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

Juan Manuel Vazquez[a], Vincent J Lynch[b]

[a]*Department of Human Genetics, 920 East 58th St, Chicago, IL, 60637*
[b]*Department of Biological Sciences, 551 Cooke Hall, Buffalo NY, 14260*

Abstract

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 [3,4]. 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' [5–7]. While it is clear that a resolution to Peto's Paradox must involve the evolution of enhanced cancer

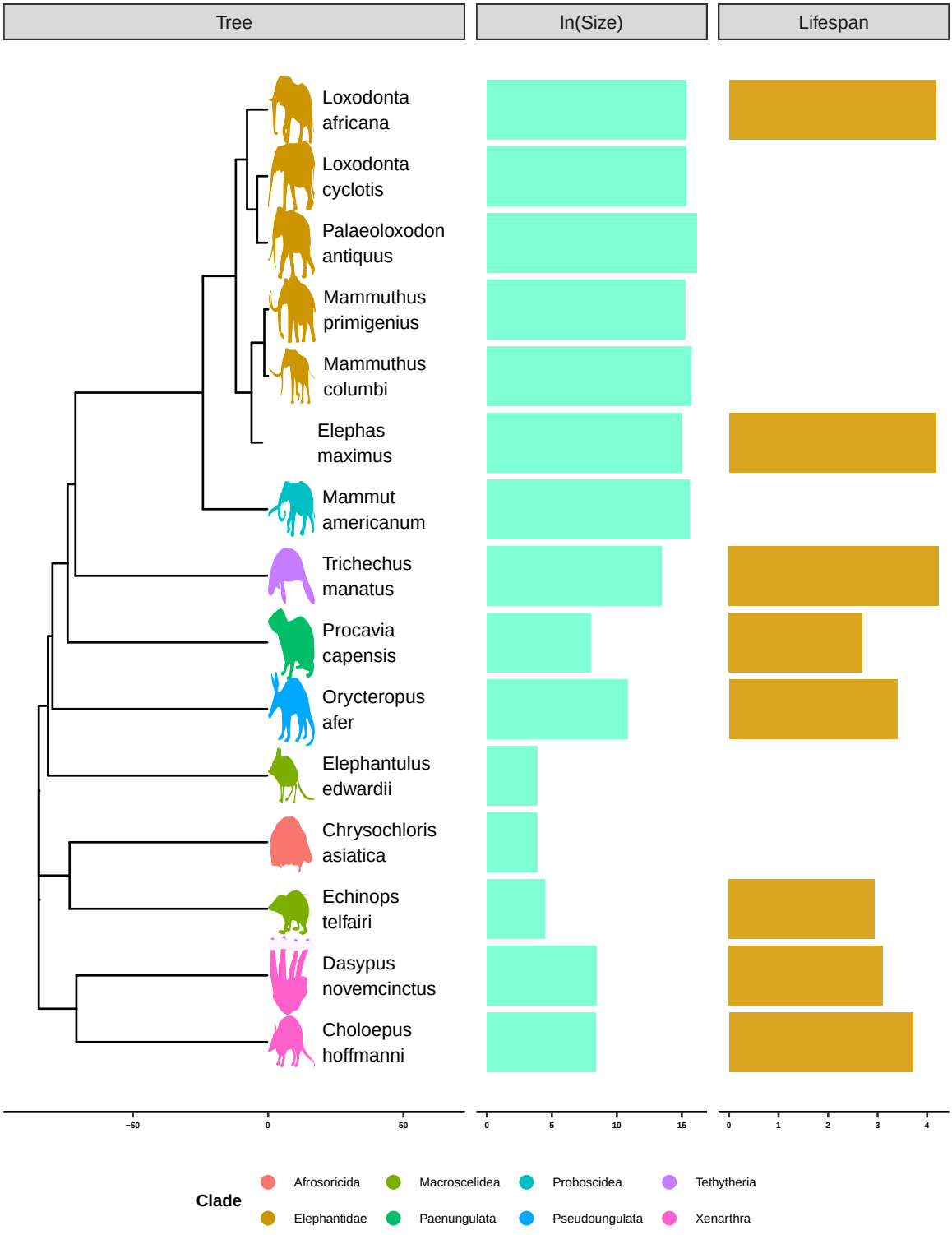

Figure 1: Atlantogenatans with sequenced genomes, body sizes, and known lifespans [@PuttickAndThomas2015; @HAGR]

protection alongside increases in body size and lifespan, the specific genetic, molecular, and cellular mechanisms that underlie this resistance have proven elusive. [8–12].

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 [13]. 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 [14]. 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 [15]. 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 [16]. These large-bodied Afrotherian lineages are nested within small bodied species (Fig. 1) [17–20], 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, among the most parsimonious routes to enhanced cancer resistance is 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* [12,21,22] as well as in studies involving a limited set of candidate genes [23,24]. 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 [17–20], 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 maximum likelihood (treating copy number as a discrete trait) 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.* [25] with the time-calibrated total evidence Afrotherian phylogeny from Puttick and Thomas [20]. While the Bininda-Emonds *et al.* [25] 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 [20] includes 77 extant species and fossil data from 39 extinct species. Therefore we replaced the Afrotherian clade in the Bininda-Emonds *et al.* [25] phylogeny with the Afrotherian phylogeny of Puttick and Thomas [20] 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") [26]. 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 [26].

Identification of Duplicate Genes

Reciprocal Best-Hit BLAT: We developed a reciprocal best hit BLAT (RBHB) pipeline to identify putative homologs and estimate gene copy numbers across species (**Figure 1A**). The Reciprocal Best Hit (RBH) search strategy is conceptually straightforward: 1) Given a gene of interest G_A in a query genome A, one searches a target genome B for all possible matches to G_A; 2) For each of these hits, one then performs the reciprocal search in the original query genome to identify the highest-scoring hit; 3) A hit in genome B is defined as a homolog of gene G_A if and only if the original gene G_A is the top reciprocal search hit in genome A. We selected BLAT [27] 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 [28,29].

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:

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

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

RecSearch Pipeline: We created a custom Python pipeline for automating RBHB searches between a single reference genome and multiple target genomes using a list of query sequences from the reference genome. For the query sequences in our search, we used the hg38 Proteome provided by UniProt [30], which is a comprehensive set of protein sequences curated from a combination of predicted and validated protein sequences generated by the UniProt Consortium. In order to refine our search, we omitted protein sequences originating from long, noncoding RNA loci (e.g. LINC genes); poorly-studied genes from predicted open reading frames (C-ORFs); and sequences with highly repetitive sequences such as zinc fingers, protocadherins, and transposon-containing genes, as these were prone to high levels of false positive hits.

After filtering out problematic protein queries (see "Query gene inclusion criteria"), 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 gene inclusion criteria: To assemble our query list, we first removed all unnamed genes from UP000005640. Next, we excluded genes from downstream analyses for which assignment of homology was uncertain, including uncharacterized ORFs (991), 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 [12,22], 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 [31,32]. 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 duplicate gene calls, we intersected our results with either gene prediction or transcriptomic evidence as a proxy for functionality.

Transcriptome Assembly: For the African Savana Elephant, Asian Elephant, West Indian Manatee, and Nine-Banded Armadillo, we generated *de novo* transcriptomes using publically-available RNA-sequencing data from NCBI SRA (Supplementary Table 1). We mapped reads to all genomes available for each species, and assembled transcripts using HISAT2 and StringTie, respectively [33–35]. 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 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. The most likely model for the evolution of copy number was inferred to be a Jukes-Cantor type model for morphological data with equal state frequencies, allowing for a proportion of invariable sites, and using a discrete Gamma model with 4 rate categories. ("MK+FQ+I+G4"). 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 [36]. For the database of pathways used in the analysis, we used

Reactome [37], Wikipathways, [38], and KEGG [39]. 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 K 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) [40,41] 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 D and lifespan t: $K \approx Dt^6$ [7,42–44]. To determine the change in K between nodes, we obtained the ratio between the cancer risk K_1 at any given node, and the cancer risk K_2 at its ancestral node, using the equation $\frac{K_2}{K_1} \approx \frac{D_1 t_1^6}{D_2 t_2^6}$. Finally, to simplify comparisons, we calculated the fold change cancer risk between a node and its ancestor as $\log_2(\frac{K_2}{K_1})$.

Results

Body size frequently and independently expands and contracts throughout Atlantogenata

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 [25] with a total evidence Afrotherian phylogeny including 77 extant and fossil data from 39 extinct species [20]. 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 [26].

Similar to previous studies of Afrotherian body size [20,26], 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 in body mass in the stem-lineage of *Proboscidea* (4114.39kg) (Figure 1B,C). The ancestral *Hyracoidea* was inferred to be relatively small 2.86kg-118.18kg, and rate accelerations were coincident with independent body mass increases in large hyraxes such as *Titanohyrax andrewsi* 429.34kg; 67.36x increase. While the body mass of the ancestral *Sirenian* was inferred to be large 61.7kg-955.51kg, a rate acceleration occurred coincident with a 10.59x increase in body mass in Stellar's sea cow. Rate accelerations also occurred coincident with 36.6x decrease in body mass in the stem-lineage of the dwarf elephants *Elephas (Palaeoloxodon) antiquus falconeri* and *Elephas cypriotes*. These data suggest that gigantism in *Afrotherians* evolved step-wise, from small to medium bodies in the *Pseudoungulata* stem-lineage, medium to large bodies in the *Tehthytherian* stem-lineage and extinct hyraxes, and from large to exceptionally large bodies independently in the *Proboscidean* stem-lineage and Stellar's sea cow (Figure 1).

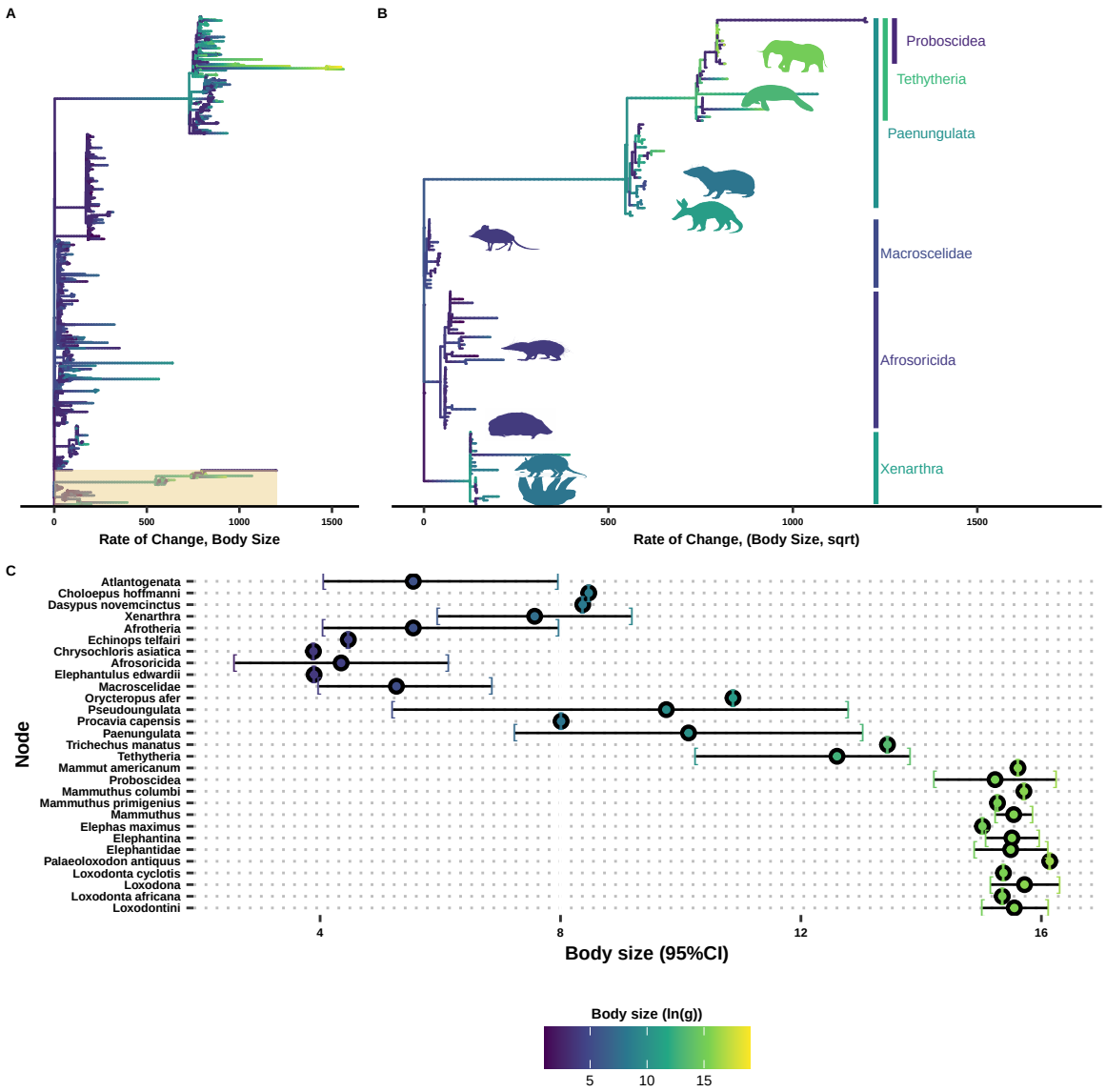

Figure 2: Body sizes rapidly and frequently expand in Eutherians, especially in Atlantogenata. **A)** Tree of Eutherian species, colored by ln(Body Size) and with branch lengths set to the rate of change in body sizes, normalized by the square root of the root branch. Atlantogenata is highlighted at the bottom. **B)** Zoom-in of (**A**) on Atlantogenata. Silhuetes for the African Elephant, West Indian Manatee, Cape Elephant Shrew, Lesser Hedgehog Tenrec, Cape Golden Mole, Nine-Banded Armadillo, and Hoffman's Two-Toed Sloth are colored by their extant body sizes, while clade labels are colored based on the common ancestor's estimated body size. **C)** Confidence interval plot for representative species and ancestral nodes.

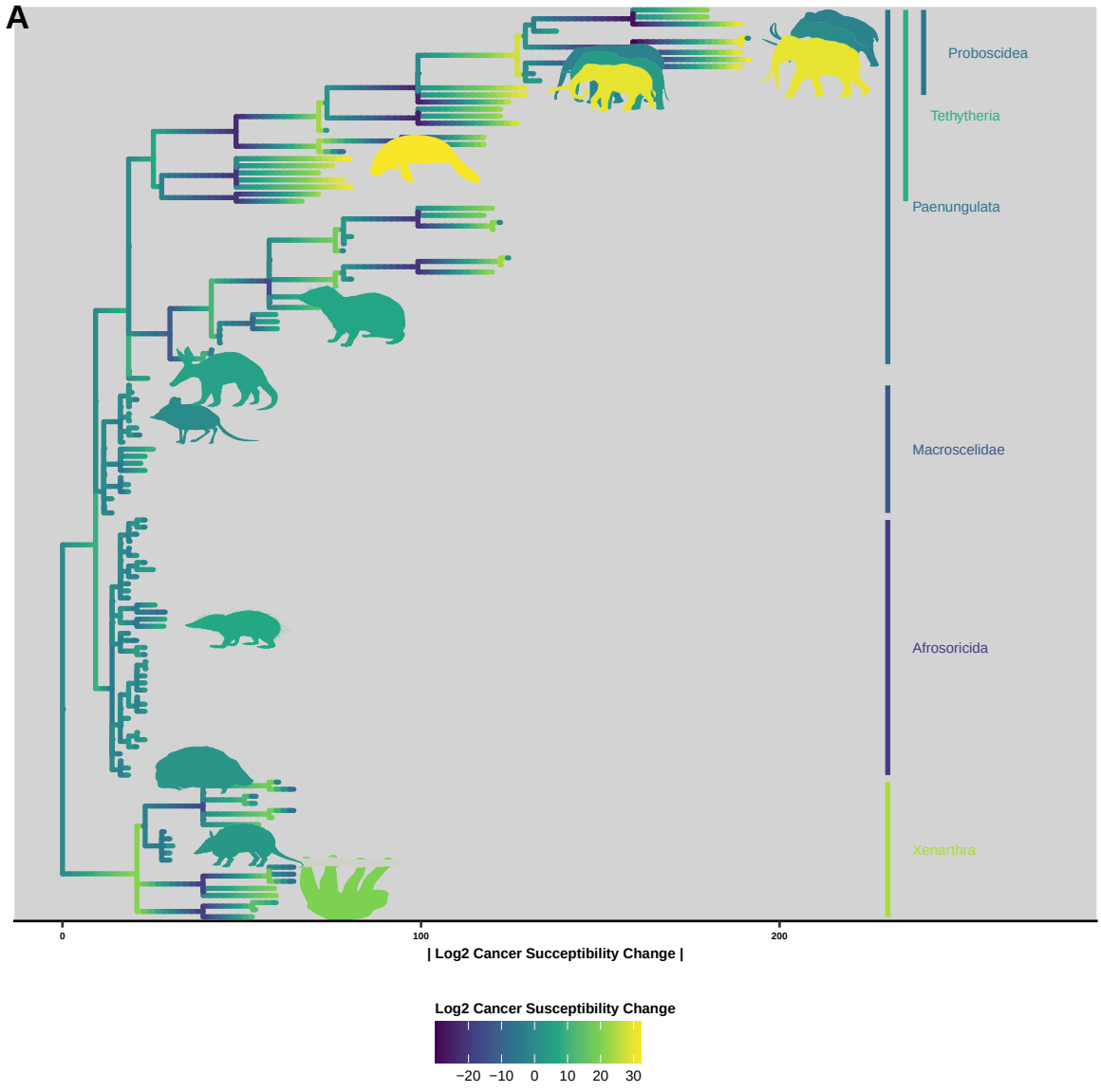

Figure 3: Cancer succecptibility across Atlantogenata

Table 1: Estimated Cancer Succeptibilty for nodes in Atlantogenata

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

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

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 [44] we estimate the intrinsic cancer risk as the product of risk associated with body mass and lifespan. Differences in cancer susceptibility K due to body mass differences between species can be approximated simply as the fold difference in body mass (D) between species [44]. The risk of developing cancer also increases in proportion to the sixth power of age and is approximated by the formula Ct^6, in which the proportionality constant C that determines susceptibility to cancer induction is multiplied by the sixth power of the age in years, t [44–46]. Thus we can estimate the intrinsic cancer risk for a species as $K \approx Dt^6$.

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 $ln(lifespan) = \beta_1 corBrownian + \beta_2 ln(Size) + \epsilon$ (Table 2). With this information in hand, we calculated K_1 at all nodes, and then estimated the fold change in cancer susceptibility between an ancestral node and a given node as $\frac{K_2}{K_1}$ (**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 2**, cancer susceptibility skyrocketed at the initial divergence of Atlantogenata, followed by a generally upwards trend. At the common ancestor of Afrotheria there is an initial 9.22-fold increase in cancer risk. In parallel to Afrotheria, cancer susceptibility 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 susceptibility 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 susceptibility, and decreases once more at the Cape Elephant Shrew (-4.61-fold). The emergence of Pseudoungulata sees the next big leap in cancer susceptibility 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 susceptibility 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 susceptibility 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 susceptibility. 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 susceptibility (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 susceptibility.

Identification and evolutionary history of gene duplications

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

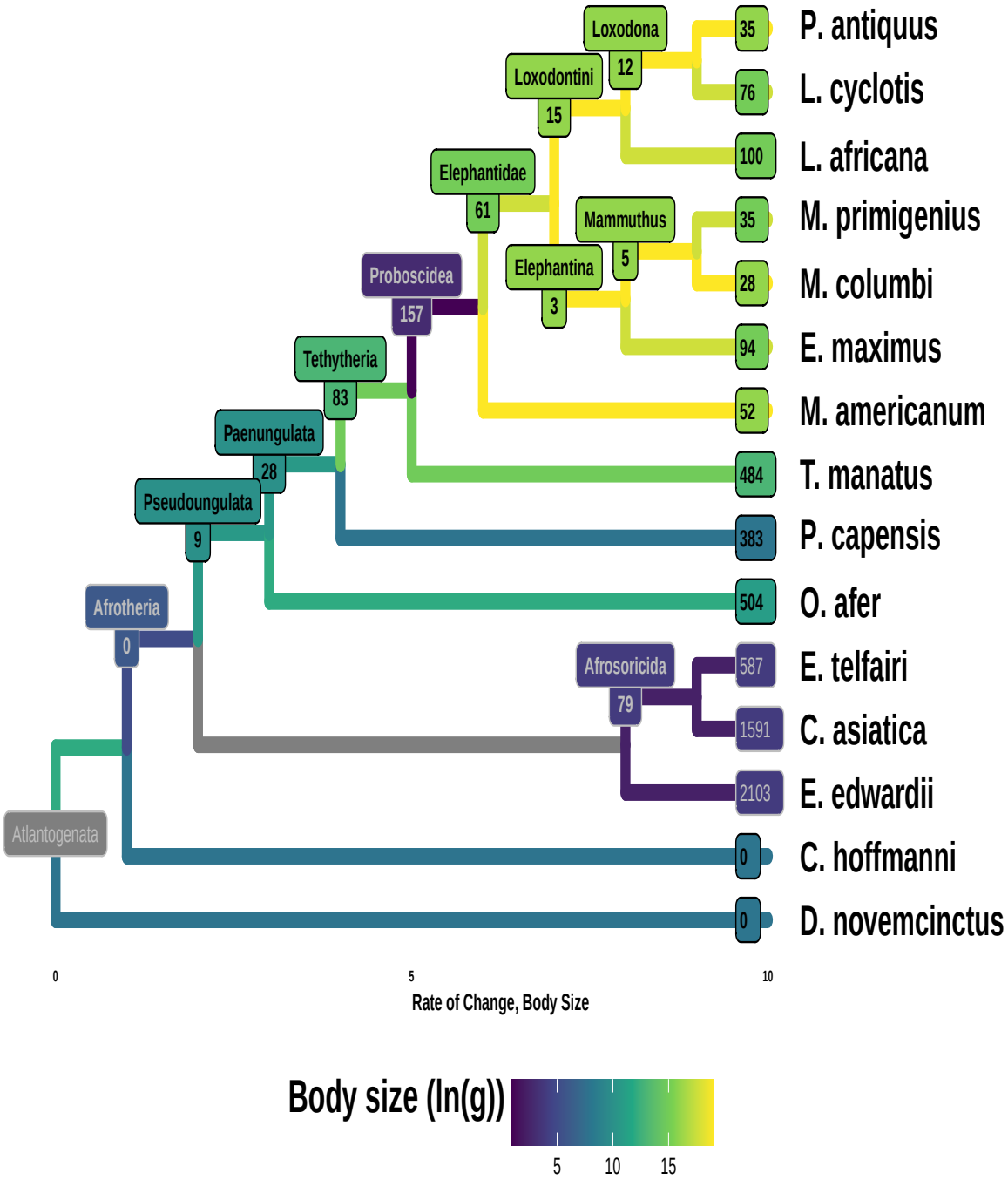

Figure 4: Gene duplications occur readily throughout Atlantogenata. A) A table summarizing duplication events in Atlantogenata. B) A tree of the species in Atlantogenata with genomes, with the number of genes that underwent an increase in copy number overlayed at each node.

Table 2: Summary of duplications in Atlantogenata

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

studies focusing on candidate gene studies, for example, have identified increased copy number of the tumor suppressors *TP53* and *LIF* in elephants [12,21–24]. Therefore, in order to test whether this was pervasive 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.

As our sequence database included various protein transcripts for each gene, in order to obtain gene-level copy number information and eliminate 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 in order to account for partial hits that may be due to incomplete coverage.

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 [31,32] 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

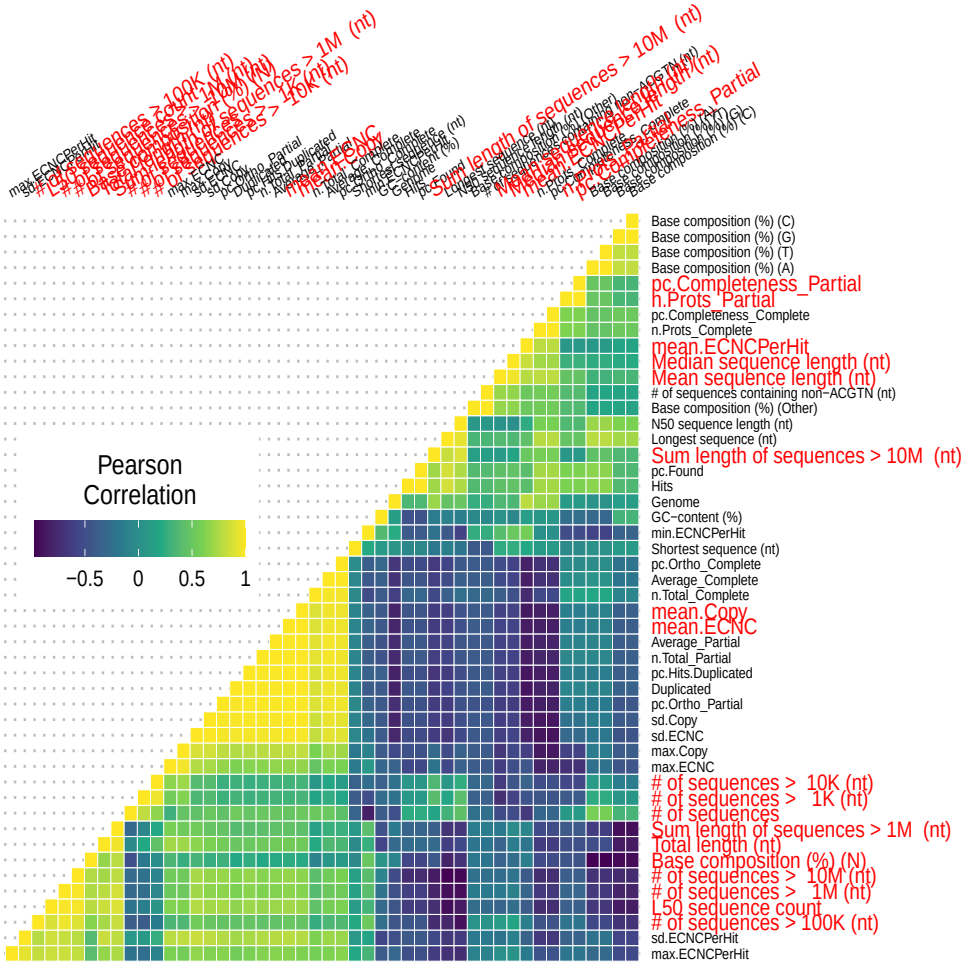

Figure 5: Correlations between genome quality metrics and ECNC metrics

Table 3: Number of pathways overrepresented among duplicated genes at different FDRs.

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

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 occurred, 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 Figure 2B.

Gene duplications events occured readily thoughout *Atlantogenata* (Figure 2B). Among the genes that increased in copy number in the elephant lineage are TP53 and LIF, as previously described; however, we find that these two genes represent a fraction of the 940 genes that are duplicated in the African Elephant overall, which accumulated over various steps through their evolution. While the extinct elephantids have acceptable genome quality metrics according to CEGMA, they are nonetheless missing a significant number of sequences; this may contribute to the low number of duplicated genes that occured in internal nodes. The number of duplicates that occur at each branch is also proportional to the density of the sampling of the clade overall, as would be expected. In branches, such as *Afrosoricida*, where the number of species is relatively minuscule compared to the size of the clade, we see many significantly larger numbers of duplications private to these species.

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

Our initial hypothesis was that genes which duplicated in lineages that experienced a growth in size would be enriched for membership in tumor suppressor pathways. Thus, we used WebGestalt and its Overrepresentation Analysis (ORA) functionality to determine what pathways were enriched in our duplicated gene sets in each branch relative to our initial query set. For our database, we used Reactome for our primary analysis, but additionally used the KEGG, Panther, Wikipathways, and Wikipathways_cancer databases using WebGestalt ORA. Going through the tree, at no FDR<0.5 is there any significant pathway representation for genes that increased in the branches leading to , or *Paenungulata*; furthermore, there are no significant pathway enrichments at FDR<0.5 for genes whose copy number did not change between branches (copy-number-stable) for , 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

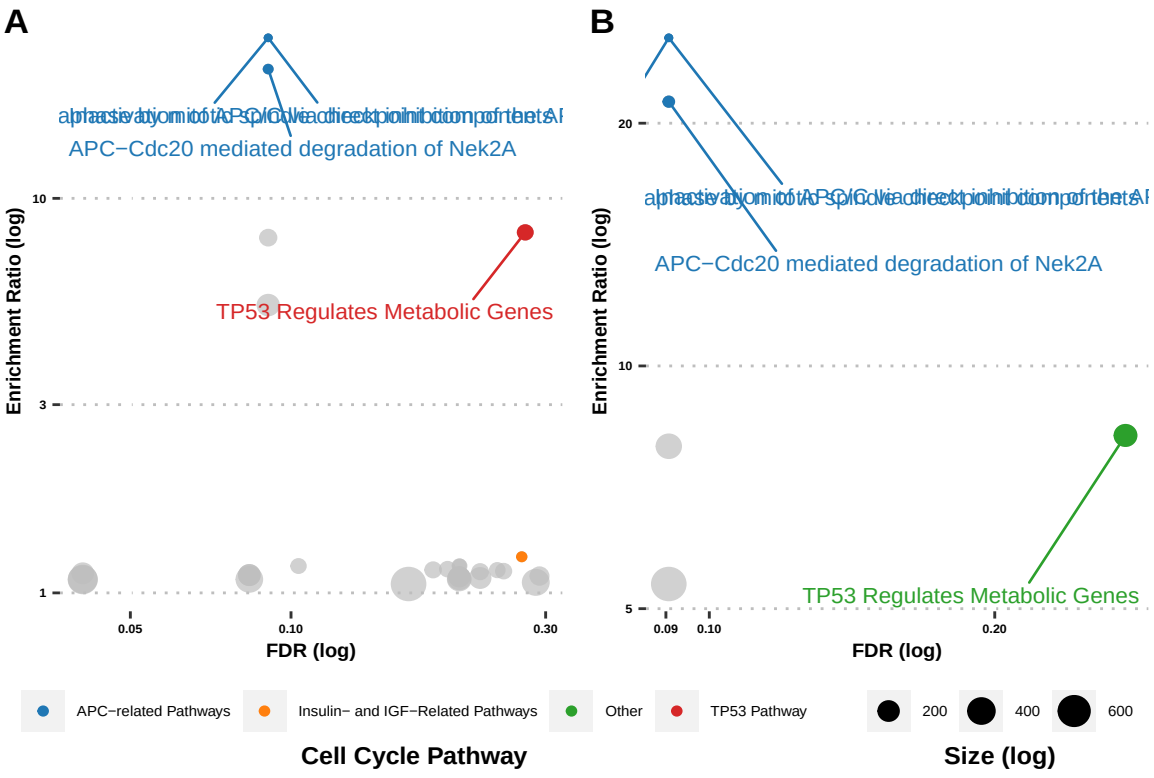

Figure 6: Overrepresentation Analysis of Duplicated Genes in Atlantogenata using Reactome Pathways.

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

The pathway enrichments for genes whose copy number either did not change, or whose copy number increased, between *Loxodona* and the African Elephant are shown in Figure 3B and 3C, respectively. 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 [47]. 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

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

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

3. Dobson JM. Breed-predispositions to cancer in pedigree dogs. ISRN veterinary science. 2013;2013: 941275. doi:10.1155/2013/941275

4. Dorn CR, Taylor DON, Schneider R, Hibbard HH, Klauber MR. Survey of Animal Neoplasms in Alameda and Contra Costa Counties, California. II. Cancer Morbidity in Dogs and Cats From Alameda County<xref ref-type="fn" rid="FN2">2</xref>. JNCI: Journal of the National Cancer Institute. 1968;40: 307–318. doi:10.1093/jnci/40.2.307

5. 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

6. Leroi AM, Koufopanou V, Burt A. Cancer selection. Nature Reviews Cancer. 2003;3: 226–231. doi:10.1038/nrc1016

7. 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

8. 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

9. 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

10. 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

11. 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

12. 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

13. 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

14. 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

15. Scheffer VB. The Weight of the Steller Sea Cow. Journal of Mammalogy. 1972;53: 912–914. doi:10.2307/1379236

16. Larramendi A. Shoulder Height, Body Mass, and Shape of Proboscideans. Acta Palaeontologica Polonica. 2015;61. doi:10.4202/app.00136.2014

17. 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

18. 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

19. 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

20. 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

21. 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

22. 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

23. 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

24. 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

25. 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

26. 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

27. Kent JW. BLAT?The BLAST-Like Alignment Tool. Genome Research. 2002;12: 656–664. doi:10.1101/gr.229202

28. 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

29. Salichos L, Rokas A. Evaluating ortholog prediction algorithms in a yeast model clade. PloS one. 2011;6: e18755. doi:10.1371/journal.pone.0018755

30. Consortium TU. UniProt: the universal protein knowledgebase. Nucleic Acids Research. 2017;45: D158–D169. doi:10.1093/nar/gkw1099

31. Nishimura O, Hara Y, Kuraku S. gVolante for standardizing completeness assessment of genome and transcriptome assemblies. Bioinformatics. 2017;33: 3635–3637. doi:10.1093/bioinformatics/btx445

32. Parra G, Bradnam K, Ning Z, Keane T, Korf I. Assessing the gene space in draft genomes. Nucleic Acids Research. 2008;37: 289–297. doi:10.1093/nar/gkn916

33. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. Nature Methods. 2015;12: 357–360. doi:10.1038/nmeth.3317

34. Pertea M, Pertea GM, Antonescu CM, Chang T-C, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nature Biotechnology. 2015;33: 290–295. doi:10.1038/nbt.3122

35. Pertea M, Kim D, Pertea GM, Leek JT, Salzberg SL. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. Nature Protocols. 2016;11: 1650–1667. doi:10.1038/nprot.2016.095

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

37. Jassal B, Matthews L, Viteri G, Gong C, Lorente P, Fabregat A, et al. The reactome pathway knowledgebase. Nucleic acids research. 2020;48: D498–D503. doi:10.1093/nar/gkz1031

38. Slenter DN, Kutmon M, Hanspers K, Riutta A, Windsor J, Nunes N, et al. WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. Nucleic Acids Research. 2017;46: D661–D667. doi:10.1093/nar/gkx1064

39. Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Research. 2000;28: 27–30. doi:10.1093/nar/28.1.27

40. Felsenstein J. Phylogenies and the Comparative Method. The American Naturalist. 1985;125: 1–15. doi:10.1086/284325

41. Martins EP, Hansen TF. Phylogenies and the Comparative Method: A General Approach to Incorporating Phylogenetic Information into the Analysis of Interspecific Data. The American Naturalist. 1997;149: 646–667. doi:10.1086/286013

42. Armitage P. Multistage models of carcinogenesis. Environmental health perspectives. 1985;63: 195–201. doi:10.1289/ehp.8563195

43. Armitage P, Doll R. The age distribution of cancer and a multi-stage theory of carcinogenesis. British Journal of Cancer. 2004;91: 6602297. doi:10.1038/sj.bjc.6602297

44. Peto R. Quantitative implications of the approximate irrelevance of mammalian body size and lifespan to lifelong cancer risk. Phil Trans R Soc B. 2015;370: 20150198. doi:10.1098/rstb.2015.0198

45. Armitage P, Doll R. The Age Distribution of Cancer and a Multi-stage Theory of Carcinogenesis. British Journal of Cancer. 1954;8: 1–12. doi:10.1038/bjc.1954.1

46. Nordling CO. A New Theory on the Cancer-inducing Mechanism. British Journal of Cancer. 1953;7: 68–72. doi:10.1038/bjc.1953.8

47. Goodman M, Sterner KN, Islam M, Uddin M, Sherwood CC, Hof PR, et al. Phylogenomic analyses reveal convergent patterns of adaptive evolution in elephant and human ancestries. Proceedings of the National Academy of Sciences. 2009;106: 20824–20829. doi:10.1073/pnas.0911239106