Pervasive duplication of tumor suppressor genes preceded parallel evolution of large bodied Atlantogenatans

Cancer is an intrinsic disease of multicellular organisms. Within a species, the size of an animal, - correlated with the individual’s number of cells - and its lifespan - correlated with increasing cellular damage over time - positively correlate with the risk any individual has to form tumors. Between species, however, we do not observe any correlation between size, lifespan, and cancer, a phenomenon that referred to as *Peto’s Paradox*. Elephants are a particularly interesting member of this class of paradoxical animals, since they are a set of large species deeply nested in a clade of smaller species, indicating a recent gain of size. Recent work has identified several individual cases of gene duplicates contributing to the increased cancer resistance of elephants, which suggests that duplication of tumor suppressor genes may play a more general role in mediating Peto’s Paradox by increasing cancer resistance in large, long-lived species. By using a Reciprocal Best-Hit BLAT search approach, we investigated copy numbers of all protein-coding genes in *Atlantogenatan* genomes to see if there is any correlation between the copy number of duplicates and changes body size along the phylogenetic tree. From an initial set of 18,011 protein-coding genes in hg38, we identified a median of 13,880 genes in *Atlantogenatan* genomes, of which a median of 940 genes are duplicated. We find that, just as body size fluctuates throughout *Atlantogenata*, genes involved in tumor suppressor pathways are also duplicated throughout the phylogenetic tree. Extant species of elephants, however, show active transcription of both canonical and duplicated copies of tummor suppressors that duplicated prior to and during their sudden increase in body size, suggesting that the duplication of tumor suppressor genes facilitates the evolution of increased body size by compensating for the increased cancer risk.

)

# Introduction

One of the major constraints on the evolution of large body sizes in animals is an increased risk of developing cancer. If all cells in all organisms have a similar risk of malignant transformation and equivalent cancer suppression mechanisms, organism with many cells should have a higher prevalence of cancer than organisms with fewer cells. Consistent with this expectation there is a strong positive correlation between body size and cancer incidence within species, for example, human cancer incidence increases with increasing adult height [1,2] and cancer incidence is positively correlated with body size in dogs [**???**,3]. There is no correlation, however, between body size and cancer risk between species. This lack of correlation is often referred to as ‘Peto’s Paradox’ [4–6]. While it is clear that a resolution to Peto’s Paradox must involve the evolution of enhanced cancer protection alongside increases in body size and lifespan, the specific genetic, molecular, and cellular mechanisms that underlie this resistance have proven elusive. [7–11].

Among the challenges for discovering how animals evolved enhanced cancer protection mechanisms is identifying lineages in which large bodied species are nested within species with small body sizes. Afrotherian mammals are generally small-bodied, similarly to the prediced common ancestor of Eutherian mammals. For example, maximum adult weights are ~70g in golden moles, ~120g in tenrecs, ~170g in elephant shrews, ~3kg in hyraxes, and 60kg in aardvarks [12]. However, while these extant species are relatively small, the fossil evidence demonstrates that their ancestral lineages reached enormous sizes. For example, while extant hyraxes are relatively small, the extinct Titanohyrax is estimated to have weighted up to ~1300kg [13]. The largest members of Afrotheria, too, are dwarfed by the size of their recent ancestors: extant cows manatees are large bodied (~322-480kg) but are relatively small compared to the extinct Stellar’s sea cow which is estimated to have weight 8000-10000kg [14]. Similarly African (4,800kg) and Asian elephants (3,200kg) are the largest living elephant species, but are dwarfed by the truly gigantic extinct Proboscideans such as Deinotherium (~132,000kg), Mammut borsoni (110,000kg), and the Asian straight-tusked elephant (~220,000kg), the largest known land mammal [15]. Remarkably these large-bodied Afrotherian lineages are nested within small bodied species (Fig. 1) [16–19], indicating that gigantism independently evolved in hyraxes, sea cows, and elephants (Paenungulates). Thus, Paenungulates are an excellent model system in which to explore the mechanisms that underlie the evolution of large body sizes and augmented cancer resistance.

Although many mechanisms can potentially resolve Peto’s paradox, the most parsimonious route to enhanced cancer resistance is likely through an increased copy number of tumor suppressors. Such an example has been seen in the case of candidate genes such as *TP53* and *LIF* [11,20,21] as well as in studies involving a limited set of candidate genes [22,23]. As these studies focus on *a priori* gene sets, however, it remains unknown whether this is a general, genome-wide trend in Afrotherian genomes; and whether such a general trend is associated with the recent increases in body size – and therefore expected cancer risk – in these species.

Here, we trace the evolution of body mass and gene copy number variation in Afrotherians in order to investigate whether gene duplications are enriched in large, long-lived species for genes involved in known tumor suppression pathways. Our estimates of the evolution of body mass, similarly to previous studies [16–19], show that large body masses evolved in a step-wise manner, with major increases in body mass in the Pseudoungulata (17kg), Paenungulata (25kg), Tethytheria (296kg), and Proboscidea (4,100kg) stem-lineages. Furthermore, we see that the ancestral body size increases in Hydracoidia and Sirenia were independent events. To study the evolution of gene copy number, we used a genome-wide Reciprocal Best BLAT Hit (RBBH) method to identify gene duplications in Afrotherian genomes, and used parsimony to infer the lineages in which those duplications occurred. We found gene duplications in lineages with increased body mass were enriched in functions related to tumor suppression, including regulation of the cell cycle, DNA damage repair, and regulation of apoptosis. These data suggest that duplication of tumor suppressors played a role in the evolution of large, long-lived in Afrotherians.

# Methods

## Ancestral Body Size Reconstruction

We built a time-calibrated supertree of Eutherian mammals by combining the time-calibrated molecular phylogeny of Bininda-Emonds *et al.* [24] with the time-calibrated total evidence Afrotherian phylogeny from Puttick and Thomas [**???**]. While the Bininda-Emonds *et al.* [24] phylogeny includes 1,679 species, only 34 are Afrotherian, and no fossil data are included. The inclusion of fossil data from extinct species is essential to ensure that ancestral state reconstructions of body mass are not biased by only including extant species. This can lead to inaccurate reconstructions, for example, if lineages convergently evolved large body masses from a small bodied ancestor. In contrast, the total evidence Afrotherian phylogeny of Puttick and Thomas [19] includes 77 extant species and fossil data from 39 extinct species. Therefore we replaced the Afrotherian clade in the Bininda-Emonds *et al.* [24] phylogeny with the Afrotherian phylogeny of Puttick and Thomas [19] using Mesquite. Next, we jointly estimated rates of body mass evolution and reconstructed ancestral states using a generalization of the Brownian motion model that relaxes assumptions of neutrality and gradualism by considering increments to evolving characters to be drawn from a heavy-tailed stable distribution (the “Stable Model”) [25]. The stable model allows for occasional large jumps in traits and has previously been shown to out-perform other models of body mass evolution, including standard Brownian motion models, Ornstein–Uhlenbeck models, early burst maximum likelihood models, and heterogeneous multi-rate models [25].

## Identification of Duplicate Genes

**Reciprocal Best-Hit BLAT:** We developed a reciprocal best hit BLAT (RBHB) pipeline to quickly identify homologs and estimate gene copy numbers (**Figure 1A**). The Reciprocal Best Hit (RBH) search strategy is conceptually straightforward: 1) Given a gene of interest in a query genome , one searches a target genome for all possible matches to ; 2) For each of these hits, one then performs the reciprocal search in the original query genome to identify the highest-scoring hit; 3) A hit in genome is defined as a homolog of gene if and only if the original gene is the top reciprocal search hit in genome . We selected BLAT [26] as our algorithm of choice, as this algorithm is sensitive to highly simliar (>90% identity) sequences, thus identifying the highest-confidence homologs while minimizing many-to-one mapping problems when searching for multiple genes. RBH performs similar to other more complex methods of orthology prediction, and is particularly good at identifying incomplete genes that may be fragmented in low quality/poor assembled regions of the genome [**???**,27].

**Effective Copy Number By Coverage:** In lower-quality genomes, many genes are fragmented across multiple scaffolds, which results in BLAT calling multiple hits when in reality there is only one gene. To compensate for this, we came up with a novel statistic, Estimated Copy Number by Coverage (ECNC), which averages the number of times we see each nucleotides of a query sequence in a target genome over the total number of nucleotides of the query sequence found overall in each target genome (Supplementary Figure 1). This allows us to correct for genes that have been fragmented across incomplete genomes, while also taking into account missing sequences from the human query in the target genome. Mathematically, this can be written as:

where is a given nucleotide in the query, is the total length of the query, is the number of instances that is present within a reciprocal best hit, and is 1 if or 0 if .

**RecSearch Pipeline** We created a custom Python pipeline for automating RBHB searches between a single reference genome and multiple target genomes using a list of query sequences from the reference genome. For the query sequences in our search, we used the hg38 Proteome provided by UniProt [28], which is a comprehensive set of protein sequences curated from a combination of predicted and validated protein sequences generated by the UniProt Consortium. In order to refine our search, we omitted protein sequences originating from long, noncoding RNA loci (e.g. LINC genes); poorly-studied genes from predicted open reading frames (C-ORFs); and sequences with highly repetitive sequences such as zinc fingers, protocadherins, and transposon-containing genes, as these were prone to high levels of false positive hits. After filtering out problematic protein queries, we then used our pipeline (Figure 1A) to search for all copies of our 18011 query genes in publicly available Afrotherian genomes, including African savannah elephant (*Loxodonta africana*: loxAfr3, loxAfr4, loxAfrC), African forest elephant (*Loxodonta cyclotis*: loxCycF), Asian Elephant (*Elephas maximus*: eleMaxD), Woolly Mammoth (*Mammuthus primigenius*: mamPriV), Colombian mammoth (*Mammuthus columbi*: mamColU), American mastodon (*Mammut americanum*: mamAmeI), Rock Hyrax (*Procavia capensis*: proCap1, proCap2, proCap2\_HiC), West Indian Manatee (*Trichechus manatus latirostris*: triManLat1, triManLat1\_HiC), Aardvark (*Orycteropus afer*: oryAfe1, oryAfe1\_HiC), Lesser Hedgehog Tenrec (*Echinops telfairi*: echTel2), Nine-banded armadillo (*Dasypus novemcinctus*: dasNov3), Hoffman’s two-toed sloth (*Choloepus hoffmannii*: choHof1, choHof2, choHof2\_HiC), Cape golden mole (*Chrysochloris asiatica*: chrAsi1), and Cape elephant shrew (*Elephantulus edwardii*: eleEdw1). For many of these species, we covered multiple assemblies in order to test the effects of assembly size and quality on our hits.

**Query protein criteria** To assemble our query list, we first removed all unnamed genes from UP000005640. Next, we excluded genes from downstream analyses for which assignment of homology was uncertain, including uncharacterized ORFs (991), LOC (63), HLA genes (402), replication dependent histones (72), odorant receptors (499), ribosomal proteins (410), zinc finger transcription factors (1983), viral and repetitive-element-associated proteins (82) and any protein described as either “Uncharacterized,” “Putative,” or “Fragment” by UniProt in UP000005640 (30724), leaving us with a final set of 37582 query protein sequences, corresponding to 18011 genes.

**Duplication gene inclusion criteria** In order to condense transcript-level hits into single gene loci, and to resolve many-to-one genome mappings, we removed exons where transcripts from different genes overlapped, and merged overlapping transcripts of the same gene into a single gene locus call. The resulting gene-level copy number table was then combined with the maximum ECNC values observed for each gene in order to call gene duplications. We called a gene duplicated if its copy number was two or more, and if the maximum ECNC value of all the gene transcripts searched was 1.5 or greater; previous studies have shown that incomplete duplications can encode functional genes, therefore partial gene duplications were included provided they passed additional inclusion criteria. The ECNC cut off of 1.5 was selected empirically, as this value minimized the number of false positives seen in a test set of genes and genomes. The results of our initial search are summarized in Figure 1B. Overall, we identified 13880 genes across all species, or 77.1% of our starting query genes.

**Genome Quality Assessment using CEGMA** In order to determine the effect of genome quality on our results, we used the gVolante webserver and CEGMA to assess the quality and completeness of the genome [**???**,**??**]. CEGMA was run using the default settings for mammals (“Cut-off length for sequence statistics and composition” = 1;“CEGMA max intron length” = 100000; “CEGMA gene flanks” = 10000, “Selected reference gene set” = CVG). For each genome, we generated a correlation matrix using the aforementioned genome quality scores, and either the mean Copy Number or mean ECNC for all hits in the genome.

## Evidence for Functionality of Identified Genes

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

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

**Gene Prediction:** We obtained tracks for genes predicted using GenScan for all the genomes available via UCSC Genome Browser: African savannah elephant (loxAfr3), Rock Hyrax (proCap1), West Indian Manatee (triManLat1), Aardvark (oryAfe1), Lesser Hedgehog Tenrec (echTel2), Nine-banded armadillo (dasNov3), Hoffman’s Two-Toed Sloth (choHof1), Cape golden mole (chrAsi1), and Cape Elephant Shrew (eleEdw1); gene prediction tracks for higher-quality assemblies were not available.

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

## Reconstruction of Ancestral Copy Numbers

We implemented a maximum likelihood method for determining the ancestral copy numbers of genes in *Atlantogenata* using IQ-Tree. For this analysis, we used an unrooted subset of our prior species tree, including only the aforementioned *Atlantogenata* species. We generated PHYLIP files containing the copy number of each gene in the highest quality genome for each species, encoding genes on a scale from 1-31+ copies as 1-9, A-V; and encoding a gene’s copy number as uncetain (“?”) when we did not identify it in the genome. We used the included tree-searching and model-testing functionality in IQ-Tree to determine the most likely topology for the species tree, and to obtain the most likely model for copy number changes in the genome. We defined the ancestral state of a node if it had greater than an 80% posterior probability.

## Pathway Enrichment Analysis

To determine which pathways were associated with duplicated genes in each species and lineage, we used WEBGESTALT to perform overrepresentation analysis (ORA) of the duplicated gene lists relative to our initial query gene list [**???**]. For the database of pathways used in the analysis, we used Reactome [**???**], Wikipathways, and Wikipathways\_cancer [**???**], and KEGG [**???**]. For the ORA, we used FDR for determining significance, and ran the analysis at FDR=0.1, FDR=0.2, FDR=0.3, and FDR=0.5.

## Lifespan Phylogenetic Least-Square Regression and Calculating Estimated Cancer Risk Thoughout Atlantogenata

In order to determine the cancer risk at each node, we first needed to calcultate ancestral lifespans at each node. To do so, we used a Phylogenetic Generalized Least-Square Regression (PGLS) [**???**,**??**] to calculate estimated ancestral lifespans across *Atlantogenata* using our estimates for body size at each node.

Next, we used a simplified multistage cancer risk model for body size and lifespan : [**???**,**??**,**??**]. To determine the change in between nodes, we obtained the ratio between the cancer risk at any given node, and the cancer risk at its ancestral node, using the equation . Finally, to simplify comparisons, we calculated the fold change cancer risk between a node and its ancestor as .

# Results

## Body size frequently and independently expands and contracts throughout *Atlantogenata*

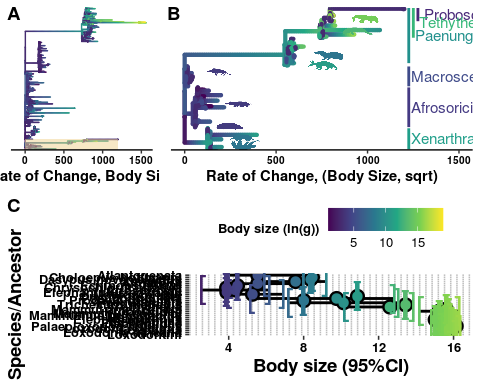


Figure 1: Body sizes rapidly and frequently expand in Eutherians, especially in Atlantogenata. **A)** Tree of Eutherian species, colored by ln(Body Size) and with branch lengths set to the rate of change in body sizes, normalized by the square root of the root branch. Atlantogenata is highlighted at the bottom. **B)** Zoom-in of (A) on Atlantogenata. Silhuetes for the African Elephant, West Indian Manatee, Cape Elephant Shrew, Lesser Hedgehog Tenrec, Cape Golden Mole, Nine-Banded Armadillo, and Hoffman’s Two-Toed Sloth are colored by their extant body sizes, while clade labels are colored based on the common ancestor’s estimated body size

A table generated by the longtable package.

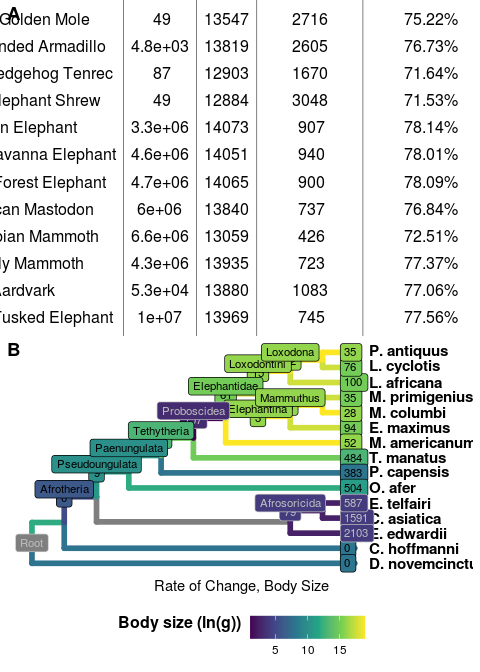
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Node | Size (log(g)) | 95% CI (Low) | 95% CI (High) | Rate (sqrt) |
| Cryptochloris wintoni | 3.13 | 3.13 | 3.13 | 5.78 |
| Amblysomus marleyi | 3.53 | 3.53 | 3.53 | 3.79 |
| Elephantulus revoili | 3.48 | 3.48 | 3.48 | 1.10 |
| Titanohyrax andrewsi | 12.97 | 12.97 | 12.97 | 0.07 |
| Titanohyrax ultimus | 14.08 | 14.08 | 14.08 | 34.61 |
| Megalohyrax sp nov | 12.52 | 12.52 | 12.52 | 7.21 |
| Elephas maximus asurus | 15.66 | 15.66 | 15.66 | 0.34 |
| Protenrec tricuspis | 1.14 | 1.14 | 1.14 | 69.75 |
| Microgale parvula | 1.16 | 1.16 | 1.16 | 33.46 |
| Microgale pusilla | 1.25 | 1.25 | 1.25 | 34.31 |
| Geogale aurita | 1.90 | 1.90 | 1.90 | 40.07 |
| Microgale longicaudata | 2.09 | 2.09 | 2.09 | 0.77 |
| Microgale brevicaudata | 2.19 | 2.19 | 2.19 | 0.60 |
| Microgale jobihely | 2.30 | 2.30 | 2.30 | 1.07 |
| Microgale principula | 2.32 | 2.32 | 2.32 | 0.17 |
| Dilambdogale gheerbranti | 2.38 | 2.38 | 2.38 | 2.21 |
| Microgale taiva | 2.47 | 2.47 | 2.47 | 0.13 |
| Microgale cowani | 2.62 | 2.62 | 2.62 | 0.57 |
| Eremitalpa granti | 3.14 | 3.14 | 3.14 | 9.65 |
| Calcochloris obtusirostris | 3.27 | 3.27 | 3.27 | 13.38 |
| Neamblysomus julianae | 3.33 | 3.33 | 3.33 | 5.72 |
| Chlorotalpa duthieae | 3.38 | 3.38 | 3.38 | 0.32 |
| Chlorotalpa sclateri | 3.54 | 3.54 | 3.54 | 0.09 |
| Macroscelides proboscideus | 3.64 | 3.64 | 3.64 | 14.17 |
| Chrysochloris stuhlmanni | 3.74 | 3.74 | 3.74 | 0.33 |
| Oryzorictes hova | 3.79 | 3.79 | 3.79 | 22.77 |
| Elephantulus myurus | 3.81 | 3.81 | 3.81 | 0.95 |
| Elephantulus brachyrhynchus | 3.81 | 3.81 | 3.81 | 0.93 |
| Elephantulus rozeti | 3.81 | 3.81 | 3.81 | 10.51 |
| Elephantulus fuscus | 3.82 | 3.82 | 3.82 | 0.68 |
| Elephantulus intufi | 3.82 | 3.82 | 3.82 | 1.15 |
| Microgale talazaci | 3.88 | 3.88 | 3.88 | 61.40 |
| Chrysochloris asiatica | 3.89 | 3.89 | 3.89 | 3.34 |
| Elephantulus edwardii | 3.90 | 3.90 | 3.90 | 0.24 |
| Carpitalpa arendsi | 3.94 | 3.94 | 3.94 | 0.45 |
| Amblysomus corriae | 3.94 | 3.94 | 3.94 | 0.98 |
| Amblysomus hottentotus | 3.98 | 3.98 | 3.98 | 0.02 |
| Elephantulus fuscipes | 4.04 | 4.04 | 4.04 | 1.93 |
| Elephantulus rufescens | 4.05 | 4.05 | 4.05 | 0.12 |
| Neamblysomus gunningi | 4.09 | 4.09 | 4.09 | 3.26 |
| Elephantulus rupestris | 4.12 | 4.12 | 4.12 | 0.32 |
| Amblysomus septentrionalis | 4.23 | 4.23 | 4.23 | 0.52 |
| Chambius kasserinensis | 4.27 | 4.27 | 4.27 | 11.84 |
| Amblysomus robustus | 4.33 | 4.33 | 4.33 | 1.38 |
| Micropotamogale lamottei | 4.36 | 4.36 | 4.36 | 2.82 |
| Echinops telfairi | 4.47 | 4.47 | 4.47 | 7.75 |
| Limnogale mergulus | 4.52 | 4.52 | 4.52 | 121.95 |
| Hemicentetes semispinosus | 4.75 | 4.75 | 4.75 | 4.68 |
| Chrysospalax villosus | 4.77 | 4.77 | 4.77 | 0.13 |
| Petrodromus tetradactylus | 5.29 | 5.29 | 5.29 | 24.61 |
| Herodotius pattersoni | 5.50 | 5.50 | 5.50 | 11.64 |
| Setifer setosus | 5.61 | 5.61 | 5.61 | 12.52 |
| Rhynchocyon cirnei | 5.86 | 5.86 | 5.86 | 3.30 |
| Metoldobotes sp nov | 5.93 | 5.93 | 5.93 | 15.94 |
| Chrysospalax trevelyani | 6.13 | 6.13 | 6.13 | 62.84 |
| Rhynchocyon petersi | 6.15 | 6.15 | 6.15 | 2.13 |
| Rhynchocyon chrysopygus | 6.28 | 6.28 | 6.28 | 0.40 |
| Potamogale velox | 6.49 | 6.49 | 6.49 | 103.04 |
| Rhynchocyon udzungwensis | 6.57 | 6.57 | 6.57 | 4.33 |
| Tenrec ecaudatus | 6.75 | 6.75 | 6.75 | 79.50 |
| Dasypus sabanicola | 7.05 | 7.05 | 7.05 | 12.18 |
| Tolypeutes matacus | 7.11 | 7.11 | 7.11 | 15.96 |
| Dasypus septemcinctus | 7.30 | 7.30 | 7.30 | 4.44 |
| Zaedyus pichiy | 7.31 | 7.31 | 7.31 | 5.54 |
| Dasypus hybridus | 7.31 | 7.31 | 7.31 | 4.05 |
| Chaetophractus villosus | 7.61 | 7.61 | 7.61 | 0.42 |
| Chaetophractus nationi | 7.67 | 7.67 | 7.67 | 0.09 |
| Heterohyrax brucei | 7.78 | 7.78 | 7.78 | 1.64 |
| Cabassous centralis | 7.92 | 7.92 | 7.92 | 0.25 |
| Seggeurius amourensis | 7.98 | 7.98 | 7.98 | 2.82 |
| Procavia capensis | 8.01 | 8.01 | 8.01 | 0.00 |
| Dendrohyrax dorsalis | 8.06 | 8.06 | 8.06 | 1.86 |
| Microhyrax lavocati | 8.13 | 8.13 | 8.13 | 0.73 |
| Bradypus tridactylus | 8.23 | 8.23 | 8.23 | 0.48 |
| Bradypus torquatus | 8.27 | 8.27 | 8.27 | 0.03 |
| Dasypus novemcinctus | 8.37 | 8.37 | 8.37 | 14.73 |
| Euphractus sexcinctus | 8.43 | 8.43 | 8.43 | 14.99 |
| Choloepus hoffmanni | 8.47 | 8.47 | 8.47 | 0.32 |
| Bradypus variegatus | 8.49 | 8.49 | 8.49 | 0.51 |
| Tamandua tetradactyla | 8.52 | 8.52 | 8.52 | 10.44 |
| Cyclopes didactylus | 8.53 | 8.53 | 8.53 | 2.15 |
| Choloepus didactylus | 8.71 | 8.71 | 8.71 | 0.64 |
| Thyrohyrax meyeri | 8.78 | 8.78 | 8.78 | 3.55 |
| Saghatherium bowni | 9.13 | 9.13 | 9.13 | 15.85 |
| Dasypus kappleri | 9.23 | 9.23 | 9.23 | 74.13 |
| Thyrohyrax domorictus | 9.30 | 9.30 | 9.30 | 1.15 |
| Dimaitherium patnaiki | 9.57 | 9.57 | 9.57 | 18.23 |
| Phosphatherium escuilliei | 9.62 | 9.62 | 9.62 | 326.23 |
| Saghatherium antiquum | 9.73 | 9.73 | 9.73 | 2.90 |
| Thyrohyrax litholagus | 10.01 | 10.01 | 10.01 | 28.58 |
| Myrmecophaga tridactyla | 10.26 | 10.26 | 10.26 | 41.03 |
| Myorycteropus africanus | 10.27 | 10.27 | 10.27 | 0.57 |
| Selenohyrax chatrathi | 10.73 | 10.73 | 10.73 | 14.99 |
| Priodontes maximus | 10.82 | 10.82 | 10.82 | 268.43 |
| Orycteropus afer | 10.87 | 10.87 | 10.87 | 6.59 |
| Antilohyrax pectidens | 10.93 | 10.93 | 10.93 | 13.69 |
| Bunohyrax fajumensis | 11.32 | 11.32 | 11.32 | 1.45 |
| Afrohyrax championi | 11.32 | 11.32 | 11.32 | 0.19 |
| Geniohyus mirus | 11.33 | 11.33 | 11.33 | 5.44 |
| Prorastomus sirenoides | 11.49 | 11.49 | 11.49 | 13.61 |
| Elephas antiquus falconeri | 11.51 | 11.51 | 11.51 | 6.12 |
| Pachyhyrax crassidentatus | 11.81 | 11.81 | 11.81 | 2.29 |
| Megalohyrax eocaenus | 11.95 | 11.95 | 11.95 | 0.24 |
| Elephas cypriotes | 12.21 | 12.21 | 12.21 | 1.90 |
| Bunohyrax major | 12.36 | 12.36 | 12.36 | 11.39 |
| Titanohyrax angustidens | 12.48 | 12.48 | 12.48 | 0.04 |
| Daouitherium rebouli | 12.80 | 12.80 | 12.80 | 0.74 |
| Arcanotherium savagei | 12.89 | 12.89 | 12.89 | 7.29 |
| Dugong dugon | 12.92 | 12.92 | 12.92 | 5.85 |
| Trichechus senegalensis | 13.03 | 13.03 | 13.03 | 0.57 |
| Trichechus inunguis | 13.08 | 13.08 | 13.08 | 0.69 |
| Protosiren smithae | 13.20 | 13.20 | 13.20 | 33.69 |
| Numidotherium koholense | 13.23 | 13.23 | 13.23 | 2.29 |
| Omanitherium dhofarensis | 13.35 | 13.35 | 13.35 | 0.03 |
| Trichechus manatus | 13.44 | 13.44 | 13.44 | 1.39 |
| Moeritherium spp | 13.82 | 13.82 | 13.82 | 5.71 |
| Phiomia spp | 13.89 | 13.89 | 13.89 | 3.64 |
| Elephas maximus | 15.02 | 15.02 | 15.02 | 5.81 |
| Barytherium spp | 15.20 | 15.20 | 15.20 | 73.58 |
| Mammuthus primigenius | 15.27 | 15.27 | 15.27 | 2.17 |
| Mammut borsoni | 16.49 | 16.49 | 16.49 | 15.33 |
| Mammuthus trogontherii | 16.38 | 16.38 | 16.38 | 16.00 |
| Loxodonta africana | 15.35 | 15.35 | 15.35 | 1.28 |
| Loxodonta cyclotis | 15.37 | 15.37 | 15.37 | 3.72 |
| Palaeoloxodon antiquus | 16.14 | 16.14 | 16.14 | 0.01 |
| Palaeoloxodon namadicus | 16.81 | 16.81 | 16.81 | 12.81 |
| Mammut americanum | 15.61 | 15.61 | 15.61 | 0.95 |
| Mammuthus columbi | 15.71 | 15.71 | 15.71 | 0.91 |
| Hydrodamalis gigas | 15.72 | 15.72 | 15.72 | 172.52 |
| Atlantogenata | 5.55 | 4.06 | 7.95 | 0.03 |
| Afrotheria | 5.55 | 4.05 | 7.96 | 0.00 |
| Afrosoricida | 4.35 | 2.58 | 6.13 | 44.49 |
| Macroscelidae | 5.27 | 3.98 | 6.85 | 2.49 |
| Pseudoungulata | 9.76 | 5.21 | 12.78 | 545.83 |
| Paenungulata | 10.13 | 7.24 | 13.02 | 4.42 |
| Tethytheria | 12.60 | 10.25 | 13.81 | 187.47 |
| Proboscidea | 15.23 | 14.22 | 16.24 | 30.28 |
| Elephantidae | 15.49 | 14.89 | 16.10 | 2.21 |
| Elephantina | 15.51 | 15.08 | 15.96 | 0.01 |
| Mammuthus | 15.54 | 15.24 | 15.85 | 0.47 |
| Loxodontini | 15.55 | 15.02 | 16.11 | 0.11 |
| Loxodona | 15.72 | 15.16 | 16.30 | 0.86 |
| Xenarthra | 7.57 | 5.96 | 9.18 | 124.94 |

To trace the evolutionary history of body mass and lifespan in Afrotherians, we built a time-calibrated supertree of Eutherian mammals combining 1,679 species from Bininda-Emonds et al [24] with a total evidence Afrotherian phylogeny including 77 extant and fossil data from 39 extinct species [19]. Fossil data from extinct species were included to ensure that ancestral state reconstructions of body mass in Afrotherians were not biased by only including extant species, which can lead to inaccurate reconstructions, for example, if lineages multiple lineages evolved large body masses from a small bodied ancestor. We jointly estimated rates of body mass evolution and reconstructed ancestral states using a generalization of a Brownian model of character evolution, which allows for occasional large jumps in traits (stable model) and out performs standard Brownian motion and Ornstein-Uhlenbeck models of character evolution [25].

Similar to previous studies of Afrotherian body size [19,25], we found that the body mass of the Afrotherian ancestor was inferred to be small (0.26kg, 95% CI: 0.31-3.01kg) and that substantial accelerations in the rate of body mass evolution occurred coincident with a 67.36x increase in body mass in the stem-lineage of *Pseudoungulata* (17.33kg), a 1.45x increase in body mass in the stem-lineage of *Paenungulata* (25.08kg), a 11.82x increase in body mass in the stem-lineage of *Tehthytheria* (296.56kg), and a 2.69x increase **(Vinnie: “14× increase”, note I tried to divide the MRCA of LoxAfer with Phosphatherium, Moeritherium, Omanitherium, Phiomia, and Mammut by their respective ancestral nodes, but never got anything that was a 14x increase…)** in body mass in the stem-lineage of *Proboscidea* (4114.39kg) **(Vinnie:“(4,100kg)”; I get 4114kg; can’t find an MRCA with that value)** (Figure 1B,C). The ancestral *Hyracoidea* was inferred to be relatively small 2.86kg-118.18kg **(Vinnie wrote (2.86-15.71kg); I get an upper range of 118.18kg)**, and rate accelerations were coincident with independent body mass increases in large hyraxes such as *Titanohyrax andrewsi* 429.34kg; 67.36x increase. While the body mass of the ancestral *Sirenian* was inferred to be large 61.7kg-955.51kg \_(Vinnie: (61-656kg); I get an upper range of 956kg)\_\_, a rate acceleration occurred coincident with a 10.59x increase in body mass in Stellar’s sea cow. Rate accelerations also occurred coincident with 36.6x decrease in body mass in the stem-lineage of the dwarf elephants *Elephas (Palaeoloxodon) antiquus falconeri* and *Elephas cypriotes*. These data suggest that gigantism in *Afrotherians* evolved step-wise, from small to medium bodies in the *Pseudoungulata* stem-lineage, medium to large bodies in the *Tehthytherian* stem-lineage and extinct hyraxes, and from large to exceptionally large bodies independently in the *Proboscidean* stem-lineage and Stellar’s sea cow (Figure 1).

## Pervasive duplication of tumor suppressors during the origins of large bodied Afrotheirans

## Warning: TODO: Label each genome with citation number

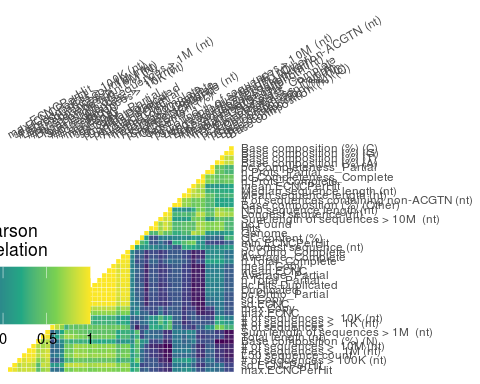


Gene duplications occur readily throughout Atlantogenata

Enhanced cancer suppression may have evolved through many mechanisms; among the most parsimonious is an increase in the copy number of genes with tumor suppressor functions. Previous studies focusing on candidate gene studies, for example, have identified increased copy number of the tumor suppressors *TP53* and *LIF* in elephants [**???**,11,21–23]. Therefore, in order to test whether this was a pervasive phenomena genome-wide in *Afrotherians*, we used a Reciprocal Best Hit BLAT (RBHB) approach to infer gene copy number in *Afrotherian* and *Atlantogenatan* genomes (Fig. 2A, Supplementary Figure 1). Because RBHB-like approaches can over-estimate copy number when genes are fragmented or incorrectly assembled across multiple scaffolds, we also inferred copy number using a complementary method that quantifies the ratio between observed and expected gene coverage per nucleotide (ECNC) (Supplementary Figure 1). By only including nucleotides from the query sequence that were observed in the target genome, we also correct for partial hits where some or all of the homologs of a gene have diverged from the human homolog.

Because our sequence database included various protein transcripts for each gene, in order to obtain gene-level copy number information and elimate any many-to-one mappings of hits, we labeled each exon of every reciprocal best hit (RBH) with the gene corresponding to the query transcript and merged all overlapping exons; next, we eliminated any many-to-one exons that resulted from the previous step. Finally, we reassembled the gene loci based on the original transcript starts and ends, and the collapsed exon data, obtaining the full sequence of each RBH locus. Genes were considered to be duplicated if its copy number via RBHB was greater than or equal to 2, and the maximum ECNC among all transcripts prior to filtering was greater than or equal to 1.50. This cutoff of ECNC was selected to account for truncated gene duplications, which have been shown to be functional in various examples [**???** examples of this].

Once we applied these criteria to the results, we obtained the results shown in Figure 2A. Our approach positively identified an average of NA of the queries we searched. We observed that the percentage of duplicated genes in non-Pseudoungulatan genomes was significantly higher: while these genomes had duplicates percentages ranging anywhere from 3.26% to 7.80%, outgroup species’ duplication rates ranged from 12.94% to 23.66%.



Given the preliminary status of the genomes for *Chrysochloris asiatica* and *Elephantulus edwardii*; and the fact that more revisions have been performed for Paenungulatan genomes; we hypothesized that genome quality may be confounding the results we’ve observed. We used CEGMA and the gVolante server [**???**,**??**] to assess the quality of all our genomes, to see if any of these metrics correlated with our metrics of gene counts and copy number. As shown in Figure 3, mean ECNC, mean Copy Number, and mean CN (the lesser of Copy Number and ECNC per gene) correlate moderately strongly with genomic quality metrics related to the length and assembly quality, like LD50 and the number of scaffolds and contigs with a length above either 100K or 1M, supporting our hypothesis that increases in copy number correlate .

In order to determine when in the phylogeny our identified duplications occured, we used maximum likelihood to reconstruct ancestral gene copy numbers across the phylogeny. We encoded the copy number of each gene as a discrete trait ranging from 0 (“?”) to 31+, and then ran the ancestral gene copy number reconstruction using the phylogenetics suite iqtree, which rapidly tests and runs the best evolutionary model for a given trait of interest. Importantly, iqtree can be set to only output states where the likelihood of the state is greater than a certain threshold. Setting the threshold to 0.8, we obtained the results shown in

Next, in order to select genes and duplicates which were likely functional, we omitted any hits that were not supported by either the gene prediction method GenScan, or by at least one transcript assembled from publically-available RNA-seq data.

We describe the number of genes that increased in each lineage in *Atlantogenata* in Figure 2. Among the genes that increased in copy number in the elephant lineage are TP53 and LIF, as previously described. Furthermore, we identify

## Duplications that occured recently in Probodiscea are enriched for tumor suppressor pathways

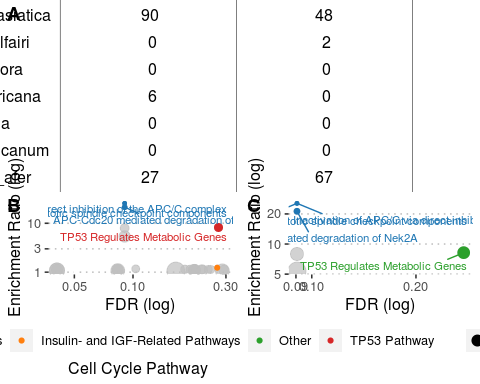


Fig 3

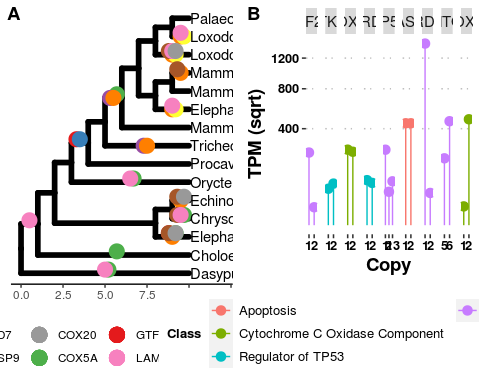
As previously stated, our inital hypothesis was that genes duplicated in lineages that experienced a growth in size would have an increased number of gene duplications that were enriched for membership in tumor suppressor pathways. Thus, we used WebGestalt and its Overrepresentation Analysis (ORA) functionality to determine what pathways were enriched in our duplicated gene sets in each branch relative to our initial query set. For our database, we used Reactome for our primary analysis, but additionally used the KEGG, Panther, Wikipathways, and Wikipathways\_cancer databases using WebGestalt ORA. Going through the tree, at no FDR<0.5 is there any significant pathway representation for genes that increased in the branches leading to , or *Paenungulata*; furthermore, there are no significant pathway enrichments at FDR<0.5 for copy-number-stable genes for , or *Afrotheria*. Note that because *Xenarthra* was selected as the outgroup, it is not possible to polarize the changes in their gene copy numbers along the tree.

For the other branches, the number of pathways that came up as significantly enriched at each FDR is shown in Figure 3A (Supplementary Figure 3). For the species with high duplication rates and lower-quality, highly-fragmented genomes, such as with *Chrysochloris asiatica* (20.05% duplicated hits) and *Elephantulus edwardii* (23.66% duplicated hits), it is unsurprising that there is an equivalently astronomical number of pathway enrichments present. In the case of these two species, their many pathway enrichments also span an incredible range of processes at every level of biology; this, in combination with the astronomically high number of copy numbers identified for these genes, further suggests a need for improvement and refinement in these genomes. In the cell cycle pathways called as significant in the genomes of these two species plus *Orycteropus afer* and *Echinops telfairi*, we see that the member duplicated genes are from the same sets families, such as the APC subunit family; the proteosome subunit families; and the protein phosphatase 2 family, among others. It is highly possible that these results reflect true expansions of these gene families, especially in the higher-quality *Orycteropus afer* genome; however, as discussed, it is also possible that it simply reflects artifcatual duplication events, and so require futher study.

The pathway enrichments for the stable - yet duplicated - copy number, and increased copy number sets of genes for the African Elephant are shown in Figure 3B-C. In support of our original hypothesis, among the few enriched pathways in the African Elephant, we see that that two tumor suppression pathways - APC Complex-related pathways, and “TP53 Regulates Metabolic Genes”- appear not only in the case of stable genes, but also in the set of newly duplicated genes. The other pathways we see in the set of recently-duplicated genes include “Functionalization of Compounds” and its daughter pathway “Xenobiotics”. Genes in these pathways serve to add functional groups to lipophylic compounds which would otherwise not be reactive in the cell, and are types of metabolic pathways. In the stable set we see enrichment of pathways such as “Neuronal Systems” and “Axon Guidance,” which fit in well with what is known about elephant biology and evolution [**???** probably]. Overall in elephants, we see enrichments within duplicated genes for pathways involved in what makes an elephant an elephant - including tumor suppressor pathways.

## Concerted duplication of TP53 and TP53-related genes towards *Probodiscea*

## Gene geneID Locus TPM loxAfr4 triManLat2  
## 1: A2M loxAfr4.11247 A2M\_1 155.29575 2 (1) 1 (1)  
## 2: A4GALT loxAfr4.15454 A4GALT\_1 17.23229 1 (1) 1 (1)  
## 3: AAAS loxAfr4.11523 AAAS\_1 68.77647 1 (1) 1 (1)  
## 4: AACS loxAfr4.15893 AACS\_1 44.25330 1 (1) 1 (1)  
## 5: AADAT loxAfr4.8279 AADAT\_1 45.40810 1 (1) 2 (1.54)  
## ---   
## 16239: ZYX loxAfr4.18095 ZYX\_1 20.48161 1 (1) 1 (1)  
## 16240: ZYX loxAfr4.18096 ZYX\_1 47.78743 1 (1) 1 (1)  
## 16241: ZZEF1 loxAfr4.2382 ZZEF1\_1 56.79362 1 (1) 1 (1)  
## 16242: ZZZ3 loxAfr4.6778 ZZZ3\_1 609.06805 1 (1) 1 (1)  
## 16243: pk loxAfr4.12065 pk\_1 157.57671 1 (1) 1 (1)



Prior studies looking at the duplication of TP53 in the African Elephant motivated further study for the enrichment of genes involved in TP53-related metabolic pathway. We traced the evolution of all genes involved in this TP53-duplicated pathway that were duplicated in the African Elephant, and used publically-available RNA-seq data to see which genes are actively expressess in living elephants. Excitingly, we found that the initial duplication of TP53 in Tethytheria, where body size expanded, was presceded by the duplication of GTF2F1 and STK11 in *Paenungulata*; and was coincided by the duplication of BRD7. [TALK ABOUT HOW THESE TWO MAY HAVE FACILITATED THE DUPLICATION OF TP53] GTF2F1 is a general transcription factor [**???** and talk more about this] STK11 is a tumor suppressor gene [**???** and talk more]

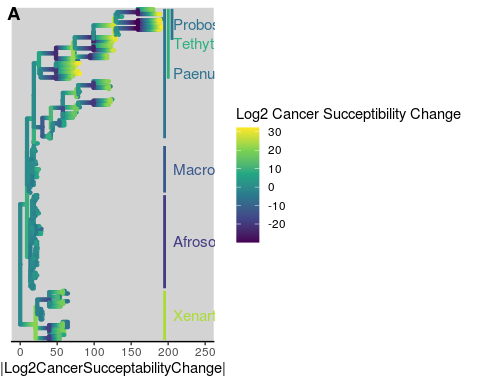
The other genes that are duplicated in the pathway are all downstream of TP53; these genes duplicated either alongside TP53 in the case of SIAH1, or subsequently in *Probodiscea*, *Elephantidae*, and in modern elephants. (Figure 4A). These genes are all expressed in RNA-seq data, suggesting that they encode functional genes in modern elephants (Fig 4B).

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

##   
## Phylogenetic Least Squares: ln(Lifespan) & ln(Body Size) Regression  
## ===============================================  
## Dependent variable:   
## ---------------------------  
## Lifespan   
## -----------------------------------------------  
## lnSize 0.100   
## (0.121)   
## t = 0.826   
## p = 0.409   
## Constant 1.943   
## (1.385)   
## t = 1.403   
## p = 0.161   
## -----------------------------------------------  
## Observations 28   
## Log Likelihood -39.636   
## Akaike Inf. Crit. 85.273   
## Bayesian Inf. Crit. 89.047   
## ===============================================  
## Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

## Warning: TODO: Caption Fig 5

## Warning: TODO: de-comment geom\_tiplab in f5a! Commented currently for offline  
## usage!



##   
## Phylogenetic Least Squares: ln(Lifespan) & ln(Body Size) Regression  
## ===============================================  
## Dependent variable:   
## ---------------------------  
## Lifespan   
## -----------------------------------------------  
## lnSize 0.100   
## (0.121)   
## t = 0.826   
## p = 0.409   
## Constant 1.943   
## (1.385)   
## t = 1.403   
## p = 0.161   
## -----------------------------------------------  
## Observations 28   
## Log Likelihood -39.636   
## Akaike Inf. Crit. 85.273   
## Bayesian Inf. Crit. 89.047   
## ===============================================  
## Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

A table generated by the longtable package.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Node | lnSize | Est. Lifespan | Est. Cancer Susceptibility (K1) | Est. Ancestral K2 | Change in K (K2/K1) | log2 Fold Change | Ancestor |
| Loxodontini | 16 | 34.38 | 1.47e+16 | 2.97e+15 | 4.940000e+00 | 2.31 | Elephantidae |
| Loxodonta africana | 15 | 65.00 | 2.47e+17 | 1.47e+16 | 1.681000e+01 | 4.07 | Loxodontini |
| Loxodona | 16 | 34.38 | 1.47e+16 | 1.47e+16 | 1.000000e+00 | 0.00 | Loxodontini |
| Loxodonta cyclotis | 15 | 31.12 | 2.97e+15 | 1.47e+16 | 2.000000e-01 | -2.31 | Loxodona |
| Palaeoloxodon antiquus | 16 | 34.38 | 1.47e+16 | 1.47e+16 | 1.000000e+00 | 0.00 | Loxodona |
| Elephantidae | 15 | 31.12 | 2.97e+15 | 1.40e+07 | 2.125534e+08 | 27.66 | Proboscidea |
| Elephantina | 16 | 34.38 | 1.47e+16 | 2.97e+15 | 4.940000e+00 | 2.31 | Elephantidae |
| Elephas maximus | 15 | 65.50 | 2.58e+17 | 1.47e+16 | 1.760000e+01 | 4.14 | Elephantina |
| Mammuthus | 16 | 34.38 | 1.47e+16 | 1.47e+16 | 1.000000e+00 | 0.00 | Elephantina |
| Mammuthus primigenius | 15 | 31.12 | 2.97e+15 | 1.47e+16 | 2.000000e-01 | -2.31 | Mammuthus |
| Mammuthus columbi | 16 | 34.38 | 1.47e+16 | 1.47e+16 | 1.000000e+00 | 0.00 | Mammuthus |
| Proboscidea | 3 | 9.41 | 1.40e+07 | 1.21e+14 | 0.000000e+00 | -23.05 | Tethytheria |
| Mammut americanum | 16 | 34.38 | 1.47e+16 | 1.40e+07 | 1.050567e+09 | 29.97 | Proboscidea |
| Tethytheria | 13 | 25.49 | 1.21e+14 | 1.01e+12 | 1.207400e+02 | 6.92 | Paenungulata |
| Trichechus manatus | 13 | 69.00 | 4.77e+16 | 1.21e+14 | 3.929800e+02 | 8.62 | Tethytheria |
| Paenungulata | 10 | 18.91 | 1.01e+12 | 1.01e+12 | 1.000000e+00 | 0.00 | Pseudoungulata |
| Procavia capensis | 8 | 14.80 | 3.13e+10 | 1.01e+12 | 3.000000e-02 | -5.01 | Paenungulata |
| Pseudoungulata | 10 | 18.91 | 1.01e+12 | 1.69e+09 | 5.967900e+02 | 9.22 | Afroinsectivora |
| Orycteropus afer | 11 | 29.80 | 4.19e+13 | 1.01e+12 | 4.167000e+01 | 5.38 | Pseudoungulata |
| Elephantulus edwardii | 4 | 10.40 | 6.90e+07 | 1.69e+09 | 4.000000e-02 | -4.61 | Afroinsectivora |
| Afrosoricida | 4 | 10.40 | 6.90e+07 | 1.69e+09 | 4.000000e-02 | -4.61 | Afrotheria |
| Chrysochloris asiatica | 4 | 10.40 | 6.90e+07 | 6.90e+07 | 1.000000e+00 | 0.00 | Afrosoricida |
| Echinops telfairi | 4 | 19.00 | 2.57e+09 | 6.90e+07 | 3.722000e+01 | 5.22 | Afrosoricida |
| Afrotheria | 6 | 12.69 | 1.69e+09 | 2.83e+06 | 5.967900e+02 | 9.22 | Atlantogenata |
| Xenarthra | 11 | 20.89 | 4.97e+12 | 2.83e+06 | 1.760358e+06 | 20.75 | Atlantogenata |
| Dasypus novemcinctus | 8 | 22.30 | 3.67e+11 | 4.97e+12 | 7.000000e-02 | -3.76 | Xenarthra |
| Choloepus hoffmanni | 8 | 41.00 | 1.42e+13 | 4.97e+12 | 2.850000e+00 | 1.51 | Xenarthra |
| Atlantogenata | 2 | 8.52 | 2.83e+06 | 2.83e+06 | 1.000000e+00 | 0.00 | Atlantogenata |
| Afroinsectivora | 6 | 12.69 | 1.69e+09 | 1.69e+09 | 1.000000e+00 | 0.00 | Afrotheria |

The dramatic increase in body mass and lifespan in some Afrotherian lineages implies those lineages evolved reduced cancer risk. To infer the magnitude of these reductions we estimated differences in cancer risk between small bodied, short-lived species and large bodied, long-lived species as well as for reconstructed ancestral Afrotherians. Following [**???**] we estimate the intrinsic cancer risk as the product of risk associated with body mass and lifespan. Differences in cancer susceptibility due to body mass differences between species can be approximated simply as the fold difference in body mass () between species [**???**]. The risk of developing cancer also increases in proportion to the sixth power of age and is approximated by the formula , in which the proportionality constant C that determines susceptibility to cancer induction is multiplied by the sixth power of the age in years, [**???**,**??**,**??**]. Thus we can estimate the intrinsic cancer risk for a species as .

In order to estimate the intrinsic cancer risk of a species, we first obtained estimates for lifespans at ancestral nodes using PGLS and the model (Figure ). With this information in hand, we calculated at all nodes, and then estimated the fold change in cancer succeptibility between an ancestral node and a given node as (**Table 4**). As body size and phylogeny are the strongest predictors of lifespan, we find that this regression is sufficiently robust without further metabolic or other covariates for our purposes herein.

As shown in **Table 4**, cancer succeptibility skyrocketed at the initial divergence of Atlantogenata, followed by a generally upwards trend. At the common ancestor of Afrotheria there is an inital 9.22-fold increase in cancer risk. In parallel to Afrotheria, cancer succeptibility increases 20.75-fold in Xenarthra. However, cancer risk slowly deflates as size decreases as one moves along the tree towards extant species, such as in Hoffman’s Two Toed Sloth (1.51-fold change) and in the Nine-banded Armadillo (-3.76-fold change).

Within Afrotheria, cancer succeptibility drops in Afrosoricida as species shrink (-4.61-fold, then stagnates for the Cape Golden Mole) - but then rises 5.22-fold towards the Lesser Hedgehog Tenrec. In parallel, Afroinsectivora does not increase in cancer succeptibility, and decreases once more at the Cape Elephant Shrew (-4.61-fold). The emergence of Pseudoungulata sees the next big leap in cancer succeptibility with a 9.22-fold increase. The Aardvark further increases 5.38-fold, while we don’t observe an increase at the common ancestor of Paenungulates. While the Rock Hyrax decreases in cancer succeptibility as expected (-5.01-fold), Tethytheria sees a sharp increase in cancer risk (6.92-fold). Within Tethytheria, the Manatee’s cancer risk increases once more 8.62-fold, while Proboscidea’s cancer risk drops precipitously along with its body size (-23.05-fold). Yet, within Proboscidea we see the biggest increases: right off the bat, we see that the cancer succeptibility of Elephantidae and the American Mastodon skyrocket by 27.66-fold and 29.97-fold, respectively. Both Elephantina and Loxodontini in Elephantidae have a 2.31-fold increase in cancer succeptibility. Within Elephantina, cancer susceptibility stays stable at Mammuthus and in the Colombian Mammoth, and slightly decreases in the Wooly Mammoth (-2.31-fold). The three extant elephants - Asian Elephant in Elephantina, the African Savana Elephant in Loxodontini, and the African Forest Elephant in Loxodona, meanwhile, have parallel and similar decreases in both size and cancer succeptibility (4.14-, 4.07-, and -2.31-fold, respectively). Neither the common ancestor of Loxodonta, nor the Straight-Tusked Mammoth see any further changes in cancer succeptibility.

# Discussion

With our results, we have demonstrated that gigantism in *Atlantogenata* was not limited to extant elephants, but rather occured at various points in the evolution of the clade; however, the hundred-fold to hundred-million-fold increases in cancer risk that is associated with these increases however poses an inate challenge in the evolution and persistance of this trait. The environmental selective pressures on body size have long been the fascination of evolutionary biologists, and the influence of climate, predation, geography, and ecological niche on body size have been well established. [Cooper & Purvis 2010, MORE CITATIONS]. Indeed, there is a general trend, known as Cope’s Rule, for body sizes of species to increase over time [Citation for Cope’s Rule]. However, for all the research on body size that has been done thus far, the mechanisms that enable a release from the negative pressure on body size exerted on cancer risk has proven more elusive.

Furthermore, we show that tumor suppressor duplications are enriched not only in large, extant species, but also in large common ancestors, and that these duplications evolved throughout the tree, rather than in concert. The ancestral body sizes of many of the subclades in *Atlantogenata* were estimated to be large, and the estimated cancer risk increases - even for the small clades like *Afroinsectivora* may explain why these TSG gene duplications occured early on, and persisted even in species that we did not expect to have high risks of cancer. However, some of our results also provide interesting evidence for a paradigm of TSG duplicaitons being a pervasive phenomenon, rather than a specific mechansim that occurs after the evolution of gigantism. At the common ancestor of *Proboscidea*, which was small relative to both its ancestors and its decendants, we see the emergence of various TSG duplications, which may have enabled the stratospheric increase in body size of modern elephants. We also identified many TSG duplication events in smaller species and lineages such as in *Chrysochloris asiatica* and *Elephantulus edwardii*, although these may have been the result of low-quality genomes.

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

The lack of a stronger signal from tumor suppressor duplications is likely a result of the strong effect size on both cancer risk and organismal toxicity that a TSG duplication would provide. The duplication of a tumor suppressor in many cases is associated with mild toxicity, although it greatly varies given the context and TSG in questionl; however, a single TSG duplication can also provide significant protection against cancer. For example, the overexpression of TP53 in mice, while protective of cancer, is associated with progeria and early death; however, if an additional copy of TP53 is introduced with its regulatory elements intact, the mice are healthy and experience normaging, while also demonstrating an enhanced response to cellular stress and lower rates of cancer. [**???** the mouse TP53 papers]. In light of this, it is fascinating that our results in the elephant lineage suggest that the duplication of TP53 regulators preceded the retroduplication and expansion of TP53, as this likely would have lowered the toxicity of the intial duplication and thus enabled it to occur. Given a suffienct selective pressure on increasing body size, it stands to reason that events like this could alleviate the negative pleiotropy of TSG duplications sufficiently to enable their persistence and allow for subsequent refinment over evolutionary time.

By combining a phylogenetic study on body size in addition to a survey of copy number across nearly all protein-coding genes, we provide a comprehensive look at the question of the role of cancer risk and body size in *Atlantogenata* that may provide broader insight to other mammalian species. Our study was initially motivated by the identification of functional duplicates of tumor suppressors, such as TP53 and LIF in elephants [**???**,11,21]. Further in support of our results, a larger candidate gene study by Caulin et al. [**???**] characterized the copy number of 830 known tumor-suppressor genes across 36 mammals and identified 382 putative duplicates, including duplicates in species with large body sizes and long life-spans. However, while candidate gene studies are useful, by their very design they are biased in determining larger patterns of evolution of traits. Without addressing these questions with a genome-wide approach, any and all insights will be inevitably limited to a fraction of the whole story.

Our results suggest that the pervasive duplication of tumor suppressors may help enable the evolution of larger body sizes by lowering the cancer risk of species, either prior to or in lockstep with increasing body size. however, this is unlikely to be the only genetic mechanism at play in this scenario. In genome-wide studies of unusually large or long-lived species such as the bowhead whale (Keane et al., 2015), Myotid bats (Seim et al., 2013; Zhang et al., 2013), naked mole rat (Kim et al., 2011), and blind mole rat (Fang et al., 2014), there were cases of overrepresentation of TSGs among duplicate genes that were outshadowed by the identification of strong signatures of positive selection at TSGs. While the evolution of regulatory and coding elements of both TSGs and other non-cannonical tumor suppressor genes have been shown to be important for mediating the cancer risk of long-lived species, there has been no attention given to the possiblity of TSG duplications providing a relaxation of possible negative pleitropy that could result from these traits. It has been well-established in the literature that genes duplication events allow for evolutionary drift in one of the copies, which frequently results in neofunctionalization or specialization of the two copies [@ now actually support this]. While this is beyond the scope of our study, a promising future direction of this work would include an evolutionary analysis of duplicated genes relative to each other to see if this has already occured between the pairs of genes we have identified.

# Acknowledgements

I would like to thank Olga Duchenko at the Aiden Lab, D.H. Vazquez for his indispensible support.

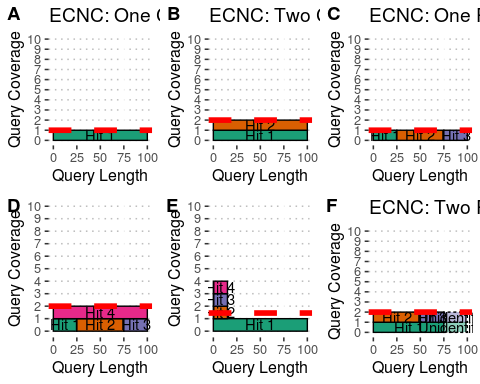
# Conflicts of Interest

The Authors have no conflicts of interest to report

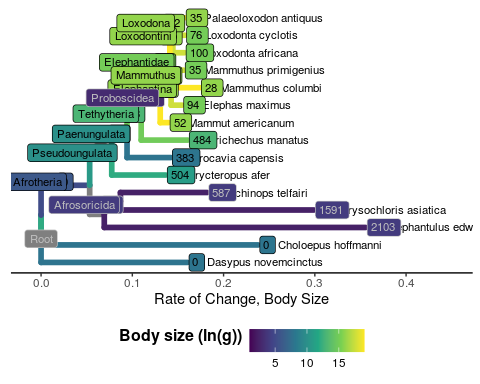
# Funding Source

We would like to thank the Department of Human Genetics at the University of Chicago for supporting this project.

# SUPPLEMENTARY FIGURES

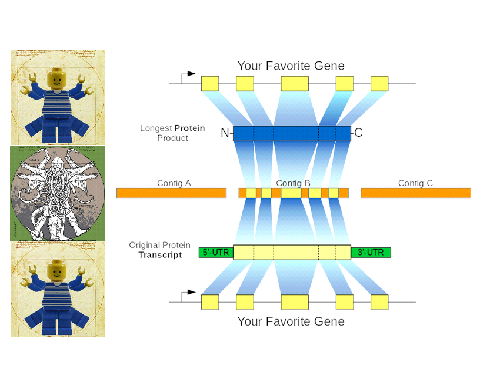


Supplementary Figure 1: Estimated Copy Number by Coverage (ECNC) consolidates fragmented genes while accounting for missing domains in homologs. **A)** A single, contiguous gene homolog in a target genome with 100% query length coverage has an ECNC of 1.0. **B)** Two contiguous gene homologs, each with 100% query length coverage have an ECNC of 2.0. **C)** A single gene homolog, split across multiple scaffolds and contigs in a fragmented target genome; BLAT identifies each fragment as a single hit. Per nucleotide of query sequence, there is only one corresponding nucleotide over all the hits, thus the ECNC is 1.0. **D)** Two gene homologs, one fragmented and one contiguous. 100% of nucleotides in the query sequence are represented between all hits; however, every nucleotide in the query has two matching nucleotides in the target genome, thus the ECNC is 2.0. **E)** One true gene homolog in the target genome, plus multiple hits of a conserved domain that span 20% of the query sequence. While 100% of the query sequence is represented in total, 20% of the nucleotides have 4 hits. Thus, the ECNC for this gene is 1.45. **F)** Two real gene homologs; one hit is contiguous, one hit is fragmented in two, and the tail end of both sequences was not identified by BLAT due to sequence divergence. Only 75% of the query sequence was covered in total between the hits, but for that 75%, each nucleotide has two hits. As such, ECNC is equal to 2.0 for this gene.

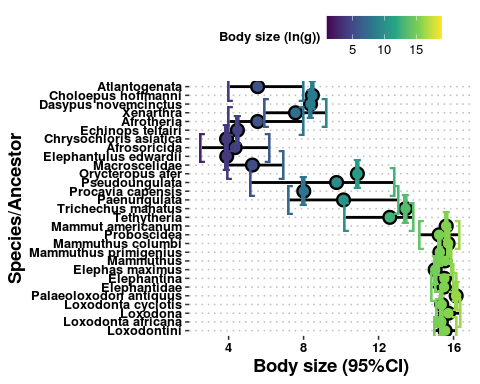


Supplementary Figure 2: Gene copy increases polarized along Atlantogenata, colored by ln(Body Size), with branch lengths equal to the change in gene copy number

## Warning: TODO: Make HQ version of this figure

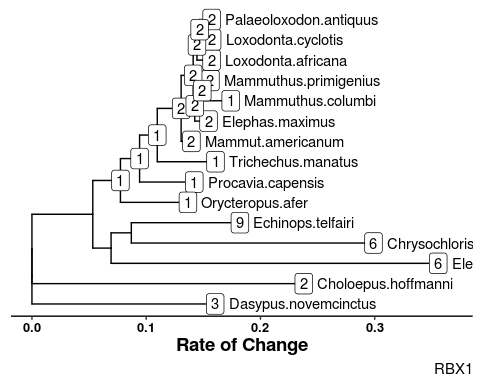


Supplementary Figure 3: Full version of RecBlat strategy, but low quality



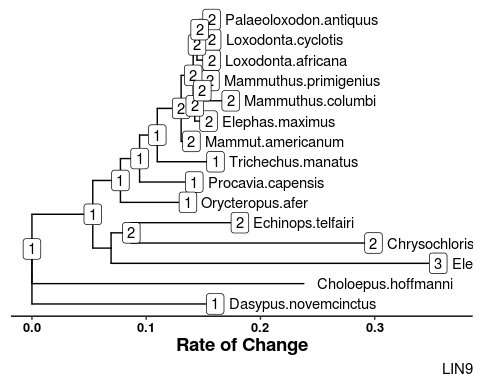
Supplementary Figure 4: Dot and Line plot for Body Sizes and such

## $RBX1



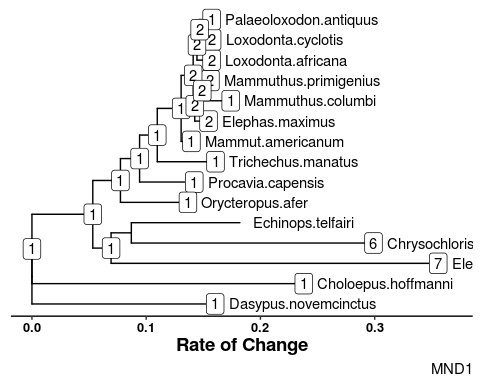
Supplementary Figure 5: All the Gene Copy Trees for interesting genes duplicated in LoxAfr4 (RIP any trees if this gets printed)

##   
## $LIN9



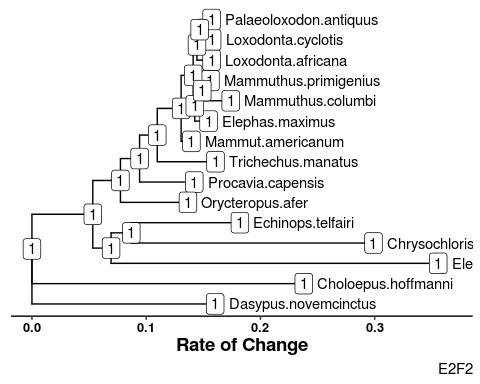
Supplementary Figure 5: All the Gene Copy Trees for interesting genes duplicated in LoxAfr4 (RIP any trees if this gets printed)

##   
## $MND1



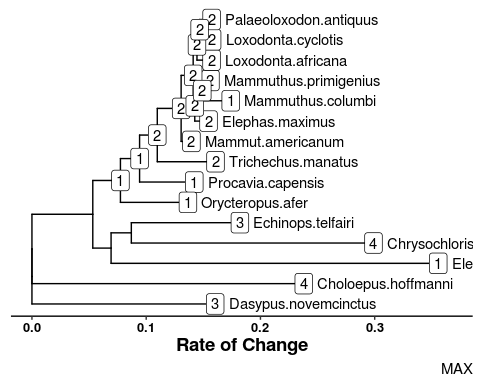
Supplementary Figure 5: All the Gene Copy Trees for interesting genes duplicated in LoxAfr4 (RIP any trees if this gets printed)

##   
## $E2F2



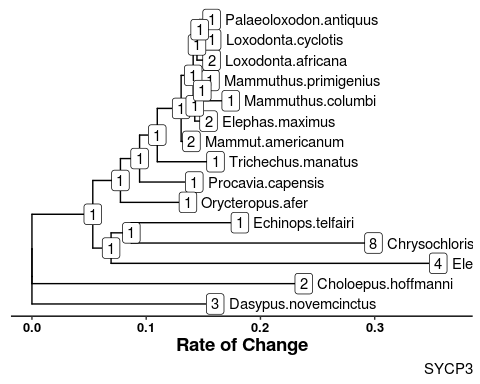
Supplementary Figure 5: All the Gene Copy Trees for interesting genes duplicated in LoxAfr4 (RIP any trees if this gets printed)

##   
## $MAX



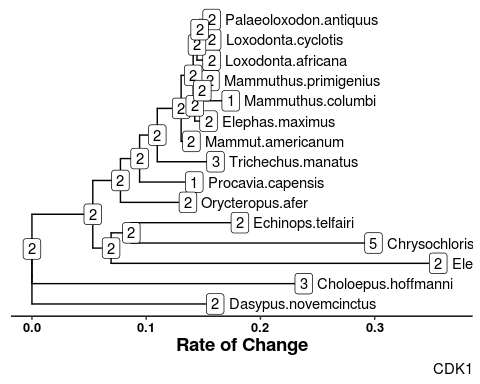
Supplementary Figure 5: All the Gene Copy Trees for interesting genes duplicated in LoxAfr4 (RIP any trees if this gets printed)

##   
## $SYCP3



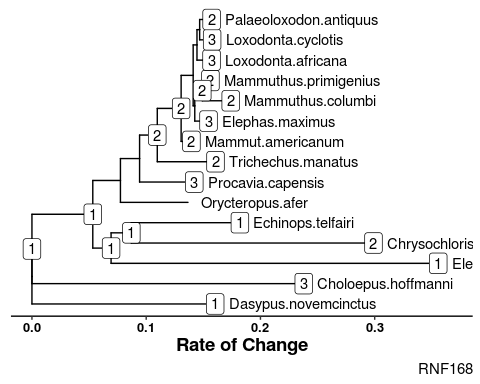
Supplementary Figure 5: All the Gene Copy Trees for interesting genes duplicated in LoxAfr4 (RIP any trees if this gets printed)

##   
## $CDK1



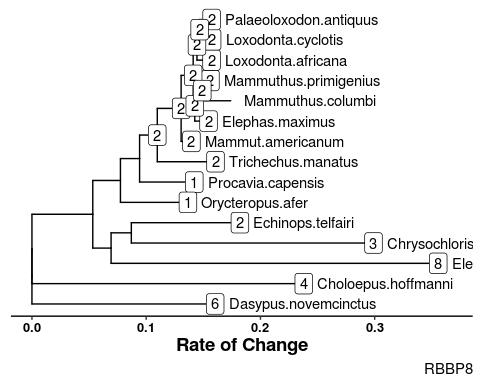
Supplementary Figure 5: All the Gene Copy Trees for interesting genes duplicated in LoxAfr4 (RIP any trees if this gets printed)

##   
## $RNF168



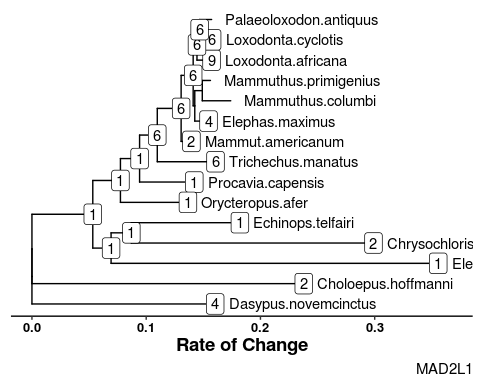
Supplementary Figure 5: All the Gene Copy Trees for interesting genes duplicated in LoxAfr4 (RIP any trees if this gets printed)

##   
## $RBBP8

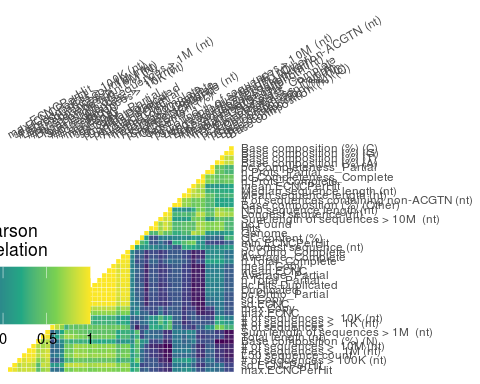


Supplementary Figure 5: All the Gene Copy Trees for interesting genes duplicated in LoxAfr4 (RIP any trees if this gets printed)

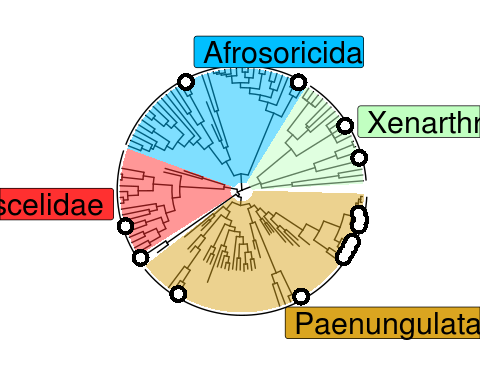
##   
## $MAD2L1



Supplementary Figure 5: All the Gene Copy Trees for interesting genes duplicated in LoxAfr4 (RIP any trees if this gets printed)



Supplementary Figure 7: Correlation matrix heat map for genome quality metrics, ECNC, RBHB Copy, and Estimated Copy Number (lesser between ECNC and RBHB)



Supplementary Figure 8: The time-calibrated Eutherian tree used as input for StableTraits

# Supplementary data

## Warning: TODO: Read in Eutheria.progress and present it to the bold.

# References

1. Green J, Cairns BJ, Casabonne D, Wright FL, Reeves G, Beral V, et al. Height and cancer incidence in the Million Women Study: prospective cohort, and meta-analysis of prospective studies of height and total cancer risk. The Lancet Oncology. 2011;12: 785–794. doi:[10.1016/s1470-2045(11)70154-1](https://doi.org/10.1016/s1470-2045(11)70154-1)

2. Nunney L. Size matters: height, cell number and a person’s risk of cancer. Proc R Soc B. 2018;285: 20181743. doi:[10.1098/rspb.2018.1743](https://doi.org/10.1098/rspb.2018.1743)

3. Dobson JM. Breed-predispositions to cancer in pedigree dogs. ISRN veterinary science. 2013;2013: 941275. doi:[10.1155/2013/941275](https://doi.org/10.1155/2013/941275)

4. Caulin AF, Maley CC. Peto’s Paradox: evolution’s prescription for cancer prevention. Trends in ecology & evolution. 2011;26: 175–82. doi:[10.1016/j.tree.2011.01.002](https://doi.org/10.1016/j.tree.2011.01.002)

5. Leroi AM, Koufopanou V, Burt A. Cancer selection. Nature Reviews Cancer. 2003;3: 226–231. doi:[10.1038/nrc1016](https://doi.org/10.1038/nrc1016)

6. Peto R, Roe F, Lee P, Levy L, Clack J. Cancer and ageing in mice and men. British Journal of Cancer. 1975;32: 411–426. doi:[10.1038/bjc.1975.242](https://doi.org/10.1038/bjc.1975.242)

7. Ashur-Fabian O, Avivi A, Trakhtenbrot L, Adamsky K, Cohen M, Kajakaro G, et al. Evolution of p53 in hypoxia-stressed Spalax mimics human tumor mutation. Proceedings of the National Academy of Sciences. 2004;101: 12236–12241. doi:[10.1073/pnas.0404998101](https://doi.org/10.1073/pnas.0404998101)

8. Seluanov A, Hine C, Bozzella M, Hall A, Sasahara THC, Ribeiro AACM, et al. Distinct tumor suppressor mechanisms evolve in rodent species that differ in size and lifespan. Aging cell. 2008;7: 813–23. doi:[10.1111/j.1474-9726.2008.00431.x](https://doi.org/10.1111/j.1474-9726.2008.00431.x)

9. Gorbunova V, Hine C, Tian X, Ablaeva J, Gudkov AV, Nevo E, et al. Cancer resistance in the blind mole rat is mediated by concerted necrotic cell death mechanism. Proceedings of the National Academy of Sciences of the United States of America. 2012;109: 19392–6. doi:[10.1073/pnas.1217211109](https://doi.org/10.1073/pnas.1217211109)

10. Tian X, Azpurua J, Hine C, Vaidya A, Myakishev-Rempel M, Ablaeva J, et al. High molecular weight hyaluronan mediates the cancer resistance of the naked mole-rat. 2013;499. doi:[10.1038/nature12234](https://doi.org/10.1038/nature12234)

11. Sulak M, Fong L, Mika K, Chigurupati S, Yon L, Mongan NP, et al. TP53 copy number expansion is associated with the evolution of increased body size and an enhanced DNA damage response in elephants. eLife. 2016;5: e11994. doi:[10.7554/elife.11994](https://doi.org/10.7554/elife.11994)

12. Tacutu R, Craig T, Budovsky A, Wuttke D, Lehmann G, Taranukha D, et al. Human Ageing Genomic Resources: Integrated databases and tools for the biology and genetics of ageing. Nucleic Acids Research. 2013;41: D1027–D1033. doi:[10.1093/nar/gks1155](https://doi.org/10.1093/nar/gks1155)

13. Schwartz GT, Rasmussen DT, Smith RJ. Body-Size Diversity and Community Structure of Fossil Hyracoids. Journal of Mammalogy. 1995;76: 1088–1099. doi:[10.2307/1382601](https://doi.org/10.2307/1382601)

14. Scheffer VB. The Weight of the Steller Sea Cow. Journal of Mammalogy. 1972;53: 912–914. doi:[10.2307/1379236](https://doi.org/10.2307/1379236)

15. Larramendi A. Shoulder Height, Body Mass, and Shape of Proboscideans. Acta Palaeontologica Polonica. 2015;61. doi:[10.4202/app.00136.2014](https://doi.org/10.4202/app.00136.2014)

16. O’Leary MA, Bloch JI, Flynn JJ, Gaudin TJ, Giallombardo A, Giannini NP, et al. The placental mammal ancestor and the post-K-Pg radiation of placentals. Science (New York, NY). 2013;339: 662–7. doi:[10.1126/science.1229237](https://doi.org/10.1126/science.1229237)

17. Springer MS, Meredith RW, Teeling EC, Murphy WJ. Technical comment on "The placental mammal ancestor and the post-K-Pg radiation of placentals". Science (New York, NY). 2013;341: 613. doi:[10.1126/science.1238025](https://doi.org/10.1126/science.1238025)

18. O’Leary MA, Bloch JI, Flynn JJ, Gaudin TJ, Giallombardo A, Giannini NP, et al. Response to comment on "The placental mammal ancestor and the post-K-Pg radiation of placentals". Science (New York, NY). 2013;341: 613. doi:[10.1126/science.1238162](https://doi.org/10.1126/science.1238162)

19. Puttick MN, Thomas GH. Fossils and living taxa agree on patterns of body mass evolution: a case study with Afrotheria. Proceedings Biological sciences / The Royal Society. 2015;282: 20152023. doi:[10.1098/rspb.2015.2023](https://doi.org/10.1098/rspb.2015.2023)

20. Abegglen LM, Caulin AF, Chan A, Lee K, Robinson R, Campbell MS, et al. Potential Mechanisms for Cancer Resistance in Elephants and Comparative Cellular Response to DNA Damage in Humans. JAMA. 2015;314: 1850–1860. doi:[10.1001/jama.2015.13134](https://doi.org/10.1001/jama.2015.13134)

21. Vazquez JM, Sulak M, Chigurupati S, Lynch VJ. A Zombie LIF Gene in Elephants Is Upregulated by TP53 to Induce Apoptosis in Response to DNA Damage. Cell Reports. 2018;24: 1765–1776. doi:[10.1016/j.celrep.2018.07.042](https://doi.org/10.1016/j.celrep.2018.07.042)

22. Caulin AF, Graham TA, Wang L-S, Maley CC. Solutions to Peto’s paradox revealed by mathematical modelling and cross-species cancer gene analysis. Philosophical transactions of the Royal Society of London Series B, Biological sciences. 2015;370: 20140222. doi:[10.1098/rstb.2014.0222](https://doi.org/10.1098/rstb.2014.0222)

23. Doherty A, Magalhães J de. Has gene duplication impacted the evolution of Eutherian longevity? Aging Cell. 2016;15: 978–980. doi:[10.1111/acel.12503](https://doi.org/10.1111/acel.12503)

24. Bininda-Emonds ORP, Cardillo M, Jones KE, MacPhee RDE, Beck RMD, Grenyer R, et al. Erratum: The delayed rise of present-day mammals. Nature. 2008;456: 274–274. doi:[10.1038/nature07347](https://doi.org/10.1038/nature07347)

25. Elliot MG, Mooers AØ. Inferring ancestral states without assuming neutrality or gradualism using a stable model of continuous character evolution. BMC evolutionary biology. 2014;14: 226. doi:[10.1186/s12862-014-0226-8](https://doi.org/10.1186/s12862-014-0226-8)

26. Kent JW. BLAT—The BLAST-Like Alignment Tool. Genome Research. 2002;12: 656–664. doi:[10.1101/gr.229202](https://doi.org/10.1101/gr.229202)

27. Altenhoff AM, Dessimoz C. Phylogenetic and functional assessment of orthologs inference projects and methods. PLoS computational biology. 2009;5: e1000262. doi:[10.1371/journal.pcbi.1000262](https://doi.org/10.1371/journal.pcbi.1000262)

28. Consortium TU. UniProt: the universal protein knowledgebase. Nucleic Acids Research. 2017;45: D158–D169. doi:[10.1093/nar/gkw1099](https://doi.org/10.1093/nar/gkw1099)