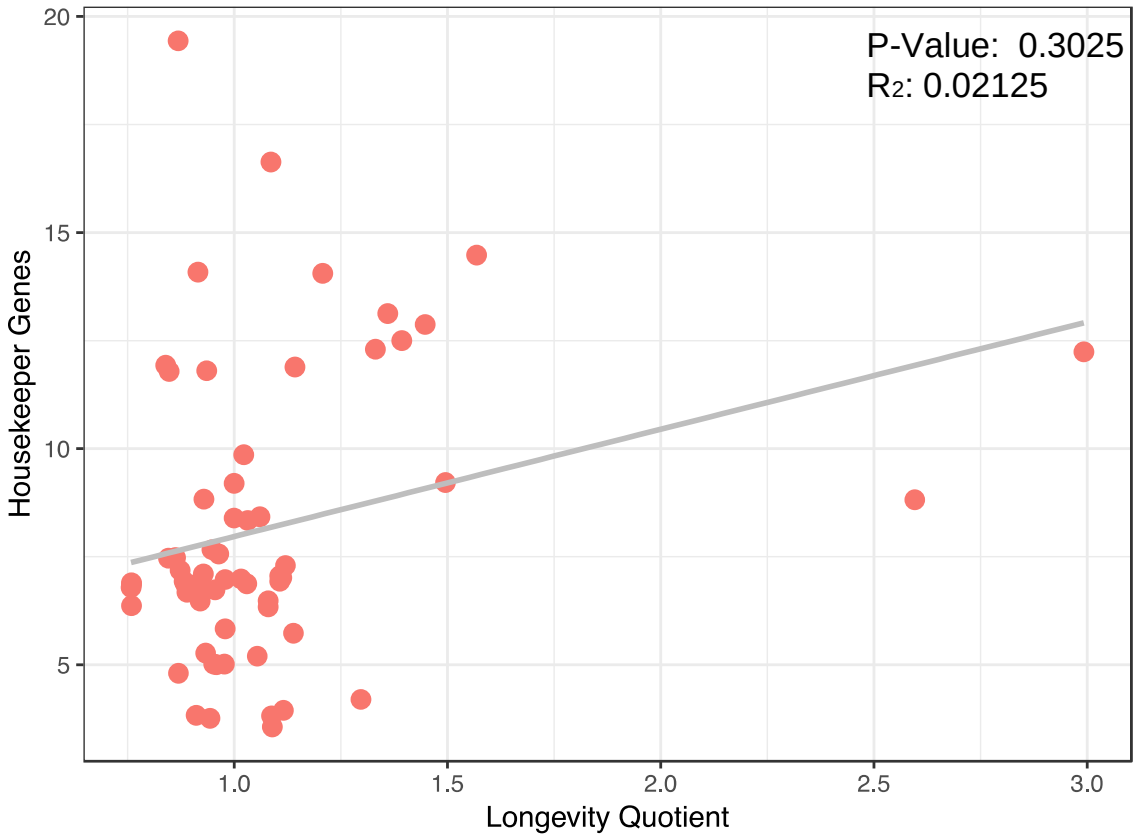

Figure 14. The relationship between longevity quotient and housekeeping genes. The relationship is calculated using a phylogenetic regression (PGLS).

CHAPTER 4

CONCLUSION

The purpose of the study was to test if patterns of gene duplications in tumor suppressor genes and oncogenes are associated with life history traits across mammals. To do this, I had to: 1) create a bioinformatic pipeline which could locate and validate orthologues and paralogues of tumor suppressor genes and oncogenes; 2) use phylogenetic regressions to determine if a species' life history and ecology links the number of gene copies; and used the results to 3) identify specific animals or life history traits that should be investigated in more depth for cancer resistance.

There is a strong positive correlation between tumor suppressor genes and oncogenes. Body mass alone cannot predict the number of tumor suppressor genes and oncogenes within a species. However, the longevity quotient can predict copy numbers in mammals. Species that have a high longevity quotient (live longer than expected for body mass) have a larger number of new tumor suppressor genes. With the use of phylogenetic regressions, I found that genes that are both germline and somatic tumor suppressor genes were the most likely to be duplicated in mammals. This suggests that there has been selection for those genes associated with selection for extending longevity, which likely requires increasing cancer resistance. I also observed higher selection on germline tumor suppressor genes compared to somatic tumor suppressor genes, perhaps because germline mutations are likely to have a larger effect on organismal fitness than a mutation in a somatic cell. My results suggest that cancer resistance should be further studied in *Choloepus hoffmanni* (two-fingered sloth), *Dasypus novemcinctus* (nine-banded armadillo) and *Heterocephalus glaber* (naked mole rat). Little research has been

conducted on cancer in the Xenarthran, *Choloepus hoffmanni*. A paper by Higginbotham et al., has determined that sloth hair carries an anti-cancer fungus (2014). However, no studies have been published to investigate the high number of tumor suppressor gene paralogues in *Choloepus hoffmanni*. The species, *Dasypus novemcinctus*, is another species from the superorder Xenarthra. Like *Choloepus hoffmanni*, the species has large numbers of paralogues in tumor suppressor genes in non-conserved genes. Cancer has been recognized in this species, but has not been extensively studied (Lee et al., 2015). In contrast to the two Xenarthrans, *Heterocephalus glaber* is justifiably famous for its cancer resistance. The excess number of tumor suppressor genes has not been appreciated until now, but these fascinating animals should be further studied to understand how those genes are, or are not, contributing to its cancer resistance.

The genes studied are known tumor suppressor genes and oncogenes in the human genome. However, these genes may not be cancer genes in the other animals I investigated. Further investigation would be needed to identify if these genes are involved in cancer in other mammals. Additionally, due to the genome assembly and coverage, the pipeline may have missed some gene copies or found gene copies that do not exist. Also, due to the shorter peptide sequences of oncogenes, the pipeline may have identified more paralogues of oncogenes compared to tumor suppressor genes. This may be resolved with the addition of more parameters in the pipeline.

My results suggest that: 1) longevity quotient can predict gene copy numbers in mammals; 2) natural selection favors duplication of tumor suppressor genes that act to prevent cancer in both the germline and somatic cells; 3) tumor suppressor genes and oncogenes are not conserved across mammals; and 4) species within the superorder

Xenarthra and *Heterocephalus glaber* should be further researched for possible cancer prevention methods.

REFERENCES

Abegglen, L. M., Caulin, A. F., Chan, A., Lee, K., Robinson, R., Campbell, M. S., ... & Jensen, S. T. (2015). Potential mechanisms for cancer resistance in elephants and comparative cellular response to DNA damage in humans. *Jama*, *314*(17), 1850-1860.

Benirschke, K. (2007, August 27). Domestic Dog: Canis familiaris. Retrieved from http://placentation.ucsd.edu/dogfs.htm.

Benirschke, K. (2007, June 19). Giraffe: Giraffa camelopardalis. Retrieved from http://placentation.ucsd.edu/giraffefs.htm.

Benirschke, K. (2007, March 21). Domestic Horse: Equus caballus. Retrieved from http://placentation.ucsd.edu/horsefs.htm.

Benirschke, K. (2007, March 21). Elephants: Elephas maximus & Loxodonta africana. Retrieved from http://placentation.ucsd.edu/elephfs.htm.

Benirschke, K. (2007, March 21). Okapi: Okapia johnstoni. Retrieved from http://placentation.ucsd.edu/okapifs.htm.

Benirschke, K. (2007, March 21). Water Buffalo: Bubalus bubalis. Retrieved from http://placentation.ucsd.edu/waterbuffalofs.htm.

Benirschke, K. (2008, April 6). Camelidae (Bactrian camel, dromedary, guanaco, llama, vicuña, alpaca): Camelus bactrianus, C. dromedarius, Lama guanicoe, L. glama, L. pacos, Vicugna vicugna. Retrieved from http://placentation.ucsd.edu/camfs.htm.

Benirschke, K. (2008, February 4). Rock hyrax - Rock dassie (rabbit) - Cony: Procavia capensis. Retrieved from http://placentation.ucsd.edu/hyraxfs.htm.

Benirschke, K. (2010, July 19). East African River Hippopotamus: Hippopotamus amphibius kiboko. Retrieved from http://placentation.ucsd.edu/hippofs.htm.

Benirschke, K. (2010, November 21). Rhesus Monkey, and some other Cercopithecidae: Macaca mulatta. Retrieved from http://placentation.ucsd.edu/macfs.html.

Benirschke, K. (2011, June 12). Walrus: Odobenus rosmarus (divergens). Retrieved from http://placentation.ucsd.edu/walrusfs.htm.

Benirschke, K. (2011, September 21). Killer Whale: Orcinus orca. Retrieved from http://placentation.ucsd.edu/killerwhalefs.htm.

Benirschke, K. (2011, September 21). Lowland Gorilla: Gorilla Gorilla. Retrieved from http://placentation.ucsd.edu/lowland_gorilla_fs.htm.

Bhagwat, M., Young, L., & Robinson, R. R. (2012). Using BLAT to find sequence similarity in closely related genomes. Current protocols in bioinformatics, 37(1), 10-8.

Boddy, A. M., Kokko, H., Breden, F., Wilkinson, G. S., & Aktipis, C. A. (2015). Cancer susceptibility and reproductive trade-offs: a model of the evolution of cancer defences. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *370*(1673), 20140220.

Boratyn GM, Schäffer AA, Agarwala R, Altschul SF, Lipman DJ, & Madden T.L. (2012) "Domain enhanced lookup time accelerated BLAST." Biol Direct. 2012 Apr 17;7:12. PubMed

Buffenstein, R. (2005). The naked mole-rat: a new long-living model for human aging research. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, *60*(11), 1369-1377.

Buffenstein, R. (2008). Negligible senescence in the longest living rodent, the naked mole-rat: insights from a successfully aging species. *Journal of Comparative Physiology B*, *178*(4), 439-445.

Caulin, A. F., & Maley, C. C. (2011). Peto's Paradox: evolution's prescription for cancer prevention. Trends in ecology & evolution, 26(4), 175-182.

Caulin, A. F., Graham, T. A., Wang, L. S., & Maley, C. C. (2015). Solutions to Peto's paradox revealed by mathematical modelling and cross-species cancer gene analysis. Phil. Trans. R. Soc. B, 370(1673), 20140222.

Dang C. V. (2012). Links between metabolism and cancer. *Genes & development*, *26*(9), 877–890. doi:10.1101/gad.189365.112

Demuth, J. P., De Bie, T., Stajich, J. E., Cristianini, N., & Hahn, M. W. (2006). The evolution of mammalian gene families. PloS one, 1(1), e85.

Dewey, T., Shefferly, N., & Havens, A. (2010). Animal Diversity Web. University of Michigan Museum of Zoology.

Dorus, S., Vallender, E. J., Evans, P. D., Anderson, J. R., Gilbert, S. L., Mahowald, M., ... & Lahn, B. T. (2004). Accelerated evolution of nervous system genes in the origin of Homo sapiens. *Cell*, *119*(7), 1027-1040.

Eisenberg, E., & Levanon, E. Y. (2013). Human housekeeping genes, revisited. *TRENDS in Genetics*, *29*(10), 569-574.

Ferlay, J., Soerjomataram, I., Dikshit, R., Eser, S., Mathers, C., Rebelo, M., ... & Bray, F. (2015). Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *International journal of cancer*, *136*(5), E359-E386.

Foley, N. M., Hughes, G. M., Huang, Z., Clarke, M., Jebb, D., Whelan, C. V., ... & Ransome, R. D. (2018). Growing old, yet staying young: The role of telomeres in bats' exceptional longevity.Science advances, 4(2), eaao0926.

Fritz, S. A., Bininda-Emonds, O. R., & Purvis, A. (2009). Geographical variation in predictors of mammalian extinction risk: big is bad, but only in the tropics. *Ecology letters*, *12*(6), 538-549.

Garner, M. M., & LaDouceur, E. E. B. (11AD, November 2019). Northwest ZooPath. Retrieved from http://www.zoopath.com/About Us.htm.

Gumal, M., Jamahari, S., Irwan, M., JANTAN-BRANDAH, C., KAMAL, M., & RAZAK-PAWI, A. (1998). The ecology and role of the large flying fox (Pteropus vampyrus) in Sarawakian rain forests–1997 report. *Hornbill*, *1*, 32-47.

Higginbotham, S., Wong, W. R., Linington, R. G., Spadafora, C., Iturrado, L., & Arnold, A. E. (2014). Sloth hair as a novel source of fungi with potent anti-parasitic, anti-cancer and anti-bacterial bioactivity. PloS one, 9(1), e84549.

HMGA2 (2019). Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information. Available from: https://www.ncbi.nlm.nih.gov/gene/8091#gene-expression

HNRNPA2B1 (2019) Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information. Available from: https:// https://www.ncbi.nlm.nih.gov/gene/3181#gene-expression

Hunt, S. E., McLaren, W., Gil, L., Thormann, A., Schuilenburg, H., Sheppard, D., ... & Cunningham, F. (2018). Ensembl variation resources. Database, 2018.

Jones, K. E., Bielby, J., Cardillo, M., Fritz, S. A., O'Dell, J., Orme, C. D. L., ... & Connolly, C. (2009). PanTHERIA: a species-level database of life history, ecology, and geography of extant and recently extinct mammals: Ecological Archives E090-184. Ecology, 90(9), 2648-2648.

Karolchik, D., Hinrichs, A. S., Furey, T. S., Roskin, K. M., Sugnet, C. W., Haussler, D., & Kent, W. J. (2004). The UCSC Table Browser data retrieval tool. *Nucleic acids research*, *32*(suppl_1), D493-D496.

Keane, M., Semeiks, J., Webb, A. E., Li, Y. I., Quesada, V., Craig, T., ... & Michalak, P. (2015). Insights into the evolution of longevity from the bowhead whale genome. Cell reports, 10(1), 112-122.

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., & Haussler, D. (2002). The human genome browser at UCSC. *Genome research*, *12*(6), 996-1006.

Kent, W. J. (2002). BLAT—the BLAST-like alignment tool. *Genome research*, *12*(4), 656-664.

Kraus, C., Thomson, D. L., Künkele, J., & Trillmich, F. (2005). Living slow and dying young? Life-history strategy and age-specific survival rates in a precocial small mammal. Journal of Animal Ecology, 74(1), 171-180.

Kriventseva, E. V., Kuznetsov, D., Tegenfeldt, F., Manni, M., Dias, R., Simão, F. A., & Zdobnov, E. M. (2018). OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologues. *Nucleic acids research*, *47*(D1), D807-D811.

Kumar, S., Stecher, G., Suleski, M., & Hedges, S. B. (2017). TimeTree: a resource for timelines, timetrees, and divergence times. *Molecular biology and evolution*, *34*(7), 1812-1819.

Lahmann, P. H., Hughes, M. C. B., Williams, G. M., & Green, A. C. (2016). A prospective study of measured body size and height and risk of keratinocyte cancers and melanoma. *Cancer epidemiology*, *40*, 119-125.

Lee, B. R., Oh, S., Lee, S. H., Kim, Y., Youn, S., Kim, Y., ... & Kim, D. Y. (2015). Squamous cell carcinoma in a nine-banded armadillo (dasypus novemcinctus). Journal of Zoo and Wildlife Medicine, 46(2), 333-334.

Lewis, K. N., Mele, J., Hornsby, P. J., & Buffenstein, R. (2012). Stress resistance in the naked mole-rat: the bare essentials–a mini-review. Gerontology, 58(5), 453-462.

Martineau, D., Lemberger, K., Dallaire, A., Labelle, P., Lipscomb, T. P., Michel, P., & Mikaelian, I. (2002). Cancer in wildlife, a case study: beluga from the St. Lawrence estuary, Québec, Canada. *Environmental health perspectives*, *110*(3), 285-292.

McDonald, J.H. (2014). Handbook of Biological Statistics (3rd ed.). Sparky House Publishing, Baltimore, Maryland.

Mossman, H. (1987). Vertebrate fetal membranes : Comparative ontogeny and morphology, evolution, phylogenetic significance, basic functions, research opportunities. New Brunswick, N.J.: Rutgers University Press.

MUC4. (2019). Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information. Available from: https:// https://www.ncbi.nlm.nih.gov/gene/4585

O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., ... & Astashyn, A. (2015). Reference sequence (RefSeq) database at NCBI: current

status, taxonomic expansion, and functional annotation. *Nucleic acids research*, *44*(D1), D733-D745.

Orme, D. et al., caper: Comparative Analyses of Phylogenetics and Evolution in R. R package version 1.0.1. (2018). Available at: https://CRAN/R-project.org/package=caper. (Accessed: 26 September 2018)

Paoloni, M. C., & Khanna, C. (2007). Comparative oncology today. The Veterinary clinics of North America. Small animal practice, 37(6), 1023-32; v.

Peto, R., Roe, F. J., Lee, P. N., Levy, L., & Clack, J. (1975). Cancer and aging in mice and men. *British journal of cancer*, *32*(4), 411.

Reinhardt, H. C., & Schumacher, B. (2012). The p53 network: cellular and systemic DNA damage responses in aging and cancer. *Trends in Genetics*, *28*(3), 128-136.

Richard, G. F., Kerrest, A., & Dujon, B. (2008). Comparative genomics and molecular dynamics of DNA repeats in eukaryotes. Microbiol. Mol. Biol. Rev., 72(4), 686-727.

Sayres, M. A. W., Venditti, C., Pagel, M., & Makova, K. D. (2011). Do variations in substitution rates and male mutation bias correlate with life-history traits? A study of 32 mammalian genomes. *Evolution: International Journal of Organic Evolution*, *65*(10), 2800-2815.

Schneider-Utaka, A. K. (2018). Retrieved from https://repository.asu.edu/attachments/208964/content/Schneider-Utaka_A_Fall_2018.pdf

Seim, I., Fang, X., Xiong, Z., Lobanov, A. V., Huang, Z., Ma, S., ... & Gerashchenko, M. V. (2013). Genome analysis reveals insights into physiology and longevity of the Brandt's bat Myotis brandtii. *Nature communications*, *4*, 2212.

Seluanov, A., Gladyshev, V. N., Vijg, J., & Gorbunova, V. (2018). Mechanisms of cancer resistance in long-lived mammals. Nature reviews. Cancer, 18(7), 433–441. doi:10.1038/s41568-018-0004-9

Sieg, A. E., O'Connor, M. P., McNair, J. N., Grant, B. W., Agosta, S. J., & Dunham, A. E. (2009). Mammalian metabolic allometry: do intraspecific variation, phylogeny, and regression models matter?. *The American Naturalist*, *174*(5), 720-733.

Siegel, R. L., Miller, K. D., & Jemal, A. (2019). Cancer statistics, 2019. CA: a cancer journal for clinicians, 69(1), 7-34.

Siepel A and Haussler D (2005). Phylogenetic hidden Markov models. In R. Nielsen, ed., Statistical Methods in Molecular Evolution, pp. 325-351, Springer, New York.

Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M., Rosenbloom, K., ... & Weinstock, G. M. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome research*, *15*(8), 1034-1050.

Stelzer, G., Rosen, N., Plaschkes, I., Zimmerman, S., Twik, M., Fishilevich, S., ... & Kaplan, S. (2016). The GeneCards suite: from gene data mining to disease genome sequence analyses. *Current protocols in bioinformatics*, *54*(1), 1-30.

Sulak, M., Fong, L., Mika, K., Chigurupati, S., Yon, L., Mongan, N. P., ... & Lynch, V. J. (2016). TP53 copy number expansion is associated with the evolution of increased body size and an enhanced DNA damage response in elephants. Elife, 5, e11994.

Tate, J. G., Bamford, S., Jubb, H. C., Sondka, Z., Beare, D. M., Bindal, N., ... & Fish, P. (2018). COSMIC: the catalogue of somatic mutations in cancer. Nucleic acids research, 47(D1), D941-D947.

Tjalma, R. A. (1966). Canine bone sarcoma: estimation of relative risk as a function of body size. *Journal of the National Cancer Institute*, *36*(6), 1137-1150.

Tollis, M., Robbins, J., Webb, A. E., Kuderna, L. F., Caulin, A. F., Garcia, J. D., ... & Palsbøll, P. J. (2019). Return to the sea, get huge, beat cancer: an analysis of cetacean genomes including an assembly for the humpback whale (Megaptera novaeangliae). Molecular biology and evolution.

Tollis, M., Boddy, A. M., & Maley, C. C. (2017). Peto's Paradox: how has evolution solved the problem of cancer prevention?. *BMC biology*, *15*(1), 60.

Tollis, M., Schiffman, J. D., & Boddy, A. M. (2017). Evolution of cancer suppression as revealed by mammalian comparative genomics. *Current opinion in genetics & development*, *42*, 40-47.

Vicens, A., & Posada, D. (2018). Selective pressures on human cancer genes along the evolution of mammals. *Genes*, *9*(12), 582.

White, M. C., Holman, D. M., Boehm, J. E., Peipins, L. A., Grossman, M., & Henley, S. J. (2014). Age and cancer risk: a potentially modifiable relationship. *American journal of preventive medicine*, *46*(3 Suppl 1), S7–S15. doi:10.1016/j.amepre.2013.10.029

Wolfram Research (2016). Amniote Life History Database. Wolfram Data Repository. https://doi.org/10.24097/wolfram.48117.data

APPENDIX A

PIPELINE RESULTS FOR CANCER GENES

CONSULT ATTATCHED EXCEL FILE

Appendix A contains the pipeline results with cancer genes. The spreadsheet includes life history data, ecology data and the number of tumor suppressor genes and oncogenes for 63 mammals. Each animal has information regarding their common name, scientific name, superorder, order, time divergence from humans (TMRCA), body mass, log-body mass, longevity, log-body mass x lifespan, longevity quotient, basal metabolic rate, habitat, biome, diet, placentation and neoplasia rate. Appendix A also contains the normalized values for caretaker genes, gatekeeper genes, total tumor suppressor genes, germline and somatic tumor suppressor genes, germline tumor suppressor genes, somatic tumor suppressor genes, oncogenes and genes that have both tumor suppressor gene and oncogene properties.

The gene type is color-coordinated. Genes with a dark green header column are considered total tumor suppressor genes. Genes with a medium green header column are considered germline and somatic tumor suppressor genes. Genes with a light green header column are considered somatic tumor suppressor genes. Genes with a blue header column are considered germline tumor suppressor genes. Genes with a yellow header column are considered oncogenes. Genes with an orange header column are genes with both tumor suppressor gene and oncogene properties.

Within the individual gene columns, white boxes are gene copy numbers from the 70% protein identity pipeline run, and red boxes are gene coy numbers from the 65% protein identity pipeline run. For PRDM1 and QKI, gene copy numbers were manually found using Ensembl and UCSC genome browser (Hunt et al. 2018; Kent et al. 2002). Orange boxes correspond to gene copies collected from Ensembl, and blue boxes correspond to copies collected from the UCSC genome browser.

APPENDIX B

PHYLOGENETIC REGRESSIONS WITH CORRECTION TESTS

CONSULT ATTATCHED EXCEL FILE

Appendix B contains the results of 106 phylogenetic regressions. The spreadsheet contains the name of the test, the lambda value, the R2 value, the adjusted R2 value, the P-value, the significance of the P-value under a Bonferroni correction test ($P < 0.00047$), the significance of the P-value under a false discovery rate test, and the corrected P-value using a false discovery rate test. An abbreviated version of the table with significant tests is found in Table 2.

APPENDIX C

PIPELINE RESULTS FOR HOUSEKEEPING GENES

CONSULT ATTATCHED EXCEL FILE

Appendix C contains the pipeline results with housekeeping genes. The spreadsheet includes the common name, scientific name, body mass, log-body mass, longevity, log-body mass x longevity, and longevity quotient for 63 mammals. The spreadsheet also includes the normalized number of housekeeping genes, and the gene copies identified for each housekeeping gene.